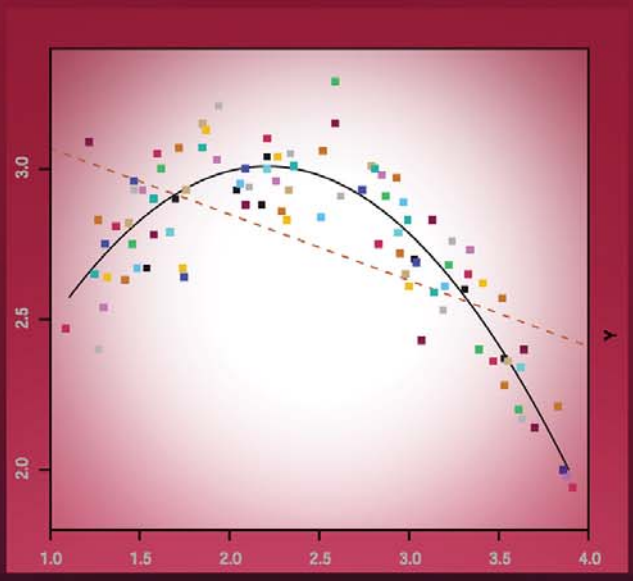
REGRESSION ANALYSIS and LINEAR MODELS

Concepts, Applications, and Implementation



Richard B. Darlington | Andrew F. Hayes

ebook

THE GUILFORD PRESS

Regression Analysis and Linear Models

Methodology in the Social Sciences

David A. Kenny, Founding Editor

Todd D. Little, Series Editor

www.guilford.com/MSS

This series provides applied researchers and students with analysis and research design books that emphasize the use of methods to answer research questions. Rather than emphasizing statistical theory, each volume in the series illustrates when a technique should (and should not) be used and how the output from available software programs should (and should not) be interpreted. Common pitfalls as well as areas of further development are clearly articulated.

RECENT VOLUMES

DOING STATISTICAL MEDIATION AND MODERATION
Paul E. Jose

LONGITUDINAL STRUCTURAL EQUATION MODELING
Todd D. Little

INTRODUCTION TO MEDIATION, MODERATION, AND CONDITIONAL PROCESS ANALYSIS: A REGRESSION-BASED APPROACH
Andrew F. Hayes

BAYESIAN STATISTICS FOR THE SOCIAL SCIENCES
David Kaplan

CONFIRMATORY FACTOR ANALYSIS FOR APPLIED RESEARCH, SECOND EDITION
Timothy A. Brown

PRINCIPLES AND PRACTICE OF STRUCTURAL EQUATION MODELING, FOURTH EDITION
Rex B. Kline

HYPOTHESIS TESTING AND MODEL SELECTION IN THE SOCIAL SCIENCES
David L. Weakliem

REGRESSION ANALYSIS AND LINEAR MODELS: CONCEPTS, APPLICATIONS, AND IMPLEMENTATION
Richard B. Darlington and Andrew F. Hayes

GROWTH MODELING: STRUCTURAL EQUATION AND MULTILEVEL MODELING APPROACHES
Kevin J. Grimm, Nilam Ram, and Ryne Estabrook

PSYCHOMETRIC METHODS: THEORY INTO PRACTICE
Larry R. Price

Regression Analysis and Linear Models

Concepts, Applications, and Implementation

Richard B. Darlington
Andrew F. Hayes

Series Editor's Note by Todd D. Little

THE GUILFORD PRESS
New York London

A Division of Guilford Publications, Inc.
370 Seventh Avenue, Suite 1200, New York, NY 10001
www.guilford.com

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data is available from the publisher.

ISBN 978-1-4625-2113-5 (hardcover)

Series Editor's Note

What a partnership: Darlington and Hayes. Richard Darlington is an icon of regression and linear modeling. His contributions to understanding the general linear model have educated social and behavioral science researchers for nearly half a century. Andrew Hayes is an icon of applied regression techniques, particularly in the context of mediation and moderation. His contributions to conditional process modeling have shaped how we think about and test processes of mediation and moderation. Bringing these two icons together in collaboration gives us a work that any researcher should use to learn and understand all aspects of linear modeling. The didactic elements are thorough, conversational, and highly accessible. You'll enjoy *Regression Analysis and Linear Models*, not as a statistics book but rather as a *Hitchhiker's Guide* to the world of linear modeling. Linear modeling is the bedrock material you need to know in order to grow into the more advanced procedures, such as multilevel regression, structural equation modeling, longitudinal modeling, and the like. The combination of clarity, easy-to-digest "bite-sized" chapters, and comprehensive breadth of coverage is just wonderful. And the software coverage is equally comprehensive, with examples in SAS, STATA, and SPSS (and some nice exposure to R)—giving every discipline's dominant software platform a thorough coverage. In addition to the software coverage, the various examples that are used span many disciplines and offer an engaging panorama of research questions and topics to stimulate the intellectually curious (a remedy for "academic attention deficit disorder").

This book is not just about linear regression as a technique, but also about research practice and the origins of scientific knowledge. The

thoughtful discussion of statistical control versus experimental control, for example, provides the basis to understand when causal conclusions are sufficiently implicated. As such, policy and practice can, in fact, rely on well-crafted nonexperimental analyses. Practical guidance is also a hallmark of this work, from detecting and managing irregularities, to collinearity issues, to probing interactions, and so on. I particularly appreciate that they take linear modeling all the way up through path analysis, an essential starting point for many advanced latent variable modeling procedures.

This book will be well worn, dog-eared, highlighted, shared, re-read, and simply cherished. It will now be required reading for all of my first-year students and a recommended primer for all of my courses. And if you are planning to come to one of my Stats Camp courses, brush up by reviewing Darlington and Hayes.

As always, "Enjoy!" Oh, and to paraphrase the catch phrase from the *Hitchhiker's Guide to the Galaxy*: "Don't forget your Darlington and Hayes."

TODD D. LITTLE
Kicking off my Stats Camp
in Albuquerque, New Mexico

Preface

Linear regression analysis is by far the most popular analytical method in the social and behavioral sciences, not to mention other fields like medicine and public health. Everyone is exposed to regression analysis in some form early on who undertakes scientific training, although sometimes that exposure takes a disguised form. Even the most basic statistical procedures taught to students in the sciences—the *t*-test and analysis of variance (ANOVA), for instance—are really just forms of regression analysis. After mastering these topics, students are often introduced to multiple regression analysis as if it is something new and designed for a wholly different type of problem than what they were exposed to in their first course. This book shows how regression analysis, ANOVA, and the independent groups *t*-test are one and the same. But we go far beyond drawing the parallels between these methods, knowing that in order for you to advance your own study in more advanced statistical methods, you need a solid background in the fundamentals of linear modeling. This book attempts to give you that background, while facilitating your understanding using a conversational writing tone, minimizing the mathematics as much as possible, and focusing on application and implementation using statistical software.

Although our intention was to deliver an introductory treatment of regression analysis theory and application, we think even the seasoned researcher and user of regression analysis will find him- or herself learning something new in each chapter. Indeed, with repeated readings of this book we predict you will come to appreciate the glory of linear modeling just as we have, and maybe even develop the kind of passion for the topic that we developed and hope we have successfully conveyed to you.

Regression analysis is conducted with computer software, and you have many good programs to choose from. We emphasize three commercial packages that are heavily used in the social and behavioral sciences: IBM SPSS Statistics (referred to throughout the book simply as "SPSS"), SAS, and STATA. A fourth program, R, is given some treatment in one of the appendices. But this book is about the concepts and application of regression analysis and is not written as a how-to guide to using your software. We assume that you already have at least some exposure to one of these programs, some working experience entering and manipulating data, and perhaps a book on your program available or a local expert to guide you as needed. That said, we do provide relevant commands for each of these programs for the key analyses and uses of regression analysis presented in these pages, using different fonts and shades of gray to most clearly distinguish them from each other. Your program's reference manual or user's guide, or your course instructor, can help you fine-tune and tailor the commands we provide to extract other information from the analysis that you may need one day.

In this rest of this preface, we provide a nonexhaustive summary of the contents of the book, chapter by chapter, to give you a sense of what you can expect to learn about in the pages that follow.

Overview of the Book

Chapter 1 introduces the book by focusing on the concept of "accounting for something" when interpreting research results, and how a failure to account for various explanations for an association between two variables renders that association ambiguous in meaning and interpretation. Two examples are offered in this first chapter, where the relationship between two variables changes after accounting for the relationship between these two variables and a third—a *covariate*. These examples are used to introduce the concept of *statistical control*, which is a major theme of the book. We discuss how the linear model, as a general analytic framework, can be used to account for covariates in a flexible, versatile manner for many types of data problems that a researcher confronts.

Chapters 2 and 3 are perhaps the core of the book, and everything that follows builds on the material in these two chapters. Chapter 2 introduces the concept of a *conditional mean* and how the ordinary least squares criterion used in regression analysis for defining the best-fitting model yields a model of conditional means by minimizing the sum of the squared residuals. After illustrating some simple computations, which are then replicated using regression routines in SPSS, SAS, and STATA, distinctions are drawn between the correlation coefficient and the regression coefficient as

related measures of association sensitive to different things (such as scale of measurement and restriction in range). Because the residual plays such an important role in the derivation of measures of partial association in the next chapter, considerable attention is paid in Chapter 2 to the properties of residuals and how residuals are interpreted.

Chapter 3 lays the foundation for an understanding of statistical control by illustrating again (as in Chapter 1, but this time using all continuous variables) how a failure to account for covariates can lead to misleading results about the true relationship between an independent and dependent variable. Using this example, the partialing process is described, focusing on how the residuals in a regression analysis can be thought of as a new measure—a variable that has been cleansed of its relationships with the other variables in the model. We show how the partial regression coefficient as well as other measures of partial association, such as the partial and semipartial correlation, can be thought of as measures of association between residuals. After showing how these measures are constructed and interpreted without using multiple regression, we illustrate how multiple regression analysis yields these measures without the hassle of having to generate residuals yourself. Considerable attention is given in this chapter to the meaning and interpretation of various measures of partial association, including the sometimes confusing difference between the semipartial and partial correlation. Venn diagrams are introduced at this stage as useful heuristics for thinking about shared and partial association and keeping straight the distinction between semipartial and partial correlation.

In many books, you find the topic of statistical inference addressed first in the simple regression model, before additional regressors and measures of partial association are introduced. With this approach, much of the same material gets repeated when models with more than one predictor are illustrated later. Our approach in this book is different and manifested in Chapter 4. Rather than discussing inference in the single and multiple regressor case as separate inferential problems in Chapters 2 and 3, we introduce inference in Chapter 4 more generally for any model regardless of the number of variables in the model. There are at least two advantages to this approach of waiting until a bit later in the book to discuss inference. First, it allows us to emphasize the mechanics and theory of regression analysis in the first few chapters while staying purely in the realm of description of association between variables with or without statistical control. Only after these concepts have been introduced and the reader has developed some comfort with the ideas of regression analysis do we then add the burden that can go with the abstraction of generalization, populations, degrees of freedom, tolerance and collinearity, and so forth. Second, with this approach, we need to cover the theory and mechanics of inference

only once, noting that a model with only a single regressor is just a special case of the more general theory and mathematics of statistical inference in regression analysis.

We return to the uses and theory of multiple regression in Chapter 5, first by showing that a dichotomous regressor can be used in a model and that, when used alone, the result is a model equivalent to the independent groups *t*-test with which readers are likely familiar. But unlike the independent groups *t*-test, additional variables are easily added to a regression model when the goal is to compare groups when holding one or more covariates constant (variables that can be dichotomous or numerical in any combination). We also discuss the phenomenon of regression to the mean, how regression analysis handles it, and the advantages of regression analysis using pretest measurements rather than difference scores when a variable is measured more than once and interest is in change over time. Also addressed in this chapter are measures and inference about partial association for sets of variables. This topic is particularly important later in the book, where an understanding of variable sets is critical to understanding how to form inferences about the effect of multicategorical variables on a dependent variable as well as testing interaction between regressors.

In Chapter 6 we take a step away from the mechanics of regression analysis to address the general topic of cause and effect. Experimentation is seen by most researchers as the gold-standard design for research motivated by a desire to establish cause–effect relationships. But fans of experimentation don't always appreciate the limitations of the randomized experiment or the strengths of statistical control as an alternative. Ultimately, experimentation and statistical control have their own sets of strengths and weaknesses. We take the position in this chapter that statistical control through regression analysis and randomized experimentation complement each other rather than compete. Although data analysis can only go so far in establishing cause–effect, statistical control through regression analysis and the randomized experiment can be used in tandem to strengthen the claims that one can make about cause–effect from a data analysis. But when random assignment is not possible or the data are already collected using a different design, regression analysis gives a means for the researcher to entertain and rule out at least some explanations for an association that compete with a cause–effect interpretation.

Emphasis in the first six chapters is on the regression coefficient and its derivatives. Chapter 7 is dedicated to the use of regression analysis as a prediction system, where focus is less on the regression coefficients and more on the multiple correlation *R* and how accurately a model generates estimates of the dependent variable in currently available or future data. Though no doubt this use of regression analysis is less common, an understanding of the subtle and sometimes complex issues that come up when

using regression analysis to make predictions is important. In this chapter we make the distinction between how well a sample model predicts the dependent variable in the sample, how well the "population model" predicts the dependent variable in the population, and how well a sample model predicts the dependent variable in the population. The latter is quantified with *shrunken R*, and we discuss some ways of estimating it. We also address mechanical methods of model construction, best known as *stepwise regression*, including the pitfalls of relinquishing control of model construction to an algorithm. Even if you don't anticipate using regression analysis as a prediction system, the section in this chapter on predictor variable configurations is worth reading, because complementarity, redundancy, and suppression are phenomena that, though introduced here in the context of prediction, do have relevance when using regression for causal analysis as well.

Chapter 8 is on the topic of variable importance. Researchers have an understandable impulse to want to describe relationships in terms that convey in one way or another the *size* of the effect they have quantified. It is tempting to rely on rules of thumb circulating in the empirical literature and statistics books for what constitutes a small versus a big effect using concepts such as the proportion of variance that an independent variable explains in the dependent variable. But establishing the size of a variable's effect or its importance is far more complex than this. For example, small effects can be important, and big effects for variables that can't be manipulated or changed have limited applied value. Furthermore, as discussed in this chapter, there is reason to be skeptical of the use of squared measures of correlations, which researchers often use, as measures of effect size. In this chapter we describe various quantitative, value-free measures of effect size, including our attraction to the semipartial correlation relative to competitors such as the standardized regression coefficient. We also provide an overview of dominance analysis as an approach to ordering the contribution of variables in explaining variation in the dependent variable.

In Chapters 9 and 10 we address how to include multicategorical variables in a regression analysis. Chapter 9 focuses on the most common means of including a categorical variable with three or more categories in a regression model through the use of *indicator* or *dummy* coding. An important take-home message from this chapter is that regression analysis can duplicate anything that can be done with a traditional single-factor one-way ANOVA or ANCOVA. With the principles of interpretation of regression coefficients and inference mastered, the reader will expand his or her understanding in Chapter 10, where we cover other systems for coding groups, including Helmert, effect, and sequential coding. In both of these chapters we also discuss contrasts between means either with or without control, including pairwise comparisons between means and

more complex contrasts that can be represented as a linear combination of means.

In the classroom, we have found that after covering multicategorical regressors, students invariably bring up the so-called *multiple test problem*, because students who have been exposed to ANOVA prior to taking a regression course often learn about Type I error inflation in the context of comparing three or more means. So Chapter 11 discusses the multiple test problem, and we offer our perspective on it. We emphasize that the problem of multiple testing surfaces any time one conducts more than one hypothesis test, whether that is done in the context of comparing means or when using any linear model that is the topic of this book. Rather than describing a litany of approaches invented for pairwise comparisons between means, we focus almost exclusively on the Bonferroni method (and a few variants) as a simple, easy-to-use, and flexible approach. Although this method is conservative, we take the position that its advantages outweigh its conservatism most of the time. We also offer our own philosophy of the multiple test problem and discuss how one has to be thoughtful rather than mindless when deciding when and how to compensate for multiple hypothesis tests in the inference process. This includes contemplating such things as the logical independence of the hypotheses, how well established the research area is, and the interest value of various hypotheses being conducted.

By the time you get to Chapter 12, the versatility of linear regression analysis will be readily apparent. By the end of Chapter 12 on nonlinearity, any remaining doubters will be convinced. We show in this chapter how *linear* regression analysis can be used to model *nonlinear* relationships. We start with polynomial regression, which largely serves as a reminder to the reader what he or she probably learned in secondary school about *functions*. But once these old lessons are combined with the idea of minimizing residuals through the least squares criterion, it seems almost obvious that linear regression analysis can and should be able to model curves. We then describe linear spline regression, which is a means of connecting straight lines at joints so as to approximate complex curves that aren't always captured well by polynomials. With the principles of linear spline regression covered, we then merge polynomial and spline regression into polynomial spline regression, which allows the analyst to model very complex curvilinear relationships without ever leaving the comfort of a linear regression analysis program. Finally, it is in this chapter that we discuss various transformations, which have a variety of uses in regression analysis including making nonlinear relationships more linear, which can have its advantages in some circumstances.

Up to this point in the book, one variable's effect on a dependent variable, as expressed by a measure of partial association such as the partial regression coefficient, is fixed to be independent of any other regressor.

This changes in Chapters 13 and 14, where we discuss *interaction*, also called *moderation*. Chapter 13 introduces the fundamentals by illustrating the flexibility that can be added to a regression model by including a cross-product of two variables in a model. Doing so allows one variable's effect—the focal predictor—to be a linear function of a second variable—the moderator. We show how this approach can be used with focal predictors and moderators that are numerical, dichotomous, or multicategorical in any combination. In Chapter 14 we formalize the linear nature of the relationship between focal predictor and moderator and how a function can be constructed, allowing you to estimate one variable's effect on the dependent variable, knowing the value of the moderator. We also address the exercise of *probing* an interaction and discuss a variety of approaches, including the appealing but less widely known Johnson–Neyman technique. We end this section by discussing various complications and myths in the study and analysis of interactions, including how nonlinearity and interaction can masquerade as each other, and why a valid test for interaction does not require that variables be centered before a cross-product term is computed, although centering may improve the interpretation of the coefficients of the linear terms in the cross-product.

Moderation is easily confused with *mediation*, the topic of Chapter 15. Whereas moderation focuses on estimating and understanding the boundary conditions or contingencies of an effect—when an effect exists and when it is large versus small—mediation addresses the question how an effect operates. Using regression analysis, we illustrate how one variable's effect in a regression model can be partitioned into direct and indirect components. The indirect effect of a variable quantifies the result of a causal chain of events in which an independent variable is presumed to affect an intermediate *mediator* variable, which in turn affects the dependent variable. We describe the regression algebra of path analysis first in a simple model with only a single mediator before extending it to more complex models involving more than one mediator. After discussing inference about direct and indirect effects, we dedicate considerable space to various controversies and extensions of mediation analysis, including cause–effect, models with multicategorical independent variables, nonlinear effects, and combining moderation and mediation analysis.

Under the topic of "irregularities," Chapter 16 is dedicated to regression diagnostics and testing regression assumptions. Some may feel these important topics are placed later in the sequence of chapters than they should be, but our decision was deliberate. We feel it is important to focus on the general concepts, uses, and remarkable flexibility of regression analysis before worrying about the things that can go wrong. In this chapter we describe various diagnostic statistics—measures of *leverage*, *distance*, and *influence*—that analysts can use to find problems in their data

or analysis (such as clerical errors in data entry) and identify cases that might be causing distortions or other difficulties in the analysis, whether they take the form of violating assumptions or producing results that are markedly different than they would be if the case were excluded from the analysis entirely. We also describe the assumptions of regression analysis more formally than we have elsewhere and offer some approaches to testing the assumptions, as well as alternative methods one can employ if one is worried about the effects of assumption violations.

Chapters 17 and 18 close the book by addressing various additional complexities and problems not addressed in Chapter 16, as well as numerous extensions of linear regression analysis. Chapter 17 focuses on power and precision of estimation. Though we do not dedicate space to how to conduct a power analysis (whole books on this topic exist, as does software to do the computations), we do dissect the formula for the standard error of a regression coefficient and describe the factors that influence its size. This shows the reader how to increase power when necessary. Also in Chapter 17 is the topic of measurement error and the effects it has on power and the validity of a hypothesis test, as well as a discussion of other miscellaneous problems such as missing data, collinearity and singularity, and rounding error. Chapter 18 closes the book with an introduction to logistic regression, which is the natural next step in one's learning about linear models. After this brief introduction to modeling dichotomous dependent variables, we point the reader to resources where one can learn about other extensions to the linear model, such as models of ordinal or count dependent variables, time series and survival analysis, structural equation modeling, and multilevel modeling.

Appendices aren't usually much worth discussing in the precis of a book such as this, but other than Appendix C, which contains various obligatory statistical tables, a few of ours are worthy of mention. Although all the analyses can be described in this book with regression analysis and in a few cases perhaps a bit of hand computation, Appendix A describes and documents the RLM macro for SPSS and SAS written for this book and referenced in a few places elsewhere in the book that makes some of the analyses considerably easier. RLM is not intended to replace your preferred program's regression routine, though it can do many ordinary regression functions. But RLM has some features not found in software off the shelf that facilitates some of the computations required for estimating and probing interactions, implementing the Johnson–Neyman technique, dominance analysis, linear spline regression, and the Bonferroni correction to the largest *t*-residual for testing regression assumptions, among a few other things. RLM can be downloaded from this book's web page at *www.afhayes.com*. Appendix B is for more advanced readers who are interested in the matrix algebra behind basic regression computations. Finally, Appendix D

addresses regression analysis with R, a freely available open-source computing platform that has been growing in popularity. Though this quick introduction will not make you an expert on regression analysis with R, it should get you started and position you for additional reading about R on your own.

To the Instructor

Instructors will find that our precis above combined with the Contents provides a thorough overview of the topics we cover in this book. But we highlight some of its strengths and unique features below:

- Repeated references to syntax for regression analysis in three statistical packages: SPSS, SAS, and STATA. Introduction of the R statistical language for regression analysis in an appendix.
- Introduction of regression through the concept of statistical control of covariates, including discussions of the relative advantages of statistical and experimental control in section 1.1 and Chapter 6.
- Differences between simple regression and correlation coefficients in their uses and properties; see section 2.3.
- When to use partial, semipartial, and simple correlations, or standardized and unstandardized regression coefficients; see sections 3.3 and 3.4.
- Is collinearity really a serious problem? See section 4.7.1.
- Truly understanding regression to the mean; see section 5.2.
- Using regression for prediction. Why the familiar "adjusted" multiple correlation overestimates the accuracy of a sample regression equation; see section 7.2.
- When should a mechanical regression prediction replace expert judgment in making decisions about real people? See sections 7.1 and 7.5.
- Assessing the relative importance of the variables in a model; see Chapter 8.
- Should correlations be squared when assessing relative importance? See section 8.2.
- Sequential, Helmert, and effect coding for multicategorical variables; see Chapter 10.
- A different view of the multiple test problem. Why should we correct for some tests, but not correct for all tests in the entire history of science? See Chapter 11.
- Fitting curves with polynomial, spline, and polynomial spline regression; see Chapter 12.
- Advanced techniques for probing interactions; see Chapter 14.

Acknowledgments

Writing a book is a team effort, and many have contributed in one way or another to this one, including various reviewers, students, colleagues, and family members. C. Deborah Laughton, Seymour Weingarten, Judith Grauman, Katherine Sommer, Jeannie Tang, Martin Coleman, and others at The Guilford Press have been professional and supportive at various phases while also cheering us on. They make book writing enjoyable and worth doing often. Amanda Montoya and Cindy Gunthrie provided editing and readability advice and offered a reader's perspective that helped to improve the book. Todd Little, the editor of Guilford's Methodology in the Social Sciences series, was an enthusiastic supporter of this book from the very beginning. Scott C. Roesch and Chris Oshima reviewed the manuscript prior to publication and made various suggestions, most of which we incorporated into the final draft. And our families, and in particular our wives, Betsy and Carole, deserve much credit for their support and also tolerating the divided attention that often comes with writing a book of any kind, but especially one of this size and scope.

RICHARD B. DARLINGTON
Ithaca, New York

ANDREW F. HAYES
Columbus, Ohio

List of Symbols and Abbreviations

Symbol	Meaning
b_0	regression constant
b_j	partial regression coefficient for regressor j
$\tilde{b}_j$	standardized partial regression coefficient for regressor j
B	number of hypothesis tests conducted
c_j	contrast coefficient for group j
Cov	covariance
$D_1, D_2 \ldots$	codes used in the representation of a multicategorical regressor
$DB(b_j)$	*dfbeta* for regressor j
df	degrees of freedom
E	expected value
e	residual
e_i	residual for case i
${}_de_i$	case i's residual when it is excluded from the model
F	F-ratio used in hypothesis testing
g	number of groups
h_i	leverage for case i
$J_1, J_2 \ldots$	artificial variables created in spline regression
k	number of regressors
LL	log likelihood
ln	natural logarithm
MD	Mahalanobis distance
MS	mean square
N	sample size
n_j	sample size of group j
p	observed significance or p-value
PE_i	probability of an event for case i
PR	partial multiple correlation
$PR(B.A)$	partial correlation for set B controlling for set A
pr_j	partial correlation for regressor j
R	multiple correlation
$R(A)$	R with regressors in set A
$R(AB)$	R with regressors in set A and set B
RS	shrunken R
r_{XY}	Pearson correlation coefficient
rel_j	reliability of regressor j

Symbol	Meaning
s_X	standard deviation of X
s_Y	standard deviation of Y
$s_{Y.X}$	standard error of estimate
SE	standard error
SR	semipartial correlation for a set
$SR(B.A)$	semipartial correlation for set B controlling for set A
sr_j	semipartial correlation for regressor j
str_i	standardized residual for case i
SS	sum of squares
$_T$	as a prefix, the true or population value of the quantity
t	t statistic used in hypothesis testing
tr_i	studentized residual for case i
t_j	t statistic for regressor j
Tol_j	tolerance for regressor j
Var	variance
Var($Y.X$)	variance of the residuals
VIF_j	variance inflation factor for regressor j
X	a regressor
$\overline{X}$	mean of X
X_j	regressor j
$X_{1.2}$	portion of X_1 independent of X_2
x	deviation from the mean of X
Y	usually the dependent variable
$\overline{Y}$	mean of Y
y	deviation from the mean of Y
$Y.1$	portion of Y independent of X_1
Z_f	Fisher's Z
Z_X	standardized value of X
Z_Y	standardized value of Y
$\overline{Y}$	mean of Y
$\hat{Y}$	estimate or fitted value of Y from a model
α	chosen significance level for a hypothesis test
α_{FW}	familywise Type I error rate
ΔR^2	change in R^2
$\hat{}$	estimated value
Π	multiplication
Σ	summation
θ_X	conditional effect of X
.	"controlling for"; for example, $r_{XY.C}$ is r_{XY} controlling for C

Contents

Data files for the examples used in the book and files containing the SPSS and SAS versions of RLM are available on the companion web page at *www.afhayes.com*.

1
Statistical Control and Linear Models

Researchers routinely ask questions about the relationship between an independent variable and a dependent variable in a research study. In experimental studies, relationships observed between a manipulated independent variable and a measured dependent variable are fairly easy to interpret. But in many studies, experimental control in the form of random assignment is not possible. Absent experimental or some form of procedural control, relationships between variables can be difficult to interpret but can be made more interpretable through *statistical control*. After discussing the need for statistical control, this chapter overviews the linear model—widely used throughout the social sciences, health and medical fields, business and marketing, and countless other disciplines. Linear modeling has many uses, among them being a means of implementing statistical control.

1.1 Statistical Control

1.1.1 The Need for Control

If you have ever described a piece of research to a friend, it was probably not very long before you were asked a question like "But did the researchers account for this?" If the research found a difference between the average salaries of men and women in a particular industry, did it account for differences in years of employment? If the research found differences among several ethnic groups in attitudes toward social welfare spending, did it account for income differences among the groups? If the research found that males who hold relatively higher-status jobs are seen as less physically attractive by females than are males in lower-status jobs, did it account for age differences among men who differ in status?

All these studies concern the relationship between an *independent variable* and a *dependent variable*. The study on salary differences concerns the

relationship between the independent variable of sex and the dependent variable of salary. The study on welfare spending concerns the relationship between the independent variable of ethnicity and the dependent variable of attitude. The study on perceived male attractiveness concerns the relationship between the independent variable of status and the dependent variable of perceived attractiveness. In each case, there is a need to account for, in some way, a third variable; this third variable is called a *covariate*. The covariates for the three studies are, respectively, years of employment, income, and age.

Suppose you wanted to study these three relationships without worrying about covariates. You may be familiar with three very different statistical methods for analyzing these three problems. You may have studied the *t*-test for testing questions like the sex difference in salaries, analysis of variance (also known as "ANOVA") for questions like the difference in average attitude among several ethnic groups, and the Pearson or rank-order correlation for questions like the relationship between status and perceived attractiveness. These three methods are all similar in that they can all be used to test the relationship between an independent variable and a dependent variable; they differ primarily in the type of independent variable used. For sex differences in salary you could use the *t*-test because the independent variable—sex—is *dichotomous*; there are two categories—male and female. In the example on welfare spending, you could use analysis of variance because the independent variable of ethnicity is *multicategorical*, since there are several categories rather than just two—the various ethnic groups in the study. You could use a correlation coefficient for the example about perceived attractiveness because status is *numerical*—a more or less continuous dimension from high status to low status. But for our purposes, the differences among these three variable types are relatively minor. You should begin thinking of problems like these as basically similar, as this book presents the *linear model* as a single method that can be applied to all of these problems and many others with fairly minor variations in the method.

1.1.2 Five Methods of Control

The layperson's notion of "accounting for" something in a study is a colloquial expression for what scientists refer to as *controlling for* that something. Suppose you want to know whether driver training courses help students pass driving tests. One problem is that the students who take a driver training course may differ in some way before taking the course from those

who do not take the course. If that thing they differ on is related to test performance, then any differences in test performance may be due to that thing rather than the training course itself. This needs to be accounted for or "controlled" in some fashion in order to determine whether the course helps students pass the test. Or perhaps in a particular town, some testers may be easier than others. The driving schools may know which testers are easiest and encourage their students to take their tests when they know those testers are on duty. So the standards being used to evaluate a student driver during the test may be systematically different for students who take the driver training course relative to those who do not. This also needs to be controlled in some fashion.

You might control the problem caused by preexisting difference between those who do and do not take the course by using a list of applicants for driving courses, randomly choosing which of the applicants is allowed to take the course, and using the rejected applicants as the control group. That way you know that students are likely to be equal on all things that might be related to performance on the test before the course begins. This is *random assignment on the independent variable*. Or, if you find that more women take the course than men, you might construct a sample that is half female and half male for both the trained and untrained groups by discarding some of the women in the available data. This is control by *exclusion of cases*.

You might control the problem of differential testing standards by training testers to make them apply uniform evaluation standards; that would be *manipulation of covariates*. Or you might control that problem by randomly altering the schedule different testers work, so that nobody would know which testers are on duty at a particular moment. That would not be random assignment on the independent variable, since you have not determined which applicants take the course; rather, it would be *other types of randomization*. This includes randomly assigning which of two or more forms of the dependent variable you use, choosing stimuli from a population of stimuli (e.g., in a psycholinguistics study, all common English adjectives), and manipulating the order of presentation of stimuli.

All these methods except exclusion of cases are types of *experimental control* since they all require you to manipulate the situation in some way rather than merely observe it. But these methods are often impractical or impossible. For instance, you might not be allowed to decide which students take the driving course or to train testers or alter their schedules. Or, if a covariate is worker seniority, as in one of our earlier examples, you cannot manipulate the covariate by telling workers how long to keep

their jobs. In the same example, the independent variable is sex, and you cannot randomly decide that a particular worker will be male or female the way you can decide whether the worker will be in the experimental or control condition of an experiment. Even when experimental control is possible, the very exertion of control often intrudes the investigator into the situation in a way that disturbs participants or alters results; ethologists and anthropologists are especially sensitive to such issues. Experimental control may be difficult even in laboratory studies on animals. Researchers may not be able to control how long a rat looks at a stimulus, but they are able to measure looking time.

Control by exclusion of cases avoids these difficulties, because you are manipulating data rather than participants. But this method lowers sample size, and thus lowers the precision of estimates and the power of hypothesis tests.

A fifth method of controlling covariates—statistical control—is one of the main topics of this book. It avoids the disadvantages of the previous four methods. No manipulation of participants or conditions is required, and no data are excluded. Several terms mean the same thing: to control a covariate statistically means the same as to *adjust for* it or to *correct for* it, or to *hold constant* or to *partial out* the covariate.

Statistical control has limitations. Scientists may disagree on what variables need to be controlled—an investigator who has controlled age, income, and ethnicity may be criticized for failing to control education and family size. And because covariates must be measured to be controlled, they will be controlled inaccurately if they are measured inaccurately. We return to these and other problems in Chapters 6 and 17. But because control of some covariates is almost always needed, and because the other four methods of control are so limited, statistical control is widely recognized as one of the most important statistical tools in the empiricist's toolbox.

1.1.3 Examples of Statistical Control

The nature of statistical control can be illustrated by a simple fictitious example, though the precise methods used in this example are not those we emphasize later. In Holly City, 130 children attended a city-subsidized preschool program and 130 others did not. Later, all 260 children took a "school readiness test" on entering first grade. Of the 130 preschool children, only 60 scored above the median on the test; of the other 130 children, 70 scored above the median. In other words, the preschool children scored worse on the test than the others. These results are shown in the "Total"

TABLE 1.1. Test Scores, Socioeconomic Status, and Preschool Attendance in Holly City

	Raw frequencies								
	Middle-class			Working-class			Total		
	A	B	Total	A	B	Total	A	B	Total
Preschool	30	10	40	30	60	90	60	70	130
Other	60	30	90	10	30	40	70	60	130

TABLE 1.2. Socioeconomic Status and Preschool Attendance in Holly City

	Percentage scoring above the median		
	Middle-class	Working-class	Total
Preschool	75	33	46
Other	67	25	54

section of Table 1.1; A and B refer to scoring above and below the test median, respectively.

But when the children are divided into "middle-class" and "working-class," the results are as shown on the left and center of Table 1.1. We see that of the 40 middle-class children attending preschool, 30, or 75%, scored above the median. There were 90 middle-class children not attending preschool, and 60, or 67%, of them scored above the median. These values of 75 and 67% are shown on the left in Table 1.2. Similar calculations based on the working-class and total tables yield the other figures in Table 1.2. This table shows clearly that within each level of socioeconomic status (SES), the preschool children outperform the other children, even though they appear to do worse when you ignore socioeconomic status (SES). We have *held constant* or *controlled* or *partialed out* the covariate of SES.

When we perform a similar analysis for nearby Ivy City, we find the results in Table 1.3. When we inspect the total percentages, preschool appears to have a positive effect. But when we look within each SES group, no effect is found. Thus, the "total" tables overstate the effect of

preschool in Ivy City and understate it in Holly City. In these examples the independent variable is preschool attendance and the dependent variable is test score. In Holly City, we found a negative simple relationship between these two variables (those attending preschool scored lower on the test) but a positive *partial* relationship (a term more formally defined later) when SES was controlled. In Ivy City, we found a positive simple relationship but no partial relationship.

By examining the data more carefully, we can see what caused these paradoxical results, known as *Simpson's paradox* (for a discussion of this and related phenomena, see Tu, Gunnel, & Gilthorpe, 2008). In Holly City, the 130 children attending preschool included 90 working-class children and 40 middle-class children, so 69% of the preschool attenders were working-class. But the 130 nonpreschool children included 90 middle-class children and 40 working-class children, so this group was only 31% working-class. Thus, the test scores of the preschool group were lowered by the disproportionate number of working-class children in that group. This might have occurred if city-subsidized preschool programs had been established primarily in poorer neighborhoods. But in Ivy City this difference was in the opposite direction: The preschool group was 75% middle-class, while the nonpreschool group was only 25% middle-class; thus, the test scores of the preschool group were raised by the disproportionate number of middle-class children. This might have occurred if parents had to pay for their children to attend preschool. In both cities the effects of preschool were seen more clearly by controlling for or holding constant SES.

All three variables in this example were dichotomous—they had just two levels each. The independent variable of preschool attendance had two levels we called "preschool" and "other." The dependent variable of test score was dichotomized into those above and below the median. The covariate of SES was also dichotomized. Such dichotomization is rarely if ever something you would want do in practice (as discussed later in section 5.1.6). Fortunately, with the methods described in this book, such categorization is not necessary. Any or all of the variables in this problem could have been numerically scaled. Test scores might have ranged from 0 to 100, and SES might have been measured on a scale with very many points on a continuum. Even preschool attendance might have been numerical, such as if we measured the exact number of days each child had attended preschool. Changing some or all variables from dichotomous to numerical would change the details of the analysis, but in its underlying logic the problem would remain the same.

TABLE 1.3. Socioeconomic Status and Preschool Attendance in Ivy City

	Raw frequencies								
	Middle-class			Working-class			Total		
	A	B	Total	A	B	Total	A	B	Total
Preschool	90	30	120	10	30	40	100	60	160
Other	30	10	40	30	90	120	60	100	100

	Percentage scoring above the median		
	Middle-class	Working-class	Total
Preschool	75	25	62
Other	75	25	38

Consider now a problem in which the dependent variable is numerical. At Swamp College, the dean calculated that among professors and other instructional staff under 30 years of age, the average salary among males was $81,000 and the average salary among females was only $69,000. To see whether this difference might be attributed to different proportions of men and women who have completed the Ph.D., the dean made up the table given here as Table 1.4.

If the dean had hoped that different rates of completion of the Ph.D. would explain the $12,000 difference between men and women in average salary, that hope was frustrated. We see that men had completed the Ph.D. *less* often than women: 10 of 40 men, versus 15 of 30 women. The first column of the table shows that among instructors with a Ph.D., the mean difference in salaries between men and women is $15,000. The second column shows the same difference of $15,000 among instructors with no Ph.D. Therefore, in this artificial example, controlling for completion of the Ph.D. does not lower the difference between the mean salaries of men and women, but rather *raises* it from $12,000 to $15,000.

This example differs from the preschool example in its mechanical details; we are dealing with means rather than frequencies and proportions. But the underlying logic is the same. In the present case, the independent variable is sex, the dependent variable is salary, and the covariate is

TABLE 1.4. Average Salaries at Swamp College

	Ph.D. completed		
	Yes	No	Total
Men	\$90,000 $n = 10$	\$78,000 $n = 30$	\$81,000 $n = 40$
Women	\$75,000 $n = 15$	\$63,000 $n = 15$	\$69,000 $n = 30$

educational level. Again, the partial relationship differs from the simple relationship, though this time both the simple and partial relationships have the same sign, meaning that men make more than women, with or without controlling for education.

1.2 An Overview of Linear Models

The examples presented in section 1.1.3 are so simple that you may be wondering why a whole book is needed to discuss statistical control. But when the covariate is numerical, it may be that no two participants in a study have the same measurement on the covariate and so we cannot construct tables like those in the two earlier examples. And we may want to control many covariates at once; the dean might want to simultaneously control teaching ratings and other covariates as well as completion of the Ph.D. Also, we need methods for inference about partial relationships such as hypothesis testing procedures and confidence intervals. Linear modeling, the topic of this book, offers a means of accomplishing all of these things and many others.

This book presents the fundamentals of linear modeling in the form of *linear regression analysis*. A linear regression analysis yields a mathematical equation—a *linear model*—that estimates a dependent variable Y from a set of predictor variables or *regressors* X. Such a linear model in its most general form looks like

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e \tag{1.1}$$

Each regressor in a linear model is given a numerical weight—the b next to each X in equation 1.1—called its *regression coefficient*, *regression slope*, or simply its *regression weight* that determines how much the equation uses values on that variable to produce an estimate of Y. These regression weights are derived by an algorithm that produces a mathematical equation or *model* for Y that best fits the data, using some kind of criterion for defining "best." In this book, we focus on linear modeling using the *least squares* criterion.

Linear modeling has many uses, among them being the process of statistical control introduced conceptually in the prior section. Linear modeling is widely used throughout the behavioral sciences, medical research and public health, business and marketing, and countless other fields. It is safe to say that one really cannot progress far in one's development as a scientist without a solid understanding of linear modeling. Most universities offer at least one and typically several courses on linear regression analysis. Indeed, it is so important that many if not most academic departments whose faculty use the scientific method regularly offer their own version of a course on linear modeling in one form or another.

The basic linear model method imposes six requirements:

1. As in any statistical analysis, there must be a set of "participants," "cases," or "units." In most every example and application in this book, the data come from people, so we use the term "participant" frequently. But case, unit, and participant can be thought of as synonymous and we use all three of these terms.

2. Each of these participants must have values or measurements on two or more variables, each of which is numerical, dichotomous, or multicategorical. Thus, the raw data for the analysis form a rectangular data matrix with participants in the rows and variables in the columns.

3. Each variable must be represented by a single column of numbers. For instance, the dichotomy of sex can be represented by letting the number 1 represent male and 0 represent female, so that the sexes of 100 people could be represented by a column of 100 numbers, each 0 or 1. A multicategorical variable with, say, five categories can be represented by a column of numbers, each 1, 2, 3, 4, or 5. For both dichotomous and multicategorical variables, the numbers representing categories are mere codes and are arbitrary. They carry no meaning about quantity and can be exchanged with any other set

of numbers without changing the results of the analysis so long as proper coding methods are used. And of course a numerical variable such as age can be represented by a column of ages.

4. Each analysis must have just one dependent variable, though it may have several independent variables and several covariates.

5. The dependent variable must be numerical. A numerical variable is something like age or income with interval properties, such that values can be meaningfully averaged.

6. Statistical inference from linear models often requires several additional assumptions that are described elsewhere in this book, such as in section 4.1.2 and Chapter 16.

Within these conditions, linear models are flexible in many ways:

1. A variable might be a natural property of a participant, such as age or sex, or might be a property manipulated in an experiment, such as which of two or more experimental conditions into which the participant is placed through a random assignment procedure. Manipulated variables are typically categorical but may be numerical, such as the number of hours of practice at a task participants are given or the number of acts of violence on television a person is exposed to during an experiment.

2. You may choose to conduct a series of analyses from the same rectangular data matrix, and the same variable might be a dependent variable in one analysis and an independent variable or covariate in another. For instance, if the matrix includes the variables age, sex, years of education, and salary, one analysis may examine years of education as a function of age and sex, while another analysis examines salary as a function of age, sex, and education.

3. As explained more fully in section 3.1.2, the distinction between independent variables and covariates may be fuzzy since linear modeling programs make no distinction between the two. The program computes a measure of the relationship between the dependent variable and every other variable in the analysis while controlling statistically for all remaining variables, including both covariates and other independent variables. Independent variables are those whose relationship to the dependent variable you wish to discuss or are the focus of your study, while covariates are other variables you wish to

control or otherwise include in the model for some other purpose. Thus, the distinction between the two determines how you describe the results of the analysis but is not used in writing the computer commands that specify the analysis or the underlying mathematics.

4. Each independent variable or covariate may be dichotomous, multicategorical, or numerical. All three variable types may occur in the same problem. For instance, if we studied salary in a professional firm as a function of sex, ethnicity, and age while controlling for seniority, citizenship (American or not), and type of college degree (business, arts, engineering, etc.), we would have one independent variable and one covariate from each of the three scale types.

5. The independent variables and covariates may all be intercorrelated, as they are likely to be in all these examples. In fact, the need to control a covariate typically arises because it correlates with one or more independent variables or the dependent variable or both.

6. In addition to correlating with each other, the independent variables and covariates may *interact* in affecting the dependent variable. For instance, age or sex might have a larger or smaller effect on salary for American citizens than for noncitizens. Interaction is explained in detail in Chapters 13 and 14.

7. Despite the names "linear regression" and "linear model," these methods can easily be extended to a great variety of problems involving curvilinear relations between variables. For example, physical strength is curvilinearly related to age, peaking in the 20s. But a linear model could be used to study the relationship between age and strength or even to estimate the age at which strength peaks. We discuss how in Chapter 12.

8. The assumptions required for statistical inference are not extremely limiting. There are a number of ways around the limits imposed by those assumptions.

There are many statistical methods that are just linear models in disguise, or closely related to linear regression analysis. For example, ANOVA, which you may already be familiar with, can be thought of as a particular subset of linear models designed early in the 20th century, well before computers were around. Mostly this meant using only categorical independent variables, no covariates, and equal cell frequencies if there were two or

more independent variables. When a problem does meet the narrow requirements of ANOVA, linear models and analysis of variance give the same answers. Thus, ANOVA is just a special subset of the linear model method. As shown in various locations throughout this book, ANOVA, *t*-tests on differences between means, tests on Pearson correlations—things you likely have already been exposed to—can all be thought of as special simple cases of the general linear model, and can all be executed with a program that can estimate a linear model.

Logistic regression, probit regression, and multilevel modeling are close relatives of linear regression analysis. In logistic and probit regression, the dependent variable can be dichotomous or ordinal, such as whether a person succeeds or fails at a task, acts or does not act in a particular way in some situation, or dislikes, feels neutral, or likes a stimulus. Multilevel modeling is used when the data exhibit a "nested" structure, such as when different subsets of the participants in a study share something such as the neighborhood or housing development they live in or the building in a city they work in. But you cannot fruitfully study these methods until you have mastered linear models, since a great many concepts used in these methods are introduced in connection with linear models.

1.2.1 What You Should Know Already

This book assumes a working familiarity with the concepts of means and standard deviations, correlation coefficients, distributions, samples and populations, random sampling, sampling distributions, standardized variables, null hypotheses, standard errors, statistical significance, power, confidence intervals, one-tailed and two-tailed tests, summation, subscripts, and similar basic statistical terms and concepts. It refers occasionally to basic statistical methods including *t*-tests, ANOVA, and factorial analysis of variance. It is not assumed that you remember the mechanics of these methods in detail, but some sections of this book will be easier if you understand the uses of these methods.

1.2.2 Statistical Software for Linear Modeling and Statistical Control

In most research applications, statistical control is undertaken not by looking at simple association in subsets of the data, as in the two examples presented earlier, but through *mathematical equating* or *partialing*. This process is conducted automatically through linear regression analysis and will

be described starting in Chapter 3. Suffice it to say now that statistical control is usually accomplished by computer software. Only the simplest linear models are practical without the aid of a computer.

Fortunately, most statistical packages that researchers have access to in one way or another include routines that conduct linear regression analysis. There are many statistical packages that can conduct regression analysis; examples include SPSS, SAS, SYSTAT, Minitab, and STATA, and most are available for Windows and MacOS. These are all commercial programs and can be quite expensive. Fortunately most universities purchase licenses for one or more of these programs that provide free or low-cost access to its faculty, staff, and students. Over the last decade, a freely available statistical language and program called R has become quite popular. It also has procedures built in that conduct the kind of analyses described in this book. R can be downloaded at no charge from *www.r-project.org*.

This book is about the principles of linear modeling, not about using software that implements the methods we describe. These principles are not software specific. We assume you already have some working familiarity with at least one statistics program capable of doing the types of analyses described in this book. In many chapters we include code for SPSS, SAS, or STATA that generates output pertinent to the analyses described. In Appendix B we offer a brief primer on the use of R for linear regression analysis. We chose to emphasize SPSS, SAS, and STATA because these programs are arguably most readily available and widely used by researchers in the social sciences, medical and health fields, business and marketing, and elsewhere. But you will not become an expert on the use of any of these programs by reading this book. It is no substitute for the documentation, a book dedicated to specific software packages, or a local expert who can guide you on its use.

SAS and R require the user to write *syntax* or *code* instructing the software which analysis is desired, which variables are playing the roles of independent and dependent variable, what options to produce in output, and so forth. SPSS is often chosen by beginners or adopted by instructors of introductory statistics classes because it has a friendly menu-based interface that allows the user to select various analyses and options by pointing and clicking on the screen rather than by typing instructions. STATA has a similar interface, though most users instruct STATA using code. For consistency, and because we believe that ultimately researchers need to be familiar with how to write code for their chosen program, we offer SPSS syntax rather than point-and-click instructions. Consult a local SPSS expert

for guidance on how to type in and execute syntax in SPSS if you are not already familiar with this.

A convention we follow in this book is to use different text colors and background for code corresponding to different programs. For SPSS code, we use white text in a black box. For example, SPSS code to produce a scatterplot, such as the one found in the beginning of Chapter 2, would appear in this book as

```
graph/scatterplot plays with points.
```

For SAS, the code will be set in black text in a white box. Thus, the corresponding SAS code to produce this scatterplot would look like

```
proc sgscatter data=golf;plot points*plays;run;
```

STATA code will appear as white text in a gray box. So corresponding STATA code would would appear as

```
twoway (scatter points plays)
```

Data files we use are archived on the web page for this book at *www.afhayes.com* in the native data file formats of SPSS and STATA as well as text files, along with code to produce corresponding data files in SAS. Throughout this book, when we provide computer instructions, we assume you already have a data file available for analysis and know how to open or generate data files in your chosen software. Therefore, we do not provide code or instructions for how to do so. We refer to data files by name using CAPITAL letters. Variables in those data files we refer to in the text using an *`italicized courier`* font. When we otherwise refer to computer code within the body of the text of this book rather than set in boxes as described above, we use the **`boldface courier`** font.

1.2.3 About Formulas

If you glance through this book, you will see many algebraic formulas. If formulas frighten you, relax. You can master the material in this book without memorizing any formulas. Most of the formulas in this book, or variations closely related to these formulas, are applied by computer programs. They are provided here merely so you will know what the program is doing. Many other formulas are for relatively uncommon problems. Still other formulas are so simple that they merely express concepts you can easily put into words. For instance, in Chapter 2 we define a deviation

score x_i as the difference between a raw score X_i and the sample mean $\overline{X}$, and another formula defines a variance Var(X) as the mean of the squared deviation scores. In this book there are only a few formulas that you can expect to ever have to actually apply to your own data by hand, without the assistance of a computer. That said, understanding the formulas is sometimes a good way of learning what the formulas represent conceptually as well as mathematically.

1.2.4 On Symbolic Representations

If you were to lay three or four statistical methods books side by side and open to chapters on a common topic, you would find a considerable lack of consistency in the symbols the authors use to refer to the same concepts. Although there would be some overlap—for instance, the use of the Greek letter mu (μ) to refer to a population mean and the Roman letter r as a reference to a correlation coefficient are both nearly universal—it would be hard to generalize your learning from one book to another if you relied entirely on symbolic representations of ideas used by one author. And two people who learn statistics from different books, who are taught by different instructors who use particular (and perhaps idiosyncratic) symbols to communicate ideas, or who come from different fields may appear to both outsiders and insiders to be speaking different languages even when talking about the same thing.

The Greek letter "beta" (β) and Roman letter b or B are examples. Some use β to refer to a population regression weight, whereas others use this symbol to refer to a sample regression weight. Still others might use β when talking about a standardized regression weight. SPSS labels some regression coefficients in its output with the word "beta" and others with B, and it is not uncommon to hear people talk about the "beta weights" or simply the "betas" in a regression model. You might be asked by someone if you aren't clear in a presentation whether you are presenting "bees or betas" from your model. The questioner may be asking whether you are reporting standardized or unstandardized regression coefficients, though others in the audience may not understand the meaning of this question as phrased if they were trained elsewhere or used a different book. Others restrict the use of Roman letters such as b or B to refer to estimates from a sample and reserve Greek symbols for parameters. But not all Greek letters refer to parameters in scientific discourse. For instance, people commonly report Cronbach's α calculated in a sample of participants who filled out

some kind of measurement instrument used to measure personality or attitude.

Ultimately, symbols are arbitrary. We can use whatever symbols we want to communicate ideas so long as we communicate what those symbols refer to when using them for the first time. Because there are relatively few conventions in the literature and books on linear modeling, we do not attempt to follow any of them. We introduce various symbols we use along the way—symbols that may at times be idiosyncratic to this book—but we always communicate what those symbols refer to unless there is a very strong convention in existence. As a reader, your job is to avoid assuming that one symbol that we use has the same meaning as this symbol when used by others. Such an assumption will inevitably result in confusion in your mind at some point.

1.3 Chapter Summary

Association between two variables X and Y can be difficult to interpret or obscured when a third variable Z is related to both X and Y. Researchers have a variety of procedural tricks they can employ to deal with such circumstances either prior to or following data collection, such as random assignment to X or other forms of randomization, case deletion, and maintaining strict control over various aspects of the research design and its administration. When none of these are possible, as is often the case, statistical control is an option and can render relationships easier to interpret and less susceptible to competing explanations. Linear modeling is one of the more frequently used procedures for statistical control by behavioral scientists, business and marketing researchers, investigators in health and medical fields, and many other disciplines. This book introduces the linear model in the form of linear regression analysis not only as a means of implementation of statistical control but also as a general and flexible tool that can be used for a variety of data-analytic tasks.

2
The Simple Regression Model

In this chapter we describe some of the key principles and ideas behind linear modeling in the form of linear regression analysis. We introduce the *simple linear regression model* as a means of estimating one variable Y from another variable X given information about the association between X and Y. Linear regression using the *least squares criterion*, or *ordinary least squares regression*, operates by figuring out how to weight X in an equation or *model* of Y so as to minimize the sum of the squared *residuals*. Residuals represent the difference between a model's estimate of Y and the actual values of Y observed in a data set. Because an understanding of the residuals from a regression model is so important to grasping the concept of statistical control, we devote an entire section to describing some of the algebraic properties of residuals as well as the process of *residual analysis*.

2.1 Scatterplots and Conditional Distributions

2.1.1 Scatterplots

Suppose as part of an organizational program to promote office morale, you and 22 of your office mates attend a retreat that involves playing miniature golf together. Upon questioning, you find that a few have never played before but most have, and some have played as many as six times before. In this golfing establishment, players are awarded points that can be used toward a discount next time based on their score at the end of a round of golf. This evening all members of your group win from 2 to 6 points.

You decide to examine the relationship between the number of points won and the number of times a player has played before. You start by numbering the people in your group (including yourself) from 1 to 23 and then recording for each player the number of previous plays and the number of points won. In Table 2.1, the identification numbers you assign

TABLE 2.1. Golfing Score and Prior Plays

ID	X	Y
1	0	2
2	0	3
3	1	2
4	1	3
5	1	4
6	2	2
7	2	3
8	2	4
9	2	5
10	3	2
11	3	3
12	3	4
13	3	5
14	3	6
15	4	3
16	4	4
17	4	5
18	4	6
19	5	4
20	5	5
21	5	6
22	6	5
23	6	6

are shown in the first column labeled ID. The number of previous plays is shown in the column labeled X, and the number of points each player won is shown in the column labeled Y.

Figure 2.1 is a *scatterplot* showing the pairs of X and Y in two-dimensional space for all 23 people in your group. The two dots on the far right of the figure represent the people with ID numbers 22 and 23. They have both played six times before ($X = 6$) and they score $Y = 5$, and $Y = 6$ points, respectively. The remaining 21 dots in the scatterplot are interpreted similarly.

2.1.2 A Line through Conditional Means

There are three people in Table 2.1 who reported having played miniature golf just once ($X = 1$) yet received 2, 3, or 4 points (Y). These three values of Y represent the *distribution* of Y of the three people who meet the condition

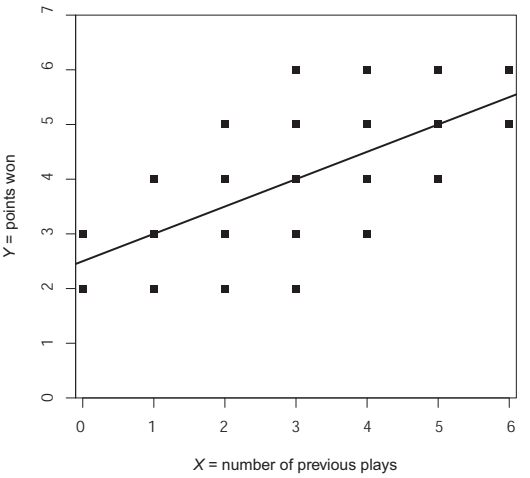


FIGURE 2.1. A simple scatterplot.

$X = 1$. As such, it is a *conditional distribution*. The mean of these three Y values is 3. We say that 3 is the *conditional mean* of Y when $X = 1$. This conditional mean is represented by the open circle over $X = 1$ toward the left side of Figure 2.2. The four people who reported playing twice ($X = 2$) have Y values of 2, 3, 4, and 5. The mean of these four values of Y is 3.5, so that is the conditional mean of Y when $X = 2$. This conditional mean is represented in Figure 2.2 by the open circle over $X = 2$. The conditional means at $X = 3$, $X = 4$, $X = 5$, and $X = 6$ are also shown as open circles. The overall mean of all 23 Y scores is called the *marginal mean* or *grand mean* of Y; in these data, the marginal mean ($\overline{Y}$) is 4.

In this example, all seven conditional means fall in a straight line. This line appears in Figures 2.1 and 2.2. This situation is called *linearity*. The number of units the line rises for each unit of X is called the *slope* of the line. The slope of the line equals the gain in Y associated with each 1-unit gain in X. As can be seen in the figures, the line rises one-half of a unit for each unit increase in X: the line is at $Y = 3$ when $X = 1$, is at $Y = 3.5$ when $X = 2$, is at $Y = 4$ when $X = 3$, and so on. Therefore, the slope of this line is 0.5. So we can say that each extra previous play is associated with a half-point average rise in points won. This answers our question about the

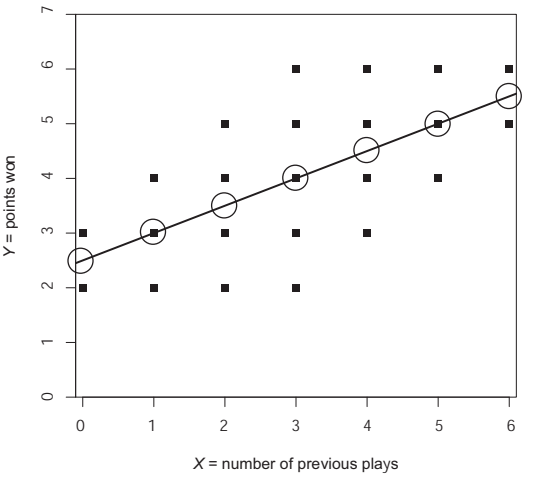


FIGURE 2.2. A line through conditional means.

relationship between points won and number of plays; each extra previous play is associated with an extra half a point won.

If the line sloped down from the upper left to the lower right, we would say that its slope was negative; if a line fell 2 units for each unit increase in X, we would say that its slope was -2. For instance, we might find a line with a negative slope if we plotted points won not against previous plays, but instead against the number of alcoholic drinks before playing.

In this example, the Y-value at which the line touches the Y-axis is called its *Y-intercept* or, more simply, just the line's *intercept*. In Figure 2.2, the line's Y-intercept is 2.5. More formally, the Y-intercept is the value of Y when $X = 0$. In Figure 2.2, the Y-axis is shown at $X = 0$, so the Y-intercept of 2.5 is indeed the point on the Y-axis at which the line crosses. In some visual depictions of association, such as when the Y-axis projects upward at a point on the X-axis that is different from zero, the Y-intercept may not correspond visually to the point at which the line crosses the Y-axis.

Like any straight line, this line can be represented by an algebraic equation. In many if not most elementary algebra textbooks, a straight line is represented by the equation $Y = b + mX$. Here we use X and Y in the same way, but instead of b and m we use b_0 and b_1, respectively. So in this book, the equation is written as $Y = b_0 + b_1X$, where b_0 is the Y-intercept and b_1 is

the slope of the line. In this example, $b_0 = 2.5$ and $b_1 = 0.5$ and so the line's equation is $Y = 2.5 + 0.5X$.

We can use the line or its equation to estimate new conditional means. For instance, if we extended the line to the right we would see that when $X = 7$, the line has a Y-value of 6. This means that 6 is the estimated conditional mean of Y when $X = 7$. You can find the same value from the line's equation by substituting $X = 7$ into the equation $Y = 2.5 + 0.5X$. We then have $Y = 2.5 + 0.5 \times 7 = 2.5 + 3.5 = 6$. Or when $X = 8$, the estimated conditional mean is $Y = 2.5 + 0.5 \times 8 = 2.5 + 4 = 6.5$. The conditional mean found this way is also the Y-value we estimate for any new player with a particular value of X. For instance, if someone played 7 times before, we estimate this this player will win 6 points. An estimated value of Y is denoted by $\hat{Y}$, pronounced "hat Y" or "Y hat." A "hat" over any value means an estimate of that value, so $\hat{Y}$ is an estimate of Y.

2.1.3 Errors of Estimate

How accurately can we estimate the Y-values (points won) from X (the number of previous times playing miniature golf)? Table 2.2 shows the $\hat{Y}$-value for each of the 23 people, as found from the regression equation or the regression line. The column labeled e shows $Y - \hat{Y}$, the *residual* or *error in estimate* for each person. These errors average 0 and always will using the methods described in this book, so the average error is not a good estimate of our accuracy. Intuitively, you would think that a measure of accuracy should not be zero (suggesting no inaccuracy) except in the circumstance where $Y = \hat{Y}$ for every person in the data. The *sum of squared errors*, also called the *sum of squared residuals* or $SS_{residual}$, is a measure of accuracy with this property. Formally,

$$SS_{residual} = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} e_i^2 \tag{2.1}$$

where Y_i is case i's actual value of Y (number of points earned), $\hat{Y}_i$ is what the equation estimates case i's Y value to be given information about case i's value of X (number of previous plays), and N is the sample size (23 in this example). The column labeled e^2 in Table 2.2 shows the squared errors for each player; its sum is 25, so the sum of squared errors or $SS_{residual}$ equals 25.

If linearity holds, as it does in this example, it can be shown that the line through the conditional means has a smaller sum of squared residuals

TABLE 2.2. Estimates and Residuals

ID	X	Y	$\hat{Y}$	e	e^2
1	0	2	2.5	−0.5	0.25
2	0	3	2.5	0.5	0.25
3	1	2	3.0	−1.0	1.00
4	1	3	3.0	0.0	0.00
5	1	4	3.0	1.0	1.00
6	2	2	3.5	−1.5	2.25
7	2	3	3.5	−0.5	0.25
8	2	4	3.5	0.5	0.25
9	2	5	3.5	1.5	2.25
10	3	2	4.0	−2.0	4.00
11	3	3	4.0	−1.0	0.00
12	3	4	4.0	0.0	0.00
13	3	5	4.0	1.0	1.00
14	3	6	4.0	2.0	4.00
15	4	3	4.5	−1.5	2.25
16	4	4	4.5	−0.5	0.25
17	4	5	4.5	0.5	0.25
18	4	6	4.5	1.5	2.25
19	5	4	5.0	−1.0	1.00
20	5	5	5.0	0.0	0.00
21	5	6	5.0	1.0	1.00
22	6	5	5.5	−0.5	0.25
23	6	6	5.5	0.5	0.25
Sum	69	92	92.0	0.0	25.00
Mean	3	4	4.0	0.0	1.09

than any other possible line that can be drawn through this scatterplot. For instance, Figure 2.3 shows an alternative line through the scatterplot. This line has equation $Y = 1X + 1$. The squared residuals and their sum are

$$1+4+0+1+4+1+0+1+4+4+1+0+1+4+4+1+0+1+4+1+0+1+4 = 42$$

This is considerably larger than the sum of 25 corresponding to the regression line defined as $Y = 2.5 + 0.5X$. As stated, no equation of the form $Y = b_0 + b_1X$ will produce a smaller sum of squared residuals when applied to these data. As such, $Y = 2.5 + 0.5X$ is the "best" equation for the relationship, conditioned on the assumption that the relationship is linear.

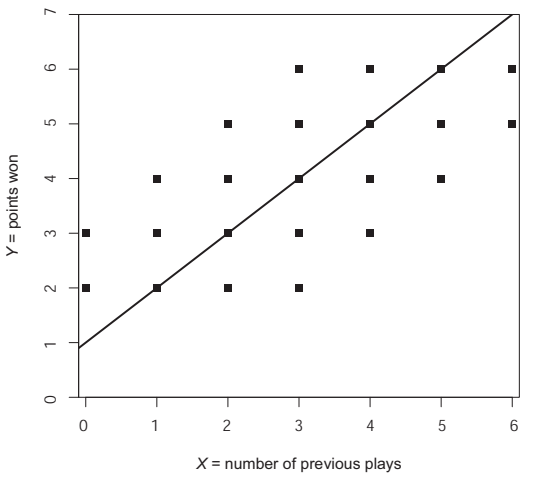


FIGURE 2.3. An alternative line with a larger sum of squared errors.

2.2 The Simple Regression Model

2.2.1 The Regression Line

Ordinarily, the data set we work with is a *sample* from a larger *population*. Because of random fluctuations from sample to sample, exact linearity (conditional means falling exactly in a straight line) hardly ever holds, though we often assume that it holds for the larger population. In fact, in a sample there may not even be any two people with exactly the same measurements on X, so the very concept of conditional means may have little or no meaning for the sample. Therefore, we need a way to derive a line and its equation that does not rely on sample values of conditional means.

The solution to this problem relies on the fact that the sum of squared residuals constructed from equation 2.1 is defined even when no two cases in the data have the same value of X. There is always one straight line that has a smaller $SS_{residual}$ than any other straight line. That line is called the *regression line*, and its equation called the *regression equation*. The regression equation consists of the *regression constant* b_0, also called the Y-intercept, and the *regression coefficient* b_1, which is the slope of the line. Because the

derivation of the regression equation is based on minimizing the sum of the squared residuals, this method is called *ordinary least squares regression* or just *OLS regression*.

2.2.2 Variance, Covariance, and Correlation

Regression lines can be computed from *covariances*. Covariances are not usually interpreted, but they are useful for computing both regression coefficients and correlations. Define x_i—a *deviation score*—as

$$x_i = X_i - \overline{X}$$

meaning that x_i is the deviation of person i's X measurement from the mean of X. A comparable deviation score, y_i, equals $Y_i - \overline{Y}$, or the deviation of person i's Y measurement from the mean of Y. The product $x_i y_i$ is the *cross-product* for person i. The cross-product is positive if person i is above the mean on both X and Y, or below on both. The cross-product is negative if person i is above the mean on one variable and below the mean on the other. The *covariance* between X and Y is the mean of the cross-products. We denote it Cov(XY). Thus,

$$\text{Cov}(XY) = \frac{\sum_{i=1}^{N}(x_i y_i)}{N} \tag{2.2}$$

where N is the sample size. An alternative formula uses the original values of X and Y rather than deviation scores:

$$\text{Cov}(XY) = \frac{N\sum_{i=1}^{N}(X_i Y_i) - (\sum_{i=1}^{N} X_i)(\sum_{i=1}^{N} Y_i)}{N^2}$$

The covariance of any variable with itself is the variable's *variance*. The variance of X we denote by Var(X). We have

$$\text{Var}(X) = \frac{\sum_{i=1}^{N}(x_i x_i)}{N} = \sum_{i=1}^{N}\frac{x_i^2}{N} \tag{2.3}$$

or in terms of original values of X rather than deviation scores:

$$\text{Var}(XY) = \frac{N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2}{N^2}$$

Like the covariance, the variance is not usually interpreted. But the variance is the square of an inherently interpretable statistic called the *standard deviation*. It is the square root of the variance:

$$s_X = \sqrt{\text{Var}(X)}$$

The standard deviation is a widely used measure of a distribution's variability or spread. As its name implies, the standard deviation is the "standard" measure of spread, because theorists have shown that in normal distributions it is less susceptible than other measures of spread to random fluctuation from sample to sample.

The *Pearson correlation coefficient*, or simply the *correlation*, between X and Y is defined as

$$r_{XY} = \frac{\text{Cov}(XY)}{s_X s_Y} \tag{2.4}$$

The correlation measures the strength of the association between X and Y; there is perfect linear association between X and Y if $r_{XY} = 1$ or $r_{XY} = -1$, whereas $r_{XY} = 0$ if X and Y are linearly independent. The sign of r conveys the *direction* of association. If r_{XY} is positive, that means cases above the mean on X tend to be above the mean on Y, and cases below the mean on X tend to be below the mean on Y. If r_{XY} is negative, then cases above the mean on one variable tend to be below the mean on the other.

Since a variance is a type of covariance and a standard deviation is the square root of a variance, equation 2.4 shows that a correlation is determined entirely by covariances.

2.2.3 Finding the Regression Line

Covariances also define the regression coefficient b_1. The formula is

$$b_1 = \frac{\text{Cov}(XY)}{\text{Var}(X)} \tag{2.5}$$

An alternative formula is

$$b_1 = r_{XY}\frac{s_Y}{s_X} \tag{2.6}$$

We can call equation 2.5 the *computing formula* and equation 2.6 the *definitional formula*. Equation 2.6 shows more clearly how b_1 relates to the familiar concepts of correlation and standard deviations, while equation 2.5 allows us to compute b_1 without taking any square roots.

If we multiply the numerator and denominator of equation 2.5 each by N, they become $\sum_{i=1}^{N} x_i y_i$ and $\sum_{i=1}^{N} x_i^2$, respectively. Thus, an alternative computing formula is

$$b_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{2.7}$$

Once b_1 has been found, we can find the intercept b_0 by the formula

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{2.8}$$

Then the equation for the regression line is

$$Y = b_0 + b_1 X \tag{2.9}$$

The estimated conditional mean for any value of X is found by substituting that value of X into equation 2.9. This equation also gives us the estimated Y for any person with that value of X. Since the symbol ˆ denotes "estimate of," equation 2.9 can be written

$$\hat{Y} = b_0 + b_1 X \tag{2.10}$$

which means that $b_0 + b_1 X$ yields an estimate of Y. Equation 2.10 is also called the *model* of Y since its purpose is to simulate or model Y as accurately as possible. The regression line found this way always passes through the location on a scatterplot with coordinates $\overline{X}$ and $\overline{Y}$.

For the data set used to construct the regression line, the line has a smaller sum of squared residuals than any other straight line one could construct. And if the sample was a random sample from a population in which linearity holds, then the slope of the regression line is an unbiased estimate of the slope of the population regression line, and any value of $\hat{Y} = b_0 + b_1 X$ is an unbiased estimate of the conditional mean for the value of X employed.

2.2.4 Example Computations

The formulas introduced in sections 2.2.2 and 2.2.3 are illustrated for the data set of 23 cases used in the miniature golf example (Table 2.1). Computers are ordinarily used to conduct the computations in regression analysis. We do this here only to illustrate the mathematics behind the computation of the regression coefficient and intercept in a regression analysis, not to give you a skill you will actually ever have to use or even want to use.

The means of X and Y are $\overline{X} = 3$ and $\overline{Y} = 4$. Thus, for person 1, the deviation scores are

$$\begin{aligned} x_1 &= 0 - 3 = -3 \\ y_1 &= 2 - 4 = -2 \end{aligned}$$

For person 1, we calculate

$$\begin{aligned} x_1^2 &= (-3)^2 = 9 \\ y_1^2 &= (-2)^2 = 4 \\ x_1 y_1 &= (-3) \times (-2) = 6 \end{aligned}$$

The entire set of raw scores, deviation scores, squares, and cross-products is shown in Table 2.3. Thus, $\text{Var}(X) = 2.96$, $\text{Var}(Y) = 1.83$, $\text{Cov}(XY) = 1.48$, and $r_{XY} = 1.48/\sqrt{2.96 \times 1.83} = 0.64$. Also, $b_1 = 34/68 = 0.5$ (from equation 2.7) and $b_0 = 4 - (34/68) \times 3 = 2.5$ (from equation 2.8).

Using the formulas above to estimate the conditional mean of Y at $X = 5$, for example, we have $2.5 + 0.5 \times 5 = 5$. This is also the estimate Y for any new person we might meet at the miniature golf course who had played five times before. Person 21 in the group has this value of X; for person 21, $X = 5$ and $Y = 6$. Thus, person 21's residual is $e_{21} = Y - \hat{Y} = 6 - 5 = 1$. This person received one point more than the regression model predicts given that he or she has played five times before.

It is also worth pointing out that the manual computations carried out here and throughout this book will not always correspond to output generated by statistical analysis software. Computers do computations to much higher precision than we do by hand. Rounding error in hand computations is generated by carrying mathematical operations only to the second or third decimal place. Furthermore, we used N as the denominator in equations 2.2 and 2.3 because we are merely *describing* the data, whereas computer software will typically use $N - 1$ assuming you are interested in *inference* to a population. As a result, if you calculated the variances of X and Y as well as their covariance, your software would probably show $\text{Var}(X) = 3.09$, $\text{Var}(Y) = 1.91$, $\text{Cov}(XY) = 1.55$.

Some textbooks define the covariance and variance using the $N - 1$ formula to eliminate the discrepancy between presented formulas and statistical software results, while others use N in the denominator as we do. When N is large, the difference is negligible, but in smaller samples the difference in the denominator will produce noticeable discrepancies. The $N-1$

TABLE 2.3. Regression Computations

Person	X	Y	x	y	x^2	y^2	xy
1	0	2	−3	−2	9.00	4.00	6.00
2	0	3	−3	−1	9.00	1.00	3.00
3	1	2	−2	−2	4.00	4.00	4.00
4	1	3	−2	−1	4.00	1.00	2.00
5	1	4	−2	0	4.00	0.00	0.00
6	2	2	−1	−2	1.00	4.00	2.00
7	2	3	−1	−1	1.00	1.00	1.00
8	2	4	−1	0	1.00	0.00	0.00
9	2	5	−1	1	1.00	1.00	−1.00
10	3	2	0	−2	0.00	4.00	0.00
11	3	3	0	−1	0.00	1.00	0.00
12	3	4	0	0	0.00	0.00	0.00
13	3	5	0	1	0.00	1.00	0.00
14	3	6	0	2	0.00	4.00	0.00
15	4	3	1	−1	1.00	1.00	−1.00
16	4	4	1	0	1.00	0.00	1.00
17	4	5	1	1	1.00	1.00	1.00
18	4	6	1	2	1.00	4.00	2.00
19	5	4	2	0	4.00	0.00	0.00
20	5	5	2	1	4.00	1.00	2.00
21	5	6	2	2	4.00	4.00	4.00
22	6	5	3	1	9.00	1.00	3.00
23	6	6	3	2	9.00	4.00	6.00
Sum	69	92	0	0	68.00	42.00	34.00
Mean	3	4	0	0	2.96	1.83	1.48

definition actually makes Var(X) a somewhat better estimator of the population variance, but this advantage is unimportant in regression analysis because the most important variance in regression is the residual variance, and a separate formula yields an unbiased estimator of it. We discuss the bias of an estimator in section 4.1.3. Take comfort that the use of N rather than $N - 1$ in these computations will produce the very same values of correlations, regression coefficients, or other regression computations that matter most.

2.2.5 Linear Regression Analysis by Computer

Reseachers use computers for the vast majority of data analysis tasks, and regression analysis is no exception. In this section we provide code for

SPSS, SAS, and STATA to generate the linear regression equation for the miniature golfing example. The data displayed in Table 2.1 can be found at *www.afhayes.com* in a file named GOLF. In this data set and analysis, the independent variable X is named `plays` and the dependent variable Y is named `points`.

The SPSS code that produces output for a linear regression estimating points earned from number of plays is

```
regression/dep=points/method=enter plays.
```

The corresponding SAS code is

```
proc reg data=golf;model points=plays;run;
```

and in STATA, try

```
regress points plays
```

The output generated by each set of commands can be found in Figure 2.4. Although each program formats its output differently, the outputs overlap considerably in the information they contain. Most of what is found in these outputs has not been discussed or even defined in this book thus far, but by the time you finish this book, you should be able to understand everything you see in these outputs.

All three outputs contain estimates of the regression constant $b_0 = 2.5$ and the regression coefficient $b_1 = 0.5$ (see the dashed boxes) that define the regression equation in the columns labeled "Unstandardized Coefficients" (SPSS), "Parameter Estimates" (SAS), or "Coef." (STATA). They also all show the sum of the squared residuals, $\sum e_i^2 = 25$, as calculated by hand in section 2.1.3. But they differ in how they label it. Whereas SPSS and STATA list it as the residual sum of squares, SAS calls it the error sum of squares.

Such discrepancies in how different programs format and label output are the norm rather than the exception. Don't let these discrepancies bother you too much. Most likely you will have a single preferred data analysis program that you rely on for the majority of the analyses that you do over much of your career. The principles we have discussed thus far and henceforth apply regardless of which statistical program you are using.

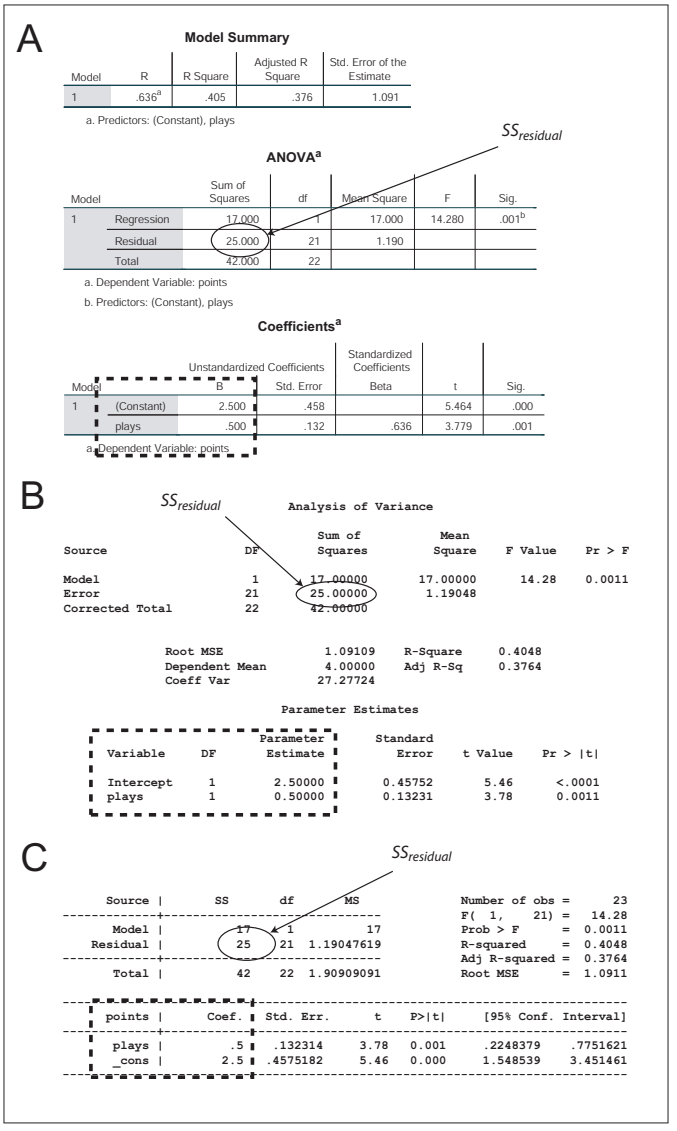


FIGURE 2.4. Regression analysis output from SPSS (A), SAS (B), and STATA (C). The regression intercept b_0 and coefficient b_1 are found in the sections highlighted by a dashed box.

2.3 The Regression Coefficient versus the Correlation Coefficient

Both r_{XY} and b_1 are measures of the relationship between X and Y; they are related by the formula $b_1 = r_{XY}(s_Y/s_X)$. As s_X and s_Y are always positive, b_1 and r_{XY} always have the same sign. Therefore, the two measures always agree on whether the relationship between X and Y is positive, negative, or zero.

The last statement applies to both samples and populations. Therefore, the hypothesis that a correlation is zero in the population is equivalent to the hypothesis that the corresponding regression coefficient is zero in the population. We will see later when discussing statistical inference that there is a t-test for testing the significance of r_{XY}, and a very different-appearing t-test for testing the significance of b_1. But in fact the two tests are equivalent, because in any one sample, both tests always give exactly the same value of t. Thus, both the null hypotheses and the tests are equivalent.

But b_1 and r_{XY} measure very different properties of a relationship. For example, it has been estimated that every cigarette smoked lowers one's life expectancy by about 11 minutes (Shaw, Mitchell, & Dorling, 2000). If this were derived from a linear regression analysis (with lifespan measured in minutes!), we would have $b_1 = -11$. Or suppose we studied the relation between hours of study for a test and number of points scored on the test. If we found $b_1 = 6$, it would mean that each extra hour of study is associated with a 6-point increase in test score. Statements like these convey information wholly lacking from statements like "The correlation between study time and test scores was +0.4," or "The correlation between smoking and lifespan is −0.2." The value of b_1 is the increase in $\hat{Y}$ associated with each 1-point increase in X.

How, then, do b_1 and r_{XY} differ? In terms of their *formulas*, they differ in primarily just one way; in terms of their *properties*, they differ in three major ways; and in terms of their *uses*, they differ in four major ways. We examine all of these differences in this section.

In terms of their *formulas*, r_{XY} is a *standardized* b_1. That is, if for each case in the data, we replaced their X_i and Y_i scores by their standardized values

$$Z_{X_i} = \frac{X_i - \overline{X}}{s_X}$$

$$Z_{Y_i} = \frac{Y_i - \overline{Y}}{s_Y}$$

and then predicted Z_Y from Z_X using linear regression, the regression coefficient for Z_X, b_1, would be r_{XY}. Another way of thinking about this is to recognize that when X and Y are standardized, s_X and s_Y both equal 1. Recalling from section 2.2.3 that $b_1 = r_{XY}(s_Y/s_X)$, this means that $b_1 = r_{XY}$ when X and Y are standardized. But the correlation between X and Y following standardization would be the same as r_{XY} without standardization.

2.3.1 Properties of the Regression and Correlation Coefficients

The first difference in the *properties* of r_{XY} and b_1 is that b_1, but not r_{XY}, is influenced by the *units* used to measure X and Y, so that b_1 is *scale-bound* but r_{XY} is *scale-free*. For instance, suppose we have a regression line estimating a child's weight from his or her age, and we switch from measuring weight in pounds to ounces. The switch will not change r_{XY} but it will multiply s_Y by 16, since there are 16 ounces in a pound. The formula $b_1 = r_{XY}(s_Y/s_X)$ shows us that b_1 is then also multiplied by 16. Also, if we switch from measuring ages in months to measuring them in years, s_X will be divided by 12, which will result in multiplying b_1 by 12. It is mathematically equivalent to say that children gain weight at an average rate of 0.2 pounds per month or 2.4 pounds per year, or 38.4 ounces per year, or 3.2 ounces per month, even though four different numbers are used to express the fact. Any of these four numbers could be the b_1 computed for the same set of children. But only one value of r_{XY} would be found.

The second difference in the properties of r_{XY} and b_1 is that r_{XY}, but not b_1, increases in absolute value with the *range* of the variables measured. For instance, suppose X is income, Y is amount saved annually for retirement, and there is a linear relationship between X and Y. If investigators A and B each study the relationship between X and Y but investigator A studies the entire nation while investigator B studies only one wealthy suburb of a particular city, then it may well be that both find the same value of b_1, but investigator A is likely to find that r_{XY} is much closer to 1 (in absolute value) than does investigator B, because X ranges more widely in investigator A's study.

Figure 2.5 illustrates this point. Figure 2.5, panel B, consists of the first three columns from Figure 2.5, panel A. In both parts, the diagonal lines pass through all conditional means, so we know without calculation that they are the regression lines. The regression line has the same slope in panels A and B. The mean of the squared residuals is also about the same in panel A as in panel B. But the positive correlation is clearly visible in

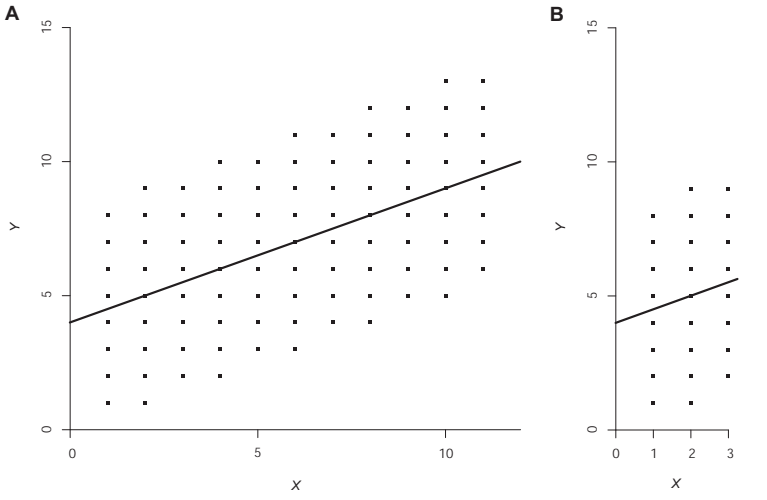


FIGURE 2.5. How range restriction affects r_{XY} without affecting b_1.

panel A but hardly noticeable in panel B; the values of r_{XY} in panels A and B are 0.54 and 0.16, respectively.

These figures relate to the formula $b_1 = r_{XY}(s_Y/s_X)$, but we can see the relation most easily by rewriting the formula as $r_{XY} = b_1(s_X/s_Y)$. The values of b_1 are equal in both panels of Figure 2.5. The value of s_Y is slightly larger in panel A than in panel B (2.90 vs. 2.43), which would tend to make r_{XY} smaller in panel A. But s_X is *much* larger in panel A (3.15 vs. 0.80), which makes r_{XY}, on the net, substantially larger in panel A than in panel B.

The third difference in the properties of r_{XY} and b_1 is that when Y is affected by *other variables* uncorrelated with X, r_{XY} is reduced but b_1 is not affected. r_{XY} is really a measure of the importance of the XY relationship relative to other factors affecting Y, whereas b_1 in a sense measures the absolute size of the relationship, ignoring other factors. For instance, if a nation had a safety campaign that greatly lowered the rate of accidental death, it would presumably not affect the regression coefficient mentioned earlier of an 11-minute reduction in life expectancy per cigarette. But the drop in accidental deaths would *raise* the correlation between smoking and lifespan (i.e., make the correlation more negative), because it would raise the importance of smoking relative to other factors affecting lifespan.

Rephrased more formally in terms of $r_{XY} = b_1(s_X/s_Y)$, where X is smoking and Y is lifespan, lowering the importance of accidental deaths would leave b_1 and s_X unchanged but would lower s_Y, thereby raising r_{XY}.

2.3.2 Uses of the Regression and Correlation Coefficients

We can summarize the difference in the *uses* of r_{XY} and b_1 by saying that b_1 is a better measure of X's *effect* on Y, especially in *experiments*, while r_{XY} is a better measure of *predictive power*, *relative importance*, and *statistical significance*. The phrase "better measure" is chosen carefully to avoid the implication that b_1 and r_{XY} are always *good* measures of these qualities. But when we find statistically that every cigarette smoked is associated with a 11-minute decrease in lifespan, that is a useful way of describing a relationship that *may* be causal.

When X is manipulated in an experiment, its range is a property of the experiment rather than a property of the natural world. For instance, suppose we correlate hours of exercise (X) with later physical fitness (Y). If an experimenter decides that participants in an experiment will exercise for 1, 3, or 5 hours, the range of $5 - 1 = 4$ hours has nothing to do with the range of hours of exercise that people would choose for themselves. Since r_{XY} is affected by range, the value we find from experimental data tells us nothing about the correlation we would find in nonexperimental data or in data from an experiment with a different range. But we would expect to find approximately the same value of b_1 in all these cases since b_1 is unaffected by range restriction.

But if we want to predict a person's grades in college from his or her grades in high school, or a worker's productivity from his or her score on an employment test, then r_{XY} measures the *predictive power* of X, meaning how accurately it estimates or predicts Y. If we wished to select the test with the greatest predictive power, we would select the test with the largest correlation with productivity. A test that happens to have a small standard deviation s_X might have a very high value of b_1, because of the formula $b_1 = r_{XY}(s_Y/s_X)$. But this would tell you nothing about how accurately X predicts Y except in the special case where $b_1 = 0$, meaning that X predicts Y with no accuracy whatsoever.

And r_{XY} is more closely related to statistical significance than b_1 is. In particular, with the proper assumptions, once you know r_{XY}, to test its statistical significance the only other information you need is the sample size N. Thus, with sample size fixed, the larger of the two correlations will

have a smaller p-value when testing the null hypothesis that the population correlation equals zero. This is not true of b_1.

2.4 Residuals

2.4.1 The Three Components of Y

After we have regressed Y on X (i.e., estimated Y from X in a regression analysis), we can partition each person's Y measurement Y_i into three components. Consider the equation

$$Y_i = \overline{Y} + (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i) \tag{2.11}$$

This is a simple algebraic identity; if you remove the parentheses and cancel values on the right, you find that the equation reduces to $Y_i = Y_i$.

The first of the three components, $\overline{Y}$, is constant for all people in the sample. Because $\hat{Y}_i$ is computed from X_i, the second component $(\hat{Y}_i - \overline{Y})$ correlates perfectly with X and is called the component of Y *explained* by X. The remaining component $(Y_i - \hat{Y}_i)$ is the *unexplained* or *independent* or *residual* portion of Y. Thus, if John's grade-point average (GPA) in school is 0.8 units above the mean and we predict from his SES that it would be 0.3 units above the mean, then John's SES explains 0.3 of his 0.8 units of deviation, and the other 0.5 is unexplained by SES.

The unexplained portion is the residual e_i in the regression of Y on X; in section 2.1.3 we defined $e_i = Y_i - \hat{Y}_i$. The explained and unexplained components are sometimes called the *model* and *error* components of Y. Thus, equation 2.11 can be rewritten as

$$Y = \overline{Y} + \text{explained component} + \text{unexplained component}$$

or as

$$Y = \overline{Y} + \text{model component} + \text{error component}$$

For instance, in the miniature golf example, the regression equation is $Y = 2.5 + 0.5X$. Furthermore, $\overline{Y} = 4$, $Y_{21} = 6$, and recall that $\hat{Y}_{21} = 5$. Thus, for person 21 the residual or error component is $Y_{21} - \hat{Y}_{21} = 6 - 5 = 1$ and the model component is $\hat{Y}_{21} - \overline{Y} = 5 - 4 = 1$. Indeed, observe that $Y_{21} = \overline{Y} + \text{model} + \text{error} = 4 + 1 + 1 = 6$.

The properties and uses of these three components will become clear gradually. The rest of this section considers one of the three components—the residual component.

2.4.2 Algebraic Properties of Residuals

Residuals have three important algebraic properties that are always true in any sample or population:

1. The mean of the residuals is exactly zero.

2. The residuals have exactly zero correlation with X. There is no tendency for the residuals to be more positive or more negative as X increases. The model component of Y is completely determined by X and correlates perfectly with X, so the residuals have zero correlation with the model component.

3. The variance of the residuals, denoted in this book as $\text{Var}(Y.X)$, is

$$\text{Var}(Y.X) = \text{Var}(Y)(1 - r_{XY}^2)$$

   which can be written as

$$\frac{\text{Var}(Y.X)}{\text{Var}(Y)} = 1 - r_{XY}^2 \tag{2.12}$$

   The left side of this equation is the variance of the residuals expressed as a proportion of the total variance of Y. We can call this the proportion of variance in Y not explained by X. The proportion explained is then 1 minus this unexplained portion, or r_{XY}^2. Therefore, we often speak of r_{XY}^2 as the proportion of the variance of Y that is explained by X.

In the miniature golfing example, $\text{Var}(Y) = 1.83$ (from Table 2.3), $\text{Var}(Y.X) = 1.09$ (from Table 2.2), and $r_{XY} = 0.64$ (from section 2.2.4). Thus, using equation 2.12,

$$\frac{1.09}{1.83} = 1 - 0.64^2$$

As mentioned earlier, we often assume that linearity holds in the population from which our sample is drawn. Other assumptions about the population, notably homoscedasticity and conditional normality, are discussed in section 4.1.2 and Chapter 16. However, the above three properties of residuals are strictly algebraic and do not depend at all on these or other assumptions.

2.4.3 Residuals as *Y* Adjusted for Differences in *X*

When we predict Y from X, we can think of the residuals in the regression as measurements on a new variable, which we can call *Y adjusted for X* or *Y corrected for X*. For instance, if Teresa's residual is 3 and Jason's is 2, then Teresa is 3 units higher on Y than we would have expected from her score on X, while Jason is only 2 units higher. Teresa is then higher than Jason on Y in relation to what we would have predicted from their scores on X. In other words, Teresa is higher than Jason on Y after adjusting or correcting for differences in their scores on X.

This use of regression residuals capitalizes on the fact that residuals have zero correlation with X. Thus, there is no tendency for a person's residual to be high just because that person scores high or low on X.

Residuals are important for understanding partial relationship, which we discuss starting in Chapter 3, but they have uses in their own right. For instance, consider the problem that organizers have when planning a 5-kilometer race for amateurs. If all entrants compete against each other, then the older runners have little chance of winning. But if a separate division is made for runners over 40, then runners in their 50s or older are still discriminated against. This problem could be solved by fitting a regression line predicting the running times from the runners' ages. The runner with the most *negative* residual (the shortest time adjusting for age differences) could be declared the winner. The disadvantage of this procedure is that no winner could be declared until all runners had completed the race, but it should help you to understand the uses and meaning of residuals. When yachts of different sizes race against each other, a procedure like this is sometimes used since larger boats go faster. The winning boat is the one that does best relative to its own predicted speed given its size. Regression can be used to derive the necessary formula for predicting speed from size.

2.4.4 Residual Analysis

We can define *residual analysis* as the process of selecting cases with highest and/or lowest residuals for scrutiny, in the hope that these will suggest insights into the factors affecting Y. We often engage in a kind of intuitive residual analysis when thinking about the accomplishments of others. For instance, we are often impressed by people who have achieved great success in spite of coming from poor family or economic conditions or having experienced a hard childhood. Similarly, we often wonder what went wrong when we learn about the mediocre accomplishments or legal trou-

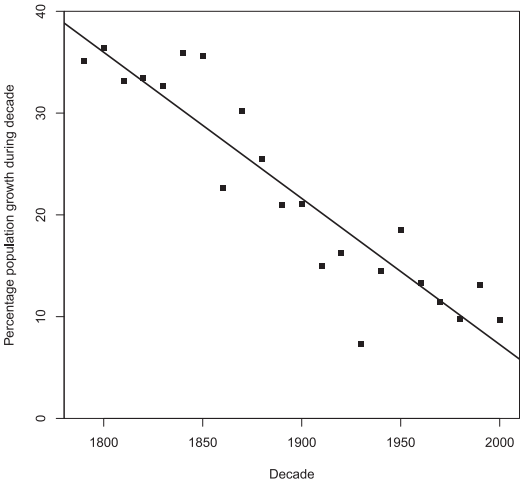


FIGURE 2.6. Percentage population growth in the United States for each decade of its history, with the regression line. (Source: U.S. Census Bureau).

bles of the children of highly successful parents who were able to provide unlimited resources to their offspring. Such people are noteworthy because they deviate in some way from expectation based on what we know about the causes of success and achievement.

As a more concrete example, investigators have selected the schools whose average student achievement is highest relative to the socioeconomic level of the school's neighborhood for further scrutiny to figure out what makes students in that school succeed. Studying those schools has then led to important insights about the factors producing successful schools—notably, for example, that the personality and determination of the principal make a surprising difference (Edmonds, 1986).

Figure 2.6 illustrates another example of residual analysis. This figure shows the percentage growth in the population of the United States for each decade of American history since the first population census in 1790. If we regress population growth (Y) on decade (X) (i.e., predict Y from X), we get the line shown. The negative regression slope indicates that the rate of population growth has slowed considerably since 1790. The line slopes

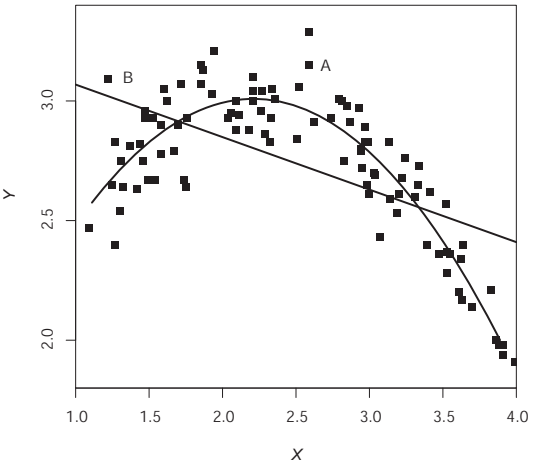


FIGURE 2.7. A violation of linearity.

down because we have plotted percentage increase in the population; if we had plotted absolute increase, the line would slope up.

An examination of this plot shows the five largest residuals to be for the decades following 1840, 1850, 1910, 1930, and 1950. Each of these is tied to an important event in history. The decade of the 1840s witnessed the great potato famine, which drove millions to America. The 1850s was a decade of political repression in Europe, which also drove large numbers to the United States. In the second decade of the 20th century, immigration was slowed by World War I, producing a negative residual. The 1930s saw the Great Depression, which lowered birthrates. And the 1950s saw the baby boom. The largest of these residuals was for the 1930s. This indicates that after correcting for the gradual slowing of the rate of population growth, the single most important event to affect population growth in the United States was the Great Depression.

Residual analysis is most meaningful if we can assume linearity and also that conditional distributions are identical except for means. Figure 2.7 illustrates a violation of linearity. In it, person A has a much larger positive residual than person B relative to the straight regression line that does not adequately describe the nonlinear association between X and Y. Thus, person A seems more extreme than person B. But person B is

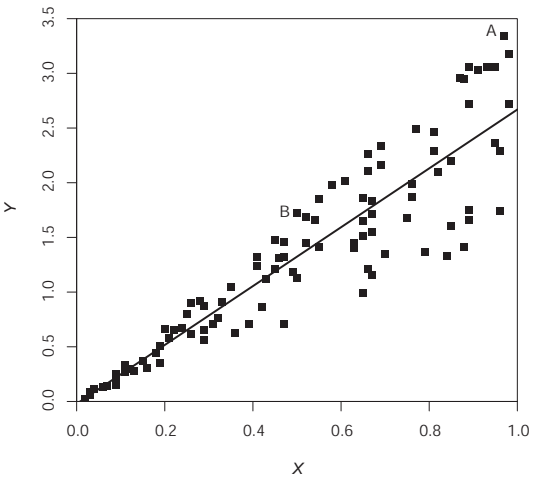


FIGURE 2.8. A violation of homoscedasticity (equality of conditional variances).

substantially farther above the true curved line than person A is. That is, person B is more extreme than person A with respect to the nonlinear model that more accurately captures the association between X and Y.

Figure 2.8 was generated by a computer program that made the conditional standard deviation of Y related to X, a phenomenon known as *heteroscedasticity*. In this example, person A is over twice as far above the regression line as person B. But when we correct properly for the different conditional standard deviations of Y, person B is farther above the regression line than person A. So the residual for person A is not especially large relative to person B when you account for differences in the variance in Y at different values of X. Thus, residual analysis is of limited value if we improperly assume linearity and equality of conditional variances, or *homoscedasticity*. But these assumptions sometimes are met or are not so severely violated that the validity of the analysis is brought into question, so residual analysis is an important tool in its own right. It is also the basis for understanding the concept of partial relationship, which is the fundamental tool of statistical control. The quantification and interpretation of measures of partial association is the topic of the next chapter.

2.5 Chapter Summary

This chapter has introduced some of the fundamental principles and concepts of linear regression, such as scatterplots, the regression equation, the intercept and slope of a regression line, regression residuals, the least squares criterion, residual analysis, and differences in the interpretation and use of Pearson's coefficient of correlation relative to the regression coefficient as a measure of association and effect. With these fundamentals mastered, you are prepared to develop an understanding and appreciation of one of the most important concepts in linear regression analysis and statistical control—*partial association*. This is the topic of the next chapter.

3
Partial Relationship and the Multiple Regression Model

This chapter extends the fundamentals of linear regression analysis introduced in Chapter 2 to models with more than one predictor variable. Measures of multivariate and partial association are derived and described, including the unstandardized and standardized partial regression weight, the partial correlation, and the semipartial correlation. We outline the process of statistical control through the partialing process and show how measures of partial association are based on the residuals from a regression analysis. We outline the differences between measures of partial association both in terms of computation and interpretation.

3.1 Regression Analysis with More Than One Predictor Variable

3.1.1 An Example

In section 1.1.3 we gave some examples of statistical control in which an independent variable and a covariate are both dichotomous. Those examples concerned the effects of preschool programs, and the difference between the salaries of male and female instructors at Swamp College. This section gives a similar example in which the independent variable, dependent variable, and covariate are all numerical. This case is much more complex, so we devote all of this chapter to this one example.

Suppose you conducted a study examining the relationship between food consumption and weight loss among people enrolled in a month-long healthy living class. The data from 10 participants can be found in Table 3.1, where X_1 is average weekly hours of exercise, X_2 is average daily food con-

TABLE 3.1. Exercise, Food Intake, and Weight Loss

ID	Exercise frequency (average weekly hours) X_1	Average daily food intake (100s of calories above recommended) X_2	Average weekly weight loss (100s of grams) Y
1	0	2	6
2	0	4	2
3	0	6	4
4	2	2	8
5	2	4	9
6	2	6	8
7	2	8	5
8	4	4	11
9	4	6	13
10	4	8	9
Mean	2	5	7.5

sumption (in 100s of calories above the recommended minimum of 1,000 calories required to maintain good health), and Y is average weight loss in hundreds of grams per week. The data file is available at *www.afhayes.com* and is named EXERCISE. In these data, Pearson's correlation between exercise frequency and weight loss is positive, as you would expect it to be: $r_{X_1Y} = 0.864$. Those who exercised relatively more during the month-long class lost relatively more weight.

Pearson's correlation between daily food consumption and weight loss is also positive, though very small, $r_{X_2Y} = 0.047$. Ignoring the fact that this relationship is tiny, the positive sign means that those who ate relatively more during the month lost relatively *more* weight than those who ate relatively less. This seems counterintuitive, if not also surprising. You'd think that eating less would translate into more weight loss. Intuition and common sense both suggest that the relationship between food consumption and weight loss should be negative.

This surprising finding can be explained by considering differences among the participants in the amount they exercised weekly. As can be seen in Table 3.1, participants 1, 2, and 3 did not exercise at all, participants 4, 5, 6, and 7 exercised 2 hours each week on average, and participants 8,

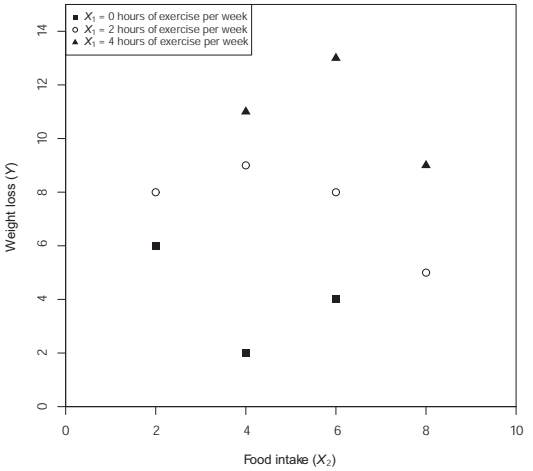


FIGURE 3.1. An example with positive simple association and negative partial association.

9, and 10 exercised 4 hours per week on average. Figure 3.1 shows weight loss plotted against food intake but uses different symbols for these three sets of participants who differ in the amount they exercised. The three participants who did not exercise at all are represented with squares, those who exercised 2 hours per week, with circles, and those who exercised 4 hours per week, with triangles.

If you look just at the squares in Figure 3.1—those who did not exercise at all—the relationship between food intake and weight loss appears negative. Turning your attention to the circles, representing those who exercised 2 hours per week, the relationship again seems negative. And again, focusing only on the triangles—participants who exercised 4 hours per week—the relationship is clearly negative. So when exercise is *held constant*, the expected negative relationship between food intake and weight loss appears. The surprising positive (albeit tiny) relationship that exists when ignoring exercise is merely due to the fact that those who consumed more also exercised the most, and those who exercised the most lost the most weight.

How can we state more precisely the relationship between food intake and weight loss when exercise is held constant? In this example, the negative relationship between these two variables is clearly visible when

examining subgroups of people equal on exercise. But you can imagine a similar data set in which no two people have exercised exactly the same amount. In that case, how could exercise be held constant? And how exactly do we measure the relationship between two variables when holding a third variable constant? The bulk of this entire chapter describes methods for answering these questions.

3.1.2 Regressors

In the previous example, we asked about the relationship between food intake and weight loss when exercise is held constant and found that the relationship is negative, as would be expected. In this example, food consumption is the independent variable (the presumed cause of weight loss or lack thereof), weight loss is the dependent variable (the presumed effect of lower food consumption), and exercise frequency is a covariate. But we could just as easily ask about the relationship between exercise frequency and weight loss, holding food consumption constant. In that case, exercise frequency could be thought of as the independent variable, weight loss the dependent variable, and food consumption the covariate.

We will see that in finding the answer to the first question using regression analysis, we automatically find the answer to the second question. For this reason, mathematically we don't need to distinguish between independent variables and covariates when estimating regression models. So we use the term *regressor* to include both independent variables and covariates in a regression analysis.

In a typical regression analysis problem, we tell the computer the name of the dependent variable in the data that is being modeled, as well as the name of one or more regressors, making no distinction between independent variable(s) and covariates. The computer then prints a measure of the partial relationship between the dependent variable and each regressor, holding constant all the other regressors. For example, using the weight-loss data set, the SPSS code below will conduct a regression analysis estimating dependent variable *wtloss* from regressors *exercise* and *food*:

```
regression/dep=wtloss/method=enter exercise food.
```

In SAS, try

```
proc reg data=exercise;model wtloss=exercise food;run;
```

and in STATA, use

```
regress wtloss exercise food
```

The computer does not care which variable is the independent variable and which is the covariate in your thinking about the problem, although it certainly requires you to specify which is the dependent variable (in STATA, the dependent variable is the first variable listed after the **regress** command). Each regressor in a regression model simultaneously functions as both independent variable and covariate. So the distinction between independent variable and covariate is not a mathematical one. Rather, the distinction enters when the data analyst looks at the computer output, interprets that output, or talks or writes about about the partial relationships that a regression analysis provides.

Because independent variables and covariates are both regressors, we usually do not even bother to distinguish between them in our notation. In this example, the regressors are X_1 and X_2 rather than, say, *IV* for independent variable and *C* for covariate. Some authors use the term *independent variables* to refer to all regressors, but we will normally use it in the more restrictive sense explained here—as the variables whose relationship to Y most interests us at the time of interpretation or when describing the results of the analysis in a talk or research report, such as an academic journal article.

3.1.3 Models

Suppose an examination of the literature on diet, exercise, and weight loss led you to the following conclusions:

1. People who eat the minimum number of calories to maintain good health and who do not exercise at all will lose about 600 grams of weight on average in a week.

2. If food intake is held constant, then each 1 hour of average daily exercise leads to an average weight loss of 200 grams per week.

3. If you think of food intake above the minimum in units of 100 calories each, then someone who eats 2,000 calories per day is consuming 10 units above the 1,000-calorie minimum. You conclude that each one unit of food intake per day (i.e., 100 calories) above the minimum to maintain good health translates to 50 grams less weight loss per week than the person otherwise would have experienced by eating only the minimum recommended calories.

From these conclusions, you could estimate how much weight a person could expect to lose who ate 1,600 calories per day (i.e., 6 units above the 1,000-calorie minimum) and exercised 2 hours daily per week. The reasoning for your estimate might go something like this:

> Someone who eats the minimum of 1,000 calories and who doesn't exercise can expect to lose about 600 grams in a week. But if such a person exercises 2 hours per week, this translates to an additional 400 grams of weight loss (200 grams per week per hour of exercise). But by eating 6 units more than the minimum (600 calories), this adds back 300 grams (6×50) to what otherwise would have been lost. So the estimated weight loss for such a person is $600 + 400 - 300 = 700$ grams per week.

This reasoning could be represented mathematically in the form of a *model* of weight loss. Letting Y be weight loss in 100s of grams, X_1 be hours of exercise per week, and X_2 be units of food intake above the 1,000-calorie minimum in 100s of grams, the mathematical model is

$$\hat{Y} = 6 + 2X_1 - 0.5X_2$$

You could apply this equation to any person to estimate how much weight that person could be expected to lose based on the amount he or she exercises and consumes per day. This is a *linear* model because the numbers by which X_1 and X_2 are multiplied when generating the estimate are simple numbers; *nonlinear models* are described in Chapter 12.

This model could be written in a more generic form as

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

where b_0 is the model constant as in a simple regression model, and b_1 and b_2 are *coefficients* or *weights* for X_1 and X_2, respectively. In this example, $b_0 = 6$, $b_1 = 2$, and $b_2 = -0.5$. We shall see shortly that b_1 and b_2 are also called the *slopes* in a geometric representation of the model. They are also often called *beta weights*, *beta coefficients*, or just the *betas* of the model. We will avoid the use of the term *beta* because, as discussed in section 1.2.4, it is used in so many ways by different people who write about regression that doing so just invites confusion.

Because the regression coefficients in a model are scale-bound, they cannot be meaningfully compared with each other. For instance, the coefficients of 2 and −0.5 for exercise and food intake might lead you to believe that exercise is more highly associated with weight loss or more important

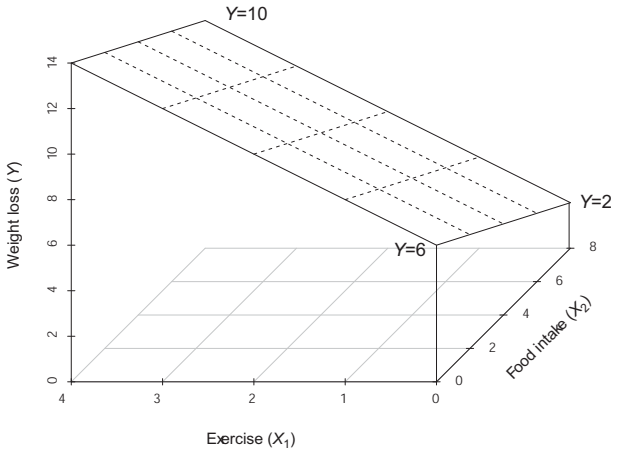


FIGURE 3.2. Plane representing the model $Y = 6 + 2X_1 - 0.5X_2$.

in the model as a predictor of weight loss than is food intake since 2 is larger in absolute value than −0.5. But had we expressed exercise time in minutes rather than hours, then b_1 would change from 2 to 0.033, which makes exercise look less important than food intake. The problem of comparing the importance of regressors is examined at length in Chapter 8.

3.1.4 Representing a Model Geometrically

In Chapter 2 we illustrated that a simple regression model of the form $\hat{Y} = b_0 + b_1X$ can be represented geometrically by a straight line. A model of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ can be represented by a plane in three-dimensional space, as in Figure 3.2. In fact, this figure represents the very model $\hat{Y} = 6 + 2X_1 - 0.5X_2$ that we have already considered. The left-hand horizonal axis represents exercise (X_1), the right-hand axis represents food intake (X_2), and the vertical axis represents weight loss (Y).

To see how this plane represents the model, consider first someone with zero on both exercise and food intake. That is, $X_1 = 0$ and $X_2 = 0$. The model estimates that this person will lose 600 grams per week (i.e., six 100-gram units) since for that person $\hat{Y} = 6 + 2 \times 0 - 0.5 \times 0 = 6$. At the near corner of the figure, you see that the plane has a Y value of 6 when both X_1 and X_2 are equal to zero.

Now consider someone for whom $X_1 = 4$ and $X_2 = 0$. Inserting these values into the model yields $\hat{Y} = 6 + 2 \times 4 - 0.5 \times 0 = 14$. At the upper left corner of the tilted plane, you can see that $Y = 14$ when $X_1 = 4$ and $X_2 = 0$. Since the plane rises 8 units, from 6 to 14, as X_1 rises 4 units from 0 to 4, its slope relative to X_1 is $8/4 = 2$. If you try this with different values of X_1 and X_2, you will find that regardless of the value of X_2 you choose, when you hold X_2 constant at that value, each 1 unit increase in X_1 results in an increase of Y of 2 units. This is the coefficient for X_1 in the model. It represents the amount Y is estimated to change as X_1 increases by 1 unit but X_2 is held constant.

Now consider someone for whom $X_1 = 0$ and $X_2 = 8$. Inserting these values into the model, we find $\hat{Y} = 6 + 2 \times 0 - 0.5 \times 8 = 2$. At the far right corner of the tilted plane, observe that $Y = 2$ when $X_1 = 0$ and $X_2 = 8$. Comparing this to the value of $Y = 6$ when $X_1 = 0$ and $X_2 = 0$, Y falls 4 units, from 6 to 2, when X_2 increases by 8 but X_1 is fixed. So the plane's slope relative to X_2 is $4/8 = -0.5$. Regardless of the value of X_1 you choose, when you hold X_1 constant at that value, each 1 unit increase in X_2 results in an increase of Y of -0.5 units or, in other words, a decrease of 0.5 units. This is the coefficient for X_2 in the model. It represents the amount Y is estimated to change as X_2 increases by 1 unit but X_1 is held constant.

In summary, a model that represents $\hat{Y}$ as a linear function of two variables, X_1 and X_2, can be represented as a plane in three-dimensional space. The plane's Y value at $X_1 = X_2 = 0$ represents the additive constant, b_0 in the model. And the plane's slopes relative to X_1 and X_2 represent the coefficients for X_1 and X_2, or b_1 and b_2.

3.1.5 Model Errors

We have imagined that you made up the model from reading the literature on weight loss. Yet you have data from 10 people in Table 3.1 who participated in a weight-loss class, each measured on the variables in the model. How could you measure the consistency between the model of weight loss you constructed and the weight loss actually experienced by these 10 people? Might it be the case that a different model using the same regressors would actually produce a more accurate estimate of weight loss? For example, perhaps setting $b_1 = 2.1$ and $b_2 = -0.6$ would produce a more accurate estimate of how much weight a person actually loses based on food intake and exercise frequency.

As in simple regression, we can measure the discrepancy between the estimates of Y generated by the model and the values of Y available in

TABLE 3.2. Estimates and Residuals in the Weight-Loss Data Set

ID	Exercise X_1	Food intake X_2	Weight loss Y	Estimate $\hat{Y}$	Residual e
1	0	2	6	5	1
2	0	4	2	4	−2
3	0	6	4	3	1
4	2	2	8	9	1
5	2	4	9	8	1
6	2	6	8	7	1
7	2	8	5	6	−1
8	4	4	11	12	−1
9	4	6	13	11	2
10	4	8	9	10	−1
Mean	2	5	7.5	7.5	0

a data set by quantifying the amount the estimates of Y depart from the actual values of Y observed. That is, the model generates $\hat{Y}_i$ for each case, which allows us to construct $e_i = Y_i - \hat{Y}_i$, the residual or error in estimate. Table 3.2 contains the data in Table 3.1 but includes two additional columns containing the estimates of Y from the model $\hat{Y} = 6 + 2X_1 - 0.5X_2$, as well as the residuals e_i, for each of the 10 cases.

We can give these data a geometric interpretation. Figure 3.3 shows the data for our 10-person sample plotted in three-dimensional space. For instance, for person 1 we have $X_1 = 0$, $X_2 = 2$, and $Y = 6$. Person 1 in this figure appears atop a stick 6 units long, whose base is at $X_1 = 0, X_2 = 2$. Other people in the figure are represented similarly.

If we put the plane and the 10 cases into the same figure, we get Figure 3.4. In this figure, each person's vertical distance between Y and the plane—which represents $\hat{Y}$—is shown by a short vertical line. The length of this line for person i is e_i, which is positive for those above the plane and negative for those below. Persons 2 and 9 are each 2 units from the plane, and all other people are just 1 unit from the plane.

As the size of these vertical lines—the errors in estimation—are determined in part by $\hat{Y}$, it follows that a different model will produce different errors in estimation for each case. A good-fitting model is one in which these errors of estimation are small. The consistency between a model and the data is determined by how small these errors tend to be aggregated

across all cases in the data. A model that is more consistent with the data will have smaller errors in estimation than a model that is less consistent with the data.

3.1.6 An Alternative View of the Model

Figures 3.2 and 3.4 represent $\hat{Y} = 6 + 2X_1 - 0.5X_2$ in three-dimensional space. Figure 3.5 represents the model in two dimensions but otherwise contains the same information. Line *AB* in Figure 3.4 falls from 6 to 2 as X_2 increases from 0 to 8. Line *IJ* in this figure falls from 14 to 10 over the same range of X_2. Lines *CD*, *EF*, and *GH* are evenly spaced between them. Figure 3.5 shows the same five lines drawn in a diagram of Y against X_2. Like their counterparts in Figure 3.4, line *AB* in Figure 3.5 falls from 6 to 2, *IJ* falls from 14 to 10, and lines *CD*, *EF*, and *GH* are evenly spaced between them. Thus, Figure 3.5 conveys all the same information as Figure 3.4; it represents the model in two-dimensional space rather than three-dimensional space.

As you can see in Figure 3.5, line *AB* applies when $X_1 = 0$, line *IJ* applies when $X_1 = 4$, and the other three parallel lines apply when X_1 equals 1, 2, and 3. The appropriate value of X_1 is written next to each sloping line in Figure 3.5. Because the five sloping lines in Figure 3.5 apply to values of X_1 1 unit apart, the vertical distance between them is the amount the plane rises as X_1 increases 1 unit. But this is the slope of the plane relative to X_1. Thus, in this representation of the model, the vertical distance between parallel lines represents the slope or apparent effect of X_1.

Figure 3.6 also conveys all the information in Figure 3.4, but in Figure 3.6, Y is plotted against X_1 instead of X_2. Like line *AI* in Figure 3.4, line *AI* in Figure 3.6 rises from 6 to 14 as X_1 increases from 0 to 4. Similarly, line *BJ* in Figure 3.6 corresponds to line *BJ* in Figure 3.4. But there the lines are for values of X_2 2 units apart, so the vertical distance between them is twice b_2.

In summary, a linear model involving two X variables can also be represented by a series of parallel lines in two dimensions whose slope equals one slope of the plane when the model is represented in three dimensions. If lines represent values 1 unit apart on the other X variable, then the vertical distance between adjacent lines represents the plane's other slope.

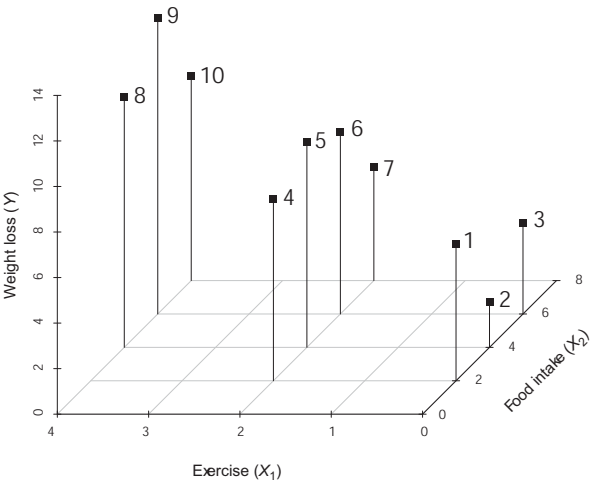


FIGURE 3.3. Ten data points plotted in three-dimensional space.

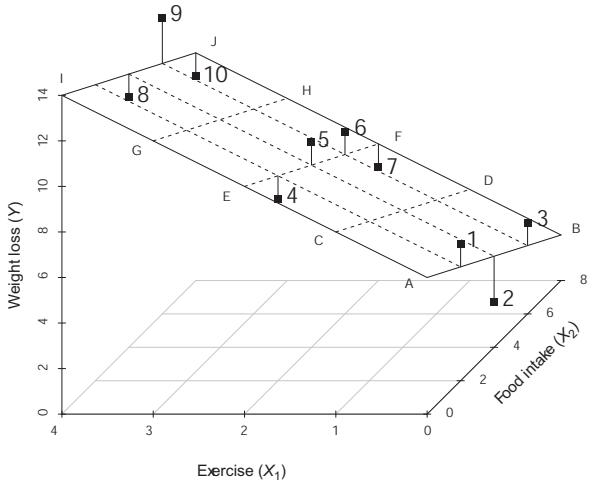


FIGURE 3.4. The data and the best-fitting plane.

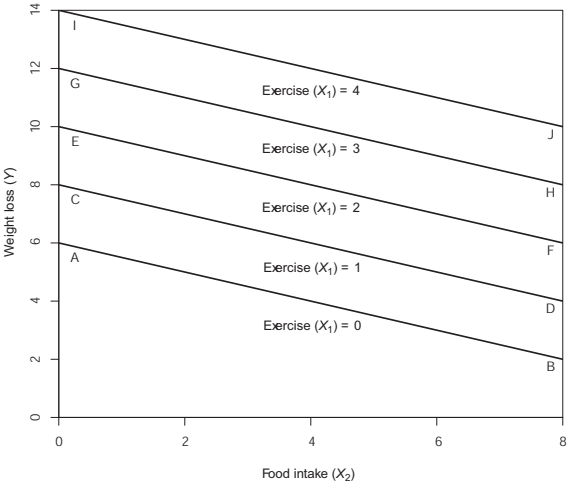


FIGURE 3.5. Another representation of the tilted plane.

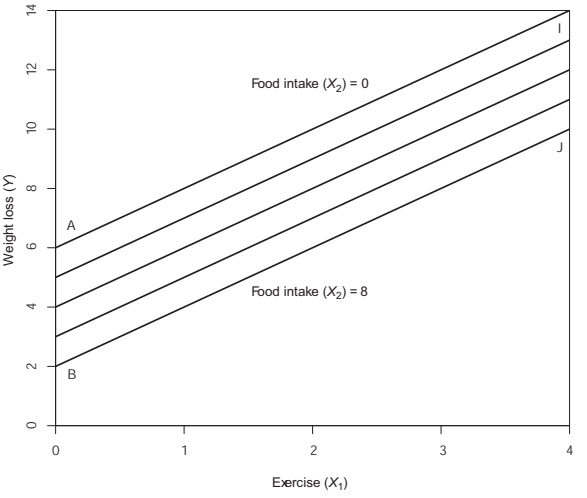


FIGURE 3.6. Still another representation of the tilted plane.

3.2 The Best-Fitting Model

3.2.1 Model Estimation with Computer Software

In practice, linear models are constructed not by piecing one together through an examination of the published literature, as we did hypothetically in the example that introduced section 3.1. Instead, linear models are constructed by using a data set in which measurements on the outcome and regressors are available and then subjecting that data to a linear regression analysis. In the case of a two-regressor model, the model is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

It is possible to derive b_0, b_1, and b_2 by hand with some formulas that are not too complex and that we provide in section 3.4.5, but we don't recommend doing so. Instead, the computations required to derive the model are almost always done by computer. A computer algorithm figures out the values of b_0, b_1, and b_2 such that the resulting model best fits the data. A linear regression analysis using the least squares criterion defines the best-fitting model as the one that minimizes the sum of the squared residuals as first introduced in section 2.1.3:

$$SS_{residual} = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} e_i^2$$

Although in that section the model had only a single regressor, the least squares criterion works regardless of the number of regressors. No modification to the math or procedure is necessary.

The sum of the squared residuals is a measure of consistency between the model and the data. A perfectly fitting model has $SS_{residual} = 0$, which occurs only when $\hat{Y} = Y$ for every case in the data. More typically, $SS_{residual}$ is greater than zero. The larger $SS_{residual}$, the less consistent the model is with the data, and the worse the fit of the model.

Figure 3.7 contains SPSS output from a linear regression analysis of the data in Table 3.1. The command to generate this output is

```
regression/statistics defaults zpp/dep=wtloss/method=enter exercise
   food.
```

which is similar to the command on page 46 but includes a few additional options to generate output that will be discussed later in this chapter and elsewhere. Comparable commands in SAS and STATA would be

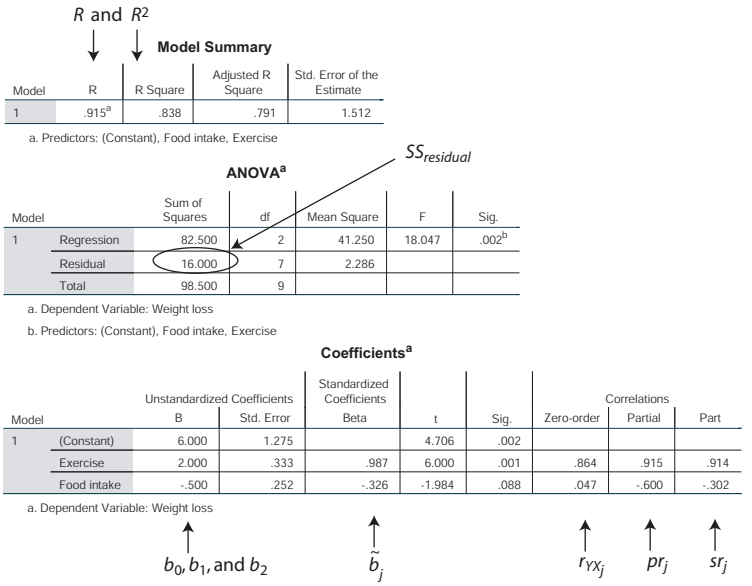


FIGURE 3.7. SPSS output from a multiple regression analysis of the weight-loss data.

```
proc reg data=exercise;model wtloss=exercise food/stb pcorr2
scorr2;run;
```

```
regress wtloss exercise food, beta
pcorr wtloss exercise food
```

which will produce output similar in content to Figure 3.7, although it will be formatted differently. We describe some but not all of the contents of this output in this chapter. By the time you get to the end of this book, you will understand what everything in this output means.

As can be seen in Figure 3.7 under the section of output labeled "Unstandardized Coefficients," the best-fitting OLS regression model is $\hat{Y} = 6 + 2X_1 - 0.5X_2$. It is no coincidence that this is the model we used in the prior section, as we made that model up knowing it would correspond to what a regression analysis would produce.

In the ANOVA summary in the middle of the output can be found $SS_{residual}$, which is 16. No values of b_0, b_1, and b_2 would produce a smaller $SS_{residual}$ than this. To verify yourself that this sum of the squared residuals

```
Dependent Variable
wtloss

Sample size
         10

Complete Model Regression Summary
          R       R-sq    Adj R-sq          F          p     SEofEst
      .9152      .8376      .7912    18.0469      .0017     1.5119

ANOVA summary table
                 SS         df         MS
Regress     82.5000     2.0000    41.2500
Residual    16.0000     7.0000     2.2857
Total       98.5000     9.0000    10.9444

Regression Model
             Coeff         se          t          p       LLCI       ULCI
constant     6.0000     1.2749     4.7062      .0022     2.9834     9.0166
exercise     2.0000      .3333     6.0000      .0005     1.2113     2.7887
food         -.5000      .2520    -1.9843      .0876    -1.0962      .0962

Simple (r), semipartial (sr), partial (pr) correlations with outcome
and standardized regression coefficients (stand)
                  r          sr          pr       stand
exercise      .8638      .9140      .9150      .9872
food          .0466     -.3023     -.6000     -.3265

******************** ANALYSIS NOTES AND WARNINGS ************************

NOTE: Level of confidence for confidence intervals:
 95.00
```

FIGURE 3.8. RLM macro output from SPSS for a multiple regression analysis of the weight-loss data.

is in fact 16, take a look at the errors in estimate in Table 3.2 (e_i) that result when the model is applied to the 10 cases' X_1 and X_2 values. Observe that when you square these 10 values of e_i and then add them up, they do sum to 16: $1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 4 + 1 = 16$.

In Appendix A we describe and document a macro for SPSS and SAS that conducts regression analysis. The SPSS version of the RLM command to conduct this analysis is

```
rlm y=wtloss/x=exercise food/stand=1.
```

The comparable command for the SAS version of RLM is

```
%rlm (data=exercise,y=wtloss,x=exercise food,stand=1);
```

(Note that an RLM command won't work without first running the RLM program that defines the macro; see Appendix A.) The output from this

RLM command can be found in Figure 3.8. As can be seen comparing the RLM output to Figure 3.7, RLM produces much of the same information as does SPSS's regression routine (as well as the regression routines in SAS and STATA).

There is no reason to use RLM for straightforward regression problems such as this. But RLM has some features and options that make some analyses easier than what comes built into SPSS's and SAS's regression procedures, and some of those features in RLM can do things that SPSS or SAS can't do at all. We describe some of these features when appropriate throughout this book.

3.2.2 Partial Regression Coefficients

The regression coefficients b_1 and b_2 are known as *partial regression coefficients*, or *partial regression slopes*. They quantify the relationship between Y and each regressor while holding all other regressors constant. More formally, b_1 quantifies the amount two cases that differ by 1 unit on X_1 are estimated to differ on Y when X_2 is held constant. Similarly, b_2 quantifies the amount two cases that differ by 1 unit on X_2 are estimated to differ on Y when X_1 is held constant.

The partial regression coefficient is one of several measures of *partial association* that one can quantify with linear regression analysis. In this section we demonstrate what it means to hold a variable constant mathematically through the process of *partialing* one variable out of another. As you will see, we can generate b_1 and b_2 without regressing Y on X_1 and X_2 simultaneously. In practice you will not have to engage in the partialing process we describe here, as it is done automatically by a computer using matrix algebra when a linear regression analysis is conducted (see Appendix D for the details). But it will help you to understand what the partialing process entails.

If X_1 and X_2 are uncorrelated, no partialing is required to estimate b_1 and b_2. In that case, you can estimate b_1 by regressing Y on X_1 alone. The regression coefficient for X_1 in this model is b_1 in the model with two regressors X_1 and X_2. By the same token, b_2 can be estimated by regressing Y on X_2 alone.

More typically, when you conduct a regression analysis the regressors are correlated to some degree, although some regressors may be more correlated with others in models with several regressors. The two regression coefficients for X_1 and X_2 can still be generated by regressing Y on X_1 and X_2 separately, but only after X_1 and X_2 have been partialed of their rela-

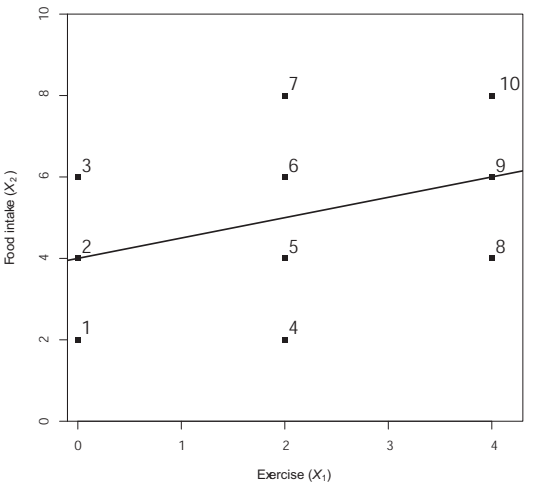


FIGURE 3.9. A scatterplot with the regression line estimating food intake (X_2) from exercise (X_1).

tionship with each other. This involves constructing new measures of X_1 and X_2 that are independent of the other.

Consider a regression model that ignores Y, in which X_2 is predicted from X_1. We'll call a regression estimating one regressor from the other regressor(s) in the model a *crosswise regression*. So imagine a crosswise regression estimating X_2 from X_1. This model will generate estimates of X_2 from X_1. Figure 3.9 depicts this for the weight-loss example in the form of a scatterplot of X_2 against X_1, along with the best-fitting OLS regression line.

The equation for this regression line is $\hat{X}_2 = 4 + 0.5X_1$. This model of X_2 generates estimates of X_2 given information about a case's value of X_1. From these estimates we could construct residuals for each case, defined as $X_2 - \hat{X}_2$. Call these residuals $X_{2.1}$. These residuals correspond to the vertical distance between each point in the scatterplot and the regression line from the model estimating X_2 from X_1. Each case's residual in the weight-loss data for this crosswise regression can be found in Table 3.3.

Consider the third person (ID = 3) in Table 3.3. This person did not exercise at all ($X_1 = 0$). As can be seen in Figure 3.9, the crosswise regression estimates this person's food intake to be $\hat{X}_2 = 4 + 0.5 \times 0 = 4$, or 400 calories above the minimum recommended. But this person actually consumed 6

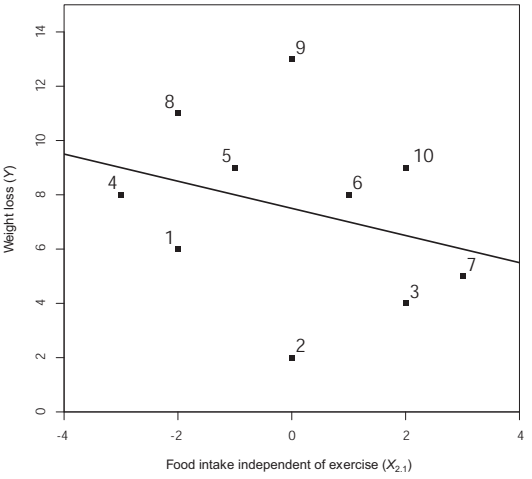


FIGURE 3.10. The regression coefficient b_2 is the slope of the line.

units above (600 calories), so this person's residual is $X_{2.1} = X_2 - \hat{X}_2 = 6 - 4 = 2$. This is the vertical distance between the point in Figure 3.9 labeled "3" and the regression line. So case 3 consumed 200 calories more than would be expected given the relationship between exercise and food intake. Using the same reasoning, the eighth person (ID = 8) consumed 2 units *fewer* (200 calories less) than would be expected given the relationship between exercise and food intake. For case 8, $X_{2.1} = X_2 - \hat{X}_2 = 4 - (4 + 0.5 \times 4) = -2$. This is the vertical distance between the point in Figure 3.9 labeled "8" and the regression line.

The residuals from a regression model are linearly uncorrelated with all regressors in the model, as first discussed in section 2.4. So the correlation between $X_{2.1}$ and X_1 is exactly zero. Verify this for yourself by correlating $X_{2.1}$ and X_1 using the data in Table 3.3. Thus, we can call $X_{2.1}$ the component of X_2 that is *independent* of X_1. Rephrased, the residuals from this crosswise regression, $X_{2.1}$, quantify the part of X_2 that is *unique* to X_2, meaning that it cannot be explained by differences between people in X_1. So we can think of $X_{2.1}$ as a new measure of food intake that provides unique information about individual differences in food intake relative to the information about food intake that could be predicted from individual differences in exercise frequency.

TABLE 3.3. Data Matrix with Some Residuals

ID	Exercise X_1	Food intake X_2	Weight loss Y	$X_{1.2}$	$X_{2.1}$	$Y.1$	$Y.2$
1	0	2	6	−1.14	−2.00	2.00	−1.29
2	0	4	2	−1.71	0.00	−2.00	−5.43
3	0	6	4	−2.29	2.00	0.00	−3.57
4	2	2	8	0.86	−3.00	0.50	0.71
5	2	4	9	0.29	−1.00	1.50	1.57
6	2	6	8	−0.29	1.00	0.50	0.43
7	2	8	5	−0.86	3.00	−2.50	−2.71
8	4	4	11	2.29	−2.00	0.00	3.57
9	4	6	13	1.71	0.00	2.00	5.43
10	4	8	9	1.14	2.00	−2.00	1.29
Mean	2	5	7.5	0.00	0.00	0.00	0.00

Now consider a regression model estimating Y from $X_{2.1}$. That is, let's regress Y on $X_{2.1}$ to generate a model predicting how much weight a person loses from information about his or her food intake that is independent of the amount he or she exercises. Figure 3.10 depicts the association between Y and $X_{2.1}$, along with the regression line estimating Y from $X_{2.1}$. The equation for this line is $\hat{Y} = 7.5 - 0.5X_{2.1}$, which you can verify for yourself from the data in Table 3.3. It is not a coincidence that the regression coefficient for $X_{2.1}$ of −0.5 corresponds exactly to b_2 in the regression model estimating Y from both X_1 and X_2. We can say that *controlling for* exercise frequency, or *accounting for differences between people in exercise frequency*, or *holding exercise frequency constant*, two people that differ by 1 unit (100 calories) in food intake are estimated to differ by 0.5 units (50 grams) in weight loss. The negative sign means that the person who takes in 1 more unit of food loses 0.5 units *less* weight.

The same procedure can be used to generate b_1. In a crosswise regression, estimate X_1 from X_2 and construct the residuals $X_{1.2} = \hat{X}_1 - X_1$. Figure 3.11 depicts this crosswise regression, along with the regression line estimating X_1 from X_2. Cases above the line have positive values of $X_{1.2}$, meaning they exercise more than expected given the relationship between food intake and exercise. Cases below the line have negative values of $X_{1.2}$, meaning they exercise less than expected given their food intake. These

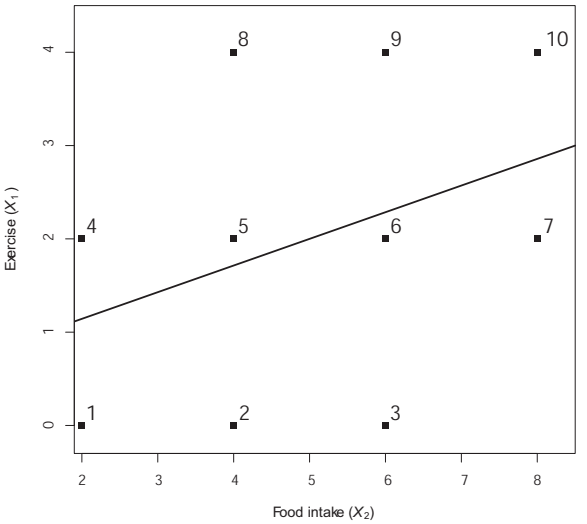


FIGURE 3.11. A scatterplot with the regression line estimating exercise (X_1) from food intake (X_2).

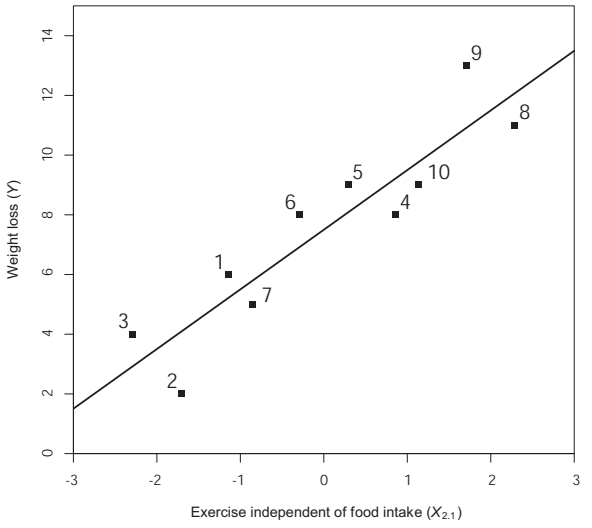


FIGURE 3.12. The regression coefficient b_1 is the slope of the line.

residuals quantify the component of X_1 that is linearly independent of X_2 or unique to X_1.

Now regress Y on $X_{1.2}$, as depicted in Figure 3.12. This generates a model of weight loss from the part of exercise frequency that is independent of food intake. The regression line is $7.5 + 2.0X_{1.2}$. The regression coefficient of 2.0 is the slope of the line in Figure 3.12, and it corresponds to b_1 in the regression model estimating Y from X_1 and X_2 simultaneously. We can say that *controlling for food intake*, *accounting for differences between people in food intake*, or *holding food intake constant*, two people that differ by 1 unit (1 hour) in exercise frequency are estimated to differ by 2 units (200 grams) in weight loss. The positive sign means that the person who exercises more loses *more* weight.

So the partial regression coefficient b_1 in the model estimating Y from X_1 and X_2 simultaneously corresponds to the regression weight estimating Y from the component of X_1 that is unique to X_1. And b_2 in the model corresponds to the regression weight estimating Y from the component of X_2 that is unique to X_2.

3.2.3 The Regression Constant

The regression constant b_0 is interpreted as the estimated value of Y when all regressors are set to zero. It is chosen so that the mean of the estimates of Y correspond to the mean of Y (i.e., such that $\overline{\hat{Y}} = \overline{Y}$). This prevents the estimated values of Y from either consistently exceeding or consistently falling below the actual values of Y. When b_1 and b_2 are known, b_0 can be found from

$$b_0 = \overline{Y} - (b_1\overline{X}_1 + b_2\overline{X}_2) \tag{3.1}$$

In this example, $\overline{Y} = 7.5$, $b_1 = 2$, $\overline{X}_1 = 2$, $b_2 = -0.5$, $\overline{X}_2 = 5$, so $b_0 = 7.5 - (2 \times 2 - 0.5 \times 5) = 7.5 - 1.5 = 6$, which corresponds to the value generated by the OLS regression output from SPSS in Figure 3.7.

If you think about what a good model should do, then equation 3.1 makes sense. If you knew someone was average on the variables used to generate $\hat{Y}$, then you'd expect the model to estimate that he or she would be average on Y as well. Furthermore, you'd expect that a sensible model should generate estimates of Y that, on average, equal $\overline{Y}$. The regression constant ensures the resulting model has these properties.

3.2.4 Problems with Three or More Regressors

We very often want to quantify the relationship between an independent variable and a dependent variable when controlling for or holding constant several covariates. For instance, we might want to examine the relationship between the success in school of adopted children and the school performance of their biological mothers while holding constant several measures of their adoptive environment—school quality, education of the adoptive parents, number of books in the home, and similar variables. This section describes the derivation of the partial regression coefficients for a problem such as this. As before, these are not actual computing directions; their purpose is to help you see the meaning of the statistics computed.

Let k be the number of regressors in a regression model. Our goal is to estimate Y from k regressors using linear regression, which will yield a model of Y that takes the form

$$\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \\ &= b_0 + \sum_{j=1}^{k} b_jX_j\end{aligned}$$

In the previous section, weight loss was regressed on (modeled as a linear function of) weekly hours of exercise (X_1) and food intake (X_2). We add a new regressor: metabolic rate (X_3). Measurements on this variable can be found in Table 3.4. Using these data, we estimate

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \tag{3.2}$$

Consider the partial relationship between Y and X_3, with X_1 and X_2 held constant. We want to know the relationship between Y and the part of X_3 that is unique to it, meaning uncorrelated with X_1 and X_2. Recognizing the residuals from a regression are uncorrelated with all regressors in the model that generates the residuals, we can easily construct a measure of X_3 that is uncorrelated with both X_1 and X_2. In section 3.2.2 we defined a crosswise regression as a regression estimating one regressor from the other regressors in the model. In this case, the crosswise regression estimating X_3 from *both* X_1 and X_2 results in a set of residuals that we denote $X_{3.12}$. The residuals from this regression, $X_{3.12}$, can be found in Table 3.4. You can verify for yourself that $X_{3.12}$ is uncorrelated with both X_1 and X_2.

A scatterplot of Y against $X_{3.12}$ can be found in Figure 3.13, with the regression line estimating Y from $X_{3.12}$ superimposed on the scatterplot.

TABLE 3.4. Data Matrix Including Metabolism

ID	Exercise X_1	Food intake X_2	Metabolism X_3	Weight loss Y	$X_{1.23}$	$X_{2.13}$	$X_{3.12}$
1	0	2	15	6	−0.82	−1.38	1.00
2	0	4	14	2	0.35	1.24	−2.00
3	0	6	19	4	−1.19	0.14	1.00
4	2	2	15	8	1.18	0.10	−2.00
5	2	4	21	9	−0.81	−1.62	2.00
6	2	6	23	8	−1.00	−0.86	2.00
7	2	8	21	5	0.63	2.38	−2.00
8	4	4	22	11	0.74	−0.76	0.00
9	4	6	24	13	0.55	0.00	0.00
10	4	8	26	9	0.37	0.76	0.00
Mean	2	5	20	7.5	0.00	0.00	0.00

The slope of this line is b_3 in equation 3.2. In this analysis, $b_3 = 0.636$. So we can say that two people who consume the same number of calories and who exercise the same amount (i.e., holding food intake and exercise constant, or controlling for food intake and exercise) but who differ by 1 unit in metabolism are estimated to differ by 0.636 units in weight loss. The positive sign for b_3 means that the person with higher metabolism is estimated to lose more weight, which is what you would expect.

We repeat this process to generate b_1 and b_2. To find b_1, estimate the crosswise regression of X_1 on X_2 and X_3. This produces the residuals denoted $X_{1.23}$ in Table 3.4. Regressing Y on these residuals $X_{1.23}$ yields a regression weight of 1.046 for $X_{1.23}$. This is b_1 in equation 3.2. So two people who consume the same number of calories and who have the same metabolic rate but who differ by 1 hour in exercise frequency are estimated to differ by 1.046 units of weight loss, with the positive sign denoting that the person who exercises more loses more weight.

Repeat this process to estimate b_2 by regressing X_2 on X_1 and X_3 to produce $X_{2.13}$ (see Table 3.4). Regressing Y on $X_{2.13}$ yields $b_2 = -1.136$. Two people who exercise the same amount and with the same metabolic rate (i.e., holding constant or controlling for these two variables) but who differ by 1 unit in food intake are estimated to differ by 1.136 units in weight loss, with the negative sign denoting that the person who consumes more loses *less* weight.

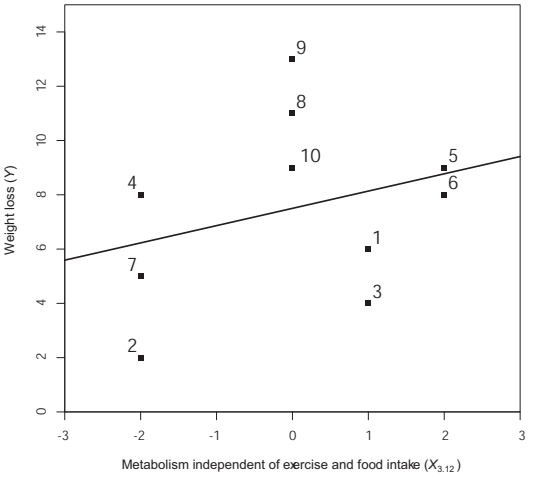


FIGURE 3.13. The regression coefficient b_3 is the slope of the line.

With the regression coefficients for X_1, X_2, and X_3 derived, the regression constant can be found. This is a direct extension of the approach used in a model with only two regressors, and it generalizes to any number of regressors. For the three regressor case,

$$b_0 = \overline{Y} - (b_1\overline{X}_1 + b_2\overline{X}_2 + b_3\overline{X}_3)$$

Inserting b_1, b_2, and b_3 as well as $\overline{X}_1 = 2$, $\overline{X}_2 = 5$, $\overline{X}_3 = 20$, and $\overline{Y} = 7.5$ into this formula yields

$$b_0 = 7.5 - [1.046(2) - 1.136(5) + 0.636(20)] = -1.634$$

So the final regression model is

$$\hat{Y} = -1.634 + 1.046X_1 - 1.136X_2 + 0.636X_3$$

This model minimizes the sum of the squared residuals, which is 7.091 here (you can verify this yourself by constructing the residuals from this model, squaring them, and adding them up). We cannot easily represent this model visually. With two regressors, the model appears in three dimensions as a

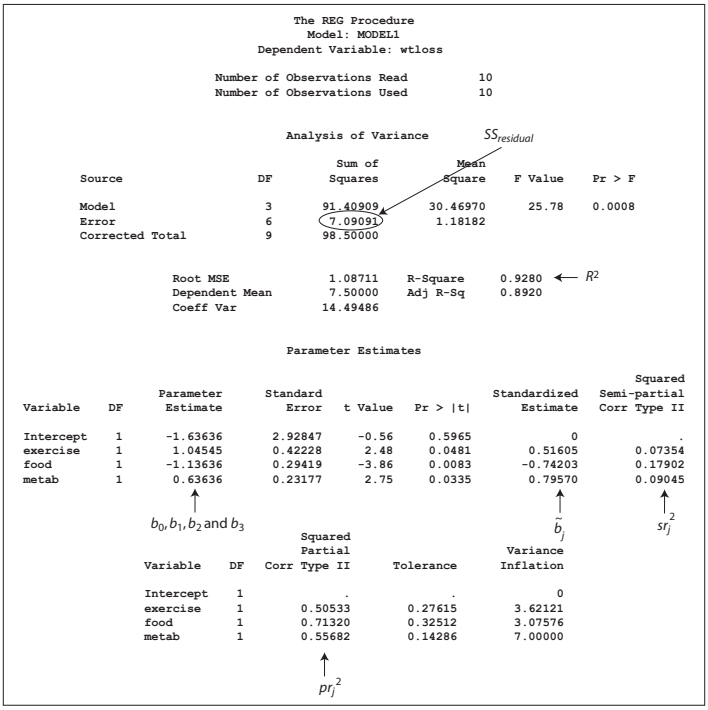


FIGURE 3.14. SAS output from a multiple regression analysis of the weight-loss data.

plane. With three regressors, the model is a "hyperplane" (a plane in space of more than three dimensions).

But as with the two regressor case, there would be no need for you to ever actually conduct these crosswise regressions, calculate the residuals, and then regress Y on these residuals to generate the regression coefficients and constant for this model. A linear regression module in any decent statistical package will find the values of b_0, b_1, b_2, and b_3 for you that minimize the sum of the squared residuals. An example output from SAS can be found in Figure 3.14. Observe that it produces the same regression constant and regression coefficients, and $SS_{residual}$ is 7.091. No other combination of values of b_0, b_1, b_2, and b_3 generate a smaller sum of squared residuals.

3.2.5 The Multiple Correlation *R*

The correlation between $\hat{Y}$ from a regression model and Y, the actual values in the data, is called the *multiple correlation coefficient* and denoted R. R is reported by almost all regression programs. R is frequently used as a measure of model fit, as it quantifies the correspondence between what the model estimates for Y and the actual values of Y it is attempting to estimate. So larger values of R correspond to better fit, unlike $SS_{residual}$, where smaller values reflect better fit.

As can be seen in Figure 3.7, $R = 0.915$ for the model estimating weight loss from food intake and exercise. When metabolism is added to the model, $R = 0.963$. These are exceptionally large values of R—higher than you would typically find in research. But it makes sense in this case, as you'd expect a combination of metabolism, the amount people eat, and how much they exercise during a given period of time would predict very well the amount of weight they lose during that period of time. Furthermore, all other things being equal, R tends to be big in very small samples such as this one. We shall see in Chapter 4 that R tends to exaggerate the fit between model and data, especially in small samples. Better measures of overall fit are introduced there.

R is never negative and would rarely be zero. It is almost always positive, but it cannot exceed 1. To understand why, consider a scatterplot showing a negative correlation between X and Y, as in Figure 3.15, panel A. In this sample there is perfect linearity, so the regression line goes through the three conditional means. Thus, $\hat{Y} = 4$ for the two cases on the left, 3 for the three cases in the middle, and 2 for the two on the right. But notice that if you put Y on the X-axis and $\hat{Y}$ on the Y-axis, as in Figure 3.15, panel B, the correlation between Y and $\hat{Y}$ is positive. Imagine the usefulness of a modeling procedure that would allow the correlation between what the model estimates and reality to be negative. Such a procedure would never be used. At its worst, a regression model will produce estimates of Y that are uncorrelated with Y, in which case $R = 0$. But as a Pearson correlation (between $\hat{Y}$ and Y), R cannot exceed 1 because Pearson's r can't exceed 1. At its best, R equals 1 if and only if $Y = \hat{Y}$ for every case in the data.

R also never falls below the absolute value of any regressor's simple correlation with Y. That is, $R \geq |r_{YX_j}|$ for all j. This is because R is the correlation between Y and a weighted sum of regressors ($\hat{Y}$) constructed such that it is maximally correlated with Y. But X_j is a weighted sum of regressors in which X_j is weighted 1 and all other regressors are weighted

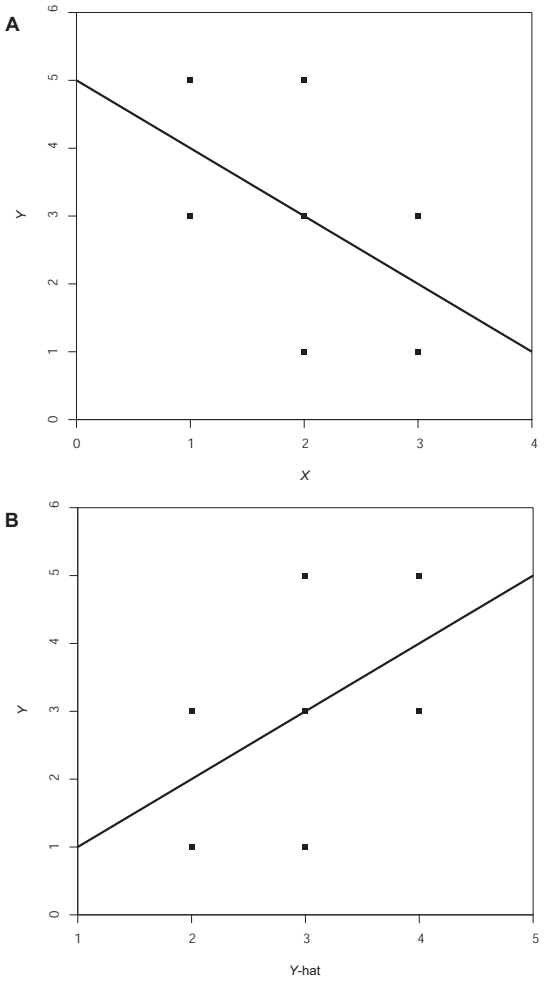


FIGURE 3.15. *R* is positive even though r_{XY} is negative.

zero. So X_j is one possible weighted sum of the regressors, and the absolute value of its correlation with Y could not exceed R.

If a regressor X_j makes absolutely no independent contribution to the prediction of Y beyond what the other regressors in the model contribute, then we will find that $b_j = 0$ and R will be the same whether X_j is in the model or not. But since the regression procedure will set b_j to zero if doing so minimizes $SS_{residual}$, then adding a variable to a regression model could *never* lower R.

3.3 Scale-Free Measures of Partial Association

As discussed in section 3.1.3, the partial regression coefficients b_j are scale-bound. Changing the units of measurement of a regressor in a model by multiplying or dividing all the measurements by a constant will change the regression coefficient for that regressor both in absolute terms and relative to other regression coefficients. This section introduces various measures of partial association that are scale-free. Their values will not change merely by multiplying or dividing a regressor by a constant.

3.3.1 Semipartial Correlation

We began this chapter by illustrating a regression model with two regressors in which Y (weight loss) was regressed on X_1 (exercise) and X_2 (food intake). We said earlier that the slopes of regression lines in Figures 3.10 and 3.12 are the partial regression coefficients for X_2 and X_1 in that model, respectively. For reasons that will soon become clear, these scatterplots are called *semipartial scatterplots*, and the correlations between Y and $X_{1.2}$, and between Y and $X_{2.1}$ are called *semipartial correlations*, although some refer to these as *part correlations*. These correlations are denoted sr_1 and sr_2. In this example, $sr_1 = 0.914$ and $sr_2 = -0.302$. More generally, we use sr_j to denote the semipartial correlation for regressor j.

Most regression programs will not produce the semipartial correlations in the output by default, but some regression programs will give them to you if you ask for them. The commands starting on page 55 include options to request the semipartial correlations as well as the partial correlations, described next. They can be found in the outputs in Figures 3.7 and 3.14. Observe that SAS reports sr_j^2 rather than sr_j. The squaring removes the sign, but the sign of sr_j is always the sign of b_j.

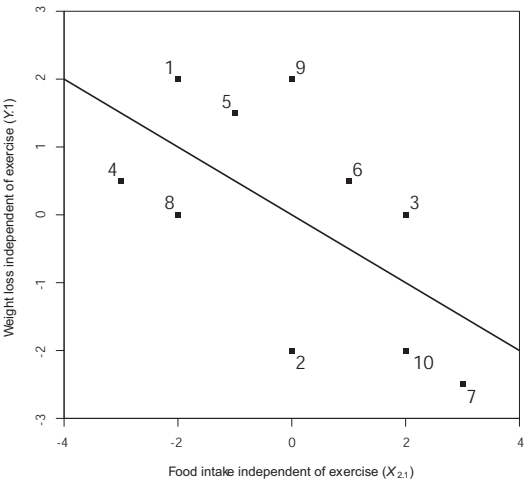


FIGURE 3.16. A partial scatterplot of the relationship between Y and X_2 controlling for X_1.

3.3.2 Partial Correlation

In section 2.4 we discussed the residuals in a model with a single regressor at length. In this chapter, thus far we have discussed residuals primarily in the context of a crosswise regression, in which a regressor is estimated from another regressor or set of regressors. That now changes. Consider that your focus is on quantifying partial association for two regressors X_1 and X_2 in a model of Y that includes only those two regressors, such as the model described in section 3.1. Let $Y.1$ denote the residuals from a model estimating weight loss (Y) from exercise frequency (X_1) alone, and let $Y.2$ denote the residuals from a model estimating weight loss from food intake (X_2) alone. Knowing that residuals are uncorrelated with all regressors in the model, $Y.1$ will be uncorrelated with X_1, and $Y.2$ will be uncorrelated with X_2. So $Y.1$ and $Y.2$ are new measures of weight loss that have been partialled out of their relationship with exercise frequency and food intake, respectively. These two sets of residuals can be found in Table 3.3.

Figure 3.16 shows a scatterplot of $Y.1$ against $X_{2.1}$. Recall that $X_{2.1}$ is the residual from a crosswise regression estimating X_2 from X_1. This figure is similar to but not the same as Figure 3.10, which plotted Y against $X_{2.1}$.

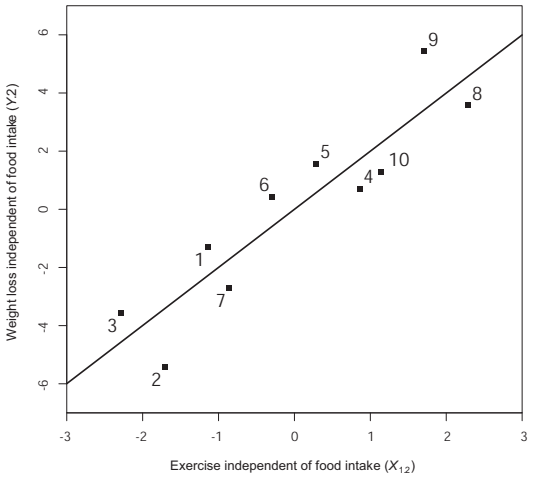


FIGURE 3.17. A partial scatterplot of the relationship between Y and X_1 controlling for X_2.

In words, Figure 3.10 plots *all* of Y against part of X_2, whereas Figure 3.16 plots part of Y—the part independent of X_1—against the the part of X_2 that is independent of X_1. Therefore, we call Figure 3.16 a *partial scatterplot* and Figure 3.10 a *semipartial* scatterplot. Figure 3.17 is also a partial scatterplot, but with the roles of X_1 and X_2 reversed compared to Figure 3.16. It plots the part of Y independent of X_2 against the part of X_1 independent of X_2.

The correlation between $Y.1$ and $X_{2.1}$ depicted in the scatterplot in Figure 3.16 is called the *partial correlation* between Y and X_2 controlling for X_1. It is the correlation between these two sets of residuals. Similarly, the correlation between the $Y.2$ and $X_{1.2}$ residuals in Figure 3.17 is the partial correlation between Y and X_1 controlling for X_2. These are denoted pr_2 and pr_1, respectively. In this example, $pr_1 = 0.915$ and $pr_2 = -0.600$. In general, we denote the partial correlation between Y and X_j controlling for all other regressors in the model pr_j. The partial regression weight b_j, the partial correlation pr_j, and the semipartial correlation sr_j always have the same sign, though they usually have different values.

It can be shown that the regression slope in a partial scatterplot always equals the slope in the corresponding semipartial scatterplot. That is, the slope of the line in Figure 3.16 is the same as the slope of the line in Figure

3.10. The same is true for the slopes in Figures 3.17 and 3.12. So we don't need to distinguish between partial slopes and semipartial slopes. We just call them partial regression slopes or partial regression weights. These are the same as the partial regression weight for X_j in a model that includes X_j and all the other regressors being controlled when the measures of partial association are constructed.

The last sentence of the prior paragraph is important. Values of sr_j and pr_j are always defined in terms of all the variables in the model. So if X_3 were included in the model along with X_1 and X_2, then sr_1 is defined as the correlation between Y and $X_{1.23}$, the portion of X_1 independent of both X_2 and X_3, and pr_1 is the correlation between $Y.23$ and $X_{1.23}$, the portions of Y and X_1 independent of X_2 and X_3.

The partial correlation has three interpretations, depending on whether it is squared or not. When unsquared, it is interpreted as an estimate of the correlation between Y and X_j when all other regressors are held constant. So it can be thought of as a correlation between Y and X_j that has been "corrected for" their shared association with other regressors. When squared it can be interpreted as the proportion by which the variance of the residuals shrinks when X_j is added to the model. Its squared value is also interpreted as the proportion of the variance in Y not explained by the other regressors in the model that can be explained by regressor X_j. We further discuss the interpretation of the semipartial and partial correlations in section 3.4.1, and we offer a fourth interpretation in section 8.3.

Like the semipartial correlation, the partial correlation for each variable in a model is not usually generated automatically by a regression program, but many will print them if you ask. We did so when Figures 3.7 and 3.14 were generated in SPSS and SAS, and you will find the partial correlations (in SPSS) or squared partial correlations (in SAS) in these outputs as a result.

3.3.3 The Standardized Regression Coefficient

A third measure of partial association is the standardized partial regression coefficient, which we denote $\tilde{b}_j$ to distinguish it from the *un*standardized regression coefficient b_j. In section 2.3 we introduced the standardized regression coefficient as the regression coefficient for X when you estimate Y from X after first standardizing both. In a model with more than one regressor, the standardized partial regression coefficient for regressor j is the regression coefficient for X_j when Y and X_j are standardized prior to

running the regression. But you can calculate it without actually doing this regression on standardized variables by using

$$\tilde{b}_j = b_j \frac{s_{X_j}}{s_Y}$$

where s_{X_j} and s_Y are the standard deviations of X_j and Y, respectively. For instance, in the two-regressor example we have been focusing on, $b_1 = 2$, $s_{X_1} = 1.549$, and $s_Y = 3.138$, so $\tilde{b}_1 = 0.987$. You can verify for yourself that $\tilde{b}_2 = -0.326$.

The standardized partial regression coefficient is interpreted just as is the unstandardized regression coefficient b_j, but the metric of discussion is standard deviations of X_j and Y rather than their original metrics. That is, two cases that differ by *one standard deviation* on X_j are estimated to differ by $\tilde{b}_j$ *standard deviations* on Y, holding all other regressors constant. Given $\tilde{b}_1 = 0.987$, we can say that two people who differ by one standard deviation in exercise frequency but who consume the same amount are estimated to differ by 0.987 standard deviations in weight loss, with the positive sign meaning that the person who exercises more is estimated to lose more weight. The sign of $\tilde{b}_j$ will always be the same as the signs of b_j, sr_j, and pr_j.

Many people use $\tilde{b}_j$ as a measure of the relative importance of a regressor in a model. That is, if b_1 is larger than b_2, then some would say that X_1 is more important in a statistical sense than is X_2 in estimating Y. But as we discuss in Chapter 8, this isn't necessarily true, and we prefer the semipartial correlation as a measure of relative importance. In addition, many people use β or spell out "beta" when talking or writing about the standardized regression coefficient. As noted in Chapter 1, β is used in many ways in statistics as well as in regression analysis, so we don't follow this semiconvention, because doing so invites confusion.

If the investigator chooses the values of a regressor or manipulates them experimentally rather than observes them naturally, then the standardized regression coefficient does not quantify anything that generalizes to the natural world. This is because the standard deviation of X_j and therefore the value of $\tilde{b}_j$ will be determined by the choice the investigator makes about the values of this regressor. Even if X_j is a nominal and dichotomous variable, such as whether a person is exposed to a message or not in an experiment in order to assess the effects of the message on something like attitudes, the value of $\tilde{b}_j$ will be determined in part by the distribution of the cases between the two groups. We discuss this in more detail in section

5.1.5. For this reason, we discourage the use of standardized regression coefficients as a measure of partial association in these circumstances.

Most regression programs will produce the standardized partial regression coefficients in output, although they won't all do so automatically. SPSS will print them whether you want them or not, whereas SAS and STATA require that you request them. They can be found in Figures 3.7 and 3.14 for the two models we have been considering in this chapter.

3.4 Some Relations among Statistics

3.4.1 Relations among Simple, Multiple, Partial, and Semipartial Correlations

We have now discussed three scale-free measures of partial relationship: pr_j, sr_j, and $\tilde{b}_j$. The three measures are often but not always similar numerically. In the weight-loss example, $sr_1 = 0.914$, $pr_1 = 0.915$, $\tilde{b}_1 = 0.987$, and $sr_2 = -0.302$, $pr_2 = -0.600$, $\tilde{b}_2 = -0.327$. The proper use of each has been a matter of some debate and confusion. We offer some of our own opinions on this topic in Chapter 8. For now, suffice it to say that each has certain good uses.

In Chapter 2 we learned that r^2_{XY} is the proportion of the variance in Y explained by X, and $1 - r^2_{XY}$ is the proportion unexplained. In similar terms, R^2 is the proportion of the variance in Y explained by the model as a whole—meaning the entire set of regressors. We will make this more explicit in Chapter 4 when introducing the regression sum of squares and total sum of squares. If R^2 is the proportion of the total variance in Y explained by the entire set of regressors, we can define a single regressor's *unique contribution* to the variance in Y explained as the amount that R^2 would drop if that regressor were removed from the model. Alternatively, it can be defined equivalently as the amount that R^2 increases when it is added to the model.

This difference in R^2 that results when regressor j is added or removed from a model is equivalent to sr^2_j, its squared semipartial correlation. In the two-regressor example modeling weight loss as a function of exercise frequency and food intake, $R^2 = 0.837$. If food intake (X_2) were dropped from the model, then only X_1—exercise frequency—would remain. In that case, $R^2 = r^2_{YX_1} = 0.746$. Thus, the contribution of X_2 to variance explained in Y is $0.837 - 0.746 = 0.091$, which is indeed the square of X_2's semipartial correlation: $sr^2_2 = 0.302^2 = 0.091$. Rephrased, the proportion of the variance

explained in weight loss increases by 0.091 when food intake is added to a model that includes only exercise frequency. Using the same computations, you will find that when exercise frequency is added to a model of weight loss that includes only food intake, R^2 increases from 0.002 to 0.837. This increase is equivalent to the squared semipartial correlation for exercise frequency: $sr_1^2 = 0.914^2 = 0.835$.

So sr_j^2 quantifies how much of the variance in Y is uniquely explained by X_j. It might be more useful to redefine sr_j^2 as the proportion of the *total* variance in Y uniquely explained by X_j to distinguish it from pr_j^2, which quantifies the proportion of the variance in Y unexplained by all the other regressors in the model that can be uniquely explained by X_j. The difference in language is subtle but important. Consider the model of weight loss that includes only exercise frequency. For this model, $R^2 = 0.746$. So the proportion of the variance in Y *not* explained by exercise frequency is $1 - 0.746 = 0.254$. We know that when food intake is added to the model, R^2 increases by 0.091, which is sr_2^2, to 0.837. So the proportion of the variance in Y that was not accounted for by exercise frequency that is accounted for by food intake is $0.091/0.254 = 0.360$. But this is the square of food intake's partial correlation: $pr_2^2 = -0.600^2 = 0.360$. Using the same logic, we can say that the proportion of the variance in weight loss not explained by food intake that is explained by exercise frequency is $pr_1^2 = 0.915^2 = 0.837$.

So sr_j^2 and pr_j^2 are both measures of the proportion of the variance in Y uniquely explained by regressor j, but they differ with respect to the reference point used for calculating the contribution. Whereas sr_j^2 gauges X_j's unique contribution to explaining variance relative to *all* of the variance in Y, pr_j^2 indexes X_j's contribution relative to the part of the variance in Y that remains unexplained by all the other regressors in the model.

Define R_k^2 as the squared multiple correlation estimating Y from all k regressors in a model and $R_{(k-j)}^2$ as the squared multiple correlation estimating Y from those same k regressors *except* for regressor j. Then the following relations hold:

$$\begin{aligned} R_k^2 &= R_{(k-j)}^2 + sr_j^2 \\ sr_j^2 &= R_k^2 - R^2{}_{(k-j)} \\ pr_j^2 &= \frac{sr_j^2}{1 - R^2{}_{(k-j)}} \end{aligned}$$

From these equations, it can be derived that a variable's partial correlation with Y usually exceeds its semipartial correlation with Y in absolute value, and its semipartial correlation can never exceed its partial correlation. The former is apparent when you consider that pr_j^2 is sr_j^2 divided by a number that is never larger than 1 and almost always smaller. This means that pr_j^2 is generally more distant from zero than is sr_j^2. So ignoring sign, a variable's partial correlation is the upper bound on its semipartial correlation.

Considering the two-regressor model as a special case and letting R^2 be the squared multiple correlation for the model containing both X_1 and X_2:

$$\begin{aligned} R^2 &= r_{YX_1}^2 + sr_2^2 \\ &= r_{YX_2}^2 + sr_1^2 \end{aligned}$$

which has as a special case when $r_{X_1X_2} = 0$,

$$R^2 = r_{YX_1}^2 + r_{YX_2}^2$$

When two or more regressors correlate highly, R^2 may fall well below the sum of the squared correlations between Y and each of the regressors. Indeed, R^2 may be only slightly larger than the largest $r_{YX_j}^2$. In fact, each regressor's unique contribution may be quite small even though the regressors when treated as a set explain a substantial proportion of the variability in Y. Such regressors are said to be *collinear*.

Although the individual values of $r_{YX_j}^2$ set a lower limit on R, they set no upper limit. Artificial examples can be created in which all values of $r_{YX_j}^2$ are small or zero but R is near 1 or even equal to 1 (Hamilton, 1987). For example, consider the data in Table 3.5 from 10 schoolchildren. In these data, X_1 is skill at softball, and X_2 is skill at basketball. The correlation between skill at softball and basketball is high, as you might expect given that some kids are more athletic and dextrous than are others: $r = 0.918$. Suppose Y is the response to the question "On a scale from 1 to 9, which sport do you prefer?", where 1 = much prefer softball, 5 = no preference, and 9 = much prefer basketball. Further suppose that a child's preference is determined primarily by the difference between his or her skill in the two sports, with every child preferring the sport in which he or she excels relative to the other. In these data, the correlation between each skill and preference is negative, but only slightly so: $r_{YX_1} = -0.209$, $r_{YX_2} = -0.195$. Yet the two skill measures almost perfectly predict preference in a linear

TABLE 3.5. Skill at Softball, Basketball, and Preference

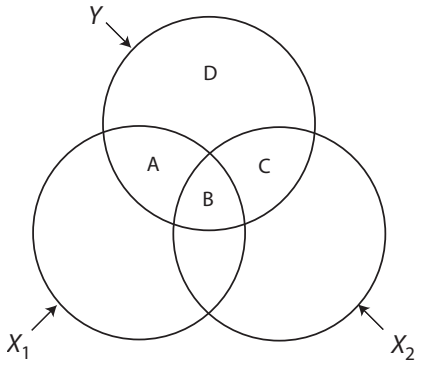
ID	Softball X_1	Basketball X_2	Preference Y
1	4	17	1
2	56	60	3
3	25	3	8
4	50	52	4
5	5	16	2
6	72	84	2
7	100	95	5
8	39	20	8
9	81	75	5
10	61	47	7

regression analysis, $R = 0.993$. To understand why, recognize that this difference in skill is one possible linear function of the two skill measures. So when sport preference is regressed onto the two skill measures, R is very high, because preference is driven by the difference in skill. The regression equation is $\hat{Y} = 3.899 + 0.198X_1 - 0.195X_2$, which is approximately $\hat{Y} = 3.899 + 0.2(X_1 - X_2)$.

We call a set of regressors *complementary* if R^2 for the set exceeds the sum of the individual values of $r^2_{YX_j}$. Thus, complementarity and collinearity are opposites, though either can occur only when regressors in a set are intercorrelated. As discussed later in section 5.3.3, either collinearity or complementarity can exist within a subset of the whole set of regressors. For instance, in a problem with several regressors, two of them could be collinear or complementary with each other but independent or nearly independent of all other regressors.

3.4.2 Venn Diagrams

In our experience teaching regression, the distinction between the partial and semipartial correlation is one of the toughest for students of regression to master. The Venn diagram in Figure 3.18 will likely help keep them straight. The Venn diagram represents relationships between variables, and the relative sizes of areas in the diagram reflect measures of squared association. Not all possible patterns of association can be depicted in a Venn diagram but many can. In the diagram in Figure 3.18 there are two



$R^2 = (A + B + C) / (A + B + C + D)$

$sr_1^2 = A / (A + B + C + D)$

$pr_1^2 = A / (A + D)$

$sr_2^2 = C / (A + B + C + D)$

$pr_2^2 = C / (C + D)$

FIGURE 3.18. A Venn diagram.

regressors, X_1 and X_2, and a dependent variable, Y. These variables are depicted here in what we call the *standard configuration*, where all three variables are correlated with each other to some degree. This is the circumstance you are most likely to encounter when conducting a regression analysis.

The total area of the Y circle in the Venn diagram is the sum of the areas labeled A, B, C, and D and can be thought of as the total variance in Y. The squared correlation between X_1 and Y is the ratio of (A + B) over (A + B + C + D). It is the proportion of the variance in Y shared with X_1. Likewise, the squared correlation between X_2 and Y is (B + C) over (A + B + C + D). And if Y were regressed on X_1 and X_2, then R^2 is the proportion of the variance in Y explained by X_1 and X_2, and it corresponds to the ratio of (A + B + C) over all of Y. Thus, $R^2 = (A + B + C)/(A + B + C + D)$.

With ratios of areas in a Venn diagram reflecting squared association, sr_1^2—the proportion of the variance in Y uniquely explained by X_1—is the proportion of the variance in Y (or A + B + C + D) shared only with X_1 (or A in the diagram). Thus, sr_1^2 is A/(A + B + C + D). Similarly, sr_2^2 is the proportion of the variance in Y shared only with X_2, or C/(A + B + C + D).

In contrast, the squared partial correlation between X_1 and Y is the proportion of the variance in Y not explained by X_2 that can be uniquely explained by X_1. The proportion of the variance in Y not explained by X_2 is A + D in the Venn diagram. Of this remaining variance in Y, the part it shares with X_1 is A, and thus pr_1^2 = A/(A + D). Using a similar logic, pr_2^2 = C/(C + D).

If the Venn diagram and the previous explanation don't help clarify the semi- and partial correlations, perhaps adding a food analogy will. Suppose you are second in line for a slice of pie. In front of you is Uncle Patrick, and behind you is your sister Amanda. Call Uncle Patrick X_2, yourself X_1, and the pie Y. Uncle Patrick gets first dibbs at eating part of the pie. Suppose he takes the area of the pie in the Venn diagram corresponding to B + C. That leaves A + D for you and Amanda and anyone who comes after her. Suppose you take A, leaving D for Amanda and everyone else.

The area you have taken can be interpreted in two ways. On the one hand, you've eaten a certain fraction of the total pie. This is A/(A + B + C + D). In the Venn diagram in Figure 3.18, this looks like about 15% of the pie, or 0.15 in proportion terms. This is sr_1^2. Amanda might complain that you got more than this, however. From her perspective, what Uncle Patrick took is long gone. She wasn't going to get a shot at this. She is more worried about the amount you take, because she is next in line. From her perspective (and in terms of Figure 3.18), you took what appears to be about 25% of the remaining pie. In proportion terms, this is 0.25, or pr_1^2.

Did you get 15% or 25% of the pie? That depends on your perspective. In the same way, sr_1^2 and pr_1^2 differ in the perspective they take about X_1's role in explaining variation in Y. sr_1^2 gauges variance uniquely explained by X_1 relative to all of Y, whereas pr_1^2 references X_1's contribution to explaining Y relative to variance in Y not already explained by X_2.

3.4.3 Partial Relationships and Simple Relationships May Have Different Signs

We have seen that b_j, sr_j, and pr_j all have the same sign. Intuition would suggest that these measures of partial association would have the same sign as the simple association between regressor X_j and Y. However, this intuition is faulty. Recall the preschool examples of section 1.1.3, where we saw that in Holly City, preschool had a negative simple relationship to later school readiness, but a positive partial relationship when SES was controlled. In Ivy City, the simple relationship was positive but the partial relationship was zero. So we have seen in two examples already that the

sign of a variable's simple relationship with Y is not necessarily diagnostic of the sign of its partial relationship with Y. Similar paradoxes can arise when variables are continuous.

In section 3.4.5 we provide formulas for various measures of partial association. We bring one to your attention here to make this point analytically. Although this discussion is framed in terms of correlations and assumes one independent variable and one covariate, it generalizes to any number of regressors and the other measures of partial association we have discussed.

The partial correlation between independent variable X_1 and Y when controlling for covariate X_2 is

$$pr_1 = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{YX_2}^2}\sqrt{1 - r_{X_1X_2}^2}} \tag{3.3}$$

Notice that the denominator of equation 3.3 cannot be negative, so we know that the sign of pr_1 is determined by the sign of the numerator of equation 3.3. The numerator contains r_{YX_1}, the simple correlation between Y and X_1. If X_2 is uncorrelated with X_1 and Y, then we know that $pr_1 = r_{YX_1}$, and thus the simple and partial correlations for X_1 have not only the same sign but are also equal. However, equation 3.3 shows that pr_1 could be opposite in sign to r_{YX_1} if $r_{YX_2} r_{X_1X_2}$ is further from zero than r_{YX_1}, which can certainly happen and often does. And there is nothing to preclude equation 3.3 from producing a nonzero value for pr_1 when $r_{YX_1} = 0$. That would occur if both r_{YX_2} and $r_{X_1X_2}$ are different from zero.

So an independent variable correlating positively with Y could get a negative partial regression weight, or vice versa, depending on the sizes and signs of the correlations between covariates and Y and between the covariates and the independent variable. And a variable that is uncorrelated with Y may nevertheless receive a nonzero regression weight. We will see in section 7.4.5 that similar paradoxes also arise when regression is used for prediction. Some people call the situations we just described *suppression*, but this term has many definitions in the statistics literature. We reserve the use of this term to the situation we describe later.

3.4.4 How Covariates Affect Regression Coefficients

Suppose you have already computed a simple or partial regression coefficient for an independent variable X, possibly in the presence of several covariates collectively labeled U, and you are considering adding one more

covariate C to the regression model. How will doing so change the regression coefficient for X? To find out precisely, you add C to the regression and see what happens to b_X. Here we give some rules that may enable you to guess at least the sign of the change in b_X without running a new regression—is the change going to be positive, negative, or zero?

The sign of the change in b_X in this scenario will be the product of the signs of two correlations: pr_C and $pr_{CX.U}$. Here, pr_C is the partial correlation between Y and C with X and U held constant, while $pr_{CX.U}$ is the partial correlation between C and X with U held constant. Of course, if there are no other covariates U, then $pr_{CX.U}$ reduces to r_{CX}.

This rule means that if either pr_C or $pr_{CX.U}$ is zero, then adding C to the regression does not change b_X at all. If the two correlations agree in sign (either both positive or both negative), then b_X increases, while if the two correlations have opposite signs, then b_X decreases. Here "increase" means moving right on the number line and "decrease" means moving left on the number line, rather than becoming more distant from zero. So either an increase or a decrease in b_X could mean getting closer to zero, depending on what b_X is before C is added.

There is no easy way to set a limit on the magnitude of the change in b_X produced by C; b_X may change in sign, and may either increase or decrease in absolute value. The simple regression formula $b = r_{XY}(s_Y)/s_X$ tells us that in simple regression, b cannot exceed s_Y/s_X in value since $-1 \leq r \leq 1$. But there is no comparable limit in multiple regression; even with variables standardized so that $s_X = s_Y = 1$, b_X could theoretically fall far above 1 or below -1, although this would tend to be rare.

3.4.5 Formulas for b_j, pr_j, sr_j, and R

We have seen that b_j, pr_j, sr_j and R can be defined in terms of residuals, and that one need not estimate a multiple regression model to generate these measures of partial association. In fact, one need not have a regression program at all, as they can be computed from nothing other than Pearson correlations and standard deviations. The formulas for three or more regressors are too complex to include here. But for two regressors X_1 and X_2, the formulas are

$$
\begin{aligned}
b_1 &= \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \times \frac{s_Y}{s_{X_1}} \\
b_2 &= \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \times \frac{s_Y}{s_{X_2}} \\
pr_1 &= \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{YX_2}}\sqrt{1 - r^2_{X_1X_2}}} \\
pr_2 &= \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r^2_{YX_1}}\sqrt{1 - r^2_{X_1X_2}}} \\
sr_1 &= \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}} \\
sr_2 &= \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}} \\
R^2 &= r^2_{YX_1} + sr^2_2 = r^2_{YX_2} + sr^2_1
\end{aligned}
$$

3.5 Chapter Summary

A multiple regression model is a linear model with more than one predictor or *regressor*. Just as in a model with only a single regressor, an ordinary least squares regression routine can derive a linear combination of regressors that minimizes the sum of the squared residuals when Y is estimated from k regressors. The resulting weights for each regressor in the model are called *partial regression weights* or *partial regression slopes*. The partial regression weight for regressor j provides information about the relationship between regressor j and Y when holding all other regressors constant, also called *statistically controlling* for those other regressors. There are several measures of partial association, including the partial regression weight, the partial and semipartial correlation, and the standardized partial regression weight. They have different interpretations, but each in some way quantifies the unique relationship between regressor j and Y.

Thus far our treatment of linear regression analysis has been entirely descriptive in nature. That is, we have focused on the estimation of a model and interpretation of various statistics that describe the association between variables in the model. In the next chapter we switch our focus away from description and direct it toward statistical inference, acknowledging that

the statistics computed in any sample are specific to that sample, and that we often want to generalize the knowledge we acquire through regression analysis to a broader population or to the process generating the data.

4
Statistical Inference in Regression

This chapter addresses statistical inference in linear regression. It begins with a conceptual overview of the primary and secondary assumptions of inference in regression analysis, followed by a deconstruction of the ANOVA summary table provided by most regression programs. Also defined are the regression and total sum of squares, degrees of freedom, and mean squares. Null hypothesis tests and interval estimates for partial regression coefficients, multiple correlations, and partial correlations are described. Significant attention is devoted to the factors that affect the standard error of a regression coefficient, as well as collinearity between regressors and its effect on inference.

4.1 Concepts in Statistical Inference

4.1.1 Statistics and Parameters

The first three chapters have couched linear regression analysis in purely *descriptive* terms. We've seen how the least squares criterion in linear regression analysis generates an estimate of Y (denoted $\hat{Y}$) that is a linear combination of regressors, with each regressor given a weight so as to maximize the correspondence between Y and $\hat{Y}$. Various statistics can be constructed along the way, such as the proportion of variance in Y explained by the model, measures of partial association, and so forth. These all describe in one way or another, in numerical terms, something about the relationship between Y and the regressors in the model in the data set.

All the quantities we've calculated to this point are sample specific. They are statistics. The term *statistic* is often used to refer to the value of some index or quantity that is a property of a particular sample. For instance, the mean weight loss of the 10 people in the exercise and weight-loss example we've been using thus far is a statistic, as is the standard

deviation of weight loss, as is R—the correlation between Y and $\hat{Y}$ in any of the models we've estimated. They are sample specific in the sense that if we calculated these quantities in exactly the same way but using a different set of 10 people, they would almost certainly be different. Indeed, if any of them were the same, we'd be witnessing something very improbable, and probably just a coincidence.

Statistics are often used as estimators of their corresponding *parameters*, sometimes called their *true values* or *population values*. *Parameter* has several meanings, but for our purpose it is sufficient to think about a parameter as the value of the corresponding statistic that you would get if you had an infinitely large sample size or, alternatively, if the cases in your analysis constitute all cases that could have been included if you had unlimited resources and time and could collect data from (in the case of people) everyone in the population you are studying. Such a sample is given a special name—it is a *census*.

In this ideal but impossible reality of infinite sample size or statistics based on a census, statistics cannot vary from sample to sample, because there would be only one possible sample. The statistics we calculate are parameters—their true values—in this case. But statistics vary from sample to sample in the real world of research because our samples are not infinite in size. And rarely do we have a census of a population. The statistics you have calculated are what they are in part just by the luck of the draw—the fact that you included *these* 10 people in your study rather than *those* 10 people. We call this *sampling variance*. Sampling variance is largely an unavoidable fact of life in data analysis, but it can be estimated, and it can be managed. We use information about sampling variance along with our statistics to conduct statistical inferences, which is the topic of this chapter. This chapter merely introduces statistical inference in regression by focusing on the more common inferential tasks and tests used by researchers. The topic also arises in later chapters.

The goal of *population inference* is to make a statement about the unknown (a parameter) from a known (the corresponding statistic). But sampling variance gets in the way. We know that our statistics vary from their true values—the population values or parameters—merely as a result of calculating the statistics on one set of cases rather than on another set. But fortunately we have an arsenal of statistical theory to help us make the leap from the known to the unknown. We assume that you've already been exposed to the theory of statistical inference at some point and have

probably already done an inference or two in your life as a scientist, so we do not go into the theory in detail in this book.

We deal with many kinds of statistics and parameters in regression analysis: multiple and partial correlations, simple and partial regression weights, residual variances, partial multiple correlations (introduced later in Chapter 5), marginal and conditional means, regression constants, and so forth. It is common to use Greek letters to refer to parameters and Roman letters to refer to statistics. If we followed this convention, you'd have to remember the names and meanings of the entire Greek alphabet. Furthermore, the same Greek symbols are used to mean different things, depending on context. To avoid this confusion, we use a nonstandard approach but one that we have found helpful. A presubscript T before a symbol or abbreviation will denote a parameter, true value, or population value (terms that mean the same thing in this book). Thus, ${}_Tb_j$, ${}_TR$, and ${}_Tpr_j$ denote, respectively, a population regression weight, a population multiple correlation, and a population partial correlation. In words, you might say "sub-T-b-sub-j" or "true p-r-sub-j."

This chapter considers inferences concerning three types of parameters:

- The multiple correlation ${}_TR$
- The simple and partial regression coefficient ${}_Tb_j$
- The simple and partial correlation ${}_Tpr_j$

and four inferential problems:

- Estimating the parameter
- Testing the null hypothesis that the parameter equals zero
- Testing other null hypotheses about the parameter
- Constructing a confidence interval for the parameter

In combination, this gives $3 \times 4 = 12$ types of inference. This chapter covers most of them in one way or another, although some are less important than others and so space is differentially allocated to these problems depending on their importance. We leave out one important family of parameters involving subsets of regressors and corresponding inferential problems. This topic is addressed in Chapter 5 when we address multidimensional sets.

Given that any correlation is determined in part by the variation of the variables in the sample, it would seem that some of these tests are of limited

use given that correlations are therefore properties of the specific sample. In particular, when the investigator determines variability of the variables (e.g., when a quantitative regressor is experimentally manipulated), all correlations are properties of the experiment itself rather than of the natural world. But it turns out that this is not such a problem when testing a hypothesis that a correlation is zero, because hypotheses about association between variables can always be stated without reference to correlations, as hypotheses about a *relationship* captured by statistics that are not sensitive to artificially induced, researcher-imposed constraints on variation, such as a regression weight.

4.1.2 Assumptions for Proper Inference

There are four assumptions made for proper statistical inference in linear regression analysis that we call the *standard assumptions* of regression theory. In our discussion we distinguish between *primary* assumptions and *secondary* assumptions, although this terminology is not standard. A primary assumption is one that, when violated, jeopardizes the very meaning of the parameter you are estimating with the model. A secondary assumption is one that, when violated, may threaten the accuracy of the inference we make about a parameter, but not the very meaning of the parameter itself. We introduce these assumptions here only conceptually. We address each of them in one form or another later in the book.

The first assumption is *linearity*, and it is a primary assumption. The assumption of linearity states that conditional means of Y fall in a straight line. Recall that a conditional mean is a mean conditioned on a value of the regressor or regressors in the model. To interpret b_j as an estimate of the amount by which two cases that differ by one unit on X_j differ on Y, we must assume that the conditional means of Y given X_j (controlling for all other regressors if others are in the model) fall in a straight line. Recall that $\hat{Y}$ is used as an estimate of the conditional mean of Y. If we can't assume that the conditional means we are attempting to estimate are linearly related to the regressor of interest, then this jeopardizes the meaning of the regression coefficient for that regressor and any inference about it. A violation of linearity would mean that the amount that two cases differing by 1 unit on X_j actually differ on Y depends on X_j. If this were the case, then the relationship between X_j and Y is not a line but, instead, a curve of some kind. If the real relationship between X_j and Y is curvilinear, any estimate of ${}_T b_j$ generated by a regression analysis will be a mischaracterization of the true association.

The second assumption goes by the tongue-twisting name of *homoscedasticity*, sometimes but less frequently spelled with a *k* as *homoskedasticity*. The homoscedasticity assumption is a secondary assumption, and it states that the conditional distributions of *Y* have equal variances. Consider a simple regression model estimating *Y* from a single regressor *X*. The homoscedasticity assumption would be violated if, in the population, the variance in the values of *Y* differed for different values of *X*. This is a secondary assumption because violating it will only influence the accuracy of the inference about certain parameters in a model. Research shows that minor violations of homoscedasticity (called *heteroscedasticity*) don't cause too much of a problem, but the assumption is important enough that you need to be aware of what heteroscedasticity is and its effects.

The third assumption concerns the shape of the conditional distributions of *Y*. For statistical inferences to be exact, these distributions must be normal. Note that this assumption pertains to the conditional distributions of *Y*, not the distribution of *Y* itself. As already mentioned, normality is a secondary assumption. Regression is one of many statistical techniques that assume normality. A very important theorem called the *central limit theorem* applies to all these techniques. It says that the larger the sample, the less important is the assumption of normality since the larger the sample size, the closer the *F*- and *t*-statistics for these methods will come to having the same shape as they would under exact normality. Thus, moderate non-normality of the conditional distributions of *Y* is a problem mainly for researchers working with sample sizes under about 30, although extreme non-normality may require somewhat larger sample sizes.

The fourth assumption is *independent sampling*, which requires that cases in the data be independent from one another on *Y*. Independence could be violated in a number of ways. For example, if you included married couples in your analysis but treated them as if they were strangers and didn't know and influence each other in some fashion, this could be a violation of independence, depending on what *Y* is. Alternatively, suppose you randomly selected four downtown office buildings and stood outside of the buildings and asked people questions as they left as part of your data collection effort. If you are measuring something that might be correlated with things that distinguish people who work in different buildings, this could be a violation of independence as well. Independence is a secondary assumption. There are special forms of linear regression that can be used when nonindependence is likely or assured due to the nature of the sampling or data collection, such as multilevel modeling.

As noted earlier, some assumption violations are more problematic than others. Violations of normality are least severe, and the effects of non-normality can usually be made even less problematic by increasing the sample size. Violations of homoscedasticity can have different effects, depending on the form of the violation. Unlike with non-normal conditional distributions of Y, more data (i.e, increasing the sample size) won't make the problem go away. Violation of independence can wreak havoc with an inference, but it may not, depending on the form of nonindependence and how pervasive it is.

As a primary assumption, violating linearity can be disastrous for inference because the regression coefficient simply isn't estimating anything that is meaningful. But it turns out that linear regression analysis can be used to model curves, so even this problem can be dealt with in some fashion. We address nonlinearity in Chapter 12.

All of these assumptions can also be framed in terms of the errors in estimation of Y. Recall that the residuals from a regression model are the errors in the estimation of Y from the regressors. The assumption of linearity implies that the conditional distributions of the errors in estimation all have means of zero. The assumptions of normality and homoscedasticity of the conditional distributions of Y translate into equivalent assumptions about the distributions of the errors in estimation. And independence means that the errors in estimation are uncorrelated with each other. When all these assumptions are combined, it is often said that inference in regression assumes errors in estimation that are *independently and identically distributed* (sometimes abbreviated *i.i.d.*) that are normal and centered at zero. When these assumptions are met, it is said that the errors in estimation are *exchangeable*. They are exchangeable because they all come from the same distribution—one with a mean of zero that is normal in shape with a given variance.

Notice that none of these assumptions imposes a requirement of *random sampling* from a population, although certain kinds of inferences do. If the cases that are included in an analysis are selected from a population through some kind of chance process, such as random selection with a known or at least calculable probability of inclusion, we say the sample is a random or *probability* sample. Most statistical texts and the tests described in those texts presume random sampling from a defined population. When a sample from a population is obtained through a probability sampling method, it is possible to make claims about attributes of the population from which the sample was derived. But this *population model* of inference

is not the only model of inference. There are other ways of thinking about sampling variance and inference that don't assume random sampling from a population. We discuss this in Chapter 16.

We can make assumptions all we want, but that doesn't mean those assumptions are true in a particular case. Fortunately, the plausibility of the assumptions when a regression analysis is applied to a research problem can be tested empirically. Some of these tests are easy to conduct; others are more tedious, time consuming, and not implemented in popular software. This can be a complex topic and entire books are dedicated to it (e.g., Berry, 1993). We address some approaches to testing these assumptions in Chapter 16. In reality, assumptions are routinely violated, sometimes with little effect, but you can never know for certain what effect a particular violation has on the validity or power of the inference you are attempting to make with your data for the specific analysis being conducted. Assumptions are worth understanding and checking, and when you have clear evidence that there could be a problem, do something about it if you can. But don't lose lots of sleep worrying too much about every assumption violation.

4.1.3 Expected Values and Unbiased Estimation

The symbol E before any statistic denotes the *expected value* of that statistic. For instance, the expected value of the sample mean of Y would be denoted $E(\overline{Y})$, and the expected value of the regression coefficient for regressor X_j would be $E(b_j)$. To understand the concept of an expected value, imagine drawing an infinite number of independent random samples of the same size from some population, computing some statistic (e.g., a sample mean or a regression coefficient) in each of these samples, then computing the mean of all those values of the statistic. That mean is defined as the expected value of the statistic. Note that it is possible for the expected value of a statistic to be impossible to ever observe. For instance, consider a population of five numbers: 1, 2, 3, 4, and 6. The expected value of the mean of a sample of three of these numbers is 3.2, but a sample of three of these numbers could never have a mean of 3.2.

If the expected value of a statistic is equal to its corresponding parameter or "true value," then the statistic is said to be an *unbiased estimator* of that parameter. But if its expected value is different from the parameter, we say the statistic is *biased*. An unbiased statistic may not and usually will not be exactly equal to the parameter in any specific sample. Bias or lack of bias is a property of the statistic over repeated sampling and estimation, not a specific estimate calculated in a specific sample.

It can be shown that when randomly sampling from a population, the sample mean of some variable Y is an unbiased estimator of the population mean of that variable, even if the distribution of Y is nonnormal and highly skewed. Furthermore, each sample regression weight b_j in a regression analysis is an unbiased estimator of its corresponding parameter ${}_T b_j$, but R is not an unbiased estimator of ${}_T R$.

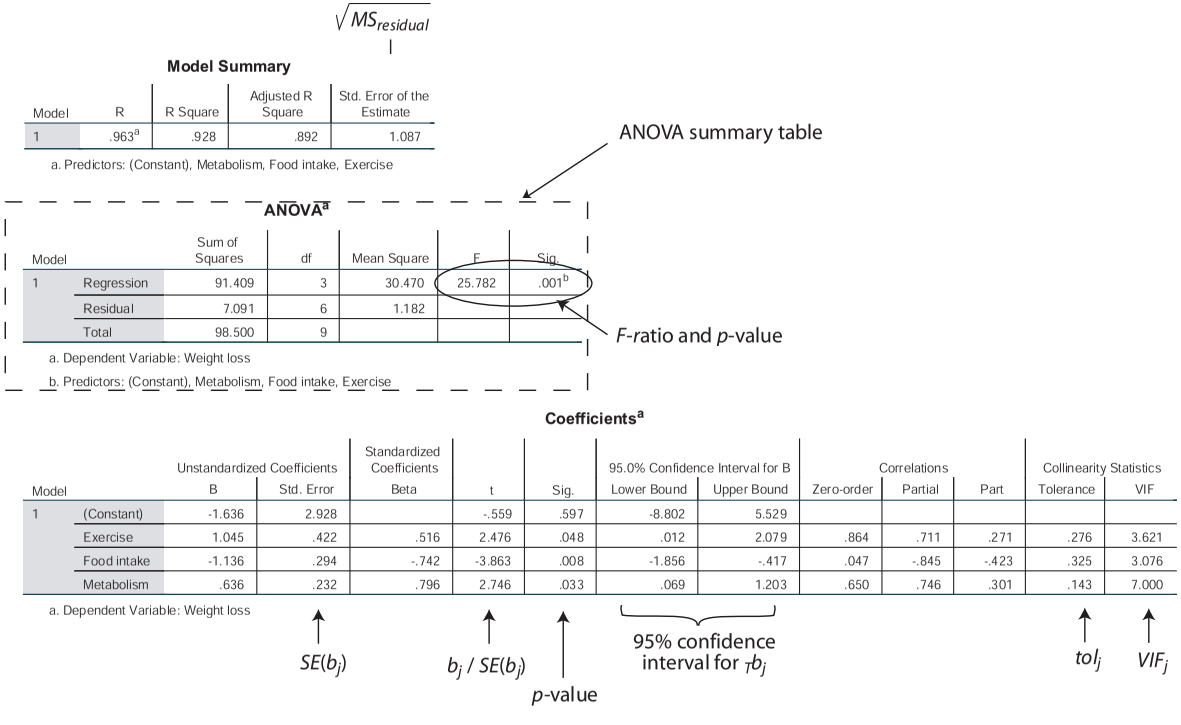
4.2 The ANOVA Summary Table

Most every statistics program that conducts linear regression analysis will produce an "ANOVA Summary Table" or something similar in its output, although not all will label it as such. Understanding this table and where the numbers come from are important for getting a grasp on regression analysis as a whole, as well as the inference process. You won't always need to look at this table to get the information you need for a particular inference, but the information you do look at can trace its source to something in this table. So in this section, we use the ANOVA table as a springboard for further discussing principles of regression and the topic of inference.

Throughout this section we rely on a linear regression analysis using the EXERCISE data file in which weight loss is the dependent variable Y and exercise frequency, food intake, and metabolism are regressors X_1, X_2, and X_3, respectively. Recall from section 3.2.4 that the best-fitting OLS regression model was $\hat{Y} = -1.636 + 1.045X_1 - 1.136X_2 + 0.636X_3$. Corresponding output from SPSS and STATA's regression procedure can be found in Figures 4.1 and 4.2, and comparable SAS output was already presented earlier in Figure 3.14. These were generated with the commands below. These outputs contain more or less the same information, but we include them all so you can easily map the discussion below to the output, regardless of which software platform you choose.

```
regression/statistics defaults zpp ci tol/dep=wtloss/method=enter
   exercise food metab.
```

```
proc reg data=exercise;model wtloss=exercise food metab/stb pcorr2
   scorr2 tol vif;run;
```



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.963[a]	.928	.892	1.087

a. Predictors: (Constant), Metabolism, Food intake, Exercise

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91.409	3	30.470	25.782	.001[b]
	Residual	7.091	6	1.182		
	Total	98.500	9			

a. Dependent Variable: Weight loss

b. Predictors: (Constant), Metabolism, Food intake, Exercise

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-1.636	2.928		-.559	.597	-8.802	5.529					
	Exercise	1.045	.422	.516	2.476	.048	.012	2.079	.864	.711	.271	.276	3.621
	Food intake	-1.136	.294	-.742	-3.863	.008	-1.856	-.417	.047	-.845	-.423	.325	3.076
	Metabolism	.636	.232	.796	2.746	.033	.069	1.203	.650	.746	.301	.143	7.000

a. Dependent Variable: Weight loss

FIGURE 4.1. SPSS output from a multiple regression analysis of the weight-loss data.

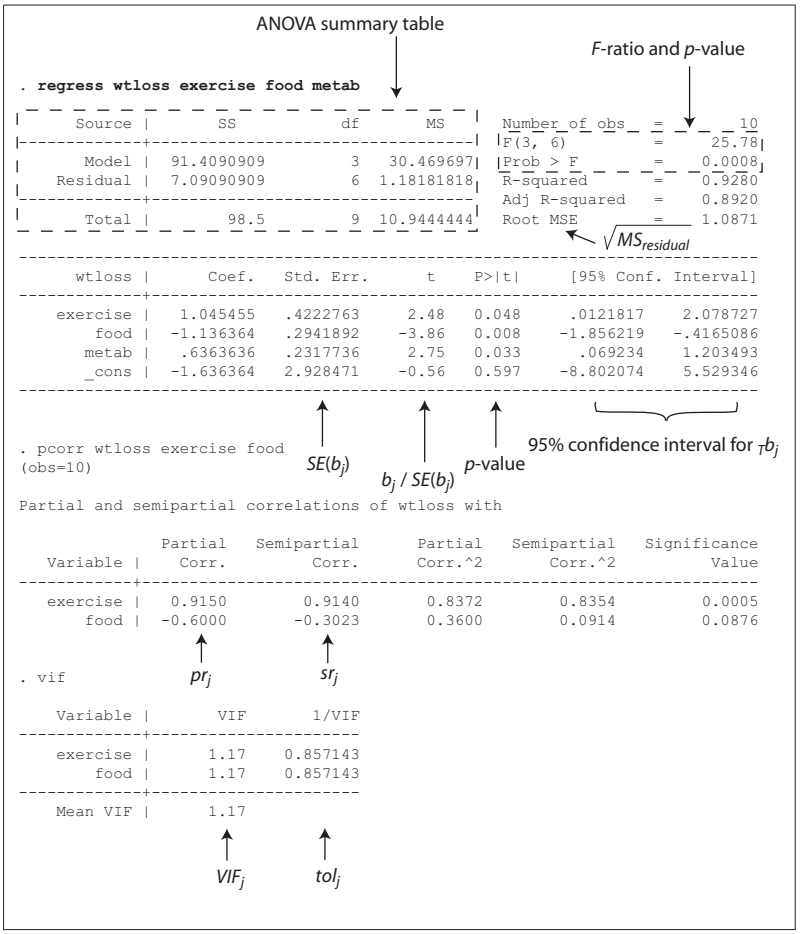


FIGURE 4.2. STATA output from a multiple regression analysis of the weight-loss data.

```
regress wtloss exercise food metab
pcorr wtloss exercise food
vif
```

4.2.1 Data = Model + Error

In section 2.4.1 the formula

$$Y_i = \overline{Y} + (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

was introduced, showing that in a regression analysis, each case's value of Y can be broken into three components: the sample mean of Y, the difference between the estimate of that case's Y from the model and the mean of Y, and the difference between the case's actual value of Y and the model's estimate of that case's value of Y. Of these three components, $\overline{Y}$ carries no information about individual differences in Y, which is ultimately what we are trying to model with regression analysis.

If we move $\overline{Y}$ to the left side of the equation, as in

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

then it can be seen more clearly that regression analysis splits how much case i's value on Y differs from $\overline{Y}$ into two components. $\hat{Y} - \overline{Y}$ is the *model* or *regression* component. It is the part of the difference between Y_i and $\overline{Y}$ that the regression analysis extracts from information about the relationship between the regressors and Y, or the part *explained by* the variables in the model. What is left over, $Y_i - \hat{Y}_i$, is the part of the difference between Y_i and $\overline{Y}$ that can't be explained by the regressors. It is the *error* component of the model. We've discussed the error component at length, as the error component is just case i's residual.

Thus, if we think of "the data" as the difference between Y and $\overline{Y}$, then we can say *data = model + error*. That is, the data for each case are the sum of model and error components. In regression-speak, $Y - \overline{Y}$ is frequently referred to as the *total* component, and we say *total = regression + residual*.

Table 4.1 shows each of these components generated from the regression of weight loss on exercise frequency, food intake, and metabolism in the columns labeled "Total," "Regression,", and "Residual." As you can see, for every case in the data, Total = Regression + Residual.

TABLE 4.1. Generating the Total, Regression, and Residual Components of the Model

Exercise	Food intake	Metabolism	Weight loss		Total		Regression		Residual	
X_1	X_2	X_3	Y	$\hat{Y}$	$Y - \overline{Y}$	$(Y - \overline{Y})^2$	$\hat{Y} - \overline{Y}$	$(\hat{Y} - \overline{Y})^2$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
0	2	13	6	5.636	−1.500	2.250	−1.846	3.473	0.364	0.132
0	4	14	2	2.727	−5.500	30.250	−4.773	22.779	−0.727	0.529
0	6	19	4	3.636	−3.500	12.250	−3.864	14.928	0.364	0.132
2	2	15	8	7.727	0.500	0.250	0.227	0.052	0.273	0.074
2	4	21	9	9.273	1.500	2.250	1.773	3.143	−0.273	0.074
2	6	23	8	8.273	0.500	0.250	0.773	0.597	−0.273	0.074
2	8	21	5	4.727	−2.500	6.250	−2.773	7.688	0.273	0.074
4	4	22	11	12.000	3.500	12.250	4.500	20.250	−1.000	1.000
4	6	24	13	11.000	5.500	30.250	3.500	12.250	2.000	4.000
4	8	26	9	10.000	1.500	2.250	2.500	6.250	−1.000	1.000
				$SS_{total} = \sum \left(Y - \overline{Y}\right)^2 =$		98.500				
				$SS_{regression} = \sum \left(\hat{Y} - \overline{Y}\right)^2 =$				91.409		
				$SS_{residual} = \sum \left(Y - \hat{Y}\right)^2 =$						7.091

4.2.2 Total and Regression Sums of Squares

In section 2.4 the residual sum of squares was introduced. In terms of the formula *data = model + error* or *total = regression + residual*, $SS_{residual}$ is the sum of the squared error or residual components. $SS_{residual}$ quantifies the discrepancy between the estimates of Y from the regression model and the actual values of Y, and it is this quantity that the least squares criterion minimizes in the process of deriving the regression coefficients. It has a lower bound of zero, which occurs only when $\hat{Y} = Y$ for every case in the data and indicates a model that perfectly fits the data.

Whereas zero is the lower bound of $SS_{residual}$, its upper bound is a quantity called the *total sum of squares*, defined as

$$SS_{total} = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

It is quite literally the sum of the squared total components. SS_{total} is essentially a measure of variability in Y. In fact, SS_{total} is the numerator of the definition of variance given in section 2.2.2:

$$\text{Var}(Y) = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N} = \frac{SS_{total}}{N}$$

For this reason, it is helpful for the sake of discussing regression to think of SS_{total} as "variance to be explained." It quantifies the variation between cases around the mean of Y. A perfectly fitting model will explain all of this variance. The closer $SS_{residual}$ is to SS_{total}, the worse the model, meaning the less of the variability between cases in Y the model explains. Note that SS_{total} is a property of the *data* and not a property of the regression model. It will be the same regardless of the number of regressors in the model, what those regressors measure, and the relationship between those regressors and Y. That is, *any model* of the same N values of Y will have the same SS_{total}. Table 4.1 contains the squared regression and total components from the weight-loss regression analysis, with their sums at the bottom. Observe that $SS_{total} = 98.5$ and for this model, $SS_{residual} = 7.091$.

The difference between $SS_{residual}$ and SS_{total} has its own name. It is called the *regression sum of squares* and is defined as

$$SS_{regression} = \sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2$$

It is the sum of the squared regression components and quantifies the discrepancy between what the model estimates for Y for each case and the mean of Y. Unlike with $SS_{residual}$, for $SS_{regression}$, the bigger the better. It quantifies how much information the variables in the model provide about how cases differ from $\overline{Y}$. The closer $SS_{regression}$ is to SS_{total}, the better the model fits the data. If the variables in the model contained no information about Y, then the most sensible estimate of Y would be $\overline{Y}$. That is, knowing nothing else about a person, your best guess as to that person's Y would be $\overline{Y}$. But if you knew something about that person related to Y, then you could use that information to deviate your guess away from $\overline{Y}$ to something that is more accurate. $SS_{regression}$ quantifies the amount the estimates of Y deviate from $\overline{Y}$.

$SS_{regression}$ and $SS_{residual}$ are just two ways of conveying the same information, because they are bound together by the fact that they add up to the total sum of squares. That is,

$$SS_{total} = SS_{regression} + SS_{residual}$$

Thus, a smaller regression sum of squares necessarily implies a larger residual sum of squares. When one goes up, the other must come down by the same amount. The squared regression components for the weight-loss analysis can be found in Table 4.1, as can $SS_{regression}$. As can be seen, $SS_{regression} = 91.409$, which when added to $SS_{residual} = 7.091$ yields $SS_{total} = 98.5$.

Given that SS_{total} represents total variance in Y and $SS_{residual}$ is variance in Y unexplained by the model, it follows that $SS_{regression}$ can be interpreted as variance in Y explained by the model. But recall from section 3.2.5 that R is the correlation between $\hat{Y}$ and Y, and R^2 is interpreted as the proportion of variance in Y explained by the model. It follows that R^2 is related to these sum of squares, as such:

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

From the weight-loss regression analysis, $R^2 = 0.928$, or in terms of sums of squares,

$$R^2 = \frac{91.409}{98.500} = 1 - \frac{7.091}{98.500} = 0.928$$

If OLS regression operates by minimizing $SS_{residual}$, then it follows that in so doing, it is also *maximizing* $SS_{regression}$ and therefore R^2, because SS_{total} is

fixed as a property of the data rather than the model. So the least squares criterion generates estimates of the regression coefficients and regression constant that maximize the correlation between Y and $\hat{Y}$.

4.2.3 Degrees of Freedom

A straight line is determined by two numbers—a Y-intercept and a slope. By choosing those numbers we can make the line fall anywhere we choose in two-dimensional space. We know that if a sample contained only two cases measured on X and Y, so that a scatterplot of Y against X contained only two dots that differed on X, then we could choose a regression constant b_0 and a regression coefficient b_1 such that the regression line $\hat{Y} = b_0 + b_1X$ passes through both dots. We are free to choose any two values of b_0 and b_1, so we have two *degrees of freedom*. A tilted plane, such as in section 3.1.4, is determined by three numbers—b_0, b_1, and b_2. By freely choosing appropriate values for these, we can make the plane pass through any three points in three-dimensional space. Therefore, the plane has three degrees of freedom. More generally, in any regression model with k regressors, we have one degree of freedom for each regressor, and an extra degree of freedom for the regression constant, making $k + 1$ degrees of freedom. When a sample size N equals $k + 1$, we know before inspecting the data that we can make the model fit the data perfectly. In other words, for any sample size $N = k + 1$, a regression model with k regressors will *almost* always produce $R = 1$, with $SS_{residual}$ exactly zero, and $SS_{regression}$ equal to SS_{total}. One exception would be when some cases have the same set of values on the regressors but different Y values.

We are rarely interested in testing a hypothesis about ${}_Tb_0$ so we usually think of a regression model as having k degrees of freedom if it contains k regressors, one for each regression coefficient b_j. We often let *df* denote "degrees of freedom," so k is the *model* or *regression degrees of freedom*, *model df*, or *regression df*. In this book we will denote it $df_{regression}$.

Knowing that a linear regression model will always fit the data perfectly with $k + 1$ cases, $k + 1$ of the cases contain no information that we can use for statistical inference. But as soon as we add another case without adding any regressors, now we know that a perfect-fitting model is not assured. That additional case gives us information we can use for making inferences about aspects of the model, whether it be an inference about ${}_TR$, ${}_Tb_j$, or any other parameter. The larger our sample size N, the better we can estimate a model's fit to the data, and the more precisely we can estimate the regression coefficients and other measures of association. Each

additional case contributes more and more information. More specifically, if the sample size is N, then the number of cases that provide data useful for inference is $N - (k + 1) = N - k - 1$. So, for instance, if $k = 3$, we know beforehand that the model will fit perfectly a sample of four cases. Thus, if $N = 10$, only the last six cases are actually useful for estimating a model's fit and making inferences, since $N - k - 1 = 10 - 3 - 1 = 6$. We call $N - k - 1$ the *residual degrees of freedom*, or $df_{residual}$. Some authors call it the *error degrees of freedom*.

We can also view $N - k - 1$ as the number of regressors or degrees of freedom we could add to the model before completely exhausting the sample's ability to tell us how well the model fits the population or to make inferences about parameters of the model. For instance, if $N = 10$ and $k = 3$, we know that if we added six regressors (and thus 6 degrees of freedom) to the model, making $k = 9$, then the model would necessarily fit the sample perfectly, so that its fit would tell us nothing about its fit to the population and we couldn't make inferences about the population or parameters of the model.

It is possible to estimate a regression model without a constant or, in other words, fixing the regression constant b_0 to zero. In that case, $df_{residual}$ is $N - k$ instead of $N - k - 1$. But such problems are relatively rare.

Recall that the total sum of squares, SS_{total}, is the sum of the regression and residual sum of squares. If you add up the regression and residual degrees of freedom, you get the *total degrees of freedom*, or df_{total}. The total degrees of freedom is one less than the sample size. That is, $df_{total} = df_{regression} + df_{residual} = k + (N - k - 1) = N - 1$.

4.2.4 Mean Squares

The regression and residual degrees of freedom both have zero as their lower bound and SS_{total} as their upper bound. We know that as you add regressors to a model, $SS_{residual}$ cannot go up and will virtually always go down, which means that $SS_{regression}$ generally goes up when regressors are added. We also know that all other things being equal, as the sample size N increases, so too will all of the sums of squares, because the more positive numbers you add, the larger that sum will get.

An ANOVA table contains a statistic that adjusts the sum of squares by dividing them by their corresponding degrees of freedom. The result

is the *mean squared total*, *mean squared regression*, and *mean squared residual*, denoted MS_{total}, $MS_{regression}$, and $MS_{residual}$. That is,

$$MS_{total} = \frac{SS_{total}}{df_{total}}$$

$$MS_{regression} = \frac{SS_{regression}}{df_{regression}}$$

$$MS_{residual} = \frac{SS_{residual}}{df_{residual}}$$

The ratio of $MS_{regression}$ to $MS_{residual}$ is important when testing a hypothesis about the multiple correlation. We address this in section 4.3. The name "mean squared" is rather unfortunate given that neither of these is an actual mean of the squared components. These statistics have the property that they are generally less influenced by adding regressors or cases to a model than are the sums of squares.

The mean squared residual, also called the *mean squared error* and often abbreviated MSE, is an unbiased estimator of the variance of the errors in estimation of Y, which we denoted Var($Y.X$) in Chapter 2. That is, suppose you wanted to know the amount, on average, $\hat{Y}$ tends to differ from Y when the model is fitted to the entire population (or in a sample of infinite size). $MS_{residual}$ is generally used in statistics as an important estimator of the square of this quantity. Its square root is called the *standard error of estimate*, which we denote $s_{Y.X}$, and it is printed as a matter of routine by many regression programs. It is an estimator of the standard deviation of the errors in estimate. As you know, means, regression coefficients, and other statistics have their own standard errors. These usually decline with sample size. But the standard error of estimate does not decline with increasing sample size, because we are estimating a value for each participant rather than a single value for the entire population.

It is unlikely you would ever do any of these computations by hand. Most every statistical program that conducts regression analysis will provide a table containing the total, regression, and residual sum of squares, degrees of freedom, and mean squares. SPSS, SAS, and STATA are not exceptions, as can be seen in Figures 4.1, 4.2, and 3.14. Furthermore, you need not memorize any these formulas, although there is no harm in doing so and you may find yourself accidentally memorizing them as your knowledge of regression expands. Table 4.2 contains a generic ANOVA

TABLE 4.2. Formulas for Entries in a Regression ANOVA Summary Table

Source	*SS*	*df*	*MS*	*F*
Regression	$\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2$	k	$SS_{regression}/k$	$MS_{regression}/MS_{residual}$
Residual	$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$	$N-k-1$	$SS_{residual}/(N-k-1)$	
Total	$\sum_{i=1}^{N}(Y_i - \overline{Y})^2$	$N-1$		

summary table that shows the formulas for each of the entries in the typical table.

4.3 Inference about the Multiple Correlation

4.3.1 Biased and Less Biased Estimation of ${}_T R^2$

If you had a census of a population, or at least a sample so large that to distinguish between the size of the sample and the size of the population would be splitting hairs, you could calculate ${}_T R^2$, the true squared multiple correlation between a set of regressors and an outcome of interest. Instead, by relying on a sample of finite size, you can only estimate ${}_T R^2$ as R^2 calculated with whatever data you have available to you.

The problem is that R^2 is a biased estimator of ${}_T R^2$. Because R^2 is the square of R, R is also a biased estimator of ${}_T R$. More specifically, R^2 tends to overestimate ${}_T R^2$. That is, $E(R^2) > {}_T R^2$. To see why, suppose you wanted to estimate the true squared multiple correlation ${}_T R^2$ of the best-fitting model estimating weight loss from exercise frequency, food intake, and metabolism, and you did so using a sample size of only four people. For reasons discussed in section 4.2.3, you know that your regression model will fit the data perfectly, because $SS_{residual} = 0$ and $R^2 = 1$ whenever the residual degrees of freedom ($N - k - 1$) equals zero. Regardless of which four people you used, you'd always get $R^2 = 1$. This generalizes to a model with any number of regressors. Whenever the residual degrees of freedom is zero, $R^2 = 1$. But if R^2 always equals 1 whenever the sample size is exactly one more than the number of regressors, then how could R^2 be an unbiased estimator of ${}_T R^2$?

It turns out that this bias is R^2 as an estimator of ${}_T R^2$ is systematically related to sample size and the number of regressors. A less biased (but still

slightly biased) estimator of R^2 adjusts R^2 by k and N. It can be found in two algebraically equivalent forms:

$$\text{Adjusted } R^2 = R^2 - \frac{k(1-R^2)}{N-k-1} = 1 - \frac{N-1}{N-k-1}(1-R^2) \tag{4.1}$$

which it turns out is also algebraically equivalent to

$$1 - \frac{MS_{residual}}{MS_{total}}$$

Adjusted R^2 is not a new measure of multiple correlation but is, rather, merely a better estimator of ${}_T R^2$. The term $k(1-R^2)/(N-k-1)$ in equation 4.1 is the adjustment to R^2 that makes it less biased. Fortunately, in most real-world regression problems, the bias in R^2 is not particularly severe so long as the sample size is sufficiently large.

There are two caveats to adjusted R^2. First, unlike R^2, adjusted R^2 can be less than zero. A squared multiple correlation cannot be less than zero, so whenever you calculate or see adjusted R^2 below zero, simply round it up to zero. Second, although adjusted R^2 is less biased than R^2, it is not completely unbiased. There are less biased estimators of ${}_T R^2$ (see Olkin & Pratt, 1958; Yin & Fan, 2001) but they are not calculated by popular statistics programs, because they have other disadvantages we shall not discuss. Unbiasedness is not the only desirable feature of a statistic, but a discussion of these other features is beyond the scope of this book.

Even though R^2 is a biased estimator, many people report R^2 and go ahead and interpret it as the proportion of the variance explained by the model. There is probably little harm in this, and it is literally the proportion of the variance in Y explained by the model *in that sample*. But you will also see adjusted R^2 reported by some instead.

Some people mistakenly interpret adjusted R^2 as the proportion of variance in Y *in the population* explained when the regression model derived *in the sample* is applied to the population. That is, imagine using the estimated regression coefficients from the sample to produce $\hat{Y}$ for every member in the population. The squared correlation between Y in the population and these estimates of Y is not adjusted R^2. This squared correlation we call *shrunken* R^2 in Chapter 7, where we discuss it further.

4.3.2 Testing a Hypothesis about ${}_TR$

The multiple correlation quantifies the fit of the model to the data, as the correlation between Y and $\hat{Y}$. We know that R and R^2 will pretty much always be greater than zero in any sample. We might want to know or report whether the variables used as predictors in the regression model explain *any* of the variation in the dependent variable in the population. This is a question about whether ${}_TR = 0$ or, equivalently, whether ${}_TR^2 = 0$.

The null hypothesis that ${}_TR = 0$ can be tested against the alternative that ${}_TR > 0$ by forming a ratio of the regression and residual components of the model or, more specifically, their mean squares. If the null hypothesis is true, the ratio

$$F = \frac{MS_{regression}}{MS_{residual}} \tag{4.2}$$

follows the $F(k, N-k-1)$ distribution. These two values in parentheses are the regression and residual degrees of freedom, respectively. This ratio is available in the ANOVA summary table produced by most regression programs, along with a p-value for the obtained F. When the null hypothesis is true, $E(F) = df_{residual}/(df_{residual} - 2)$, which is about 1 unless the sample size is small. The larger the obtained F, the smaller the probability of the obtained R or R^2 if the null hypothesis is true. If this probability, the p-value, is smaller than some chosen level of significance for the test then the null hypothesis is rejected. In the model estimating weight loss from food intake, exercise frequency, and metabolism, $F(3, 6) = 30.470/1.182 = 25.782$, which has a p-value of .001 (see the SAS, SPSS, and STATA outputs in Figures 3.14, 4.1, and 4.2). We can reject the null hypothesis. There is some association between weight loss and a linear combination of food intake, exercise frequency, and metabolism in the population.

The F-ratio can be derived without actually calculating mean squares. It turns out that equation 4.2 is equivalent to

$$F = \frac{df_{residual} \times R^2}{k(1 - R^2)}$$

Although it might seem strange to use R^2, which is a biased estimator of ${}_TR^2$, in the computations here, this is not a problem. The sampling distribution of F and resulting p-value takes into account the bias in R^2.

4.4 The Distribution of and Inference about a Partial Regression Coefficient

4.4.1 Testing a Null Hypothesis about ${}_T b_j$

Unlike R, each regression weight b_j is an unbiased estimator of the corresponding true weight ${}_T b_j$. That is, $E(b_j) = {}_T b_j$. This unbiasedness does not depend on any of the secondary assumptions being met (i.e., normality, independence, and homoscedasticity).

Next to each value of b_j, most statistical programs include $SE(b_j)$, the estimated standard error of b_j. It will usually appear under the label "Standard Error," "Std. Error", or something similar. The standard error estimates the amount b_j tends to vary around ${}_T b_j$ due to random sampling. More specifically, $SE(b_j)$ estimates the standard deviation of the sampling distribution of b_j when taking a random sample of size N from the population and estimating the regression model. Its square is called the sampling variance of b_j. The accuracy of $SE(b_j)$ as an estimator of ${}_T SE(b_j)$ does require normality, homoscedasticity, and independence. The formula for $SE(b_j)$ is examined at length in sections 4.4.3 and 4.4.4 and then again in section 17.1.2. We first examine the uses of $SE(b_j)$.

The value of t usually printed next to b_j is the ratio of b_j to $SE(b_j)$ and is used when testing the null hypothesis that ${}_T b_j = 0$. The obtained t has a corresponding p-value that quantifies the probability of getting the obtained value of b_j or something more different from zero assuming that ${}_T b_j = 0$. This p-value comes from the t distribution with $df = N - k - 1$, which is $df_{residual}$ for the model as a whole. This p-value, often labeled "Sig.," "Pr(> |t|)," or the like, is a two-tailed significance level associated with the printed t. For instance, in the model estimating weight loss from exercise, food intake, and metabolism, we have $b_1 = 1.045$ and $SE(b_1) = 0.422$, so $t = 1.045/0.422 = 2.476$. In a sample of size 10 and with three regressors, $df_{residual} = 10 - 3 - 1 = 6$, and $p = .048$. So we can conclude at an α level of .05 that ${}_T b_1 \neq 0$; holding food intake and metabolism constant, there is a positive relationship between exercise frequency and weight loss. The obtained value of $b_1 = 1.045$ is too far from zero to be credibly attributed to "chance."

With the exception of the final interpretation, all this information is contained in most regression analysis outputs, as in Figures 4.1, 4.2, and 3.14. Typically, the p-value will be rounded to three or four decimal places, depending on the software. So if you ever see a p-value listed as "0.000," this does not mean that $p = 0$. Some programs allow you to change the number

of decimal places displayed. See the documentation for the program you are using.

Occasionally you may wish to test a null hypothesis that ${}_T b_j$ equals some value other than zero. In this example, for instance, $b_1 = 1.045$, which means that each daily hour of exercise is associated with 1.045 units (i.e., 104.5 grams) of weekly weight loss when food intake and metabolism are held constant. If you wanted to show that ${}_T b_1$ is greater than 1, you could test that hypothesis with the formula

$$t = \frac{b_j - \text{null value of } b_j}{SE(b_j)}$$

In this example, we have $t = (1.045 - 1)/0.422 = 0.107$, $df = 6$, $p = .460$ (one-tailed), so you cannot claim that ${}_T b_1$ is greater than one. You would usually have to rely on a table of critical values of t or a computer algorithm to find the p-value when testing a null hypothesis other than zero, as most statistical packages only provide p-values for testing a *nil hypothesis* (i.e., a null hypothesis that the parameter equals zero). A table of critical values of t can be found in Appendix C.

4.4.2 Interval Estimates for ${}_T b_j$

You can also use $SE(b_j)$, along with a t-value from a table of critical values of t to construct an *interval estimate*, *confidence limit*, or *confidence interval* for ${}_T b_j$ based on the estimate b_j.

$$\text{Confidence limit} = b_j \pm \text{tabled } t \times SE(b_j)$$

For instance, for a two-tailed 95% confidence interval, the critical value of t when $df_{residual} = 6$ is 2.447. Thus, in this example, a 95% confidence limit for ${}_T b_1$ is $1.045 \pm 2.365 \times 0.422 = 0.012$ to 2.078. Your preferred software package may also provide confidence limits for regression coefficients, making it unnecessary to find the proper tabled value of t and do the computations manually. For instance, as can be seen in Figure 4.1, the confidence interval for ${}_T b_1$ produced by SPSS is 0.012 to 2.079, which agrees almost exactly with these hand computations.

If you are concerned about errors in only one direction—either overestimation or underestimation of ${}_T b_j$—you may use a one-sided confidence interval. For instance, we may be more anxious to avoid overestimating the effect of exercise than underestimating it, because we want to say "Exercise is at least this effective." Thus, we want a lower bound but not necessarily

an upper bound. In that case, for a one-sided 95% confidence interval, use the tabled t-value at the 0.05 level. That value is 1.943, resulting in a lower bound estimate of $1.045 - 1.943 \times 0.422 = 0.225$. The upper limit of the interval estimate, however, is $+\infty$. The price paid for reducing the error in underestimation is no bound on the upper limit.

4.4.3 Factors Affecting the Standard Error of b_j

The size of the standard error of b_j directly influences the p-value for b_j when testing a hypotheses about a variable's unique effect on Y or its contribution to explaining variance in Y. It also determines the width of an interval estimate of ${}_T b_j$. The smaller the standard error of b_j, the less values of b_j vary from sample to sample, and the more accurate is b_j's estimate of ${}_T b_j$. The square of the standard error of b_j, called b_j's *sampling variance*, is a function of four quantities:

$$SE^2(b_j) = \frac{MS_{residual}}{N \times \text{Var}(X_j) \times (1 - R_j^2)} \tag{4.3}$$

where $\text{Var}(X_j)$ is the variance of regressor X_j and $(1 - R_j^2)$ is the squared multiple correlation estimating regressor X_j from the other $k - 1$ regressors in the model. We will later define $(1 - R_j^2)$ as X_j's *tolerance*, and its inverse as X_j's *variance inflation factor*. The square root of equation 4.3 is the standard error of b_j. This is perhaps one of the more important formulas in regression, and much can be learned from understanding and dissecting it. In this section we describe how each of these four quantities relates to the size of $SE(b_j)$.

The most intuitive of the entries in this formula is the sample size N. It is in the denominator of the standard error formula, so as N increases, $SE(b_j)$ decreases. This is consistent with what is generally well known by anyone who has taken a course in statistics. All other things being equal, the more data one has, the less statistics vary around their corresponding parameters, meaning the more accurate those statistics are as estimators.[1]

The numerator has only a single quantity: the mean squared residual. $MS_{residual}$ is directly tied to the size of $SS_{residual}$ as well as to the multiple correlation R. Given that the least squares criterion seeks to minimize $SS_{residual}$, it also minimizes $MS_{residual}$ while also maximizing R. $MS_{residual}$ is in the numerator of this formula, so the smaller $MS_{residual}$ (i.e., the smaller

[1]Equation 4.3 is the definitional formula for $SE(b_j)$. Most likely, your regression program uses $df_{residual}$ in place of N in the computation of regressor j's standard error.

$SS_{residual}$ and the larger R), the smaller the standard error for b_j; indeed, the smaller $MS_{residual}$, the smaller the standard errors for all k regression coefficients. In short, all other things being equal, better fitting models yield regression coefficients with smaller standard errors.

At its extreme, consider a regression model with a single predictor, and imagine that ${}_TR = |{}_Tr_{XY}| = 1$. In such a world, a scatterplot of Y against X will be a straight line in any sample you take from this population, and the regression coefficient for X will be the same in every sample. That means that the regression coefficient for X does not vary from sample to sample, meaning that the standard error of the regression coefficient should be zero. Indeed, in any sample from such a population, $SS_{residual}$ and $MS_{residual}$ will be zero, which when plugged into equation 4.3 yields $SE(b_1) = 0$, as you would expect. This generalizes to a regression model with any number of predictors, under the condition that ${}_TR = 1$.

The appearance of $\text{Var}(X_j)$ in the denominator means that $SE(b_j)$ decreases as the variance of X_j increases. The reason for this is made clear by considering a simple example. Imagine you are interested in building a linear model of life expectancy predicted from the amount a person smoked during the course of life. Suppose you have available a person's age at death (Y) and the number of cigarettes the person smoked, on average, each week (X), and you have these data for a large sample of people who died at a local hospital. The regression coefficient for X in this model would probably be negative, meaning those who smoked more died younger. This regression coefficient could be thought of as a measure of the effect a single additional cigarette per day would have on how much sooner a person will die. It should make sense that you could estimate this effect much more precisely if you have a sample that contains people who differ widely in the amount they smoked during life relative to one in which the people are very similar to each other in their smoking frequency. In other words, you'd expect the standard error for the regression coefficient for cigarettes smoked in this model to be much smaller in a sample that is very heterogeneous in smoking frequency, meaning in a sample that is more variable on this regressor. Equation 4.3 reflects this.

The last component of this formula, $1 - R_j^2$, quantifies the proportion of the variance in X_j that cannot be explained by the other regressors in the model. As this quantity increases, $SE(b_j)$ decreases. As $1 - R_j^2$ is important in its own right, we dedicate an entire section to this statistic next.

4.4.4 Tolerance

In most uses of regression analysis, the regressors are correlated with each other to some degree, some perhaps more than others. A statistic called a variable's *tolerance* quantifies how correlated a regressor is with the other regressors in the model. Each regressor has a tolerance, which we denote Tol_j. Imagine estimating regressor X_j from the other $k-1$ regressors. In section 3.2.2 we introduced this idea, called a *crosswise regression*, when deriving various measures of partial association. Call the squared multiple correlation from this crosswise regression R_j^2. Regressor X_j's tolerance is

$$Tol_j = 1 - R_j^2$$

As R_j^2 quantifies the proportion of variance in X_j explained by the other regressors in the model, Tol_j quantifies the proportion of the variance in X_j that is unexplained by the other regressors. Thus, Tol_j is a measure of the *independence* of X_j from the other regressors. If X_j is independent of the other regressors, then $Tol_j = 1$. If it is entirely dependent on the other regressors, then $Tol_j = 0$, a condition known as a *singularity*. Tol_j also measures X_j's *collinearity* with other regressors, with low tolerance indicating high collinearity. We introduced collinearity in section 3.4.1 without formally defining it. Two regressors are high in collinearity if they are highly correlated. Thus, when two regressors that are highly correlated are in the same model, each of their tolerances will be small because R_j^2 for each of these regressors will be large.

Nonindependence between X_j and other regressors raises R_j, which lowers Tol_j, which raises $SE(b_j)$. The reason for this is perhaps seen most easily when two regressors are both dichotomous. For instance, suppose you want to study attitudes toward Saturday classes in a college's student body; in particular you want to examine the effects of biological sex and residential status (living on campus or off campus) on this attitude. Suppose that the difference between the mean attitude scores of residential and nonresidential students (uncontrolled for sex) is large and significant, as is the difference between the mean attitudes of male and female students uncontrolled for residence. If there were no other important unmeasured covariates, then you could conclude that attitude toward Saturday classes must be affected by either residence or sex, or both.

But suppose most women live on campus and most men live off campus, and you want to distinguish between the effects of residence and sex. Your intuition tells you correctly that your ability to make this distinction

depends on the numbers of off-campus women and on-campus men. The smaller these frequencies, the harder it is to distinguish between these two effects; at the extreme in which there are no off-campus women or on-campus men, you have no ability at all to distinguish between the two effects.

To perform significance tests that distinguish between these two effects, you must run a regression predicting attitude from residence and sex simultaneously. Then the regression coefficient for residence controls for sex, and vice versa, so these regression coefficients distinguish between the two effects. But these regression coefficients are unbiased estimates of the true effect sizes, regardless of how few off-campus women or on-campus men are in the sample, so long as there are some. The difficulty in distinguishing between the two effects shows up not as smaller expected coefficients but as higher standard errors of those coefficients.

But also, the smaller the number of off-campus women and on-campus men, the higher the correlation between the variables coding residence and sex. Thus, the higher the correlation between two regressors, the higher are both values of $SE(b_j)$. Similar effects operate when regressors are numerical. The inclusion of Tol_j in the formula for $SE(b_j)$ reflects this effect; the more highly any regressor X_j correlates with the other regressors, as measured by the crosswise squared multiple correlation R_j, the lower Tol_j and the higher $SE(b_j)$.

The denominator of the formula for $SE(b_j)$ contains $\text{Var}(X_j)$ and the tolerance for X_j, which is the proportion of the variance in X_j not explained by the other regressors in the model. When multiplied together, $\text{Var}(X_j) \times Tol_j$ is the unique portion of the variance in X_j, meaning that variance in X_j not shared with other regressors. So equation 4.3 could be written as

$$SE^2(b_j) = \frac{MS_{residual}}{N \times \text{Var(unique portion of } X_j)}$$

This formula, too, is intuitively reasonable. For instance, suppose we want to study the degree to which assertiveness is influenced by upbringing—specifically, by encouragement of assertiveness in the child by the parents. We might want to control for sex of the child when examining the effect of encouragement. But if the study is performed in a society where assertiveness is strongly encouraged in boys and strongly discouraged in girls, then the variable measuring encouragement by the parents will have little variance that is independent of sex, so its effect on later assertiveness cannot be measured very accurately. That is, across the sample, the extent to which a

parent encouraged the child to be assertive will be highly correlated with sex, so encouragement will have little unique variance. This raises the standard error of its regression coefficient.

The standard error formula reflects the fact that only so much intercorrelation between regressors can be "tolerated" by the mathematics. As the tolerance of X_j is in the denominator of the equation, it can't be zero, as this would make the entire denominator zero and division by zero is not allowed in mathematics. When this happens, most regression programs will either fail to execute and generate an error, or will attempt to solve the problem by automatically removing a variable or variables from the model so that no regressor's tolerance is zero. Thus, it might be helpful to remember that a small tolerance is less desirable by thinking "regression has zero tolerance for zero tolerance."

There is another way of expressing the standard error formula that is helpful in understanding a related statistic called a regressor's *variance inflation factor*, sometimes abbreviated VIF. Consider

$$\begin{aligned} SE(b_j) &= \sqrt{\frac{1}{1-R_j^2}}\sqrt{\frac{MS_{residual}}{N \times \text{Var}(X_j)}} \\ &= \sqrt{VIF_j} \times \sqrt{\frac{MS_{residual}}{N \times \text{Var}(X_j)}} \end{aligned}$$

which separates tolerance—actually its inverse—from the other factors that affect the standard error of regressor X_j. The inverse of a variable's tolerance is its variance inflation factor, VIF_j. It quantifies the amount the sampling variance of the regression coefficient for regressor X_j is increased due to its correlation with other regressors in the model. Its square root can be interpreted as the factor increase in the standard error of b_j that results from the correlation between X_j and the other regressors. For example, if the squared multiple correlation between X_j and the other regressors is 0.60, then $Tol_j = 1 - 0.60 = 0.40$, and VIF_j is $1/0.4 = 2.5$. So the standard error of X_j's regression coefficient is $\sqrt{VIF} = \sqrt{2.5} = 1.58$ times larger than it would be if X_j were uncorrelated with the other regressors.

The quality of a regression program is determined in part by whether it has a procedure or option for generating the tolerance or variance inflation factor of a regressor. As can be seen in Figures 4.1, 4.2, and 3.14, SPSS, SAS, and STATA are quality programs by that metric. The tolerance for metabolism is the lowest of the three at 0.143, meaning that only 14.3%

of its variance is unique to it, so its variance inflation factor is the largest. This means that its standard error is raised more by its correlation with the other regressors than are the standard errors of exercise frequency and food intake.

You will find that some authors offer guidelines for deciding whether collinearity is too large for a regression analysis to be conducted or whether certain variables should be removed from the model. Such recommendations are often based on looking at the variance inflation factors to see if any of them are larger than some value, with 10 being one often recommended cutoff. The logic is that if a variable's VIF is large, the analysis can't be trusted, because one or more of the variables are too highly correlated with others, standard errors end up huge, and power of hypothesis tests are low.

We don't find such rules of thumb particularly helpful and don't use them ourselves. There is nothing magical about 10 or any other arbitrary rule of thumb when it comes to determining whether an analysis is meaningful or can be trusted. And as the formula for $SE(b_j)$ shows, other things can offset the effect of collinearity. If you have the luxury of increasing the sample size, this is the easiest way of offsetting the effect of collinearity on standard errors. Another alternative is adding regressors to the model correlated with Y but not with the other regressors. This will lower $MS_{residual}$ and the standard errors of all the regression coefficients. Further, it may be that all collinearity is confined to covariates, so the important independent variables have low standard errors. The collinearity then does no damage at all.

4.5 Inferences about Partial Correlations

4.5.1 Testing a Null Hypothesis about ${}_{T}pr_j$ and ${}_{T}sr_j$

If b_j is statistically significant using the test described in section 4.4.1, it can be said that there is a linear relationship between X_j and Y when all other regressors are held constant. By the same token, if b_j is not statistically different from zero, then the corresponding claim is that X_j and Y are not linearly related, holding all other regressors constant. Such a claim generalizes to partial and semipartial correlations as well. If one claims that ${}_{T}b_j \neq 0$, one can also claim that ${}_{T}pr_j$ and ${}_{T}sr_j \neq 0$. Conversely, a claim that ${}_{T}b_j$ is not different from zero leads to the corresponding claim that ${}_{T}pr_j$ and ${}_{T}sr_j$ are not different from zero. So separate tests are not required for partial

regression coefficients, partial correlations, and semipartial correlations when the null hypothesis is that the parameter equals zero.[2]

In fact, there is a direct correspondence between the size of t for b_j and the size of pr_j and sr_j. Although many statistical packages will provide sr_j and pr_j in regression output, they can be computed from the t-value for b_j and a few other statistics. For pr_j,

$$|pr_j| = \sqrt{\frac{t_j^2}{t_j^2 + df_{residual}}} \tag{4.4}$$

where t_j is the t-statistic for the regression coefficient for regressor X_j. Equation 4.4 doesn't give the sign of pr_j, but its sign always matches the sign of b_j. So if b_j is negative, then so too is pr_j. For sr_j,

$$sr_j = t_j \sqrt{\frac{1 - R^2}{df_{residual}}} \tag{4.5}$$

You will not usually have to do these computations by hand. We provide equations 4.4 and 4.5 to make the point that if the p-value for b_j comes from the t-value, and the t-value can be converted into the partial or semipartial correlation for X_j, then the p-value for a test of the null hypothesis that ${}_Tpr_j = 0$ or ${}_Tsr_j = 0$ must have the same p-value as the does the test that ${}_Tb_j = 0$.

4.5.2 Other Inferences about Partial Correlations

Inferences other than whether ${}_Tpr_j$ or ${}_Tsr_j$ equals zero are more problematic. The problem is that, for reasons explained later in this section, inferences about partial and semipartial correlations are considerably less robust to assumption violations than other statistics. A test is robust when it produces a valid inference even when its assumptions are violated.

In section 4.1.2 we defined a *primary assumption* as one whose violation jeopardizes the very meaning of the parameter in question, while violations of secondary assumptions merely threaten the accuracy of our inferences about that parameter. The latter violations can sometimes be overcome by using larger or more representative samples, while the former cannot. As discussed, linearity is the only primary standard assumption. But the most

[2]It is worth keeping in mind that a failure to reject the null hypothesis does not imply that the null hypothesis is actually true. It merely means that the evidence available is not sufficient to reject the null from the realm of possibility.

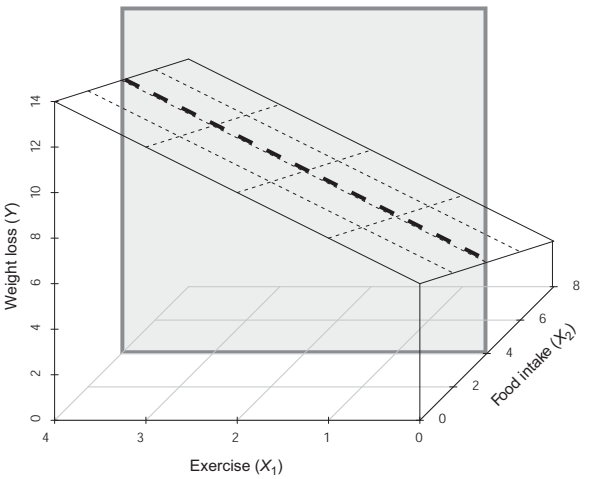


FIGURE 4.3. A slice defining a bivariate conditional distribution of Y and X_1 with X_2 held constant at 4.

obvious uses of partial correlations require primary assumptions that are not among even the *secondary* assumptions of most regression theory.

The most common inferential use of the partial correlation pr_j is to estimate the *conditional correlation* between Y and X_j—the correlation between Y and X_j in a subpopulation in which all other regressors are held constant at some fixed value. There may be infinitely many such correlations—one for every possible combination of values on the other regressors. To visualize a conditional correlation, consider Figure 4.3. It shows not only the familiar titled plane from Chapter 3 but also a vertical plane at the value $X_2 = 4$. In the example using the tilted plane, only 3 of the 10 sample cases fell in the plane $X_2 = 4$ (see Figure 3.3). But we can imagine that in the population the number of cases at $X_2 = 4$ is large. The correlation between Y and X_1 in the plane $X_2 = 4$ is a conditional correlation. When scientists say "the correlation between Y and X_j with other variables held constant," they ordinarily mean a conditional correlation of just this type. Thus, the primary value of pr_j as an estimator lies in its ability to estimate such a correlation.

But this use of pr_j requires bivariate conditional normality as a primary assumption (see Figure 4.4 for a visual depiction). Violation of this assumption is not trivial. In one artificial population satisfying all the standard assumptions but not bivariate conditional normality, each conditional cor-

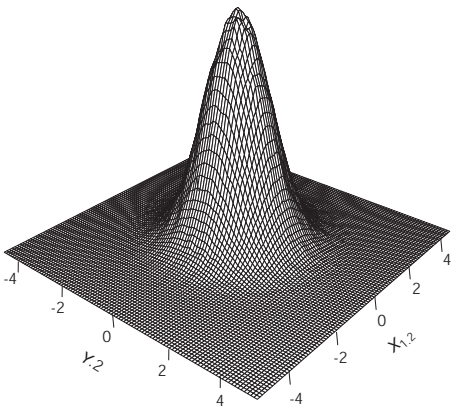


FIGURE 4.4. A visual depiction of bivariate normality of X_1 and Y conditioned on X_2.

relation between Y and X_1 holding X_2 constant was 0.32, but the *population* value ${}_Tpr_1$ was 0.85. Thus, pr_j can grossly mischaracterize the true conditional correlation, even in a very large and representative sample. Even more extreme examples can be created. Bivariate conditional normality can never be satisfied if X_j is categorical since then the crosswise regression predicting X_j can never attain linearity.

All this said, bivariate conditional normality may be met in some circumstances. When so, a meaningful interval estimate for ${}_Tpr_j$ can be constructed by using the *Fisher r-to-Z transformation*. This transformation is often used when constructing a confidence interval for ${}_Tr$, but its use generalizes to partial correlations. The method involves the following steps:

1. Translate pr_j into a Fisher Z using the formula

$$Z_f = 0.5 \times ln\left(\frac{1 + pr_j}{1 - pr_j}\right)$$

   The symbol *ln* denotes a natural logarithm. A table of values of Z_f can be found in Appendix C.

2. Find the standard error of Z_f from the formula

$$SE(Z_f) = \frac{1}{\sqrt{N - k - 2}}$$

   where k is the number of regressors in the model.

3. From a table of normal probabilities (see Appendix C), select the value of z for the desired confidence level. For example, for a two-tailed confidence interval at the 95% level, set $z = 1.96$.

4. The confidence limits on Z_f are then found by the formula

$$\text{Upper and lower bound} = Z_f \pm z \times SE(Z_f)$$

5. Denoting these upper and lower bounds ζ, translate each of these limits back to pr with the formula

$$pr = \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1}$$

where e is *Euler's constant*, which is about 2.718282. Most scientific calculators have an e button.

For instance, the partial correlation between exercise frequency and weight loss controlling for food intake and metabolism is $pr_1 = 0.711$. This translates to $Z_f = 1.006$. Since $N = 10$ and $k = 3$, $SE(Z_f) = 1/\sqrt{10 - 3 - 2} = 0.447$. For a two-tailed 95% confidence interval, we find confidence limits of $1.006 \pm 1.96 \times 0.447 = 0.130$ to 1.882. These translate back into partial correlations of 0.129 and 0.955, which is a 95% confidence interval for ${}_T pr_1$. This is a very wide interval indeed, and shows that we really know very little about ${}_T pr_1$ except that it is probably positive.

To test the null hypothesis that pr_j equals some specified nonzero value, translate both the observed pr_j and the null value into Fisher Z's from step 1, denoting the latter Z_{f0}. Then compute

$$Z = \frac{Z_f - Z_{f0}}{SE(Z_f)}$$

and derive a p for Z from a table of standard normal probabilities.

This same procedure can be used to construct an interval estimate or test a nonzero null hypothesis for ${}_T r$ by substituting r for pr and using $\sqrt{N-3}$ in the denominator of the standard error of Z_f rather than $\sqrt{N-k-2}$.

4.6 Inferences about Conditional Means

You may want to draw an inference about a conditional mean. Remember that a linear regression model is a model of conditional means. When we generate $\hat{Y}$ from a regression model, we are generating a conditional mean.

In this section, we describe the construction of the standard error for a conditional mean. This can be used to generate a confidence interval for or test a hypothesis about the conditional mean.

Let G be a combination of regressor scores, and define $\hat{Y}_G$ as the estimate of Y from the model at point G. Point G need not necessarily be a case in the sample; there may or may not be a point G in the data. For instance, in the regression in section 3.1 we predicted weight loss from exercise frequency (X_1) and food intake (X_2). Let G represent the point at which $X_1 = 3$ and $X_2 = 7$. The regression model in that example was $\hat{Y} = 6.0 + 2.0X_1 - 0.5X_2$, and so $\hat{Y}_G = 6.0 + 2.0(3) - 0.5(7) = 8.5$, or 850 grams of weight loss per week.

The estimate of Y at point G, $\hat{Y}_G$, has two different interpretations. It is the estimated Y for any case (and remember such a case may not exist in the data) with values on the regressors corresponding to point G. It is also an estimator of the population conditional $\overline{Y}$ at point G. Let ${}_T\overline{Y}_G$ denote this true conditional mean. If we wanted to construct a confidence interval for ${}_T\overline{Y}_G$ or test a hypothesis about its value, we would need to know $SE(\hat{Y}_G)$, the standard error of $\hat{Y}_G$. Unfortunately, $SE(\hat{Y}_G)$ depends heavily on G. Different G points have different values of $SE(\hat{Y}_G)$. Fortunately, there is a simple way of finding $SE(\hat{Y}_G)$ for any G with little computational effort. The trick is to make the regression constant b_0 equal $\hat{Y}_G$. When this is done, $SE(b_0)$ from the regression program is equal to $SE(\hat{Y}_G)$.

Remember that in a regression model, b_0 is $\hat{Y}$ when all regressors are set to zero. To make b_0 equal $\hat{Y}_G$, subtract the values of the regressors that define point G from all regressor values in the data. For example, we have defined point G as $X_1 = 3$ and $X_2 = 7$, so we construct two new variables $X_1' = X_1 - 3$ and $X_2' = X_2 - 7$. Now we regress Y on X_1' and X_2'. In this model, b_0 is $\hat{Y}$ when $X_1' = 0$ and $X_2' = 0$. But these correspond to $X_1 = 3$ and $X_2 = 7$, so b_0 from this model is $\hat{Y}_G$, and $SE(b_0)$ is $SE(\hat{Y}_G)$.

When this method was applied to the weight-loss example, the resulting model was $\hat{Y} = 8.5 + 2.0X_1 - 0.5X_2$. Notice that the regression coefficients for X_1' and X_2' are the same as b_1 and b_2 prior to subtracting the values defining G from the regressors. But b_0 is now 8.5 rather than 6.0, and this new regression constant equals $\hat{Y}_G$. The regression program shows $SE(b_0) = 0.683$, which is also $SE(\hat{Y}_G)$. If you ask for it (and perhaps even if you don't, depending on the program you are using), you can get a 95% confidence interval for ${}_Tb_0$, which is also a 95% confidence interval for ${}_T\overline{Y}_G$. In this example, we get $6.885 \leq {}_T\overline{Y}_G \leq 10.115$.

You could also test a null hypothesis about ${}_T\overline{Y}_G$. Suppose we want to test the null hypothesis that ${}_T\overline{Y}_G = 6.0$ against the alternative that it is not

6.0. The t-statistic for $\hat{Y}_G$ is $t = (8.5 - 6)/0.683 = 3.660$, which is distributed as t on $df_{residual} = 7$. This is statistically significant two-tailed at the $\alpha = 0.001$ level.

You may find yourself wanting to conduct an inference for G when G is defined by a specific value or values of some but not all of the regressors. For instance, what if you defined G as $X_1 = 3$ but didn't care what X_2 was? Unfortunately, you have to define X_2 as something, because G must be defined using all the regressors. In such a case, a sensible thing to do is define G by setting all the other regressors you don't care about to their sample means (even if one or more of those other regressors is dichotomous).

This centering strategy works with any regression program, but some regression programs have built-in features for generating linear combinations of regressors along with standard errors and other information for inference. For example, the command to generate $\hat{Y}_G$ and $SE(\hat{Y}_G)$ for the example just described using SPSS would be

```
glm wtloss with exercise food/print=parameters/lmatrix all 1 3 7.
```

The RLM macro for SPSS and SAS discussed in Appendix A has a similar option for conducting inferences for linear combinations of regressors. See your preferred program's documentation for guidance.

4.7 Miscellaneous Issues in Inference

4.7.1 How Great a Drawback Is Collinearity?

As discussed in section 4.4.4, high collinearity increases the standard errors for regression coefficients in a linear regression analysis, but this is often a far less serious problem than researchers fear. There are three reasons for this.

First, collinearity affects only the power of tests on regression coefficients—not their validity. The standard errors of the partial regression slopes are increased for collinear variables. This widens confidence intervals for ${}_T b_j$ and makes it harder to find statistically significant values of b_j. But a significant value of b_j is just as conclusive when collinearity is present as when it is absent.

Second, collinearity often affects only a few of the regressors. If those affected are merely covariates, then values of $SE(b_j)$ are not raised for any of the independent variables. For instance, suppose you have several measures of SES and cannot decide which to use as a covariate. If you

decide to avoid the problem by using them all, little or no harm results from the fact that they may be highly collinear. $SE(b_j)$ will then be high for those covariates but not for the independent variables. (The related problem of excessive number of covariates is examined in sections 4.7.3 and 17.1.3; it, too, is less serious a problem than widely believed.)

Third, although collinearity reduces the power of tests of the individual regressors that are highly correlated, collinearity does not reduce the power of a test on the effect of the set of regressors as a whole. For instance, if you want to test whether SES affects a dependent variable, you can do so efficiently even if SES is measured by several highly collinear variables. Such tests are described in section 5.3.3. Methods for discovering collinear sets are discussed in section 17.3.2.

Another common misconception about collinearity is that it is somehow a problem specific to regression analysis, and that more advanced statistical methods can eliminate the problem either currently or some day in the future. But the problem is essentially that when two variables are highly correlated, it is harder to disentangle their effects than when the variables are independent. This is simply an unalterable fact of life; it can't be avoided with more sophisticated methods. The solution lies not in more clever analytic methods, but in straightforward devices such as larger sample sizes or experimental manipulation of the variables.

But the effects of collinearity are very important to keep in mind when testing competing theories using linear regression. You might want to show that your theoretically important regressor X_1 explains variation in Y after accounting for a regressor X_2 representing an alternative theoretical orientation. That competing theory may make the opposite prediction—that X_2 and not X_1 uniquely accounts for variation in Y. Suppose you find that X_1's regression coefficient is statistically significant in a model estimating Y from X_1, X_2, and a set of covariates, but in this model X_2's regression coefficient is not statistically significant. You might be inclined to celebrate, but this isn't a fair comparison of theories if X_2 is more highly correlated with the covariates than is X_1. This higher collinearity would affect the standard error of b_2 more than b_1, resulting in a widening of the confidence interval for ${}_Tb_2$ and lowering the power of the hypothesis test for ${}_Tb_2$ relative to ${}_Tb_1$.

4.7.2 Contradicting Inferences

Tests described in this chapter for inference about ${}_TR$ and ${}_Tb_j$ can produce what seem to be conflicting results. It is not uncommon for a researcher to

find that R is not statistically different from zero even though one or more of the regression coefficients is. This leads to the apparent contradiction that no linear combination of regressors explains variation in Y even though one or more of the regressors in the model does. Conversely, one could find that none of the regression coefficients are statistically different from zero, yet a hypothesis test on the multiple correlation leads to the inference that ${}_TR > 0$. Thus, while none of the individual regressors is uniquely related to Y, when considered as a set they explain variation in Y.

These tests are testing different null hypotheses. There is nothing in the mathematics that requires them to produce internally consistent results. Usually, such conflicts are due to differences in the power by which the methods test their corresponding hypotheses. As a general rule, hypothesis tests that lead to vague claims are conducted with more power than hypothesis tests that lead to specific claims. This is consistent with day-to-day life. A detective may be certain that a burglary has occurred even though he or she may be unable to claim who committed the crime. You may be certain that your keys were misplaced, but you may not know whether it was you, your spouse, or one of your children who set them down in some errant location.

We saw that collinearity between regressors increases standard errors, yet collinearity does not lower the power of hypothesis tests on the entire set of variables in the model. As a result, it can be harder to come away from an analysis being able to claim that a specific regressor is related to Y than it is to make the much less specific claim that at least one of them is related to Y. Yet the inclusion of a set of regressors in a model that is in reality uncorrelated with Y increases sampling variance in the estimation of ${}_TR$ but not necessarily the standard error of regression coefficients for regressors that *are* correlated with Y. As a result, one might be left with a specific conclusion about a given regressor's unique relationship with Y even though the model, by a hypothesis-testing standard, does not explain more variance in Y than can be expected by chance.

In general, if a test of a specific null hypothesis is significant, and that test survives corrections for multiple tests described in Chapter 11, we can reject both that null hypothesis and any broader null hypothesis in which the specific null is imbedded. These issues are discussed further in Chapter 11.

4.7.3 Sample Size and Nonsignificant Covariates

To minimize standard errors and maximize the power of tests, should you limit the number of covariates or delete nonsignificant covariates from the model? Generally not. If a covariate correlates with no other variables in the analysis, then its inclusion lowers the power of tests on independent variables by the same amount as the loss of one case from the sample—both lower $df_{residual}$ by one. Thus, when sample sizes are moderate or large, little power is lost by adding a few extra covariates that turn out to be independent of other variables. A covariate X_i that correlates highly with an independent variable X_j can substantially increase $SE(b_j)$ and thus lower power of tests for the regression coefficient for X_j. But that very fact is usually evidence that it would be invalid to arbitrarily exclude covariate X_i from the analysis. The collinearity between X_i and X_j will also raise $SE(b_i)$, perhaps making it nonsignificant. Therefore, the nonsignificance of a covariate's regression weight is not a good reason for deleting it from the model. If you felt controlling for a variable is necessary for the sake of accurate inference about the phenomenon being modeled, that doesn't change just because its p-value is not small enough to reject the null hypothesis that its regression weight is zero.

A common but misleading rule of thumb is that a regression analysis should not contain more variables than one-tenth the sample size. But the most important tests are usually on values of b_j, and the power of those tests is determined not by the ratio N/k suggested by this rule of thumb but by $df_{residual}$, which usually is $N - k - 1$, which, of course, is determined by $N - k$. Thus, the power of the most important tests is determined by the difference between the sample size N and the number of regressors k, not their ratio (e.g., Green, 1991). When $N - k = 40$, a two-tailed test on b_j at the 0.05 level has power of 0.80 if ${}_Tpr_j = 0.43$, so one simple rule of thumb is that unless the effects of interest are believed to be quite large, the sample size should exceed k by 40 or more. But large samples allow a great variety of potentially useful analyses that are not practical in small samples, so the overriding rule is simply that larger samples are better than smaller ones. The ratio of cases to regressors has little relevance to inference in regression analysis.

4.7.4 Inference in Simple Regression (When $k = 1$)

Most books on regression analysis discuss inference in simple regression prior to extending the principles to models with more than one regressor.

We see regression with a single predictor as just a special case of multiple regression in which $k = 1$, and all the tests mentioned so far thus apply to this special case. The test on b_j in section 4.4 and the test on R described in section 4.3 apply to simple regression. The two tests then test the same null hypothesis: the hypothesis of no association between Y and the single regressor X. Further, the two tests are equivalent. The F found in testing R will always equal the square of the t found when testing b_1, and the two-tailed p of the t-test will be equal to the p in the F-test. Both tests are also equivalent to the following t-test on a simple correlation r for testing the null hypothesis that ${}_T r = 0$:

$$t = \sqrt{\frac{N-2}{1-r^2}}, \quad df = N - 2$$

In section 4.5.2 we provide a method for constructing a confidence interval for ${}_T pr_j$ that can also be used for constructing an interval estimate for ${}_T r$.

4.8 Chapter Summary

This chapter is dedicated to statistical inference in linear regression. After discussing the distinction between statistics and parameters, we covered the primary (linearity) and secondary (normality, homoscedasticity, and independence of errors in estimation) assumptions of valid inference, saving a more detailed discussion of these assumptions for Chapter 16. We addressed how to test hypotheses about the multiple correlation coefficient and various measures of partial association, such as the partial regression coefficient and partial and semipartial correlation. Whereas the partial regression coefficient is generally an unbiased estimator of its corresponding parameter, the multiple correlation is a biased estimator, but an adjustment reduces this bias.

Regression is particularly well-suited to answering questions about the relationships between correlated regressors and a dependent variable when other regressors are held constant. There is a limit, however, to how correlated regressors can be before damage is done to the inferential process. Collinearity can decrease the precision in estimation of measures of partial association, but as the formula for the standard error of a regression coefficient shows, high collinearity can be offset by other things under an investigator's control. As a result, collinearity—which is a fact of life and

not something with effects that can be eliminated by more sophisticated methods—is not necessarily as problematic as some people believe.

The prior chapters of this book are dedicated to building the foundation of your understanding of linear regression analysis principles, with an emphasis on statistical control, measuring partial association, and testing hypotheses using regression. The next chapter builds upon the foundation just laid by addressing additional topics in statistical control and linear models, such as dichotomous regressors, regression to the mean, and assessing the contribution of conceptually overlapping sets of variables to model fit and explaining variance in a dependent variable.

5
Extending Regression Analysis Principles

This chapter expands on the principles of linear regression and partial association introduced in Chapters 2 and 3. We show how linear regression does not require that all regressors be numerical; one or more can be dichotomous without modifying any of the underlying logic or mathematics of the modeling process. In this case, a linear model can be used to estimate the mean of *Y* for the two groups coded with the dichotomous regressor, with or without equating the groups on other regressors. The following section introduces *regression to the mean* and how linear models properly deal with the phenomenon in ways that other methods of analysis do not. We then generalize the measures of partial association explained in Chapter 3 to sets of regressors before closing with a look at how linear regression analysis can be extended and foreshadowing certain problems that we address later in the book.

5.1 Dichotomous Regressors

5.1.1 Indicator or Dummy Variables

The methods already described can easily be adapted to include dichotomous regressors such as a participant's biological sex or to which of two conditions a participant in an experiment is assigned. The use of categorical independent variables together with numerical covariates often goes by the name *analysis of covariance* (ANCOVA), but in fact ANCOVA is just a special case of the more general linear model discussed in this book. ANCOVA can be conducted with any regression program.

To illustrate, we add a person's sex to the model of weight loss that was used throughout Chapter 3. Suppose our sample of 10 people includes four women and six men. The sexes are distributed as shown in Table 5.1, with

TABLE 5.1. The Weight-Loss Data Set Including Sex (with Males Coded 0 and Females Coded 1)

ID	Exercise X_1	Food intake X_2	Metabolism X_3	Sex X_4	Weight loss Y
1	0	2	15	0	6
2	0	4	14	0	2
3	0	6	19	0	4
4	2	2	15	1	8
5	2	4	21	1	9
6	2	6	23	0	8
7	2	8	21	1	5
8	4	4	22	1	11
9	4	6	24	0	13
10	4	8	26	0	9
Means	2	5	20		7.5

males coded 0 and females coded 1. A dichotomous variable coded in this fashion is called an *indicator* or *dummy variable*. Any two numerical values can be used, but it is good and convenient for the purpose of interpretation to get into the habit of coding a dichotomous variable with two codes that differ by only 1 unit (e.g., 0 and 1; −0.5 and 0.5; and so forth).

For simplicity, we shall first consider the simple regression of weight loss (Y) on sex (X_4). A scatterplot is shown in Figure 5.1. We know from section 2.1.2 that if conditional Y means fall in a straight line, the regression line will pass through them. But in this figure there are only two conditional Y means. The straight line connecting those two conditional means is therefore the regression line.

5.1.2 Estimates of Y Are Group Means

If you do the calculations, you will find that the six men lost 7 units of weight per week on average (i.e., 700 grams), whereas the four women lost 8.25 units (825 grams) of weight on average per week. When Y is regressed on a dichotomous regressor, the resulting values of $\hat{Y}$ that the model generates correspond exactly to these two means. You can verify for yourself that regressing weight loss (Y) on sex (X_4) yields $\hat{Y} = 7.00 + 1.25X_4$ as the best-fitting ordinary least squares (OLS) regression model. Plugging zero (males) into this equation for X_4 yields $\hat{Y} = 7.00 + 1.25(0) = 7.00$,

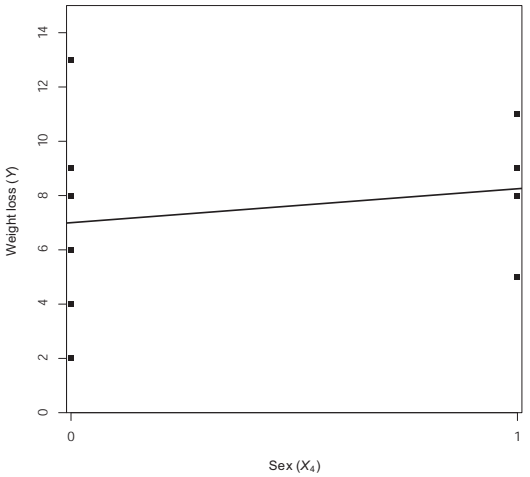


FIGURE 5.1. A scatterplot of weight loss against sex for the weight-loss data set.

which is $\overline{Y}$ for males. Similarly, plugging in one (females) yields $\hat{Y} = 7.00 + 1.25(1) = 8.25$, or $\overline{Y}$ for females. Observe that when using 0 and 1 to code groups, b_0 corresponds to $\overline{Y}$ for the group coded 0, and as discussed in section 5.1.3, b_1 is the difference between the group means.

A regression analysis with a dichotomous predictor generates the group means regardless of the two values used to code the two groups. For instance, suppose that rather than using 0 and 1 to code sex, we had used 1 and −1 for males and females, respectively. In that case, regressing weight loss on sex yields $\hat{Y} = 7.625 - 0.625X_4$. This equation yields $\hat{Y}$ values of 7.00 for males ($X_4 = 1$) and 8.25 for females ($X_4 = -1$), just as the regression model did when using 0 and 1 to code sex.

To anthromorphize regression analysis, it is fairly smart. Knowing only whether a case in the data file is male or female, it figures out that the best guess of Y for males is the mean of the males, and the best guess of Y for females is the mean of the females. It doesn't care how you code sex. It will figure out a regression coefficient and regression constant that reproduces the two group means regardless of your decision.

5.1.3 The Regression Coefficient for an Indicator Is a Difference

We have defined the slope of a line linking regressor X to Y as the amount Y rises for a 1-unit increase in the regressor. But in Figure 5.1, there is only one such increase in the regressor—the increase from 0 to 1. Thus, the slope equals the difference between the two conditional Y means. The means for males and females are 7.00 and 8.25, respectively, so their difference is 1.25, which is the regression coefficient in a linear regression estimating Y from X_4. In this example, the regression coefficient for X_4 is the difference between the mean weight loss for males versus females.

In sections 4.4.1 and 4.7.4 we presented a significance test for the null hypothesis that a true simple or partial regression coefficient equals zero. When there is only one regressor and it is dichotomous, that test is equivalent to the familiar two-group t-test for the null hypothesis that the group means are equal. In the present example, the difference between the two means is 1.25. Using the ordinary formula for the standard error of a difference between means and assuming equal population variances, we estimate the standard error of the difference to be 2.221. By the t-test, the significance of the difference is tested by computing $t = 1.25/2.221 = 0.562, df = N - 2 = 8, p = .59$ (two-tailed). The regression test in section 4.7.4 would give exactly the same values of t and p.

You need not code two groups with 0 and 1. When you do so, the regression coefficient for that variable is the difference between conditional Y means. This will be true even when using different codes, so long as they differ by 1 unit. We used males = 0 and females = 1 for X_4, so they do differ by 1 unit. The regression coefficient for X_4 would be the same if we used -0.5 and 0.5 instead. But if you chose different numerical codes that differed by more than 1 unit, such as -1 and 1, then this would change the regression coefficient. In this example, the regression coefficient would be one-half of the difference between conditional Y means. More generally, if the two groups are coded by values that differ by δ units—*whatever* their values—then the regression coefficient will be $1/\delta$ times the difference between the conditional Y means. Changing the coding of the groups to *any* two arbitrary values will not change the results of the inferential tests for the regression coefficient, and the model will still exactly generate the two group means.

Similar principles apply when the dichotomous regressor is one of several. In multiple regression analysis, the regression coefficient for a dichotomous regressor coded 0 and 1 can be interpreted as the Y difference between the two groups, adjusted for differences between groups on the

other regressors. For instance, when all four regressors (exercise, food intake, metabolism, and sex) are used in the current example, we find that the regression coefficient for sex is −0.404. This means that after adjusting or correcting for differences between the four women and six men on exercise, food intake, and metabolism, the women lost on average about 0.4 units or 40 grams *less* than the men per week—less because women were coded as 1 and men were coded 0 and the regression coefficient for sex is negative. By the test introduced in section 4.4.1, this difference is not statistically significant.

5.1.4 A Graphic Representation

In Figure 3.5 in section 3.1.6, we saw an example in which the regression slope of X_1 equals the vertical distance between parallel lines of best fit for X_2 where the lines represented X_1 values 1 unit apart. If there is a dichotomous regressor scored 0 and 1, then the adjusted slope or adjusted difference for the dichotomous regressor can be interpreted as the vertical distance between two parallel lines or planes of best fit for the other regressors. For instance, if we regress Y on the numerical regressors of exercise (X_1) and food intake (X_2) and the dichotomous regressor of sex (X_4), we find $b_1 = 2.216$, $b_2 = -0.584$, $b_4 = -1.008$ and $b_0 = 6.571$, so the model is

$$\hat{Y} = 6.571 + 2.126X_1 - 0.584X_2 - 1.008X_4$$

When $X_4 = 0$, this equation reduces to

$$\hat{Y} = 6.571 + 2.126X_1 - 0.584X_2$$

whereas when $X_4 = 1$, it reduces to

$$\begin{aligned} \hat{Y} &= 6.571 + 2.126X_1 - 0.584X_2 - 1.008 \\ &= 5.563 + 2.126X_1 - 0.584X_2 \end{aligned}$$

Each of these models can be represented by a tilted plane. Putting both planes in the same figure gives Figure 5.2, in which the upper tilted plane represents the model for men and the parallel plane just below it represents the model for women. The vertical distance between the two planes is 1.008, which is b_4. This is the difference between the average weight loss of the four women and six men in the sample, adjusted for differences between them on exercise and food intake.

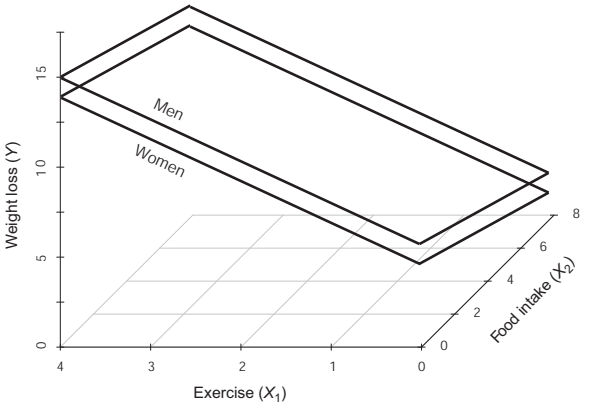


FIGURE 5.2. A model of weight loss that fits separate but parallel planes for men and women.

5.1.5 A Caution about Standardized Regression Coefficients for Dichotomous Regressors

As introduced in section 3.3.3, a regression coefficient can be expressed in standardized or unstandardized form. Regression coefficients are in standardized form if one estimates Y from one or more regressors when Y and all regressors are first standardized (converted to variables with mean 0 and standard deviation 1). We introduced $\tilde{b}$ as our notation for the standardized regression coefficient to distinguish it from the unstandardized regression coefficient b.

Although it is not uncommon for investigators to report a regression model in standardized form, we discourage the reporting of standardized regression coefficients for dichotomous regressors. There are two problems with doing so. First, standardization destroys the convenient interpretation of the regression coefficient as a mean difference. Remember that the regression coefficient for X in a regression model corresponds to the estimated difference in Y between two cases that differ by 1-unit on X (adjusting for other regressors if others are in the model). So when expressed in unstandardized form, the regression coefficient for a dichotomous regressor when the groups are coded by a 1-unit difference is equal to the difference between the group means on Y. However, this will not be true for the standardized regression coefficient, because the two groups will not differ by 1 standard deviation on the variable coding groups following standardization. That is, $\tilde{b}$ cannot be interpreted as a mean difference regardless of how X is coded.

Given that substantive questions involving dichotomous regressors in a regression model almost always focus on differences between groups, we think it makes more sense to keep a dichotomous regressor coded in a form that yields a regression coefficient with a mean difference interpretation.

Second, to make matters worse, $\tilde{b}$ is affected by both the difference between group means on Y and the distribution of the cases across the two groups. To illustrate, consider a dichotomous regressor X coded 0 and 1. The standard deviation of X is $\sqrt{P_0 \times P_1}$, where P_0 is the proportion of the individuals in the sample who are members of the group coded $X = 0$ and P_1 is the corresponding proportion of the individuals in the sample who are members of the group coded $X = 1$. So when the sample is equally split between the two groups, $P_0 = 0.5$ and $P_1 = 0.5$, meaning that $s_X = \sqrt{0.5 \times 0.5} = 0.5$. But when the sample is split, say 4:1 in favor of one group, meaning $P_0 = 0.80$ and $P_1 = 0.20$, then $s_X = \sqrt{0.80 \times 0.20} = 0.40$. In the former case, the two groups differ by 1 unit on X prior to standardization but 2 *standard deviations* on X after standardization. In the latter case, the two groups of course still differ by 1 unit on X before standardization, but they differ by 2.5 standard deviations on X after standardization.

The implications of this are important. Suppose that the two groups differ by 2 points on Y on average, and $s_Y = 4.00$, meaning that they differ by half a standard deviation on Y on average. If X is coded 0 and 1 and the groups are equally split between the two groups (i.e., $P_0 = P_1 = 0.5$) and you estimate $\hat{Y} = b_0 + b_1 X$, then $b_1 = 2.00$, but $\tilde{b}_1 = 0.25$. This should make sense; b_1 corresponds to the difference between group means on Y, as discussed earlier. And remember that in the standardized regression model, the regression coefficient for dichotomous predictor X is the estimated number of standard deviations on Y by which two cases are estimated to differ who differ by 1 standard deviation on X. The two groups differ by 2 standard deviations on X, so $\tilde{b}_1$ is one-half of the difference between means on standardized Y following standardization of X.

But suppose that instead your sample heavily favored one group, with four times as many cases in group $X = 0$ than in group $X = 1$, but everything else was otherwise the same, with the group means differing by 2 units on Y and $s_Y = 4.00$. So the two group means differ by the same amount on Y in either standardized or unstandardized form as when the groups were equal in size. Whereas b_1 in this regression will still be 2.00, which is the difference between the group means on Y, $\tilde{b}_1$ is not 0.25 as before but, rather, 0.20. So the same mean difference on Y has resulted in a smaller value of $\tilde{b}_1$,

but b_1 is unaffected. Furthermore, b_1 still has its convenient mean difference interpretation, but $\tilde{b}_1$ does not.

This means that $\tilde{b}_1$ can't be interpreted sensibly without considering information about the distribution of the sample across the two groups, whereas b_1 does not require this information. In addition, any attempt at generalizing $\tilde{b}_1$ to the population must condition that inference on a population with that same split on the dichotomous variable. If your sample represents the population with respect to the distribution of X, no problem, but often various processes at work can result in a loss of representativeness of the distribution of one variable relative to the population (e.g., if some members of the population are more likely to refuse to participate in the study, or if you have intentionally oversampled one group). But b_1 doesn't have such a constraint on generalization. Finally, you can't compare standardized regression coefficients in the same model that includes a dichotomous regressor but estimated in two different samples that differ in the distribution of the dichotomous variable, even if the distribution on Y is exactly the same in the two samples. If the two samples differ in variance in Y as well as the distribution of cases across the two groups coded with X, comparability is even further reduced. But b_1 does not suffer from this lack of comparability. (Note that this applies to comparisons of standardized regression coefficients for numerical regressors as well).

This unfortunate property of standardized regression coefficients for dichotomous regressors generalizes to models with multiple regressors, although the argument is made slightly more complex by the fact that the distribution of cases across the two groups also affects the correlation between the dichotomous variable and the other regressors. Regardless, if you want to report a standardized regression model, we recommend you standardize all variables manually but leave the dichotomous regressors in their original, unstandardized form, prior to estimating standardized Y from the regressors. When you do so, the proper regression coefficients to interpret are the ones identified in your regression program output as *unstandardized* coefficients (which are actually standardized coefficients for those regressors that were standardized prior to analysis).

5.1.6 Artificial Categorization of Numerical Variables

ANOVA is a staple in the curriculum of undergraduate and graduate programs in psychology and related disciplines. As described later in Chapter 9, ANOVA is just a special case of a more general linear model that is the topic of this book. Unfortunately, researchers accustomed to thinking in

ANOVA terms often take numerical variables and categorize them prior to analysis, probably in order to fit the data to the analytical approach to which they are most comfortable or accustomed. For instance, rather than analyzing age as a continuum, a researcher might classify people into age groups such as 20s, 30s, 40s, and so forth, and treat this as a categorical variable using ANOVA. Or perhaps a media effects researcher measures television viewing frequency in number of hours per day but then classifies people into light, moderate, and heavy viewers based on how many hours they report.

Typically, such categorization of people based on numerical data is done arbitrarily. It may be based on predetermined criteria or on the distribution of the numerical data. For instance, a light television viewer might be defined as someone who watches less than 1 hour per day, a moderate viewer as between 1 and 3 hours, and a heavy viewer as someone who watches more than 3 hours per day. Alternatively, the researcher might attempt to divide the sample based on the distribution, such that those in the lower third of the distribution are classified as light, the middle third as moderate, and the highest third as heavy viewers. Even more coarse categorization is possible and often undertaken, such as the construction of high and low groups based on whether a person is above or below the sample mean or median.

Such categorization is no doubt most typically undertaken because the researcher is familiar with ANOVA but not its more general linear model form that doesn't require people to be put into groups. Such categorization is not necessary using the procedures described in this book. If one's hypothesis is that people who watch more television are likely to be less healthy, as measured by something like the body mass index (BMI), it is not necessary and can even be damaging to the analysis to classify people into groups of TV viewing frequency and then conduct an ANOVA comparing the group means. Better would be simply to regress BMI on the number of hours the person reports watching TV and examine various measures of linear association. Doing so generally is more valid and more powerful because categorization throws away information about association, treats people who are very similar as if they are maximally different, and increases measurement error. There is a large literature admonishing researchers not to artificially categorize variables that are or can be measured numerically or continuously (Cohen, 1983; Irwin & McClelland, 2002; Kuss, 2013; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993; Rucker, McShane, & Preacher, 2015). We largely agree with that literature.

With this in mind, there are a few circumstances in which artificial categorization of numerical variables can be justified. One such circumstance is when a quantitative measurement procedure results in a severely bimodal distribution, with very little variation around two points. For instance, you might ask people to place themselves on a 10-point scale with respect to how much (1 = *not at all*, 10 = *very much*) they support a particular government proposal to reduce the potential threats of global climate change. Suppose you find that most everyone responds with either a 1 or 10, with a few people giving a 2 or 9, and still fewer responding with 3 or 7, with no one giving a 4, 5, or 6 response. In that case, there probably isn't much to be gained by recognizing the distinction between a response of 1 and 2, or between a 9 and 10. The results of the measurement procedure reflect that there really are just two groups of people—those who support the proposal and those who don't. It would probably be safe in this case just to code people dichotomously as opponents or proponents of the proposal.

Another circumstance in which categorization based on a numerical variable could be sensible is when the variable you are really interested in is dichotomous by nature but you have measured it in such a way that you have more information than you need to use. For example, perhaps you have asked people in a questionnaire how many cigarettes a week they typically smoke. If all you really care about for the sake of analysis is whether someone is a smoker or not, it would be sensible to classify anyone who responds zero as a nonsmoker, and consider anyone who gives a response other than zero as a smoker.

Finally, in some fields, whether or not a measurement exceeds a threshold is considered important for theoretical or applied reasons, and you may want to honor what is commonplace in that field. For example, in clinical psychology, a person might be considered eligible for a particular diagnosis if he or she has at least five symptoms from a list of many possible symptoms. Although you could do the analysis using how many symptoms the person reports as a regressor, it may be clinically more meaningful or yield results that are easier to apply if you simply categorized people as eligible or not for the diagnosis prior to the analysis.

But don't take these decisions lightly. The consequences of artificial categorization are potentially severe. Make sure you think through the decision, and be prepared to tell consumers of your research your justification for discarding information that you had prior to analysis. As a general rule, avoid the use of mean or median splits or other arbitrary approaches to creating categories out of quantitative variables. Assume that your reader is

aware of the pitfalls of artificial categorization and will be skeptical of your decision. Thinking about it this way will force you to ponder your decision and make an informed one.

5.2 Regression to the Mean

5.2.1 How Regression Got Its Name

In its most general form, regression to the mean applies whenever two variables X and Y are correlated less than perfectly in a sample of cases. The principle asserts that if we select a subsample of cases with extreme measurements on X, then the subsample's mean on Y will almost always be less extreme than its mean on X—that is, it will *regress* toward the mean of the total sample. This section explains why and shows how the phenomenon can lead the unwary researcher into a variety of errors that are avoided by the proper use of linear models. When Sir Francis Galton first noticed the phenomenon in the late 19th century, he considered it so important that the linear models he used came to be known as *regression* models.

5.2.2 The Phenomenon

Galton noticed the phenomenon when he studied the heights of a large sample of middle-aged men and their grown sons. He observed that most of the older men who were above average in height had sons shorter than they, while most of the older men who were shorter than average had sons taller than they. In other words, the heights of the sons were regressing toward the mean height. This regression seemed to imply that the younger generation was more homogeneous in height than the older generation. But this conclusion was not supported; the standard deviation of the sons' heights was found to be almost exactly equal to the standard deviation of the fathers' heights. This equality of standard deviations seem to contradict the regression; how could both be true?

We know that this paradox was caused not by any peculiar properties of the English or father–son pairs or height, but by the very general statistical phenomenon of regression to the mean. For simplicity, consider a sample of 19 father–son pairs and their heights, depicted in the scatterplot in Figure 5.3. The solid line is not the best-fitting regression line but rather the line of equality, meaning that this line represents father–son pairs with identical heights (i.e., $Y = X$). Dots above this line represent father–son pairs with

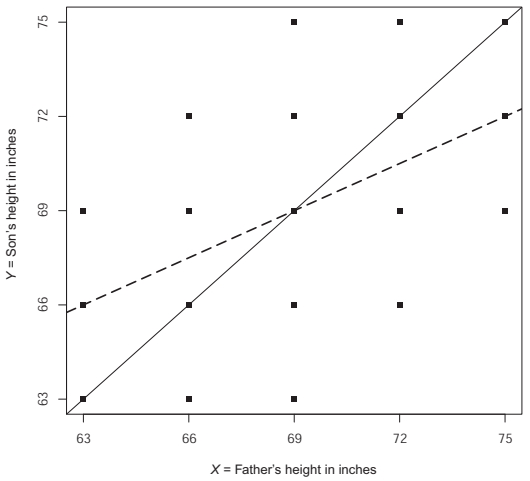


FIGURE 5.3. Sample data set illustrating regression to the mean.

sons taller than their fathers (i.e., $Y > X$). Dots below this line correspond to father–son pairs with fathers taller than their sons (i.e., $Y < X$).

Now observe a number of things. First, the marginal distributions of the heights of fathers and sons are identical. That is, each distribution has three heights in inches of 63, four heights of 66, five heights of 69, four heights of 72, and three heights of 75. Thus, they have the same mean and the same standard deviation. Second, only five of the sons are exactly the same height as their fathers. More often, the son is either taller than the father (7 pairs) or shorter than the father (7 pairs).

The average height of the fathers is 69 inches, as is the average height of their sons. Consider the seven fathers who are below average in height—the seven leftmost dots in Figure 5.3. Notice that most of their sons are taller (4 of 7), and only one of the sons of these relatively shorter fathers is shorter than his father. Furthermore, these sons are, on average, taller than their fathers and closer to overall mean height than their fathers. Consider the three fathers who are 63 inches tall. Their sons are, on average, 66 inches tall. Similarly, the four fathers who are 66 inches tall have sons who are, on average, 67.5 inches tall.

Consider next the seven fathers who are *above* average in height. These are the seven rightmost dots in Figure 5.3. Most of their sons are shorter

than they are (4 of 7), and only one is taller. Consequently, the sons of these seven fathers are, on average, shorter than their fathers and closer to the mean height, on average, than their fathers. Examining the three tallest fathers first—each of whom is 75 inches tall—notice that their sons are 72 inches on average. And the four fathers who are 72 inches tall have sons who are, on average, only 70.5 inches tall.

This example has all the features of the paradox that Galton observed: Most taller-than-average fathers have sons shorter than they, and most shorter-than-average fathers have sons taller than they. Yet sons are not less variable in height than are fathers, as the standard deviations of sons' and fathers' heights are the same.

Notice as well that this effect is symmetrical, in that most of the tallest *sons* have fathers shorter than they, and most of the shortest sons have fathers taller than they. That would lead you to think that the standard deviation of the sons' heights exceeds that of the fathers' heights, while the original paradox led you to believe that the opposite was true. Yet neither is true. The standard deviations of the two distributions are the same.

Regression to the mean can also be described in terms of the *gain* or *difference* score $Y - X$. So in this example, $Y - X$ would be the difference between the height of a son and the son's father. A positive difference means the son is taller than the father, and a negative difference means the son is shorter than the father. Regression to the mean implies that the difference $Y - X$ is negatively related to X. That is, the difference between the height of the son and the father is negatively correlated with the height of the father. The 19 father–son pairs are depicted again in Figure 5.4 but with $Y - X$, the difference between the son and father's height, on the Y-axis rather than just the son's height. As can be seen, the correlation between $Y - X$ and X is negative; here, Pearson's $r = -0.50$. The fathers who are below average in height tend to have sons who are taller than they are are (i.e., $Y - X > 0$), and the fathers who are above average in height tend to have sons who are shorter than they are (i.e., $Y - X < 0$).

When framed in terms of difference scores, regression to the mean becomes very important to acknowledge whenever you conduct research that involves measuring change over time. Consider a sample of people whose depression is measured at a certain time, such as before therapy begins, then again, perhaps after several sessions of psychotherapy, using the same depression inventory. Unless you force it to be otherwise through the sampling procedure, there will be differences between people in depression at time 1. That is, some people will be more depressed than others; there

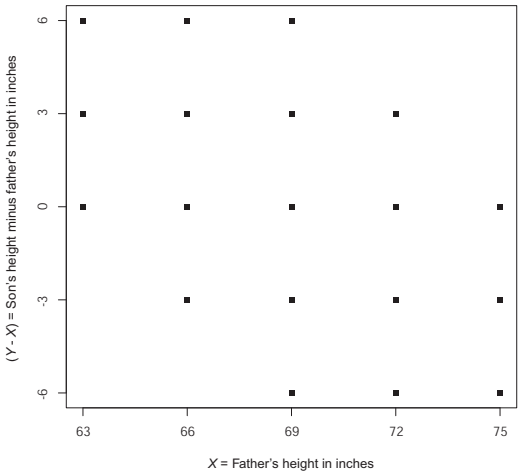


FIGURE 5.4. The negative correlation between gain and *X*.

will be some variation. Furthermore, depression scores at time 1 and time 2 are likely to be positively correlated, but not perfectly.

Now suppose that the therapy is not effective, meaning that, on average, people are just as depressed after therapy than they were before. Further suppose that therapy does not affect variation in depression at time 2 relative to time 1 (i.e., the standard deviations at both times are about the same). In that case, unless depression at time 1 and time 2 are perfectly correlated, regression to the mean implies that people who are relatively more depressed at time 1 are expected to improve at time 2, and people who are relatively less depressed at time 1 are likely to be worse at time 2. In other words, the correlation between depression at time 1 and improvement in depression from time 1 to time 2 following therapy is likely to be negative. Failure to acknowledge regression to the mean could lead to the interpretation that therapy helps people who are worse off prior to therapy but harms people who are not so bad to start with.

5.2.3 Versions of the Phenomenon

In both the father–son example and the depression and therapy example, the means and standard deviations were the same. However, regression to the mean is more general than this, and equality of the means or standard

deviations is not a requirement for regression to the mean to occur. Given this, it is worth acknowledging various forms that regression to the mean can take. Consider four conditions:

1. Two variables X and Y correlate imperfectly.
2. X and Y are measured in equal units.
3. X and Y have equal standard deviations.
4. X and Y have equal means.

When all four conditions hold, as in the two prior examples, we can say that in a subsample of people selected for their high values of X, their values on Y will tend to be below their X values. Similarly, in a subsample of people selected for their low values on X, their values on Y will tend to be above their X values.

If we discard the fourth condition that X and Y have the same means, then we can still say that in a subsample of people scoring below the mean on X (or above the mean on X), their values of Y will tend to be *closer to the overall mean* of Y than their values of X are to the overall mean of X. The first three conditions are the minimum conditions for asserting confidently that difference correlates negatively with X. For instance, if it is true that "the rich get richer and the poor get poorer," then gain in wealth correlates *positively* with initial wealth—but then inequality of wealth is increasing, meaning that standard deviations of wealth at the two times are not the same.

If we further discard the second and third conditions, then it is no longer true that a subsample selected for their relatively high (or relatively low) values of X will necessarily have values of Y that are closer to $\overline{Y}$ than their values of X are to $\overline{X}$. But even given only condition one—imperfect correlation between X and Y—we can still say that cases that are extreme in the distribution of X relative to $\overline{X}$ will tend to be less extreme in the distribution of Y relative to $\overline{Y}$. For instance, let X and Y be height and weight. These variables are not measured on the same scale, so it is not even meaningful to ask if their means and standard deviations are the same. But it is still meaningful to ask whether a specific person is more extreme on height than on weight. For this case, regression to the mean says that if you select the most extreme people on either variable, most of them will probably be less extreme on the other variable.

5.2.4 Misconceptions and Mistakes Fostered by Regression to the Mean

Regression to the mean can be easily overlooked, even by experts in linear models. A failure to recognize this phenomenon results in some common misconceptions and potential mistakes. One of these we have already addressed, and that is assuming that the standard deviation of Y must be less than that standard deviation of X. We have seen the fallacy of this conclusion.

An additional mistake is incorrectly accounting for regression to mean when attempting to estimate a person's measurement on one variable from his or her measurement on another. If a person is 1 standard deviation below the mean on one variable X, it is all too easy to assume if X and Y are correlated that this person is likely to be around 1 standard deviation below the mean on Y as well. But recall from Chapter 2 that in a simple regression model of the form $\hat{Y} = b_0 + b_1 X$,

$$b_1 = r_{XY}\frac{s_Y}{s_X}$$

If X and Y have been standardized, they have standard deviations of 1, so $b_1 = r_{XY}$. Furthermore, since standardized variables have means of zero, we also have

$$b_0 = \overline{Y} - b_1\overline{X} = 0 - b_1 \times 0 = 0$$

Therefore, when X and Y are standardized,

$$\hat{Y} = 0 + b_1 X = r_{XY} X$$

So the best estimate of Y for someone who is 1 standard deviation below the mean on X, meaning $X = -1$, is not $Y = -1$ but $Y = -r_{XY}$, or r_{XY} standard deviations below the mean. More generally, if someone is k standard deviations from the mean on X, meaning $X = k$, the best guess for how many standard deviations from the mean that person is on Y is $k \times r_{XY}$. Regression was named so in part for its ability to estimate accurately the amount of regression to the mean in problems like this.

A failure to recognize regression to the mean can lead a person to invent some kind of causal mechanism to explain the phenomenon when it is adequately explained by chance variation. For instance, suppose in a statistics course, students take a midterm and a final exam. If the two exam scores do not correlate perfectly, which they likely will not, then students with the highest scores on the midterm are not likely to be highest on the

final as well. Similarly, those who do worst on the midterm are not likely to be worst on the final. If you happened to notice this in a set of exam scores, you might think that those who scored highest on the midterm must not have studied as hard for the final, while those who scored lowest on the midterm worked especially hard to overcome their deficits produced by their first exam performance. But, in fact, regression to the mean can account for such a pattern. Whenever two variables are less than perfectly correlated, extreme measurements on one variable tend to be paired with measurements on the other that are less extreme.

Finally, analysts not familiar with regression to the mean often make the mistake of using the difference between measurements of the same variable over time as a dependent variable in experiments lacking random assignment, without also including the initial measurement as a covariate. As discussed in section 5.2.5, there is little gained in the use of such a difference score relative to just using the second measurement as the dependent variable and modeling it using the first measurement as a covariate.

5.2.5 Accounting for Regression to the Mean Using Linear Models

Linear modeling formally accounts for regression to the mean by generating estimates of Y that will be less extreme from $\overline{Y}$ than X is from $\overline{X}$. To illustrate, consider the data in Figure 5.3. Regressing sons' heights (Y) on fathers' heights (X) yields

$$\hat{Y} = 34.5 + 0.5X$$

This regression equation is depicted graphically with the dotted line in Figure 5.3. Consider the three fathers who are 63 inches tall, meaning 6 inches below the mean height of fathers. The regression equation estimates that their sons are, on average, 34.5 + 0.5(63) = 66 inches tall, which is 3 inches below the mean of sons' heights. This is also the conditional mean of the heights of the sons of 66-inch-tall fathers depicted in Figure 5.3. Similarly, this model estimates that the height of the son of a 72-inch-tall father (3 inches above the mean height of fathers) is 34.5 + 0.5(72) = 70.5 inches, which is only 1.5 inches above the mean height of sons. Recalling that standard deviations of fathers' and sons' heights are the same in this example, it is clear that this equation generates estimates of Y that are less extreme than X.

This property of the regression equation can also be seen by expressing it in terms of standardized heights and correlations. The standard deviations of the heights of sons and fathers are both 3 inches, and the correlation

between their heights is 0.50. Thus, fathers who are 2 standard deviations below the mean height (63 inches) are estimated to have sons who are only $0.5(-2) = -1$ standard deviation from the mean height of sons, or 1 standard deviation below the mean, which is 63 inches. And fathers who are 1 standard deviation above the mean (72 inches) are estimated to have sons who are $0.5(1) = 0.5$ standard deviations above the mean height of sons, or 70.5 inches. So the heights of sons are estimated to be less extreme from the mean than their fathers' heights.

A linear model also properly accounts for regression to the mean when a variable Y is measured over time and interest is in predicting change in Y. For example, suppose you wonder whether a driver training course is more effective in reducing the accident rate among bad drivers who have been compelled to take such a course by a traffic court than it is among ordinary drivers not so compelled. So you obtain accident rates of two groups of people before taking the course, those compelled to take the course because of their tendency to get in accidents and those who simply chose to do so. After the course is completed and some time passes, you obtain the accident rates of these same people. Regression to the mean implies that you are likely to find that the poor drivers compelled to take the course improved more than those who took it by choice, because the poor drivers were likely extreme in the distribution of accident rate prior to taking the course. Remember that regression to the mean says that extremes on one variable tend to be paired with less extremes on the other when the variables are not perfectly correlated. Poor drivers have less room to get worse and more room to improve than ordinary drivers. This is regression to the mean.

Analytically, the problem can be handled with a linear model. Let Y_2 be the postcourse accident rate, Y_1 be the accident rate prior to taking the course, and X be a dichotomous variable coding whether ($X = 1$) or not ($X = 0$) the person was required by a court to take the course. It would be unwise to estimate the effect of the training course using a simple model of the difference score $Y_2 - Y_1$:

$$Y_2 - Y_1 = b_0 + b_1 X + e$$

In this model, b_1 is typically interpreted as an estimate of the average difference in change in accident rate between those compelled to take the course and those who choose to take it. To those familiar with ANOVA, the test of significance for b_1 is mathematically identical to a test of interaction between time and condition in a mixed factorial analysis of variance with

time as a within-subjects factor and condition as a between-subjects factor. And the *F*-ratio for this interaction is the square of the *t*-statistic for b_1 when testing the null hypothesis that ${}_T b_1 = 0$. We discuss interaction in detail in Chapters 13 and 14.

But this model (whether labeled a regression analysis or a mixed ANOVA) will fail to account for regression to the mean, and the differential effectiveness of the course as estimated by b_1 is likely to be biased. One way of recognizing this is to consider that this model can be reexpressed in identical form as

$$Y_2 = b_0 + b_1 X + b_2 Y_1 + e, \text{ where } b_2 = 1$$

This is an *improper linear model* of postcourse accident rate, in the sense that you are fixing the weight given to prior accident rate to $b_2 = 1$ in the model rather than letting the least squares algorithm figure out how to best weight prior accident rate in order to minimize the sum of the squared residuals. If X is correlated with Y_1, as it likely to be in this example (i.e., those who are compelled to take the course by a traffic course presumably have worse accident rates prior to the course than those who choose to take it), the difference between what b_2 would be if properly estimated compared to when it is fixed to 1 will get absorbed at least in part into b_1, which biases the estimate of the effect of being compelled to take the course on postcourse accident rate.

The proper approach to answering the question as to whether those compelled to take the course improve more than those not forced to take it is to estimate postcourse accident rate Y_2 from X while using precourse accident rate Y_1 as a covariate, as in

$$Y_2 = b_0 + b_1 X + b_2 Y_1 + e$$

and letting your OLS regression program figure out how to weight Y_1 rather than fixing b_2 to 1.

An alternative might have occurred to you. Since the difference in accident rate before and after taking the course seems like the most direct measure of change, why not keep the difference score as the dependent variable but use precourse accident rate as a covariate? That is, how about if we estimate

$$Y_2 - Y_1 = b_0 + b_1 X + b_2 Y_1 + e \tag{5.1}$$

This would account for the fact that those whose accident rates are high are likely to improve more, as expected by regression to the mean. Any change

independent of prior accident rate will show up in b_1, in accordance with the principles introduced in Chapter 3.

But it is obvious once you see it that b_1—the quantity of direct interest to the question—will be the *same* regardless of whether Y_2 or $Y_2 - Y_1$ is the dependent variable, so long as Y_1 is used as a covariate. Equation 5.1 can be rewritten in identical form as

$$Y_2 = b_0 + b_1 X + (b_2 + 1)Y_1 + e$$

This means that b_2 in a model of Y_2 with Y_1 as a covariate is simply one more than b_2 in a model of the difference score $Y_2 - Y_1$ when Y_1 is a covariate (cf. Allison, 1990; van Breukelen, 2013). But one of the morals of this section is that the choice of dependent variable is of no consequence, as b_1 is of primary interest, and it will be the same in each of these models. So although there is no harm in using the difference score as the dependent variable, mathematically there is no advantage to doing so. But, and this is the second moral, it is mistake to model change over time (pre to post) without including premeasurement as a covariate. Doing so fails to account for regression to the mean, and this can bias the estimate of the effect of other variables in the model if they are correlated with premeasurement.

5.3 Multidimensional Sets

Many measures of interest can be thought of as multidimensional. For instance, we could probably construct a simple self-report attitude measure containing a couple of questions that most people would agree measure at least roughly what they mean by the term *politically liberal*. But if we wish to assess this in a more comprehensive manner, we might create several different scales or use several different things that measure an aspect of what it means to be politically liberal. For instance, we might think about measuring people with respect to not only the liberalness of their political perspective on matters of foreign policy, but also on economic policy and various social issues. We could also ask them how many times they have voted for the candidate in a major political race who could be described as liberal. Or we could quantify someone's SES by using his or her income, his or her occupational prestige, and how many years of education he or she has received.

To say that some variable Y is unrelated to a set of other variables is to say that it is uncorrelated with *all* the variables in the set. For instance, to say that a measure of political ideology is unrelated to SES as operationalized

above is to say that it is uncorrelated with all three of the variables in the set used to define SES. More formally, this means that the multiple correlation R in a model estimating political ideology from the three variables in the SES set is equal to zero.

A single analysis may contain two or more sets of variables. For instance, we might have a set of regressors we call the *demographics set*, such as age, sex, and race, and a set we call the SES set, which includes such things as income, occupational prestige, and education. Or one could have a *miscellaneous set* of variables that do not squarely fit into other sets of variables used in the model.

Using the indices of *setwise partial association* described in this section, one would not need to aggregate and reduce all these variables into a single score but, rather, they could be treated as distinct measures of separate but related things by quantifying the relationship between the set and some outcome variable of interest, while controlling for other variables in the analysis.

5.3.1 The Partial and Semipartial Multiple Correlation

In section 3.4.1 we said that regressor j's *unique contribution* to a regression can be defined as the amount R^2 would drop if the regressor were removed from the analysis. Alternatively, it is the amount by which R^2 increases when it is added to a model without it. We saw that this unique contribution equals 0 if and only if $sr_j = 0$, which implies that the other measures of partial relationship (b_j, $\tilde{b}_j$, and pr_j) also equal 0. These ideas can be extended to sets of variables.

Let A and B denote two *sets* of variables. For example, let set A include several demographic measures, while set B includes several SES measures. We shall define the *partial multiple correlation* $PR(B.A)$ as the multiple correlation between Y and set B with all the variables in set A held constant. To be precise, if m_B is the number of variables in set B, imagine m_B separate regressions in which each of these B variables is regressed on (predicted from) all the variables in set A. The residuals in all these regressions give us m_B variables that we can call the *portions* or *components* of B *independent of* A. Imagine also regressing Y onto the set of A variables. The residuals for this model are the portion of Y independent of A. Finally, imagine regressing these residuals for Y onto the portions of set B independent of A (i.e., the residuals from estimating each of the B variables from all of the A variables). The multiple correlation in this regression is $PR(B.A)$. If set B

has only one variable, the value of $PR(B.A)$ equals the absolute value of the ordinary partial correlation.

We can similarly define a *semipartial multiple correlation* $SR(B.A)$ as the correlation between *all* of Y and the portions of set B independent of set A. Again, if set B has only one variable, $SR(B.A)$ reduces to the absolute value of the ordinary semipartial correlation.

We do not actually have to compute all these residuals to calculate $PR(B.A)$ and $SR(B.A)$, as they are easily calculated from other statistics. Define $R(A)$ as the multiple correlation from a model estimating Y from only set A variables and define $R(AB)$ as the multiple correlation from a model estimating Y from variables in both set A and B. Then

$$SR(B.A)^2 = R(AB)^2 - R(A)^2$$

while

$$\begin{aligned} PR(B.A)^2 &= \frac{SR(B.A)^2}{1 - R(A)^2} \\ &= \frac{R(AB)^2 - R(A)^2}{1 - R(A)^2} \end{aligned}$$

These formulas are consistent with previous interpretations of partial and semipartial correlations. $SR(B.A.)^2$ is the *unique contribution of set B* to the regression. If we think of Y as standardized to unit variance, then $SR(B.A)^2$ is the proportion of the Y variance explained by the variables that define set B independent of A. It is also the amount R^2 increases when the variables in set B are added to a model of Y including set A variables as regressors.

The denominator $1 - R(A)^2$ of the last ratio is the proportion of the Y variance unexplained by A, so $PR(B.A)^2$ is the proportion of remaining variance in Y (i.e., the proportion not accounted for by A) that can be uniquely explained by set B. A similar interpretation of pr_j^2 was given in section 3.3.

A close examination of these formulas reveals that $PR(B.A)^2$ and $SR(B.A)^2$ differ in their reference point for gauging variance explained by set B. Whereas $SR(B.A)^2$ indexes set B's unique contribution to the model as relative to *all* the variance of Y—whether explained or unexplained by set A—$PR(B.A)^2$ indexes set B's contribution as relative to only the variance in Y that remains unaccounted for by set A.

The formulas also show clearly another useful fact about a partial or semipartial multiple correlation: To say that $PR(B.A) = 0$ or $SR(B.A) = 0$ is

to say that the multiple correlation between Y and set A equals that between Y and sets A and B together. Adding the regressors in set B to a model that includes the regressors in set A does not increase the regression model's ability to explain variance in Y.

To illustrate these computations, suppose we want to estimate how much of the variance in weight loss can be explained by things that can more easily be controlled by a person—exercise (X_1) and food intake (X_2)—after accounting for things that might affect weight loss that are harder or impossible for a person to control—metabolism (X_3) and biological sex (X_4). Define set A as those factors not under personal control and set B those things under a person's control. If we first regress weight loss on set A, we get $\hat{Y} = -3.669 + 0.529X_3 + 1.470X_4$, $SS(A)_{regression} = 46.734$, $R(A)^2 = 0.474$. When the variables in set B are added, the result is $Y = -0.967 + 1.151X_1 - 1.133X_2 + 0.6X_3 - 0.404X_4$, $SS(AB)_{regression} = 91.703$, $R(AB)^2 = 0.931$. We don't need the regression sum of squares in these computations but we use them later so we include them here.

Given these statistics,

$$\begin{aligned} SR(B.A)^2 &= R(AB)^2 - R(A)^2 \\ &= 0.931 - 0.474 \\ &= 0.457 \end{aligned}$$

$$\begin{aligned} PR(B.A)^2 &= \frac{SR(B.A)^2}{1 - R(A)^2} \\ &= \frac{R(AB)^2 - R(A)^2}{1 - R(A)^2} \\ &= \frac{0.931 - 0.474}{1 - 0.474} \\ &= 0.869 \end{aligned}$$

So we can say that of these four factors, the ones most under a person's control (food intake and exercise) uniquely explain about 45.7% of the variance in weight loss, holding constant the ones less under control (sex and metabolism). We can also say the factors more under personal control explain about 86.9% of the variance in weight loss that remains after accounting for the factors less under control.

5.3.2 What It Means If $PR = 0$ or $SR = 0$

In the SES example, the word *correlation* strongly suggested positive simple correlations, but that is not a necessary property of multiple and partial multiple correlations. For instance, suppose we want to examine the relationship between a city's rate of homelessness and both poverty rate and availability of public housing. Four sociologists might have very different ideas as to how these variables relate to homelessness rates. One might guess that the homelessness rate is related primarily to the poverty rate, while a second supposes that it is related primarily to the availability of public housing. A third might advance the prediction that homelessness is related to the *difference* between the poverty rate and the availability of public housing expressed as a proportion of the city's population. Yet a fourth might argue that only two-thirds of poor people want to live in public housing, so that homelessness might be related to the difference between the availability of public housing and two-thirds of the poverty rate.

These views are quite different, but all imply that there is a correlation between homelessness rate and *some* linear function of poverty rate and public housing availability, and so all imply a nonzero multiple correlation between the homelessness rate and the other two variables. If the multiple correlation were found to be zero, this would contradict all four of these views plus many others involving other specific combinations of the two variables. Thus, a hypothesis test on a single multiple correlation of the type introduced in Chapter 4 can test an entire array of specific hypotheses about combinations of regressors. This is true even if some of the hypotheses involve negative correlations.

A related set of questions might be answered by a partial multiple correlation. Homelessness might be higher in larger cities or in cities with warmer winter climates. Suppose we want to see whether homelessness relates to poverty and public housing while controlling for city size and average winter temperature. We could then include a city's population and average winter temperature in set A, include poverty rate and public housing availability in set B, and examine the relationship between homelessness and set B with set A held constant.

5.3.3 Inference Concerning Sets of Variables

In section 5.3.1 we saw that the contribution of a set of variables to a model's ability to explain variance in Y can be estimated with the partial or semipartial multiple correlation. It is often of interest to test whether

this contribution is statistically different from zero. That is, one might ask whether the set of regressors B contributes to the model's fit after controlling for the regressors defining set A. This boils down to a question about whether ${}_TSR(B.A)$ or ${}_TPR(B.A)$ is equal to zero or, equivalently, whether ${}_TR(AB) = {}_TR(A)$. Reframed in terms of change in R^2, we want to test whether the increase in R^2 for the model that results when set B regressors are added to a model that already contains set A regressors is statistically different from zero. We don't need separate tests for ${}_TSR$ or ${}_TPR$ because if one is zero, so is the other, and if a test described below leads to rejection that one is zero, it leads to a corresponding claim that the other is also not zero.

Such a hypothesis is frequently tested using the method of *hierarchical entry* of regressors into a model, which is what we did to estimate the squared partial and semipartial multiple correlations earlier, but by following up the computations with some inferential statistics. For instance, we saw that as a set, exercise and food intake (set B) accounts for about 45.7% of the variance in weight loss, holding constant metabolism and sex (set A). That is, $SR(B.A)^2 = 0.457$. Is this proportion—which is equivalent to the change in R^2 when set B variables are added to a model containing set A—statistically different from zero?

To answer this question, $SR(B.A)$ is converted to a statistic with a known sampling distribution under the null hypothesis that ${}_TSR(B.A) = 0$. That statistic is an F-ratio, calculated as

$$F = \frac{R(AB)^2 - R(A)^2}{1 - R(AB)^2} \times \frac{df_{residual}}{m_B} \tag{5.2}$$

where m_B is the number of regressors in set B and $df_{residual}$ is the residual degrees of freedom for the model including set A and set B regressors. Observe that the numerator of the first term above is just $SR(B.A)^2$. Equation 5.2 can be expressed equivalently in terms of $PR(B.A)$ as

$$F = \frac{PR(B.A)^2}{1 - PR(B.A)^2} \times \frac{df_{residual}}{m_B} \tag{5.3}$$

Under a true null hypothesis, this F-ratio is distributed as $F(m_B, df_{residual})$. A p-value can be found by computer or using a table of critical values of F for a desired level of significance (see Appendix C).

In this example, for instance, we have $R(AB)^2 = 0.931$, $R(A)^2 = 0.474$, $PR(B.A)^2 = 0.869$, $m_B = 2$, and $df_{residual} = 5$. Equation 5.2 gives

$$F = \frac{0.931 - 0.474}{1 - 0.931} \times \frac{5}{2} = 16.558$$

and equation 5.3 gives

$$F = \frac{0.869}{1 - 0.869} \times \frac{5}{2} = 16.583$$

These are in fact equivalent formulas, but they seem to generate slightly different results here due to rounding error introduced by doing the computations to only three decimal places. The critical $F(2, 5)$ for rejection of the null hypothesis at a 0.05 level of significance is 5.786, so the p-value is less than .05. We reject the null hypothesis at the $\alpha = .05$ level and claim that food intake and exercise frequency explain some of the variation in weight loss after accounting for the effects of sex and metabolism.

The F-statistic was described as an F-ratio above because it is a ratio of mean squares, just as is the F-ratio for the test that ${}_TR = 0$ described in section 4.3. To see how, remember that $SS_{regression}$ is sensitive to how much of the variance in Y is explained by the model. When the set A regressors are in the model, $SS_{regression} = SS(A)_{regression} = 46.734$, but when the regressors in set B are added, $SS_{regression}$ increases to $SS(AB)_{regression} = 91.703$. This increase in the regression sum of squares when set B variables are added to a model with set A variables already in it we denote $SS(B.A)_{regression}$. In this case, $SS(B.A)_{regression} = SS(AB)_{regression} - SS(A)_{regression} = 91.703 - 46.734 = 44.969$. Expressed in terms of the *mean increase per regressor*, we get $MS(B.A)_{regression}$, which is $44.969/2 = 22.485$. More generally, $MS(B.A)_{regression} = SS(B.A)_{regression}/m_B$, where m_B is the number of regressors in set B.

Equations 5.2 and 5.3 are mathematically equivalent to

$$F = \frac{MS(B.A)_{regression}}{MS_{residual}}$$

where $MS_{residual}$ is the mean squared residual from the model containing set A and set B regressors. This is a ratio of mean squares. In this example,

$$F = \frac{22.485}{1.359} = 16.545$$

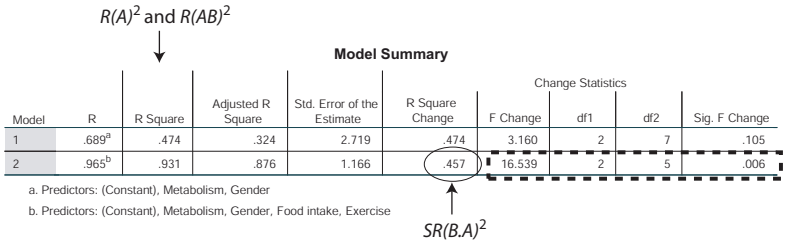


FIGURE 5.5. SPSS output from a test of the contribution of a set of predictors to a model.

which again differs slightly from the results using the earlier formulas due to rounding error in hand computation.

It is worth pointing out that the computation of $SS(B.A)_{regression}$ offers another means of calculating $SR(B.A)^2$, as

$$\begin{aligned} SR(B.A)^2 &= \frac{SS(B.A)_{regression}}{SS_{total}} \\ &= \frac{44.969}{98.5} \\ &= 0.457 \end{aligned}$$

This should make sense, as SS_{total} is variance in Y to be explained, and $SS(B.A)$ is variance explained in Y uniquely by set B regressors. So their ratio is the proportion of the variance in Y uniquely explained by the regressors in set B, which is one definition of a squared semipartial correlation.

These computations are not difficult to do, but doing so is a little bit tedious, and hand computation can introduce rounding errors in any formula applied manually. Fortunately, there is no need to do any of these computations manually, as most statistical packages have a means of implementing this test automatically and to a higher degree of precision. The code for SPSS, SAS, and STATA to conduct this test for this example can be found below, and Figure 5.5 provides an excerpt of the relevant part of the output from SPSS. All three programs yield $F(2, 5) = 16.54, p = .006$, which is very close to our hand computations but more accurate.

```
regression/statistics defaults change/dep=wtloss/method=enter metab
    sex/method=enter exercise food.
```

```
proc reg data=exercise;model wtloss=metab sex exercise food;
test exercise=0,food=0;run;
```

```
regress wtloss metab sex exercise food
test exercise food
```

This procedure works even when there is only one variable in set B. But when interest is in whether one regressor improves the model to a statistically significant degree, it is easier to simply look at the t- and p-values for the regression coefficient for that regressor. The t-value for that regressor would be equal to the square root of the F-statistic calculated using the procedure above, and the F- and t-statistics will have the same p-value. These are mathematically equivalent tests when $m_B = 1$.

As discussed in section 4.4.4, collinearity between two regressors occurs when they correlate highly, and this correlation diminishes the unique contribution of each to accounting for variation in Y. In larger problems, collinearity may pervade an entire set of regressors. Three measures of political liberalism might correlate so highly with each other that when they are used together in a regression, any one might contribute nearly as much to the regression as the entire set of three. They then form a *collinear set*—a set whose members all correlate highly and thereby lower each other's unique contribution. Occasionally, collinearity may pervade all the regressors in an analysis, but the usual situation is for it to pervade merely one set of regressors, not affecting others in the same analysis.

Whereas collinearity between variables that define a set increases the standard errors of the regressors in that set, this is not a problem when testing whether the set improves the fit of the model or its ability to explain variance in Y. In fact, it is not unusual to find a situation in which none of the regression coefficients in the set is statistically different from zero, yet the set as a whole significantly improves the fit of the model. The proper conclusion in such cases is that one or more members of the set is related to Y even though we cannot say specifically which one. This is a vague conclusion, but such conclusions are sometimes the best you can do with collinear sets.

5.4 A Glance at the Big Picture

With these principles of linear regression analysis outlined, now is a good time to step back and take a look at how linear regression analysis can be

further extended to deal with more complex analytical questions and also introduce some problems you might encounter.

5.4.1 Further Extensions of Regression

This chapter has described several extensions of the basic type of linear model introduced in Chapter 3. More are still to come. Chapter 6 examines the use of linear models in conjunction with random assignment. In Chapters 9 and 10 we shall see that a *multicategorical regressor* can be treated as a set of dichotomous variables. For instance, the multicategorical variable of religion (Protestant, Catholic, Jewish, or Other) can be thought of as a set of yes–no questions, as in "Are you Protestant?," "Are you Catholic?," "Are you Jewish?" A variable treated in this way is sometimes called a *factor* in ANOVA, but as we will see, factors in the context of regression analysis are just a set of dichotomous regressors.

In Chapter 12 we will see ways of fitting *curves and curved surfaces* rather than straight lines and planes to a data set. A great many curvilinear relationships can be fitted by the techniques described in Chapter 12. Still more can be fitted by more complex extensions of regression analysis not covered in this book.

Extensions of the ANOVA concept of *interactions* (differences between differences) are treated in Chapters 13 and 14. Interaction allows you to fit models in which one regressor's relationship with Y depends on another regressor in the model. For example, the relationship or partial relationship between wealth and political conservativism might be higher among men than among women, or it might decrease with increasing years of education. And Chapter 15 introduces the concept of *indirect effect* in which one variable affects another through a third. Additional chapters deal with various issues and controversies in regression, such as testing assumptions (Chapter 16), assessing the importance of regressors (Chapter 8), and statistical power (Chapter 17)

5.4.2 Some Difficulties and Limitations

Some significant difficulties that can arise in regression analysis, as well as various limitations, are handled with varying degrees of ease and controversy. We list them here to assure you early on that statisticians and others who study linear models for a living have thought rather deeply about these problems. You may find it convenient to use this section as a checklist when planning an analysis. Of course, no list can cover all conceivable

problems, and this list is not meant as a substitute for careful thought and a little common sense.

1. *Undercontrol* is the basic problem we have considered from the very beginning of this book—a failure to control all relevant variables that make it difficult to interpret the correlation betweeen X and Y.

2. *Overcontrol* means destroying an otherwise valid design, analysis, or interpretation by including as covariates any variable(s) affected by the dependent variable Y; see section 17.3.4.

3. *Singularity* is the inability to compute regression coefficients because at least one regressor is perfectly predictable from other regressors; see section 17.3.3.

4. *Nonlinearity* occurs when a curved line or surface fits the data better than a straight line or plane; see Chapter 12.

5. *Interaction* or *moderation* occurs when one regressor's relationship with dependent variable Y depends on the value of another regressor in the model; see Chapters 13 and 14.

6. *Heteroscedasticity* means that conditional distributions of dependent variable Y do not have the same standard deviation. This can destroy the validity of ordinary methods of statistical inference in regression or reduce their power; see Chapter 16, especially section 16.2.3.

7. *Non-normality of errors in estimation* means that the conditional distributions of $Y - \hat{Y}$ are not normal; see Chapter 16.

8. *Outliers* are cases with values on the dependent variable Y that are extreme given their pattern of measurements on the regressors; see Chapter 16.

9. *Leverage points* are cases with very unusual patterns of scores on regressors. They do not destroy the validity of regression by themselves, but they make it more difficult to detect outliers; see Chapter 16.

10. *Influential outliers* are cases that heavily affect one or more partial regression coefficients. Most often these are both leverage points and ordinary outliers; see Chapter 16.

11. *Noninterval scaling* arises when a 1-unit difference between two cases on a variable has different meanings depending on where on the scale the comparison between the cases is made; see section 17.3.5.

12. *Missing data* occur when a case has known scores on some variables and unknown scores on others because the variable was not measured for that case. An example would be when someone fails to answer a question on a survey when responses to that question serve as one of the regressors in the model; see section 17.3.6.

13. *Measurement error* is perhaps the most common and important problem of all. It occurs when variables are simply not measured accurately; see section 17.2.

5.5 Chapter Summary

This chapter concludes our introduction to the fundamentals of the mechanics of linear regression analysis. We showed that the principles of estimation and interpretation described in the first three chapters generalize to linear models that include a dichotomous regressor. The regression coefficient for a dichotomous regressor corresponds to a mean difference in Y when the two groups are represented by codes on the regressor that differ by 1 unit. Although mean differences are a common metric of effect in research, we discourage researchers from artificially dichotomizing numerical regressors, as doing so is not necessary and can damage an analysis.

Whenever two variables X and Y are less than perfectly correlated, regression to the mean will occur, meaning that cases relatively high or low on X will tend to be less extreme on Y than they are on X. Linear regression analysis properly handles this phenomenon by generating estimates of Y that are correspondingly less extreme than X. The use of difference scores in linear models does not properly account for regression to the mean.

The partial and semipartial correlations introduced in Chapter 3 generalize to sets of variables. These setwise measures of partial association quantify the relationship between a set of regressors and a dependent variable while holding all other regressors in the model constant.

In the next section of this book, we further advance your understanding of regression by considering topics such as multicategorical regressors, nonlinear relationships, and models that allow one regressor's effect on Y to depend on another in the model. But before diving into these areas,

given our treatment of regression analysis thus far, it is worth now revisiting some of the issues brought up in Chapter 1 that justify the value of statistical control relative to other forms of control, as well as its limitations. That is the topic of Chapter 6.

6
Statistical versus Experimental Control

The prior chapters have described the fundamentals of linear modeling using linear regression analysis and some applications of statistical control. This chapter considers the advantages and disadvantages of statistical control relative to experimental control—its principal competitor—especially random assignment. Although experimentation through random assignment is often placed on an empirical pedestal, it too has some disadvantages unique to it, as well as some shared with statistical control. After a brief philosophical treatment of an extreme philosophy of causation we call *manipulationism*, this chapter ends with a discussion of how random assignment and statistical control can be used in tandem to strengthen each other.

As the preceding few chapters have illustrated, statistical control and various measures of partial association are useful means of establishing whether X and Y are associated in spite of their covariation with other variables that cloud the interpretation of their simple association. But they are not a panacea to all problems one might face, and they are limited relative to random assignment. That said, random assignment has its limitations too, which can be overcome in part through statistical control. Given the limitations and strengths of both random assignment and statistical control, they can complement each other when used together. This chapter addresses some of these limitations, strengths, and advantages of using both random assignment and statistical control.

This is primarily a chapter of lists. We begin with some limitations of statistical control that are overcome by random assignment. Then we describe things often confused with random assignment, as well as some nonstandard forms of random assignment that are sometimes not correctly recognized as random assignment. We follow this discussion with several limitations common to statistical control and random assignment, and then a few problems often produced by or associated with random assignment in

particular. A philosophical discussion of the concept of causation follows, and we end with a set of reasons for combining statistical control with random assignment when both are practical.

6.1 Why Random Assignment?

6.1.1 Limitations of Statistical Control

Statistical control has some major limitations. Perhaps the biggest of these limitations is that the list of potential covariates is often endless. Consider any study comparing two groups of people on some dependent variable. If we control statistically for age, race, sex, and income, then a critic may ask why we have not controlled for education, IQ, and political ideology. No matter how many variables are controlled statistically, we always know that the two groups are still likely differ in some nonrandom way we do not fully understand, simply because the cases in one group are in that group and others are not. In the experimental design literature, this is a threat to internal validity called *selection*.

Although it might seem that one could just indiscriminately control for everything one can conceive, it is possible to go overboard; hence, the second limitation. A phenomenon called *overcontrol*, described in section 17.3.4, means that a single covariate added improperly can invalidate an otherwise sound design and analysis.

A third problem is that we often approach an analysis under the assumption that X causes Y. Our analysis is undertaken to establish whether the association exists after controlling for various other variables that represent alternative explanations for the association. But even if we could control all the right variables and no others, statistical control still does not tell us whether X affects Y or whether Y affects X. An association that survives the test of statistical control still provides no information about causal order.

Fourth, statistical control requires not only that we measure covariates but also that we measure them accurately. Covariates such as SES, IQ, and various measures of personality or attitude, for example, are never measured with complete accuracy. Even if we are measuring validly (i.e., measuring what we claim we are measuring), we probably are not measuring with perfect reliability (i.e., with no random measurement error). Thus, a covariate can never be fully controlled. We address this problem further in Chapter 17.

Finally, statistical control cannot distinguish causation from correlation produced by excluding certain cases from the population. For instance, if we found a correlation between age and wealth, we might at first say that age must have caused the wealth, because wealth cannot influence the date one was born. But if poor people die younger on the average, then there is a different sense in which wealth produces age. This is *correlation by selective exclusion*, better known in the experimental design literature as a threat to internal validity known as *attrition* or *mortality*.

These limitations imply some (but not all) assumptions necessary to infer causation from statistical control: that we have controlled all necessary variables, that we have controlled no variables that would distort the relationship, that Y does not affect X, that covariates are measured accurately, and that correlation is not produced by selective exclusion.

6.1.2 The Advantage of Random Assignment

Random assignment can provide an elegant way around these limitations. An important outcome of successful random assignment is that all covariates that are properties of participants, such as a person's age or education, are validly controlled without even being measured. If a study's validity depends upon the control of such covariates, then random assignment guarantees this validity more surely than does statistical control.

Random assignment can give valid control of covariates even for very small samples. *Validity* means that the effect of an independent variable will be statistically significant at, say, the 5% level only 5% of the time when the null hypothesis is true. Designs with random assignment can have this property even if the sample size is very small.

We might imagine an argument against random assignment that goes like this: Even with random assignment to treatment and control groups, there is a 5% chance that the two groups will differ on average age at the 5% level of significance. There is also a 5% chance that the two groups will differ on average educational level. Likewise for income, race, sex, IQ, and other covariates. When we consider that the number of potential covariates is infinite, the two groups certainly differ significantly on at least one. Then what have we gained by random assignment?

This argument is answered by recalling that while there may be many covariates, there typically are few, and often there is only one dependent variable. If assignment is random, then the probability is only .05 that the two groups will differ significantly (at the .05 level) *on the dependent variable* by chance alone.

Therefore, random assignment, even with no attention paid at all to covariates, can produce valid hypothesis tests more surely than the most exhaustive and careful statistical control. If this argument seems incomplete to you, there is good cause: Wait until we consider the limitations of random assignment in section 6.2.

6.1.3 The Meaning of Random Assignment

Random assignment is sometimes confused with other things, so that random assignment is sometimes claimed to be part of a design when in fact it is not. On the other hand, there are some forms of random assignment that may not be correctly recognized as random assignment. We begin first with things sometimes confused with random assignment.

Mere Manipulation. Random assignment requires us to manipulate participants (except in unusual cases such as lotteries); but manipulation, though a necessary condition, is not a sufficient condition for random assignment. We are manipulating participants even when we just require them to sit still and listen to our directions, but that does not mean we have randomly assigned them to anything.

Other Types of Randomization. As mentioned in section 1.1.2, random assignment must not be confused with other types of randomization. A researcher may randomize the order of presentation of stimuli, or randomize which form of a test each participant takes, but that does not constitute random assignment unless these are the variables or effects under study.

Forced Equality of Cell Frequencies. There are occasions when an investigator may choose to force the number of people in different conditions to be the same across all conditions. This actually makes random assignment more difficult, and special procedures must be followed. For instance, suppose an investigator is using a 2×2 design and flips a coin twice for each participant, using the first flip to place the participant in a row and a second flip to place him or her in a column of the design. Then assignment is random, but the four cells will probably end up with unequal cell frequencies.

One way to ensure an equal number of cases in each cell while retaining true random assignment is to assign each participant a number several digits long from a random number table. The experimenter can then rank the participants by the random numbers (breaking any ties randomly) and place several participants in cell 1, an equal number in cell 2, and so forth.

Random But Nonindependent Assignment. The phrase *random assignment* is really shorthand for *random and independent assignment*. If we choose

two schools for an experiment involving a new curriculum and flip a coin to see which school gets the curriculum, then, speaking literally, we have assigned all students in each school randomly to the experimental or control group. But the students will not be assigned independently to the two groups, so we do not have *random assignment* as that phrase is used here. Unlike true random assignment, this design does not allow us to assume that the two groups have been equated on covariates.

Random Sampling. *Sampling* refers to the process by which study participants are selected from a larger population for inclusion in a study or analysis, while *assignment* refers to the process by which the selected cases are allocated to positions on the independent variable(s). For instance, if there are 40 people in a class and 20 are randomly placed in one experimental condition and 20 are randomly placed in a control group, then some people might say we have randomly selected the groups. But in our terminology, this is an example of *random assignment without random sampling*. We have used the entire population (the class) in the analysis rather than selecting part of it. But there is random assignment, because after the sampling (or, in this case, nonsampling) is conducted, students are randomly assigned to conditions.

Despite the uniformity with which introductory statistics textbooks assume random sampling, it is well known among statisticians that valid statistical inferences may sometimes be drawn without random sampling—and in fact without either random sampling or random assignment. For instance, suppose a college hired 100 new professors in the 1990s and another 100 in the 2000s. If women made up 10 of the new professors in the 1990s and 30 in the 2000s, a statistical test could be conducted to determine whether this difference can be attributed to chance. But there is no hint of either random sampling or random assignment.

As mentioned earlier, there are forms of random assignment that are sometimes not recognized as such:

1. Random assignment is usually employed with categorical independent variables, but it may also be employed with numerical independent variables. For instance, participants may be randomly assigned to various hours of practice on a task.

2. If a process is truly random, it need not be under our control. If we study the differences between winners and losers in a fair lottery, we can assume that assignment has been random. But if we study the differences between lottery players and others, there has been no random assignment.

3. Although random assignment is usually easier in laboratory experiments than in field studies, it can be done in field studies and may be absent in laboratory studies. For instance, if we are studying the differences in responses to male and female pollsters and a random process determines whether any given respondent is contacted by a male or female pollster, then assignment is still random. On the other hand, if those who are available to participate in a laboratory study in the morning are placed in the experimental group, while those who can participate in the afternoon are placed in the control group, then assignment is not random.

4. If a study has two or more independent variables, assignment might be random on some and not on others. For instance, a 2×2 design might cross sex with an experimental versus a control group. Whereas individual men and women may be randomly assigned to experimental or control conditions, they certainly aren't randomly assigned to be male or female.

5. If a study has three or more conditions, assignment might be random among some of the conditions but not others. For instance, in a stressful experiment, all participants in poor health might be placed in a control condition while others are randomly assigned to the remaining groups. Then the advantages of random assignment apply to some comparisons but not to others.

6.2 Limitations of Random Assignment

6.2.1 Limitations Common to Statistical Control and Random Assignment

Despite the advantages of random assignment, there are several problems it fails to solve.

First, we never know for certain what *facet* of an independent variable has produced an observed effect. For instance, was the effect of a new school curriculum due to the curriculum itself or to the particular teachers who implemented it?

Second, even when we have both random assignment and random selection from some population, we will always have nonrandom selection from the population to which the results will be applied. If we establish an effect on human populations through an experiment run in April, then by October all the people in the population are 6 months older, some have died,

others have been born, and everyone has no doubt changed in one way or another. The attacks on Pearl Harbor in 1941 and the World Trade Center and Pentagon in September 2001 changed the way Americans perceive the world in 1 day, affecting many aspects of their lives. A beautifully randomized experiment on the most effective advertisement for a particular product in August 2001 might have little relevance now. Although this is perhaps an extreme example, the point is still valid. Populations change in many ways over time, and as they change, the generalizability of the findings of some study conducted in the past may be weaker.

But even if we cannot automatically generalize an experiment's results to the future, doesn't an experiment at least establish a causal conclusion for the precise population and moment of the experiment? This brings us to the third insoluble problem: We never know whether a conclusion applies to everyone in a population or only to some subpopulation. Thus, even if we were to show, via a randomized experiment on human participants, that smoking causes cancer, we should qualify the result with the phrase "in at least part of the population." For all we would know even after this experiment, there may be some unidentified subpopulation in which smoking *prevents* cancer. Another way to state this point is that we never know whether we have identified all important *moderator variables*—a variable that affects the relationship between the independent and dependent variable—so that, for instance, the relationship is positive for people high on the moderator and negative for people low on the moderator.

The fourth limitation is that we can never study all possible *side effects* of a particular experimental treatment or intervention. We may show conclusively that a motivational program lowers the number of teenagers who drop out of high school, but we merely assume rather than know for sure that it doesn't have some negative effects later in time. Or it may be that the effect observed is only short term and requires the presence of various features inherent in the program in order for those effects to appear. For instance, perhaps people exposed to such programs adapt so that they never do anything positive without special incentives that are rarely available in the real world.

The fifth limitation is that with numerical dependent variables, neither ordinary nor "distribution-free" statistical methods allow us to draw truly firm conclusions about the difference between two groups (e.g., experimental treatment vs. control) without making some assumptions about the shape of the population distribution of the thing measured. For instance, suppose the dependent variable is measured on a 100-point scale and the

population distribution in one group is positively skewed, with most of the measurements bunched up in middle of the scale, but with a long tail extending out to 100. Further suppose that the opposite is true for the other group, with most measurements in the middle of the scale and a long tail extending down to 0. Examples of this sort can be constructed in which the population mean of the experimental group is lower than the mean of the control population, but even "distribution-free" tests lead to the claim that the experimental group's mean is higher. For maximum relevance to policy and practical conclusions, we usually want our conclusions to concern means. But distribution-free tests either reach conclusions about medians or other statistics or, more commonly, yield "nonparametric" conclusions, which are vague because they don't describe parameters such as a mean or median. Thus, a conclusion about the difference between two means cannot actually be reached without some assumptions about the shapes of the distributions of the variables being compared, contrary to the pretense that random assignment allows us to avoid assumptions that are even slightly questionable.

Sixth, whether or not random assignment is used, we never find ultimate causation. Rather, we must always assume that causation operates indirectly through intermediate variables. For instance, the independent variable of a child's parents' attitudes about the importance of education could not magically and directly make the child's grades better. Even if it is the case that differences between parents' attitudes do in some way cause differences in how their children perform in school, this effect must work through some indirect mechanism, such as making the child study harder, or through resources the parents give, such as a quiet place to study, which in turn results in better performance. All causal effects operate through something. So establishing an effect through random assignment only can tell whether an effect exists, not how that effect operates.

Seventh, in the social and behavioral sciences, almost any study that takes more than an hour or two of a participant's time will suffer from some kind of nonrandom attrition. Participants may fail to return for the final session in which the dependent variable is measured, for example. Or some may misunderstand some directions they were given during the experiment, so no measurement of the dependent variable that is meaningful can be obtained. As a result, the equating of groups through random assignment on all covariates may be destroyed if the experimenter is forced to nonrandomly delete participants from one or both groups. Even if such attrition occurs with the same frequency in each group, this offers no as-

surance that those removed are comparable. Participants may drop out of one group because they find it too time-consuming, and participants in the other may drop out because it is boring and they don't see the point of the study. Any claim that the two groups of dropouts are equivalent must be based on some kind of analysis of the association between dropout status and other variables, but the whole point of random assignment is to avoid the limitations inherent in such mathematical methods.

In summary, random assignment does in principle allow us to make firmer statements than statistical control about the effect of an independent variable on a dependent variable. But even if there is no attrition, we are still unsure about the operative facets of the independent variable, the population to which the results apply, moderators, side effects, and ultimate causation.

6.2.2 Limitations Specific to Random Assignment

The title of this section is not meant to imply that the problems discussed here are unique to random assignment. But these problems differ from those in section 6.2.1 in that these are more likely to be serious under random assignment.

The most obvious problem with random assignment is that it is often illegal, immoral, impractical, or impossible. It is simply impossible to manipulate a person's age, race, or genetic sex. Or if we wanted to study the effect of college attendance on income at age 40, random manipulation of college attendance would require us to determine randomly who attends college and who does not. This would be impractical, immoral, and perhaps illegal. Often, the more important the independent and dependent variables in a study, the less practical it is to use random assignment. But as mentioned in section 1.1.2, random assignment may be impractical even in laboratory experiments with animals. You may have no control over the length of time a mouse looks at a stimulus, but you may be able to measure that time accurately.

A second problem is that participants in randomized experiments often know they are being manipulated. This may change their responses. If they resent the manipulation or their placement in a control group, and if that resentment makes them more likely to drop out of the study, then random assignment can actually make groups *less* comparable after nonrandom attrition has occurred.

The third major problem associated with random assignment is not so much a problem of using it yourself as it is of evaluating research in which

others have used it. The words *random assignment* have such blinding effects on journal editors, reviewers, and granting agencies that there is enormous temptation for a researcher to use this phrase when it is not completely accurate. A research assistant in an educational experiment may feel a special sympathy for a particular child and stretch the rules by making sure that the child ends up in a particular group in the study in which students are expected to excel rather than a different condition. When the chief scientist learns of this lapse in proper procedure later, he or she may be reluctant to drop the magic words *randomly assigned* from the research report. When researchers or their assistants are questioned in detail about every single participant in a study, the words *random assignment* may turn out to cover up a variety of methodological sins.

6.2.3 Correlation and Causation

The limitations of statistical control described in section 6.1.1 are the basis for the familiar saying, "Correlation does not imply causation." In this expression, the word *correlation* refers not specifically to the correlation coefficient but, instead, to any type of statistical analysis based entirely on observation rather than experimental manipulation. Another version of this saying was offered by Holland (1986, p. 959): "No causation without manipulation." These two expressions are intended to apply not just to the words *cause* and *effect* but to the very concept of causation, thereby affecting our usage of dozens or hundreds of words implying causation, such as *produce*, *increase*, *harm*, *prevent*, and so forth. We use the term *manipulationism* to denote the viewpoint expressed by these sayings.

The manipulationist position helps us remember the limitations of statistical control described in section 6.1.1, but it should not be overstated. Before discussing manipulationism itself, let's briefly review the very basis of knowledge. How do you know this book really exists? Perhaps you and your friends or your instructor merely *think* that you see it and feel it. The fact is that you cannot prove this book exists. But there is no doubt that the simplest and most plausible hypothesis to explain your observations is that the book does in fact exist. Essentially all knowledge about the external world is of this type. It is not something we know beyond all doubt, but is rather the simplest and most plausible way to explain our observations. This includes virtually all knowledge that it never even occurs to us to doubt, such as the existence of the objects we see around us everyday. With this in mind, let's examine several successively more modest versions of manipulationism. We agree only with the last of them.

An extreme form of manipulationism (denied even by most manipulationists) holds that real causal knowledge always requires random assignment. That extreme position would deny most of the knowledge that we routinely accept about the physical sciences, which rarely engage in experimentation through random assignment. It seems more reasonable to accept a hypothesis as supported by a single event predicted by researchers but never before observed. For instance, when the first atomic bomb was successfully exploded, few doubted anything about the causal process at work producing the explosion, even though there had been no random assignment to conditions. The probability is near zero that it would have occurred just at the exact time and location that it actually did by chance alone when it had never occurred anywhere else in recorded history. It is safe to assume that the thought and work of scientists who produced the atomic bomb are responsible for the explosion at that moment and location rather than just happenstance.

Can we say then that causal knowledge requires random assignment, with the exception of a small number of events such as this or ones like it? That is the standard manipulationist position. But it is hard to imagine anyone taking this seriously. Most of us go to work to keep our jobs or professional status, study for important exams we take, drive carefully to avoid accidents, dress before going outside to avoid embarrassment or arrest, and lock our doors at night or when we leave the house so as to avoid being a victim of a burglary. Yet, most likely, neither you nor anyone else has ever experimentally examined whether people are more likely to be burglarized when doors are left unlocked, whether arrest is more likely when wandering the streets naked than when not, or whether people who show up to work regularly are more likely to keep their jobs. None of the causal associations implied in the beliefs that influence this behavior have ever been put to experimental test in the manner that such a perspective requires. By this principle, a manipulationist professor would be comfortable with the rationale that you decided not to study for his or her exam because there have been no experiments showing that studying for his or her exams is effective.

Is manipulationism then tenable for "scientific knowledge"? Can we say that causal relations are not proven "scientifically" unless they are demonstrated by experimental manipulation and random assignment? If so, this principle would reduce fields like biology, geology, meterology, astronomy, and even astrophysics to pseudosciences or quasi-sciences, because these fields are based almost entirely on observation rather than on

manipulation of the phenomena under study. A meterologist does not generally engage in experimentation to examine the effects of ocean currents on weather but rather observes data from available sources and attempts to examine how variations in current speeds, temperature, and depth is related to variations in certain weather events. An investigator building a weather satellite or slicing tissues for microscopic examination may be manipulating the immediate objects at hand, but in a broader sense the process is one of observation rather than experimentation in the manipulationist sense.

At its extreme, this position would mean that scientists frequently conduct their business in nonscientific ways. For example, a chemist might avoid using certain beakers that his or her observation suggests are more likely to break at higher temperatures. But absent planned experimental evidence, such experience garnered through nonexperimental methods should not influence the behavior of the scientist, yet no doubt scientists make decisions about how to run their laboratories through such informal experiences all the time. And at its most insidious, such a principle leads people to doubt conclusions that are overwhelmingly consistent with data collected through nonexperimental methods, such as how humans are modifying the climate of the planet.

Might manipulationism then be further narrowed to the social and biological sciences, which are the areas in which scientists really debate these questions? But scientists have never established experimentally on human participants the life-saving effects of seat belt use or the toxic effects of leaded gasoline exhaust. But seat belts are required in most developed countries and many less developed ones, and leaded gasoline has been banned. Few would doubt the wisdom of these policies even though this wisdom has never been put to experimental test. In the legislative battles for these reforms against fierce industry opposition, we can certainly be thankful for the efforts of the advocates of reform who weren't afraid to utter simple sentences such as "All the evidence suggests that seat belts prevent unnecessary deaths." This is that point at which we should recall the nature of knowledge as described earlier.

Could we then just say that in the social and biological sciences, causal relationships verified by random assignment experiments are established more firmly than other causal relationships? But is that really true? Sections 6.2.1 and 6.2.2 mentioned some rather fundamental limitations that may apply even to research with random assignment. The concentration of fluoride in drinking water varies widely around the world, with towns just

a few miles apart sometimes differing drastically in fluoride concentration. Detailed statistical analyses of these differences can show the preventive effect of fluoride on tooth cavities. It would be at least as accurate to summarize such studies with the simple causal sentence, "Research shows that fluoride prevents cavities in children," as it would be to summarize a typical randomized study on the effects of two algebra curricula with the sentence, "Curriculum A has been shown to be more effective than curriculum B."

Finally we might just say that random assignment is a very important feature of a study, and if random assignment was feasible even if not carried out, then even a very brief report of a study in the newspaper or a university press release should mention whether or not it was used. This is a version of manipulationism we can agree with. Given that some reports of studies are just a sentence long, we need a single phrase that conveys this point. Of course "random assignment" is not available. A phrase like "experimentally verified causal relationship" seems reasonable. Anyone is free to invent a shorter phrase. But manipulationists shouldn't try to monopolize the entire concept of causation with their narrow conceptualization of what evidence is required unless they are genuinely uncertain that careless driving causes accidents or are willing to streak past a police station naked.

6.3 Supplementing Random Assignment with Statistical Control

We have been speaking as if statistical control and random assignment are simply competitors. But there are several reasons for supplementing random assignment with linear models whenever possible. The first is really a set of reasons that requires no additional discussion—the existence of nonrandom attrition and all the real-world limitations of random assignment that we described in section 6.2. This section is dedicated to three reasons not yet discussed.

6.3.1 Increased Precision and Power

One reason for supplementing random assignment with linear models follows more directly than the prior reason from principles of statistics and logic. This reason is a potentially large gain in the precision with which an independent variable's effect is estimated and the power of hypothesis

TABLE 6.1. Data From a Hypothetical Study Examining the Effect of a New Therapy on Symptoms of Posttraumatic Stress

ID	Posttest Y	Pretest X_1	Therapy X_2	Gain $Y - X_1$
1	2	1	1	1
2	4	3	1	1
3	6	7	1	−1
4	6	10	1	−4
5	9	13	1	−4
6	10	17	1	−7
7	12	19	1	−7
8	6	1	0	5
9	7	5	0	2
10	9	7	0	2
11	9	9	0	0
12	12	13	0	−1
13	12	16	0	−4
14	15	19	0	−4

tests on that effect. This advantage increases with the correlations between covariates and the dependent variable within levels of the independent variable.

As an illustration, consider the data in Table 6.1 from a hypothetical study of the effect of a proposed therapy for posttraumatic stress disorder (PTSD) conducted on 14 military veterans who experienced combat. Half of the veterans were randomly assigned to experience 6 weeks of the proposed therapy ($X_2 = 1$) and the other half received 6 weeks of a traditional therapy ($X_2 = 0$). Each was pretested prior to therapy with respect to their PTSD symptoms (X_1), and an identical posttest assessment of symptoms was administered upon the completion of therapy (Y). These data are represented graphically in Figure 6.1, with the seven veterans in the experimental therapy group denoted with solid squares and the seven in the traditional therapy group denoted with hollow circles. In these data, the mean pretest scores are exactly equal in the two groups, although the major points below are valid without this condition.

The diagonal lines in the figure represent the regression lines estimating posttest from pretest in each of the two groups. The slopes of the two regression lines are almost identical—even the most discerning eye could not detect any difference in them. Furthermore, a formal test of the difference

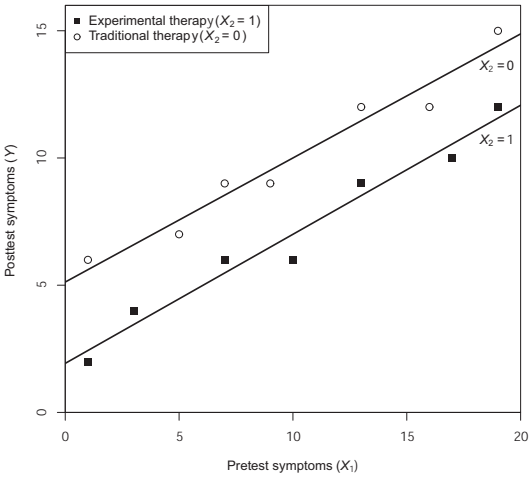


FIGURE 6.1. A scatterplot with regression lines estimating posttherapy PTSD symptoms from pretherapy symptoms.

between the slopes of the two lines that we introduce in Chapter 13 reveals no statistically significant evidence of difference between the slopes of these lines. Therefore, the vertical distance between the two regression lines can be construed as a rough estimate of the effect of the experimental therapy, which in this case appears to be about 3 points on the posttest PTSD assessment relative to the traditional therapy. Your intuition tells you after even a cursory look at Figure 6.1 that the differences between treatment and control groups at the end of therapy probably can't be attributed to chance. The regression line for the experimental therapy group is considerably lower than the line for the traditional therapy group. Furthermore, every one of the experimental therapy cases is closer to the regression line for the experimental therapy group than to the traditional therapy line, while every one of the traditional therapy cases is closer to the traditional therapy line.

We can formally estimate the size of the effect of the proposed therapy—the distance between the two regression lines—using a linear model estimating posttest PTSD scores from pretest PTSD scores and the variable coding which therapy the veteran received. Doing so yields

$$\hat{Y} = 5.019 + 0.498X_1 - 3.000X_2$$

Consistent with the figure, the regression coefficient for X_2 is -3.000. That is, holding constant pretest PTSD symptoms, veterans who received the experimental therapy are estimated to be about 3 points lower in PTSD symptoms than those who experienced traditional therapy. Using the formulas and inferential procedures introduced in Chapter 4, this effect is statistically significant at any sensible α-level, $t(11) = -8.326, p = .00004$. The standard error of the regression coefficient for X_2 measures how precisely this effect is estimated. In this case, the standard error for the treatment effect is 0.360. All other things being equal, a smaller standard error translates into a more precise estimate. In this case, the treatment effect is estimated quite precisely.

Given that random assignment would be expected to equate the two groups in their PTSD symptoms before therapy (and in this case it exactly did), a much simpler analysis taught to any student of introductory statistics or research methods might have occurred to you. Why not just conduct a two-sample t-test to test the significance of the differences in means between the two groups in their symptoms at the conclusion of therapy? Doing so effectively ignores the information about each case's horizontal placement in Figure 6.1 (i.e., pretest PTSD) and uses only its vertical placement. If this information were all we had, our noncomputational intuitive testing would lead us to be considerably less certain about the size or even the existence of the treatment effect since the experimental therapy posttest PTSD scores range from 6 to 15, and the traditional therapy group's scores overlap them substantially, ranging from 2 to 12. An independent groups t-test confirms intuition. The difference between the means is 3.000, just as was found with the linear regression analysis, with the experimental therapy group lower on average in posttest PTSD symptoms, but this difference is not statistically different from zero, $t(12) = -1.680, p = .119$. This difference in significance is attributable entirely to the much larger standard error for the estimated effect. In the independent groups t-test, the standard error of the estimated mean difference is 1.786. In other words, the effect is estimated with much less precision than when the analysis included pretest PTSD symptoms as a covariate.

Yet another alternative might have occurred to you. Why not instead perform a t-test on change in PTSD symptoms between pretest and posttest? In other words, use the difference score $Y - X_1$ as the dependent variable in a run-of-the-mill t-test? It is hard to intuit the result of this test merely by inspecting Figure 6.1. But if you do the computations, you will find that the average change in symptoms for those assigned to the traditional

therapy was 0 (i.e., no change on average), whereas the the average change for those given the experimental therapy was −3 (i.e., an improvement in symptoms). But the difference between these differences is not statistically different from zero, $t(12) = 1.667, p = .121$. So the effect is the same as in the prior analysis—a 3-point difference in symptom change pre- to post-therapy, but this effect is estimated rather imprecisely, as the standard error of the difference between these differences is 1.799. As noted in section 5.2.5, this is equivalent to testing for an interaction between time and therapy in a 2 (time: PTSD symptoms pre and posttherapy) × 2 (experimental vs. traditional therapy) mixed factorial analysis of variance, with time as a within-subjects factor and therapy type as a between-subjects factor.

These two alternatives are really special forms of a two-regressor linear model in which the regression coefficient for pretest PTSD symptoms is constrained to either zero or one. The covariate adjustment approach involves the estimation of b_0, b_1, and b_2 so as to minimize the sum of the squared residuals in a model of the form

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

When doing so, the regression coefficient for pretest PTSD symptoms (b_1) is 0.498, $SS_{residual} = 4.998$, and $R = 0.985$. The first alternative, which ignores pretest PTSD symptoms entirely, is equivalent to estimating this same model but constraining b_1 to zero. Not surprisingly, that model does not fit nearly as well as the model that derives the optimal value of b_1 by the least squares criterion, $SS_{residual} = 134.000$, $R = 0.436$. The analysis using the difference between posttest PTSD symptoms and pretest symptoms can be written in the form of a linear model as

$$Y - X_1 = b_0 + b_2X_2 + e$$

which is equivalent to

$$Y = b_0 + 1X_1 + b_2X_2 + e$$

In other words, the difference score approach is like the covariance adjustment approach but fixing b_1 to 1 rather than letting the mathematics derive the optimal weight for pretest PTSD symptoms. This model, not surprisingly, does not fit as well as the model that optimizes fit by letting the data inform the estimate of b_1, $SS_{residual} = 136.000$, $R = 0.434$.

The main point of this example is that random assignment combined with covariate adjustment can result in more precise estimates of the effect

of independent variables of interest. This follows directly from the discussion of the factors that influence the standard error of a regression coefficient in section 4.4.3. Including covariates that are related to Y but unrelated to manipulated independent variables through the random assignment procedure lowers $MS_{residual}$ and thereby lowers standard errors for manipulated variables (and possibly nonmanipulated regressors as well). A corollary point is that the use of a difference score is rarely a good substitute for an $SS_{residual}$-optimized linear model. For the reasons described in our discussion of regression to the mean (see section 5.2), it is rarely accurate to assume that when a posttest is predicted from a pretest, its regression slope will be 1.0. And ignoring pretest PTSD symptoms would be legitimate if it were unrelated to posttest symptoms, holding therapy method constant. But in this example, clearly, such an assumption is unwarranted, as you'd expect veterans relatively higher in the distribution of PTSD symptoms before treatment to still be relatively higher following therapy, regardless of the form that therapy takes. Figure 6.1 confirms this expectation. In either case, the result is an estimate of the effect of therapy that is much less precise and a hypothesis test that is lower in power compared to when pretest is used as a covariate.

This does not mean, however, that one always gains power by indiscriminately adding covariates to a model with random assignment. The more strongly a covariate affects the dependent variable Y, the more power is gained from controlling it. But if a covariate has absolutely no effect on Y, one actually loses a little power by adding it to the model. The power lost is the same as what is lost by randomly discarding one case from the sample, so the loss is usually small unless the sample size is tiny. But even this small loss suggests that one should not indiscriminately add dozens of extra covariates to the model just because they happen to be in the data set.

6.3.2 Invulnerability to Chance Differences between Groups

An additional reason for supplementing random assignment with statistical control is that statistical control increases a conclusion's invulnerability to criticism based on chance differences between groups on other variables correlated with the dependent variable. For an extreme example, consider a design like the one in the prior section with pretest and posttest measures of PTSD symptoms from military veterans randomly assigned to experience either a new therapy for PTSD or a traditional therapy already in wide use. Suppose we find a statistically significant posttest mean difference between the two groups in PTSD symptoms using a test whose validity is

not challenged. But we happen to notice that every person's PTSD score on the posttest, regardless of which therapy he or she experienced, exactly equals his or her score on the pretest prior to therapy, so that the difference between groups is just as significant on pretest as on posttest! Although random assignment makes such a scenario unlikely, it doesn't make it impossible. Given such a "failure" of random assignment to equate the groups on pretest PTSD symptoms, the significant difference on posttest no longer gives us any confidence at all that the experimental therapy works any better or worse than the traditional therapy. Rather, the hypothesis most consistent with the data is that the new therapy has no effect at all, that nobody's PTSD symptoms change from pretest to posttest, and that just by chance, the groups differed on the pretest.

Once we agree that in this extreme example there is some doubt about the treatment's effectiveness, we must ask how extreme an example must be to raise similar doubts. Perhaps we should be concerned about all significant differences between groups on covariates, despite the argument of section 6.1.2 against this position. But we can avoid the whole problem by using linear models along with random assignment. The problem arises because we presume that the covariates correlate with the dependent variable in the population, so that if by chance we draw a sample in which the covariates correlate with the experimental manipulation as well, then we must presume the sample correlation between the manipulation and the dependent variable is at least partly spurious. But as described in Chapter 3, linear models examine the relation of the dependent variable to an artificially constructed independent variable that is uncorrelated with all other regressors in the model. This constructed variable is exactly uncorrelated with all covariates in the sample studied, not merely in some hypothetical population. This eliminates the problem. Even in this extreme example, regression would estimate the treatment effect to be zero, which is the estimate supported by intuition.

6.3.3 Quantifying and Assessing Indirect Effects

Even if random assignment establishes that X affects Y, it does nothing to establish how the effect works. But when random assignment to values of X is combined with linear modeling, it is possible to quantify how that effect operates if one has measured one or more mechanism or mediator variables M presumed to be causally located between X and Y. For instance, consider an experiment in which half of the participants are told a mental test indicates they should be especially good at solving problems of a certain

type, whereas the other half are given no such information. Suppose that those told they are good at solving such problems are found to persist longer in trying to solve such problems, which are in fact impossible, than those not told they are good at solving them. Was their persistence produced by an increase in self-confidence (M_1) or by a desire to please the experimenter who had just complimented him or her (M_2)? A *mediation analysis* using linear modeling is a common approach to answering this question. Such an analysis requires, among other things, estimating Y from both X and the proposed mediator or mediators M. Doing so is part of the procedure required for the computation of the direct and indirect effects of X on Y. We save a discussion of mediation analysis for Chapter 15.

6.4 Chapter Summary

Experimental and statistical control are both powerful weapons in the empiricist's arsenal. Random assignment has as its primary advantage its ability to equate groups on all conceivable covariates, measured or unmeasured, and this strengthens causal conclusions. Even so, random assignment has some limitations. Statistical control can be an effective fallback when random assignment is not feasible or practical, but one must not use it indiscriminately. Although it may be harder for the researcher to muster faith in a causal conclusion based only on statistical control, we disagree with the manipulationist position that random assignment is required in order to make causal claims. When possible, we recommend designs that include random assignment, but analyzed using statistical control as well, for doing so can yield more precise estimates and more powerful tests.

7
Regression for Prediction

Regression is very frequently used in causal analysis with the goal of examining if there is a relationship between a regressor *X* and dependent variable *Y* when all other regressors are held constant. Another use of regression is the production of an equation to generate a prediction of *Y* for cases not used to generate the equation in the first place. With this use of regression, emphasis is on building a regression model that is accurate in the sense of predicting *Y* well for those cases. This chapter addresses issues in the use of regression for prediction, such as how to select a model from among a large set of possible models, and how to estimate how well a model fits data when applied to a population. Also addressed is how a variable's contribution to a prediction of *Y* is related to its correlation with *Y* and other variables.

7.1 Mechanical Prediction and Regression

7.1.1 The Advantages of Mechanical Prediction

The need to predict one variable from others arises in many contexts. College admissions officers may want to predict the future GPA of a student in college if he or she were admitted. A firm hiring new employees may want to predict employee performance if hired. An insurance company may want to predict the remaining lifespan of someone applying for a life insurance policy, or whether a person is likely to be in a car accident if given an auto insurance policy. A parole board may want to predict whether a prison inmate will commit more crimes if released. The U.S. Secret Service may want to predict whether someone who has threatened a government official is likely to take action and attempt to carry out the threat. A social worker may want to determine whether a person is likely to commit suicide

or otherwise harm him- or herself based on information obtained during a court-mandated interview of the person.

Most people would assume that such predictions are left to experts in the domain, and that there is no substitute for expert judgment. However, research has shown that even when human judges follow the right rules when attempting to make predictions of this sort, they often follow them inconsistently, so that they make one prediction one day but make a different prediction based on the same information on another day or for another case with identical information. For this and other reasons, many studies comparing predictions of human judges and predictions made from mathematical formulas have found that mechanical, mathematical approaches tend to outperform human judges far more often than the opposite. For a discussion of this point, related research, and some counterarguments, see Dana and Thomas (2006), Grove, Zald, Lebow, Snitz, and Nelson (2000), Kleinmuntz (1990), Litwack (2001), Meehl (1957), and Wiggins (1973).

It is understandable how some might take offense to using a formula or computer algorithm to make decisions about people—decisions that have consequences on their lives and thus should be made with people who are informed about the specifics of the case at hand and can be compassionate and reasoned in their approach. But computers are particularly good at these kinds of things because they have no stake in the decision and can apply the rule, once it is constructed, consistently and without bias, exception, and (for better or worse) emotion. We still need people to figure out what things should go into a cold, calculating, mechanical prediction system or algorithm, in the same way that computers can't predict the weather without human beings telling the computer what to use to make the prediction. But once human beings have decided what inputs to put into the formula, the research suggests it is best to then leave people out of the process and let the computer aggregate the information to make the actual prediction in any particular case.

7.1.2 Regression as a Mechanical Prediction Method

There are many ways that a computer could make a prediction of some variable Y from a set of input variables X. Regression analysis is only one of those ways, but it is the one we focus on in this chapter. It can be particularly good when it has a lot of data upon which to generate the decision, meaning that the formula was constructed from a large number of cases used to calibrate the prediction system. For small samples, there

are alternatives that we cannot begin to discuss here; see Wiggins (1973) and Darlington (1978).

Consider a very simple case in which a college admissions office is considering the relative importance of tests of verbal and mathematical ability, which we'll call X_1 and X_2, respectively, in predicting an applicant's success in college. One admissions officer might think that verbal ability is more important than mathematical ability, for without being able to speak and write effectively, a student will have a hard time succeeding in a rigorous academic environment. If the goal is to predict, say, college GPA at the end of the first year, this admissions officer would presumably give more weight to X_1 when making the prediction, and might even ignore X_2 entirely. This officer might choose to admit only those students who achieve a certain score on X_1. But a second officer might argue that there is no reason to weight one more than another or discard information in X_2: They can just be weighted equally. This officer would just add up X_1 and X_2 and decide to admit a student based on his or her score on this sum. Still another officer might agree with the first but not be willing to completely discount mathematical ability. Perhaps this officer would argue that verbal ability should be given twice as much weight, and so the criterion for admission should be based on achieving some value on $2X_1 + X_2$.

So here we have three fairly simple mechanical rules for choosing which applicants to admit. The first officer advocates $1X_1 + 0X_2$ (which is the same as just using X_1, ignoring X_2 entirely), the second advances $X_1 + X_2$ as the prediction rule, and the third recommends the use of $2X_1 + X_2$. Each of these would require some kind of decision criterion for determining whether a student should be admitted, such as a certain score on this favored index.

Regression is well suited to dealing with problems such as this. Not only can regression analysis be used to generate a formula to produce the prediction, but it can also help to resolve questions involving how to best weight X_1 and X_2. Suppose the goal is to predict first-year GPA, which will be Y in the model. To do so, take a sample of current students or students from the past who have already completed 1 year and thus have first-year GPAs that are known. Presumably these students' records also contain their scores on X_1 and X_2, which we call *predictors* in this context. Then use a regression analysis to estimate Y from predictor variables X_1 and X_2 based on these cases with known first-year GPAs. The regression formula will generate $\hat{Y}$, which is a weighted sum of X_1 and X_2, that is maximally correlated with Y in the data available. With this optimal weighted sum of X_1 and X_2 constructed, it can be applied to future students who have

applied for admission to generate a prediction for their GPAs at the end of the first year. Perhaps the admissions office decides then to admit those students whose predicted first-year GPA is at least 3.0.

In this example, you may have assumed that X_1 and X_2 are measured on roughly equal scales, such as is the case with the verbal and quantitative sections of the Scholastic Aptitude Test used in the United States for college admissions. If X_1 and X_2 were not on comparable metrics, then the selection rules advocated by officers two and three above wouldn't weight these predictors appropriately. For instance, if X_1 were on a 200 to 800 scale and X_2 were on a 1 to 10 scale, then $X_1 + X_2$ would not equally weight these two variables, because the sum would be determined almost entirely by X_1. Of course, if the tests had different scales, this would have complicated the discussion between the admissions officers.

But regression requires absolutely no assumption that different predictors are measured on comparable scales. For instance, X_1 and X_2 might be tests of verbal and mathematical ability that are on the same scale, but the regression might also include an X_3, high school GPA, which may be on a 1 to 4 scale. The regression formula will still find the weighted sum of X_1, X_2, and X_3 that yields the best prediction of Y in the sample of cases available. Regression compensates for a predictor's low range by giving it a larger regression coefficient in the model. For instance, if we changed the range of high school GPA from 1 to 4 to 100 to 400 by multiplying all the GPAs by 100, it would not change the predictions at all but would simply make b_3 one-hundredth the size it would be otherwise.

7.1.3 A Focus on *R* Rather Than on the Regression Weights

There are a great many differences between regression's use in prediction problems such as this and when it is used in causal analysis. In prediction we refer to *predictors* and the *criterion* instead of regressors and the dependent variable. There is no distinction between independent variables and covariates—the terms are not used. The word *validity* in prediction problems refers to the accuracy with which Y can be predicted. Variables are included or excluded from the analysis primarily based on availability and ease of measurement. Interest focuses on the multiple correlation R and related statistics, with rather little emphasis on the individual predictors in the model. And when attention *is* directed at a specific predictor rather than at how well the model as a whole estimates Y, focus is directed on its predictive power, meaning how much it contributes to increasing R, rather than on its regression coefficient. Indeed, a valid regression analysis

used for prediction need not involve tests of significance for the individual regression coefficients or their predictive power. More important is *cross-validation* or a close substitute, which addresses how well the prediction formula predicts when it is used on a data set different from the one used to generate the equation in the first place. This is not something ordinarily undertaken in causal analysis. Modified versions of regression analysis, such as ridge regression and stepwise regression, are appropriate for prediction but not for causal analysis.

7.2 Estimating True Validity

7.2.1 Shrunken versus Adjusted *R*

In section 4.3.1 we described how R is a biased estimator of ${}_TR$ and that adjusted R^2 (equation 4.1) is a better estimator of ${}_TR^2$ than is R^2. Recall that R quantifies the correlation between Y and $\hat{Y}$ in the sample data when $\hat{Y}$ is derived from the sample regression weights for the predictors. It tells us how well the variables in the model predict Y in the sample. By contrast, ${}_TR$ is the correlation between Y and $\hat{Y}$ using the population data and using the population regression weights for each of the predictors to generate $\hat{Y}$. Thus, there are two ways in which ${}_TR$ is a population value relative to R. It is derived using the population of data on the variables in the model. It is also based on the population regression weights for the variables rather than the sample regression weights.

If our goal is to estimate how well the variables in a regression model would predict Y if we had the population of data available to generate the model (and thus we would not have to estimate the regression coefficients—they could be known exactly), then we care about ${}_TR$. However, in applied prediction problems, we are interested in a different quantity, which we denote ${}_TRS$ for *shrunken R*. (We use *RS* rather than *SR* because we already used *SR* to refer to a semipartial multiple correlation in section 5.3.1). ${}_TRS$ quantifies how well the regression model generated from a sample predicts Y when the sample regression model is applied to the population. In other words, if the regression model *based on the sample regression weights* is used to produce an estimate of Y in the population of data, then ${}_TRS$ is the resulting correlation between Y and $\hat{Y}$. This tells us about the predictive power of the sample regression weights when the model is used to generate estimates of Y in the population. We care about this in an applied prediction problem, because when making predictions, the model being used to generate the

TABLE 7.1. Shrunken R versus R and ${}_TR$

		Compute $r_{Y\hat{Y}}$ in:	
		Sample	**Population**
Compute b's in:	**Sample**	R	${}_TRS$
	Population		${}_TR$

prediction is derived from the sample regression coefficients generated in a single sample.

If the preceding two paragraphs are confusing after a first reading, we recommend reading them again. If the distinction between R, ${}_TR$, and ${}_TRS$ remains unclear, Table 7.1 may help. In this table we define these three quantities based on the whether the weights are generated in a sample or using the population, and whether the correlation between Y and $\hat{Y}$ is computed in the sample or the population.

A parameter is fixed, though typically unknown. It must be estimated using some data, but doing so does not change the parameter. A parameter does not vary from sample to sample, so it is independent of the sample drawn. So whereas R varies from sample to sample, ${}_TR$ does not. ${}_TRS$ is a fixed quantity, like a parameter, but unlike ${}_TR$, there is no single value of ${}_TRS$, because it is determined by the sample regression weights used to generate the estimate of Y in the population. Different samples will generate different sample regression weights, and therefore different prediction models, which generate different estimates of Y when applied to the population. Thus, two investigators who use different samples to produce a model to estimate Y in a population will have different models. They may each want to estimate ${}_TRS$, but the quantity they are attempting to estimate is probably not going to be the same, because ${}_TRS$ depends on the sample used to generate the model.

Since ${}_TR$ is defined as $r_{Y\hat{Y}}$ when the population weights ${}_Tb_j$ are used to compute $\hat{Y}$, no other set of weights can yield a higher value of $r_{Y\hat{Y}}$ in the population. Thus, ${}_TRS$ cannot possibly exceed ${}_TR$ and will usually fall below it, often by a substantial margin. This is very important when we are interested in prediction; adjusted R is sometimes used as an estimate of ${}_TRS$, but instead it estimates the completely different value ${}_TR$. ${}_TRS$ also almost always falls below R; this is known as *validity shrinkage*.

7.2.2 Estimating $_TRS$

$_TRS$ can be estimated in at least four different ways. We address each of these ways in this section.

Cross-Validation. The first method, *cross-validation*, requires two samples of data with measurements on all predictors, as well as Y. The two samples may be independently drawn random samples from the same population or one large sample that is randomly split into two subsamples. Regression weights b_j are calculated in the first sample, then the regression model based on these regression weights is used to generate $\hat{Y}$ in the second sample. The correlation between Y and $\hat{Y}$ in the second sample is used as an estimate of $_TRS$.

A validity estimate based on cross-validation is not $_TRS$ itself. Rather, the correlation between Y and $\hat{Y}$ in the second sample is an approximately unbiased estimator of $_TRS$. Validity estimates constructed in this fashion are subject to the same sampling errors as any correlation. When used to estimate $_TRS$, the ordinary multiple correlation R suffers from both bias and random variability; cross-validation removes the bias but not the random variability.

To handle this problem, the method described in section 4.5.2 based on Fisher's r-to-Z transformation for constructing a confidence interval for a correlation can be applied to generate an interval estimate for $_TRS$. A little experimentation with the r-to-Z method shows that if much faith is to be put in cross-validity figures, cross-validation samples must be moderately large. For instance, suppose we decide in advance that if our estimate of $_TRS$ turns out to be 0.4, we would like to be 95% confident that $_TRS$ is at least 0.3. It turns out that the necessary cross-validation sample (i.e., the second sample in the method described above) size is 211. If the first sample contained 100 cases, then over 300 in total would be needed. So cross-validation with far smaller samples should be viewed as more of a negative test than a positive one. That is, with smaller samples, a low cross-validity means we can have little faith in the predictions made from the regression, but a high cross-validity may not mean we can have much faith in them.

Double Cross-Validation. The second approach to estimation of $_TRS$ is an extension of cross-validation called *double cross-validation*. With this approach, the total available sample is divided in half, a regression model is estimated in each half, and then each of these regression models is used to estimate $\hat{Y}$ in the other half. This generates two correlations between $\hat{Y}$ and Y. The average of these two correlations is then used as a conserva-

tive estimate of the validity of the regression model generated in the total sample.

Leave-One-Out. A third method called the *leave-one-out* method eliminates nearly all the conservatism of ordinary and double cross-validation. We describe two leave-one-out approaches. Imagine you have N cases and you regress Y on a set of k predictors, but after excluding the first case in the data. Using the regression model based on these $N - 1$ cases, you then generate an estimate of Y for the first case discarded from this analysis. Then repeat this process by returning the first case to the data but discarding the second case, estimating Y from the predictors, and using this model to generate an estimate of Y for the second case. Repeat this for a total of N times, at which point you will have estimated Y for each case in the data based on a model that excludes it. At the end of this procedure, you have N values of Y and N values of $\hat{Y}$. You could use any measure of correspondence between $\hat{Y}$ and Y as a cross-validity estimate, such as the correlation between Y and $\hat{Y}$. This is a bona fide estimate of ${}_{T}RS$.

Although it may seem like this procedure would take a lot of computer time, this turns out not to be a problem. Computers are quite fast these days, and conducting thousands, tens of thousands, or even hundreds of thousands of regressions is a fairly routine computational problem accomplished by modern computers quite rapidly. Furthermore, it turns out that it isn't even necessary to conduct all N leave-one-out regressions, because it can be shown that the error in prediction made by leaving out case i from the regression is simply the residual for case i from a model containing all N cases divided by $1 - h_i$, where h_i, defined in section 16.1.3, is also derived from the regression based on all N cases. So in fact, only one regression based on all N cases is required to implement the leave-one-out method. In section 7.2.3 we show how this can be done in SPSS in a few lines of code.

One problem with this approach is that it can yield negative estimates of ${}_{T}RS$. This is in part due to the fact that the higher the Y of the omitted case, the lower is $\overline{Y}$ when that case is excluded. This will tend to increase the size of the errors produced by this method. When applied to random data, where the correlation between Y and $\hat{Y}$ should be zero, this method typically yields negative estimates of the correlation between Y and $\hat{Y}$. A second leave-one-out approach corrects this problem by computing the other $N - 1$ values of Y calculated from the regression based on $N - 1$ cases, computing the mean and standard deviation of Y of those $N - 1$ cases, and expressing $\hat{Y}_i$ in standardized form using that mean and standard

deviation. This too can be done in far less computing time than one would think.

Formulaic Methods. A fourth means of estimating ${}_TRS$ is through one of many formulas that have been advanced for this purpose. Yin and Fan (2001) have reviewed and studied a variety of these methods and find, as do we, that the method proposed by Browne (1975) works quite well, at least when the predictors are multivariate normal. This estimator is based in part on adjusted R^2 as calculated using equation 4.1 and denoted below as R_a^2. If adjusted R^2 is zero, then don't bother with the computations below and consider $RS = 0$. But if adjusted $R^2 > 0$, then calculate

$$\rho^4 = R_a^2 - \frac{2k(1 - R_a^2)^2}{N - k - 1}$$

If $\rho^4 > 0$, use ρ^4 in equation 7.1 to produce RS. But if $\rho^4 \leq 0$, then consider $RS = 0$.

$$RS = \sqrt{\frac{(N - k - 3)\rho^4 R_a^2}{R_a^2(N - 2k + 2) + k}} \tag{7.1}$$

Table 7.2 shows values of RS calculated using equation 7.1 for various ratios of sample size and number of predictors. To produce this table, RS was calculated for each of the 96 values of k from 5 to 100; the table shows the minimum and maximum values yielded by the formula for a fixed N/k ratio for different values of adjusted R. Observe that when both adjusted R and N/k are small, RS is substantially smaller than adjusted R. This shrinkage itself shrinks as adjusted R is increased, or as N/k is increased. Shrinkage is noticeably smaller but still not zero, with as many as 20 or 30 times more cases than predictors.

All four of these methods—ordinary and double cross-validation, leave-one-out, and the Browne method—yield approximately unbiased estimates of ${}_TRS$. Ordinary cross-validation gives a fairly exact method—the Fisher r-to-Z method—for finding confidence limits on ${}_TRS$. We have found that when ${}_TRS$ is low, the standard error of the leave-one-out methods is about $\sqrt{2/N}$; the number of predictors k has very little effect. When ${}_TRS$ is high, the standard error is smaller than $\sqrt{2/N}$, so $\sqrt{2/N}$ is conservative. The leave-one-out methods and the Browne (1975) method give very similar estimates of ${}_TRS$ when the distributions of the regressors are approximately normal and both estimates are positive; the major difference between the two is that the leave-one-out method can yield negative estimates of ${}_TRS$, while the Browne method cannot. Thus, $\sqrt{2/N}$ can be taken as a conservative

TABLE 7.2. Values of Expected Shrunken *R* Using the Browne (1975) Estimator

		Adjusted *R*						
N/*k*		0.2	0.3	0.4	0.5	0.6	0.7	0.8
2	Min	.020	.030	.075	.162	.302	.453	.611
	Max	.090	.137	.185	.241	.355	.487	.638
3	Min	.020	.049	.119	.255	.414	.560	.705
	Max	.088	.131	.202	.313	.435	.567	.707
4	Min	.020	.067	.160	.327	.469	.603	.734
	Max	.087	.133	.237	.352	.475	.604	.737
5	Min	.025	.085	.217	.366	.499	.624	.749
	Max	.085	.153	.261	.377	.499	.627	.753
6	Min	.030	.098	.252	.390	.515	.637	.759
	Max	.084	.168	.278	.395	.518	.642	.763
8	Min	.040	.141	.292	.418	.536	.653	.770
	Max	.086	.190	.302	.420	.541	.658	.774
10	Min	.048	.179	.315	.433	.549	.663	.776
	Max	.099	.205	.318	.437	.554	.667	.780
12	Min	.058	.201	.329	.443	.557	.669	.780
	Max	.108	.216	.329	.448	.562	.673	.783
15	Min	.073	.222	.341	.454	.565	.676	.785
	Max	.119	.228	.344	.459	.570	.679	.787
20	Min	.106	.241	.354	.465	.574	.682	.789
	Max	.132	.242	.358	.469	.578	.685	.790
25	Min	.127	.252	.363	.472	.579	.685	.791
	Max	.141	.253	.366	.476	.582	.688	.792
30	Min	.140	.258	.368	.476	.583	.688	.792
	Max	.148	.261	.372	.480	.585	.690	.794

estimate of the standard error of *RS* when calculated using equation 7.1 when the distributions of the regressors are approximately normal.

7.2.3 Shrunken *R* Using Statistical Software

Almost all statistical software that does regression analysis generates *R* and adjusted *R*. Few if any widely used programs produce estimates of *RS* using the methods described in section 7.2.2. In Appendix A we document a macro called RLM for SPSS and SAS that conducts regression analysis that includes a number of features not ordinarily available in these two programs, including the production of three measures of *RS*. The example output in Figure 7.1 is from a regression based on 340 people esti-

```
Outcome Variable
pknow

Complete Model Regression Summary
          R       R-sq   Adj R-sq          F          p    SEofEst
      .5120      .2621      .2511    23.7311      .0000     3.7843

Shrunken R estimates
     Browne     LvOut1     LvOut2
      .4922      .4863      .4942

ANOVA summary table
                 SS         df         MS
Regress   1699.2879     5.0000   339.8576
Residual  4783.2857   334.0000    14.3212
Total     6482.5735   339.0000    19.1226

Regression Model
              Coeff         se          t          p       LLCI       ULCI
constant     7.0122      .5872    11.9416      .0000     5.8571     8.1673
natnews       .1765      .0885     1.9954      .0468      .0025      .3506
npnews        .4141      .0759     5.4558      .0000      .2648      .5634
locnews      -.4511      .1029    -4.3839      .0000     -.6535     -.2487
talkrad       .6596      .1660     3.9725      .0001      .3330      .9862
pdiscuss      .4579      .0844     5.4244      .0000      .2918      .6239

Simple (r), semipartial (sr), and partial (pr) correlations with outcome
                 r         sr         pr
natnews       .1481      .0938      .1085
npnews        .2983      .2564      .2861
locnews      -.1065     -.2061     -.2333
talkrad       .2609      .1867      .2124
pdiscuss      .3465      .2550      .2845
```

Estimates of $_{T}RS$

FIGURE 7.1. Output from the RLM macro for regression analysis.

mating scores on a test of political knowledge from measures of exposure to various sources of information about politics (e.g., newspaper reading, talking about politics, watching the local news; the data can be found at *www.afhayes.com* and is called POLITICS). The variables used or the substantive interpretation of the model is not important here. But observe that in addition to R and adjusted R, the output produces both leave-one-out estimates as well as Browne's estimate of RS from equation 7.1. In this case, they are all very similar, and not much smaller than adjusted R given that the N/k ratio is over 60.

This macro does produce the first leave-one-out estimate of $_{T}RS$ described in section 7.2.2, but it is worth pointing out how easy it is to generate in at least some statistical programs. Let $_{d}e_i$ be the difference in $\hat{Y}_i$ when a case is excluded versus included from the model. Let $\hat{Y}_i$ be case i's estimate of Y when all cases are in the analysis, and $\hat{Y}_{i,\text{not}\,i}$ be case i's estimate of Y when it is excluded. That is, $\hat{Y}_{i,\text{not}\,i}$ is generated by regressing Y on the set

of predictors after excluding case i, and then the resulting model is used to generate $\hat{Y}$ for that excluded case. ${}_d e_i$ is defined as

$$ {}_d e_i = \hat{Y}_i - \hat{Y}_{i,\text{not } i} $$

Though not obvious from this formula, ${}_d e_i$ is equivalent to the difference in case i's residual when the case i is used to estimate the model versus when it is not. That is, ${}_d e_i = (Y_i - \hat{Y}_i) - (Y_i - \hat{Y}_{i,\text{not } i})$, which simplifies to $\hat{Y}_i - \hat{Y}_{i,\text{not } i}$. With ${}_d e_i$ calculated,

$$ \hat{Y}_{i,\text{not } i} = \hat{Y}_i - {}_d e_i $$

Since $\hat{Y}_i$ is easily constructed, and with ${}_d e_i$ automatically calculated by a regression routine, it is easy to then construct $\hat{Y}_{i,\text{not } i}$ and correlate that with Y. The resulting correlation is an estimate of ${}_T RS$ using the first leave-one-out method.

The code for SPSS below does this for the same data set that generated Figure 7.1. In this code, **pred** and **dfit** are options to produce Y_i and ${}_d e_i$ that are added automatically by SPSS to the data set as *pre_1* and *dfi_1*. The next line of code constructs $\hat{Y}_{i,\text{not } i}$. The correlation between Y (*pknow* in the data) and the new variable constructed named *yhatnoti* by this code is the estimate of ${}_T RS$ using the first leave-one-out method. The last line of the code generates this correlation.

```
regression/dep=pknow/method=enter natnews npnews locnews talkrad
   pdiscuss/save pred dfit.
compute yhatnoti=pre_1 - dfi_1.
correlations variables = pknow yhatnoti.
```

7.3 Selecting Predictor Variables

We have seen that ${}_T RS$ is nearly always below ${}_T R$, which in turn is nearly always below R. Although regression analysis finds a set of regression weights b_j that maximizes R, other methods may yield predictions of Y that are more accurate for the population.

When k predictor variables are available for use in predicting Y, regression yields the model with the highest R, but that model may not be the best when the goal is to maximize ${}_T RS$. With k variables available for use in predicting Y, there are a great many possible models that could be constructed from this set of k when you consider models that exclude one or more of these k variables. Perhaps a better model contains only a subset of these k. It is not immediately obvious how to choose from among

the possible alternative candidate models. Adding a variable to a model almost always raises R, meaning that larger models (i.e., those with more predictors) will generally have larger values of R, but the addition of some of those predictors may not actually improve the validity of the model as measured by ${}_TRS$. Indeed, the more variables in the model, the larger the validity shrinkage.

One of the simplest ways of addressing validity shrinkage is to reduce the number of variables used as predictors in the model, as validity shrinkage increases with the number of predictor variables used to estimate Y in the sample. The methods for selecting predictor variables discussed here not only address validity shrinkage but also can be used to reduce a large number of potential predictors to a smaller and potentially more manageable number that could be used when the model is applied to future cases. Thus, these methods can also have some practical value, even if none of them can be used to establish that any one model is better beyond a reasonable doubt.

The methods described here are crude in comparison to other prediction methods available in the literature on psychometric theory. Their major advantage over those other methods is that they can be applied with ordinary regression computer programs. However, the methods we discuss can be quite satisfactory if N is large relative to k. More complex methods are discussed in books on psychometric theory as well as in Darlington (1978), including some that that can be effective when N is small and k is large.

7.3.1 Stepwise Regression

Stepwise regression exists in two primary forms: *forward* and *backward*. They are both based on the the goal of maximizing the correlation between Y and $\hat{Y}$ while using as few predictors as needed to do so. In order to accomplish this, some kind of decision rule is needed to determine whether adding or removing a predictor changes R to a degree worth concerning oneself about, since we know that adding or removing predictor variables will, except in very unusual circumstances, change R to some extent. Most regression programs that can do stepwise regression give the data analyst some control over the criterion used for deciding whether to add or remove a variable.

Forward Stepwise Regression. In its simplest form, a forward stepwise regression starts by selecting, from a set of k predictor variables determined by the data analyst, the one predictor with largest absolute correlation with Y; that is, $|r_{YX_j}|$. We will call this variable P_1 to denote that it was the first

predictor selected. This leaves $k - 1$ remaining variables for consideration for addition to the model. The stepwise procedure then chooses from these remaining $k - 1$ variables that which most increases R when it is added to the model containing only P_1. Call this P_2 for the second predictor selected. This procedure continues until all k variables have been added, producing a model with all k predictor variables. This method generates k models, and the data analyst can choose that model that balances large R with a small number of predictors. Variables added later generally contribute less to increasing R, so those variables added later may not be worth keeping in the prediction model, especially if it is difficult or otherwise burdensome to collect the data on those variables in future cases to which the model is likely to be applied.

The procedure just described is often carried out not by constructing all k models but, instead, by using a statistical significance test for deciding whether to add a variable to a model or to stop the selection process entirely. In the first step, a variable P_1 is chosen only if correlated with Y by a statistical significance criterion, such as a p-value of less than .05. If no variables among the k are significantly correlated with Y, the procedure stops with no predictors. Assuming one is found, the second step chooses a predictor variable from the remaining $k - 1$ that increases R the most *and* to a statistically significant degree (meaning that the p-value for its regression weight is less than or equal to .05) when it is added to a model containing only P_1. If no such variables exist, the procedure stops with only P_1 as the sole predictor. But if one is found, the process adds P_2 to the model discovered in this step. This process continues until no more variables that haven't yet been added increase R to a statistically significant degree or the number of candidate predictors is exhausted. Note that the first procedure could be thought of as a version of this more refined procedure but using a p-value of less than 1 as the decision criterion for adding a variable to the model.

This procedure can be refined further by allowing a variable already selected to be removed at a later step. At later steps in the forward selection process, it is possible for a variable that significantly increased R at a prior step to become nonsignificantly associated with Y, with other variables in the model added after it. In this case, removing the variable wouldn't significantly lower R, and so it becomes a candidate for removal. Thus, this refinement of forward stepwise regression is really a combination of forward and *backward* stepwise selection. We address backward stepwise selection next.

Backward stepwise regression. Whereas forward selection starts with no predictors and then builds the model by adding predictors one at a time, backward stepwise selection begins with a model with all k predictors and then removes them one at a time using some kind of criterion for removal. In forward stepwise regression, a variable is added at a step if it increases R the most. In backward selection, at each step a variable is removed that *lowers R the least* relative to others still in the model. This process continues until only one variable remains.

In practice, a statistical significance test often is used to determine whether or not a variable is removed at a certain step or whether the process terminates. Unlike with forward selection, a variable is kept in a model if removing it would significantly lower R (which is equivalent to saying that adding it to a model without it would increase R to a statistically significant degree). If more than one variable meets this test, the one that least reduces R is the winner for removal at that step. The process terminates when no variables can be removed without significantly lowering R.

Most statistics programs that conduct linear regression analysis can do forward or backward stepwise variable selection. For instance, suppose you are trying to build a model of first-year college GPA that maximizes R without including a bunch of unnecessary predictors (which would increase validity shrinkage). Suppose among six candidate predictors are scores on the verbal, analytical, and writing sections of the Scholastic Aptitude Test, GPA in courses taken in years 3 and 4 of high school, and the number of sports a person played in high school. The SPSS command below would conduct a forward stepwise variable selection, using the default of a p-value of .05 for variable entry.

```
regression/dep=cgpa/method=forward vsat asat wsat gpa3 gpa4 sports.
```

Comparable commands in SAS and STATA are

```
proc reg;
   model cgpa=vsat asat wsat gpa3 gpa4 sports
   /selection=forward slentry=0.05;
run;
```

```
stepwise, pe(.05):  regress cgpa vsat asat wsat gpa3 gpa4 sports
```

In SPSS, changing the selection method to backward stepwise is as simple as changing **forward** in the command above to **backward**. In SAS or STATA, code for backward selection for this problem would be

```
proc reg;
    model cgpa=vsat asat wsat gpa3 gpa4 sports
    /selection=backward slstay=0.05;
run;
```

```
stepwise, pr(.05):  regress cgpa vsat asat wsat gpa3 gpa4 sports
```

7.3.2 All Subsets Regression

The number of possible models from a set of candidate predictors is typically quite large even for fairly simple problems. Specifically, if you are considering k variables as possible predictors in your model, there are 2^k possible models if you include the model with no predictors at all. This number can be huge; when k is 5, 10, 15, or 20, 2^k is 32, 1,024, 32,768, and 1,048,576, respectively.

When using forward or backward stepwise selection, only a small subset of the possible models one could construct from the k candidate predictors is evaluated. To see why, consider you are using forward selection with $k = 10$ candidate predictors, and the method chooses variable 2 as the first predictor to include. Unless you allow for removal of variable 2 later in the process, you know that any of the 512 models that excludes variable 2 among the 1,024 possible models is ruled out for consideration in future steps. Yet the optimal model in terms of maximizing R while minimizing the number of predictors may very well be one of those 512 models. Once the second variable is chosen, still more of the possible models are ruled out from further consideration even though one of those may be the best model. A similar logic applies to backward selection.

All subsets regression gets around this problem by considering all of the possibilities. Using this method, a statistic called *Mallow's* C_p is often used to choose the "best" of the 2^k possible models. But in our opinion, there is no clear rationale for using C_p. A superior strategy would be to use any of the estimates of ${}_TRS$ described in section 7.2.2. If you want to give all subsets regression a try, it is available as an option in STATA and SAS, as well as in the RLM macro discussed in Appendix A.

7.3.3 How Do Variable Selection Methods Perform?

Once we have used stepwise or all subsets regression to select a prediction model, the estimates of ${}_TRS$ described earlier no longer apply, and p-values and confidence intervals lose their meaning and usefulness. This is like

regression to the mean; once we have selected a model with the highest value of R from among many models, we can no longer have confidence that its true value is as high as it appears.

Numerous studies have been conducted that compare the performance of different methods of variable selection, including those discussed here. To test these methods, researchers can generate samples from populations mathematically defined where certain variables are known to be useful in increasing R and others are not. The methods are evaluated relative to others by examining how often they select the best model (which, of course, is defined by the researcher and thus known), and how often they choose models that include useless predictors or accept models that don't include at least some of the useful predictors. This literature has shown that all methods are vulnerable to some extent to choosing the wrong model. For specifics of some of these findings, see Derksen and Keselman (1992) and Flack and Chang (1987), among many others.

Furthermore, different methods often select different "best" models when applied to the same data. To consider how this sometimes can happen, revisit the example in section 3.4.1. Recall from that example that while neither skill at baseball nor skill at basketball significantly predicted preference for basketball over baseball on their own, when both were put in the model, $R = 0.97$. In this example, forward selection relying on a statistical significance criterion for variable selection would select as the best model of preference the one with *no predictors at all*, as neither variable significantly predicts preference. Yet backward selection would choose the model with *both* predictors as best, because removing either performance variable would significantly lower R.

In this case, we know that backward selection is getting it right. In general, you'd expect backward selection to do better than forward selection, because it would be less susceptible to problems produced by collinearity between predictor variables, as in this example, that would lead forward selection to reject predictors that should be in the model. Yet the literature also shows that backward selection can select models that include useless predictors, just as forward and all subsets can, and this is a problem given that the goal of these methods is to find a model that maximizes the correlation between Y and $\hat{Y}$ while minimizing the number of predictors.

Given the documented problems of these selection methods, it is hard to have too much faith in the model that any one of them derives in a specific circumstance as the "best" of possible models given the k candidate predictors used. As a general rule, you can assume that the model selected

by any of these methods will tend to *overfit* the available data and offer a more optimistic prognosis of its fit to future data than is likely to be realized.

To understand the problem of overfitting, suppose you were trying to estimate the mean of some population and you took 1,000 samples of 100 from this population. If you selected the largest mean of the 1,000 sample means as your estimate of the population mean, you would almost certainly end up overestimating the population mean. Or suppose you calculated the correlation between two variables in 10 samples and used the highest correlation as your estimate of the population correlation in a research article. It is likely that in doing so, you would be overestimating the size of the population correlation.

Automated variable selection methods suffer from this same kind of problem to varying degrees, with all subsets regression being most vulnerable. Regression is good at extracting information from the observed relationships between predictors and criterion. Its job is to do so such that estimation of Y in the sample cannot be improved upon without fundamentally changing the form of the model in some fashion. When you have conducted many regressions with different subsets of variables from a set of k candidates (with all subsets being the most extreme form), the model that looks best in the sample is not likely to be quite as good when it is applied to future data. It may be that the model chosen is the best of the candidate models considered (though it may not be), but it is very likely not to fit future data as well as it fit the data used to construct it.

But these methods can be still be useful if their limitations are acknowledged and proper precautions are taken. The goal of these methods is to produce a model that makes good predictions of Y when applied to cases not used to construct the model. When it is applied, examine how well the model performs using some of the methods discussed elsewhere in this chapter, and be willing to modify the model later if new data suggest the original model does not perform as well as expected (and most likely, it will not). The construction of a useful prediction model should be an iterative process that is informed by new data, which feeds into modifications that are then validated with the next round. Such a process will, over time, likely result in a model that performs better than the model initially used.

Resist the temptation to employ an automated variable selection method when regression is being used for causal analysis and statistical control. Variable selection methods are more appropriate when research emphasis is directed toward maximizing the correlation between Y and $\hat{Y}$. In causal analysis, much greater emphasis is placed on regression coefficients

and measures of partial association rather than on the multiple correlation. When the focus is on the relationship between Y and a particular independent variable or set of independent variables, then, of course, those independent variables should always be in the model and not a candidate for exclusion using a variable selection algorithm. It is likely when doing causal analysis that you have reason to believe that any associations observed may be attributable to other variables that correlate with the independent and the dependent variable. When a critic, existing research, or theory suggests certain variables should be controlled, then those covariates should be in the model too, and not on the chopping block for potential exclusion using a variable selection method.

7.4 Predictor Variable Configurations

Although focus is on R rather than on the regression weights when using regression analysis for prediction, the regression weights certainly do matter for the prediction. If a predictor variable j receives no weight in the prediction formula, then it is not used in generating $\hat{Y}$, meaning that variation in predictions is not determined at all by predictor j. This is equivalent to excluding variable j from the prediction formula entirely. Conversely, the more predictor variable j's weight deviates from zero, the more variation in predictor j maps onto variation in $\hat{Y}$.

In a simple two-variable model predicting Y from X_1 and X_2, the formula for the regression weight for X_1 is provided in section 3.4.5. We repeat it below for convenience.

$$b_1 = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \times \frac{s_Y}{s_{X_1}} \tag{7.2}$$

In this discussion we can ignore s_Y/s_{X_1}, as these merely scale the regression weight for X_1 in terms of the measurement scales of X_1 and Y. Examining equation 7.2, it is apparent that whereas b_1 is clearly determined in part by X_1's simple correlation with Y (r_{YX_1}), it is also determined by the correlation between Y and X_2 (r_{YX_2}), as well as the correlation between X_1 and X_2 ($r_{X_1X_2}$). So we can't say that X_1 will necessarily receive a nonzero regression weight just because it is correlated with Y. We also cannot say that a regression weight of zero for X_1 implies no correlation between X_1 and Y. It is possible for X_1 to receive a weight of zero or nearly so even if X_1 is highly correlated with Y. It's weight could even differ in sign from its simple correlation,

meaning that a variable that is positively correlated with Y could be given a *negative* weight in a prediction formula.

In this section and the next, we discuss several configurations of intercorrelation between predictors and between predictors and Y. Although we couch this discussion in terms of prediction, they are also relevant to the use of regression for causal analysis. The main configurations are *independence*, *redundancy*, *complementarity*, and *suppression*. The latter three can be either *partial* or *complete*. Thus, there are seven configurations in all: independence (which can only be complete), partial and complete redundancy, partial and complete complementary, and partial and complete suppression.

The four "complete" configurations all require certain exact equalities (e.g., a certain correlation may have to be exactly zero), while the three "partial" configurations merely require certain inequalities (e.g., a certain correlation may have to be negative). Thus, the three partial configurations are the ones usually observed. But the four complete configurations can help you understand the others, and they do occur occasionally. Partial redundancy is by far the most common of all seven configurations, and so we start with it.

7.4.1 Partial Redundancy (the Standard Configuration)

In section 3.4.2 we introduced the Venn diagram as a visual aide to understanding the distinction between the semipartial and partial correlation. In a Venn diagram, areas of overlap between variables depicted as circles represent correlation or shared variance. Remember that we also said that Venn diagrams can't depict every situation that can happen mathematically. This remains true, as will be seen. We again use the Venn diagram to illustrate some of the configurations that can be depicted visually in this manner.

In that Venn diagram, which we replicate in Figure 7.2, panel A, the variables were in the standard configuration, or what we are calling partial redundancy here. In this configuration, the predictors are correlated with each other, and correlated predictor variables typically (though not necessarily) duplicate each other's function to some extent in the role they play in estimating Y. The intercorrelation between predictors is visually depicted in this Venn diagram as the area B + E. When two variables are intercorrelated, it is often the case that one variable contributes less to explaining variation in Y in the presence of the other variable in the model. For instance, suppose you want to estimate future performance from two

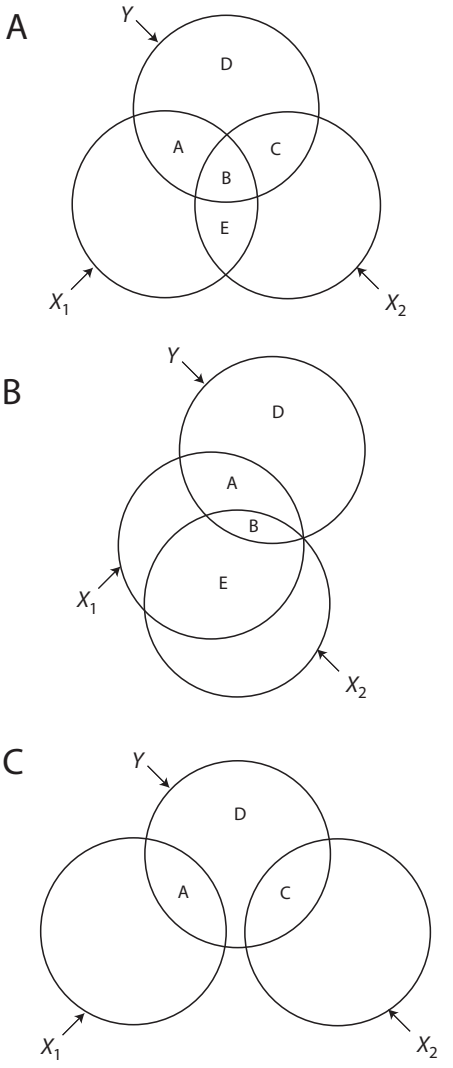


FIGURE 7.2. Venn diagrams depicting partial redundancy (A), complete redundancy (B), and independence (C).

tests of verbal ability administered in high school. Most likely, the two test scores will be somewhat correlated, perhaps moderately or even highly so. Using them both to predict future performance may result in less accurate predictions than if we used one of the measures of verbal ability and something else correlated with Y but less correlated with verbal ability, such as mathematical ability. The two measures of verbal ability are likely to be partially redundant. Although there may be some redundancy between verbal and mathematical ability, their redundancy is likely to be less than the redundancy between the two measures of verbal ability.

7.4.2 Complete Redundancy

Imagine pushing the X_1 and X_2 circles closer together in Figure 7.2, panel A. As you do so, redundancy between X_1 and X_2 would increase as B + E increases in size, and the areas A and C shrink relative to B. At its extreme, if X_1 and X_2 completely overlap, then they are completely redundant. However, complete overlap between the X_1 and X_2 circles in the Venn diagram (meaning $|r_{X_1X_2}| = 1$) is not the definition of complete redundancy. Rather, X_2 is completely redundant with X_1 if adding X_2 to a model containing X_1 does not increase R at all. In the situation just described, this would be true. If the X_1 and X_2 circles overlapped completely, then no information about variability in Y (i.e., no increase in prediction accuracy) is acquired by adding X_2 to a model already containing X_1. That is, there is no part of X_2 that uniquely overlaps with Y. But panel B of Figure 7.2 also depicts redundancy. Even though X_1 and X_2 are not perfectly correlated (i.e., they don't completely overlap), adding X_2 to a model containing X_1 would not increase R, because X_2 shares nothing with Y that is unique to it relative to X_1.

In regression analysis, complete redundancy is evidenced by a variable receiving a regression weight of zero even though it is correlated with Y. For example, suppose you are trying to predict college GPA (Y) from high school GPA and a test of academic ability, such as the Scholastic Aptitude Test. Suppose the test of academic ability correlates positively with Y, but when high school GPA is controlled, that correlation drops to zero. In other words, in a group of people with the same high school GPA, there is no correlation between the test of academic ability and college GPA. Thus, the test is useless as a predictor of college GPA once high school GPA is considered. This means that R is the same with or without the test of academic ability in the model when high school GPA is in the model.

This is complete redundancy. The test of academic ability is completely redundant with high school GPA.

7.4.3 Independence

When all predictors are uncorrelated, a very simple relation holds:

$$R^2 = r^2_{YX_1} + r^2_{YX_2} + \cdots + r^2_{YX_k} \tag{7.3}$$

That is, R^2 is the sum of squared simple correlations between Y and the k predictors, or what we are calling their squared validities in this context. This situation is depicted in the Venn diagram in Figure 7.2, panel C. Equation 7.3 makes it useful to think of each squared correlation as a "proportion of variance explained" when predictors are independent. In this special case, proportions of variance simply add up, so that two variables accounting for 20 and 30% of the variance in Y will together account for 50% of the variance in Y. In Chapter 8 we will see that it is misleading to say that a variable accounting for 20% of the variance in Y is only two-thirds as "important" as one accounting for 30%. But provided they are not misinterpreted, the additive nature of proportions of variance provides a convenient way to talk about differences among variables.

7.4.4 Complementarity

In section 3.4.1 we defined *complementarity* as any situation in which the unique contribution of a set of regressors exceeds the sum of their individual unique contributions. In prediction, we distinguish between ordinary complementarity and a type of complementarity called *suppression*. In common usage in prediction contexts, only ordinary complementarity is even called complementarity, and we continue that usage here. In a prediction context such as that discussed in this chapter, we call two variables complementary if $R^2 > (r^2_{YX_1} + r^2_{YX_2})$ in the absence of the suppression effects described in section 7.4.5. As we will see in section 7.4.6, these conditions imply that $r_{X_1X_2} < 0$ even though neither predictor correlates negatively with Y.

Under independence, R^2 equals the sum of the squared validities, but under complementarity, R^2 may be far above that sum. A Venn diagram cannot represent this, and if this seems strange or impossible to you, recall the example from section 3.4.1, though that was not the kind of complementarity we are discussing here. For instance, imagine that success as a trial lawyer is determined primarily by scholarly ability and acting ability.

TABLE 7.3. A Data Set Illustrating Perfect Complementarity

X_1	X_2	Y
1	8	9
2	9	11
3	6	9
4	7	11
5	5	10
6	3	9
7	4	11
8	1	9
9	2	11

Imagine that the correlation between these two abilities is highly negative. Then either trait alone may be rather poor at predicting success as a trial lawyer, even though the two together could predict success excellently.

A numerical example illustrates this point. In Table 7.3 are scores on a variable X_1 that are simply the first nine integers. Values on X_2 are chosen so that they are highly negatively correlated with X_1 but not perfectly (here, $r_{X_1X_2} = -0.933$). Y is defined as $X_1 + X_2$, and thus we know Y can be predicted perfectly by this linear combination. Because X_1 and X_2 correlate so negatively, their sum cannot correlate highly with either; here, $r_{YX_1} = r_{YX_2} = 0.183$. But we know that $R = 1$ (verify this for yourself, if you desire). Thus, this example illustrates complete complementarity. In this case, $r^2_{YX_1} + r^2_{YX_2} = 0.067$, which is well below $R^2 = 1$.

7.4.5 Suppression

Imagine a multiple-choice history test (X_1) that is an excellent measure of knowledge of history (Y), except that the test is highly speeded so that people who can read faster tend to do better. Suppose we have a measure of the test takers' reading speed (X_2). Even though reading speed is positively correlated with performance (i.e., $r_{YX_2} > 0$), there is a surprising way that we could use it. If two people scored equally on the history test but person A scored higher than person B on reading speed, then we might make the following argument: Person B's low reading speed has disadvantaged B relative to A, yet B still scored as high as A on the test, so B probably knows more history than A. In other words, among people equal in performance

on the test, we'd guess that the lower the person's reading speed, the more knowledge that person has of history.

But this is equivalent to giving reading speed a negative regression weight even though it correlates positively with Y. For instance, if $\hat{Y} = 6 + 5X_1 - 2X_2$, then holding test score X_1 constant, the lower a person's reading speed X_2, the *higher* we estimate his or her history knowledge Y relative to someone who scored higher on X_2.

This illustrates a rule we can state more generally: If X_2 is a good measure of the sources of error in X_1 (e.g., reading speed), then by giving X_2 a negative weight, we may be able to predict Y very accurately even though neither X_1 nor X_2 alone does so. We are using X_2 to subtract out, correct for, or *suppress* the sources of error in X_1. In that case X_2 is called a *suppressor variable*.

In an extreme case, X_1 could have very large sources of error such as reading speed, but if X_2 measured those sources of error perfectly, we could completely correct for them and get perfect prediction, even though neither X_1 nor X_2 correlated highly with Y. Thus, we would have *complete suppression*. As a fictitious example, imagine a sample of adults, each of whom was the older child in a two-child family. Suppose a personality trait Y is completely determined by the person's age at which his or her sibling was born. Then Y will correlate little with the person's age and little with the age of the person's younger sibling, but it will correlate perfectly with the difference between the two.

You can imagine how users of a regression model for prediction might be reluctant to take advantage of suppression they may find, even though using suppressor variables would enhance accuracy of prediction. Few administrators of schools, government agencies, or corporations using regression models for personnel selection or related prediction tasks would survive public knowledge of the fact that they select people with the lowest scores on certain tests. But insights gained by the discovery of suppressors can be used to modify other measures in the prediction formula to reduce the influence of things such as reading speed, "test-wiseness," or other individual differences that generate sources of error in test scores.

7.4.6 How These Configurations Relate to the Correlation between Predictors

The seven configurations described in this section seem to represent a mind-boggling array of possibilities. But if we think about it from the perspective of pairs of variables X_1 and X_2 and there being two correlations r_{YX_1} and

r_{YX_2} that are fixed, then the seven configurations simply represent different values and ranges of $r_{X_1X_2}$. To explore this topic further, we must first consider the possible range of $r_{X_1X_2}$ for fixed values of r_{YX_1} and r_{YX_2}.

It should not surprise you to learn that if two variables are correlated 0.9 with each other, then it would be impossible for one to be correlated 0.9 with a third variable but for the other to be uncorrelated with that third variable. The material in this section requires understanding which combinations of correlations are possible and which are impossible. Therefore, for the case of two predictor variables, we now develop a formula showing the possible range of $r_{X_1X_2}$ as a function of r_{YX_1} and r_{YX_2}.

We know that any correlation—simple, partial, or semipartial—cannot fall above 1 or below −1. Define $r_{12.Y}$ as the partial correlation between X_1 and X_2 when controlling for Y (we've never discussed a partial correlation between predictors holding Y constant, but that doesn't mean such a correlation can't or doesn't exist) is

$$r_{12.Y} = \frac{r_{X_1X_2} - r_{YX_1}r_{YX_2}}{\sqrt{1 - r_{YX_1}^2}\sqrt{1 - r_{YX_2}^2}}$$

A close look at this formula shows that if r_{YX_1} and r_{YX_2} are fixed, $r_{12.Y}$ increases as $r_{X_1X_2}$ increases. Given that no correlation can be larger than 1 or smaller than −1, we can find the range of possible values of $r_{X_1X_2}$ by setting $r_{12.Y}$ to 1 and −1 and solving for $r_{X_1X_2}$. When doing so, we find that the limits on $r_{X_1X_2}$ are

$$r_{YX_1}r_{YX_2} \pm \sqrt{(1 - r_{YX_1}^2)(1 - r_{YX_2}^2)}$$

For instance, if $r_{YX_1} = 0.5$ and $r_{YX_2} = 0.3$, then $-0.676 < r_{X_1X_2} < 0.976$. So under these constraints on r_{YX_1} and r_{YX_2}, $r_{X_1X_2}$ can be almost but not perfectly correlated positively, but $r_{X_1X_2}$ cannot be as negative as it can be positive.

Having now derived the possible values of $r_{X_1X_2}$ given their correlations with Y, we can now discuss how these configurations relate to the size of $r_{X_1X_2}$. This discussion requires recognition that it is always possible to reflect a predictor variable (i.e., reverse its scoring direction, making low values high and high values low) so that it has a positive or zero correlation with Y. We assume this has been done for both X_1 and X_2. But this still allows $r_{X_1X_2}$ to be negative.

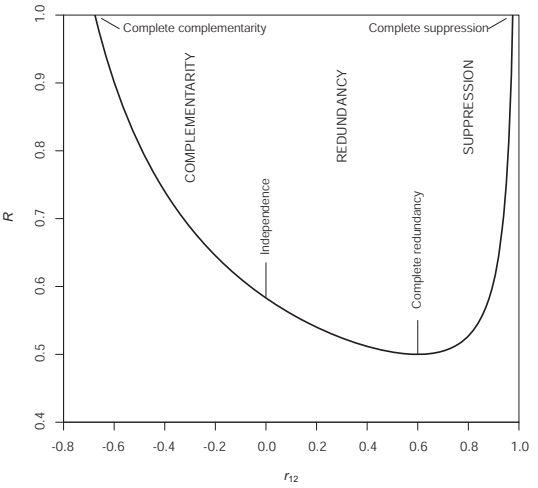


FIGURE 7.3. Complementarity, redundancy, and suppression as a function of $r_{X_1X_2}$ when $r_{YX_1} = 0.5$ and $r_{YX_2} = 0.3$.

Figure 7.3 shows how R relates to $r_{X_1X_2}$ when r_{YX_1} and r_{YX_2} are fixed at 0.5 and 0.3, respectively. In section 3.4.5 we saw that

$$R^2 = r^2_{YX_1} + sr^2_2 = r^2_{YX_1} + \frac{(r_{YX_2} - r_{YX_1}r_{X_1X_2})^2}{1 - r^2_{X_1X_2}}$$

Figure 7.3 was constructed using this formula to compute values of R for various values of $r_{X_1X_2}$ when $r_{YX_1} = 0.5$ and $r_{YX_2} = 0.3$. We shall use this figure to show how $r_{X_1X_2}$ relates to the seven configurations.

Two of the seven configurations have already been related to $r_{X_1X_2}$; we said that *independence* implies $r_{X_1X_2} = 0$ and *complementarity* implies $r_{X_1X_2} < 0$. These configurations have been marked in Figure 7.3.

We can define *suppression* as the case in which either b_1 or b_2 is negative even though r_{YX_1} and r_{YX_2} are both non-negative. We shall consider the case in which b_2 is negative and then derive the result for b_1 by analogy. We have

$$b_2 = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \times \frac{s_Y}{s_{X_2}} \tag{7.4}$$

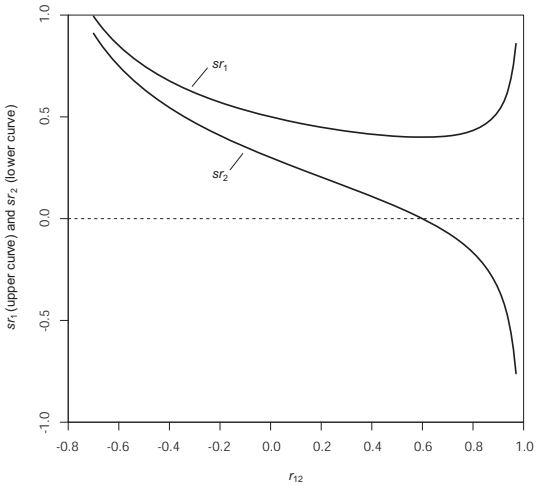


FIGURE 7.4. sr_1 and sr_2 as a function of $r_{X_1X_2}$ when $r_{YX_1} = 0.5$ and $r_{YX_2} = 0.3$.

But s_Y, s_{X_2}, and $1 - r^2_{X_1X_2}$ are all non-negative, so b_2 is negative whenever $r_{YX_2} - r_{YX_1}r_{X_1X_2}$ is negative. This, in turn, occurs if and only if $r_{X_1X_2} > (r_{YX_2}/r_{YX_1})$. A symmetrical argument shows that b_1 can be negative only if $r_{X_1X_2} > (r_{YX_1}/r_{YX_2})$. We have assumed that r_{YX_1} and r_{YX_2} are non-negative, so we conclude that suppression occurs only if $r_{X_1X_2}$ is positive and above the lower of r_{YX_1}/r_{YX_2} and r_{YX_2}/r_{YX_1}. In this example, these two values are 1.67 and 0.6, so suppression occurs if $r_{X_1X_2} > 0.6$. Figure 7.4 shows the values of sr_1 and sr_2 for each value of $r_{X_1X_2}$; notice that sr_2 is negative in the suppression region.

Complete redundancy occurs for X_2 if its optimum weight is zero. At the point of complete redundancy, R reaches its lowest possible value for fixed values of r_{YX_1} and r_{YX_2}; it equals the higher of these two validities. We see from equation 7.4 for b_2 that $b_2 = 0$ if $r_{YX_2} - r_{YX_1}r_{X_1X_2} = 0$, which implies $r_{X_1X_2} = r_{YX_2}/r_{YX_1}$. In this example, that means $r_{X_1X_2} = 0.3/0.5 = 0.6$. This value is also shown in Figure 7.3. Lower positive values of $r_{X_1X_2}$ yield the standard configuration or *partial redundancy*.

As $r_{X_1X_2}$ becomes more negative with r_{YX_1} and r_{YX_2} fixed, the predictors become ever more complementary, until at the most negative possible values of $r_{X_1X_2}$ *complete complementarity* is achieved (at the far left in Figure

7.3). At the right, as $r_{X_1X_2}$ increases, suppression becomes ever larger, until at the maximum possible value of $r_{X_1X_2}$ we have *complete suppression*.

Thus, in summary, as $r_{X_1X_2}$ rises from its smallest to its largest possible value, we move through the seven configurations of complete and partial complementarity, independence, partial and complete redundancy, and partial and complete suppression.

7.4.7 Configurations of Three or More Predictors

When there are three or more predictors, we can define *suppression* as any case in which a variable receives a significant negative weight when it has a positive or zero correlation with Y. *Complementarity* can be defined as any significant negative correlation between predictors when both have a positive or zero correlation with Y. As before, complete complementarity or suppression occurs if $R = 1$ when no simple correlation involving Y is 1, and complete redundancy occurs if some regression weight is exactly zero. Independence occurs if the correlation between all predictors is zero. The standard configuration of *partial redundancy* can be defined as the absence of the other six configurations. We may be able to predict perfectly well without even knowing whether nonstandard configurations exist in the data. But understanding these configurations can help explain unexpected findings such as negative regression weights for variables positively correlated with Y or surprisingly high values of R.

7.5 Revisiting the Value of Human Judgment

At the beginning of this chapter we said that predictions made solely by expert human judgment are quite consistently less accurate than those made by mechanical methods such as regression because of the unreliability in the judgment process. That is, human judges don't always apply the same decision rule when using the available information. In one case, a variable might be given no weight, whereas in another it might be given a lot of weight. Experts might say that the complexity of human behavior requires flexibility in how information is weighted, depending on the circumstance. But that is generally the source of the problem, not the solution to it.

One way to remove the unreliablity is to ask expert judges not to make decisions in every specific case, but rather to assign weights to the various predictor variables using whatever system their expertise or intuition dictates and using that weighting system for all cases. Dawes (1979, 1988) and others have argued that predictions made by such a system are often at least

as accurate as those made by regression, especially if the sample available for the regression is small. And this method is also simpler, in that it does not require a sample of cases to develop a mechanical prediction system in the first place.

A choice between these two methods cannot be made simply by seeing which method has usually worked better on other problems in the past, because the relative merits of the two approaches obviously depend on factors such as the sample size available for the regression and the degree to which relationships among the variables studied are genuinely understood by the expert judges. But we know on theoretical grounds that a pure human judgment approach has the same problem as forward and mixed stepwise regression—there is no upper limit to the predictive power that may be overlooked by ignoring available data. Therefore, what is needed is a way to combine the advantages of regression and human judgment in such a way as to overcome the sampling errors in regression if those errors are large and human judgments are accurate, but to use the regression results if they turn out to add predictive power to inaccurate human judgments. There is a way to do this that we discuss below. Darlington (1978) discusses some others.

The first step is to make a composite variable for each case being predicted using the relative weights for each variable an expert (alone, by committee, or whatever system) judges to be optimum. Call this X_{SC} for a "subjective composite." This should be created before examination of any of the actual regression weights that a regression procedure generates. X_{SC} may be a composite of all k predictors, or it may exclude some of them.

If you are thinking the next step would be to include X_{SC} in the regression analysis of Y with the k predictors to see what kind of weight it gets, your thinking fails to acknowledge a problem with this. Because X_{SC} is a linear combination of the k predictors, including X_{SC} in a regression model along with the k predictors would produce a singularity. A singularity occurs when one predictor is a perfect linear combination of the other predictors in the model. But that is exactly what X_{SC} is. Therefore, one of the other predictors must be removed before X_{SC} can be added. Furthermore, the variable removed must be one of those with a nonzero weight in X_{SC}, or else the singularity will remain after the variable is removed.

So the next step is to momentarily ignore X_{SC} and estimate Y from all k predictors. From the variables with nonzero weights in X_{SC}, remove the one with the t closest to zero and replace it with X_{SC}. As you already know, the one removed is the one whose contribution to the regression

prediction is least statistically significant. It is also the one whose removal will lower R the least. Since this variable is the one least useful even in the absence of X_{SC}, it seems extremely likely that it will not be missed in any regression that includes X_{SC}, which itself includes the deleted predictor. After making this replacement, apply any of the methods of this chapter, such as backward stepwise regression, to this model.

If the subjective guesses used in the construction of X_{SC} were accurate, it is very likely that the regression finally selected by this process is the regression that includes only X_{SC}. Thus, this process enjoys the advantages of subjective judgment—the avoidance of unnecessary reliance on sample data that may be unreliable—while still allowing the sample data to play a role if the subjective judgments turn out to be inaccurate.

7.6 Chapter Summary

Regression has a practical, applied role to play in prediction problems. With a sample of cases measured on a set of predictor variables, as well as some criterion variable of interest, regression analysis can be used to derive a regression equation that produces an optimally weighted sum of the predictors that correlates maximally with the criterion. The regression formula can then be applied to cases not originally used to derive the equation in the first place, in order to generate a prediction of Y for those cases, knowing only their scores on the predictor variables. Using a regression equation in this fashion gets around some of the inaccuracy created when human judges attempt to make predictions relying only their expertise or intuition. We still need human beings to decide what predictor variables to consider as candidates in the regression equation, but once that is done, research suggests it is best to remove humans from the process of generating the prediction itself.

Regression analysis is so good at its job of constructing the regression weights that it tends to "overfit" the data, modeling not only the systematic processes linking predictors to criterion but also all the idiosyncracies of the sample itself. As a result, the correlation between what the model estimates for Y and the actual model of Y when applied to cases not used to generate the regression model—*shrunken R*—tends to be smaller than R. Shrunken R can be estimated in a variety of ways, such as cross-validation, leave-one-out methods, and various analytical formulas.

The reduction in the predictive accuracy of a sample-derived regression model when applied to new data, what is called *validity shrinkage*, can be

managed in part by a number of variable selection methods. Methods such as forward and backward stepwise regression and all subsets regression attempt to maximize the correlation between $\hat{Y}$ and Y, while minimizing the number of predictors. The fewer predictors that are used in a regression model applied to future cases, the less validity shrinkage tends to be. These variable selection methods have documented problems, however, in that they too tend to overfit the data. But they can be useful in some contexts if their limits are understood and it is recognized that none of them is likely to be selecting the best or correct model by some objective standard.

Whether a variable receives a zero or a nonzero weight in a prediction formula will depend on more than just its correlation with Y. There are many configurations of intercorrelation between predictors and between predictors and criterion—redundancy, independence, complementary, and suppression—that give rise to some counterintuitive phenemona, such as when a regressor is uncorrelated with Y yet receives a nonzero weight in the regression equation, or when R is very large when all the correlation between Y and the predictors in the model are small.

It is possible to combine subjective human judgment and mechanical prediction methods. This can be done by constructing a subjective composite of predictor variables using human judgment or intuition, and using it as a predictor in variable selection problem, after first acknowledging the singularity this will produce. Such a method may help to determine whether human judgment produces the best prediction, should be ignored, or can be supplemented by information derived by a mechanical prediction system such as regression analysis.

8
Assessing the Importance of Regressors

In this chapter we address the topic of assessing importance, either in absolute or relative terms, of a regressor in a regression model. Although the impulse to label one regressor's relationship with Y as important and another's as less so is very strong, how to quantify the importance of a simple or partial relationship is a controversial topic in statistics. Although researchers often use standardized regression coefficients or squared measures of association as indices of a variable's importance, we provide some arguments against doing so. In this chapter we offer some recommendations, but mostly opinions, about how to think about, define, and measure the importance of a regressor in a regression analysis.

We have seen that there is little ambiguity concerning the interpretation of the partial regression coefficient b_j in a regression analysis. Given the single assumption of linearity, ${}_T b_j$ is the average difference in Y associated with a 1-unit difference in X_j when other regressors are held constant. The sample regression coefficient b_j estimates this effect. But b_j cannot be used to compare the importance of regressors because it is a scale-bound metric. If X_j is measured in inches, then changing its metric to feet will multiply b_j by 12 but will not change its importance in absolute terms or relative to other regressors in the model. If our goal is to assess the importance of a regressor, the distance between b_j and zero will not accomplish that goal, at least not without thinking about the scale of measurement, nor will it allow us to compare the importance of regressors in a model. A scale-free measure is needed.

There are three basic scale-free measures: the standardized regression coefficient $\tilde{b}_j$, the partial correlation pr_j, and the semipartial correlation sr_j. The latter two are often squared; as mentioned in section 3.4.1, sr_j^2 is the

proportion of the variance in Y uniquely explained by X_j, and pr_j^2 is the same value expressed as a proportion of the variance in Y unexplained by the other regressors in the model.

Thus, we have at least four questions: Which, if any, of the three basic scale-free measures should be used as a measure of a variable's importance? Should they be squared? Are there better measures of importance? And what do we mean by *importance* anyway? There is no consensus among scientists and statisticians on the answers to these questions. In this chapter, we offer some of our opinions on the topic of the importance of regressors, acknowledging that importance cannot usually be distilled down to just one number. Kelley and Preacher (2012) offer a good discussion of variable importance under the label of "effect size," pointing out the many confusing definitions that exist and how it is nearly impossible to settle on a single measure of importance.

8.1 What Does It Mean for a Variable to Be Important?

Null hypothesis significance testing is the framework most researchers use for establishing whether or not an effect in some study exists. Although it has its many critics, and alternatives such as Bayesian methods exist, it is likely to remain dominant into the foreseeable future. Typically, the null hypothesis tested is that there is no effect, meaning two variables are not associated, or two groups don't differ from each other. A "statistically significant" relationship allows for a claim of the existence of an effect, but it says nothing about the size of its effect or its practical or theoretical importance. Regression analysis can be used to generate an estimate of the size of some regressor's effect and test hypotheses about whether a relationship exists between a regressor and a dependent variable. But whether that effect is large or small, important or trivial, is largely in the eye of the beholder. These judgments are highly context-dependent and will vary between areas of inquiry and even within an area, depending on who is making the judgment.

8.1.1 Variable Importance in Substantive or Applied Terms

Researchers probably could not approach their jobs day in and day out if they felt that their work wasn't important. Everyone wants to feel like the work he or she does matters in some way. But what does it mean to

say that one's work is important? And if one decides that one's work is important at least in principle or on the face of it, who is to say that the research one has done has revealed important findings? From a funding perspective, which research is worth funding, and which is not? We can all agree that taxpayer dollars probably should not be spent on things that aren't important. But who is to make that judgment? And how do we know if a researcher's findings, if funded, will be important? Basic research is often undertaken without obvious application, and governments regularly fund basic research. The importance of research is sometimes determined only later, once someone finds an application for it.

It is probably obvious to you that statistics can have little to say about judgments of importance of this variety, whether those judgments are about entire fields of inquiry or specific findings, or whether a line of work is worth funding. We start this chapter by stating the obvious, because we think it is important to keep in mind that when we use the term *important*, we are not using it with these kinds of questions in mind.

8.1.2 Variable Importance in Statistical Terms

Statistics can and does offer a contribution to science with respect to how researchers can think about and quantify the importance of a variable in abstract quantitative terms that is free of value judgment. For instance, the importance of a variable could be indexed with respect to how much of the variability in the dependent variable Y it explains since, after all, accounting for variation between Y and $\overline{Y}$ is essentially what regression algebra does. The more important variables in a model explain more of the variability in Y. A variable could be deemed especially important if it explains lots of the variability, even in the presence of other variables in the model. But complicating this way of conceptualizing importance is the fact that variability can be quantified in more than one way, such as by squaring deviations from the mean or using the standard deviation.

Importance could also be quantified as the amount that including a regressor in a model lowers the error in estimation of Y. The most important variables in a regression would be those that lower the error in estimation the most relative to when those variables are not in the model. If your goal is to estimate Y accurately and you care how much in error your estimates of Y tend to be, then using a measure of importance that is sensitive to this would be desired.

Importance could also be measured by the amount Y changes as a result of changing the regressor a certain amount. Regressors that "move"

Y more are more important than those that move Y little. So a variable that, if changed by a certain amount, results in a larger change in Y than some other regressor when changed by a comparable amount, then that first variable could be deemed more important.

Such approaches are related mathematically, and all may be useful in one way or another, but perhaps more so in some contexts than others. There is no best way of defining importance in a purely numerical sense that is independent of context, but there are some ways that can be avoided and others that may be more useful across contexts. We make a few recommendations later in the chapter.

If you are new to the business of science, it won't be long before you encounter various rules of thumb circulating about how to label a variable's effect as "small," "moderate," or "large" based on some metric. These rules take a variety of forms depending on the metric being used (e.g., a standardized mean difference or proportion of variance explained), and not all rules of thumb are consistent. We don't see much value in these rules of thumb and don't use them ourselves. If someone tells you that a correlation of 0.3 is too small an effect to concern yourself with, or you should care only about large effects that explain at least 10% of the variance in Y, take that advice or criticism with a grain of salt.

8.2 Should Correlations Be Squared?

We have discussed in various places already (e.g., sections 2.4.2, 3.4.1, and 4.2.2) that r^2_{XY} can be interpreted as the proportion of the variance in Y explained by X. If X is a set of regressors rather than just one, then R^2 for the model estimating Y from this set is interpreted as the proportion of variance in Y explained by the model or the set of regressors. If an investigator reported that the correlation (simple or multiple) in his or her study between X and Y is 0.1 but statistically different from zero, no doubt a critic would bring up that X explains a mere 1% of the variance in Y, and this is hardly news worth disseminating.

Similarly, a researcher advocating theory B over theory C might show that when theory B's regressors are added to a model of Y that already contains theory C's regressors, the squared multiple correlation increases more than can be explained by chance, and this supports the relevance of theory B in explaining individual differences in Y. We saw in sections 3.4.1 and 5.3.1 that when a variable or variable set is added to a model, that variable's or set of variables' squared semipartial correlation with Y is

interpreted as the proportion of the variance in Y uniquely explained by that variable or variable set. It is the change in R^2 that results when those variables are added to the model, and a researcher may expect criticism if it seems too small to observers and consumers of that research.

It is almost taken as gospel that the square of a correlation is a measure of the importance of the relationship between two variables. Researchers routinely square r_{XY} to derive the proportion of variance in Y explained by X, they report R^2 in their regression tables in research articles, and they talk about improvement in the fit of a model in terms of the amount R^2 increases when a variable or set of variables is added to a model. Thinking of relationships in terms of proportions of variance explained is extremely entrenched among scientists and statisticians. But doing so can be very misleading, and result in understated claims and perhaps unwarranted criticisms about the role a variable plays in explaining Y. In this section we document how so. We are not the first to put this perspective in writing. See, for example, Cronbach and Gleser (1965), D'Andrade and Dart (1990), and Ozer (1985) for related discussions.

8.2.1 Decision Theory

In decision theory, the importance of the relationship between two variables is often defined as proportional to the *expected gain* from taking advantage of the relationship relative to ignoring it. For instance, if we were to use a psychological test to decide which of two kinds of therapy to administer to patients in a psychiatric hospital, then the expected gain from using that test to make the decision might be measured in the number of days by which the average patient's hospital stay is shortened or lengthened as a result. The test's value or importance would be defined as proportional to the number of days the average hospital stay is shortened by use of the test. So if using the test to choose method A or B cuts the number of days in the hospital in half relative to when the test is not used, that would be twice as valuable as if using the test cut the stay of the average patient's hospitalization by only 25%. Of course, other metrics could be used, such as dollars saved by the patient or the hospital. It makes no difference, though using dollars generalizes the point we make in this section to other areas such as business and marketing. Regardless, thinking about importance in these terms yields a ratio scale of importance. If using the relationship between X and Y when making a decision cuts costs from \$2,000 to \$1,000 dollars (or from 20 days to 10 days), this is twice as important as if it cuts costs from \$2,000 to only \$1,500 (or from 20 days to 15 days).

In this section we show how a decision theory approach yields no single relationship between a correlation and importance. But if we had to pick the most common single relationship, it would be most accurate to say that the importance of a correlation is proportional to r rather than r^2. This point is widely accepted among psychometricians; see, for instance, Cronbach and Gleser (1965). But it seems this is hardly known outside of this circle (except perhaps among decision theorists), so we shall develop it in some detail with a few examples.

As a first example, suppose 19 people apply for a small number of positions at your company, and you give them each a test known to be correlated with ability to do the job for which they are applying. With only seven spaces available, naturally you select the seven who score highest, and the 12 lower scorers are rejected. In real life, of course, we would not know their actual ability to do the job until they were hired. But let's imagine that for these 19 people, their actual abilities are known. Let Figure 8.1 represent a scatterplot depicting the observed test scores of these 19 people (X) against their actual ability to do the job (Y).

In these data, the correlation between test score and ability is $r_{XY} = 0.50$. Furthermore, the mean of the actual ability of these 19 people is $\overline{Y} = 3.0$. So if you didn't use the test at all and just randomly selected seven of these 19 to hire, then the expected ability of the seven you hire is 3.0. On the other hand, if your test were correlated $r = 1.0$ with actual ability rather than only 0.5, then your test would lead you to select the seven people with the highest ability, meaning scores of 4 or 5 on Y. These are the seven cases in Figure 8.1 at the top of the plot. The mean of their actual ability is $(4 + 4 + 4 + 4 + 5 + 5 + 5)/7 = 31/7$. Thus, the use of this hypothetical perfect test raises the expected mean of the ability of the seven selected from 3, if the test were not used at all, to $31/7 = 4.429$. This is a gain of $10/7$, or 1.429.

However, you aren't using this hypothetical perfect test but, rather, a test that is not perfectly correlated with ability. You select the seven highest scorers on this imperfect test. These are the seven people scoring $X = 4$ or $X = 5$ on the test, found in Figure 8.1 at the right end of the scatterplot. The mean of the *actual* ability (Y) of these seven is $(2 + 3 + 4 + 5 + 3 + 4 + 5)/7 = 26/7 = 3.714$. This is a gain of $5/7$ or 0.714 in the ability of those selected relative to if no test were used. This gain of $5/7$ is exactly half of the $10/7$ gain that would result if you used the hypothetical perfect test. But r_{XY} is 0.5, which is exactly one half of 1.0. So in this example, the value of the test is proportional to r, not to r^2.

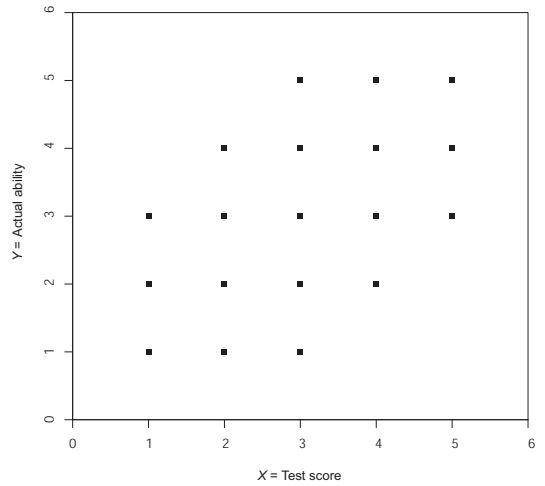


FIGURE 8.1. A scatterplot of observed test score of ability against actual ability.

Consider a second example. Suppose you are taking two different courses, A and B. In each course, a study has been conducted examining the relationship between course grade (Y) and hours of study in that course (X) as recorded in student diaries. Both studies have adequately controlled for various confounds such as a student's ability and previous knowledge of course material. In course A, $r_{XY} = 0.3$ and in course B, $r_{XY} = 0.6$. Further suppose that the standard deviations of study time (s_X) are the same in the two courses, as are the standard deviations of grade (s_Y). We might want to know whether studying is more important in one course than another by examining the sizes of the regression weight for study time in a linear model estimating Y from X in each of the courses. This regression weight tells us the expected gain in course grade resulting from one additional hour of study time.

From section 2.2.3, these regression weights are equal to $r_{XY}(s_Y/s_X)$, but given that the standard deviations of X and Y are the same across the two courses, then the ratio of their standard deviations is the same. That means that the ratio of the two regression weights is the same as the ratio of the two correlations between X and Y. That is, given that r_{XY} is twice as large in course B than in course A, so too is the regression weight relating Y to

X in course B than in course A. That means that 1 hour of additional study time has twice the impact on grade in course B as in course A. If your objective is to maximize your course grade, you'd be better off studying the additional hour in course B than course A, because the expected gain in grade is double relative to studying more in course A. But the important point here is that the ratio of expected gains between the two courses is the ratio of the two correlations, not the ratio of the two squared correlations, which is 4 (i.e., $0.6^2/0.3^2 = 4$). So in this example, a correlation twice as high equates to twice the importance, not four times.

If you are gambler, this third example will appeal to you. You flip a nickel (worth 5 cents) and a dime (worth 10 cents) simultaneously and repeatedly. On each trial, if the dime comes up heads, you win 10 cents, tails you win nothing. Similarly, if the nickel comes up heads, you win 5 cents, tails you win nothing. So on each trial, you have a 25% chance of winning 15 cents (heads on both), a 25% chance of winning 10 cents (heads on dime not on nickel), a 25% chance of winning 5 cents (heads on nickel not on dime), and a 25% chance of winning nothing (heads on neither). Call your winnings on each set of flips W.

You know that in a long sequence of trials, the results of the nickel will be independent of the results from the dime because the outcome of the flip of the dime has no influence on the outcome of the flip of the nickel. And since the winnings you receive over a long sequence of trials is determined entirely by the outcome of these two coins, the percentage of your winnings attributable to the nickel and to the dime must sum to 100%. It can be shown that the dime accounts for 80% of the variance in your total winnings, while the nickel accounts for only 20%. In correlational terms, imagine a dummy variable D coded 1 for the dime if it comes up heads and 0 if it comes up tails and a similar dummy variable N for the nickel. In a long series of flips, you'd find that $r^2_{DW} = 0.8$ and $r^2_{NW} = 0.2$.

This makes it seem like the dime is four times as important as the nickel in determining your winnings. But in ordinary language and thinking, we would say that the dime affects your winnings twice as much not four times as much, because the dime is worth twice as much as the nickel when it comes up heads. Importantly, notice that the unsquared correlations between W and each of the two dummy variables are $r_{DW} = 0.894$ and $r_{NW} = 0.447$, which is a 2 to 1 ratio, same as the value of a dime relative to a nickel.

TABLE 8.1. Five-Year Mortality Rate

	Alive 5 years after diagnosis?		
	No ($Y = 0$)	Yes ($Y = 1$)	Total
Received drug ($X = 1$)	90 45%	110 55%	200
Received placebo ($X = 0$)	110 55%	90 45%	200

8.2.2 Small Squared Correlations Can Reflect Noteworthy Effects

The previous examples show that sometimes the size of an effect is better represented by an unsquared correlation than a squared correlation. We now address a slightly different matter, and that is whether small squared correlations are necessarily small or unimpressive effects.

Suppose you were suffering from a serious illness, and mortality statistics show that for every 100 people diagnosed with this illness, 55 die within 5 years of diagnosis. In other words, 45%, or less than half, are still alive after 5 years. This sounds like a fairly depressing prognosis. But suppose a randomized clinical trial for a new drug shows that this drug reduces the death rate within 5 years after diagnosis to 45 out of 100. Rephrased, of those who take the drug, 55% are still alive after 5 years. This may still sound depressing, but at least the drug offers some hope.

These statistics are reflected in Table 8.1 for 400 people, half of whom were given the drug and half of whom were given a placebo. Let X be a variable coded 0 for those who did not take the drug and 1 for those who did, and let Y be a variable coded 1 if still alive in 5 years and 0 if dead in 5 years. In this table, r_{XY} is 0.10, meaning that the drug accounts for only $r^2_{XY} = .01$ or 1% of the variance in death within 5 years. Though statistically significant, in variance-explained terms, this seems like a pretty small effect. Even if you were a researcher who advocates squared correlations as measures of effect size, you would probably still take the drug if offered. What would you have to lose, after all? But these statistics don't seem to offer you much hope of prolonging your life with the drug. If the side effects were more than minor, you might even contemplate not taking it.

But if thought of in terms of lives saved with the drug, the effect of this drug is actually fairly large. Given the findings from the clinical trial, which necessarily involves random assignment to condition (drug or no drug), we can infer that if the 200 in the bottom row of Table 8.1 were given the drug, then 20 of those 110 who would have otherwise died with 5 years would still be alive. In other words, we would expect the drug to spare the lives of 20 of 110 people, or about 18% of those would have died within 5 years by not taking the drug. Yet the drug accounts for only 1% of the variance in mortality rate in 5 years. This tiny effect in variance-explained terms is actually a fairly substantial effect when thought of in terms of percentage of lives saved.

Rosenthal and Rubin (1982) provide a more thorough discussion of effect size for problems similar to this one, labeling r_{XY} the *binomial effect size display*, or BESD. In this example, BESD = 0.10, which corresponds to a difference of 10% between the "success rates" (i.e., being alive after 5 years) in the two groups. They offer a generalization of the measure to continuous variables, and argue that at least for some problems like this, squaring r can make large effects seem considerably less impressive. Their brief discussion is worth reading. Also see Abelson (1985) for a related discussion in the context of sports.

8.2.3 Pearson's r as the Ratio of a Regression Coefficient to Its Maximum Possible Value

When talking about the size of an effect or the importance of a variable, it is natural to couch the discussion relative to zero; that is, no effect or "zero importance." But the meaningfulness of the distance between some effect or measure and zero will depend on how large that effect or measure can get. To say you walked half the distance from your house to work means something different if your house is 1 mile away from work relative to whether it is 20 miles away. This is one reason why r_{XY} is generally more useful as a measure of association than the covariance between X and Y. The correlation is scaled to be between −1 and 1, whereas the covariance's upper bound is always scaled by the metrics of measurement of X and Y. A "large" covariance could be either large or small, depending on the variance of the variables.

The same is true for regression coefficients. Although a regression coefficient of zero clearly is meaningful, anything other than zero needs to be interpreted relative to how large it *could* be hypothetically or theoretically. For instance, if each hour of foreign language vocabulary training expands

a student's working vocabulary by 10 words, this could be a large or small effect depending on how you think about it. We might be able to imagine some superefficient training method that adds 1,000, or even 10,000 words for each hour, so for all we know, the actual training method might be only one-hundredth or one-thousandth as effective as some other method that maybe does or could exist. So how large a measure of association is relative to the maximum it could be needs to be considered when interpreting an effect's importance. Sometimes this is impossible, but not always.

It is possible to calculate the maximum possible association between two variables if we think of certain quantities as fixed. For instance, in a certain sample, suppose we have calculated that the standard deviation of annual income is \$20,000 and that the standard deviation of years of education is 4 years. Now suppose we regress income on education and find $b_1 = 2{,}000$. So each year of education translates into an additional \$2,000 in income. Is this a big effect or a small effect?

To answer this question, at least in a statistical sense, suppose the correlation between these two variables was perfect (i.e., $r_{XY} = 1$) and we regressed income (Y) on education (X). The simple regression formula

$$b_1 = r_{XY}\frac{s_Y}{s_X}$$

tells us that the regression coefficient when estimating income from education would be $1 \times 20{,}000/4 = 5{,}000$. Because r_{XY} cannot exceed 1, that is the highest the regression coefficient could be given these standard deviations of income and education. If we think of those standard deviations as fixed, then $b_1 = 5{,}000 \times r_{XY}$ and b_1 is proportional to r_{XY}. Therefore, r_{XY} can be interpreted as the ratio of b_1 to its maximum possible value given the standard deviations of Y and X. So if the correlation between income and education were actually $r_{XY} = 0.4$ rather than 1, that means $b_1 = 2{,}000$. In other words, the regression coefficient for education is 0.4 times or 40% of its maximum possible value of $5{,}000$ given the two standard deviations.

A similar but more complex rule applies to multiple regression. Suppose we think of all statistics in the sample as fixed except statistics measuring the partial association between Y and a particular regressor X_j. These conditions fix a maximum possible value of b_j, and pr_j equals the ratio between the actual value of b_j and that theoretical maximum. This is a useful interpretation of pr_j which does not require the limiting conditions mentioned in section 4.5.

This point can be expressed in a formula. We regress Y on k regressors, one of which, X_j, we think of as the independent variable and the rest are covariates. Let C (for "covariates") denote the set of $k-1$ regressors. Let $V(Y.C)$ be the mean of the squared residuals when Y is regressed on the $k-1$ covariates. You can think of $V(Y.C)$ as variance in Y not explained by the covariates. It can be shown that

$$pr_j = b_j s_{X_j} \sqrt{\frac{Tol_j}{V(Y.C)}}$$

This formula shows that if we think of s_{X_j}, Tol_j (the proportion of the variance in X_j not explained by the covariates) and $V(Y.C)$ as fixed, then given that pr_j cannot exceed 1, pr_j equals the ratio between b_j and its maximum possible value with the same values of those three statistics.

So the maximum size that b_j can possibly be can be as useful a reference against which to compare b_j as is zero. If a language training program increases vocabulary by 10 words for every hour (i.e., $b_1 = 10$), and the correlation between hours of training and vocabulary is 0.67, then you can say that given the standard deviations of hours and vocabulary observed in the data, the largest b_1 could be in a regression estimating vocabulary from hours of training is 15. This means that the observed effect is two-thirds of its maximum possible value given the available information. If you had controlled for a set of covariates in a multiple regression and pr_j was 0.5, then the observed effect of an hour of training is one-half of its maximum possible value given the observed variability in hours or training and how much of the observed variance in hours of training and vocabulary is explained by the covariates.

8.2.4 Proportional Reduction in Estimation Error

Now that you may be starting to believe that r is a more sensible measure of importance than r^2, we provide an example illustrating the opposite. Suppose that your goal is to estimate a future college student's GPA (Y). You have two measures available to you: performance on the Scholastic Aptitude Test (SAT, which we denote X), and high school GPA (W). Suppose that research shows that $r_{XY} = 0.3$ and $r_{WY} = 0.6$. There would be no argument that high school GPA is a better predictor in this example, though we could debate how to quantify just how much better. Using the squared correlation as the metric, GPA is four times as important as SAT, because it explains four times more variance in college GPA. But the

prior argument suggests that the ratio of the unsquared correlations better reflects their relative importance. By that measure, SAT is only two times more important.

One way of measuring the quality of a prediction system is how large the errors in estimation tend to be. The standard error of estimate first introduced in section 4.2.4 is widely used as a measure of this. In large samples, this is very close to the standard deviation of the residuals, and it is estimated as $s_{Y.X} = \sqrt{MS_{residual}}$. Because the least squares criterion minimizes $MS_{residual}$, it follows that it also minimizes $s_{Y.X}$. The smaller $s_{Y.X}$, the "better" the model, in the sense that the model generates estimates of Y that are closer to Y than some other model of the same Y with a bigger $s_{Y.X}$.

In a model with a single predictor X of Y, the standard error of estimate is related to r_{XY} by the formula

$$s_{Y.X} = s_Y \sqrt{(1 - r_{XY}^2)} \tag{8.1}$$

Now suppose that you wanted to guess the college GPA of a set of applicants, but you had no information available about their high school GPA or SAT scores. In that case, your best guess for every applicant would be that their college GPAs will be average. That is, your model would $\hat{Y} = \overline{Y}$. This is equivalent to using a predictor with no correlation with Y, and so, from equation 8.1, $s_{Y.X} = s_Y$ for this model. That is, the standard error of estimate is just the standard deviation of Y.

How much would this error in estimation be reduced by using information about the students' SAT scores? Earlier, you were told that $r_{XY} = 0.3$, so the standard error of estimate would be $s_Y \sqrt{1 - 0.3^2} = 0.954 s_Y$. That is, the standard error of estimate is 95.4% of the size of the standard deviation of Y. This could be expressed in a different way as the proportional reduction in the standard error of estimate that results when using the relationship between X and Y to estimate Y, known as the *coefficient of forecasting efficiency*:

$$E = 1 - \sqrt{1 - r_{XY}^2} \tag{8.2}$$

(The fact that we use E to denote expected value in earlier chapters does not complicate notation since we do not mention the coefficient of forecasting efficiency after this section.) E ranges between 0 and 1, with a number closer to 1 reflecting a more "important" relationship in a statistical sense. In this example, $E = 1 - \sqrt{1 - 0.3^2} = 0.046$, meaning that the error in estimation

is reduced by 4.6% by using SAT scores relative to when not. Observe that this number is even smaller than the squared multiple correlation. If explaining only 9% of the variance in college GPA is unimpressive to you, you'd be even more unimpressed with reducing the error in estimation by only 4.6%.

How does high school GPA fare by this standard? Earlier you were told that $r_{XW} = 0.6$. By equation 8.1 and 8.2, the standard error of estimate if high school GPA were used to predict college GPA would be $0.800s_Y$, which represents a proportional reduction in error of $E = 0.200$, or a 20% reduction in the size of the errors in estimation relative to when predicting the mean college GPA for every application. While 20% is not small by some standards, it certainly is smaller than 36%, which is the percent of the variance in Y explained by W.

In this example, using high school GPA to estimate college GPA reduces the error in estimation by 20%, whereas using SAT reduces the error in estimation by 4.6%. The ratio of the forecasting efficiencies of these two predictors is $0.2/0.046 = 4.3$. Observe that this is much closer to the ratio of their squared correlations $(0.36/0.09) = 4$ than their unsquared correlations $(0.6/0.3) = 2$. This is contrary to our other examples, where the ratio of the unsquared correlations better reflected the gain that results from using one measure to predict Y compared to another measure.

8.2.5 When the Standard Is Perfection

Sometimes we expect very accurate prediction and naturally focus not on our ability to predict better than chance but on whatever errors remain. For instance, you understandably would not be particularly impressed if a weather predictor was able to forecast very well the temperature in a particular location on the globe based on latitude, longitude, day of the year, and time of the day. Across many predictions, it wouldn't be at all surprising to find his or her predictions correlated 0.95 with actual temperature readings across the globe.

Suppose a meteorologist proposed a new weather mechanism that when utilized in the prediction process further increases this correlation from 0.95 to, say, 0.98. Even though this is a tiny increase, it is a drop in the error of estimation from small to nearly zero, and this seems noteworthy and perhaps even impressive. So an increase of 0.03 in the size of a correlation need not be seen as small. It depends on the reference against which it is being compared. An increase from 0.95 to 0.98 seems much more impressive than the same increase in an absolute sense from 0 to 0.03.

This reflects a problem that is not unique to regression and correlation. If a certain training program raises the success rate on some task from 10 to 20%, we would naturally say that the success rate was doubled. But if the training increased the success rate from 98 to 99%, a much smaller increase, we could justifiably note that the training cut the failure rate in half. By the same reasoning, there is a sense in which a correlation of 0.99 is much higher than one of 0.98, not just 1% higher. So a correlation needs to be interpreted relative to a certain standard, and a correlation of zero is not necessarily the only meaningful standard. Sometimes, in fact, it may not be a meaningful standard at all.

8.2.6 Summary

The examples in this section illustrate that *importance*, by at least some definitions of the word, is very often proportional to r rather than r^2. But we have seen that there are some meaningful measures of importance, such as the coefficient of forecasting efficiency, where the opposite is true. Regardless, small or moderate correlations are not necessarily unimportant. On the contrary, small correlations are often more important or impressive than is often realized. Squared correlations fit remarkably well into algebra and much statistical theory, as many of the chapters in this book illustrate, but algebraic simplicity and elegance do not imply substantive meaningfulness.

8.3 Determining the Relative Importance of Regressors in a Single Regression Model

We now turn to the problem of comparing the importance of two or more regressors in the same regression model. It is common for an investigator to build a model estimating a dependent variable from several regressors, with the goal of determining which variable or variables are most important, or somehow ranking their relative importance on some kind of quantitative metric. For example, suppose the dependent variable is GPA at the end of the first year of college, and the regressors include sex, education level of the parents, high school GPA, performance on a college entrance exam, and the quality of the student's high school. When accounting for their intercorrelations, which variable best accounts for individual differences in college performance?

This is a complex problem, and there have been many treatments of this topic in the regression analysis literature. Suffice it to say that there is no single way of comparing the importance of two regressors that is ideal in all circumstances, but we feel some measures and approaches are better than others. Later we discuss a method called *dominance analysis* that is computationally burdensome and involves the estimation of many regression models but can be done fairly easily with a computer. We dedicate all of section 8.4 to it. Here we offer some opinions about some metrics that are readily available from a regression analysis.

We can rule out from the beginning and for most problems any approach that uses the unstandardized regression coefficient b_j as a measure of the importance of X_j relative to another regressor X_i. As discussed already in sections 2.3.1 and 3.1.3, as well as at the beginning of this chapter, b_j is a scale-bound metric, meaning that the regression coefficients for two regressors in a model generally cannot be meaningfully compared to each other, because changing the metric of measurement of one may change the relative sizes of their regression coefficients. So unless variables being compared are measured on the same scale, their regression coefficients cannot be compared, and judgments of relative importance cannot be made using them.

8.3.1 The Limitations of the Standardized Regression Coefficient

Perhaps the most widely used measure of relative importance is $\tilde{b}_j$, the standardized regression coefficient. Recall from sections 2.3 and 3.3.3 that the standardized regression coefficient is the regression weight a regressor would receive if X_j and Y were standardized prior to estimating the regression. Most regression programs, when requested or by default, provide the standardized regression coefficient for all regressors. It is commonly believed that the most important regressors are those with larger values of $\tilde{b}$ in absolute value. Because standardization places all regressors on the same measurement metric (one with a mean of zero and a standard deviation of one), according to widely held belief, this eliminates the problem with the use of the unstandardized regression coefficient b_j as a measure of importance, and so standardized regression coefficients can be directly compared to each other. In this section we take a contrary position and argue that $\tilde{b}_j$ should not be used as a measure of importance.

The standardized regression coefficient, $\tilde{b}_j$, can be interpreted as the expected difference in standard deviations of Y between two cases that are one standard deviation apart on X_j but are the same on all other regressors

in the model. But it seems to have been overlooked that if X_j correlates highly with other regressors, it is rare, even nearly impossible, for two such cases to even exist. As an extreme example, suppose X_1 and X_2 are correlated 0.99. Then the unique variance of each variable is $1 - 0.99^2 = 0.0199$. In a crosswise regression predicting X_2 from X_1, the standard deviation of the conditional distribution of X_2 is only $\sqrt{0.0199} = 0.141$. Since $1/0.141 = 7.09$, two people who are equal on X_1 but one standard deviation apart on X_2 are actually 7.09 standard deviations apart on the conditional distribution of X_2 in which they both fall. Rarely would you find two cases that many standard deviations apart. Even in a sample of a thousand it would be uncommon.

But now suppose you have a third regressor X_3 in the model with X_1 and X_2 that is uncorrelated with both X_1 and X_2. There would be no difficulty at all in finding two people who are one standard deviation apart on X_3 but equal on X_1 and X_2. In a model with X_1, X_2, and X_3 as regressors, $\tilde{b}_3$ would understate the relative importance of X_3 compared to X_2 and X_1, because $\tilde{b}_3$ reflects only a small fraction of the total range of X_3, while $\tilde{b}_1$ and $\tilde{b}_2$ in effect reflect the total possible range of those variables—or even more than the possible range—given that other variables are held constant.

8.3.2 The Advantage of the Semipartial Correlation

There is a simple solution to this problem with the standardized regression coefficient as a measure of importance. This solution takes into account the standard deviations of the *conditional* distributions of the regressors. For X_j, the ratio of the standard deviation of its conditional distribution (i.e., when other regressors are held constant) to the standard deviation of its unconditional or marginal distribution (i.e., without holding other regressors constant) is the square root of its tolerance, $\sqrt{Tol_j}$ (recall from section 4.4.4 that the *tolerance* of a variable is the proportion of the variance in X_j that is not explained by the other regressors in the model). The importance of X_j based on $\tilde{b}_j$ is overstated by the reciprocal of this amount. For instance, in the example in section 8.3.1, the importance of X_2 is overstated relative to X_3 by a factor of $1/\sqrt{Tol_2} = 7.09$. We can correct this problem by multiplying each value of $\tilde{b}_j$ by $\sqrt{Tol_j}$. Thus, the corrected measure of a variable's importance is $\tilde{b}_j \sqrt{Tol_j}$.

But it can be shown that $\tilde{b}_j \sqrt{Tol_j}$ is equivalent to sr_j, regressor X_j's semipartial correlation. This gives us another interpretation of sr_j as the expected difference in standardized Y between two people who are equal on all regressors except X_j and who differ on X_j by the standard deviation

of the *conditional* distribution of X_j. For any two regressors X_i and X_j, the ratio sr_i/sr_j equals the ratio of these expected differences.

A variable's semipartial correlation is available in output in most regression programs, and we've now given three different formulas (in the prior paragraph, as well in sections 3.4.1 and 3.4.5) for deriving it. Although you would typically let a computer calculate a variable's semipartial correlation, there is still another formula you could use that requires only regressor j's t from the regression (i.e., the ratio of b to its standard error), as well as the multiple correlation estimating Y from all k predictors including regressor j. That formula is

$$sr_j = t_j \sqrt{\frac{1 - R^2}{N - k - 1}}$$

We mention this formula because it illustrates how t_j can be thought of as a measure of importance with ratio qualities. Given that R, N, and k are the same for all variables in a regression, it follows that the ratio of two regressors' t-ratios is equivalent to the ratio of their semipartial correlations. That is, $t_i/t_j = sr_i/sr_j$. So if you are convinced that the semipartial correlation is a sensible measure of importance, and thus the ratio of two semipartial correlations is meaningful as a measure of relative importance, then you can just take the ratio of two regressors' t-statistics as an equivalent measure of relative importance with ratio qualities.

8.3.3 Some Equivalences among Measures

For many other measures of partial association or mathematical derivatives of those measures, the rest of this discussion could be quite lengthy, but it turns out not to have to be. It is often the case that a researcher wants to know whether one variable is more important than another, with little regard for *how much* more. But it makes little difference which measure we discuss below is used, because the vast majority of measures of partial association rank regressors in the same order (when ignoring their signs, for measures that need not be positive). That is, the regressor that is most important by any one measure will be most important by all of them. We therefore think of these measures as a family and call them measures of the *unique contribution of* X_j to the regression. These measures are

1. pr_j and pr_j^2.
2. sr_j and sr_j^2.

3. t_j or F_j, the value of t or F used to test the significance of b_j. Since *df* is equal for t and F and $F = t^2$, the regressors with the highest value of t or F will also have the smallest p-values for b_j.

4. The change in R, R^2, adjusted R^2, or adjusted R^2 when X_j is deleted from the regression.

5. Unique *SS*: The increase in $SS_{residual}$ or the decrease in $SS_{regression}$ when X_j is deleted from the regression.

But sometimes a researcher wants to be able to make ratio-type claims, such as regressor X_j is twice as important as regressor X_i. It is important to remember that most of these measures do not have similar proportionality or ratio properties. For instance, if pr_1 is twice pr_2, that does *not* mean that sr_1 is twice sr_2. But proportionality does hold for some of these measures; sr_j^2, F_j, unique *SS*, and change in R^2 are all proportional to each other, and sr_j and t_j are proportional to each other. Thus, whether one can say that X_j is twice or three times as important, as X_i will depend on the measure used, and that claim will not necessarily generalize to other measures of importance.

8.3.4 Eta-Squared, Partial Eta-Squared, and Cohen's *f*-Squared

In section 8.2 we said that although it is common for researchers to use squared measures of association to quantify the size of an effect, squared measures of association can be misleading as measures of a variable's effect. Yet use of squared measures is pervasive both when quantifying the size of an effect in an absolute sense and relative to the effects of other variables in a model. This section introduces a third measure, Cohen's f^2 (Cohen, 1988), which is often used or reported in regression analysis as a measure of a variable's effect on Y. We are not advocating its use, but because it is used by many, you should be familiar with it. It is similar to sr^2 and pr^2 in some ways and quite different from them in others.

We have said that sr_j^2 quantifies the proportion of the variance in Y uniquely explained by X_j. In terms of the Venn diagrams in Figure 3.18 or Figure 7.2, panel A, sr_1^2 = A/(A + B + C + D) and sr_2^2 = C/(A + B + C + D). In the ANOVA literature, sr^2 is mathematically equivalent to an effect size measure called *eta-squared*, symbolized η^2. The partial correlation pr_j^2, by contrast, is the proportion of the variance in Y that is unaccounted for by the other regressors in the model that can be uniquely explained by X_j. In

the Venn diagram, pr_1^2 = A/(A + D) and pr_2^2 = C/(C + D). In the ANOVA literature, pr^2 goes by the name *partial eta-squared*, or *partial* η^2.

Cohen's f^2 for regressor X_j is a ratio of the proportion of the variance in Y uniquely explained by X_j to the proportion of the variance in Y unexplained by *any* variable in the model. Suppose you have estimated Y from two regressors X_1 and X_2 as

$$\hat{Y} + b_0 + b_1 X_1 + b_1 X_2 \tag{8.3}$$

which results in R^2. In this model, $f_j^2 = sr_j^2/(1 - R^2)$. In terms of the Venn diagram, f_1^2 = A/D and f_2^2 = C/D. For instance, from the weight-loss example from Chapter 3, where X_1 is exercise frequency and X_2 is food intake, $sr_1^2 = 0.835$, $sr_2^2 = 0.091$, and $R^2 = 0.838$, so $f_1^2 = 0.835/(1 - 0.838) = 5.154$ and $f_2^2 = 0.091/(1 - 0.838) = 0.562$. The variance in weight loss uniquely explained by exercise frequency is over five times larger than the variance not accounted for by food intake and exercise frequency as a set. And the variance in weight loss uniquely explained by food intake is about half as large as the variance not accounted for by food intake or exercise frequency as a set.

We say f^2 is a ratio rather than a proportion because although it can't be smaller than zero, f^2 has no upper bound. A proportion must be between 0 and 1. If R^2 is large enough, f^2 can be greater than 1 and perhaps much greater, as in the weight-loss example just presented. So one important difference between sr^2, pr^2, and f^2 is that whereas sr^2 and pr^2 are bound between 0 and 1 and have a proportion of variance explained interpretation, f^2 does not. Thus, if you use or see people report f^2, do not interpret this like you would a squared correlation. It does not have such an interpretation.

All three of these measures of effect size have something in common. They all index X_j's effect partly in terms of the variance in Y that can be uniquely explained by X_j. These are areas A and C for X_1 and X_2 in the Venn diagram. But they differ with respect to the reference against which that explained variance is compared. For sr_j^2, the reference is *all* of the variance in Y (the area A + B + C + D). For pr_j^2, the reference is the variance in Y that is not explained by X_j (the area A + D or C + D). And for f_j^2, the reference is the variance in Y not explained by any of the regressors (area D).

Cohen's f^2 shares a limitation with pr^2 that is not a property of sr^2. An investigator unsatisfied with the size of an independent variable's effect can hunt for variables to add to the model that are correlated with Y but uncor-

related or weakly correlated with the independent variable and covariates. Doing so will increase power of the test of the independent variable (see sections 6.3.1 and 17.1.2 for a discussion of this point), but it will also increase pr^2 and f^2 for the independent variable, because adding that new regressor will take a bite out of area D in the Venn diagram while doing little to the sizes of A or C. But adding this additional regressor would do little to sr^2. So sr^2 is less easily manipulated by an investigator motivated to report a large effect than are pr^2 and f^2.

An investigator doesn't need to be consciously unscrupulous to gain this advantage of reporting partial η^2 (i.e., pr^2) or f^2 rather than η^2 (i.e., sr^2). Consider two investigators who have conducted exactly the same experiment using the same sample size, with some variables that are manipulated identically and with random assignment, so that they are all uncorrelated. Perhaps investigator A sensibly justifies including a few additional regressors in the model correlated with Y, whereas investigator B doesn't think to include any covariates or simply decides not to include any. If random assignment was effective, then all these additional regressors should be uncorrelated or nearly so with the manipulated variables. In that case, all other things being equal, we would expect investigator A to find bigger effects of the manipulated variables than investigator B if they reported effect size as pr^2 or f^2. But we would expect their effects to be very similar if they reported sr^2 as their measure of effect size, because it wouldn't be affected by the inclusion of the additional regressors. Of course, sampling variance will by itself produce *some* differences in the observed effect sizes. But effect sizes for investigators A and B for the manipulated variables would be the same if they used sr^2 rather than pr^2 or f^2.

Some of this discussion focused on the differences between sr^2, pr^2, and f^2 in a model with two regressors. But we could think of X_1 and X_2 as sets of variables, and everything we have said would generalize to measures of multivariate partial association and multivariate f^2, substituting SR^2 and PR^2 into the discussion and formulas. We have also couched our discussion in terms of squared measures. But this discussion applies to their unsquared counterparts. So the arguments we present here are yet another reason for preferring the semipartial correlation as a measure of a variable importance over many other measures that are commonly used.

8.3.5 Comparing Two Regression Coefficients in the Same Model

At the beginning of this chapter and in section 8.3 we stated that the regression coefficient b_j cannot be used to assess the relative importance of

regressors because b_j is a scale-bound measure of association. The rank ordering of regression coefficients can be changed by changing the scale of measurement for one or more variables, such that if we found that $b_1 > b_2$, it is likely we could rescale X_1 so that $b_1 < b_2$. That means that b_1 and b_2 aren't comparable, and it would not be meaningful to describe X_1 as more or less important than X_2 depending on the relative sizes of their regression coefficients.

But sometimes two or more regressors are measured on comparable scales so their regression coefficients are comparable. For instance, two regressors predicting weight loss might be hours spent running per week and hours spent swimming per week, or they might be calories consumed before 5 in the afternoon and calories consumed after 5. When two regressors are measured in the same units, as in these examples, it is meaningful to ask whether one has a larger effect on Y, such as weight loss, than the other. By regressing weight loss on hours spent running, hours spent swimming, and additional covariates if desired, one can determine whether the regression coefficient for hours spent running is the same or different than the regression coefficient for hours spent swimming. If we found a difference between these two regression coefficients, we would conclude that an additional 1 hour of swimming does not have the same effect on weight loss as an additional 1 hour of running.

It is possible to compare the regression coefficients for two regressors X_1 and X_2 that are both in a model of Y. This can be done even if the two regressors are measured on different scales, but the comparison probably wouldn't be particularly meaningful in that case. The easiest way to do this is to first create two new variables, one that is one-half the *sum* of X_1 and X_2, and the other that is the one-half the *difference* between X_1 and X_2. We will call these two new regressors X^+ and X^-, respectively. That is,

$$X^+ = 0.5(X_1 + X_2)$$
$$X^- = 0.5(X_1 - X_2)$$

With X^+ and X^- constructed, regress Y on X^+ and X^-, as well as any covariates you want to hold constant. In that regression analysis, the regression coefficient for X^- will be equal to difference between the regression coefficients for X_1 and X_2 from the regression estimating Y from X_1 and X_2. The standard error for the regression coefficient X^- along with its t-value, p-value, and confidence interval can be used for inference about the difference between the regression coefficients for X_1 and X_2.

Now that we have described this in words, we put it in symbolic form. Our model of Y is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \tag{8.4}$$

and we want to test the null that ${}_T b_1 = {}_T b_2$ or construct a confidence interval for that difference. We do so by computing $X^+ = 0.5(X_1 + X_2)$ and $X^- = 0.5(X_1 - X_2)$ and then estimate the model

$$\hat{Y} = b_0 + b_1^+ X^+ + b_2^- X^-$$

It can be shown that in this model, b_1^- is equal to $b_1 - b_2$ from equation 8.4, and so a test that ${}_T b_1^- = 0$ is equivalent to the test that ${}_T b_1 = {}_T b_2$. Likewise, a confidence interval for ${}_T b^-$ is a confidence interval for ${}_T b_1 - {}_T b_2$.

We illustrate using data from a national survey of residents of the United States. The data set is named POLITICS and it can be downloaded from this book's web page at *www.afhayes.com*. The participants in this study were asked a set of questions used to quantify their knowledge of politics, politicians, and the political process (`pknow`). In addition, various measures of frequency of exposure (in days per week) to various sources of news were measured, including reading the newspaper (`npnews`) and watching a national network news broadcast (`natnews`). Under the assumption that knowledge is caused by exposure to information in the news, we ask whether reading the newspaper has the same effect on knowledge acquisition as does watching the national network news broadcast. We can answer this question using these data because newspaper reading and watching televised news are scaled on the same metric: days per week.

Call Y political knowledge, call X_1 days per week reading the newspaper, and call X_2 days per week watching the national network news broadcast. We include age (X_3) and sex (X_4) as covariates. Regressing Y on X_1 through X_4 yields

$$\hat{Y} = 8.795 + 0.371 X_1 + 0.155 X_2 + 2.245 X_3 - 0.009 X_4 \tag{8.5}$$

with $R = .401$ and $SS_{residual} = 5442.070$. In this model, $b_1 = 0.371$, meaning that holding sex, age, and days per week watching the national network news broadcast constant, two people who differ by 1 day in how often they read the newspaper are estimated to differ by 0.371 units in their knowledge. But the effect of watching the national network news on knowledge appears smaller. Here, $b_2 = 0.155$, meaning that holding sex, age, and days per week reading the newspaper constant, two people who differ by 1 day in how

often they watch the national network news broadcast are estimated to differ by 0.155 units in their knowledge.

To test the equality of the two regression coefficients, we construct X^+ and X^- as described above and then regress Y on these sums and differences as well as sex and age. In SPSS, the code is

```
compute sum=0.5*(npnews+natnews).
compute diff=0.5*(npnews-natnews).
regression/dep=pknow/method=enter sum diff sex age.
```

Or in SAS and STATA, use

```
data politics;set politics;sum=0.5*(npnews+natnews);
   diff=0.5*(npnews-natnews);run;
proc reg data=politics;model pknow=sum diff sex age;run;
```

```
gen sum=0.5*(npnews+natnews)
gen diff=0.5*(npnews-natnews)
regress pknow sum diff sex age
```

The resulting model is

$$\hat{Y} = 8.795 + 0.525X^+ + 0.216X^- + 2.245X_3 - 0.009X_4$$

with $R = .401$ and $SS_{residual} = 5442.070$. So the fit of this model is the same as the fit of the model using X_1 and X_2 as regressors, and it generates the same values of $\hat{Y}$. But notice that the regression coefficient for X^- is 0.216, which is the same as $b_1 - b_2 = 0.371 - 0.155$ from equation 8.5. From the regression output, the standard error of the regression coefficient for X^- is 0.130, and so $t(335) = 1.666, p = .097$. We can't reject the null hypothesis that watching the network news and reading the newspaper have the same effect on political knowledge.

A little algebra shows why this works. We defined X^+ as $0.5(X_1 + X_2)$ and X^- as $0.5(X_1 - X_2)$. If we solve for X_1 and X_2 in terms of X^+ and X^-, we get $X_1 = (X^+ + X^-)$ and $X_2 = (X^+ - X^-)$. Therefore,

$$\begin{aligned}
\hat{Y} &= b_0 + b_1X_1 + b_2X_2 \\
&= b_0 + b_1(X^+ + X^-) + b_2(X^+ - X^-) \\
&= b_0 + b_1X^+ + b_1X^- + b_2X^+ - b_2X^- \\
&= b_0 + (b_1 + b_2)X^+ + (b_1 - b_2)X^-
\end{aligned}$$

which means that no matter what b_1 and b_2 turn out to be, the linear function $b_0 + b_1X_1 + b_2X_2$ will be exactly replicated by the function $b_0 + (b_1 + b_2)X^+ + (b_1 - b_2)X^-$. Therefore, the hypothesis ${}_Tb_1 = {}_Tb_2$ is equivalent to the hypothesis that the true regression coefficient for X^- is zero.

This trick to comparing two regression coefficients from the same model is easy to employ. But some statistical packages have this test built in, though in somewhat disguised form. We can think of $b_2 - b_1$ as a weighted linear combination of regression coefficients, just as is $\hat{Y}$. In this case, the linear combination is $0(b_0) + 1(b_1) - 1(b_2) + 0(b_3) + 0(b_4)$, which reduces to $b_1 - b_2$. Your computer software may be able to construct a standard error for this linear combination, which can be used to generate a p-value or a confidence interval.

SPSS has such a feature. The code to conduct this test is

```
glm pknow with npnews natnews sex age/print = parameters/
   lmatrix all 0 1 -1 0 0.
```

The sequence of five numbers following **`lmatrix all`** is the weights for the regression constant and b_1 through b_4 in that order.

The RLM macro for SPSS and SAS has a comparable feature. In SPSS, the RLM code to conduct this test is

```
rlm y=pknow/x=npnews natnews sex age/contrast=0,1,-1,0,0.
```

Or in the SAS version, use

```
%rlm (data=politics,y=pknow,x=npnews natnews sex age,
   contrast=0 1 -1 0 0);
```

The RLM documentation in Appendix A provides some detail about the **`contrast`** option in RLM. See your preferred program's documentation for information about whether it can perform this test.

8.4 Dominance Analysis

Dominance analysis (Azen & Budescu, 2003; Budescu, 1993) is a means of rank ordering the regressors in a model with respect to importance as defined by improvement in the fit of the model. It is based on how one regressor relative to another contributes to increasing R, which coincides with the amount of the variance in Y a regressor explains or how much it shrinks the error in estimation. Actually, dominance analysis as described by Azen and Budescu (2003) and Budescu (1993) relies on the relative

increase in R^2 rather than R, but since we are rank-ordering the variables, it doesn't matter whether we use an increase in R or R^2.

Remember that sr_j^2 quantifies the amount R^2 increases when X_j is added to a model without it. Given this, it sounds as if dominance analysis merely defines a regressor's importance as sr^2 and so rank-orders regressors with respect to the absolute value of their semipartial correlations. However, this is not quite what dominance analysis does, because it assesses regressor j's contribution to model fit in competition with another regressor *i when that competing regressor i is not in the model*. So unlike sr_j from the full model with all k regressors, dominance analysis uses the increase in fit due to regressor j relative to regressor i in a model that otherwise includes only the other $k-2$ regressors. Furthermore, regressors i and j compete against each other in *all possible subset* models that contain some or all of those $k-2$ regressors.

Suppose that in your regression model with k regressors, you want to know whether regressor j is more important than regressor i. Define ΔR_j as the amount R increases when regressor j is added to a model that contains regressor set A, where set A is defined as some *subset* of the remaining $k-2$ regressors with i not included in set A. Similarly, define ΔR_i as the amount R increases when regressor i is added to a model that contains that same regressor set A, with A not including regressor j. So these two regressions have the regressors in set A in common, with set A containing neither regressor i nor j. In other words, they differ only with respect to whether i or j is included in the model.

Three relations between ΔR_j and ΔR_i are possible: $\Delta R_j > \Delta R_i$, $\Delta R_j < \Delta R_i$ or $\Delta R_j = \Delta R_i$. That is, R may increase more when j is added to regressor set A relative to when i is added, R may increase more when i is added to regressor set A relative to when j is added, or the increase in R may be the same. In the first case, we'd say that regressor j is more important than regressor i. In the second case we would claim i is more important than regressor j. In the third case we would say that regressor i and j are equally important. Note that given that ΔR_j and ΔR_i are constructed from models that differ only with respect to the inclusion of regressor i or j, we can also just compare the size of the two multiple correlations rather than the change in the multiple correlations.

This competition between i and j is undertaken for all possible models defined by subsets of the regressors in set A, including the subset containing no regressors. There are 2^{k-2} such subsets of $k-2$ regressors. For example, if $k = 5$ regressors, when we remove i and j because they are in competition with each other, then there are three remaining regressors, and so $2^{5-2} = 8$

subsets of these three regressors one can construct. If these regressors are X_1, X_2, and X_3, then the eight subsets are (1) no regressors, (2) X_1 only, (3) X_2 only, (4) X_3 only, (5) X_1 and X_2, (6) X_1 and X_3, (7) X_2 and X_3, and (8) X_1, X_2, and X_3. ΔR_j and ΔR_i is constructed for each of these eight models, resulting in eight comparisons. The outcome of these comparisons determines the extent to which variable j is deemed more important than regressor i.

8.4.1 Complete and Partial Dominance

In a dominance analysis, regressor j is said to *dominate* regressor i if $\Delta R_j > \Delta R_i$ for each and every one of these 2^{k-2} comparisons based on subsets of the $k-2$ other regressors. In other words, if in every model containing some subset of the $k-2$ regressors not being compared R increases more when j is added to the model compared to when regressor i is added to the model, then regressor j dominates regressor i and regressor j is deemed more important than regressor i.

In the language of dominance analysis as introduced by Budescu (1993), dominance is an all-or-none property. Regressor j either dominates i entirely or completely, or it does not. But it seems worth acknowledging that regressor j may enhance prediction or reduce error in the estimation of Y in *most* of the subset models. We call such a scenario *partial* dominance and say that regressor j *partly dominates* regressor i, by our definition, if $\Delta R_j > \Delta R_i$ in more than half of the subset model comparisons. Of course, in some circumstances, it may be that $\Delta R_j > \Delta R_i$ in as many subset models as does $\Delta R_i > \Delta R_j$. In that case, neither j nor i dominates the other either partially or completely.[1]

Clearly, a claim of complete dominance of one regressor over another is a punchier conclusion than a claim of partial dominance. Complete dominance means that the completely dominant regressor adds more to prediction accuracy or explaining variance in Y than does any regressor it completely dominates; thus the dominant regressor is more important by these measures. But partial dominance means that in at least some of the possible subset models, the less dominant regressor actually explains more variance or does better at reducing prediction inaccuracy. This leaves the question as to which of the two is more important open to debate.

[1] Azen and Budescu (2003) introduce finer degrees of dominance than the *complete* versus *partial* distinction we make. See their discussion of *conditional* dominance and *general* dominance.

8.4.2 Example Computations

We illustrate these computations using the POLITICS data first described in section 8.3.5. In addition to a measure of each participant's political knowledge (`talkrad`) and days per week reading the newspaper (`npnews`) and watching the national network news broadcast (`natnews`), the data set includes days per week watching a local news broadcast (`locnews`) and how much the participant reports listening to political talk radio (`talkrad`).

We treat political knowledge as the dependent variable Y and determine whether listening to political talk radio is more or less important than watching the national network news broadcast in explaining individual differences in political knowledge. We do this in the context of a full model that includes all four sources of information as regressors. So $k = 4$. We'll call political talk ratio regressor j and national network news use regressor i. The two remaining regressors, reading the newspaper and watching the local news broadcast, are defined as set A.

It is important to note that we can't use the test described in section 8.3.5 to determine the relative importance of listening to political talk radio and watching the national network news, because the measurement scales for `natnews` and `talkrad` are different. Specifically, listening to political talk radio is the participant's average response on an ordinal scale to two questions about how often he or she listens to political talk radio and how much attention he or she pays when listening. But watching the network news is measured as number of days per week the person watches the broadcast.

With set A defined as two regressors, there are $2^{4-2} = 4$ subsets of these two regressors. Those four sets can be found in the rows of Table 8.2. For each each subset, we calculate R three times, first regressing Y on just the variables in the A subset, then regressing Y on the subset as well as regressor j, then regressing Y on the subset as well as regressor i. Importantly, we do *not* calculate R when both regressor i and j are in the model. With these computations done we can derive ΔR_j and ΔR_i in each of the four subsets. Table 8.2 shows these computations.

As can be seen, in all four models defined by subsets of newspaper reading and local news use, adding talk radio use to the model increases R more than does watching the national network news. Never does the addition of watching the national network news improve model fit more than listening to political talk radio. So talk radio use completely dominates watching the national network news in explaining variation in political knowledge.

TABLE 8.2. Relative Improvement in Fit for Dominance Computations

Set A subset	R	Adding i `talkrad` R	Adding j `natnews` R	ΔR_j	ΔR_i
None	—	.261	.148	.261	.148
`npnews`	.298	.393	.310	.095	.012
`locnews`	.106	.288	.237	.182	.131
`npnews` and `locnews`	.338	.430	.373	.092	.035

8.4.3 Dominance Analysis Using a Regression Program

Remember that the full regression model contains k regressors, and interest is in rank-ordering the relative importance of these k regressors, not just two of them. Our discussion thus far has been restricted to one pair of regressors i and j. But in a model with k regressors, there are $k(k-1)/2$ possible pairs of regressors i and j. In order to rank-order the relative importance of the k regressors, these comparisons need to be done for *all* of these possible pairs of regressors. Except for fairly simple models where k is small (5 or 6 at most), this is a lot of computation and comparisons to keep track of.

Dominance analysis is a very computationally tedious task and so should be left to a computer. As far as we are aware, you won't find dominance analysis implemented in any software off the shelf. Azen and Budescu (2003) provide a SAS macro that can conduct a more sophisticated version of dominance analysis—*quantitative dominance analysis*—than we have described. The RLM macro described in Appendix A has an option for the *qualitative* dominance analysis we discuss. An example output from the RLM macro for SPSS can be found in Figure 8.2, generated with the command

```
rlm y=pknow/x=natnews npnews locnews talkrad/subsets=1/dominate=1.
```

The section relevant to the dominance analysis can be found toward the bottom under the heading "Dominance matrix." This matrix contains four

rows and columns because there are four regressors.[2] The elements in the dominance matrix, all between 0 and 1, are the proportion of the subset models in which adding the variable in the row to the subset increases R more than does adding the variable in the column to the subset. An entry of 1 indicates complete dominance of the variable in the row over the variable in the column, and a number greater than 0.5 but less than 1 indicates partial dominance. In general, the entries in row i and column j and row j and column i will add to 1, except in the case where there are ties.

Remember that in a dominance analysis, regressors i and j being compared are never in the model at the same time. So with four regressors as in this example, the subset models contain no more than two regressors, because the subset excludes the variables in the row and the column of the table. In this case, there are four possible submodels, one with no regressors, two with one regressor, and one with two regressors. For this reason, all entries are either 0, 0.25, 0.5, 0.75, or 1, which are the only proportions possible when an integer between 0 (inclusive) and 4 is divided by 4.

Examining the dominance matrix reveals that reading the newspaper completely dominates the other three sources of news. In all subset models, adding newspaper reading frequency increases R more than does adding any other source. Following newspaper news use, political talk radio is next most dominant, as it is completely dominant over exposure to local news and the national network news. Local news use comes next, but it only partially dominant over watching the national network news broadcast. In only three of the four (and hence the 0.75 entry) subset models did adding local news use increase R more than did adding national network news. Thus, from the dominance analysis, the importance of these four sources can be ranked in the order newspaper > talk radio > local network news > national network news. This is the same rank ordering one could get looking at the absolute values of the semipartial correlations or t-values, though this wouldn't necessarily always be the case.

The dominance matrix is constructed from an all subsets regression (introduced in section 7.3.2, though in that section we discounted its usefulness in the context of prediction and model selection). The RLM macro can do all subsets regression. In Figure 8.2 the results of all subsets re-

[2]Budescu (1993) does not discuss such a table in the article using this term. So far as we are aware, this term is our invention, but any user of a dominance analysis would probably have to invent such a table to make sense of his or her own analysis. We feel it is a sensible way of representing the results of what Budescu (1993) calls a "qualitative" dominance analysis.

```
Dependent Variable
 pknow

Sample size
        340

Complete Model Regression Summary
          R       R-sq   Adj R-sq          F          p    SEofEst
      .4440      .1971      .1875    20.5629      .0000     3.9416

ANOVA summary table
                   SS         df         MS
Regress    1277.8934     4.0000   319.4733
Residual   5204.6801   335.0000    15.5364
Total      6482.5735   339.0000    19.1226

Regression Model
              Coeff         se          t          p       LLCI       ULCI
constant     8.5566      .5349    15.9959      .0000     7.5043     9.6088
natnews       .2086      .0919     2.2687      .0239      .0277      .3895
npnews        .4717      .0783     6.0258      .0000      .3177      .6257
locnews      -.4441      .1072    -4.1439      .0000     -.6549     -.2333
talkrad       .8324      .1697     4.9044      .0000      .4986     1.1663

Simple (r), semipartial (sr), and partial (pr) correlations with outcome
                  r          sr         pr
natnews       .1481      .1111      .1230
npnews        .2983      .2950      .3127
locnews      -.1065     -.2029     -.2208
talkrad       .2609      .2401      .2588

*************************************************************************

All subsets regression results
 natnews  npnews locnews talkrad       R
   .0000   .0000  1.0000   .0000   .1065
  1.0000   .0000   .0000   .0000   .1481
  1.0000   .0000  1.0000   .0000   .2373
   .0000   .0000   .0000  1.0000   .2609
  1.0000   .0000   .0000  1.0000   .2793
   .0000   .0000  1.0000  1.0000   .2878
   .0000  1.0000   .0000   .0000   .2983
  1.0000  1.0000   .0000   .0000   .3097
  1.0000   .0000  1.0000  1.0000   .3318
   .0000  1.0000  1.0000   .0000   .3379
  1.0000  1.0000  1.0000   .0000   .3735
   .0000  1.0000   .0000  1.0000   .3935
  1.0000  1.0000   .0000  1.0000   .3949
   .0000  1.0000  1.0000  1.0000   .4299
  1.0000  1.0000  1.0000  1.0000   .4440

*************************************************************************

Dominance matrix
          natnews  npnews locnews talkrad
natnews      .000    .000    .250    .000
npnews      1.000    .000   1.000   1.000
locnews      .750    .000    .000    .000
talkrad     1.000    .000   1.000    .000
```

Proportion of submodels in which the inclusion of the regressor in the row results in a larger increase in R than inclusion of the regressor in the column.

FIGURE 8.2. SPSS RLM macro output showing the dominance matrix and all subsets regression results.

gression are displayed above the dominance matrix. In this section of the output, possible models with at least one predictor define the rows, a 1 indicates the variable in the column is in that model, whereas 0 indicates that variable is not in that model. The multiple correlation for that model is in the "R" column. If you were to look carefully and with considerable concentration, you would observe that there are four sets of subset models that are the same with respect to the inclusion or exclusion of `locnews` and `talkrad` yet include `natnews` or `npnews` but not both. For example, one of these sets is defined by rows 2 and 7, where both `locnews` and `talkrad` are not in the model. Another set is defined by rows 3 and 10, where `locnews` is in the model but `talkrad` is not. In each of these four sets, R is larger when `npnews` is included and `natnews` is excluded than when `natnews` is included and `npnews` is excluded. This is why newspaper reading frequency completely dominates watching the national network news and its entry in the dominance matrix is 1.

This is not the case for local news and national network news. There are also four sets of subset models that are the same with respect to the inclusion or exclusion of `npnews` and `talkrad` yet include `natnews` or `locnews` but not both (e.g., rows 1 and 2; rows 5 and 6). In three of these R is larger when `locnews` is included and `natnews` is excluded, but in one of them, R is larger when `natnews` is included and `locnews` is excluded. This translates into 0.75 in the local news row and national network news column, and 0.25 in the national network news row and local news column. Local news use only partially dominates national network news use.

8.5 Chapter Summary

The importance of a regressor in a regression model can be framed in either substantive or applied terms, or in abstract quantitative terms. In substantive or applied terms, importance is a value judgment, and statistics has little to say about matters of personal or social values. But statistics can be used to inform judgments relevant to such values. Unfortunately there is no single way of quantifying the importance of an effect in a regression analysis that can or should be applied to all circumstances. It is tempting to just wash your hands of the problem by relying on rules of thumb about what constitutes a large or a small effect, but arbitrary guidelines such as those you find in some books about effect size are not useful, in our opinion.

Squaring measures of simple or partial association as a means of quantifying the importance or size of an effect is deeply ingrained in practice.

When interest is merely in rank-ordering the size of effects in a regression analysis, little harm is done by squaring simple, partial, or semipartial correlations as measures of relative importance. But importance is as often if not more often proportional to unsquared correlations rather than squared correlations. We provided some examples illustrating that squaring relationships can result in counterintuitive or understated claims about the size or magnitude of a relationship in statistical and substantive or applied terms. We also discussed how zero is not the only meaningful reference point for evaluating the size of a relationship or effect.

The standardized partial regression coefficient is perhaps the most widely used measure of relative importance or relative size of the effect of a regressor. But this measure has problems that make it harder to recommend than the semipartial correlation, our preferred metric. Many measures of variable importance rank-order the variables in the same way, so it makes little difference which is used when that is the goal. Dominance analysis is an interesting approach to assessing relative importance in a statistical sense, and we provide an entire discussion and simple way of conducting a dominance analysis using SPSS or SAS.

In all examples of regression analysis provided thus far in this book, a regressor was either dichotomous or a quantitative dimension of some kind. But researchers often want to use regressors that code membership in one of several groups. How to properly represent multicategorical variables for use in a regression analysis is the topic of the next two chapters.

9
Multicategorical Regressors

In all discussions and examples of linear regression analysis thus far, regressors have been either quantitative variables or dichotomous. But multicategorical variables—variables that are categorical but with more than two categories—can be used as regressors if special procedures are employed to represent group membership. This chapter describes regression analysis with multicategorical variables. It begins by introducing indicator coding of groups, followed by a discussion of the mathematical parallels between linear regression analysis with multicategorical variables and single-factor or "one-way" ANOVA. Comparing groups while statistically controlling for other variables that the groups may (or may not) differ on is the next topic, followed by a discussion of the linkage between linear regression analysis and analysis of covariance.

Regression analysis is used to estimate a dependent variable Y from a set of regressors. The regressors often are numerical, but we saw in Chapter 5 that a dichotomous variable can also be used as a regressor. When the two groups that a dichotomous regressor represents are coded with numbers that differ by only one unit (e.g., as 0 and 1, or -0.5 and 0.5), its regression coefficient quantifies the difference between the group means on Y. When other regressors are in the model, the dichotomous regressor's regression coefficient quantifies the difference between the group means on Y when all other regressors are held constant.

Often you will want to include a regressor (or two or three) in a regression model that represents membership in one of *several* distinct groups. For instance, perhaps you want to include religion (Catholic, Jewish, Protestant, Muslim) or occupational category (manual laborer, office worker, retail sales, professional, etc.) as a regressor in your model. Or perhaps you have data from an experiment that includes a control condition and three or four experimental treatment conditions.

These variables might be numerically coded in your data, much like when using 0 for females and 1 for males. For instance, maybe ethnicity is coded 1 for Caucasians, 2 for Asians, 3 for Hispanics, and 4 for everyone else. But these numbers are arbitrary codes. They carry no quantitative information. Although this is true for dichotomous regressors as well, we saw in Chapter 5 that this is not a problem when there are only two groups. But when a categorical variable has more than two categories, we cannot just use these arbitrary numerical codes as a regressor in a regression model. Doing so will generally yield nonsense.

This chapter addresses how to properly represent categorical variables such as these so that they can be used in a regression model. We will call a variable that is categorical and codes more than two groups a *multicategorical* variable. A multicategorical variable is sometimes called a *factor*, a term often used in ANOVA. In this chapter we discuss the use of a multicategorical variable or factor in a regression model that is strictly nominal and thus of *kind*, such as the ethnicity or religion examples above. But the methods described in this chapter could also be used for any regressor that is categorical but ordinal. An example would be level of education, which might include the categories "no high school diploma," "high school diploma but no university or college coursework," "some university or college coursework," "university or college degree," "some postgraduate coursework," and "postgraduate degree." But we focus on nominal multicategorical variables in this chapter, saving a discussion of ordinal multicategorical variables in regression analysis for Chapter 10.

A regression analysis with just one multicategorical regressor is essentially a one-way ANOVA, which you may have already learned about elsewhere. We demonstrate how so in this chapter. It may seem like the method discussed here is really just a roundabout way of doing ANOVA, but a regression-based version of ANOVA is much more versatile. For example, your multicategorical regressor might be religion, but you could include ethnicity, biological sex, income, various measures of certain social attitudes, or any other conceivable variable in the model too.

9.1 Multicategorical Variables as Sets

In section 5.1.1 we saw that a dichotomous variable can be represented with an *indicator* or *dummy variable*: a variable taking only one of two values. Dichotomous regressors such as this can be legitimately included in a regression model as is. You might have several dichotomous variables

as regressors in a linear model, such as biological sex, whether or not a person has a university or college degree, whether or not he or she is currently married, and so forth.

A multicategorical variable can be represented with a *set* of indicator or dummy variables—variables with the values 0 and 1—and we introduce a system for coding a categorical variable based on dummy variable sets. As will be seen, if a multicategorical variable has g categories, it takes a set of $g - 1$ variables (dummy variables or something else) to code the g categories. So although we think of a dimension such as ethnicity as a single variable, if that dimension includes more than two categories, it requires more than one regressor to represent it. We use the term *compound variable* to refer to a multicategorical variable represented with a set of variables in a regression model. Although ethnicity might have four categories and therefore requires three regressors in the model to represent it, it is still only a single variable in our thinking about it.

9.1.1 Indicator (Dummy) Coding

Suppose you included a question in a survey like

> What is your current marital status (choose only one)?
>
> (a) Married
>
> (b) Divorced
>
> (c) Single
>
> (d) Widowed

You may choose to code peoples' responses to this question with the numbers 1, 2, 3, and 4, to represent the four responses. Thus, in your data, you would have a single column containing a person's marital status as coded arbitrarily, with the numbers 1 through 4 representing the person's response to the question.

Table 9.1 contains a hypothetical data set containing 20 cases with marital status represented in this way in the column labeled X_1. The data file is available from *www.afhayes.com* and is named MARRIED. The column labeled Y is a measure of how satisfied the person is with life at this moment (`satis` in the data file), on a 1 to 100 scale, based on other questions in the survey. The remaining columns will be explained later.

Perhaps you want to know whether there is a relationship between marital status and life satisfaction and so you regress life satisfaction (Y)

TABLE 9.1. Marital Status, Life Satisfaction, Income (in Thousands of Dollars), and Sex (0 = Female, 1 = Male)

ID		*mstatus* X_1	D_1	D_2	D_3	D_4	*satis* Y	*income* X_2	*sex* X_3
1	Single	3	0	0	1	0	85	53	0
2	Divorced	2	0	1	0	0	80	65	1
3	Widowed	4	0	0	0	1	72	54	0
4	Widowed	4	0	0	0	1	60	35	0
5	Married	1	1	0	0	0	92	73	1
6	Single	3	0	0	1	0	88	75	1
7	Divorced	2	0	1	0	0	74	57	0
8	Divorced	2	0	1	0	0	84	59	0
9	Single	3	0	0	1	0	88	60	0
10	Married	1	1	0	0	0	82	52	0
11	Single	3	0	0	1	0	76	47	1
12	Windowed	4	0	0	0	1	78	60	0
13	Married	1	1	0	0	0	78	63	1
14	Married	1	1	0	0	0	93	66	1
15	Divorced	2	0	1	0	0	73	51	0
16	Married	1	1	0	0	0	80	53	1
17	Single	3	0	0	1	0	75	44	0
18	Divorced	2	0	1	0	0	85	61	1
19	Widowed	4	0	0	0	1	76	55	0
20	Married	1	1	0	0	0	88	55	1

on marital status (X_1) using a simple linear regression model $\hat{Y} = b_0 + b_1 X_1$. If you did so, you'd find the best-fitting model is $\hat{Y} = 89.104 - 3.725X_1$, with $R = 0.536$. This model would lead to the claim that married people ($X_1 = 2$) are estimated to have a life satisfaction of $\hat{Y} = 89.104 - 3.725 \times 2 = 81.654$ units on average, single people ($X_1 = 3$) are estimated to have a life satisfaction of $\hat{Y} = 89.104 - 3.725 \times 3 = 77.929$ units on average, and the correlation between actual and estimated life satisfaction is 0.536. You might also say that single people are estimated to be 3.725 units less satisfied than those who are divorced. This is $b_1 = -3.725$—the estimated difference in Y between two cases that differ by one unit on X_1. From the negative value of b_1, it seems that as marital status increases, life satisfaction decreases.

These claims are all nonsense. The use of the codes 1, 2, 3, and 4 for the different marital status groups was arbitrary. And the choice to use numbers rather than letters or some other symbol was just as arbitrary

as which numbers were used to code the groups. Most important, even though there is nothing even quantitative about a person's marital status, the regression math is treating these "values" of marital status as if they carry quantitative information about distinctions between people in their absolute degrees of marital status, and then it uses information about the relationship between these arbitrary numbers and life satisfaction to derive the regression model. A multicategorical variable should not be included in a regression model in this manner. Doing so will yield nonsense. An alternative approach to representing groups is required.

To understand the alternative approach, you could imagine a different way of asking a person about marital status. Rather than asking the person to choose from one of four response options, you could ask four yes/no questions, as such:

> Are you currently. . .
>
> . . . married? Yes_ No_
> . . . divorced? Yes_ No_
> . . . single? Yes_ No_
> . . . neither married, divorced, nor single? Yes_ No_

You wouldn't actually need to ask the last question, because saying no to the first three implies a yes answer to the last question, and saying yes to any of the preceding questions implies saying no to the last.

Recognizing the redundancy of the last question, assume you only asked the first three questions. When marital status is asked in this way, you might enter the data with three variables that are dichotomous, set to 1 if the person said yes to the question and 0 if the person said no. The three columns in Table 9.1 labeled D_1, D_2, and D_3 represent how each person presumably would have responded to these three questions based on their marital status, had they been asked about marital status in this manner rather than in the first way. Notice that someone who said no to all three questions has a zero for D_1, D_2, and D_3. In that sense, the fourth question is not needed. If we had asked this question and coded the response with D_4, as in Table 9.1, then we would know that $D_4 = 1 - (D_1 + D_2 + D_3)$ since a person would not (or at least should not) say yes to more than one of the questions. So D_4 contains no information about marital status not already provided by D_1, D_2, and D_3.

Although these two approaches are essentially equivalent in the information they yield, they are not equivalent in some important ways. The

first format of the question requires less time to ask and probably less time to answer. The first format also produces data that are much easier to enter if you are entering manually by typing them into a program for analysis. Whereas the first approach requires entering only a single variable (i.e., a single column in one's data set) with the values 1 through 4, the second approach requires several variables and hence much more typing and potential for data entry errors. But most important for the sake of regression analysis, marital status as represented with a single variable based on the first format could not be used as predictor in a regression model, as discussed earlier, whereas the data based on the second format can be used legitimately in a regression analysis.

We know that it is legitimate to include dichotomous variables as regressors in a linear regression model. If D_1, D_2, and D_3 contain all the information about marital status contained in X_1, then we can estimate a person's life satisfaction from marital status by using D_1, D_2, and D_3 as regressors, which is legitimate, rather than X_1, which is not. Doing so with the data in Table 9.1 yields

$$\hat{Y} = 71.500 + 14.000D_1 + 7.700D_2 + 10.900D_3$$

and $R = 0.649$. This regression model does contain three regressors, but keep in mind that really it is a model of life satisfaction from a single compound variable—marital status. Thus, in our thinking about the problem, we are predicting Y from only one variable, even though the model literally does contain three regressors.

This system of coding groups is called *indicator coding* or *dummy coding*. In this example, there were $g = 4$ groups. For reasons to be made clear soon, we use only three rather than four dummy variables to code a multicategorical variable with four groups. Recall that we didn't have to ask the last question in the set of four, as it contains no new information, so we need not and indeed *cannot* include D_4 in the model along with the other three dummy variables. More generally, a multicategorical variable with g categories requires $g - 1$ dummy variables (or $g - 1$ variables of some other kind, as discussed in Chapter 10) to represent membership in the g categories.

A set of indicator variables to represent a compound variable with g categories can be constructed with a simple algorithm. To code membership in one of g groups, use $g - 1$ dummy variables $D_1, D_2, \ldots, D_{g-1}$. Set $D_j = 1$ for all cases that are in group j and set all remaining $g - 1$ $D_{j,(i \neq j)}$ to 0. This algorithm is illustrated in Table 9.2.

TABLE 9.2. Indicator Coding of g Categories

Group	D_1	D_2	$\cdots$	D_j	$\cdots$	D_{g-1}
1	1	0	$\cdots$	0	$\cdots$	0
2	0	1	$\cdots$	0	$\cdots$	0
$\vdots$						
j	0	0	$\cdots$	1	$\cdots$	0
$\vdots$						
$g-1$	0	0	$\cdots$	0	$\cdots$	1
g	0	0	$\cdots$	0	$\cdots$	0

9.1.2 Constructing Indicator Variables

We can perform a regression analysis estimating life satisfaction from marital status using indicator coding of marital status without having to inconvenience our respondents using the awkward, second question format described earlier. We can ask participants using the simpler first format and then enter the data as if the second format was used. But more typically, a researcher would enter the data as a single variable with arbitrary numerical codes for the groups, as with X_1 in Table 9.1, and then write computer code that produces the indicator codes.

Using SPSS, for example, with marital status and life satisfaction in the data as variables named *mstatus* and *satis*, the code below constructs D_1, D_2, and D_3 and then regresses life satisfaction on marital status. The resulting output can be found in Figure 9.1.

```
compute d1=(mstatus=1).
compute d2=(mstatus=2).
compute d3=(mstatus=3).
regression/dep=satis/method=enter d1 d2 d3.
```

Corresponding code in SAS and STATA is

```
data married;set married;
d1=(mstatus=1);d2=(mstatus=2);d3=(mstatus=3);run;
proc reg data=married;model satis=d1 d2 d3;run;
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.649[a]	.421	.313	6.550

a. Predictors: (Constant), d3, d2, d1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	500.050	3	166.683	3.885	.029[b]
	Residual	686.500	16	42.906		
	Total	1186.550	19			

a. Dependent Variable: satis

b. Predictors: (Constant), d3, d2, d1

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	71.500	3.275		21.831	.000
	d1	14.000	4.228	.833	3.311	.004
	d2	7.700	4.394	.433	1.752	.099
	d3	10.900	4.394	.613	2.481	.025

a. Dependent Variable: satis

FIGURE 9.1. SPSS output from a multiple regression analysis estimating life satisfaction from marital status represented with three indicator variables.

```
gen d1=(mstatus==1)
gen d2=(mstatus==2)
gen d3=(mstatus==3)
regress satis d1 d2 d3
```

These are not the only ways to construct indicator codes. Different programmers choose to write code in different ways; you may already have a better approach in mind that is more efficient or otherwise more elegant. And some programs have commands built in to produce indicator codes automatically. Check your program's manual.

9.1.3 The Reference Category

In this example, the "Widowed" category was treated differently than all other categories; it had no indicator variable of its own in the regression model. Most systems of coding multicategorical variables require that one category be treated differently from the others. When using indicator coding, this category is called the *reference category* or *base category*, for reasons that will be made clear in section 9.1.6. If we perform the coding

by hand or write the code to do so ourselves, as above, we have a number of options concerning the reference category. If your program produces codes automatically, it will probably have its own system for designating the reference category.

What happens if we fail to select a reference category when using indicator coding and instead use g dummy variables as regressors, one for each of the g categories? In the example above, this would mean including D_4 in the regression model along with D_1, D_2, and D_3. In section 9.1.1 we saw that D_4, the dummy variable for the widowed, could be perfectly predicted from the equation $D_4 = 1 - (D_1 + D_2 + D_3)$. So D_4 contains no information about marital status not already contained in the combination of values of D_1, D_2 and D_3 for each case in the data. That means that the multiple correlation between D_4 and the remaining indicator variables is 1, so its tolerance is zero. But recall from section 4.4.4 that regression has "zero tolerance for zero tolerance." This is a *singularity*, and it is not allowed in the regression mathematics.

In fact, every indicator variable will have a tolerance of 0 and a crosswise multiple correlation of 1 if all four indicators are included in the regression equation. To see why, recognize that every indicator variable can be predicted perfectly from the other three dummy variables. For example, $D_1 = 1 - (D_2 + D_3 + D_4)$, $D_2 = 1 - (D_1 + D_3 + D_4)$, and so forth. This means that the standard error of the regression coefficient for every indicator would be infinity as a result of this disturbing and destructive singularity.

The simplest way to avoid this problem is to identify one category as the reference category and omit its indicator variable from the regression model, as we did in the example. Thus, for a multicategorical variable with g categories, we need $g - 1$ indicator variables, not g. We could use any category we want as the reference category; it makes no difference, but some choices may be more convenient than others when it comes to substantive interpretation of the regression coefficients. For example, if the multicategorical variable consists of three categories—two experimental groups and a control group—then it is usually convenient for interpretation to use the control group category as the reference. Or if there is a catchall "other" category, this is often a sensible choice for the reference. But mathematically it makes no difference. The choice will affect only the regression coefficients and the regression constant and their standard errors. The choice of reference won't change how well the model fits the data or the estimates of Y the model generates. Mathematically, regression models using different

TABLE 9.3. Satisfaction with Life in Four Groups

Group (j)	Marital status	$\overline{Y}_j$	s_{Y_j}	n_j
1	Married	85.500	6.380	6
2	Divorced	79.200	5.541	5
3	Single	82.400	6.427	5
4	Widowed	71.500	8.062	4

reference categories will be the same; one is just a *reparameterization* of the other. They are identical but just package the information about differences between the groups in different ways.

9.1.4 Testing the Equality of Several Means

Indicator coding can be used in complex analyses with many regressors, but we first consider its simplest use: to test the equality of the means of a set of groups. For instance, suppose we want to test whether people of different marital status differ, on average, in how satisfied they are with life. Table 9.3 contains the mean and standard deviation of life satisfaction for each of the four marital status groups, using the data in Table 9.1. Defining $\overline{Y}_j$ as the mean life satisfaction for group j and n_j as the sample size in group j, you can see in Table 9.3 that $\overline{Y}_1 = 85.500$, $\overline{Y}_2 = 79.200$, $\overline{Y}_3 = 82.400$, and $\overline{Y}_4 = 71.500$, with the numerical subscripts referring to married, divorced, single, and those who are widowed, respectively. These means are based on sample sizes of $n_1 = 6$, $n_2 = 5$, $n_3 = 5$, and $n_4 = 4$.

From a purely descriptive perspective, it seems that people who are married are most satisfied on average, followed by single people, then divorced people, with those who are widowed being least satisfied. That said, there is quite a bit of variation in satisfaction between people of the same marital status, so it is hard to say just looking at the means whether these differences are statistically significant. We'll address that question in this section using output from the regression analysis in Figure 9.1.

Consider the meaning of the correlation between the three dummy variables used as regressors and Y, which we'll denote as r_{D_1Y}, r_{D_2Y}, and r_{D_3Y}. If you correlated D_1 with Y, you'd get $r_{D_1Y} = 0.438$. This positive value reflects that fact that the mean life satisfaction of 85.500 for the six married people ($D_1 = 1$) is higher than the mean life satisfaction of 78.143

for those who are not married ($D_1 = 0$). Had the married peoples' mean satisfaction been lower, then r_{D_1Y} would be negative. Using this same logic, $r_{D_2Y} = -0.086$, which is negative because the five divorced people are less satisfied on average (79.200) than the 15 who are not divorced (80.733), and $r_{D_3Y} = 0.154$, which is positive because the five single people are more satisfied on average (82.400) than the 15 who are not single (79.667).

Suppose that all four group means were exactly equal in the data, meaning that the four groups all had exactly equal average life satisfaction. In that case, then r_{D_1Y}, r_{D_2Y}, and r_{D_3Y} would all be exactly zero. But if D_1, D_2, and D_3 were the sole predictors in a regression model of Y and they are all uncorrelated with Y, then the multiple correlation would be zero. Thus, R is zero if (and only if) all group means are equal. Any departure from zero for any of the correlations between Y and any of the indicator variables coding groups would translate into $R > 0$.

This argument has been made using a sample of only 20 people, but it applies to the population as well. The true multiple correlation ${}_TR$ equals zero if, and only if, all g population group means are equal. This means that the F-ratio for testing the null hypothesis that ${}_TR = 0$ first introduced in section 4.3.2 can be used as a test of the null hypothesis that the g population means are the same.

Figure 9.1 shows computer output from a regression of life satisfaction on D_1, D_2, and D_3. As discussed in section 4.3.2, the F-ratio in the ANOVA summary table is used to test the null hypothesis that ${}_TR = 0$. In this example, $F(3, 16) = 3.885, p = .029$. That null hypothesis can be rejected. So we can conclude that married people, divorced people, single people, and those who are widowed differ on average in their satisfaction with life. Reframed, we can say that there is a nonzero association between marital status and life satisfaction. Marital status and life satisfaction are related to each other in this sample more than can be explained by just chance.

Earlier it was noted that g indicators cannot be used as regressors in the model, as this will produce a singularity. The group whose indicator is left out of the analysis ends up as the reference category, and it makes no difference which group that is. This also extends to the result of the test of the null hypothesis that ${}_TR = 0$. We could have, for instance, decided to specify married people as the reference category by regressing life satisfaction on D_2, D_3, and D_4. Doing so would produce the same F-ratio and the same p-value (and many other things that are the same). Try this for yourself, and you will see that happily, your conclusion about differences between the groups does not depend on the group you choose.

```
. oneway satis mstatus, tabulate

           |          Summary of satis
   religion |        Mean   Std. Dev.       Freq.
------------+------------------------------------
    married |        85.5   6.3796552           6
   divorced |        79.2   5.5407581           5
     single |        82.4   6.4265076           5
    widowed |        71.5   8.0622577           4
------------+------------------------------------
      Total |       80.35   7.9025312          20

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups           500.05      3   166.683333      3.88      0.0292
 Within groups           686.5      16     42.90625
------------------------------------------------------------------------
    Total              1186.55     19        62.45
```

FIGURE 9.2. STATA output from a one-way ANOVA examining differences between marital status groups in life satisfaction.

9.1.5 Parallels with Analysis of Variance

If you have studied one-way ANOVA, you know that standard ANOVA output contains a summary table that typically has at least 10 entries. Three of these are sums of squares (between, within, and total), three are degrees of freedom (between, within, and total), two are mean squares (between and within), one is an F-ratio, and the remaining one is a p-value. Some statistical programs will include more or less than this, but these 10 are most typical. The SPSS, SAS, and STATA code to conduct a one-way ANOVA testing for differences between these four groups in life satisfaction can be found below.

```
oneway satis by mstatus/statistics descriptive.
```

```
proc anova data=married;class mstatus;model satis = mstatus;
   means mstatus;
run;
```

```
oneway satis mstatus, tabulate
```

The STATA output can be found in Figure 9.2. Output from other programs will contain basically the same information, although the format will of

course be different. Most important for our discussion is the ANOVA summary table containing the sums of squares, mean squares, and so forth. As can be seen, we would reject the null hypothesis of equality of the four group means, $F(3, 16) = 3.88, p = .029$.

Everything in the ANOVA summary table should look familiar, for these are the same statistics generated by a regression analysis using indicator variables. Compare the section of the regression output in Figure 9.1 (page 250) used to test the null that ${}_TR = 0$ to the summary table from the one-way ANOVA in Figure 9.2. As can be seen, $SS_{regression} = SS_{between}$, $SS_{within} = SS_{residual}$, $MS_{between} = MS_{regression}$, the F-ratios are equivalent, and so forth. Mathematically, one-way ANOVA is just a special case of regression analysis with a single multicategorical variable represented with a set of codes for group membership (in this case, indicator variables). If you have a program capable of doing linear regression, you don't need a one-way ANOVA routine. You can do one-way ANOVA with linear regression.

9.1.6 Interpreting Estimated Y and the Regression Coefficients

With g groups, the linear regression equation estimating Y from $g - 1$ indicator variables coding group membership is

$$\hat{Y} = b_0 + b_1 D_i + \cdots + b_{g-1} D_{g-1}$$

In this example, the model is (from Figure 9.1)

$$\hat{Y} = 71.500 + 14.000 D_1 + 7.700 D_2 + 10.900 D_3 \tag{9.1}$$

and thus $b_0 = 71.500$, $b_1 = 14.000$, $b_2 = 7.700$, and $b_3 = 10.900$.

Equation 9.1 can be used to generate an estimate of a person's satisfaction from his or her marital status by plugging in the values of D_1, D_2, and D_3 corresponding to his or her marital status. Each group has a unique combination of values of D_1, D_2, and D_3—unique to that group relative to the other three groups—but *not* unique for each person within a group. Quite the contrary, the pattern of values of D_1, D_2, and D_3 are the same for every person in the same group. Thus, the regression model produces the same estimate of life satisfaction for every person in the same group.

Because there are four groups coded with the indicator variables, there are only four unique patterns of the indicators, so the model generates only four values of $\hat{Y}$, one for each group. These are produced by plugging the corresponding indicator values for each group into equation 9.1, as below:

$$\begin{aligned}
\text{Married:} \quad \hat{Y} &= 71.500 + 14.000(1) + 7.700(0) + 10.900(0) = 85.500 \\
\text{Divorced:} \quad \hat{Y} &= 71.500 + 14.000(0) + 7.700(1) + 10.900(0) = 79.200 \\
\text{Single:} \quad \hat{Y} &= 71.500 + 14.000(0) + 7.700(0) + 10.900(1) = 82.400 \\
\text{Widowed:} \quad \hat{Y} &= 71.500 + 14.000(0) + 7.700(0) + 10.900(0) = 71.500
\end{aligned}$$

Notice that $\hat{Y}$ for each group corresponds to $\overline{Y}$ for that group in Table 9.3. So the best-fitting linear regression model weights each indicator in such a manner that $\hat{Y}$ for cases in group j equals $\overline{Y}_j$.

In this analysis, the widowed group is the reference group because its indicator (D_4) is left out of the model. In this model the regression constant b_0 corresponds to $\overline{Y}$ for this reference category. Indeed, observe that $\hat{Y}$ for this group is $b_0 = 71.500$, which corresponds to $\overline{Y}_4$ in Table 9.3. This is generally true when using indicator coding of a multicategorical variable. The regression constant is $\hat{Y}$ for the group whose indicator is excluded from the model. But we'll see in section 9.2.3 that this is true only if the model doesn't include other regressors.

To see what each regression coefficient b_j equals when using indicator coding, consider in generic form the model we estimated:

$$\hat{Y} = b_0 + b_1 D_1 + b_2 D_2 + b_3 D_3 \tag{9.2}$$

We know that $\hat{Y}$ for each group equals that group's mean of Y. We also know that for every person, except those in the reference category, one D_j is 1 and all the others are 0. Now consider a married person, for whom $D_1 = 1$ and $D_2 = D_3 = 0$. We know that for married people, from equation 9.2, $\hat{Y} = \overline{Y}_1$. We also know that $b_0 = \overline{Y}_4$. Substituting all these values into equation 9.2, we have

$$\begin{aligned}
\overline{Y}_1 &= \overline{Y}_4 + b_1(1) + b_2(0) + b_3(0) \\
\overline{Y}_1 &= \overline{Y}_4 + b_1
\end{aligned}$$

which can be written as

$$b_1 = \overline{Y}_1 - \overline{Y}_4$$

So b_1 corresponds to the difference between $\overline{Y}$ for married people and $\overline{Y}$ for the reference category—the widowed. In these data, $b_1 = 14.000$, which is indeed exactly equal to 85.500 − 71.500 from Table 9.3.

The same reasoning leads to the following derivations

$$b_2 = \overline{Y}_2 - \overline{Y}_4$$
$$b_3 = \overline{Y}_3 - \overline{Y}_4$$

and the general principle:

> When indicator coding is used in a model with no other regressors, b_j equals the difference between the mean of the corresponding group coded with D_j and the mean of the reference group.

The t- and p-values for each b_j can be used to test a hypothesis about the difference between the mean of group j and the reference group, or the standard error for b_j can be used to construct a confidence interval around the population mean difference. This information is available in standard regression output. As can be seen in Figure 9.1, we can say that the mean difference of 14.000 units in life satisfaction between married people and the widowed is statistically significant, $t(16) = 3.311, p = .004$. There is no statistically significant difference in the average life satisfaction of the divorced relative to the widowed, $b_2 = 7.700, t(16) = 1.752, p = .099$. But single people are more satisfied with life, on average, than the widowed ($b_3 = 10.900$ units higher), $t(16) = 2.481, p = .025$

We see now that even though the reference category lacks its own indicator variable in the model, it is not ignored at all. Rather, the reference category provides a standard of comparison for the remaining groups (which is why it is called the reference category). You may also see now why indicator coding can be especially useful when analyzing a study from a design with one control group. By making the control group the reference, the regression analysis gives not only a test of the null hypothesis that all groups are equal on average on the dependent variable, but also a set of specific comparisons between the means of each group and the control condition.

You may be wondering why all this matters, given that we already have ANOVA to handle these problems. The indicator-variable approach is a bit more complex, but we see next that it can easily be used in models with other independent variables or covariates, whether or not they correlate with the multicategorical variable.

9.2 Multicategorical Regressors as or with Covariates

We have introduced the principles of coding multicategorical variables and interpretation of regression analysis results using a simple model with only a single multicategorical independent variable represented with indicator codes. In practice, you will often want to estimate Y from a multicategorical variable as well as additional regressors. Those other variables could be numerical, dichotomous, even additional compound variables—sets of indicator codes for another multicategorical variable. There are at least two major reasons you may find yourself wanting to do this, and we discuss each of these in this section.

9.2.1 Multicategorical Variables as Covariates

In section 3.1.2 we distinguished between an independent variable and a covariate, pointing out that a linear regression algorithm in a computer makes no distinction between them. They are merely regressors in a linear model. Your primary research interest may be in estimating the relationship between a numerical or dichotomous independent variable and Y, but you want to control for a multicategorical variable. That is, you may be thinking of the multicategorical variable as a *covariate* rather than as your primary independent variable of interest. Differences between groups defined by the multicategorical variable on the independent variable or Y may make it harder to interpret the relationship between the independent variable you care about and Y. Statistical control helps to disentangle association that may be causal and association that is due to other processes.

Consider the relationship between life satisfaction (Y) and income (X_2). In the data in Table 9.1, $r_{X_2Y} = 0.755$, which is statistically different from zero, $p < .001$. In regression terms, when satisfaction Y is regressed on income X_2, the model is $\hat{Y} = 44.032 + 0.638X_2$. So two people who differ by \$1,000 in income are estimated to differ by 0.638 units in life satisfaction, with the person with more income estimated as more satisfied with life.

Although establishing association does not establish cause–effect, this relationship is at least consistent with income possibly causing higher satisfaction. But the reverse is also possible of course. Maybe being satisfied leads people to make more money. Alternatively, the association could be *spurious*, due to differences between people with different marital status in their income and satisfaction. Maybe being married or divorced, for example, leads to world views that influence both life satisfaction and income.

Indeed, we have already established that the four groups differ in satisfaction with life. Although a formal test of differences in mean income between the four groups results in a nonsignificant effect, $F(3, 16) = 0.859, p = .482$, the pattern of the means is fairly consistent with the satisfaction differences. In these data, the married people have the highest income (mean = \$60,333) and are also most satisfied , those who are widowed make less on average (mean = \$51,000) and are least satisfied, and the divorced and single are in between on both income (means = \$58,600 and \$55,800, respectively) and satisfaction.

Linear regression analysis will not allow us to distinguish between competing explanations for direction of cause, at least not with data like these. But it can help us establish whether the relationship between income and satisfaction exists among people of the same marital status (i.e., holding marital status constant). We can't enter X_1 into the regression model as is, for reasons discussed in section 9.1.1. But we can include three indicator codes representing the four groups into a model along with income to examine the partial association between income and life satisfaction controlling for marital status. Doing so with a linear regression model yields

$$\hat{Y} = 44.032 + 9.00D_1 + 3.628D_2 + 8.329D_3 + 0.536X_2$$

The regression coefficient for income in this model is 0.536, with a standard error of $0.113, t(15) = 4.728, p < .001$ (see Figure 9.3). Two people that differ by \$1,000 dollars in income but are of the same marital status are estimated to differ in 0.536 units in life satisfaction. This is a smaller difference than when marital status was not controlled, but it is not zero. The relationship observed between income and satisfaction with life is not explained entirely by any income and satisfaction differences between people who differ in marital status. When marital status is held constant, the positive relationship observed absent statistical control persists.

In this analysis (or any comparable analysis), it makes no difference which group is coded as the reference group. Here, the widowed is the reference group. But the regression coefficient for income is unaffected by that choice. Indeed, since we aren't interpreting the regression coefficients for D_1, D_2, and D_3 (though we could and will in section 9.2.2), the idea of a "reference group" really doesn't have much meaning in this application, because we aren't comparing the satisfaction of the different marital status groups in this analysis. Marital status is just a covariate in our interpretation of the results, an interpretation that focuses on the partial association between income and satisfaction.

```
                          The REG Procedure
                            Model: MODEL1
                       Dependent Variable: satis

                 Number of Observations Read          20
                 Number of Observations Used          20


                          Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     4      910.84337      227.71084      12.39    0.0001
Error                    15      275.70663       18.38044
Corrected Total          19     1186.55000


          Root MSE              4.28724    R-Square     0.7676
          Dependent Mean       80.35000    Adj R-Sq     0.7057
          Coeff Var             5.33571


                       Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1       44.17806        6.16407       7.17      <.0001
d1            1        8.99991        2.96263       3.04      0.0083
d2            1        3.62850        3.00215       1.21      0.2455
d3            1        8.32852        2.92696       2.85      0.0123
income        1        0.53572        0.11332       4.73      0.0003
```

FIGURE 9.3. SAS output from a regression examining the effect of income on life satisfaction controlling for marital status.

This approach generalizes to any number of covariates. We could just as easily control for one or more additional multicategorical variables by including their indicators in the model. Or we could control for dichotomous or numerical regressors along with other multicategorical variables. It makes no difference to the regression math how many covariates are included. Regression analysis will result in a regression weight for income, along with information that can be used for inference about the relationship between income and satisfaction, holding all those other variables in the model constant.

9.2.2 Comparing Groups and Statistical Control

A second reason for estimating a model with both a multicategorical variable and additional variables—dichotomous, numerical, or multicategorical in any combination—is to examine whether mean differences in Y between groups defined by the independent multicategorical variable exist after statistically equating those groups on other variables in the model. For example, we might want to know whether three different forms of treat-

ment for PTSD result in differences in the number of symptoms experienced 6 months later after adjusting for differences between the therapy groups in the severity of the traumatic experience suffered, age of the person, and support from the person's spouse or family.

This kind of adjustment or statistical control is particularly important when groups aren't constructed through random assignment. But as discussed in section 6.3.1, statistical control can be useful even in experiments with random assignment to groups. We know random assignment will tend to equate the groups on all other variables at the time point of random assignment, including potential covariates. But when covariates are related to Y but not to the independent variable, adjusting for the covariates can increase the power of tests of group difference and enhance precision in the estimate of those differences.

In section 9.2.1 we examined the relationship between income and life satisfaction, holding marital status constant. But we could also examine differences in average life satisfaction between people who differ in marital status when holding income constant. Recall that in that section, we regressed Y on D_1, D_2, D_3, and X_2 and focused on measures of partial association between X_2 and Y along with inferential tests. The indicator variables in that model were indicator codes representing marital status.

This same model provides estimates of the mean difference in income between pairs of marital status groups if they were equal in income. In generic form, the model is

$$\hat{Y} = b_0 + b_1 D_1 + b_2 D_2 + b_3 D_3 + b_4 X_2 \tag{9.3}$$

We discussed earlier that $b_4 = 0.536$, which is statistically different from zero. But we don't care about that now. We are now thinking of income (X_2) as a covariate and marital status as the independent variable. Our interest now is the differences between the groups in satisfaction when income is held constant, not the association between income and satisfaction when marital status is held constant. Our focus therefore shifts away from b_4 toward b_1, b_2, b_3, and related statistics and tests. Figure 9.4 provides output from SPSS for the entire model, as well as additional regression analysis output, generated with the command on page 264. This output provides the context for the discussion that follows.

In section 5.3.1 we introduced the partial and semipartial multiple correlations, which quantify the association between a set of variables B and some outcome when controlling for a set of one of more covariates A. Inference about the variables in set B can be conducted by examining whether

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.755[a]	.571	.547	5.320	.571	23.917	1	18	.000
2	.876[b]	.768	.706	4.287	.197	4.240	3	15	.023

a. Predictors: (Constant), income

b. Predictors: (Constant), income, d3, d2, d1

$SR(B.A)^2 = 0.197, F(3,15) = 4.240, p = .023$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	677.028	1	677.028	23.917	.000[b]
	Residual	509.522	18	28.307		
	Total	1186.550	19			
2	Regression	910.843	4	227.711	12.389	.000[c]
	Residual	275.707	15	18.380		
	Total	1186.550	19			

a. Dependent Variable: satis

b. Predictors: (Constant), income

c. Predictors: (Constant), income, d3, d2, d1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
		B	Std. Error	Beta			
1	(Constant)	44.032	7.521		5.855	.000	MODEL *A*
	income	.638	.131	.755	4.891	.000	
2	(Constant)	44.178	6.164		7.167	.000	
	income	.536	.113	.634	4.728	.000	
	d1	9.000	2.963	.535	3.038	.008	MODEL *AB*
	d2	3.628	3.002	.204	1.209	.246	
	d3	8.329	2.927	.468	2.845	.012	

a. Dependent Variable: satis

FIGURE 9.4. SPSS output from a regression examining the effect of marital status on life satisfaction controlling for income.

the fit of the model of Y improves when the variables in set B are added to the model of Y that includes set A variables. This is exactly the problem we confront here. Think of income as set A containing only one variable, X_2, and marital status as set B, which includes three indicator variables D_1, D_2, and D_3 coding groups. If the groups don't differ in average life satisfaction when income is held constant, this means that D_1, D_2, and D_3 add no information about variability in life satisfaction relative to what is already explained by income. But if the groups differ in life satisfaction when income is held constant, then including information about which marital status group a person belongs in should improve the fit of the model.

So we have two models of Y we are comparing here. One of them includes income, X_2, as the sole predictor. Call this model A. As already discussed, estimation of that model yields

$$\hat{Y} = 44.032 + 0.638X_2$$

with $R^2(A) = 0.571$. The second model includes income as well as marital status, represented by three indicator codes. Call this model AB. Estimation of this model yields:

$$\hat{Y} = 44.178 + 9.000D_1 + 3.628D_2 + 8.329D_3 + 0.536X_2$$

with $R^2(AB) = 0.876$. Observe that this is the same as the model from the analysis described in section 9.2.2 when marital status was conceptualized as a covariate rather than the independent variable (remember that the regression math makes no distinction). The difference between $R^2(AB)$ and $R^2(A)$ we defined in section 5.3.1 as the squared semipartial multiple correlation between Y and B controlling for A: $SR^2(B.A) = R^2(AB) - R^2(A) = 0.768 - 0.571 = 0.197$. This is also called the *incremental change in* R^2, sometimes denoted ΔR^2. It is an estimate of ${}_TSR^2(B.A)$, the proportion of the variance in Y uniquely attributable to the variables in set B when the variables in set A are held constant.

To test the null hypothesis that ${}_TSR^2(B.A) = 0$ against the alternative hypothesis that ${}_TSR^2(B.A) > 0$, use equation 5.2, which yields an F-ratio that can be compared to the critical F in an F-table. Alternatively, let a computer do the work for you. In SPSS, the command below produced the output in Figure 9.4, which generates both models, as well as a test of difference in the fit of the two models.

```
regression/statistics defaults change/dep=satis/method=enter income
   /method=enter d1 d2 d3.
```

Comparable code in SAS and STATA to conduct the test would be

```
proc reg data=married;
   model satis=income d1 d2 d3;
   test d1=0,d2=0,d3=0;run;
```

```
regress satis income d1 d2 d3
test d1 d2 d3
```

As can be seen in Figure 9.4 in the section highlighted with the dashed box, $SR^2(B.A) = \Delta R^2 = 0.197$, $F(3, 15) = 4.240, p = .023$. We can reject the null hypothesis. The interpretation is that when you account for differences between the four groups in their average income, the groups differ from each other in average satisfaction in life to a statistically significant degree. That is, adding information about marital status to information about income provides some additional information about a person's life satisfaction relative to when marital status is just ignored entirely.

This latter interpretation relates to an alternative way of thinking about the null hypothesis we are testing. If there is no difference between the g groups on average on Y when the covariate is held constant, then this implies that all of the $g - 1$ true regression weights for the indicator codes representing the g group are all zero. That is, when the variables in set B represent membership in one of the g groups and set A represents one (or more variables being held constant, ${}_TSR^2(B.A) = 0$ implies that all $g - 1$ values of ${}_Tb_{D_j} = 0$. But if ${}_TSR^2(B.A) > 0$, this implies that at least one ${}_Tb_{D_j} \neq 0$. So by rejecting the null hypothesis that ${}_TSR^2(B.A) = 0$, we can also reject the null hypothesis that all $g - 1$ values of ${}_Tb_{D_j} = 0$.

9.2.3 Interpretation of Regression Coefficients

In section 9.2.1 we interpreted b_4, which is also the regression coefficient for income in model AB from section 9.2.2, as the estimated difference in satisfaction people of the same marital status who differ by \$1,000 in income. But we have not yet interpreted the regression coefficients for D_1, D_2, and D_3 or the regression constant b_0. We do so now.

In section 9.1.6, before covariates were introduced in our discussion of multicategorical regressors, the regression coefficient for D_j when using indicator coding of groups was interpreted as the difference in the mean

of Y between the group coded with D_j and the mean of the reference group. That interpretation still applies, but with the addition of *holding the covariate(s) constant* or *statistically controlling for the covariate(s)*. For example, suppose we imagine a group of people with the same income of $50,000 but who are of different marital status. Two people with different marital status will differ in their pattern of values on D_1, D_2, and D_3, but they will be the same on X_2, which we have fixed at 50. Consider a married person (any married person) who makes $50,000. The model

$$\hat{Y} = 44.032 + 9.000D_1 + 3.628D_2 + 8.329D_3 + 0.536X_2$$

generates

$$\begin{aligned}\hat{Y} &= 44.178 + 9.000(1) + 3.628(0) + 8.329(0) + 0.536(50)\\ &= 44.178 + 9.000(1) + 0.536(50)\\ &= 79.978\end{aligned}$$

as his or her estimated satisfaction with life. But for someone who is widowed, the model generates

$$\begin{aligned}\hat{Y} &= 44.032 + 9.000(0) + 3.628(0) + 8.329(0) + 0.536(50)\\ &= 44.178 + 0.536(50)\\ &= 70.978\end{aligned}$$

as his or her estimated satisfaction with life. The difference between these is 9 satisfaction units, which is the regression coefficient for D_1 in the model. But observe from this math that it makes no difference whatsoever what value of income you use in the model. Regardless of the choice, the difference between the estimates of Y for these two groups is 9 units. So b_1 quantifies the estimated difference in Y between the group set to 1 on D_1 and the reference group when X_2 is held constant. A hypothesis test or confidence interval leads to a corresponding inference about the difference between the true means. As can be seen in Figure 9.4, this estimated difference of $b_1 = 9.000$ satisfaction units is statistically different from zero, $t(15) = 3.038, p = .008$. It seems that married people and those who are widowed with the same income differ in satisfaction, with married people more satisfied by an estimated 9 units on the scale.

Regression coefficients b_2 and b_3 can be interpreted similarly. The regression coefficient for D_2 is $b_2 = 3.828$. This is the estimated difference in life satisfaction between a group of divorced people and the widowed

with the same income. This difference of 3.828 units in satisfaction is not statistically different from zero, $t(15) = 1.209, p = .246$. But single people and the widowed with the same income do differ in average satisfaction, with the singles estimated as more satisfied on average. The estimated difference in satisfaction is $b_3 = 8.329$ units, $t(15) = 2.845, p = .012$.

The inclusion of an additional regressor in the model as a covariate has changed the interpretation of b_0 relative to a model without covariates. In section 9.1.6, b_0 was $\overline{Y}$ for the reference group—the group set to 0 on all $g - 1$ indicator variables. To see what b_0 now quantifies, consider that for anyone in the reference group (widowed), $D_1 = D_2 = D_3 = 0$, so the model expressed in equation 9.3 becomes

$$\begin{aligned}\hat{Y} &= b_0 + b_1(0) + b_2(0) + b_3(0) + b_4X_2 \\ &= b_0 + b_4X_2\end{aligned}$$

which means that $\hat{Y}$ equals b_0 when $X_2 = 0$. So b_0 is the estimate of Y for the reference group when the covariate equals zero. In terms of this example, b_0 estimates the average life satisfaction of people with no income who are widowed.

9.2.4 Adjusted Means

When we have a categorical independent variable and one or more other regressors, *adjusted means* are the estimated means on Y of the various groups for people who are at the mean on all other regressors. In a sense, when comparing groups while statistically controlling for a covariate, one is essentially testing the equality of the adjusted means, and when interpreting differences between groups after adjusting for the covariate, it is conventional to base the interpretation on the adjusted means rather than on the unadjusted means in Table 9.3.

The adjusted means are calculated from the regression model by plugging in each group's pattern of values on the indicator codes into the regression equation while setting the covariate to the sample mean, disregarding group. In this example, the mean income is $\overline{X}_2 = 56.900$. So the adjusted mean life satisfaction for married people ($D_1 = 1, D_2 = 0, D_3 = 0$) is

$$\text{adj. } \overline{Y}_1 = 44.178 + 9.000(1) + 3.628(0) + 8.329(0) + 0.536(56.9) = 83.676$$

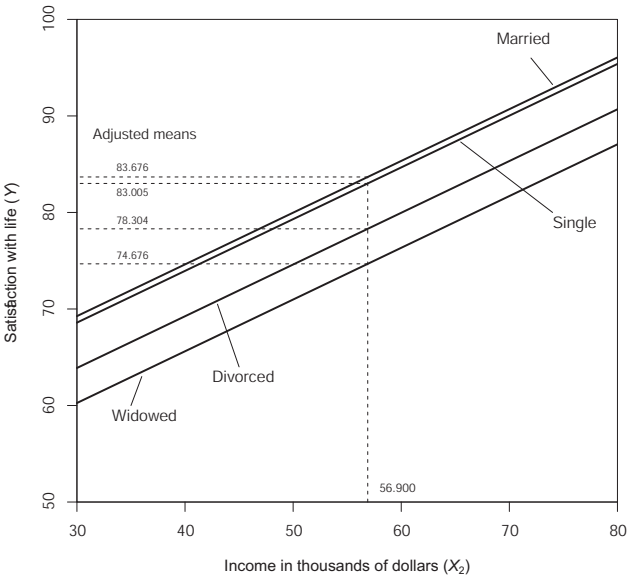


FIGURE 9.5. A visual representation of adjusted means.

The adjusted means for divorced, single, and widowed are calculated similarly:

$$\text{adj. } \overline{Y}_2 = 44.178 + 9.000(0) + 3.628(1) + 8.329(0) + 0.536(56.9) = 83.005$$
$$\text{adj. } \overline{Y}_3 = 44.178 + 9.000(0) + 3.628(0) + 8.329(1) + 0.536(56.9) = 78.304$$
$$\text{adj. } \overline{Y}_4 = 44.178 + 9.000(0) + 3.628(0) + 8.329(0) + 0.536(56.9) = 74.676$$

These adjusted means are represented graphically in Figure 9.5. This figure depicts the regression model, which as can be seen is four parallel lines relating the covariate to Y. The slope of each of these lines is the regression coefficient for the covariate, which in this example is 0.536. The adjusted mean for a group is found by projecting from the sample mean of the covariate up to a group's line relating the covariate to Y and then horizontally across to the Y-axis. The point at which the projection crosses the Y-axis is that group's adjusted mean.

Whereas the common slope for the lines in Figure 9.5 corresponds to the regression coefficient for the covariate, the vertical distances between

certain pairs of lines correspond to the regression coefficients for the indicators. For example, the vertical distance between the line for married and the line for widowed is 9.000 units, which corresponds to b_1. Observe that this is also equivalent to the difference in the adjusted means for these two groups, since these lines differ by 9.000 units in distance at the mean of the covariate (and, indeed, any other point on the covariate scale). Similarly, $b_2 = 3.628$ is the distance between the married line and the widowed line, or the difference between the adjusted means of these two groups; b_3 is interpreted equivalently for the comparison between single and widowed.

In principle, one could construct adjusted means by setting the covariate to any desired value, but it is conventional to use the sample mean for the adjustment. For instance, one could construct each group's adjusted mean if the average income of each group was equal to \$40,000 by using $X_2 = 40$ in the computations above rather than the sample mean. This would not influence the relative differences between the adjusted group means, but rather would merely shift them all up or down in value by the same constant amount. This can be seen easily in Figure 9.5 by imagining where the dotted lines would cross the Y-axis if the line projected up from $X_2 = 40$ rather than 56.9. Of course, the vertical distance between the lines would not change, meaning that the regression coefficients for the indicators can be interpreted as the difference between group-adjusted means regardless of the value of the covariate.

9.2.5 Parallels with ANCOVA

You may already be familiar with ANCOVA, which is an extension of ANOVA that allows for the comparison between groups while holding one or more covariates fixed. In this section we show that the test described in section 9.2.2 generates the same result as one would get from an ANCOVA. An important point we make is that this is neither coincidental nor surprising, because mathematically they are the same test.

Figure 9.6, panel A, contains an ANCOVA summary table from STATA examining differences between marital status groups in their satisfaction controlling for income. This output was generated with the command

```
anova satis mstatus c.income
```

Comparable summary tables can be obtained from SPSS and SAS with the commands

```
unianova satis by mstatus with income.
```

```
proc glm data=married;
class mstatus;model satis=mstatus income;run;
```

This summary table looks similar to the one in Figure 9.2, but it contains an additional line for income. Observe that each of the two effects (the effect of marital status and the effect of income) has a corresponding sum of squares, degrees of freedom, mean square, *F*-ratio, and *p*-value. Books that discuss ANCOVA often provide complex formulas for deriving the entries in this table. But ANCOVA is just a form of multiple regression, and all these entries can be generated with a regression analysis program. We focus on derivation of the sums of squares for the effects, because many of the rest of the entries are functions of these sums of squares.

Suppose we estimated satisfaction from income and marital status, with marital status represented with three indicator codes. This model has four regressors, and $SS_{regression}$ = 910.843, $SS_{residual}$ = 275.707, which add up to SS_{total} = 1186.550. These sums of squares can be found in the ANCOVA summary table, and they can also be found in the regression analysis output in Figure 9.3 from section 9.2.1.

Figure 9.6, panel B, visualizes these quantities in the form of a Venn diagram. The area of the Y circle corresponds to SS_{total}. The area in the Y circle is the sum of the areas labeled A, B, C, and D, so we can also say that A + B + C + D = 1186.550. This is all of the variance in life satisfaction available to be explained. The variance in Y that is explained by both marital status and income is the area of overlap between the life satisfaction circle and the income and marital status circles, or A + B + C. This corresponds to $SS_{regression}$ = 910.843, denoted $SS_{Mstatus+Income}$ in Figure 9.6. The variance in Y not accounted for by income and marital status is D, or $SS_{residual}$ = 275.707.

Now consider two regression models of life satisfaction Y, one that contains only marital status as its sole predictor and one that contains only income as its sole predictor. The regression and residual sums of squares for these two models we denote $SS_{Mstatus}$ and SS_{Income}, respectively, in Figure 9.6. In terms of the Venn diagram, the regression sum of squares for the model with marital status as the only predictor is the sum of areas A + B = 500.050. Similarly, the regression sum of squares for the model with income as the only predictor is the sum of areas B + C = 677.028.

We know that when both marital status and income are in the model, $SS_{regression} = SS_{Mstatus+Income}$ = A + B + C = 910.843. If martial status were

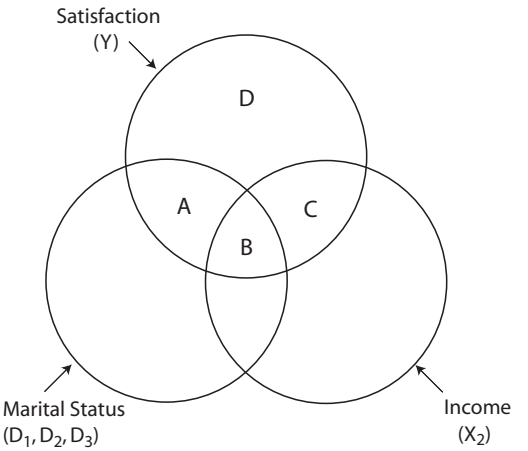
A

```
. anova satis mstatus c.income

      Number of obs =      20     R-squared     =  0.7676
      Root MSE      = 4.28724     Adj R-squared =  0.7057

     Source |  Partial SS    df       MS           F     Prob > F
------------+----------------------------------------------------
      Model |  910.843367     4  227.710842      12.39     0.0001
            |
   mstatus |  233.815753     3  77.9385843       4.24     0.0234
    income |  410.793367     1  410.793367      22.35     0.0003
            |
  Residual |  275.706633    15  18.3804422
------------+----------------------------------------------------
     Total |     1186.55    19       62.45
```

B



$SS_{Total} = A + B + C + D = 1186.550$

$SS_{Mstatus + Income} = A + B + C = 910.843$

$SS_{Mstatus} = A + B = 500.050$

$SS_{Income} = B + C = 677.028$

$SS_{Mstatus.Income} = A = 233.816$

$SS_{Income.Mstatus} = C = 410.793$

$SS_{Residual} = D = 275.707$

FIGURE 9.6. STATA output from a one-way ANCOVA examining differences between marital status groups in life satisfaction with income as a covariate (panel A) and sums of squares in Venn diagram form (panel B).

removed from the model, $SS_{regression}$ would drop from A + B + C = 910.843 to B + C = SS_{Income} = 677.028. This difference corresponds to area A in the Venn diagram, which we denote $SS_{Mstatus.Income}$ in Figure 9.6. In sum of squares terms, the area of A is 910.843 − 677.028 = 233.815. This is the sum of squares for marital status in the ANCOVA summary table in Figure 9.6. Dividing 233.815 by the number of regressors in the model representing marital status (the three indicator variables, which is the degrees of freedom for this effect), yields a mean square of 77.939 and an *F*-ratio of 4.240 when this mean square is divided by 18.380, the mean squared residual for the model containing both income and marital status. These can also be found in the ANCOVA summary table.

But notice that this *F*-ratio of 4.240 corresponds exactly to the *F*-statistic from the test that the squared semipartial multiple correlation for marital status equals zero discussed in section 9.2.2 and corresponding regression analysis output in Figure 9.4. This should make sense, because area A corresponds to the proportion of the area in Y (variance available to be explained in life satisfaction) that is uniquely attributable to marital status, meaning the amount the squared multiple correlation in a model estimating income increases when marital status is added to a model of life satisfaction that already contains income as a predictor. That quantity we have defined as a squared semipartial multiple correlation. So the test discussed in section 9.2.2 is mathematically equivalent to the results from an ANCOVA.

A similar mathematical logic results in the part of the regression sum of squares in the model with income and marital status as regressors that is uniquely attributable to income. This is area C in the Venn diagram. It is equal to $SS_{Mstatus+Income}$, from the model with both income and marital status (A + B + C = 910.843), minus $SS_{Mstatus}$, which is $SS_{regression}$ from a model with just marital status as a predictor (A + B = 233.816). That difference is 410.793, which we denote $SS_{Income.Mstatus}$ in Figure 9.6, which is the sum of squares for income in the ANCOVA summary table. The mean square and *F*-ratio follows from the same mathematics.

9.2.6 More Than One Covariate

All of this discussion about comparing groups while holding another variable constant applies without modification to more than one covariate. To test for differences between g groups on Y when holding constant more than one variable, simply include those other variables in the regression model along with the $g-1$ indicator variables coding group. A comparison of the fit of a linear model of Y that includes the covariates and the $g-1$

indicators relative to one that excludes those indicators provides a test of the null hypothesis of equality of the group means when all the covariates are held constant.

For example, we can compare the mean satisfaction of the four marital status groups while holding income and sex (X_3 in Table 9.1) constant. The resulting model is

$$\hat{Y} = 42.200 + 10.318D_1 + 4.140D_2 + 8.949D_3 + 0.575X_2 - 2.016X_3 \quad (9.4)$$

with $R^2 = 0.777$. When the three marital status indicators are excluded from the model, $R^2 = 0.584$. The difference in these squared multiple correlations is $\Delta R^2 = 0.193$, which is the squared semipartial multiple correlation between marital status and life satisfaction when sex and income are statistically controlled. Marital status uniquely explains 19.3% of the variance in life satisfaction. This is statistically different from zero, $F(3, 14) = 4.042, p = .029$, using the test described in section 9.2.2.

The adjusted means are calculated from equation 9.4 for each group using each group's pattern of values of D_1, D_2, and D_3 in the model and substituting the mean income $\overline{X}_2 = 56.900$ and "mean sex" ($\overline{X}_3 = 0.450$) for X_2 and X_3, respectively. Doing so yields $\overline{Y}_1 = 84.328$, $\overline{Y}_2 = 78.150-74.010 =$, $\overline{Y}_3 = 82.959$, and $\overline{Y}_4 = 74.010$ for married, divorced, single, and widowed, respectively. Using the mean of sex in equation 9.4 may seem strange given that the numerical codes of 0 and 1 for females and males are arbitrary. But doing so is completely legitimate mathematically. If this bothers you, you can compute the adjusted means for males and for females by repeating these computations, using $X_3 = 0$ for females and $X_3 = 1$ for males instead of $X_3 = 0.450$. This will generate eight adjusted means, four for males of different marital status and four for females of different marital status.

The regression constant $b_0 = 42.200$ estimates the mean life satisfaction of males who are widowed with no income. The regression coefficient $b_4 = 0.575$ is the estimated average difference in life satisfaction between two people of the same sex and marital status who differ by \$1,000 in income, $t(14) = 4.571, p < .001$, while $b_5 = -2.016$ is the estimated average difference in life satisfaction between men and women equal in income and of the same marital status, but this is not statistically significant, $t(14) = 0.762, p = .459$. The remaining regression coefficients for the three indicator variables estimate the mean difference in life satisfaction between the reference group of widowed and the group coded with that indicator, holding income and sex constant. So married people of a given income are more satisfied with life than those who are widowed of that same sex and income by $b_1 = 10.318$

units, $t(14) = 2.976, p = .010$, and single people are more satisfied on average than those who are widowed by $b_3 = 8.949$ units, $t(14) = 2.907, p = .011$, holding sex and income fixed. Among divorced people and widowed of the same sex and income, there is no statistically significant difference in average life satisfaction, $b_2 = 4.140, t(14) = 1.328, p = .205$.

9.3 Chapter Summary

Multicategorical variables—variables that are categorical with more than two categories—can be used as regressors in a linear regression model when properly represented using some kind of coding system. This chapter introduced indicator coding as way of representing a multicategorical variable. A test of the null hypothesis that the g groups don't differ on average on Y can be conducted by testing whether ${}_TR = 0$ when Y is regressed on the set of group codes. This test is the regression analysis equivalent of a one-way ANOVA, and regression and ANOVA will produce identical results. But regression analysis is much more flexible and versatile than ANOVA, which assumes that all variables are categorical. Regression analysis can be used with any mix of continuous, dichotomous, or multicategorical variables.

When an analyst seeks to compare more than two groups on some dependent variable that may also differ on other variables, ANCOVA is frequently used. But ANCOVA is just a special form of linear regression analysis with group as the independent variable and the other variables treated as covariates. In that case, a test of multivariate partial association between indicator variables coding group and Y provides a test of the difference between group means if the groups were equal on all covariates.

In the next chapter, we show that indicator coding is only one of many ways of representing groups. Different systems of coding groups exist that provide not only tests of differences between groups, as with indicator coding, but also allow the analyst to address questions involving association between an ordinal multicategorical variable and a dependent variable, either with or without statistical control.

10
More on Multicategorical Regressors

This chapter builds on the discussion of multicategorical regressors in Chapter 9 by introducing several additional methods of coding groups. Two of these methods, sequential and Helmert coding, are particularly useful when the multicategorical regressor is categorical and ordinal. We also discuss statistical tests of complex contrasts of means, both with and without covariates.

We described in Chapter 9 how a categorical variable representing $g \geq 3$ groups can be used as a regressor in a linear model if it is represented with $g - 1$ indicator variables. The gth group does not require its own indicator because it would not contain any information about group membership not already contained in the $g - 1$ indicators in the model. When using indicator coding, the group not given an indicator serves as the reference group, and regression coefficients for the indicators quantify the difference in Y between the group coded with an indicator and the reference group.

Using this system of coding groups, regression analysis can be used to compare g group means either with or without additional variables in the model serving as covariates. In this chapter, we discuss some other methods of coding groups that produce mathematically identical models, in that they fit just as well and produce the same estimates of Y, yet yield regression coefficients with different interpretations. We also introduce some methods for conducting complex contrasts between group means, formed by combining group means together in various ways to test whether one set of group means, when aggregated, differ from another group mean or set of means.

TABLE 10.1. Age and Willingness to Self-Censor

ID	Age cohort	`cohort`	`wtsc` (Y)
1	Baby boomer	3	2.75
2	Pre-baby boomer	4	3.50
3	Pre-baby boomer	4	2.75
4	Baby boomer	3	2.25
5	Generation X	2	4.00
⋮	⋮	⋮	⋮
457	Baby boomer	3	2.87
458	Pre-baby boomer	4	2.75
459	Generation X	2	3.00
460	Baby boomer	3	2.50
461	Generation Y	1	2.75

10.1 Alternative Coding Systems

Indicator coding is only one of many ways of representing a multicategorical variable in a linear regression model. Two of these alternatives, *sequential coding* and *Helmert coding*, are particularly useful when the groups can be ordered relative to each other on the variable used to define the groups (though these two coding methods can be used for strictly nominal categories as well). Another alternative called *effect coding* is similar to indicator coding but changes the reference against which the groups are compared.

We rely on a data set containing the responses of 461 people living in the United States and the United Kingdom to a set of questions on a survey administered through the Internet. An excerpt from the data file (named WTSC and downloadable from this book's web page at *www.afhayes.com*) can be found in Table 10.1. The variable in the column labeled `wtsc` is scores on an instrument called the Willingness to Self-Censor Scale (Hayes, Glynn, & Shanahan, 2005). This instrument measures how reluctant versus willing a person is to express his or her opinion publicly when the person believes others hold a different opinion. Higher scores reflect a greater willingness to self-censor one's opinion expression. This is the dependent variable Y in all analyses in this chapter.

The data set also contains an ordinal categorical variable coding a respondent's age named `cohort`. The data were returned from the data collection company with each respondent classified into one of four age

TABLE 10.2. Willingness to Self-Censor in Four Age Cohorts

Group (j)	Age cohort	$\overline{Y}_j$	SD_{Y_j}	n_j
1	Generation Y (born after 1985)	3.201	0.494	38
2	Generation X (born 1966 – 1985)	3.111	0.622	149
3	Baby boomer (born 1945 – 1965)	2.857	0.468	173
4	Pre-baby boomer (born before 1945)	2.802	0.454	101

cohorts. Ordinally lowest in age is "Generation Y," the youngest group and born after 1985, and is coded cohort = 1 in the data. The data collection occurred in 2009, so all Generation Y respondents were 23 years old or younger (no one under 18 participated in the study). Following Generation Y is Generation X, coded cohort = 2, a group containing people born between 1966 and 1985 and thus between the ages of 24 and 43. Next comes the baby boomers born between 1945 and 1965 (cohort = 3, between 44 and 64 years old). The ordinally highest group in age is the pre-baby boomers. They were born before 1945, all at least 65 years old, and coded cohort = 4. Thus, in terms of age, pre-baby boomers > baby boomers > Generation X > Generation Y.

Each group's mean willingness to self-censor can be found in Table 10.2. As can be seen, it appears that the relationship between age and willingness to self-censor is negative, as successive increments up the ordinal age scale correspond to a lower mean willingness to self-censor.

Let's regress willingness to self-censor on age cohort using the indicator coding system introduced in Chapter 9 at the top of Table 10.3. This system codes Generation Y, Generation X, and baby boomers with D_1, D_2, and D_3, and pre-baby boomers are the reference category. The resulting model can be found at the top of Table 10.4. You can verify for yourself that this model generates the group means as its estimates for Y for the four groups. A test of the null hypothesis that ${}_TR = 0$ can be rejected, $F(3, 457) = 12.207, p < .001$. That is, the four age groups differ in their average willingness to self-censor.

10.1.1 Sequential (Adjacent or Repeated Categories) Coding

Sequential coding would most typically be used when the groups can be ordered on the variable that defines them and interest is in examining

TABLE 10.3. Four Ways of Coding Age Cohort and the Group Means Defined in Terms of the Regression Coefficients and Regression Constant

Age cohort by increasing age	D_1	D_2	D_3	Mean of Y
Indicator coding				
Generation Y	1	0	0	$\overline{Y}_1 = b_0 + b_1$
Generation X	0	1	0	$\overline{Y}_2 = b_0 + b_2$
Baby boomer	0	0	1	$\overline{Y}_3 = b_0 + b_3$
Pre-baby boomer	0	0	0	$\overline{Y}_4 = b_0$
Sequential coding				
Generation Y	0	0	0	$\overline{Y}_1 = b_0$
Generation X	1	0	0	$\overline{Y}_2 = b_0 + b_1$
Baby boomer	1	1	0	$\overline{Y}_3 = b_0 + b_1 + b_2$
Pre-baby boomer	1	1	1	$\overline{Y}_4 = b_0 + b_1 + b_2 + b_3$
Helmert coding				
Generation Y	−3/4	0	0	$\overline{Y}_1 = b_0 - \frac{3}{4}b_1$
Generation X	1/4	−2/3	0	$\overline{Y}_2 = b_0 + \frac{1}{4}b_1 - \frac{2}{3}b_2$
Baby boomer	1/4	1/3	−1/2	$\overline{Y}_3 = b_0 + \frac{1}{4}b_1 + \frac{1}{3}b_2 - \frac{1}{2}b_3$
Pre-baby boomer	1/4	1/3	1/2	$\overline{Y}_4 = b_0 + \frac{1}{4}b_1 + \frac{1}{3}b_2 + \frac{1}{2}b_3$
Effect coding				
Generation Y	1	0	0	$\overline{Y}_1 = b_0 + b_1$
Generation X	0	1	0	$\overline{Y}_2 = b_0 + b_2$
Baby boomer	0	0	1	$\overline{Y}_3 = b_0 + b_3$
Pre-baby boomer	−1	−1	−1	$\overline{Y}_4 = b_0 - b_1 - b_2 - b_3$

TABLE 10.4. Estimating Willingess to Self-Censor from Age Cohort Using the Coding Systems in Table 10.3

		Coeff.	*SE*	*t*	*p*
Indicator coding					
(pre-baby boomers as reference)					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.802	0.052	53.933	< .001
D_1	b_1	0.399	0.099	4.019	< .001
D_2	b_2	0.310	0.067	4.603	< .001
D_3	b_3	0.055	0.065	0.846	.398
Sequential coding					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	3.201	0.085	37.797	< .001
D_1	b_1	−0.090	0.095	−0.944	.346
D_2	b_2	−0.254	0.058	−4.361	< .001
D_3	b_3	−0.055	0.065	−0.846	.398
Helmert coding					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	−0.278	0.089	−3.133	.002
D_2	b_2	−0.282	0.054	−5.240	< .001
D_3	b_3	−0.055	0.065	−0.846	.398
Effect coding					
(pre-baby boomers uncoded)					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	0.208	0.066	3.133	.002
D_2	b_2	0.119	0.042	2.841	.005
D_3	b_3	−0.136	0.040	−3.376	.001

TABLE 10.5. Sequential Coding of g Categories

Group	D_1	D_2	D_3	$\cdots$	D_{g-1}
1	0	0	0	$\cdots$	0
2	1	0	0	$\cdots$	0
3	1	1	0	$\cdots$	0
4	1	1	1	$\cdots$	0
$\vdots$					
g	1	1	1	$\cdots$	1

how $\overline{Y}_j$ changes as the ordinal predictor variable increases by one step. Like indicator coding, sequential coding relies on dummy variables. When using sequential coding with g groups, we set D_j to 1 for cases that are members of a group ordinally higher than position j on the variable defining groups; otherwise, we set to D_j to 0.

Table 10.5 provides a general representation of sequential coding with g ordered groups, and Table 10.3 provides the sequential codes for coding four groups as in this example. In this case, we set D_1 to 1 for anyone older than Generation Y, and Generation Y gets $D_1 = 0$. Moving up the ordinal age scale, D_2 is set to 1 for anyone older than Generation X, and Generations X and Y receive $D_2 = 0$. Finally, anyone older than the baby boomers (the pre-baby boomers) receives a code of $D_3 = 1$, and all others get 0 on D_3.

Regressing willingness to self-censor on the set of three sequential codes yields the regression model in Table 10.4. As can be seen, the model is

$$\hat{Y} = 3.201 - 0.090D_1 - 0.254D_2 - 0.055D_3$$

and has exactly the same R (and thus the same $SS_{residual}$ and other measures of fit) as when indicator coding was used. The outcome of the test as to whether ${}_TR = 0$ is the same as well, with the same F-ratio, degrees of freedom, and p-value. We can reject the null hypothesis and conclude that the groups differ in their average willingness to self-censor. Furthermore, plugging values of D_1, D_2, and D_3 into the model generates the four group means, just as does the model based on indicator coding:

$$\begin{aligned}
\hat{Y}_1 &= 3.201 - 0.090(0) - 0.254(0) - 0.055(0) = 3.201 = \overline{Y}_1 \\
\hat{Y}_2 &= 3.201 - 0.090(1) - 0.254(0) - 0.055(0) = 3.111 = \overline{Y}_2 \\
\hat{Y}_3 &= 3.201 - 0.090(1) - 0.254(1) - 0.055(0) = 2.857 = \overline{Y}_3 \\
\hat{Y}_4 &= 3.201 - 0.090(1) - 0.254(1) - 0.055(1) = 2.802 = \overline{Y}_4
\end{aligned}$$

So mathematically, this model is the same as the model based on indicator coding of groups. It produces the same estimates of Y, it fits identically, and it yields the same p-value when testing the null hypothesis that ${}_TR = 0$.

But there is an obvious difference between the two models in the regression coefficients and the constant. This is because these now quantify something different. Recall that with indicator coding, b_0 is $\overline{Y}$ for the reference group, and b_j is the mean difference in Y between the group receiving 1 on D_j and the reference group. But in sequential coding, b_0 is $\overline{Y}$ for the group ordinally lowest on the variable defining the groups, and b_j is the mean difference in Y between the group in ordinal position j and the group one ordinal position *lower*. In other words, b_j is the difference in means between categories that are ordinally adjacent on the variable defining groups. And the t- and p-value tests the null hypothesis that these two means are equal.

To see how this works, consider that for Generation Y,

$$\begin{aligned}
\overline{Y}_1 &= b_0 + b_1 0 + b_2 0 + b_3 0 \\
\overline{Y}_1 &= b_0
\end{aligned}$$

and for Generation X,

$$\begin{aligned}
\overline{Y}_2 &= b_0 + b_1 1 + b_2 0 + b_3 0 \\
\overline{Y}_2 &= b_0 + b_1.
\end{aligned}$$

But $b_0 = \overline{Y}_1$ and so

$$\overline{Y}_2 = \overline{Y}_1 + b_1$$

which can be rewritten as

$$b_1 = \overline{Y}_2 - \overline{Y}_1$$

So b_1 is the mean difference in Y between the two groups that are ordinally lowest in age. In this example, $b_1 = -0.090$, which is indeed $\overline{Y}_2 - \overline{Y}_1 = 3.111 - 3.201$ (from Table 10.2). These means are not statistically different

from each other, $t(457) = -0.944, p = .346$. Generation X is no more or less willing to self-censor, on average, than Generation Y.

This same reasoning leads to the derivation that b_2 is the mean difference Y between the groups in the second and third ordinal position. In this case, this is the baby boomers versus Generation X. For baby boomers,

$$\overline{Y}_3 = b_0 + b_1 1 + b_2 1 + b_3 0$$
$$\overline{Y}_3 = b_0 + b_1 + b_2$$

but $b_0 + b_1 = \overline{Y}_2$, and so

$$\overline{Y}_3 = \overline{Y}_2 + b_2$$

and isolation of b_2 results in

$$b_2 = \overline{Y}_3 - \overline{Y}_2.$$

In this example, $b_2 = -0.254$, which is $\overline{Y}_3 - \overline{Y}_2 = 2.857 - 3.111$. Baby boomers are less willing to self-censor, on average, than Generation X, $t(457) = -4.361, p < .001$. Following this same logic leads to the conclusion that pre-baby boomers do not differ significantly from baby boomers, on average, in their willingness to self-censor, $b_3 = -0.055, t(457) = -0.846, p = .398$. Observe that $b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857$.

It should be apparent why sequential coding can also be called *adjacent categories* coding. It would be the coding system to use if you are interested in comparing how Y changes with incremental increases in the ordinal multicategorical predictor represented with the $g-1$ dummy variables. This could be especially useful when the g categories can be ranked on some a priori basis on some dimension such as cost or difficulty in implementation. For instance, perhaps five drugs differ in the amount they cost. Each b_j quantifies the increase in Y associated with each additional cost increase, and hypothesis tests formally examine whether the increase in Y associated with an additional step up in cost is statistically significant.

It might be apparent already to you that sequential coding does not require that the multicategorical variable be ordinal. It could be used for a nominal multicategorical variable as well if you strategically "ordered" the nominal categories in such a way that the regression coefficients that result quantify the mean differences of interest.

TABLE 10.6. Helmert Coding of Three or Four Ordinal Categories

Ordinal position (low to high)	D_1	D_2	D_3	Mean of Y
$g = 3$ groups				
1	−2/3	0	—	$\overline{Y}_1 = b_0 - (2/3)b_1$
2	1/3	−1/2	—	$\overline{Y}_2 = b_0 + (1/3)b_1 - (1/2)b_2$
3	1/3	1/2	—	$\overline{Y}_3 = b_0 + (1/3)b_1 + (1/2)b_2$
$g = 4$ groups				
1	−3/4	0	0	$\overline{Y}_1 = b_0 - (3/4)b_1$
2	1/4	−2/3	0	$\overline{Y}_2 = b_0 + (1/4)b_1 - (2/3)b_2$
3	1/4	1/3	−1/2	$\overline{Y}_3 = b_0 + (1/4)b_1 + (1/3)b_2 - (1/2)b_3$
4	1/4	1/3	1/2	$\overline{Y}_4 = b_0 + (1/4)b_1 + (1/3)b_2 + (1/2)b_3$

10.1.2 Helmert Coding

Sequential coding results in regression coefficients that quantify the difference between means for groups ordinally adjacent to each other on the variable defining groups. An alternative coding system useful for ordinal multicategorical variables is Helmert coding. This method of coding groups results in regression coefficients that quantify the difference in means between one group and the mean of the means of all groups ordinally higher on the multicategorical variable defining groups.

Table 10.6 shows a set of Helmert codes for three as well as four groups, and Table 10.7 provides the general algorithm for constructing codes for five or more groups. Regressing willingness to self-censor on D_1, D_2, and D_3 using the Helmert codes in Table 10.3, the resulting model is (see Table 10.4)

$$\hat{Y} = 2.993 - 0.278D_1 - 0.282D_2 - 0.055D_3$$

with the same R as when groups were coded with indicator or sequential codes, and the same F- and p-values for testing the null hypothesis that ${}_TR = 0$. And the model generates $\hat{Y}$ values that correspond to the groups means:

TABLE 10.7. Helmert Coding of g Categories, $g \geq 5$

Ordinal position (low to high)	D_1	D_2	D_3	$\cdots$	D_{g-1}
1	$-(g-1)/g$	0	0	$\cdots$	0
2	$1/g$	$-(g-2)/(g-1)$	0	$\cdots$	0
3	$1/g$	$1/(g-1)$	$-(g-3)/(g-2)$	$\cdots$	0
$\vdots$					
$g-1$	$1/g$	$1/(g-1)$	$1/(g-2)$	$\cdots$	$-1/2$
g	$1/g$	$1/(g-1)$	$1/(g-2)$	$\cdots$	$1/2$

$$\hat{Y}_1 = 2.993 - 0.278(-3/4) - 0.282(0) - 0.055(0) = 3.201 = \overline{Y}_1$$
$$\hat{Y}_2 = 2.993 - 0.278(1/4) - 0.282(-2/3) - 0.055(0) = 3.111 = \overline{Y}_2$$
$$\hat{Y}_3 = 2.993 - 0.278(1/4) - 0.282(1/3) - 0.055(-1/2) = 2.857 = \overline{Y}_3$$
$$\hat{Y}_4 = 2.993 - 0.278(1/4) - 0.282(1/3) - 0.055(1/2) = 2.802 = \overline{Y}_4$$

So mathematically, this model is no different than any other model of the groups means we have estimated in that it generates the same estimates of Y and fits exactly the same. But the regression coefficients are different, because they quantify different things now.

Tables 10.3 and 10.6 contains the formulas used to derive each group's mean from the regression model, and the computations above are completed for this example. As can be seen,

$$\overline{Y}_1 = b_0 - (3/4)b_1 \tag{10.1}$$

What cannot be seen quite as easily is that the mean of the three means for the groups ordinally higher than group 1 on the variable defining the groups is

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = \frac{b_0 + b_1/4 - 2b_2/3}{3} + \frac{b_0 + b_1/4 + b_2/3 - b_3/2}{3} + \frac{b_0 + b_1/4 + b_2/3 + b_3/2}{3}$$

which, happily, reduces to a much simpler form:

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = b_0 + (1/4)b_1 \tag{10.2}$$

Substraction of equation 10.1 from equation 10.2 yields

$$(b_0 + b_1/4) - (b_0 - 3b_1/4) = \frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$
$$b_1 = \frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$

and so b_1 quantifies the difference between $\overline{Y}_1$ and the average of $\overline{Y}_2$, $\overline{Y}_3$, and $\overline{Y}_4$. Indeed, observe in this example that

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1 = \frac{3.111 + 2.857 + 2.802}{3} - 3.201 = -0.278 = b_1$$

The t-statistic and p-value are used to test whether these two means are statistically different. In this example, we can conclude that generations older than Generation Y are less willing to self-censor on average than are members of Generation Y, $b_1 = -0.278, t(457) = -3.133, p = .002$.

Similar derivations lead to similar interpretations of b_2 and b_3:

$$b_2 = \frac{\overline{Y}_3 + \overline{Y}_4}{2} - \overline{Y}_2$$
$$b_3 = \overline{Y}_4 - \overline{Y}_3$$

which is indeed the case in this example:

$$b_2 = \frac{2.857 + 2.802}{2} - 3.111 = -0.282$$
$$b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857 = -0.055$$

Generations older than Generation X are less willing to self-censor on average than are members of Generation X, $b_2 = -0.282, t(457) = -5.240, p < .001$, but there is no statistically significant difference in average willingness to self-censor between baby boomers and pre-baby boomers, $b_3 = -0.055, t(457) = -0.846, p = .398$. Notice that b_3 is the same with Helmert coding as it was in section 10.1.1 when using sequential coding.

So when using Helmert coding, b_j, the regression coefficient for D_j, estimates the difference between $\overline{Y}_j$ and the unweighted mean of means for all groups ordinally higher than group j on the variable defining groups. The t- and p-values can be used to test the null hypothesis that these means are equal.

Thus far we have neglected the regression constant, b_0. In this example, $b_0 = 2.993$, which is equal to the mean of the four group means:

$$b_0 = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4}$$
$$b_0 = \frac{3.201 + 3.111 + 2.857 + 2.802}{4}$$
$$b_0 = 2.993$$

More generally, when using Helmert coding in this fashion (and assuming no other regressors are in the model, as in this case), the regression constant is the unweighted mean of all the group means.

A variation on Helmert coding is *reverse* Helmert coding. With reverse Helmert coding, the regression coefficient b_j quantifies the difference between the mean of Y for the group in ordinal position j on the variable defining groups and the unweighted average mean of Y for all groups ordinally *lower* than position j. Although reverse Helmert coding has a different name, there is no need here to provide detail about how the codes are constructed. This is because you can mimic reverse Helmert coding by using ordinary Helmert coding of the ordinal categories as in Tables 10.3, 10.6, and 10.7, but after first ordering the groups on the variable that defines them from high to low rather than low to high.

Helmert coding can be useful even when the multicategorical variable is nominal. For example, perhaps you have conducted an experiment with four conditions that consist of a control group and three experimental treatment conditions, with the treatment being a manipulation of a variable that is not quantitative in any sense of the word. If you use numerical codes for the four conditions strategically, then you can use Helmert coding (though it wouldn't generally be called this) to set up a set of comparisons between the mean of group 1 (say, the control group) versus the mean of the three treatment groups, the mean of the first treatment group versus the mean of the other two treatment groups, and the mean of the second treatment group versus the mean of the third treatment group.

10.1.3 Effect Coding

Effect coding is a minor variation on indicator coding, but the reference against which group means are compared changes. Recall that in indicator coding, one of the g groups receives a code of zero on all indicator variables, and that group ends up the reference group against which all other group means are compared.

With effect coding, a set of $g - 1$ variables D_j are constructed just as in indicator coding, but the group left "uncoded" is set to -1 on all D_j rather 0, as in Table 10.8. Thus, the $g - 1$ D_j variables are no longer dummy variables, as they contain three values (0, 1, or -1) rather than only two. This minor change in coding has an important effect on the interpretation of the regression coefficients and the constant relative to indicator coding.

When willingness to self-censor is regressed on D_1, D_2, and D_3 using the effect coding system for age in Table 10.3, the resulting model is (see Table 10.4)

$$\hat{Y} = 2.993 + 0.208D_1 + 0.119D_2 - 0.136D_3$$

with the same R as when groups were coded with indicator, sequential, or Helmert codes, and the same F- and p-value for testing the null hypothesis that ${}_TR = 0$. And the model generates $\hat{Y}$ values that equal the group means:

$$\hat{Y}_1 = 2.993 + 0.208(1) + 0.119(0) - 0.136(0) = 3.201 = \overline{Y}_1$$
$$\hat{Y}_2 = 2.993 + 0.208(0) + 0.119(1) - 0.136(0) = 3.111 = \overline{Y}_2$$
$$\hat{Y}_3 = 2.993 + 0.208(0) + 0.119(0) - 0.136(1) = 2.857 = \overline{Y}_3$$
$$\hat{Y}_4 = 2.993 + 0.208(-1) + 0.119(-1) - 0.136(-1) = 2.802 = \overline{Y}_4$$

So mathematically, this model is no different than any other model of the groups means we have estimated, in that it generates the same estimates of Y and fits exactly the same.

With indicator coding, b_j quantifies the difference between the mean of the group coded by D_j and the reference group. But with effect coding, b_j is the difference in the mean of group j and the mean of all g group means. That is,

$$b_j = \overline{Y}_j - \frac{\overline{Y}_1 + \overline{Y}_2 + \cdots + \overline{Y}_g}{g}$$

and the t- and p-values for each b_j tests the null hypothesis that $\overline{Y}_j$ equals the mean of all group means. Assuming no additional variables are in

TABLE 10.8. Effect Coding of g Categories

Group	D_1	D_2	$\cdots$	D_j	$\cdots$	D_{g-1}
1	1	0	$\cdots$	0	$\cdots$	0
2	0	1	$\cdots$	0	$\cdots$	0
$\vdots$						
j	0	0	$\cdots$	1	$\cdots$	0
$\vdots$						
$g-1$	0	0	$\cdots$	0	$\cdots$	1
g	-1	-1	$\cdots$	-1	$\cdots$	-1

the model, the regression constant is that unweighted mean of all g group means. For example, from Table 10.2

$$b_0 = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = \frac{3.201 + 3.111 + 2.857 + 2.802}{4} = 2.993$$

and

$$b_1 = \overline{Y}_1 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 3.201 - 2.993 = 0.208$$

$$b_2 = \overline{Y}_2 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 3.111 - 2.993 = 0.119$$

$$b_3 = \overline{Y}_3 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 2.857 - 2.993 = -0.136$$

all of which correspond to the model coefficients from Table 10.4. We can conclude that Generation Y is more willing to self-censor than average, $t(457) = 3.133, p = .002$, as is Generation X, $t(457) = 2.841, p = .005$. But baby boomers are less willing to self-censor than average, $t(457) = -3.376, p = .001$

Missing from this analysis is a comparison of the pre-baby boomers to the average. This finding is sacrificed by the requirement that only $g - 1$ variables coding group can be used in the model. But this comparison can be obtained by rerunning the analysis, setting a different group to receive the -1 codes on all D_j.

10.2 Comparisons and Contrasts

10.2.1 Contrasts

Many questions about differences between group means can be phrased as questions about *contrasts*. The simplest type of contrast is a *pairwise comparison*, which is the difference between two means. Some of the coding systems described in sections 9.1.1 and 10.1 produce regression coefficients and hypothesis tests that yield pairwise comparisons, such as the $g-1$ comparisons between each group mean and a reference group mean when using indicator coding or between group means for groups ordinally adjacent on the ordinal, multicategorical variable when using sequential coding.

More complex contrasts involve more than two means. For instance, perhaps an investigator is entertaining the efficacy of five different therapies for the treatment of depression. Perhaps methods 1, 2, and 3 are all based on principles of theory A about how people think and feel, and methods 4 and 5 are based on a second theoretical orientation B, perhaps one that also includes the use of medication. One might want to know whether clients who are treated with one of the theory A methods differ, on average, in depression 6 months later than clients treated with one of the theory B methods. In this example, the mean depression of those those treated by theory A could be expressed as $(\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3)/3$ and the mean depression of those treated by theory B would be $(\overline{Y}_4 + \overline{Y}_5)/2$. The difference between these,

$$\frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3}{3} - \frac{\overline{Y}_4 + \overline{Y}_5}{2}$$

is a more complex comparison involving several group means rather than a simple pairwise comparison.

Any contrast, whether a pairwise comparison or more complex, can be expressed as a weighted sum of means of the form

$$\text{Contrast} = \sum_{j=1}^{g} c_j \overline{Y}_j \tag{10.3}$$

where c_j is the *contrast coefficient* for group j and $\sum c_j = 0$. For instance, the complex contrast above can be expressed as

$$\text{Contrast} = (1/3)\overline{Y}_1 + (1/3)\overline{Y}_2 + (1/3)\overline{Y}_3 + (-1/2)\overline{Y}_4 + (-1/2)\overline{Y}_5 \tag{10.4}$$

which is a weighted sum of means as in equation 10.3 with $c_1 = c_2 = c_3 = 1/3$ and $c_4 = c_5 = -1/2$.

There is no requirement that all contrast coefficients be nonzero, *so long as they sum to zero.* For example, a pairwise comparison among a set of five means that compares only the means of groups 1 and 2 can be written as

$$\text{Contrast} = (1)\overline{Y}_1 + (-1)\overline{Y}_2 + (0)\overline{Y}_3 + (0)\overline{Y}_4 + (0)\overline{Y}_5$$

which is in the form of equation 10.3 with $c_1 = 1$, $c_2 = -1$ and $c_3 = c_4 = c_5 = 0$. Observe that this simplifies to $\overline{Y}_1 - \overline{Y}_2$.

We can multiply contrast coefficients by a constant without affecting the results of statistical tests of contrasts. This can be especially convenient when a coefficient in fractional form cannot be expressed in decimal form without some rounding or loss of precision (such as $1/3 = 0.33333\ldots$). For instance, the complex contrast in equation 10.4 can be expressed as

$$\text{Contrast} = (2)\overline{Y}_1 + (2)\overline{Y}_2 + (2)\overline{Y}_3 + (-3)\overline{Y}_4 + (-3)\overline{Y}_5$$

which results from multiplying all contrast coefficients by six. The resulting contrast will be six times larger as a result, but the standard error as generated by the formula in section 10.2.2 will also be six times larger to compensate, so the p-value from a hypothesis test is unaffected. We'll see this illustrated in section 10.2.3.

To illustrate the computations, let's use contrast coefficients to generate a contrast of the average willingness to self-censor of Generation Y compared to everyone else. That is,

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$

We'll also construct a contrast comparing the average willingness to self-censor of Generation X and Generation Y relative to baby boomers and pre-baby boomers, which is

$$\frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2}$$

For the former contrast, we use contrast coefficients of $c_1 = -1$, $c_2 = 1/3$, $c_3 = 1/3$, and $c_4 = 1/3$, for Generation Y, Generation X, baby boomers, and

pre-baby boomers, respectively. Notice these add up to zero, as required for a proper contrast. Applying equation 10.3 yields

$$\begin{aligned}\text{Contrast} &= (1/3)\overline{Y}_2 + (1/3)\overline{Y}_3 + (1/3)\overline{Y}_4 - \overline{Y}_1 \\ &= (1/3)3.111 - (1/3)2.857 - (1/3)2.802) - (1)\overline{Y}_1 \\ &= -0.278\end{aligned}$$

which is interpreted to mean that Generation Y is estimated to be, on average, 0.278 units higher in willingness to self-censor than the average willingness to self-censor of Generation X, baby boomers, and pre-baby boomers. Note that this contrast is identical to b_1 from the regression model when using Helmert coding of groups (see Table 10.4).

The second contrast requires contrast coefficients of $c_1 = 1/2$, $c_2 = 1/2$, $c_3 = -1/2$, and $c_4 = -1/2$, respectively, which add to zero as required. Applying equation 10.3 produces

$$\begin{aligned}\text{Contrast} &= (1/2)\overline{Y}_1 + (1/2)\overline{Y}_2 + (-1/2)\overline{Y}_3 + (-1/2)\overline{Y}_4 \\ &= (1/2)3.201 + (1/2)3.111 - (1/2)2.857 - (1/2)2.802 \\ &= 0.327\end{aligned}$$

So Generations X and Y are estimated as, on average, 0.327 units higher in willingness to self-censor than baby boomers and pre-baby boomers.

10.2.2 Computing the Standard Error of a Contrast

For inference, we need an estimate of the standard error of the contrast. When regression analysis is used to emulate analysis of variance with g groups, some simple hand computations yield the standard error using only the contrast coefficients, group sample sizes, and $MS_{residual}$ from the regression. The formula is

$$SE(\text{contrast}) = \sqrt{MS_{residual} \sum_{j=1}^{g} \frac{c_j^2}{n_j}} \tag{10.5}$$

With the standard error for a contrast computed, a test of significance for the null hypothesis that the contrast equals zero can be conducted using

$$t = \frac{\text{Contrast}}{SE(\text{contrast})}$$

and generating a p-value using the $t(df_{residual})$ distribution. Alternatively, a confidence interval can be constructed in the usual way as the point estimate plus or minus t_{crit} standard errors, where t_{crit} is from a table of critical values of t for an interval corresponding to a certain degree of confidence (see Appendix C).

For the contrast comparing the mean of Generation Y against the mean of the other three group means, applying equation 10.5 using the means and group sample sizes in Table 10.2 and the $MS_{residual}$ from any of the regression models in Table 10.4 results in

$$\begin{aligned} SE(\text{contrast}) &= \sqrt{0.273\left(\frac{(-1)^2}{38} + \frac{(1/3)^2}{149} + \frac{(1/3)^2}{173} + \frac{(1/3)^2}{101}\right)} \\ &= 0.089 \end{aligned}$$

and so $t(457) = -0.278/0.089 = -3.133, p < .001$. This contrast of means is statistically significant. Notice that the standard error of this contrast is identical to the standard error of b_1 in the model of Y using Helmert coding. So clearly, given that this contrast is just a comparison of the ordinally lowest age group against all others, much work is saved conducting this contrast by just using Helmert coding and regressing willingness to self-censor on the Helmert codes.

None of the coding systems described in section 10.1 yield the second contrast results comparing the mean of the means of Generation X and Generation Y to the mean of the means of baby boomers and pre-baby boomers. The estimated standard error of this contrast is

$$\begin{aligned} SE(\text{contrast}) &= \sqrt{0.273\left(\frac{(1/2)^2}{38} + \frac{(1/2)^2}{149} + \frac{(-1/2)^2}{173} + \frac{(-1/2)^2}{101}\right)} \\ &= 0.058 \end{aligned}$$

and so $t(457) = 0.327/0.058 = 5.638$, which is statistically significant, $p < .001$.

10.2.3 Contrasts Using Statistical Software

These computations need not be conducted by hand if you have a statistics program capable of doing them. Most good programs can these days. You will probably find options for conducting contrasts in your program's ANOVA routine rather than its regression module, as historically it is in

Descriptives

WTSC: Willingness to Self-Censor

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Generation Y	38	3.2013	.49403	.08014	3.0389	3.3637	2.12	4.37
Generation X	149	3.1117	.62201	.05096	3.0110	3.2124	1.25	5.00
Baby boomer	173	2.8573	.46793	.03558	2.7871	2.9275	1.87	4.50
Pre baby boomer	101	2.8020	.45420	.04519	2.7123	2.8916	1.75	4.00
Total	461	2.9558	.54086	.02519	2.9063	3.0053	1.25	5.00

ANOVA

WTSC: Willingness to Self-Censor

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9.983	3	3.328	12.207	.000
Within Groups	124.581	457	.273		
Total	134.564	460			

Contrast Coefficients

COHORT: Age cohort

Contrast	Generation Y	Generation X	Baby boomer	Pre baby boomer
1	-3	1	1	1
2	.5	.5	-.5	-.5

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
WTSC: Willingness to Self-Censor	Assume equal variances	1	-.8329	.26584	-3.133	457	.002
		2	.3269	.05762	5.674	457	.000
	Does not assume equal variances	1	-.8329	.25241	-3.300	44.897	.002
		2	.3269	.05551	5.888	125.437	.000

FIGURE 10.1. SPSS output from a one-way ANOVA with two contrasts.

the context of ANOVA that contrasts are usually introduced in textbooks and classrooms.

In SPSS, for example, the command below will conduct an analysis of variance testing for a difference in mean willingness to self-censor between the four age cohorts, while also conducting the two contrasts described earlier by specifying the appropriate contrast coefficients following the **`contrast`** option.

```
oneway wtsc by cohort/contrast -3 1 1 1/contrast 0.5 0.5 -0.5 -0.5
   /statistics descriptive.
```

In this command, the coefficients for the first contrast were multiplied by 3. SPSS's ONEWAY module does not allow fractions such as "1/3" in the contrast line, and 1/3 cannot be represented in decimal form exactly.

Observe from the output in Figure 10.1 that the resulting contrast is three times larger than when computed using fractional coefficients, but the standard error is also three times larger. As a result, the t-ratio and p-value are the same as when fractional coefficients are used. Because 1/2 can be represented exactly in decimal form, there is no need to multiply the coefficients by a constant for the second contrast.

Comparable code for SAS is

```
proc glm data=wtsc;
   class cohort;model wtsc=cohort;means cohort;
   contrast '1 vs 2 3 4' cohort -3 1 1 1;
   contrast '1 2 vs 3 4' cohort 0.5 0.5 -0.5 -0.5;
run;
```

In SAS you must provide a name for the contrast in quotes, as above, prior to listing the coefficients. SAS will produce the contrasts in the form of F-ratios with 1 and $df_{residual}$ degrees of freedom, along with a p-value corresponding to the test of the null that the contrast equals zero.

SPSS's UNIANOVA module has some options built in to do contrasts that correspond to the coding systems described in this chapter. For instance, the command below conducts a one-way ANOVA while also producing output for contrasts equivalent to those generated by Helmert, sequential (**repeated**), and indicator (**simple**) coding of groups.

```
unianova wtsc by cohort/emmeans=tables(cohort)/contrast (cohort)=
   helmert/contrast (cohort)=repeated/contrast (cohort)=simple.
```

Consult your preferred program's documentation to see if it is capable of doing comparable analyses.

10.2.4 Covariates and the Comparison of Adjusted Means

Adjusted means were introduced in sections 9.2.4 and 9.2.6 as estimates of group means if all groups were average on a covariate or covariates. We saw in those sections that the regression coefficients for indicator codes can be interpreted as differences between adjusted means whenever a covariate is included in the model along with the codes for groups, and hypothesis tests or confidence intervals used for inference.

Covariates can be included in a model when groups are represented with any coding system, including sequential, Helmert, and effect coding.

When covariates are included, the interpretation we gave to the regression coefficients in section 10.1 apply to adjusted means rather than to the unadjusted means.

To illustrate, we examine differences between the four age groups in willingness to self-censor, with shyness used as as a covariate and using Helmert coding of age cohort. Research shows that people who are relatively higher in willingness to self-censor are also relatively higher on measures of shyness (Hayes et al., 2005), so it is worth examining whether the differences in willingness to self-censor exist independent of any differences between groups in their average shyness. A measure of shyness was included in the survey and is available in the data file, so it is a simple matter to adjust for shyness by simply including it as an additional regressor in the model. The resulting regression equation is

$$\hat{Y} = 2.187 - 0.163D_1 - 0.133D_2 + 0.028D_3 + 0.281X_1 \qquad (10.6)$$

where X_1 is shyness. Corresponding regression output (from SAS, though SPSS and STATA output provide the same information) can be found in Figure 10.2. Applying the test discussed in section 9.2.2 results in $SR^2 = \Delta R^2 = .019, F(3, 456) = 4.256, p = .006$. So the groups differ on average in willingness to self-censor even after accounting for differences between them in shyness.

Setting shyness to the sample mean (in the data, $\overline{X}_1 = 2.832$) and plugging the Helmert codes into equation 10.6 generates the adjusted mean willingness to self-censor for each group:

$$\hat{Y}_1 = 2.187 - 0.163(-3/4) - 0.133(0) + 0.028(0) + 0.281(2.832) = 3.105$$
$$\hat{Y}_2 = 2.187 - 0.163(1/4) - 0.133(-2/3) + 0.028(0) + 0.281(2.832) = 3.031$$
$$\hat{Y}_3 = 2.187 - 0.163(1/4) - 0.133(1/3) + 0.028(-1/2) + 0.281(2.832) = 2.884$$
$$\hat{Y}_4 = 2.187 - 0.163(1/4) - 0.133(1/3) + 0.028(1/2) + 0.281(2.832) = 2.912$$

The computations described in section 10.1.2 but substituting the adjusted means for the unadjusted means reveals that the regression coefficients quantify differences between adjusted means (or means of adjusted means):

$$
\begin{aligned}
b_1 &= \frac{\hat{Y}_2 + \hat{Y}_3 + \hat{Y}_4}{3} - \hat{Y}_1 \\
&= \frac{3.031 + 2.884 + 2.912}{3} - 3.105 \\
&= -0.163 \\
b_2 &= \frac{\hat{Y}_3 + \hat{Y}_4}{2} - \hat{Y}_2 \\
&= \frac{2.884 + 2.912}{2} - 3.031 \\
&= -0.133 \\
b_3 &= \hat{Y}_4 - \hat{Y}_3 \\
&= 2.912 - 2.884 \\
&= 0.028
\end{aligned}
$$

Standard errors for these differences are available in regression output, along with t- and p-values and confidence intervals if desired. As can be seen in Figure 10.2, holding shyness constant (at the mean or any other value else), Generation Y is more willing to self-censor than those older; $t(456) = -2.107, p = 0.036$; and Generation X is more willing to self-censor than those older; $t(456) = -2.748, p = .006$; but baby boomers and pre-baby boomers do not differ significantly in willingness to self-censor; $t(456) = 0.495, p = .621$.

Complex contrasts between means were introduced in section 10.2.1. A contrast is a weighted sum of means, with the weighting determined by a group's contrast coefficient. Although equation 10.3 can be applied to adjusted means, the standard error of a contrast involving weighted means cannot be calculated using equation 10.5. The proper formula is complex, especially when more than one covariate is in the model. It is best to leave the production of the standard error for a complex contrast involving adjusted means to a computer.

In SPSS, the command below produces a complex contrast comparing the adjusted mean of Generation X to the mean of the adjusted means of all other groups, as well as a contrast comparing the mean of the adjusted means for Generations X and Y against mean of the adjusted means for baby boomers and pre-baby boomers. See a discussion of this latter contrast in section 10.2.1 for how the contrast coefficients are selected. The result of the first contrast is identical to the estimate and hypothesis test for b_1

```
                              The REG Procedure
                                Model: MODEL1
                            Dependent Variable: wtsc

                        Number of Observations Read         461
                        Number of Observations Used         461


                                Analysis of Variance

                                             Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       4       41.00281        10.25070      49.96    <.0001
Error                     456       93.56104         0.20518
Corrected Total           460      134.56385


                 Root MSE              0.45297    R-Square     0.3047
                 Dependent Mean        2.95577    Adj R-Sq     0.2986
                 Coeff Var            15.32479


                              Parameter Estimates

                          Parameter       Standard
Variable      DF          Estimate          Error    t Value    Pr > |t|

Intercept      1           2.18674        0.07018      31.16      <.0001
d1             1          -0.16317        0.07744      -2.11      0.0357
d2             1          -0.13263        0.04826      -2.75      0.0062
d3             1           0.02826        0.05713       0.49      0.6211
shy            1           0.28122        0.02287      12.30      <.0001


                Test 1 Results for Dependent Variable wtsc

                                        Mean
           Source             DF        Square    F Value    Pr > F

           Numerator           3       0.87325       4.26    0.0056
           Denominator       456       0.20518
```

FIGURE 10.2. SAS output from a regression estimating willingness to self-censor from age cohort controlling for shyness.

in the example above when Helmert coding was used to code groups. The second shows that the means of these adjusted means are statistically different, contrast = −0.170, $t(456) = 3.295$, $p < .01$. SAS produces the result in the form of an F-ratio rather than a t-statistic.

```
glm wtsc by cohort with shy/emmeans=tables(cohort)/
    lmatrix cohort -1 1/3 1/3 1/3/lmatrix cohort -0.5 -0.5 0.5 0.5.
```

In SAS the comparable commands are

```
proc glm data=wtsc;
    class cohort;model wtsc=cohort shy;lsmeans cohort;
    contrast '1 vs 2 3 4' cohort 3 -1 -1 -1;
```

```
    contrast '1 2 vs 3 4' cohort 0.5 0.5 -0.5 -0.5;
run;
```

10.3 Weighted Group Coding and Contrasts

Section 10.1 described various methods for coding a multicategorical variable that produced regression coefficients that correspond to a comparison between two means. For indicator and sequential coding, the regression coefficients for each code quantified a difference between the means of two and only two groups. But for Helmert and effect coding, the regression coefficients quantified the difference between one group mean and an *unweighted* mean of the means of two or more groups.

For example, when Helmert coding was used in section 10.1.2, the regression coefficient of −0.278 for D_1 quantified the difference in mean willingness to self-censor between Generation Y (3.201) and the mean of the three means for the groups older than Generation Y (2.923). The mean for everyone older than Generation Y was constructed as

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = \frac{3.111 + 2.857 + 2.802}{3} = 2.923$$

This is an unweighted mean, in that ignores the differences in sample sizes between the three groups that contribute to it. Notice from Table 10.2 that there are 149 in the sample from Generation X, 173 baby boomers, and 101 pre-baby boomers. So the 101 pre-baby boomers contribute as much to the construction of this mean of means as the 173 baby boomers, even though there are substantially fewer pre-baby boomers in the data. If this bothers you, then read this section, where we describe versions of Helmert, effect coding, and contrasts that acknowledge differences between the group sample sizes whenever one of the means being compared is formed as a mean of means.

10.3.1 Weighted Effect Coding

We saw in section 10.1.3 that the youngest three cohorts (Generation Y, Generation X, and baby boomers) differ from average in their willingness to self-censor, where *average* was defined as the mean of the four group means. But this was an unweighted average of the four group means, meaning it ignored the fact that the four age cohorts differ in size. If you want the average against which each mean is compared when using effect

TABLE 10.9. Weighted Effect Coding of g Categories

Group (j)	D_1	D_2	$\cdots$	D_j	$\cdots$	D_{g-1}
1	1	0	$\cdots$	0	$\cdots$	0
2	0	1	$\cdots$	0	$\cdots$	0
$\vdots$						
j	0	0	$\cdots$	1	$\cdots$	0
$\vdots$						
$g-1$	0	0	$\cdots$	0	$\cdots$	1
g	$-n_1/n_g$	$-n_2/n_g$	$\cdots$	$-n_j/n_g$	$\cdots$	$-n_{g-1}/n_g$

coding to incorporate group size, you can use *weighted* effect coding. It requires replacing the −1 codes used in effect coding with ratios of group sizes. More specifically, if n_g is the sample size for the group coded −1 on all D_j, replace the −1 for D_j with $-n_j/n_g$. See Table 10.9.

For example, in these data there are 38 people from Generation Y, 149 people from Generation X, 173 baby boomers, and 101 pre-baby boomers. Thus, we change the −1 values for D_j for the pre-baby boomers to $D_1 = -38/101$, $D_2 = -149/101$, and $D_3 = -173/101$ (see Table 10.10). Regressing willingness to self-censor on these weighted effect codes yields

$$\hat{Y} = 2.956 + 0.246D_1 + 0.156D_2 - 0.098D_3$$

(see Table 10.11). As with all the other coding systems used in section 10.1, the model fits the same, and it reproduces the group means. Now b_0 is the weighted mean of means (which is equivalent to just calculating the average of Y ignoring age cohort entirely):

$$b_0 = \frac{n_1\overline{Y}_1 + n_2\overline{Y}_2 + n_3\overline{Y}_3 + n_4\overline{Y}_4}{n_1 + n_2 + n_3 + n_4}$$

$$b_0 = \frac{38(3.201) + 149(3.111) + 173(2.857) + 101(2.802)}{461}$$

$$b_0 = 2.956$$

and b_j is the difference between the mean Y for the group receiving $D_j = 1$ and the weighted mean of all g group means on Y:

$$b_1 = \overline{Y}_1 - 2.956 = 3.201 - 2.956 = 0.246$$
$$b_2 = \overline{Y}_2 - 2.956 = 3.111 - 2.956 = 0.156$$
$$b_3 = \overline{Y}_3 - 2.956 = 2.857 - 2.956 = -0.098$$

As when using unweighted effect coding, we conclude that Generation Y is more willing to self-censor than average, $t(457) = 3.026, p = .003$, as is Generation X, $t(457) = 4.433, p < .001$, whereas baby boomers are less willing to self-censor than average, $t(457) = -3.139, p = .002$.

10.3.2 Weighted Helmert Coding

Weighted Helmert coding is comparable to Helmert coding, in that it generates regression coefficients that compare the mean Y of one group to the mean Y of all groups ordinally higher on the variable coding groups. However, for weighted Helmert coding the mean of Y for all groups higher than ordinal position j is a weighted mean rather than an unweighted mean.

There is no way of representing how to generate weighted Helmert codes with a simple algorithm in table form as in Table 10.7. Construction of weighted Helmert codes requires matrix algebra. But an understanding of matrix algebra is not required to implement this coding system using the syntax we provide at the end of the section. However, you do need to know how to construct the matrix that is used as input into the syntax.

The first step is the construction of a $g \times (g - 1)$ matrix that takes the form in Table 10.12, where g is the number of groups and n_{j+} is the sum of the sample sizes for groups in ordinal position j or higher on the variable defining the groups. That is,

$$n_{j+} = \sum_{i=j}^{g} n_i$$

For example, from the size of the age cohorts in the willingness to self-censor data (see Table 10.2), $n_{2+} = n_2 + n_3 + n_4 = 149 + 173 + 101 = 423$,

TABLE 10.10. Weighted Effect and Helmert Coding and the Group Means Defined in Terms of the Regression Coefficients and Regression Constant

Age cohort by increasing age	D_1	D_2	D_3
Weighted effect coding			
Generation Y	1	0	0
Generation X	0	1	0
Baby boomer	0	0	1
Pre-baby boomer	−38/101	−149/101	−173/101
Weighted Helmert coding			
Generation Y	−.7500000000	−.0141843972	−.0656934307
Generation X	.2500000000	−.6619385343	−.0656934307
Baby boomer	.2500000000	.3380614657	−.4343065693
Pre-baby boomer	.2500000000	.3380614657	.5656934307

TABLE 10.11. Estimating Willingness to Self-Censor from Age Cohort Using the Coding Systems in Table 10.10

		Coeff.	*SE*	*t*	*p*
Weighted effect coding (Pre-baby boomers uncoded) $R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.956	0.024	121.550	< .001
D_1	b_1	0.246	0.081	3.026	.003
D_2	b_2	0.156	0.035	4.443	< .001
D_3	b_3	−0.098	0.031	−3.139	.002
Weighted Helmert coding $R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	−0.268	0.088	−3.026	.003
D_2	b_2	−0.275	0.053	−5.172	< .001
D_3	b_3	−0.055	0.065	−0.846	.398

TABLE 10.12. Construction of the Input Matrix for Weighted Helmert Coding

	Column				
	1	2	3	$\cdots$	$g-1$
Row					
1	-1	0	0	$\cdots$	0
2	n_2/n_{2+}	-1	0	$\cdots$	0
3	n_3/n_{2+}	n_3/n_{3+}	-1	$\cdots$	0
$\vdots$					
$g-1$	n_{g-1}/n_{2+}	n_{g-1}/n_{3+}	$n_{g-1}/n_{(g-1)+}$	$\cdots$	-1
g	n_g/n_{2+}	n_g/n_{3+}	$n_g/n_{(g-1)+}$	$\cdots$	1

$n_{3+} = n_3 + n_4 = 173 + 101 = 274$, and $n_{4+} = 101$. So with $g = 4$ groups as in this example, the 4×3 matrix would be

$$\begin{array}{ccc} -1 & 0 & 0 \\ n_2/n_{2+} & -1 & 0 \\ n_3/n_{2+} & n_3/n_{3+} & -1 \\ n_4/n_{2+} & n_4/n_{3+} & 1 \end{array}$$

or, in terms of the group sample sizes in the four age cohorts,

$$\begin{array}{ccc} -1 & 0 & 0 \\ 149/423 & -1 & 0 \\ 173/423 & 173/274 & -1 \\ 101/423 & 101/274 & 1 \end{array}$$

Once this matrix is constructed, it is manipulated through matrix algebra to produce a $g \times (g-1)$ matrix that contains the $g-1$ sets of weighted Helmert codes for the g groups, where rows correspond to groups and columns are the codes D_1, D_2, and so forth. In this example, the resulting matrix is

$$\begin{array}{rrr} -.7500000000 & -.0141843972 & -.0656934307 \\ .2500000000 & -.6619385343 & -.0656934307 \\ .2500000000 & .3380614657 & -.4343065693 \\ .2500000000 & .3380614657 & .5656934307 \end{array}$$

which are the codes for D_1, D_2, and D_3 found in Table 10.10. Regressing willingness to self-censor on D_1, D_2, and D_3 using these weighted Helmert codes yields the following model:

$$\hat{Y} = 2.993 - 0.268D_1 - 0.274D_2 - 0.055D_3$$

(see Table 10.4) with the same R as when any other coding system is used, as well as the same F- and p-values for testing the null that ${}_TR = 0$. And the model generates $\hat{Y}$ values that correspond to the group means.

So mathematically, this model is no different than any other model of the groups means we have constructed so far, in that it generates the same estimates of Y and fits exactly the same. But now the regression coefficient for D_j quantifies the difference between $\overline{Y}_j$ and the weighted mean of the means of Y for all groups coded higher than j on the variable quantifying the groups:

$$b_1 = \left(\frac{149\overline{Y}_2}{423} + \frac{173\overline{Y}_3}{423} + \frac{101\overline{Y}_4}{423}\right) - \overline{Y}_1 = 2.933 - 3.201 = -0.268$$

$$b_2 = \left(\frac{173\overline{Y}_3}{274} + \frac{101\overline{Y}_4}{274}\right) - \overline{Y}_2 = 2.837 - 3.111 = -0.274$$

$$b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857 = -0.055$$

The t-statistic and p-value for each regression coefficient tests the null hypothesis that the difference between the corresponding true means is equal to zero. As can be seen comparing the results when using unweighted to weighted Helmert coding (Tables 10.4 and 10.11), the results are very similar in this case, although this won't always be true. Generation Y self-censors less on average than those older, and Generation X self-censors on average more than those older, but baby boomers self-censor no more on average than pre-baby boomers.

The matrix computations are very tedious to do by hand. Fortunately, many good statistics programs have built-in features to do matrix computations. The SPSS code below takes the input matrix, implements the matrix algebra, and outputs the matrix of weighted Helmert codes. You can then construct D_1, D_2, and D_3 using **`if`** and **`compute`** commands. See the *Syntax Reference Manual* for guidance or consult a local expert.

```
matrix.
compute m={-1.0000, 0.0000, 0.0000;
```

```
            149/423,-1.0000, 0.0000;
            173/423,173/274,-1.0000;
            101/423,101/274, 1.0000}.
compute d=m*inv(t(m)*m).
print d.
end matrix.
```

In SAS, matrix operations can be conducted using PROC IML, which is an optional package. Check your installation. The comparable code in SAS is

```
proc iml;
m={-1.000000000 0.000000000 0.000000000,
     0.352245862 -1.000000000 0.000000000,
     0.408983451 0.631386861 -1.000000000,
     0.238770685 0.368613138 1.000000000};
d=m*inv(m`*m);
print d;
quit;
```

In STATA, try

```
mata
m=(-1.00000,0.00000,0.00000\
    149/423,-1.0000,0.00000\
    173/423,173/274,-1.0000\
    101/423,101/274, 1.0000)
d=m*luinv(m'*m)
d
end
```

10.3.3 Weighted Contrasts

Use of weighted Helmert codes generates regression coefficients and tests of significance, some of which can be interpreted as *complex contrasts*, a term and method introduced in section 10.2.1. However, in that discussion, the contrast involved a comparison of unweighted means. For example, in that section, we compared average willingness to self-censor among

Generation Y and Generation X to average willingness to self-censor among baby boomers and pre-baby boomers:

$$\begin{aligned}\text{Contrast} &= \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} \\ &= 3.156 - 2.830\end{aligned}$$

Those two means being compared were unweighted, because the mean of Generations Y and X was constructed merely by taking the arithmetic average of $\overline{Y}_1$ and $\overline{Y}_2$, ignoring that there are many fewer Generation Y in the sample than Generation X. Similarly, the mean of the baby boomers and pre-baby boomers was was constructed as the mean of $\overline{Y}_3$ and $\overline{Y}_4$, ignoring differences in sample size.

Weighted versions of these two means of means would give weight to Generation X relative to Generation Y, and to baby boomers relative to pre-baby boomers, in proportion to differences in their sample sizes. So rather than 3.156 for the combination of Generations X and Y, their weighted mean would be

$$\frac{38\overline{Y}_1 + 149\overline{Y}_2}{187} = \frac{38(3.201) + 149(3.111)}{187} = 3.129$$

Notice that this is closer to the mean of Generation X than Generation Y, because Generation X contributes more data to the mean. Similarly, the weighted mean for the combination of baby boomers and pre-baby boomers would be

$$\frac{173\overline{Y}_3 + 101\overline{Y}_4}{274} = \frac{173(2.857) + 101(2.802)}{274} = 2.837$$

rather than 2.830, which is closer to the mean of baby boomers, because its sample size is larger than the pre-baby boomers.

Complex contrasts can be conducted that compare weighted means to each other by using the relative sample sizes of the groups, as in the example computations above. Define a *contrast grouping* as a set of groups being combined in a contrast. A contrast always involves two, and only two, contrast groupings. In this example, contrast grouping 1 is the group defined as Generation X and Generation Y, and contrast grouping 2 is the group defined as baby boomers and pre-baby boomers. Now define n_{group_1} as the sum of the sample sizes of the groups that define contrast grouping 1, and n_{group_2} as the sum of the sample sizes of the groups that define contrast grouping 2. So in this example, $n_{group_1} = 38 + 149 = 187$ and

$n_{group_2} = 173 + 101 = 274$. Finally, define λ_j as the ratio of group j's sample size to the sample size of its corresponding contrast grouping. In this case,

$$\begin{aligned}\lambda_1 &= n_1/n_{group_1} = 38/187\\ \lambda_2 &= n_2/n_{group_1} = 149/187\\ \lambda_3 &= n_3/n_{group_2} = 173/274\\ \lambda_4 &= n_4/n_{group_2} = 101/274\end{aligned}$$

With the $g = 4$ values of λ calculated, a weighted contrast is constructed as

$$\text{Contrast} = \sum_{j=1}^{g} c_j \lambda_j \overline{Y}_j \tag{10.7}$$

and its standard error estimated as

$$SE(\text{contrast}) = \sqrt{MS_{residual} \sum_{j=1}^{g} \frac{\left(c_j \lambda_j\right)^2}{n_j}} \tag{10.8}$$

where c_j is the contrast coefficients for group j as defined in section 10.2.1. The ratio of the contrast to its standard error is distributed as $t(df_{residual})$, and a p-value can be constructed using the t-distribution for testing a null hypothesis about the contrast (e.g., that the two weighted means are equal, meaning their difference is zero).

In this example, $c_1 = c_2 = 0.5$ and $c_3 = c_4 = -0.5$. Application of equation 10.7 yields

$$\begin{aligned}\text{Contrast} &= 0.5(38/187)(3.201) + 0.5(149/187)(3.111) -\\ &\quad 0.5(173/274)(2.857) - 0.5(101/274)(2.802)\\ &= 0.146\end{aligned}$$

and equation 10.8 generates

$$\begin{aligned}SE(\text{contrast}) &= \sqrt{0.273\left[\frac{\left(0.5\frac{38}{187}\right)^2}{38} + \frac{\left(0.5\frac{149}{187}\right)^2}{149} + \frac{\left(-0.5\frac{173}{274}\right)^2}{173} + \frac{\left(-0.5\frac{101}{274}\right)^2}{101}\right]}\\ &= 0.025\end{aligned}$$

Their ratio is $t = 0.146/0.025 = 5.840$, which has an exceedingly tiny two-tailed p-value derived from the $t(457)$ distribution. The null hypothesis of equality of the weighted means is rejected.

Notice that in this example, the contrast calculated above is actually one half of the difference between the weighted means rather than the difference itself:

$$\begin{aligned}\text{Contrast} &= 0.5(38/187)\overline{Y}_1 + 0.5(149/187)\overline{Y}_2 - \\ &\quad 0.5(173/274)\overline{Y}_3 - 0.5(101/274)\overline{Y}_4 \\ &= 0.5\left[\left(\frac{38}{187}\overline{Y}_1 + \frac{149}{187}\overline{Y}_2\right) - \left(\frac{173}{274}\overline{Y}_3 + \frac{101}{274}\overline{Y}_4\right)\right] \\ &= 0.5(3.129 - 2.837) \\ &= 0.146\end{aligned}$$

However, so too is the estimated standard error one-half of the standard error of the difference between the weighted means, so the result of the inference is unaffected. If this bothers you, simply multiply both by two when reporting. This correction would be important if reporting a confidence interval for the difference between weighted means, because you would want the confidence interval to be in the metric of the difference, not one-half the difference.[1]

These computations can be done by most statistical programs that allow you to specify contrast coefficients in an ANOVA procedure, and these will be done more accurately than the hand computations illustrated above. In the unweighted contrast example from section 10.2.3, we put c_j in the computer code to produce the contrast. But now, we use $c_j\lambda_j$ for the contrast coefficients instead. So in SPSS, the code to conduct this contrast would be

```
oneway wtsc by cohort/contrast 0.101604 0.398396 -0.315693 -0.184307
   /statistics descriptive.
```

or in SAS, use

```
proc glm data=wtsc;
   class cohort;model wtsc=cohort;means cohort;
   contrast '1 2 vs 3 4' cohort 0.101604 0.398396 -0.315693 -0.184307;
run;
```

[1]It is not generally true that equation 10.7 will produce one-half the difference between weighted means. Whether or not equation 10.7 produces the weighted mean difference or some multiple of it will depend on the values of c_j used.

In this example, each value of $c_j\lambda_j$ input into the code could be multiplied by 2 to rescale the contrast to the mean difference metric rather than one-half the difference.

10.3.4 Application to Adjusted Means

Weighted effect and weighted Helmert coding will produce regression coefficients that correspond to differences between weighted adjusted means when covariates are included in the model. You may be tempted to do complex weighted contrasts between adjusted means using the procedure described in section 10.3.3, substituting adjusted means for $\overline{Y}_j$. But equation 10.8 does not produce a proper estimate of the standard error of a contrast between weighted adjusted means. The computer-assisted procedures described in section 10.2.3 can be used instead, so long as the contrast coefficients fed to the computer algorithm are multiplied by the appropriate weights (i.e., use $c_j\lambda_j$ rather than c_j) to produce the contrast of interest.

10.4 Chapter Summary

In this chapter we introduced and illustrated several ways of coding a multicategorical variable so that it can be used as a regressor in a regression model. These methods, including sequential coding, Helmert coding, and effect coding, yield models that are mathematically equivalent to the model generated when indicator coding is used. Of these methods, sequential and Helmert coding are particularly useful when the multicategorical variable represents an ordinal dimension. But regardless of the method of coding used, the choice one makes about how to code groups does not affect the fit of the model or the estimates of Y is produces. Furthermore, the choice does not affect the test of the null hypothesis that the g groups don't differ on average on Y, and using regression analysis results in the same inference produced by ANOVA and ANCOVA. However, the method of coding groups will change the regression constant and the regression coefficients and how they are interpreted.

Complex contrasts between means is a staple topic in analysis of variance books, but it is still appropriate in a regression analysis book such as this because ANOVA is just a special case of linear regression analysis, and contrasts can be conducted using output from a regression analysis. Another topic commonly introduced in the context of ANOVA is the *multiple test problem*—the positive correlation between the number of tests conducted and the probability of making at least one Type I error. In the

next chapter we address the multiple test problem and its relevance not only to comparing groups but also to regression analysis more generally.

11
Multiple Tests

A multiple regression analysis contains many inferential tests, such as the test that the multiple correlation is zero, a test for each partial regression coefficient, and maybe a test or two examining the contribution of a set of predictors to improving the fit of the model. Indeed, in any scientific report, usually there are many tests conducted, whether that takes the form of many regression analyses or some combination of other statistical procedures, each with its own set of tests. Given that each inference carries with it the possibility of a Type I error, the more tests that are conducted, the greater the likelihood of reporting as real at least one effect that is not. In this chapter we introduce this *multiple test problem* by way of illustration and then outline the Bonferroni method as a simple approach to dealing with it when one has conducted multiple hypothesis tests. As we discuss in the last section of the chapter, the multiple test problem introduces many interesting philosophical questions. It becomes clear after thinking about some of these questions that whether one should correct for all hypotheses tests to account for the multiple test problem, or just some of them, or none of them, depends on many things that are hard to quantify, such as the plausibility that all null hypotheses are true, how well established the research area is, and the logical independence of the hypotheses being tested.

In Chapters 9 and 10 we showed that a multicategorical variable can be used as a regressor in a linear model if properly represented with a *set* of regressors. We showed how linear regression analysis mimics the results you would get if you conducted an ANOVA or ANCOVA comparing the group means on Y. Furthermore, the regression coefficients for the variables coding groups can be interpreted as tests of differences between certain means or sets of means.

A reader with a background in ANOVA might have observed that we failed to address one topic in these two chapters that is a staple in ANOVA textbooks and classrooms: the problem of multiple tests. The multiple test problem is usually discussed in the context of comparing group means

following a statistically significant result from ANOVA. Because rejection of the null hypothesis of no difference between the group means leads to only a vague conclusion, follow-up tests are typically used to find the source of the differences. But this usually involves more than one hypothesis test, and the more hypothesis tests you conduct (or the more confidence intervals you construct), the more likely you are to claim an existence of a difference between means that reflects only chance or random noise.

As we discuss in this chapter, the multiple test problem is pervasive and surfaces almost any time we analyze data, regardless of the statistical method used. It is not specific to ANOVA-type problems. Rarely does an investigator conduct only a single test in a study. And these days, research articles frequently contain more than one and often several studies, each of which contains several hypothesis tests. Hypothesis tests abound in any research report, and the likelihood of a decision error—at least one Type I error—in the collection is much larger than an investigator might realize and be willing to tolerate if something isn't done about it. We offer some of our thoughts on how to approach thinking about the multiple test problem.

11.1 The Multiple Test Problem

11.1.1 An Illustration through Simulation

We illustrate the multiple test problem using a computer to conduct a simulation. The SPSS code below creates a new data set containing 10 variables named *X1*, *X2*, and so forth, through *X10*. Each variable contains a sample of size 100 from a normally distributed population. In the population, these 10 variables are all linearly uncorrelated. That is, $_{\mathcal{T}}r_{X_iX_j} = 0$ for all i and j. So when we test the null hypothesis that a specific correlation is zero, we know that the null hypothesis is true going in. That means we are in a situation scientists aren't usually in, where we know the truth we are trying to discern from a hypothesis test. That means we can tell if the procedure leads us astray in our decision, and how often it does so.

```
new file.
input program.
loop rep=1 to 100.
end case.
end loop.
end file.
```

```
end input program.
do repeat x=x1 to x10.
compute x=rv.normal(0,1).
end repeat.
correlations variables = x1 to x10.
regression/dep=x10/method=enter x1 to x9.
```

In SAS and STATA, the equivalent programs can be written more concisely:

```
data multtest;
array x {10} x1-x10;
do j = 1 to 100;do i = 1 to 10;x{i} = rand("Normal");end;output;end;
run;
proc corr data=multtest;var x1-x10;run;
proc reg data=multtest;model x10=x1-x9;run;
```

```
clear
drawnorm x1-x10,n(100)
pwcorr x1-x10,sig
regress x10 x1-x9
```

If you run this program, it will produce as output a 10×10 matrix of correlations as well as a linear regression analysis estimating X_{10} from the nine other X variables. We focus for now on the correlation matrix. Figure 11.1 is an example of the correlation matrix from the SPSS version of the program, though yours will look different because your sample of 100 will be different than the one that generated this matrix. Each cell in the matrix contains an estimate of ${}_{T}r_{X_iX_j}$, along with a p-value for testing the null hypothesis that ${}_{T}r_{X_iX_j} = 0$. There are 45 such correlations and p-values corresponding to the $10(10-1)/2 = 45$ possible pairs of two variables. The cells below the diagonal are the same as the corresponding cells above the diagonal, because $r_{X_iX_j} = r_{X_jX_i}$. That is, each correlation between X_i and X_j and corresponding p-value for the hypothesis test is found twice in the matrix.

Notice in Figure 11.1 that even though we know that these variables are all mutually uncorrelated in the population, we would reject the null hypothesis of no correlation at the .05 level of significance for two of correlations. These are highlighted in the figure. Of course, you probably observed something different when you ran the program. Perhaps none of the p-values in your 10×10 matrix are less than .05. Or maybe only one of

Correlations

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	Pearson Correlation	1	-.147	.145	-.049	-.185	.039	.120	-.117	.041	-.240
	Sig. (2-tailed)		.144	.149	.630	.065	.703	.233	.245	.682	.016
	N	100	100	100	100	100	100	100	100	100	100
X2	Pearson Correlation	-.147	1	-.035	.019	.131	.007	-.093	-.178	.128	.069
	Sig. (2-tailed)	.144		.726	.850	.193	.942	.359	.077	.203	.495
	N	100	100	100	100	100	100	100	100	100	100
X3	Pearson Correlation	.145	-.035	1	-.038	-.117	-.023	-.008	.030	-.027	-.064
	Sig. (2-tailed)	.149	.726		.711	.248	.822	.938	.766	.789	.525
	N	100	100	100	100	100	100	100	100	100	100
X4	Pearson Correlation	-.049	.019	-.038	1	.007	.100	-.067	-.030	.063	-.079
	Sig. (2-tailed)	.630	.850	.711		.944	.320	.508	.768	.537	.433
	N	100	100	100	100	100	100	100	100	100	100
X5	Pearson Correlation	-.185	.131	-.117	.007	1	-.052	-.079	-.231	.136	-.084
	Sig. (2-tailed)	.065	.193	.248	.944		.608	.434	.021	.177	.407
	N	100	100	100	100	100	100	100	100	100	100
X6	Pearson Correlation	.039	.007	-.023	.100	-.052	1	-.017	-.093	-.150	.009
	Sig. (2-tailed)	.703	.942	.822	.320	.608		.869	.357	.136	.932
	N	100	100	100	100	100	100	100	100	100	100
X7	Pearson Correlation	.120	-.093	-.008	-.067	-.079	-.017	1	-.073	.046	.000
	Sig. (2-tailed)	.233	.359	.938	.508	.434	.869		.473	.652	.999
	N	100	100	100	100	100	100	100	100	100	100
X8	Pearson Correlation	-.117	-.178	.030	-.030	-.231	-.093	-.073	1	-.125	.087
	Sig. (2-tailed)	.245	.077	.766	.768	.021	.357	.473		.217	.390
	N	100	100	100	100	100	100	100	100	100	100
X9	Pearson Correlation	.041	.128	-.027	.063	.136	-.150	.046	-.125	1	.085
	Sig. (2-tailed)	.682	.203	.789	.537	.177	.136	.652	.217		.401
	N	100	100	100	100	100	100	100	100	100	100
X10	Pearson Correlation	-.240	.069	-.064	-.079	-.084	.009	.000	.087	.085	1
	Sig. (2-tailed)	.016	.495	.525	.433	.407	.932	.999	.390	.401	
	N	100	100	100	100	100	100	100	100	100	100

FIGURE 11.1. A matrix of correlations between 10 independent random normal variables in a sample size of 100.

them is. Or maybe three or even four or five are. More likely than not, as we discuss, at least one of them is.

The first lesson to learn here is that Type I errors do happen. Even though we know that all 45 of the population correlations are zero, we would claim that two of them are not if we took Figure 11.1 at face value. These represent Type I errors—claiming an effect exists when in reality no effect exists. Of course, in your own data analyses, you don't know whether or not a particular null hypothesis is true. You assume it is true when you calculate a *p*-value, but that assumption may be wrong. And you'll never know whether a particular decision you make when you conduct a hypothesis test is correct or incorrect.

The second lesson requires that you run the program many times. Each time you run it, take notice of three things. First, focus on the correlation between X_1 and X_2 and record whether the *p*-value for its hypothesis test is less than .05. Do the same for the three correlations between X_1, X_2, and X_3, noting whether any of them are statistically significant at the .05 level. Finally, look at the entire matrix of correlations and record whether or not any of the 45 correlations in the matrix is statistically significant at the .05 level. Repeat this as many times as you care to, but at least 20 or so.

If your experience is typical, what you will likely have observed after many executions of this program is that the correlation between X_1 and X_2 is only rarely statistically significant. It is more common for at least one of the three correlations between X_1, X_2, and X_3 to be statistically significant, and it is even more common for at least one of the 45 correlations to be statistically significant. Just how often this happens in your case will depend on the number of times you executed the program.

When we did this 1 million times with the aid of a computer, in 49,833 of the runs the correlation between X_1 and X_2 was statistically significant at the .05 level. This is about what you would expect. We know that when we test a true null hypothesis, the probability of a Type I error is .05 when using an $\alpha = .05$ criterion for deciding between reject and fail to reject. So in 1 million tests of the null hypothesis that X_1 and X_2 is zero you would to expect to reject this true hypothesis about 50,000 times. Our estimate of the Type I error rate from this simulation is right on at about .05.

But even though we know that none of the correlations between X_1, X_2, and X_3 are different from zero in the population, we found that in 142,530 of the 1 million runs of the program, at least one of these three correlations was statistically significant at the .05 level. This is far more than the 50,000 expected when we focused on only one of the correlations. And in the entire matrix, at least one of the correlations was statistically significant in a whopping 901,254 of the runs, or 90% of the time!

So from this simulation, we estimate that the probability of incorrectly rejecting at least one of the true null hypotheses when testing the correlations between X_1,X_2, and X_3 in a sample of size 100 is around 0.142, even though we tested each null hypothesis the .05 level. And when we look at the whole set of 45 tests, with each test conducted at the .05 level, the probability of incorrectly rejecting at least one of the true null hypotheses is around 0.901. It is nearly certain to happen.

So the second lesson to take away from this illustration is that when you conduct a bunch of hypothesis tests, the probability of making at least one Type I error in the set increases with the size of the set. So the more tests of a true null hypothesis that you conduct, the more likely you are to make a mistake at least once and claim an effect exists that does not in reality. This is the multiple test problem.

11.1.2 The Problem Defined

When we test a null hypothesis, if the null hypothesis is true and the test is valid, the probability of incorrectly rejecting the null hypothesis is the level

of significance α we are using for the test; most researchers use $\alpha = .05$. But when we test many null hypotheses each at the α level of significance and all of the null hypotheses we test are true, the probability that at least one of those decisions is a Type I error is higher than α. And the more tests you do, the greater the probability that at least one of the decisions is a Type I error.

Define α_{FW} as the probability of making at least one Type I error in a set or *family* of B hypothesis tests (FW stands for *familywise*), and define α as the constant level of significance used to reject the null in each of the B tests. The multiple test problem is reflected in the following:

$$\alpha_{FW} \geq \alpha$$

So when you conduct more than one statistical test, the likelihood that you make at least one Type I error in the set of B tests—the *familywise Type I error rate* or *familywise* α—is generally larger than the probability that any one specific decision is a Type I error.

11.1.3 The Role of Sample Size

You might think that the inflation of the Type I error rate when many hypothesis tests are conducted would depend on sample size. For instance, you might have heard some people say that you can reject any null hypothesis if you have a big enough sample. That would suggest that the problem would be worse in large samples, because large samples are more likely to produce statistically significant effects. Alternatively, you may believe that because estimates vary more from sample to sample in small samples, the problem would be worse in small samples since it is easier to find big effects in small samples. Or maybe the problem wouldn't be as bad, because it is harder to get small p-values in small samples.

But in fact neither of these is true. Type I errors are generally no more or less probable in large samples than in small ones, and so the inflation of the Type I error rate when multiple hypothesis tests are conducted is not determined by sample size. You can demonstrate this for yourself by repeating the exercise in section 11.1.1 but changing the "100" in the line of code to a bigger or smaller number. We repeated this simulation twice (with 1 million repetitions each time), once with a sample size of 500 and once with a sample size of 20, and the results were largely identical. Sample size doesn't matter because the derivation of the p-value for a null hypothesis test incorporates sample size. So the extent of the multiple test problem

is not influenced by sample size. You can't make it lessen or go away by increasing or decreasing the sample size. What matters is the number of tests conducted, not the sample size.

11.1.4 The Generality of the Problem

The multiple test problem is usually first discussed in statistics books in the context of ANOVA and the comparison of means. Rejection of the null hypothesis that g means are equal is a vague conclusion that typically initiates a hunt for the source of the differences. One common procedure is to conduct all possible pairwise comparisons between the g means in an attempt to find where the differences between the means resides. Alternatively, one can conduct a set of focused comparisons between specific means or sets of means. At this point, students of ANOVA are often introduced to a mind boggling assortment of approaches to conducting pairwise comparisons that go by such names as Scheffe's test, Tukey's HSD test, Dunnett's test, the Neuman–Keuls method, or the Games–Howell approach, among many others. These are all attempts to deal with the multiple test problem in one way or another, with varying degrees of success. Many journal articles, book chapters, and even entire books exist on the topic of pairwise comparisons between means.

But the multiple test problem is more general than this. We saw in section 11.1.1 that it surfaces when we look at the hypothesis tests found in a matrix of correlation coefficients. In Chapters 9 and 10 we discussed that regression can be used to conduct an ANOVA and ANCOVA and also provides a set of comparisons between means or combination of means depending on how the groups are represented in the coding system employed. When covariates are included, we also get a measure of partial association between the covariate and the outcome. Each line of a table of regression includes a test of the null hypothesis that the true regression coefficient is zero, which corresponds to a null hypothesis test for the difference between group means or the partial association between Y and the covariate.

Indeed, a regression analysis typically includes several hypothesis tests, such as for each regression coefficient, for the multiple correlation, and the change in the multiple correlation when variables are entered into the model hierarchically. The various stepwise entry procedures discussed in section 7.3.1 are often conducted using a hypothesis testing approach to determine what variables to enter or remove and when. Although we didn't talk about the regression output generated by the simulation code in

section 11.1.1, if you take a look at the regression coefficients printed below the correlation matrix, you will observe that some of them are statistically significant now and then, even though we know there is no relationship between X_{10} and any of the nine regressors. In 350,677 of our 1 million runs of the simulation, or about 35% of the time, at least one of the partial regression coefficients was statistically significant at the .05 level.

The problem also surfaces whenever you use more than one analytical method to analyze a data set. For example, you may use confirmatory factor analysis for one part of the study to examine or test a measurement model. This process typically involves hypothesis testing. Once the measurement model is established, you may then seek to examine how the measured variables relate to each other by piecing them together in the form of a path diagram and estimating the model coefficients. This process also involves hypothesis testing. As the number of hypothesis tests builds up as you progress through the analysis, the likelihood of making a Type I error increases.

Finally, if you think you can avoid Type I error inflation by using confidence intervals or alternative approaches to inference such as Bayesian methods, think again. The problem is not caused by any mathematics that is particular to hypothesis testing. The more confidence intervals you construct, the more likely at least one of them will not cover the true value of the parameter being estimated. Bayesian credible intervals are just as susceptible to the problem. Any time multiple inferential procedures are employed when analyzing data from a study, the more likely you are to report at least one effect as real that is not, or report an interval (confidence or credible) that is inaccurate in some way.

As we discuss in section 11.3, you can get really philosophical about the problem and start pondering questions like whether you should worry about the multiple test problem across a set of studies you have conducted on a common topic or all the studies you have conducted in your laboratory. And if you are going to ponder these questions, why stop there? For instance, perhaps all the investigators in your department should team up and come up with some kind of plan for dealing with what is undoubtedly a large number of Type I errors being committed collectively by all the researchers in the department in a given semester, or a given year, or even in the history of the department. But as we discuss later, if you are inclined to worry about the multiple test problem, it is possible to worry more than you really need to.

11.1.5 Do Omnibus Tests Offer "Protection"?

You may have learned that one way of avoiding the multiple test problem is to condition the hunt for effects on evidence of an effect to be found. For example, you may have read about or been told that if the F-ratio from an ANOVA comparing g means is statistically significant, then this affords "protection" from making Type I errors when you start comparing specific means to each other. By this logic, the significant ANOVA result means there is some difference between the means, so you can go ahead and search all you want for it without worrying about the multiple test problem. A related notion in regression analysis says that if R is not statistically different from zero, then you can't interpret the hypothesis tests on the individual regression coefficients, but if R is different from zero by a hypothesis test, then this protects you from making Type I errors when looking at the inferential statistics for the regression coefficients.

But suppose that you've conducted an experiment with five groups, your F-ratio is statistically significant, but in reality four of the means are the same, and one of the means differs from the those four. With five groups, there are 10 pairwise comparisons between means you can do, and for four of them, the null is false, and for six of them the null is true. Since you are doing six tests of a true null hypothesis, the probability you will make at least one Type I error in that set of six is higher than α, even though you have correctly rejected the null hypothesis that the five means are the same. So a statistically significant ANOVA result does not protect you from making an excessive number of Type I errors in the follow-up tests.

11.1.6 Should You Be Concerned about the Multiple Test Problem?

Should *you* be concerned about Type I error inflation? We can't answer that question for you. Some scientists worry about the multiple test problem more than others do. It has even been suggested that this difference is like religious differences, in that no amount of argumentation will change anyone's mind. However, we assume that when you write a scientific article, you want it to be convincing to the broadest possible audience, and to all the reviewers who review it before publication. Or if you read an article, you may want to convince others of the accuracy and importance of its conclusions. Sometimes the article will not address the multiple test problem, yet the people you want to convince may include some who take the problem seriously. Several methods for managing multiple tests are

quite simple and yet deal effectively with the problem. Some don't require access to all the raw data, and can be done using only the summary statistics appearing in a published article. Therefore, any scientist who deals with statistics will benefit from having some knowledge of these methods, either when publishing or when assessing work by others. We dedicate the next section to the simplest and most versatile of these methods.

11.2 The Bonferroni Method

When we test a null hypothesis at the α-level of significance using a valid hypothesis-testing procedure, the probability of failing to reject the null hypothesis if it is true is $1 - \alpha$, and the probability of incorrectly rejecting the null hypothesis if true is α. So if the null hypothesis is true and we use $\alpha = .05$ as the level of significance for testing the null, then the probability of correctly failing to reject the null is $1 - .05 = .95$, and the probability of a Type I error is .05.

Applied to the multiple test problem, we seek to test a family of B null hypotheses while ensuring that the probability of failing to reject all B hypotheses if they are all true is $1 - \alpha_{FW}$. Rephrased, we want to test all B null hypotheses knowing that if they are all true, then the probability of incorrectly rejecting *at least one* is no more than α_{FW}.

The Bonferroni method is by far the easiest and most versatile approach to dealing with the multiple test problem. It has two variants that are mathematically identical and a third that is slightly less conservative, which we discuss in section 11.2.4. One variant is based on a modification of the upper bound on a p-value that is considered statistically significant. The other variant is an adjustment to the p-value from a hypothesis test prior to making a decision with it. Mathematically, these are equivalent, but the latter is a bit more flexible and easier to implement.

With the Bonferroni method, the probability of at least one Type I error in a set of B hypothesis tests is no higher than α_{FW} if each null hypothesis in the set is tested at the α_{FW}/B level of significance. For instance, if you conduct $B = 5$ null hypothesis tests and you want the probability of at least one Type I error in the set to be no higher than $\alpha_{FW} = .05$, then use an $\alpha = .05/5 = .01$ level of significance for rejecting each null hypothesis. That is, for each test, reject its null hypothesis only if $p \leq .01$. So this version of the Bonferroni method modifies α downward as a function of the number of tests, and the null hypothesis is rejected only if the p-value from the test is less than this smaller α.

A mathematically equivalent procedure is to compute a "corrected" p-value, which we will denote p_c, and compare this to the desired α_{FW}. Using this variant, $p_c = B \times p$, where p is the ordinary p-value from the test. In this context, B is sometimes called the "Bonferroni correction factor." So if you have conducted five tests, then an original p-value of .01 translates to $p_c = .01 \times 5 = .05$. Thus, for $\alpha_{FW} = .05$ and five tests are conducted, an original p-value of .003 would be considered statistically significant after a Bonferroni correction factor of 5 is applied, because $p_c = .003 \times 5 = .015$, which is less than .05. But an original p-value of .02 is not statistically significant after a Bonferroni correction of 5 because $p_c = .02 \times 5 = .10$, which is greater than .05.

11.2.1 Independent Tests

The reasonableness of the Bonferroni approach is shown most easily for independent tests. The concept of independence is explained more fully in section 11.2.2. Suppose that five investigators all test the same null hypothesis, and the null hypothesis is actually true. What is the probability that at least one of them will incorrectly reject the null hypothesis at the .05 level?

This question is most easily answered by calculating the probability that they will all correctly fail to reject the null hypothesis. We know that for each investigator, the probability of correctly failing to reject the null is $1 - .05 = .95$. From the multiplicative law of probability for independent events, the probability that all five would fail to reject the null is $(1 - .05)^5 = .95^5 = .774$. So the probability that at least one investigator will incorrectly reject the null is $1 - .774 = .226$. The logic of the Bonferroni method above estimates this probability as $5 \times .05 = 0.25$.

More generally, for a set of B independent hypothesis tests conducted at the same α level,

$$\alpha_{FW} = 1 - (1 - \alpha)^B \tag{11.1}$$

Using the Bonferroni method, if we wanted α_{FW} to be .05, then we test each null hypothesis using $\alpha = .05/5 = .01$. Applying equation 11.1 yields

$$\alpha_{FW} = 1 - (1 - .99)^5 = 1 - .99^5 = 0.049$$

which is pretty close. The formula αB will always be larger than what equation 11.1 yields, so for independent tests, the Bonferroni method is conservative, meaning that it overestimates α_{FW} as produced more exactly

by equation 11.1. For instance, when B is 10, the Bonferroni method yields $\alpha_{FW} = .5$, whereas equation 11.1 yields .401.

In practice, our interest typically is not in estimating α_{FW} but, instead, fixing α_{FW} to something small and adjusting α accordingly or calculating p_c. Equation 11.1 can be rearranged so as to isolate α, which gives the cutoff for p for rejecting the null:

$$\alpha = 1 - (1 - \alpha_{FW})^{1/B} \tag{11.2}$$

So for $B = 10$ and $\alpha_{FW} = .05$, equation 11.2 yields $\alpha = .00512$. The simpler Bonferroni formula, which says to reject the null only if $p \leq \alpha_{FW}/B$, generates $\alpha = .05/10 = .005$, which is close.

The result of α_{FW}/B is always smaller than what equation 11.2 generates, which means that the Bonferroni method is conservative for independent tests. But even when B is large, the approximation is close for small values of α_{FW}, which is typically the only scenario we care about in practice. For instance, for $B = 50$ tests, the Bonferroni method requires $p \leq .001$ to reject the null for each test to ensure $\alpha_{FW} = .05$. Applying equation 11.1 yields $\alpha_{FW} = 1 - (1 - .001)^{50} = .0488$, which is pretty close. In our opinion, the simplicity of the Bonferroni method is worth the tradeoff for its slight inaccuracy and conservatism.

11.2.2 The Bonferroni Method for Nonindependent Tests

If a regression contains k regressors, then the tests on the k regression coefficients are not statistically independent, because the regressors are not independent. Or if a categorical regressor has g categories, there are $g(g - 1)/2$ possible pairwise comparisons between means, and these tests are not independent either. The comparison of $\overline{Y}_1 - \overline{Y}_2$ is not independent of $\overline{Y}_1 - \overline{Y}_3$, because anything that influences $\overline{Y}_1$ will influence both comparisons. Although this would not be true for $\overline{Y}_1 - \overline{Y}_2$ versus $\overline{Y}_3 - \overline{Y}_4$, it is still not true that the tests are independent, because they both use the same $MS_{residual}$ in the construction of the standard error and p-value. So a chance random fluctuation downward of $MS_{residual}$ relative to its true value will raise both t-values for these comparisons. Because of this nonindependence between tests, it would seem that the Bonferroni method would not be relevant.

However, it is relevant. The Bonferroni method is based on an equation called the *Bonferroni inequality*. It states that for two numbers a and B, and if $0 < a < 1$ and $B > 1$, then

$$1 - (1 - a)^B < aB \tag{11.3}$$

If we substitute α for a and still call B the number of tests, then equation 11.3 can be rewritten as

$$1 - (1 - \alpha)^B < \alpha B \tag{11.4}$$

But earlier we said the Bonferroni method estimates α_{FW} as no higher than αB. So equations 11.1 and 11.4 are identical except for the $<$ rather than $=$. Earler we saw that α_{FW} is $1 - (1 - \alpha)^B$ for independent tests, and so equation 11.4 conveys what we said in section 11.2.1, when we said that the Bonferroni method overestimates α_{FW}.

Ryan (1960) showed that if we replace $<$ in equation 11.4 with $\leq$, then it applies to nonindependent tests as well. Ryan also showed that the overestimation of α_{FW} by αB is small whenever αB is small, which is the only case we would care about in practice. So the simple Bonferroni method is an accurate but slightly conservative approach for both independent and nonindependent tests.

The conservativism of the Bonferroni method is positively correlated with the extent of the nonindependence of the tests. We can think of nonindependence as a continuum from -1 to 1. Consider an extreme form of negative nonindependence. Suppose you conducted two one-tailed tests of the correlation between X and Y using a single sample. The first is a test of the null that ${}_T r_{XY} \leq 0$, and the second is a test of the null that ${}_T r_{XY} \geq 0$. We know that when applied to the same r_{XY}, the t-statistics for these tests will be equal but opposite in sign. Thus, their correlation is -1. If each test was conducted at the .025 level and both nulls are true, which can only happen if ${}_T r_{XY} = 0$, then the probability of an incorrect rejection for at least one of these tests is estimated as not exceeding $2 \times .025 = 0.05$ by the Bonferroni method. An equivalent test would be a single two-tailed test of the null that ${}_T r_{XY} = 0$. If the null is true, then the probability of false rejection is .05. So in this case, the Bonferroni method applied to two one-tailed tests that are perfectly negatively correlated gives the correct α_{FW}, not an overestimate. A two-tailed test can actually be thought of as a Bonferroni correction applied to a one-tailed test, using a Bonferroni factor of 2 to correct for the fact that an effect or difference could have come out in the opposite direction from what was observed.

It can be shown that as the nonindependence moves from -1 to 1, the conservativism of the Bonferroni method increases, and is at its maximum when the tests are perfectly positively correlated. For instance, if three tests are perfectly positively correlated, then if one null is rejected, so too are the other two nulls rejected. So if three perfectly correlated tests are conducted at the $\alpha = .05$ level, the probability of at least one false rejection of the

TABLE 11.1. Estimated α_{FW} from the Simulation

B	Uncorrected	Bonferroni
1	.0497	—
3	.1423	.0490
9	.3497	.0487
45	.9018	.0460

null hypothesis in the three tests is .05, not the much larger $.05 \times 3$. But remember that the Bonferroni method gives an upper bound of α_{FW}. That is why it is generally conservative. The upper bound may be and typically will be larger than the actual α_{FW}.

11.2.3 Revisiting the Illustration

We introduced the multiple test problem in section 11.1.1 with an example and simulation. Recall from that illustration that when you focused only on the correlation between X_1 and X_2, the true null hypothesis of no correlation was rarely rejected. In 1 million runs, incorrect rejection occurred only about 5% of the time, which is what would be expected for a test that is valid. But when examining the three correlations between X_1, X_2, and X_3, all 45 of the correlations between X_1 through X_{10}, or the nine partial regression coefficients when estimating X_{10} from the other nine variables, at least one effect in the set was statistically significant far more often than 5% of the time in our 1 million trials. These results are summarized in Table 11.1 in the column labeled "uncorrected."

We applied the Bonferroni method during the simulation as well. The Bonferroni method took care of the problem quite satisfactorily, as can be seen in Table 11.1. In all three scenarios the estimated α_{FW} over the one million replications was near and never above .05 after correction of the p-values to compensate for the number of tests conducted. The fact that the estimates are below .05 reflect the conservatism of the Bonferroni method.

11.2.4 Bonferroni Layering

The Bonferroni method applies the same Bonferroni correction factor B to all B tests to produce a corrected p-value, p_c. An alternative and slightly less conservative approach is *Bonferroni layering*, also known as *Holm's sequential*

rejection procedure (Holm, 1979). To apply layering to a set of B tests, find the one with the smallest p-value and multiply that p-value by B. If the resulting p_c is less than your desired α_{FW}, consider this result statistically significant. Then multiply the next smallest p-value by $B-1$ and compare it to α_{FW}. If p_c is smaller than α_{FW}, consider this result statistically significant. Continue with this procedure, each time reducing the Bonferroni correction factor by 1, until the first time that p_c is above α_{FW}. When that happens, declare that test as nonsignificant, as well as all other tests with p-values not yet corrected.

The idea behind layering is that when you single out the most significant result from a set, you need to be most harsh in the p-value correction for that test to compensate for the fact that it was selected post hoc. Once that result is removed from the set of B tests, you are then going back into the remaining set of $B-1$ tests and selecting the next most significant result for correction. This second correction doesn't need to be quite as harsh as the first, but it still needs substantial correction because, again, it is being selected post hoc for examination.

More generally, the jth-most significant result among B tests is the most significant result in a set of $B+1-j$ results, so to layer in the Bonferroni method, we multiply the jth-most significant p among B by a Bonferroni factor of $B+1-j$. For instance, if the most significant three results out of 10 tests yield p-values of .0012, .0038, and .0092, then the corrected p-values are $.0012 \times 10 = .012$, $.0038 \times 9 = .034$, and $.0092 \times 8 = .074$. Only the first two are significant at the .05 level after correction, and we don't apply any more corrections to the remaining p-values, as these are all considered nonsignificant.

11.2.5 Finding an "Exact" *p*-Value

Most researchers use a statistical package of some kind for data analysis, and most statistical packages produce p-values for various hypothesis tests as a matter of routine. But usually the output from a statistical package shows the p-value to only three or four decimal places of accuracy. When implementing the Bonferroni method, we often need a more precise p-value than is shown in computer output. For example, suppose you want to apply a Bonferroni correction factor of 25 to a p-value to generate p_c, and the uncorrected p is shown in your output as .002, because your program rounds output to the third decimal place. So .002 could reflect a p-value as small as .0015 or as large as but not quite .0025. If you want α_{FW} to be no greater than .05, the exact value of p matters because $.0015 \times 25$ is less than

.05, but .0024 × 25 is greater than .05. And as will be seen in later chapters, there are times when you will want to apply a Bonferroni correction as large as N, the sample size for the study. If N is large, such as 1,500, you need to know p very precisely in order to apply the correction. Multiplying something with rounding error by something large will produce something with even more rounding error in absolute terms.

Fortunately, most statistical packages have options for changing the number of decimal places of resolution generated in output. Check your program's documentation. If yours does not, there is a good chance that it has a number of functions built in that you can access for generating p-values from other statistics in the output. The application of these algorithms will yield some rounding error, because the input you feed it (e.g., a t-statistic from the output) will also contain some rounding error. But what these algorithms generate even then will still be more precise than what your output is giving you if it doesn't let you see p-values to more than three or four decimal places of accuracy.

The first four lines of SPSS code below generate p-values (two-tailed) corresponding to $Z = 4$, $t(20) = 4$, $F(3, 30) = 4$, and $\chi^2(2) = 4$, respectively. You can modify the code for different values of Z, t, F, or χ^2 for different degrees of freedom by changing the relevant part of the code. Before executing this code, you have to have a data file open. Alternatively, you can start with a new data set that you first populate with a single observation and a single arbitrary value for a single variable arbitrarily named.

```
compute pz=2*(1-cdf.normal(4,0,1)).
compute pt=2*(1-cdf.t(4,20)).
compute pf=1-cdf.f(4,3,30).
compute pchi=1-cdf.chi(4,2).
format pz pt pf (F16.8).
execute.
```

The results of these operations will be produced as new variables in the data set. In this case, the code produces variables `pz`, `pt`, `pf`, and `pchi` that are populated with the values .00006334, .00070352, .01651537, and .13533528, respectively. The 8 in **`F16.8`** tells SPSS to show eight decimal places of resolution. Change this to something larger if desired. These p-values are displayed to eight decimal places of resolution, but can't really be considered "exact" even at this level of resolution, because these functions themselves generate only approximations of p-values, not exact values.

In SAS, the code below accomplishes these computations and prints the p-values in the output window:

```
data pprob;
pz=2*(1-cdf('normal',4));
pt=2*(1-cdf('t',4,20));
pf=1-cdf('F',4,3,30);
pchi=1-cdf('chisquare',4,2);
run;
proc print data=pprob;var pz pt pf pchi;run;
```

In STATA, the command window can be used sort of like a calculator. This set of commands will display these four p-values in the output window, without producing or modifying any data file:

```
scalar pz=2*(1-normal(4))
scalar pt=2*ttail(20,4)
scalar pf=Ftail(3,30,4)
scalar pchi=chi2tail(2,4)
display pz,pt,pf,pchi
```

Consult your program's user manual for specific details on how these functions work and what computational algorithms they implement.

11.2.6 Nonsense Values

The Bonferroni method can produce values of p_c or upper bounds on α_{FW} that are greater than 1. For instance, if each of $B = 30$ tests is conducted at the .05 level, then $\alpha B = .05 \times 30 = 1.5$. But α_{FW} is a probability and so can't be larger than 1. Similarly, suppose that the original p-value for a test is .04 but you have conducted 30 tests. The corrected p-value is $p_c = .04 \times 30 = 1.2$. But probabilities can't be larger than 1. In cases like this, simply truncate p_c or α_{FW} at 1. In application it makes no difference. If $p_c > \alpha$, then you don't reject the null hypothesis. It doesn't matter how much larger p_c is relative to α in such cases. Similarly, we don't usually approach a multiple testing situation by estimating and upper bound on α_{FW}. Rather, we fix the upper bound to some small value such as .05 and ask how small does p have to be uncorrected in order to reject the null.

11.2.7 Flexibility of the Bonferroni Method

The Bonferroni method is extremely flexible and can be applied to any set of tests, not just tests involving means. The tests may be either independent or nonindependent. They may be part of the same study conducted by an

investigator, from a set of studies conducted by the same investigator, or even different studies conducted by different investigators. The original tests can be one- or two-tailed. They can be from parametric or nonparametric tests, in any combination. They can be from different types of test based on different sampling distributions, such as *t*-tests, chi-square tests, *F*-tests, and so forth.

11.2.8 Power of the Bonferroni Method

Intuitively, the power of the Bonferroni method might seem very poor. After all, if obtaining $p < .05$ seems a lofty goal under the best of circumstances, then finding something smaller than .05 must be even harder. We have already said that it tends to be conservative. But it is easy to fail to notice how small a *p*-value actually is, because many statistics programs will show small *p*-values as .000. Consider a fairly ordinary finding of $r_{XY} = 0.4$ based on a sample size of 100. The two-tailed *p*-value for such a result is .0000374, which would be statistically significant at the .05 level even with a Bonferroni correction of 1,336.

It is true that when the Bonferroni method is used instead of a method dedicated to comparing means, it tends to be more conservative than those methods whose validity is unquestioned. But the difference isn't that large, and the lower power of the Bonferroni method seems like a reasonable price to pay given its enormous flexibility. If the conservativism of the Bonferroni method still bothers you, you might consider an alternative general method such as the one described by Benjamini and Hochberg (1995, 2000) that is a bit more powerful but not as easy to implement.

11.3 Some Basic Issues Surrounding Multiple Tests

It is almost impossible to perform or even evaluate scientific research without facing the problem of multiple tests, and one cannot consider multiple tests without facing difficult philosophical questions. For example, why correct for multiple tests at all? After all, regardless of the number of tests performed, won't it still be true that, in the long run, only 5% of all true null hypotheses will be mistakenly rejected at the .05 level? At the other extreme, if we decide to correct for multiple tests in some fashion, why don't we make even more conservative corrections that one might ordinarily contemplate? If you are worried about multiple tests when evaluating

the result of a single experiment or study, why not correct for all the tests you do when you publish a multiple-study article? But why stop there, given you are likely to be reporting many studies over many articles in your career. Should you correct for all the tests you have done to date or will ever do? And what about all the tests being done in a particular area, such as social psychology, or indeed, the entire history of science? The Bonferroni method isn't the only way to deal with the multiple test problem, but until we can answer broad questions like these, it seems unlikely that we will agree on narrower philosophical questions concerning specific methods of correcting for Type I error inflation. This section attempts to answer some of these questions, at least from our perspective.

11.3.1 Why Correct for Multiple Tests at All?

The argument for correcting for multiple tests relies heavily on the concept of a composite null hypothesis (CNH). A CNH is the hypothesis that two or more simple null hypotheses are all true. Familiar types of CNH include the hypothesis tested with ANOVA that three or more means are equal, and the hypothesis in regression that ${}_{T}R = 0$, which implies that ${}_{T}r_{YX_j} = 0$ for every j.

CNHs play a very important role in the scientific process, because tests on them can be more powerful than tests on the specific hypotheses nested within them. CNHs usually lead to only vague conclusions (e.g., g means are not the same; at least one regression coefficient is not zero), but vague conclusions are more easily reached than specific ones. Rejection of a CNH is a vague conclusion; it asserts that there is at least one real effect nested within the CNH without telling us which one or ones.

But rejection of any specific null hypothesis within a CNH implies rejection of the CNH itself. Suppose we had an experiment with five conditions and we did all 10 possible pairwise comparisons between two means using independent groups t-tests rather than a one-way ANOVA. If we reject the null hypothesis tested with any of the t-tests, then we thereby must reject the CNH that all five means are equal. Thus if we perform 1,000 experiments without correcting for the fact that there are, say, five means being compared in each, far more than 5% of 1,000 CNHs will falsely be rejected at the .05 level if we don't correct for the fact that we are doing 10 t-tests in each experiment. Thus, it is simply not true, as hinted above, that even with no corrections, only 5% of all true null hypotheses will be rejected.

CNHs define whole areas of science. For instance, the assertion "No baldness lotions work" is a CNH. But folk medicine around the world mentions hundreds of lotions for baldness. We would not want to adopt rules that made rejections of a CNH almost certain if we simply tested enough simple hypotheses within the CNH. That is why we need corrections for multiple tests—a correction transforms the test of a simple hypothesis into a valid test of the CNH within which it is nested.

11.3.2 Why Not Correct for the Whole History of Science?

A CNH can be nested within a broader CNH. For example, in a 4×6 factorial ANOVA, the CNH that all four means defining the first factor are equal is nested within the broader CNH that all 24 means are equal. The CNH that humans lack mental telepathy is nested within the broader CNH that all species lack it. All null hypotheses in all areas of science are nested within the broadest CNH of all—the hypothesis that every statistically significant result in the history of science is a Type I error, resulting from the very large number of tests scientists have performed in world history—"All science is bunk."

Section 11.3.1 implied that the broader the CNH, the vaguer is the conclusion and the easier it should be to reject the CNH. This is true; the CNH, "All science is bunk," is easier to reject than you might imagine. Consider that polls of the public are undertaken weekly if not daily, and it is not uncommon for these polls to be based on 1,500 or so people. Suppose a poll of 1,500 people who were asked whether they have a favorable or unfavorable opinion of the current U.S. President shows a 60–40 split, with 60% having a favorable opinion and 40% an unfavorable opinion. A test of the null that the public is equally split (i.e., 50–50) would lead to a rejection of the null hypothesis at any level of significance you would fathom using. But what Bonferroni correction would have to be applied to make the p-value no longer statistically significant at, say, the .05 level? Might it be 100, or 1,000, or maybe even 10,000? The answer may amaze you. The two-tailed p-value for this test is about 1 in 100 trillion. So it would take a Bonferroni correction of about 5 trillion to make this p-value .05. It is probably fair to say that no more than 5 trillion null hypothesis tests have ever been performed in the entire history of science. So this one very modest poll result is sufficient for us to reject the null hypothesis that "all science is bunk." Indeed, many everyday scientific results based on moderately large samples would allow us to do so. For instance, a fairly

ordinary correlation of .5 in a sample of 211 cases is significant at the same level as the above example: 1 in 100 trillion.

A similar argument applies to CNHs that are narrower than this one but still very broad—such as the CNH that "all social psychology is bunk," because all significant results in that area can be explained by the sheer number of hypothesis tests. Because the number of tests ever performed in this area is far less than the number performed in all of science, we need an even less impressive result to reject it—and results at the level of significance required to do so are common in social psychology. In this way, we work down a hierarchy of nested CNHs. Except in perhaps a few small areas of research, this line of reasoning usually allows a researcher to conclude that broadest CNH he or she may need to consider is the one spanning a particular experiment. That is why we do not need to correct for the whole history of science, or, usually, for any tests at all outside of our present study.

11.3.3 Plausibility and Logical Independence of Hypotheses

Suppose that the investigators in a given lab work in different areas of a well-established discipline, and in a given month five studies are conducted, one by each investigator in his or her own particular area. One of the results is statistically significant at the .02 level, and the other four have large p-values. If the one result were corrected for the other four by the Bonferroni method, it would no longer be statistically significant. But most scientists would assert that such correction is unreasonable because these five tests are "independent."

But consider a slight variation of this problem. Suppose five investigators at different universities all conducted the same study, tested the same null hypothesis, and one rejected the null with $p = .02$ and the other four investigators had large p-values. Most scientists would consider it quite reasonable to correct the one significant result for the other four.

What is the difference between these two scenarios? We assert that the impulse to correct for multiple tests in the second scenario but not in the first stems from the fact that the five investigators in the first scenario are testing logically independent hypotheses, whereas the five investigators in the second scenario are testing hypotheses that are not logically independent.

We shall define two or more hypotheses as logically independent if firm knowledge of the truth or falsehood of one hypothesis would not change our opinion of the plausibility of the other hypotheses. Consider the hypothesis that in New York State more women than men will vote for the

Democratic candidate in the next presidential election. By changing the state, we can generate 49 other forms of that hypothesis. Those 50 hypotheses are certainly distinguishable; some may be true and others false. But are they logically independent? No; at least we don't think so. Learning that more women than men had voted Democratic in New York State would, for most people, increase the plausibility of the hypothesis that the same result will be found in other states. Similarly, knowing that one investigator testing a certain hypothesis failed to reject the null hypothesis would probably affect our beliefs about the likelihood that a different investigator testing that same hypothesis would be able to reject the null. In both of these cases, these hypotheses are not logically independent.

But five investigators testing different hypotheses in their own areas of inquiry are testing logically independent hypotheses, or at least more independent than five different investigators testing the same hypothesis. Knowing the outcome of one investigator's study would provide no (or, at least, less) information, leading us to change our beliefs about the plausibility of the hypotheses the other investigators are testing.

It may be apparent to you that there is some subjectivity in the determination as to whether the hypotheses in a set are logically independent. Suppose two investigators in the same lab were studying different phenomena, such as the effect of traumatic experiences on satisfaction with one's relationships, and the effect of different types of therapies for treating PTSD. But two investigators in another lab were comparing the effects of a particular type of therapy on various mental states, with one investigator studying depression and the other studying anxiety. On the surface it would seem that the hypotheses the first two investigators are testing are logically independent, more so than the latter two.

But does knowing that a method of therapy does not work for depression really provide information about the likelihood of it working or not for anxiety? That depends on your perspective. Knowing that it doesn't work for one symptom may lead you to believe it is not likely to work with others, but anxiety and depression are different psychological experiences. Who is to say that there are any implications whatsoever about the effectiveness of a therapeutic method for one kind of psychological state on its effectiveness on other psychological states or symptoms?

The relevance of logical independence is more apparent in the context of the *plausibility* of a CNH. Some CNHs are more plausible than others. Consider the CNH that experiencing trauma has no effect on any aspect of a person's life, and the CNH that a exposure to trauma does not affect

cognitive reasoning. Most would think the former CNH is much less plausible than the latter. Remember that the truth of the CNH implies the truth of all specific hypotheses nested within it. So all the null hypotheses tested by investigators examining the relationship between exposure to trauma and cognitive reasoning are nested within the broader CNH that trauma has no effect on any aspect of a person's life. Likewise, the specific null hypothesis tested by an investigator examining the effect of trauma on something like the ability to solve anagrams is nested within the CNH that the trauma does not affect cognitive reasoning.

Consider now the *multiplicative law of probabilities*, which states that when two or more events are independent, the probability that all the events will occur equals the product of the individual probabilities. Suppose 100 logically independent specific hypotheses about the effect of trauma are tested that are nested under the CNH that exposure to trauma has no effect on any aspect of a person's life. And further suppose that for each specific hypothesis test, we can all agree that the probability that the null is true for that test is very high, say 0.95. By the multiplicative law, if the tests are independent, then the probability that all 100 of the specific null hypotheses are true is $0.95^{100} = .006$. In other words, the probability that the CNH is true is only .006, even though we are pretty sure for each test that its null hypothesis is true. Reframed, we can be pretty sure that *at least one* of the specific nulls is actually false; the probability of that is $1 - 0.006 = 0.994$. This means that we can probably reject the CNH outright, without any testing needed, because it is simply too implausible.

But this logic applies to logically independent hypotheses. If the hypotheses the 100 investigators tested aren't logically independent, such as if they were all testing the same null hypothesis using the same methodology, then we can't just multiply the probabilities of the specific null hypotheses being true together in this fashion. The probability that they are all incorrect may be quite high, perhaps as high as 0.95 if the tests were perfectly positively correlated. The CNH is more plausible when the hypotheses are logically independent, and we should be concerned about falsely rejecting it if enough tests are done (as in the balding lotions example).

But recognizing the implausibility of the CNH has a bearing on the severity of the multiple test problem we face. Remember that the computations we went through in sections 11.1.1 and 11.2 assume that all null hypotheses are true. In other words, it assumes that the CNH is true. But we just said that the CNH may be implausible. If it isn't plausible, why would we worry about falsely rejecting it, and why would we want to build

in multiple test correction into our specific tests that assumes the implausible null hypothesis is true? It isn't that we shouldn't worry at all about multiple tests. It may be that there is some subset of specific hypotheses that, when combined, constitute a plausible CNH. But the size of that subset would determine the kind of correction we apply to compensate for multiple tests, not the broader set of all specific null hypotheses under the implausible CNH. Regardless, the former is smaller than the latter, so by applying a Bonferroni correction corresponding to the latter, we may be overcorrecting.

An argument could also be made that when an investigator conducts a set of hypothesis tests, the CNH that all nulls the investigator tests are true is often not particularly plausible. Most scientists don't just conjure hypotheses out of thin air. Rather, the questions they ask of their data, and the studies they design to answer those questions, usually reflect a reasoned argument based on their training and knowledge of the substantive area they are studying, existing relevant literature that leads them to make predictions about what they should find if there is any truth to the past literature, the theories prior research supports, and so forth. Most scientists are not so stupid that they are likely to get it wrong for *every* hypothesis they test in a study. It seems that there is a good chance that that a scientist's informed, theory-derived reasoning is correct for at least one of those hypotheses being tested. We may not know which one, but they probably aren't *all* wrong. This suggests that it would be overcompensating to apply a multiple test correction that assumes all specific null hypothesis are true. This just doesn't seem very plausible.

Thinking about the multiple test problem from the perspective of plausibility of the CNH suggests that we should be less concerned about the problem when conducting research in an established area than in a new area. By *established*, we mean in an area for which there is already some evidence of an effect. When a new research area appears, there may be a period where we should entertain the possibility that all apparent phenomena reported in the area are caused by the multiplicity of hypothesis tests conducted in the area. But after an area becomes established and accepted as legitimate, we worry less about later related findings in the area being Type I errors.

So if researchers have already established that a particular form of therapy works in treating some psychological conditions, it seems less plausible that it would not work at treating other conditions than it would seem if there wasn't already some evidence that the therapy works. So if you

repeatedly conduct a study on its effectiveness on some different condition, or conduct a single study of its effectiveness on several conditions, you should be less worried about the multiple test problem given that its effectiveness has already been established for at least some psychological conditions. This does not mean that you need not worry at all about multiple tests. But the CNH may not be very plausible when doing research in a well-established area relative to when doing work in a new area.

11.3.4 Planned versus Unplanned Tests

Suppose that on January 1 I declared that I was going to win the lottery at least twice this year. If I played every day, you might or might not be surprised if I did actually win twice. Your surprise would of course depend on the probability of winning on any particular day. If the likelihood of winning on any given day were, say, 1 in a million, winning even once might surprise you, and you might think that perhaps I had some kind of clairvoyance if I did indeed win twice. But if the odds of winning on any day were higher, such as 1 in 20, then winning twice wouldn't be particularly surprising. Indeed, in that case, you'd be surprised if I *didn't* win at least twice if I played every day.

But suppose that I were more specific in my forecast, and I declared that I was going to win on February 14th and on May 8th. Now if that actually happened, you would understandably be impressed, even if the odds of winning on any given day were as high as 1 in 20. And if the odds of winning on any given day were only 1 in a million, you would have a hard time convincing yourself that I wasn't clairvoyant. Cheating or the gamble being somehow fixed in my favor may be the only other plausible explanation for my success at forecasting the 2 days such a low probability event would occur.

Researchers usually conduct research with a particular objective in mind and particular predictions about what they will find that are tested statistically once the data are available. That is, researchers often approach their jobs expecting certain things to happen, in the same way as if I were to say I was going to win the lottery on February 14th and May 8th. In science, the predictions take the form of particular null hypotheses the scientist expects to be rejected when put to the test. Suppose an investigator makes three predictions, and all three null hypotheses he or she predicted would be rejected are in fact rejected. It seems implausible at this point that all of those null hypotheses are actually true (the CNH) given that the investi-

gator forecasted in advance of seeing the data and doing the analysis that they were going to be rejected.

But suppose the investigator did not forecast these results in advance and instead merely reported that only these three results were statistically significant out of many hypothesis tests the researcher decided to conduct when exploring the data looking for statistically significant effects. Now it seems much more plausible that all three of these null hypotheses could actually be true, and we would rightfully be more concerned about the multiple test problem than we would if the investigator predicted these results in advance. This is similar to how you would feel if I were to win the lottery a few times in a year when the odds of winning any given time were high, such as 1 in 20, relative to if I were to call out the specific days I was going to win and that actually happened.

So our thinking about the multiple test problem clearly has to be different when evaluating the results of tests that are anticipated in advance relative to just discovered after a round of "data snooping." Some use the terms *planned* and *unplanned* to distinguish between these two testing scenarios, though these terms seem unsatisfying, because we could plan to mine our data looking for statistically significant results. *A priori* and *post hoc* testing probably is a better set of terms, though still not perfect.

Some feel that for hypothesis tests that are planned or a priori, we need not worry about the multiple test problem or apply any multiple test correction to results. Though we are sympathetic to the position that the multiple test problem is less of a problem for a priori hypothesis tests, this seems too extreme a position for our tastes. By this argument, it would be to a researcher's advantage to formulate a prediction for every conceivable test one could conduct with one's data. But if one were to actually conduct every conceivable test, it is certainly true that some Type I errors would be made if some kind of correction for multiple tests is not made. Yet we should not correct as much for a small set of hypothesis tests relevant to specific predictions we make in advance relative to when we conduct a large number of tests for no reason other than to just see what turns out to be statistically significant.

Planned or a priori tests are usually the ones that researchers care the most about, because they test specific predictions or hypotheses that motivated the research in the first place. Unplanned or post hoc tests usually are of less interest, in that they are frequently exploratory and often suggest themselves after seeing the data. Rosenthal and Rubin (1984) offer a sensible approach to multiple test correction that weights sets of tests by

their interest value. The idea is to partition the desired α_{FW} into the part for tests of high interest and the part for tests of less interest. In principle, one could partition further, such as high, moderate, and low interest, but for simplicity, we assume only two partitions. Using this approach, a different multiple test correction is applied to the two sets, depending on the number of tests in each.

For example, suppose that you plan on doing three hypothesis tests that directly motivated the research you conducted. If you didn't conduct these tests, you wouldn't be answering the very questions you designed the study to answer. But perhaps there are five other tests that are worth conducting, even though they didn't motivate the research in the first place. For those three you care most about, you might set α_{FW} for this set of three to .03 and apply a Bonferroni correction factor of 3 to each of the p-values. That is, you reject the test's null hypothesis only if its uncorrected p is no higher than than .01, meaning its corrected p is no higher than .03. For the five other tests that are of less interest, you conduct those tests such that $\alpha_{FW} = .02$, which is .05 minus the .03 you already gave to the three tests of most interest to you. So for each of these five tests, you reject the null hypothesis only if its uncorrected p is no greater than $.02/5 = .004$, meaning its corrected p is no larger than .02.

Observe that by this strategy, the three tests you care most about are conducted less conservatively than if you corrected all tests by a Bonferroni correction of 8, which would mean you reject a test's null hypothesis only if its uncorrected p is less than .00625, regardless of the interest value of the hypothesis being tested. So for those three tests, you are gaining a little power, because the p-value doesn't have to be quite as small to reject the null. For those tests you don't care as much about, you are being more conservative than you otherwise would be, but that is a small price to pay given that you don't care as much about those tests and you gain a little power for those tests you care about. Importantly, the probability of at least one Type I error across the eight test remains acceptably low at just about .05.

The one difficulty with this approach is how to weight the tests, meaning which ones you care more about and how to partition α_{FW} across the interesting and less interesting tests. The decision is subjective, but one can usually come up with sensible and defensible strategy for making the choice. See de Cani (1984) for a discussion.

11.3.5 Summary of the Basic Issues

Most solutions to the multiple test problem are predicated on the assumption of a true CNH. We assume a true CNH when applying the Bonferroni correction of B to account for the fact that one has conducted B hypothesis tests. But if the CNH is implausible, this means that it is very unlikely that all of the specific null hypotheses nested under the CNH are true, but it may be plausible that some smaller set of fewer than B specific null hypotheses are all true. Thus, indiscriminantly multiplying each p-value by B or dividing α_{FW} by B can be an overcorrection when the CNH is implausible. When the set of specific hypotheses are logically independent, the plausibility of the CNH declines as the number of specific hypotheses nested within the CNH increases. It also declines the more established a research area becomes. Thus, the multiple test problem is not as much of a problem, meaning correction doesn't have to be as strict, for logically independent hypotheses or in research areas that are well established. We need to be particularly concerned about the multiple test problem when the specific null hypotheses being tested are logically nonindependent, when the research area is new, and when exploring data and testing hypotheses not based on a priori predictions.

Using a Bonferroni correction of B when adjusting test results for the multiplicity of tests conducted is perfectly valid even when tests are logically independent and the CNH is implausible. The only harm in doing so is conservatism, meaning reduced power to detect false specific null hypotheses. Although a less strict multiple test correction may be justified in these circumstances, there is no empirical way of choosing that less conservative correction. But you can separate the tests into those that are more versus less interesting or important to your research objectives, and partition α_{FW} across these sets with a separate Bonferroni correction in each set. This can reduce the conservativism of the ordinary Bonferroni method for those hypothesis tests the results of which you care more about.

11.4 Chapter Summary

In this chapter we addressed the problem associated with conducting multiple hypothesis tests. If all null hypotheses in a set being tested are true and each is tested at the α level of significance, then the probability of at least one Type I error in the set—the familywise α or α_{FW}—is generally higher than α. If the number of tests B is large enough, then it may be almost certain that one of the conclusions resulting from the tests is a Type I error.

But this assumes that the CNH is true—that *all* of the B null hypotheses tested are true.

The Bonferroni method is a simple, albeit somewhat conservative approach to dealing with the multiple test problem. Its conservativism is a small price to pay given its enormous flexibility, how easily it is implemented, that it works for both independent and nonindependent tests, and that it can be applied to any hypothesis test that yields a p-value regardless of the specific statistical method being used. The Bonferroni method involves multiplying the p-value from a hypothesis test by the number of tests conducted and evaluating this "Bonferroni-corrected" p relative to the desired α_{FW}, rejecting the null only if the corrected p-value is less than α_{FW}. When this procedure is applied and the CNH is true, the probability of at least one Type I error in the set is no higher than α_{FW}.

Many investigators religiously invoke worries about the multiple test problem whenever a set of tests is conducted. The extreme form is to always apply the most strict version of the Bonferroni method, multiplying all p-values by B. This works, but this extreme position fails to acknowledge that there are various factors, some that can't be objectively quantified, that influence how concerned we should be about the multiple test problem. These factors include how plausible the CNH is, whether the tests in the set are logically independent, how well established the research area is, and the interest value of the hypotheses relative to the purpose of the study. Although never correcting for multiple tests may be hard to justify, applying a Bonferroni correction of B to all tests can be an overcorrection of the problem.

12
Nonlinear Relationships

Assuming linearity between two variables when modeling their relationship often results in reasonably good models that are useful and easy to interpret. But sometimes we have reason to believe a relationship is not linear, or the evidence compels us to accept that it is not. In spite of its name, linear regression analysis can be used to model relationships that are better described with curves than with straight lines. In this chapter we discuss reasons you might choose to fit a curve to a relationship rather than a straight line, and we show how to detect nonlinearity visually as well as using polynomial regression. We also give a brief overview of spline regression, an interesting extension of regression analysis that allows for chaining of line or curve segments to capture complex forms of nonlinearity. We end with a discussion of transformations, often used to make nonlinear relationships approximate linear ones.

12.1 Linear Regression Can Model Nonlinear Relationships

Relationships between variables are sometimes better described with curves than with straight lines. A graph showing world population on the vertical axis against time on the horizontal axis would constantly curve upward, with the growth accelerating rapidly with time. Human height against age rises more slowly during childhood than in the early teen years but levels off later. A plot of "commitment to democracy" versus the extent to which a person identifies as politically conservative versus politically liberal might show greater commitment among those in the middle of the ideology continuum than among those on either the liberal or the conservative end of the spectrum. Desire to acquire more money might be especially high among people who have very little, slowly drop off as in-

come increases, and perhaps climb again among people who are already very wealthy.

It may come as a surprise that a statistical technique called *linear* regression analysis can be used to fit curves. It can. In this chapter, we show some of the ways this is done.

12.1.1 When Must Curves Be Fitted?

In a scatterplot of Y against X, sometimes you can see that a curve better describes a relationship than does a straight line. It may be that you could easily draw a curve freehand through the scatterplot that seems to fit better than any straight line that a regression program would generate. But there are times when you need to go beyond this informal means of representing curvilinearity. These include

- When you must estimate Y from X.
- When you want to test whether the relationship is curvilinear against the null hypothesis that it is linear.
- When you must estimate the value of X at which Y is maximized or minimized, such as the amplification volume at which a person's speech is perceived clearest, or the length of rest breaks that maximizes productivity.
- When you must correct for a nonlinear relationship between Y and a covariate when studying the relationship between Y and independent variable X.

Consider the data represented by the scatterplot in Figure 12.1. It is obvious that no straight line adequately characterizes the relationship between X and Y. The best-fitting regression line of the form $Y = b_0 + b_1X$ is superimposed on the scatterplot. The equation for this line is $\hat{Y} = 3.289 - 0.220X$. It is the best-fitting line by the least squares criterion. In this example, $R = 0.591$, $SS_{residual} = 6.003$, and we know that no equation of this form would result in a smaller $SS_{residual}$ or larger R.

But consider a *quadratic* equation of the form $Y = b_0 + b_1X + b_2X^2$. This is the equation for a parabola. The equation $\hat{Y} = 1.254 + 1.597X - 0.359X^2$ is superimposed on Figure 12.1, which is the best-fitting parabola for these data. Just looking at the plot, it obviously fits much better than the linear model. Statistics confirm the better fit, as $R = 0.905$ and $SS_{residual} = 1.666$ for this equation, which was found simply by regressing Y on X and X^2. R is

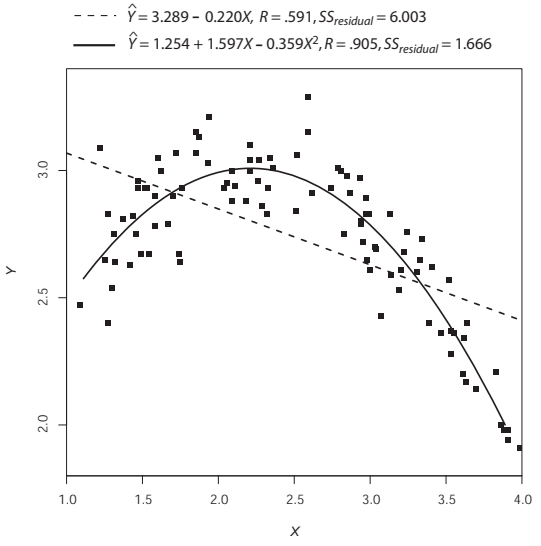


FIGURE 12.1. The best-fitting linear and quadratic model for these data.

much larger and $SS_{residual}$ is much smaller for the quadratic model than the linear model. So we have produced a better-fitting equation relating Y to X by adding the regressor X^2 to the model. Thus, linear regression analysis can be used to fit parabolas to data. Indeed, it can be used to fit other kinds of functions to data that are curves or semblances of curves.

This example illustrates each of the four points above. If your goal was to generate an estimate or prediction of Y from X, clearly you would do better using the model with X^2 than you would the model without it. You could also statistically compare the fit of the linear model to the nonlinear model to formally test whether the relationship is better described as curvilinear rather than linear. This would be the same as testing the null hypothesis that the regression coefficient for X^2 is equal to zero. Using calculus, you can derive that the estimated peak in Y occurs when $X = 2.224$; for those with a calculus background, the first derivative of the equation for $\hat{Y}$ with respect to X is $1.597 - 2 \times 0.359X$, which is equal to zero when $X = 2.224$. And suppose that X was a covariate. The procedures we described in Chapter 3 and elsewhere could result in improper control of X if you assumed that the relationship between Y and X was linear. But using

X^2 along with X as regressors in the model along with your independent variable of interest may reduce or eliminate this problem.

This latter point is worth developing further. Let the covariate be labeled C, and let X and Y be independent and dependent variables, respectively. Imagine that C has a mean near zero (either naturally or because you have made it so), meaning that C and C^2 are uncorrelated or nearly uncorrelated (we develop this point in section 12.2.4). Now suppose that Y is determined entirely by C^2 as $Y = C^2$. And further suppose that C also entirely determines X in the same way: $X = C^2$. Thus, $Y = X$, and both correlate zero or nearly so with C. If you failed to control for the curvilinear effect of C on Y, you would mistakenly conclude that X determines Y completely, when it actually has no effect at all, because Y is determined entirely by C.

The distortion in the apparent effect of X on Y occurs in this example because the relationship between X and C mirrors that between Y and C. But even in the absence of this, failure to control for curvilinear effects of covariates can distort results in the opposite direction by increasing $MS_{residual}$, which makes it harder to identify effects of X on Y that actually do exist, because all other things being equal, standard errors for regression coefficients are larger when $MS_{residual}$ is larger (recall equation 4.3).

12.1.2 The Graphical Display of Curvilinearity

When there are no covariates to complicate matters, a simple scatterplot depicting the relationship between two variables can be very useful both for seeing that a relationship is curvilinear and for discerning the nature of the curvilinearity. To take a few examples from the Roman alphabet, a scatterplot depicting nonlinearity between X and Y, with Y on the vertical axis and X on the horizonal axis, may look something like an L, with a sharp, rapid drop in Y as X increases, but a flattening of Y as X increases further. Or it could look like a U, with Y higher on the extremes of X than in the moderate values of X. The inverse of this would be a lowercase n, with Y lower in the extremes of X but higher in the middle of X. A J-shaped relationship would appear with Y relatively flat with increases in X with a sharp spike upward in Y once X reaches a certain value. Other forms of nonlinearity that are possible may not look like letters from the alphabet.

However, it is more difficult than you might think to depict or discern a nonlinear relationship between X and Y when there are covariates. If we have an independent variable X, a dependent variable Y, and one or more covariates C, and if there is a curvilinear relation between X and Y

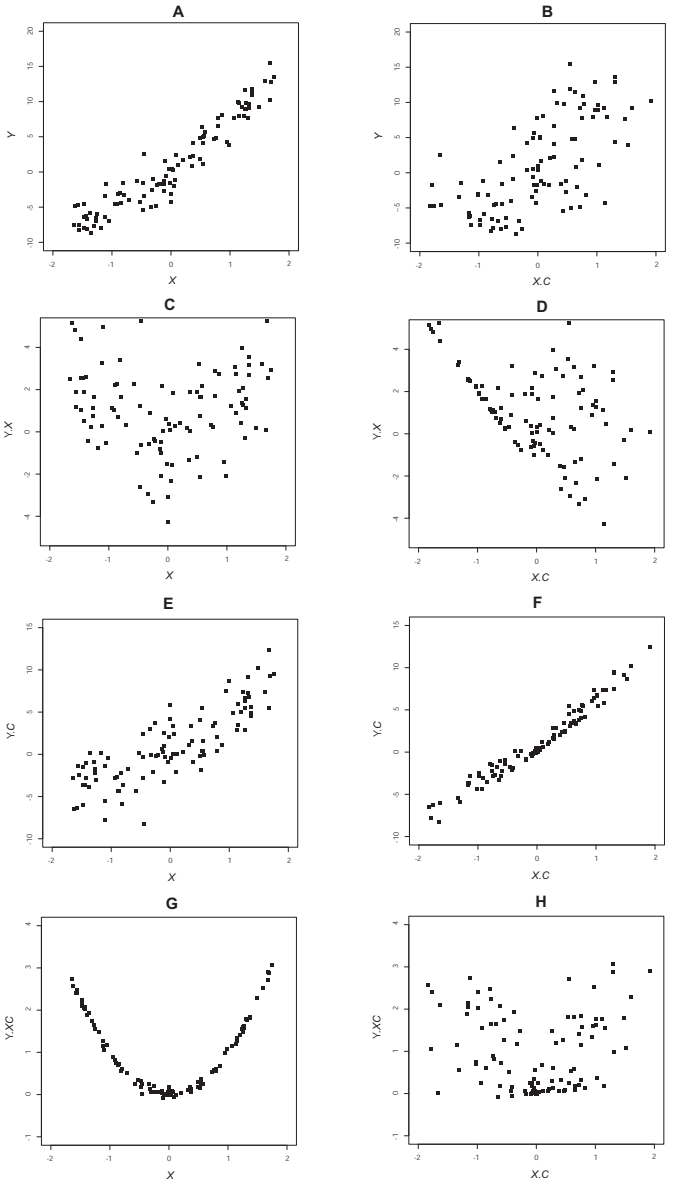


FIGURE 12.2. Eight possible scatterplots of *Y* against *X* with a covariate *C*. Plot G is the *residual scatterplot.*

when covariates are controlled, then there are no fewer than eight different scatterplots that we might think would display this curvilinearity. This is because there are four "forms" of Y we might consider: Y itself, the portion of Y independent of X ($Y.X$), the portion of Y independent of C ($Y.C$), and the portion of Y independent of X and C ($Y.XC$). In Chapter 3 we discussed that these portions of Y are residuals from a regression (e.g., $Y.C$ is the residual from a regression estimating Y from C). There are also two forms of X we might consider: X itself, and the portion of X independent of C ($X.C$). By combining the four forms of Y with the two forms of X, we can generate eight different scatterplots that we might imagine would display any curvilinearity between X and Y. And indeed any of these eight will work if X is independent of C and neither X nor C has any linear effect on Y. But abandoning any one of these three conditions can make the curvilinearity invisible, or nearly so, in four of these eight scatterplots, abandoning a second condition makes it invisible, or nearly so, in two more, and abandoning a third makes it invisible, or nearly so, in one more. The only scatterplot that is impervious to violations of all three conditions is the *residual scatterplot*, which is the plot of $Y.XC$ against X.

This point is illustrated in Figure 12.2, which shows these eight scatterplots for a sample with two regressors. This artificial data set is fairly typical, except that Y was defined as an exact nonlinear function of X and C to make any nonlinearity as visible as possible. The exact definition of Y used was $Y = 5X + 1X^2 + 10C$, meaning X is nonlinearity related to Y when C is controlled. Curvilinearity is clearly visible only in the residual scatterplot, which is plot G in the lower left corner ($Y.XC$ against X). The semipartial scatterplot (Y against $X.C$) described in section 3.3.1 is plot B, and the partial scatterplot ($Y.C$ against $X.C$) described in section 3.3.2 is plot F. They can hide even substantial nonlinearity. Curvilinearity is barely visible in plots C ($Y.X$ against X) and H ($Y.XC$ against $X.C$) but is crystal clear in the residual scatterplot.

Residual scatterplots provide the best graphical method for detecting nonlinearity and discovering its nature, but they have a major limitation that creates the need for nongraphical methods. One limitation of any graphical approach is the inefficiency of the human eye in detecting nonlinearity. This is illustrated in Figure 12.3. If you didn't know otherwise, you would probably think that the relationship between X and Y depicted there is linear. Yet in these data, nonlinearity is statistically significant at the .01 level and can easily be detected by polynomial regression introduced in section 12.2, even though that nonlinearity is essentially invisible

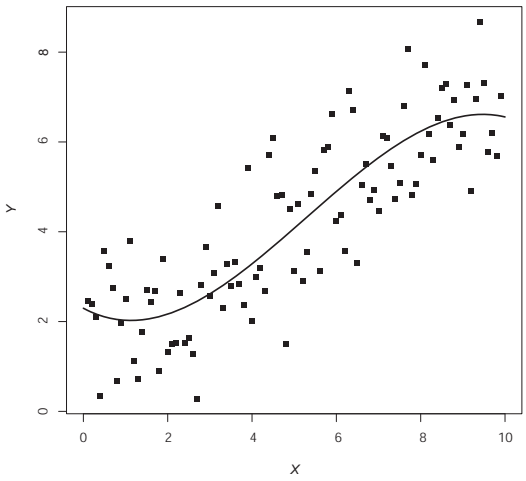


FIGURE 12.3. Real nonlinearity is sometimes hard to see in a scatterplot.

to the eye. This can be particularly problematic when there are nonlinear relations among regressors. In that situation, nonlinearity between one regressor and Y may be totally invisible even in a residual scatterplot.

An alternative problem is the tendency for the human mind to see patterns among even a random dispersion of dots. That is, you might think you see nonlinearity, but that nonlinearity is not actually present when formally tested. But whether it is failing to see real nonlinearity, or interpreting linearity as if it were nonlinearity, nongraphical methods are a good addition to and typically even better than graphical methods that rely on the subjective assessments of the perceiver. We cover some nongraphical methods in the next two sections.

12.2 Polynomial Regression

12.2.1 Basic Principles

Polynomial regression fits curves to data by using regressors that are successive powers, such as X, X^2, X^3, and so forth. The "order" of the polynomial is defined by the largest power in the polynomial. Figure 12.4 graphically depicts four equations relating Y to X. The linear equation is the one we have focused on throughout most of this book, where Y changes by the

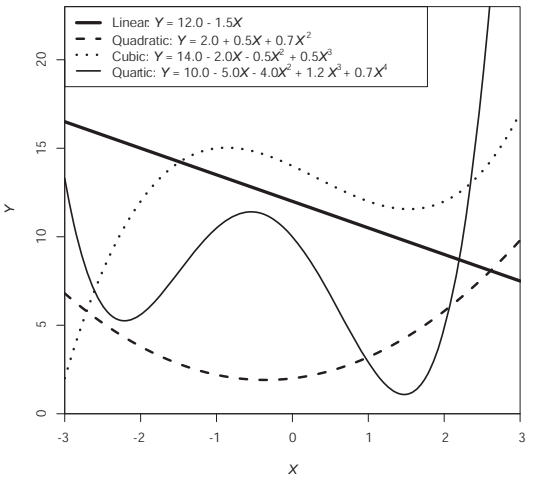


FIGURE 12.4. Some example polynomial models of the relationship between *X* and *Y*.

same amount as X increases by a fixed amount. A quadratic polynomial would take the form $Y = b_0 + b_1X + b_2X^2$ and is thus of "second order," because two is the largest power of X. A quadratic polynomial allows only a single "bend" in the relationship between X and Y, as in Figure 12.4. Adding a third power of X (thus yielding a "third-order" polynomial) results in a *cubic model*: $Y = b_0 + b_1X + b_2X^2 + b_3X^3$. This model allows for two bends in the curve, as can be seen in Figure 12.4. It would be exceedingly rare when using polynomial regression to add more than a third power of a variable to a model, but it is possible. Figure 12.4 depicts a quartic model, which by definition has a fourth power and thus is of the form $Y = b_0 + b_1X + b_2X^2 + b_3X^3 + b_4X^4$. This function allows three bends in the curve.

As Figure 12.4 depicts, the higher the order of the polynomial for X, the more complex the curve relating X to Y can be. The shape of the curve is also determined by the regression coefficients given to each of the powers of the variable. A characteristic of a polynomial of second order or higher is that the amount Y changes as X changes by a fixed unit depends on the starting point of X. So adding one unit to X will have a different effect on the amount Y changes depending on the value of X at which you start. Indeed, this is an informal definition of a curvilinear relationship between X and Y.

Some people criticize polynomial regression as excessively mechanical. Such critics argue that one should choose a curve whose shape makes scientific sense, which a polynomial may not. This is certainly good practice when possible. But polynomial regression can do a decent job representing curvilinear relationships that may not conform exactly to other kinds of functions (e.g., a logarithmic function; see section 12.4). Polynomial regression is also very versatile, because the shape a polynomial takes can be modified substantially by the amount of weight each power receives in the generation of Y, and your regression program will figure out how to weight each power in order to minimize $SS_{residual}$ and thus maximize the correlation between Y and $\hat{Y}$. Although it may be true that very few nonlinear relationships are truly parabolic, taking a U or inverted U shape, some nonlinear relationships between X and Y can be well described with a quadratic function within the domain of measurement of X.

Polynomials can also be nice ways of dealing with nonlinearity in covariates. Even if the relationship between an independent variable X and a dependent variable Y is linear, when those variables relate nonlinearly to a covariate C, it is important to allow for that nonlinearity in order to properly visualize and estimate the partial association between X and Y. We wouldn't typically care if the polynomial is a substantively or theoretically meaningful representation of the nonlinear relationship between a covariate and independent and dependent variables if it does a good job at capturing that nonlinearity and thereby affords a better adjustment for constructing measures of partial association between key variables in your analysis.

Polynomial regression is often used as a means of testing for nonlinearity in the relationship between X and Y. Because polynomials can describe such a wide range of curves, a test of nonlinearity can be conducted by determining if adding successive powers or sets of powers of X improves the fit of the model to a statistically significant degree. The test described in section 5.3.3 can be used for this purpose. We will see an example of this in section 12.2.2.

When a variable X is included as a regressor along with various powers of that variable, we usually think of that set of variables as a compound variable representing X. So, for example, if you think that age is nonlinearly related to something like attitudes toward gun control, you could use age as well as age^2 and perhaps even age^3 as regressors in the model. Any test involving age would involve all three of these. For instance, you could test whether gun control is related to age while controlling for sex

and income by adding age, age^2, and age^3 to a model of gun control that already contains income and sex. An improvement in fit as indexed by a statistically significant increase in R is evidence of a partial relationship between age and gun control, without imposing the assumption that this relationship is linear. But ordinarily, you would start with age and then decide whether adding powers of age improves the fit of the model, because it is easier to interpret linear relationships, and we wouldn't want to add an unnecessary complexity to a model unless the data (or relevant theory or past literature) suggested it was necessary to do so.

You would almost never include higher powers of a regressor in a model without including all of the lower powers as regressors as well. Consider, for example, the equation $Y = 2 + 3X^2$. This equation contains the second power of X but not the first. As a result, the line for this equation must pass through the point $X = 0, Y = 2$. This is very restrictive and not likely to be consistent with your data. When you include X, the function is no longer so restricted. Notice that $Y = 2 + 3X^2$ could be written in equivalent form as $Y = 2 + 0X + X^2$. Leaving X out of the model but including X^2 is like forcing the regression coefficient for X to be zero, and this is not likely to fit the data as well as if you let X's regression coefficient be something else. It is better to let your regression program figure out how to weight X in tandem with X^2 rather than imposing this constraint on the estimation process.

12.2.2 An Example

We illustrate polynomial regression using the POLITICS data file, which comes from a nationally representative survey of people living in the United States at the time of data collection. The dependent variable Y is score on a test of political knowledge (`pknow`), and we will estimate political knowledge from frequency of use of traditional news sources (X), named `news` in the data file. Participants in the study were asked three questions about how many days (0 through 7) during the typical week they read the newspaper, watch the national network news broadcast, and watch their local televised news broadcast. Responses to these three questions were averaged to produce the measure of traditional news use. We will look for evidence of nonlinearity between news use and political knowledge, while holding constant the respondent's age (C_1), sex (C_2), and SES (C_3, labeled `ses` in the data, defined as the average of the person's standardized education level and income).

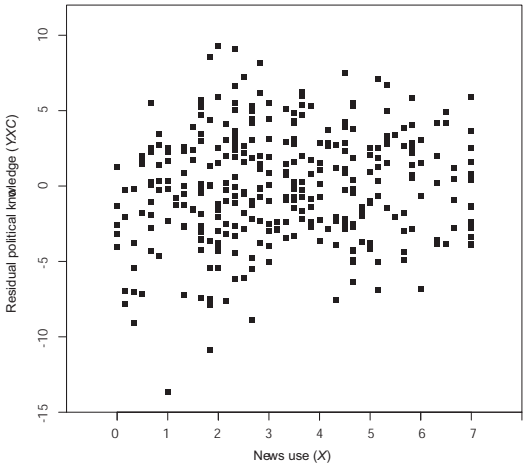


FIGURE 12.5. A residual scatterplot depicting the association between political knowledge and news use. The residuals are departures from estimated knowledge from a model that includes sex, age, SES, and news use.

Regressing political knowledge on news use X and the covariates, but without any higher powers of news use (without X^2, X^3, etc.), yields $R = 0.566$. The regression coefficient for news use is 0.265 and statistically significant, $t(335) = 2.214, p = .028$, indicating that people who use the news more frequently know more about politics. More specifically, two people who differ by 1 day in their typical news use but are equal on the covariates are estimated to differ by 0.265 units in their knowledge, with the more frequent news user being more political knowledgeable. But the meaningfulness of this depends on the partial relationship being linear.

Figure 12.5 is a residual scatterplot depicting the relationship between covariate-adjusted political knowledge and unadjusted news use. You can probably see some evidence of nonlinearity, as the residuals appear to be larger (more positive) in the center of the X distribution than in the extremes of X. But we should do a formal test.

When the square of news use is added to the model, the resulting model is

$$\hat{Y} = 7.168 + 1.372X - 0.156X^2 + 0.022C_1 + 1.720C_2 + 2.472C_3 \qquad (12.1)$$

The regression coefficient for the square of news use is statistically significant, $t(334) = -2.807, p = .005$. This test is equivalent to the change in the fit of the model when the square of news use is added to the model. Without X^2, $R^2 = 0.320$, but with X^2, $R^2 = 0.336$. This is a statistically significant increase, $F(1, 334) = 7.879, p = .005$. The increase in R^2 of .016 is the proportion of the variability in political knowledge uniquely attributable to the *square* of news use. If we wanted the proportion attributable uniquely to news use, we'd have to look at difference in the squared multiple correlations between a model that excludes news use *and* the square of news use, because news use is a compound variable in this model. Doing so, along with a test of significance as described in section 5.3.3, yields a difference of 0.026 in the two model R^2s, $F(2, 334) = 6.441, p = .002$. So news use uniquely accounts for about 2.6% of the variance in political knowledge.

Figure 12.6 visually depicts equation 12.1. This figure was generated by setting C_1, C_2, and C_3 to their sample means[1] and plotting estimated political knowledge for many values of news use (X and therefore X^2). As can be seen, holding age, SES, and sex constant, political knowledge is estimated as higher among those moderate in their news use, with more extreme users (less or more) estimated as lower in political knowledge. As you can see, the curvilinear effect is quite large even though it was barely visible in the partial scatterplot of Figure 12.5.

Just to make sure more complex curvilinearity is not missed, the cube of news use (X^3) was added to the model that includes news use and its square. The cubed term was not significant, meaning that adding it to the quadratic model does not improve the fit of the model to a statistically significant degree.

12.2.3 The Meaning of the Regression Coefficients for Lower-Order Regressors

We define a *global* property of a model as a property of the entire model, while a *local* property applies to only part of the model. For instance, a straight line relating X to Y has the same slope at all points, so the slope, estimated by the regression coefficient for X, is a global property of the model. But a curve defined by quadratic model that includes X and X^2 as regressors has different slopes at different points and may even slope downward in some sections but upward in others. Thus, the slope of

[1]It is legitimate to use the sample mean of a dichotomous variable when generating a plot such as this, even if the mean has no inherent meaning. In this case, sex is coded 0 for females and 1 for males, so the mean is the proportion of the sample that is male.

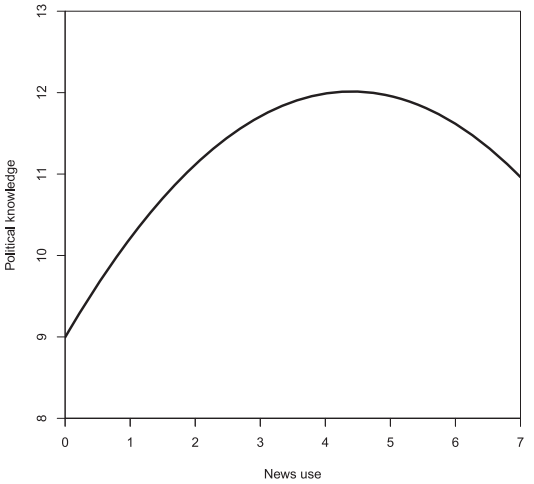


FIGURE 12.6. A quadratic polynomial model of political knowledge from news use frequency.

a curve defined by a quadratic model is a local property of the model. But a quadratic model is either concave, with a slope that becomes more positive as X increases, or convex, meaning that the slope is becoming more negative as X increases. So the concavity or convexity of a quadratic model is a global property of the model.

In a quadratic equation, it can be shown that if the regression coefficient for X^2 is positive, then the function is concave, but if this coefficient is negative, then the function is convex. It can also be shown that this regression coefficient measures the curvature of the relationship between X and Y, defined as the difference between the $\hat{Y}$ value of at any X point and the average of the two $\hat{Y}$ values corresponding to the values of X one unit to the left and one unit to the right. For instance, if $\hat{Y} = 2X^2$ and we arbitrarily use $X = 5$, then $\hat{Y} = 32$, 50, and 72 when X is 4, 5, and 6, respectively. We then have $(32 + 72)/2 - 50 = 2$, which is the coefficient for X^2. We would find the same value of 2 if we chose any other X value besides 5. Thus, the coefficient for X^2 measures a global property of the model, and we shall call X^2 a *global term* in the regression.

On the other hand, it can be shown that the coefficient of X in a quadratic equation measures the slope of the curve at the single point where $X = 0$. Readers who know calculus can see why this is so; if $\hat{Y} = b_0 + b_1X + b_2X^2$, then the first derivative of this function is $d\hat{Y}/dX = b_1 + 2b_2X$, which equals b_1 when $X = 0$. In the political knowledge example, $b_1 = 1.372$, and you can see by inspecting Figure 12.6 that this is about the slope of the parabola where it meets the Y-axis, when $X = 0$. Therefore, we call X a *local term*, since its regression coefficient measures a local property of the model.

This logic applies to higher-order polynomials, though understanding it requires knowledge of some calculus. For example, in a cubic model, the regression coefficient for X^3 is a global property of the model, but the regression coefficients for X and X^2 are local properties. In calculus terms, the first derivative of a cubic model $\hat{Y} = b_0 + b_1X + b_2X^2 + b_3X^3$ is $d\hat{Y}/dX = b_1 + 2b_2X + 3b_3X^2$. The first derivative is the slope of the curve at given point X, and you can see that if you set X to 0 in the equation for the first derivative, then you get b_1. Thus, b_1 is the slope of the curve when $X = 0$; thus, it is a local property of a cubic regression model.

The second derivative of a cubic model is $2b_2 + 6b_3X$. The second derivative quantifies how *quickly* and in what *direction* the slope is *changing* at a point X. This is sometimes called the acceleration of the function. If the second derivative is positive, that means that the slope is increasing as X is increasing in value. But if the second derivative is negative, that means that the slope is decreasing in value as X is increasing. The larger the second derivative ignoring sign, the faster the slope is changing. In this case, if you set X to 0 in the equation for the second derivative, you get $2b_2$. So b_2 is one-half of the speed at which the slope is changing at the point $X = 0$. This makes b_2 a local property in a cubic regression model.

12.2.4 Centering Variables in Polynomial Regression

A variable is *mean-centered* by subtracting its mean from all measurements, creating a new variable with a mean of zero. A variable can be mean-centered relative to its sample mean, or relative to its population mean if that happens to be known. There are two reasons why you might choose to mean-center X in a polynomial regression involving powers of X.

First, if $\overline{X}$ is high relative to s_X, then the successive powers X, X^2, X^3, and so on, might correlate with each other so highly that rounding error is produced, or you will reach the lower limit on the tolerance for a regressor that your regression program allows. For instance, in a sample of size $N = 5$ containing the values of X equal to 1,000, 1,001, 1,002, 1,003,

and 1,004, the correlation between X and X^2 is 0.99999983, which may be large enough to start introducing nontrivial rounding error into some regression computations. This can be corrected by centering X around its mean before computing the powers of X. If we subtract 1,002 from these five measurements (which is their mean), they become −2, −1, 0, 1, and 2, and now the correlation between X and X^2 is exactly zero. This can reduce computational problems and allow your regression program to estimate the model.

The second reason for mean-centering X before computing powers of X is that the regression coefficient for X is then the effect of X on Y at the mean of X, instead of when $X = 0$. This is likely to be more interpretable. A proof of this point was given in section 12.2.3.

Mean-centering a variable has no effect on regression coefficients for regressors or correlations when only first-order terms are used (e.g., X itself). But the situation with polynomial regression is more complex. Measures of simple relationship, such as correlations or simple regression coefficients, are affected for all but the first-order terms. For instance, if five measurements on X are 1, 2, 3, 4, and 5, then the five values of X^2 are 1, 4, 9, 16, and 25. But if we subtract 5 points from X before computing X^2, the new X values are −4, −3, −2, −1, and 0, and the new values of X^2 are 16, 9, 4, 1, and 0. Thus, the cases having the highest values on X^2 originally now have the lowest values. This, of course, will change the correlation between X^2 and other variables.

Measures of unique contribution for X or one of its powers, such as b_j, pr_j, sr_j, and the values of t or F that test their significance, are affected by centering for all but the highest power term. This is illustrated in Figure 12.7. Consider curve A. Its equation is $Y = 11.75 - 5.50X + 0.75X^2$. The slope of this curve is negative at $X = 0$ because Y is decreasing as X increases past zero. Thus, the regression coefficient for X is negative (see section 12.2.3). If we subtract 6 from X, then the curve shifts and becomes curve B in Figure 12.7, which has the same shape as curve A but is shifted horizontally in space. The equation for this curve is $Y = 5.75 + 3.50X + 0.75X^2$. Its slope at $X = 0$ is the coefficient for X, which is now positive because Y is increasing as X increases past $X = 0$. So centering X has changed the value of the regression coefficient for X, but the regression coefficient for X^2 is unaffected.

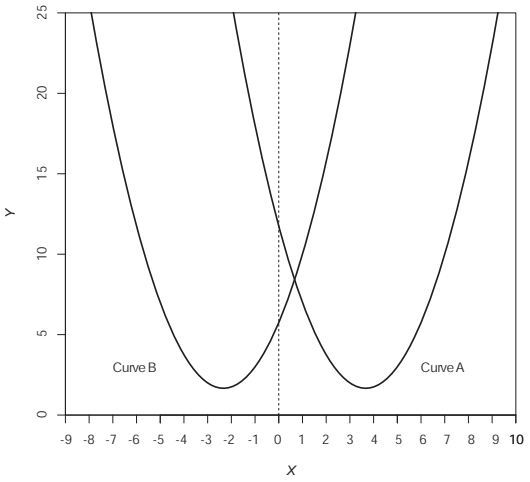


FIGURE 12.7. The effect of centering X on the regression coefficient for X in a quadratic model.

12.2.5 Finding a Parabola's Maximum or Minimum

Suppose you have estimated a model of the form $Y = b_0 + b_1X + b_2X^2$. The model could contain additional regressors as well, without changing the discussion that follows. However, the model should not include any regressors formed as the product of X and some other variable. The reasons for using a regressor that is a product of variables is discussed starting in Chapter 13.

In such a model, the value of X that either maximizes or minimizes Y (when all other variables are held constant, if the model contains additional regressors) is

$$X = \frac{-0.5b_1}{b_2} \tag{12.2}$$

Readers familiar with calculus will recognize this as the value at which the first derivative of Y with respect to X is equal to zero. The first derivative of a function of X with respect to X quantifies the amount Y is changing as X changes at a particular value of X. In a parabola, there is a point at which Y stops increasing or decreasing with changes in X and then "reverses course," such that if it was increasing with X, it now begins to decrease, or if it were decreasing with X, it now begins to increase. This point is either the minimum or maximum value. If the sign of b_2 is positive, then this

value is a minimum. If the sign of b_2 is negative, then this is a maximum value. But keep in mind that this point may not be within the range of the observed data.

To illustrate, in the political knowledge example in section 12.2.2 we had $b_1 = 1.372$ and $b_2 = -0.156$. Applying equation 12.2 gives $X = -0.5(1.372)/-0.156 = 4.397$. So we can say that holding constant education, age, sex, and SES, political knowledge is at its peak among those who use traditional news sources a bit over 4 days per week. We know it is a maximum and not a minimum because b_2 is negative, and we can also tell this from Figure 12.6.

12.3 Spline Regression

The scatterplot in Figure 12.8 depicts the association between two variables X and Y. As can be seen, the relationship is complex, with Y increasing with increasing X in some ranges of X, but decreasing Y with increasing X in other ranges. After reading section 12.2, you might think a quartic function would fit these data well. This would involving estimating Y from X, X^2, X^3, and X^4. Doing so results in $\hat{Y} = 5.161 + 6.217X - 0.845X^2 + 0.037X^3 - 0.001X^4$, and $R = 0.883$. This function is depicted in Figure 12.8 with the curve running through the scatterplot. It is apparent that even though R is fairly large, there is quite a bit of room for improvement. Observe that the vast majority of residuals are positive when X is between about 6 and 13, most are negative when X is between about 16 and 23, most are again positive between 23 and 26, and then again mostly negative beyond 26. This model is consistently underestimating Y in some ranges of X but overestimating Y in other ranges.

Spline regression is an alternative to polynomial regression. *Segmented regression* might be a better term, as the methods we discuss here all focus on fitting a set of models to various segments of the relationship between X and Y. But we will stick with the traditional term *spline regression*. Spline regression can model complex curves and do many other things, such as fitting lines with different slopes in different ranges of X. It can also be used when Y is expected to abruptly jump up or down at a specific value of X.

In this section, we introduce the fundamentals of spline regression, focusing first on *linear* spline models, which approximate a complex curve with a set of straight lines that are connected at joints. After describing these fundamentals, we discuss polynomial spline regression, which connects polynomials at joints. Polynomial splines are more versatile and therefore

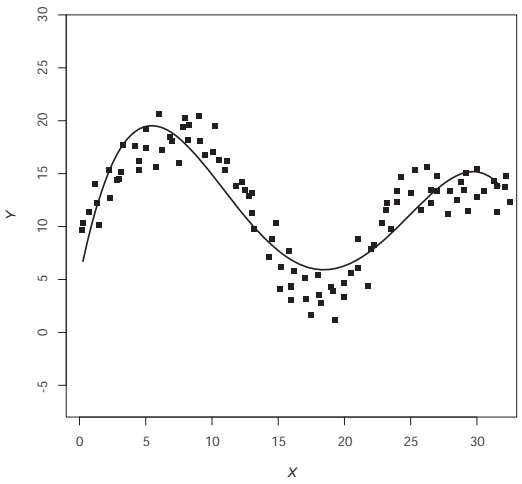


FIGURE 12.8. A scatterplot depicting a complex relationship and a quartic model superimposed.

more useful than linear splines, but you may find occasion to use linear splines, and it is easier to understand polynomial spline models by first learning how linear spline models work.

As a category of methods rather than a single method, spline regression includes more complex variants than we describe here. For a discussion of some of these more complex variants and their applications, see Ahlberg, Nilson, and Walsh (1967), Greville (1969), and Marsh and Cormier (2002). We focus only on methods that can be applied with an ordinary regression program.

12.3.1 Linear Spline Regression

In its simplest form, linear spline regression is a method for fitting to data a jagged line, like the solid line in Figure 12.9. Observe that this "curve" is formed by the four line segments that are joined together. By increasing the number of line segments, even extremely complex shapes can be fitted. The user of linear spline regression chooses the values of the regressor X but not the Y values that define the "joints" in a spline model. In Figure 12.9, these are marked J_1, J_2, and J_3. These could be chosen after examining a scatterplot, as we did in this example, or they could be chosen before

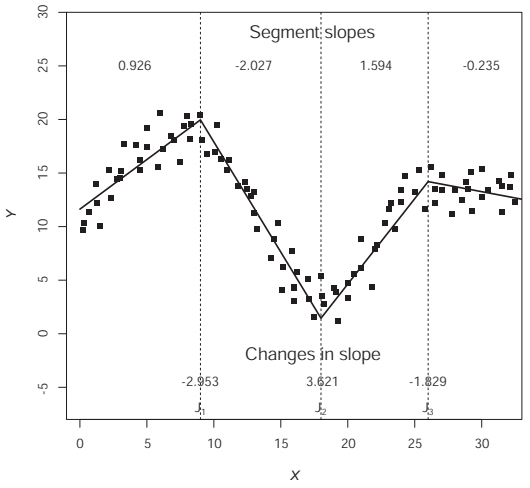


FIGURE 12.9. A linear spline regression model with three joints.

examining the data if you had an a priori basis for expecting a change in the relationship between X and Y at some values of X.

Spline regression using linear splines essentially estimates the slope of each line segment relating X to Y by computing the slope of the first segment and then the *change* in the slope at each joint. In Figure 12.9, the slopes of the four line segments displayed are 0.926, −2.027, 1.594, and −0.235, so a spline regression would estimate the changes in slope at J_1, J_2, and J_3 as −2.953, 3.621, and −1.829, respectively. These changes in slopes will be manifested in the regression solution as the regression weights for artificial variables created based on values of X.

To see how this is achieved, consider Figure 12.10. Line segment A, which applies when $X \leq 4$, is defined by the equation $Y = 1.00 + 1.00X$. Line segment B applies when $X > 4$, and it is defined by $Y = 13.00 - 2.00X$. That is,

$$Y = 1.00 + 1.00X \text{ when } X \leq 4 \text{ (segment A)}$$
$$Y = 13.00 - 2.00X \text{ when } X > 4 \text{ (segment B)}$$

Suppose we used the formula for line segment A to estimate all the Y values, regardless of whether or not X was greater than 4. As can be seen in Figure 12.10, doing so fits the Y values of the first four points perfectly,

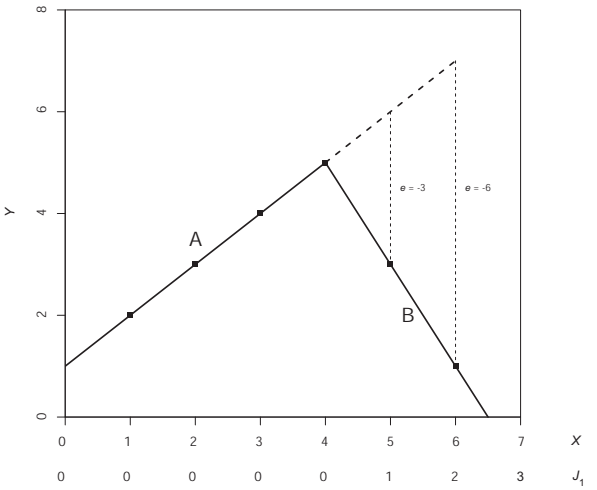


FIGURE 12.10. Why spline regression works.

but extending it beyond $X = 4$ (the dotted section of the line representing a continuation of line segment A) overestimates the next two Y values by 3 and 6 units, respectively. What we want to do is find a way of integrating the equations above into one equation that applies regardless of X.

Here is how we do it. Suppose we create a variable J_1 set to 0 when X is 4 or less, but set to $X - 4$ when $X > 4$. These values of J_1 can be found on the horizontal axis in Figure 12.10 below the values of X. Let e be defined as the errors in the estimation of Y from the equation for line segment A: $1.00 + 1.00X$. Notice that $e = 0$ when $J_1 = 0$, $e = -3$ when $J_1 = 1$, and $e = -6$ when $J_1 = 2$. In other words, $e = -3 \times J_1$. So an equation that perfectly fits the Y data would be

$$\hat{Y} = 1.00 + 1.00X - 3.00J_1 \tag{12.3}$$

This is the equation for the jagged line AB, and it integrates the equations for segments A and B into one equation. Observe that the line segment A has a slope of 1.00, and line segment B has a slope of −2.00, so the difference between these slopes is −3.00, which is the coefficient for J_1 in equation 12.3. So by creating the variable J_1 in this fashion, we were able to model the amount the slope of the line relating X to Y changes once X is higher than the location defining the joint. This example is atypical in that we do not

normally achieve a perfect fit. But once the X values of the joints have been selected, the regression program will fit the jagged line that minimizes the sum of the squared residuals and maximizes R.

Returning to the more complex example in Figure 12.9, using this approach we can fit a series of line segments with different slopes for different ranges of X but connected to each other at the joints. In Figure 12.9 there are three joints at X values of 9, 18, and 26. So we construct three new variables defined as X minus the joint location, but conditioned on X exceeding that joint value. If X does not exceed the joint value, then that variable is set to zero. In this example,

$$\begin{aligned}
&\text{if } X > 9, \text{ then } J_1 = X - 9 \text{ else } J_1 = 0\\
&\text{if } X > 18, \text{ then } J_2 = X - 18 \text{ else } J_2 = 0\\
&\text{if } X > 26, \text{ then } J_3 = X - 26 \text{ else } J_3 = 0
\end{aligned}$$

Once J_1, J_2, and J_3 are created, then regressing Y on X, J_1, J_2, and J_3 yields the equation

$$\hat{Y} = 11.628 + 0.926X - 2.953J_1 + 3.621J_2 - 1.829J_3 \tag{12.4}$$

The regression weight for X is the slope of the first line segment, and the values of b_j for J_1, J_2, and J_3 equal the changes in slope at joints J_1, J_2, and J_3. As in other forms of regression, regression programs routinely provide a test of significance for each b_j. When b_j represents a change in slope, we are testing the null hypothesis of no change in slope. In this example, these changes in slope at each joint are all statistically significant. Joints with nonsignificant changes may be deleted, given that the results in such a case suggest that there is no change in the size or direction of the association at that joint.

For this model, $R = 0.948, F(4, 95) = 209.101, p < .001$. This is a decent improvement from the quartic model (recall that in that model, $R = 0.883$), and as can be seen by comparing Figure 12.8 and Figure 12.9, the linear spline model does a better job estimating Y across the range of X. As discussed in section 4.3.2, the F-ratio for this model tests the null hypothesis that ${}_TR = 0$. This can be interpreted as a test of the null hypothesis of no relationship between X and Y, where X is a compound variable consisting of X itself as well as J_1, J_2, and J_3. We can test whether the relationship between X and Y is linear against the null hypothesis that it is nonlinear using the method in section 5.3.3. First estimate a model of Y from X alone.

Then add the J variables to the model. A statistically significant increase in R means that the linear spline model fits better than the ordinary linear model. In this case, the model with just X has $R = 0.263$, whereas the model with X and the three J variables has $R = 0.948$. This is a statistically significant increase, $F(3, 95) = 257.351, p < .001$. The linear spline model fits better than the simple linear model.

To better understand how this test works, consider that the model with just X as a regressor is equivalent to the spline model but with the constraint that all the regression weights for the J variables are equal to zero, meaning no change in slope at the joints. If the spline model fits better, then allowing for at least one joint with a change in slope produces a better-fitting model.

But we cannot use this test to compare the fit of this spline model to the quartic model. This test works only when the model with more variables (the spline model) contains all the same variables as the model with fewer variables (the quartic model), plus at least one extra variable. The J variables are not the same as the X^2, X^3, and X^4 variables, so we can't formally test the significance of the difference in fit of these two models.

However, there is an alternative approach that can be used to assess the relative value of the polynomial (X^2, X^3, and X^4) and spline terms (the J regressors). Combining these two models as

$$\hat{Y} = b_0 + b_1X + b_2J_1 + b_3J_2 + b_4J_3 + b_5X^2 + b_6X^3 + b_7X^4 \tag{12.5}$$

yields $R = .952$ when applied to these data. We can ask how much the polynomial terms add to fit by removing them from equation 12.5 and seeing if fit is significantly worse (which is the same as asking whether adding the polynomial terms to the linear spline model significantly improves fit). We already know that the linear spline model has $R = 0.948$. When the polynomial terms are added to the model, the test from section 5.3.3, which is appropriate here, does not quite achieve statistical significance, $F(3, 92) = 2.564, p = .059$. But when only the linear spline terms are removed from equation 12.5, the result is the quartic model, and we know that for this model $R = 0.883$. This reduction in fit relative to the combined model is statistically significant using this same test, $F(3, 92) = 41.034, p < .001$. That is, the inclusion of the linear spline terms significantly improves the fit of the model relative to when Y is modeled as a quartic function of X.

12.3.2 Implementation in Statistical Software

Although spline regression is not built into any commonly used statistical software packages of which we are aware, it can be implemented with any regression program. Assuming *X* is in your data and named as such, the SPSS code below constructs the three *J* variables in the four-segment linear spline model described in section 12.3.1 and then estimates the model.

```
compute j1=0.
compute j2=0.
compute j3=0.
if (x>9) j1=x-9.
if (x>18) j2=x-18.
if (x>26) j3=x-26.
regression/dep=y/method=enter x j1 j2 j3.
```

Assuming the data reside in a file named SPLINE, the comparable SAS code is

```
data spline;set spline;j1=0;j2=0;j3=0;
   if (x>9) then j1=x-9;if (x>18) then j2=x-18;if (x>26) then j3=x-26;
run;
proc reg data=spline;
   model y=x j1 j2 j3;
run;
```

and in STATA, use

```
gen j1=0
gen j2=0
gen j3=0
replace j1=x-9 if x>9
replace j2=x-18 if x>18
replace j3=x-26 if x>26
regress y x j1 j2 j3
```

The RLM macro documented in Appendix A has an option for linear spline regression. The user specifies the location of the joints, and RLM constructs all of the necessary *J* variables and then estimates the model. For instance, the SPSS RLM command below is comparable to the SPSS code above.

```
rlm y=y/x=x/spline=9,18,26.
```

See Appendix A for information on the use of the **spline** option in RLM.

12.3.3 Polynomial Spline Regression

Linear spline regression works as means of modeling nonlinearity, because any curve can be approximated by a set of line segments tied together at joints. The more joints you include, the better the approximation to the curvilinearity, in the same way that an octagon approximates a circle better than does a pentagon. But one restriction of linear spline regression is that between joints, the relationship between X and Y is fixed to be linear. As a result, the curve ends up jagged, with "elbows" at the joints and potentially very abrupt shifts in slope at the joints. A polynomial model doesn't have this problem, but a polynomial may not fit the relationship between X and Y as well, as in this example.

Polynomial spline regression combines the strengths of both polynomial and linear spline regression while eliminating the largest weakness of each. This procedure fits a polynomial rather than a straight line within each segment of the regressor. In principle, one could model the relationship between joints with a polynomial of any order, but we focus only on parabolic models (i.e., involving X^2) between joints, because this is usually sufficient. This will allow for different models of the relationship between X and Y in the segments, but will produce a smooth curve (rather than a line) between joints, with sets of smooth curves tied together at the joint points and no jaggedness at the joints.

When we fitted straight lines between joints, we constructed new variables defined as a set of one or more new variables quantifying whether and by how much X exceeded a particular joint value. To fit polynomials between the joints, we follow a similar procedure except that the new variables are higher powers of X conditioned on X exceeding the joint value. So if you want to fit a parabola between joint values, then the new variable will be set to 0 if X is less than or equal to the joint value, but if X exceeds the joint value, then set the new variable to the square of how much X exceeds that joint value. For instance, if X ranged between 0 and 20 and you placed a joint at 10, then J_1 would be set to 0 unless $X > 10$. If $X > 10$, then J_1 would be set to $(X - 10)^2$. You could include a higher additional power if desired, such as $(X - 10)^3$, if you wanted to fit a cubic function, although in practice, squares will usually suffice. You would typically also

include these same powers of X in the model to allow a polynomial of the same order for the first segment.

The scatterplot in Figure 12.11 is the same as the scatterplot in Figures 12.8 and 12.9, with the quartic model superimposed as a dashed line. This is a nice smooth curve, but as discussed already, its fit leaves something to be desired. The solid line depicts a polynomial spline model, the splines defined by second powers of X. Clearly, this does a better job describing the relationship between X and Y. We now describe how this model was constructed and estimated.

Examination of the scatterplot suggests that an inverted parabola may characterize the relationship between X and Y for values of X below 11. Between 11 and 16, the relationship appears linear or nearly so. Between 16 and 22, we can see what appears to be an upright parabola, but the left side of an inverted parabola appears to describe the relationship between X and Y between the values of 22 and 25. Finally, for X higher than 25, the relationship between X and Y looks linear or nearly so. So we define the five segments of the range of X with X values of 11, 16, 22, and 25. These are depicted in Figure 12.11.

With these five segments defined, we then create four J variables set to zero unless X exceeds the joint value. If X exceeds the joint value, then the J variable is set to the square of the amount X exceeds that joint. The algorithm for constructing these four J variables is

$$\begin{aligned}
&\text{if } X > 11, \text{ then } J_1 = (X - 11)^2 \text{ else } J_1 = 0\\
&\text{if } X > 16, \text{ then } J_2 = (X - 16)^2 \text{ else } J_2 = 0\\
&\text{if } X > 22, \text{ then } J_3 = (X - 22)^2 \text{ else } J_3 = 0\\
&\text{if } X > 25, \text{ then } J_4 = (X - 25)^2 \text{ else } J_4 = 0
\end{aligned}$$

We then regress Y on X and X^2 (which fits a parabola to the first segment), as well as J_1, J_2, J_3, and J_4. The resulting model is

$$\hat{Y} = 8.688 + 2.884X - 0.207X^2 + 0.148J_1 + 0.513J_2 - 0.972J_3 + 0.507J_4 \quad (12.6)$$

with $R = 0.956$. It is represented by the solid line in Figure 12.11. Observe it is a smooth curve, with no jaggedness at the joints as occurs when using linear splines. The fit of this model is clearly superior to the quartic model, and it is not obvious in looking at the scatterplot how this model could be changed to improve it further. All of the regression coefficients in this model are statistically significant, with p-values below .0001.

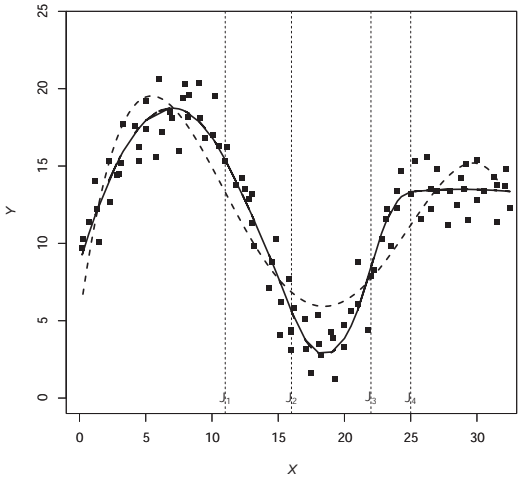


FIGURE 12.11. A quartic model (dashed line) and a polynomial spine model with four joints (solid line).

The code in section 12.3.2 can easily be modified to produce the *J* variables based on the algorithm above. For example, in SPSS, you can use

```
compute j1=0.
compute j2=0.
compute j3=0.
compute j4=0.
compute xsq=x*x.
if (x>11) j1=(x-11)*(x-11).
if (x>16) j2=(x-16)*(x-16).
if (x>22) j3=(x-22)*(x-22).
if (x>25) j4=(x-25)*(x-25).
regression/dep=y/method=enter x xsq j1 j2 j3 j4.
```

You may find with some statistics programs that the correlation between the regressors is sufficiently large that the model won't estimate, or the program may remove one or more of the regressors to deal with the near singularity (see section 17.3.3). If this occurs, center *X* around the mean of *X* (i.e., subtract $\overline{X}$ from all *X* values) and set up the joints and *J* variables using this transformed *X*. This likely will raise the tolerances of the regressors to more acceptable levels and may allow your program to estimate the model.

In section 12.3.1, we saw that the regression coefficients for the J variables quantify the difference in the slope relating X to Y between adjacent segments. In this quadratic spline regression model, the regression coefficients for the J variables quantify the change in curvilinearity of the relationship between X and Y between adjacent segments. Mathematically, this corresponds to the change in the regression coefficient for the squared term between adjacent segments. To see how this works, consider that for all values of $X \leq 11$, $J_1 = J_2 = J_3 = J_4 = 0$. So equation 12.6 reduces to $\hat{Y} = 8.688 + 2.884X - 0.207X^2$. This is the model relating X to Y when $X \leq 11$. The regression coefficient for X^2 is -0.207.

For the next segment defined as $11 < X \leq 15$, $J_1 = (X - 11)^2$ and $J_2 = J_3 = J_4 = 0$, so equation 12.6 simplifies to $\hat{Y} = 8.688 + 2.884X - 0.207X^2 + 0.148(X - 11)^2$. A little algebra results in

$$\begin{aligned}\hat{Y} &= 8.688 + 2.884X - 0.207X^2 + 0.148(X - 11)^2 \\ &= 8.688 + 2.884X - 0.207X^2 + 0.148(X^2 - 22X + 121) \\ &= 26.744 - 0.372X - 0.059X^2\end{aligned}$$

Thus, when $11 < X \leq 16$, the model relating Y to X is $\hat{Y} = 26.744 - 0.372X - 0.059X^2$. The regression coefficient for X^2 is -0.059, which is a change of 0.148 relative to the regression coefficient for X in the segment defined by $X \leq 11$. Notice that the regression coefficient for J_1 is 0.148. It is statistically significant from zero. If it were not statistically significant, then J_1 could be excluded from the model, because this would mean that allowing for a shift in the curvilinearity of the relationship between X and Y at this joint does not improve the fit of the model to a statistically significant degree.

Using this same logic and algebra for each segment produces a quadratic model for each segment, with the regression coefficient for successive J variables quantifying the difference in the regression coefficient for X^2 in a given segment relative to the prior segment. These changes add up cumulatively, so you can derive the weight for X^2 for any segment by starting with the regression coefficient for X^2 and then adding up the regression coefficients for each successive J variable, stopping once you reach the desired segment. For example, the regression coefficient for X^2 for the fourth segment ($X > 22 \leq 25$) is $0.207 + 0.148 + 0.513 - 0.972 = -0.518$. You can see in Figure 12.11 that, indeed, the model in this segment looks like the left half of a downward pointing parabola, consistent with a negative coefficient for X^2. And for the segment defined as $X > 25$, the weight for X^2

is $0.207 + 0.148 + 0.513 - 0.972 + 0.507 = -0.011$. This too is consistent with Figure 12.11. Observe that the regression line is nearly straight in the last segment, as you would expect for a polynomial model with such a small weight for the squared term.

12.3.4 Covariates, Weak Curvilinearity, and Choosing Joints

In the examples of spline regression we have described, there were no covariates, and we chose where to locate the joints by eyeballing a scatterplot. Covariates are easily added to a spline regression model simply by including them as regressors, and no modification to the procedure is needed. But we saw in section 12.1.2 that nonlinearity in the partial association between X and Y may be hard to see unless you construct the right scatterplot. And in real data, nonlinearity in the simple or partial association between X and Y may be so weak that it can't be detected with the eye even in the proper scatterplot. In such cases, you may not be able to eyeball a scatterplot and figure out where to locate the joints.

We don't have any silver bullet solutions to this problem, but is important to acknowledge the problem exists. If your sample size is sufficiently large, one option is to use a large number of joints equally dispersed across the range of X and then estimate a linear or polynomial spline model as discussed here. As you know, the p-values for the regression coefficients for the J variables can be used to decide whether a change in slope or curvilinearity is needed at specific joints. If not, those joints can be deleted. You can iteratively apply this procedure, adding or removing joints until you settle on a model that is satisfying to you. There are more advanced versions of spline regression that don't require the joints to be specified by the analyst but, rather, are derived mathematically from the data. You can read about some of these methods in the literature on spline regression, including the references we provided earlier in this chapter.

When you choose joints by eyeballing a scatterplot or using an exploratory method such as that just described, the concern is overfitting the data. Choosing joints by examining the data will tend to increase the variance explained by X. When X is a covariate, this produces a conservative bias into tests on independent variables. But if X is an independent variable, then the bias is toward exaggerating the importance of X in explaining variation in Y, and the nonlinearity captured by your spline model may not replicate in another sample.

But joint values need not always be chosen arbitrarily or by exploring the data and looking at scatterplots for visual evidence of transitions in the

relationship. You may have some a priori basis for choosing certain joint values. For example, if X were time and Y were something like a stock price, you might know that at a certain point in time (perhaps even a point in time of your choosing), some event happened that you think would change the trajectory for Y, making it increase or decrease in a particular manner that is different from what it was before that point in time. Or perhaps X is score on some kind of psychological test, such as a test of depression. If you assume, believe, or hypothesize that the relationship between depression and some dependent variable of interest is different for people who are below a certain score on the test relative to those who are above it, then that score would be natural choice for a joint in a spline regression model.

12.4 Transformations of Dependent Variables or Regressors

The natural relationship between two variables may be nonlinear, but sometimes nonlinear relationships can be made linear or nearly so by some kind of *transformation* of one of the variables. There are many kinds of transformations, but we focus on *monotonic* transformations here. A transformation is monotonic if the original and transformed values have the same rank order, such that the highest value on the original variable is the highest after transformation, the second highest original value is the second highest transformed value, and so forth. Technically, we should distinguish between positive and negative monotonic transformations. What we have described just now is positive monotonic. A negative monotonic transformation exactly reverses the ranks, so that the highest original value is the lowest transformed value, the second highest original is the second lowest transformed value, and so forth. Unless we say otherwise, when we say *monotonic* assume we are talking about positive monotonic.

Monotonic transformations of a variable may produce as many as three benefits at once. The first we have already discussed: Two variables may be nonlinearly related in their original form, but linear if one or both is transformed monotonically. This often simplifies interpretation of regression results. The second benefit is that a transformation may improve the prediction of one variable from another. Third, they can make residuals more normally distributed. Normality of the errors in estimation (manifested as residuals in a specific analysis) is an assumption of linear regression analysis.

12.4.1 Logarithmic Transformation

A logarithmic transformation can be used when the importance of the difference between two values is judged to be proportional to their ratio rather than their absolute difference. For instance, if we are studying the effect of an animal's size on some feature of its behavior or structure, we might consider the difference between body weights of 100 and 200 kilograms to be no more important than a difference between 1 and 2 kilograms. In absolute terms, a difference of 100 kilograms is 100 times larger than a difference of 1 kilogram, but both ratios are 2:1. Whereas weight may be nonlinearly related to many things (e.g., brain size), a logarithmic transformation may make the relationship linear. Or if the difference between incomes of $50,000 and $100,000 has the same average effect on attitudes toward wealth as the difference between $10,000 and $20,000, then income will have a nonlinear relationship with attitude, but a logarithmic transformation can make the relationship linear.

Only positive numbers have logarithms, but there are many kinds of logarithms. The most commonly used logarithms are the common logarithm, also called a base 10 log and often denoted *log*, and the natural logarithm or base e log, most often denoted as *ln*. The common logarithm of a number X is the power of 10, which equals X. For instance, the common logarithms of 10, 100, and 1,000 are, respectively, 1, 2, and 3, because $10^1 = 10$, $10^2 = 100$, and $10^3 = 1{,}000$.

Whereas a common logarithm is a power of 10, a natural logarithm is a power of e, where e is approximately 2.718828. Like the number pi, e cannot be written exactly. The natural logs of 10, 100, and 1,000 are, respectively, 2.30259, 4.60517, and 9.21034, because $e^{2.30259} = 10$, $e^{4.60516} = 100$, and $e^{9.21034} = 1,000$. Natural logarithms are proportional to common logarithms; for any number X, the natural logarithm of X equals approximately 2.302589 times the common logarithm of X.

An interesting property of natural logarithms is that when two numbers A and B are nearly equal, the difference between their natural logarithms approximately equals the proportional difference between them. For instance, 63 is 5% larger than 60, and their natural logarithms are 4.1431 and 4.0943, which differ by .0488, which is close to 0.05. Thus, if the weights of two animals differed by 0.05 on a natural logarithm scale, you would know without calculation that one was about 5% heavier than the other. As two numbers approach equality, this relationship approaches exactness. For instance, the natural logarithms of 1,000 and 1,001 differ by .0009995, which to four significant digits is .001, or 1/1,000.

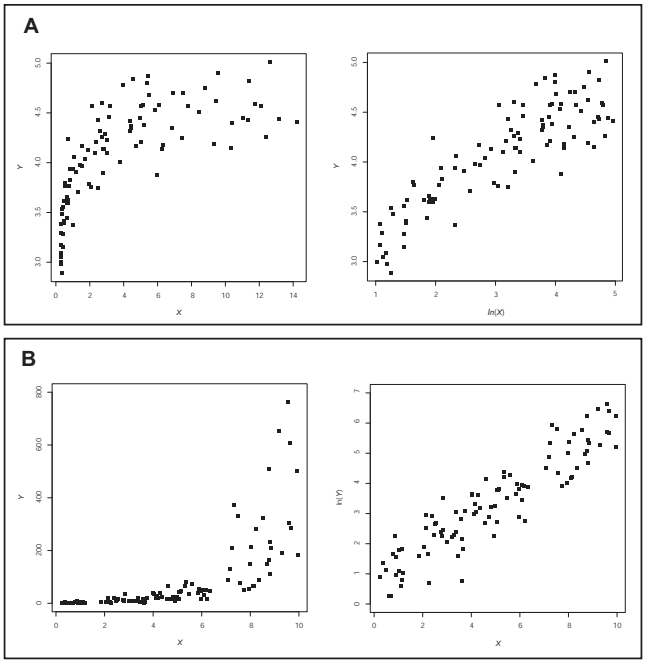


FIGURE 12.12. A log tranformation of *X* and *Y* can turn a nonlinear relationship into a linear relationship.

When a scatterplot depicting the association between two variables appears nonlinear, for some forms of nonlinearity a logarithmic transformation of *X* or *Y* may make the relationship more linear. If small changes in *X* result in large positive changes in *Y* at first but then the size of the change in *Y* levels off as *X* increases, as in Figure 12.12, panel A, on the left, then a logarithmic transformation of *X* may reduce or eliminate the nonlinearity. The scatterplot on the right of Figure 12.12, panel A, depicts the association between *X* and *Y* after a natural log transformation of *X*. As you can see, the relationship appears more linear after transformation than before.

But if small changes in *X* result in little changes in *Y* at first, but the change in *Y* with a change in *X* accelerates rapidly, as in Figure 12.12, panel B, on the left, then a logarithmic transformation of *Y* rather than *X* may reduce the nonlinearity. A scatterplot of the natural log of *Y* against *X* can be seen in the scatterplot on the right side of Figure 12.12, panel B; the

relationship between X and Y following the transformation now appears to be linear rather than nonlinear.

When using a logarithmic transformation, we often don't have to make distinctions between the different forms. If the common logarithms of X are linearly related to another variable Y, then the natural logarithms will be also. Thus, if we say that a logarithmic transformation makes a relationship linear, we need not specify which type of logarithm. But when reporting the results of an analysis that uses a transformation, it is a good idea to be explicit about what transformation was employed.

12.4.2 The Box–Cox Transformation

Box and Cox (1964) describe a family of transformations that includes logarithmic transformations as special cases. In this approach, one chooses a constant m, which may be any positive or negative real number (i.e., not zero). Then one transforms the original variable X to a transformed variable X_T by the equation

$$X_T = \frac{X^m - 1}{m} \tag{12.7}$$

In practice, you can try different values of m and see which one is best by some criterion of interest, such as making some extreme scores less extreme, improving linearity, or eliminating the need for an interaction (a concept introduced in Chapter 13). Although we use X in equation 12.7, the transformation can be applied to dependent variables, independent variables, or covariates.

Figure 12.13 displays the results of the transformation for $0 < X \leq 5$ for different values of m. The dashed line corresponding to $m = 1$ reflects no transformation (actually, when $m = 1$, $X_T = X - 1$). $X = 1$ is a pivoting point in the transformation, and what happens to the relative sizes of X after transformation depends on the distance from 1 and the value of m.

Define *measurement expansion* as making differences between values of X larger after the transformation, and define *measurement compression* as making differences between values of X smaller after transformation. Given these definitions, setting $m > 1$ results in measurement expansion when $X > 1$, with the expansion larger with higher values of m. But when $X < 1$, measurement compression is the result. But when $m < 1$, the transformation has the opposite effect on X. When $X > 1$, measurements are compressed, with greater compression occurring with smaller values of m. But when $X < 1$, measurement expansion occurs.

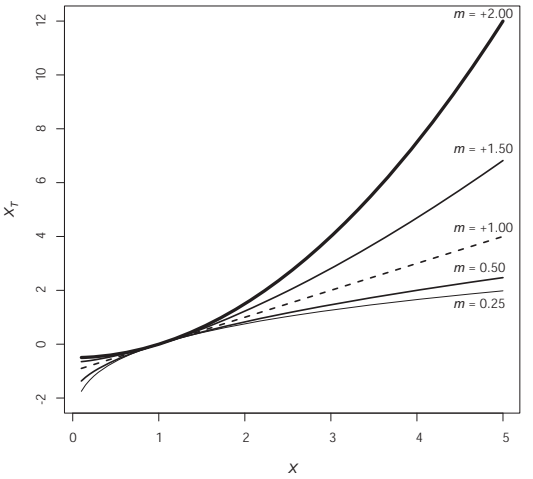


FIGURE 12.13. The Box–Cox tranformation as a function of *m*.

By making m arbitrarily close to zero (either positive or negative), we can make the Box–Cox transformation approach arbitrarily close to a logarithmic transformation. Thus, we can think of a logarithmic transformation as the special case of the Box–Cox transformation in which $m = 0$.

A Box–Cox transformation requires all measurements on the original variable to be positive since a negative number cannot be raised to a noninteger power. But if all measurements are negative we lose no information by replacing the original measurements with their absolute values before making the transformation. Thus, the requirement really is that all measurements have the same sign. This usually means that all measurements in the *population* must have the same sign, not just the measurements themselves, because inferences to the population have no meaning if some measurements in the population cannot be transformed.

Could one add points to a variable to make all its measurements have the same sign? Theoretically, a Box–Cox transformation is scientifically meaningful only if the original scale is a ratio scale—a scale with a meaningful zero point, so that it is meaningful to talk abut the ratios of two measurements. Thus, for instance, height and weight are ratio scales, but an attitude scale running from 1 to 9, with 1 denoting "very negative" and 9 denoting "very positive," is not. But in practice this restriction is not very important when using Box–Cox, because the effect of changing m is

often very similar to adding a constant to X before transformation. For instance, consider five cases scoring 1, 2, 3, 4, and 5 on a variable X. If we set $m = 0.5$, these five values transform to 0, 0.8284, 1.4641, 2.0000, and 2.4721, respectively. Thus, the second, third, and fourth transformed values are, respectively, 33.5, 59.2, and 80.9% of the distance from the first transformed value to the last. But if we add 7.841 to each of the original scores, then apply a Box–Cox transformation using $m = -1$, the percentages are nearly identical to the previous ones, now being 32.6, 59.2, and 81.3%. Thus, using $m = 0.5$ is nearly equivalent to adding 7.841 to each score and then using $m = -1$. Since trying different values of m is often very similar to adding different positive and negative constants to the original scores before making the transformation, the original zero point does not seem particularly sacrosanct.

12.5 Chapter Summary

Linear regression analysis can be used to model relationships between variables even when those relationships are not linear. It is always worth checking for nonlinearity by constructing a scatterplot, but it is important to construct the right scatterplot. The residual scatterplot is the best choice for detecting nonlinearity between X and Y when a model contains covariates. In a residual scatterplot, the residuals in the estimation of Y from X and the covariates are plotted against X. But even with the help of a residual scatterplot, the human eye is not very good at detecting relationships, so such eyeballing should be accompanied by some kind of formal analysis of nonlinearity.

Polynomial regression analysis is a versatile approach to testing for nonlinearity between X and Y, as well as modeling nonlinear relationships. This method involves estimating Y from X and successive powers of X, such as X^2 and, if desired, X^3 and (rarely) X^4. A statistically significant regression coefficient for one of the higher powers of X implies nonlinearity, as does an incremental increase in the fit of the model when one or more powers of X is added. Interpretation of the regression coefficients is complex and aided with an understanding of calculus. Most important is that in a model with a power of X higher than 1, the regression coefficient for X is a local term of the model and quantifies the relationship between X and Y when $X = 0$. Higher-order terms are interpreted in terms of changes in rates of changes of Y as X is changing.

Spline regression can be used to fit a jagged line to data. Any curve can be approximated by a set of jagged lines, and sometimes a spline model will fit better than a polynomial, because spline models better capture abrupt shifts in the relationship between X and Y as X increases or decreases. Spline and polynomial regression can be combined into polynomial spline regression. This involves estimating and tying together polynomials at various points in the distribution of X, thereby increasing the complexity of the kinds of curves that can be estimated.

Some nonlinear relationships can be made linear or nearly so through the use of a transformation, and transformations can sometimes help in meeting the other assumptions of regression. Logarithmic transformations of X and Y can be used in different circumstances, depending on the form of nonlinearity. A logarithmic transformation is a special form of the more general Box–Cox transformation. Using this transformation, the analyst selects an exponent in the function that produces the most appealing transformation, as defined by how well it makes a nonlinear relationship linear or removes skew or heteroscedasticity in the errors in estimation, for instance.

One could define a nonlinear relationship as one in which the relationship between X and Y depends on X. This could be thought of as a special kind of *moderation*, the topic of the next two chapters. With nonlinearity, X moderates its own effect on Y. In the following chapter we introduce how to build flexibility into a regression analysis by allowing the effect of a regressor X on Y to vary linearly with another regressor in the model.

13
Linear Interaction

This chapter relaxes the assumption built into all models discussed thus far that one regressor's effect on the dependent variable, expressed by its regression coefficient, is independent of the other regressors in the model. When one variable's effect depends on another, we say that the two variables *interact* or that one variable *moderates* the other's effect. We address the fundamentals of linear interaction, which allows one regressor's effect to be a linear function of another regressor. Unlike in ANOVA books and classes, where researchers are often first exposed to the concept of interaction, there is no requirement that all the variables interacting be categorical. Using any regression program, one can estimate and test hypotheses about interaction between numerical, dichotomous, or multicategorical variables in any combination.

13.1 Interaction Fundamentals

13.1.1 Interaction as a Difference in Slope

In all examples of linear regression analysis thus far, a regressor's effect, expressed in the form of its regression coefficient, is fixed to be invariant across values of the other regressors in the model. For instance, in a model of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2$, a 1 unit change in X_1 changes $\hat{Y}$ by the same amount, independent of the value of X_2. So when we say that X_1's effect equals b_1 when controlling for X_2, we are saying that when X_2 is held fixed, changing X_1 by 1 unit changes $\hat{Y}$ by b_1 units *regardless* of the value at which X_2 is held fixed. This idea was expressed visually in Figure 3.6, which represents X_1's effect as a set of parallel lines. The same is true for X_2 and its effect. In this model, a change of a given number of units in X_2 has the same effect on $\hat{Y}$ regardless of the value of X_1, as in Figure 3.5. When

plotted in three-dimensional space, such a model looks like a plane (as in Figure 3.2).

But one regressor's effect on the dependent variable may depend on another regressor in the model. When this happens, we say that the two regressors *interact*, or that there is an *interaction* between the two regressors in their effect on Y. Readers familiar with factorial ANOVA know about interaction as a difference in *simple effects*. For instance, in a two-condition experiment, it may be that the effect of the treatment relative to the control on the dependent variable is some value among men (the simple effect of treatment in men) but a different value among women (the simple effect of treatment in women). In that case, we'd say that experimental condition and sex interact. In other words, the average difference between the two conditions is different between men and women. So interaction means that *differences are different*.

We can generalize this definition of interaction to the general linear model by defining interaction as a *change in one regressor's relationship with Y when another regressor changes*. This definition subsumes the ANOVA definition as a special case. We saw in section 5.1.3 that a regression coefficient for a dichotomous regressor is a kind of difference, so a difference between differences in regression terms is akin to a change in a regression coefficient for one regressor as another regressor changes.

Another term commonly used to convey the concept of interaction is *moderation*. In this example, we would say that the effect of experimental condition (treatment vs. control) is *moderated* by sex, or that sex moderates the relationship between experimental condition and the dependent variable. So sex is a moderator variable, and experimental condition is often called the *focal predictor*. More generally, if X_2 moderates the relationship between X_1 and Y, this means that changing X_1 has a different effect on Y depending on the value of X_2. In this description, X_1 is the focal predictor, and X_2 is the moderator. But as will be seen in section 13.1.6, we can flip X_1 and X_2 and think of X_1 as the moderator and X_2 as the focal predictor without changing the mathematics of the model.

13.1.2 Interaction between Two Numerical Regressors

This broad definition of interaction applies when both variables are numerical, as in the following examples:

- Most physicians believe that the higher a patient's blood pressure, the greater the negative effect of being overweight on life expectancy.

This hypothesis specifies an interaction between blood pressure and weight in affecting life expectancy.

- If the relationship between political liberalism and SES is positive among people who are relatively younger but negative among people who are relatively older, then age and SES interact in affecting liberalism.

- If five persuasive political messages differ in their appeal to emotion versus reason, the most emotional appeals may be most effective in changing opinions among less educated voters, while the least emotional appeals may be most effective among more educated voters. If so, then education and strength of emotional appeal interact in affecting opinion change.

Interaction between two numerical regressors X_1 and X_2 is illustrated in Figure 13.1. Each line in this figure represents the regression of Y on X_1 for a particular value of X_2. It shows that the relationship between Y and X_1 is negative when X_2 equals 0, 1, or 2; it is zero when X_2 equals 3, and it is positive when X_2 equals 4 or 5. The effect of X_1 on Y for a particular value of X_2 goes by various names, such as a *simple effect*, a *conditional effect*, or a *simple slope*. If X_2 is numerical, then there may be infinitely many conditional effects of X_1; but if X_2 is categorical with g categories, then there are just g conditional effects of X_1.

Notice that the Y-intercepts in Figure 13.1 change as X_2 changes. But that was also true of the series of parallel lines in Figure 3.6, when there was no interaction between X_1 and X_2. The defining feature of interaction when depicted in visual form is nonparallel regression lines, or lines with different slopes, such as in Figure 13.1.

13.1.3 Interaction versus Intercorrelation

Interaction between regressors is sometimes confused with intercorrelation between regressors. So you might read a statement like "b_X was reduced to near zero by the addition of covariate C to the regression, because C interacts with X." But the writer clearly means to say that C *correlates* with X. The sentence in quotation marks says that the effect of X on Y (b_X) is affected by the *inclusion* of C, while *interaction* means that the effect of X on Y depends on the *value* of C. The interaction between X and C describes how X and C relate to Y when considered jointly, so it can be computed only when Y values are known and available. But the correlation between X and

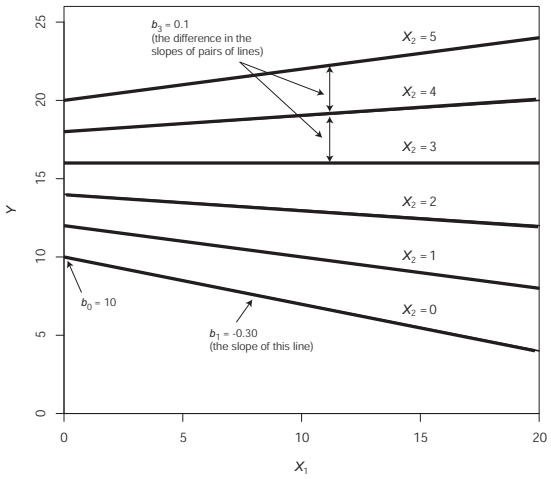


FIGURE 13.1. Simple linear interaction between X_1 and X_2.

C can be computed even if there is no third variable Y. When the addition of a covariate C to a model changes the effect of X on Y, there still is only one number that describes how changing X changes Y, and this applies to all cases. But interaction means that the amount changing X changes Y depends on the value of C. Depending on whether C is categorical or numerical, there may be a few such effects of X on Y, or a infinite number of them.

13.1.4 Simple Linear Interaction

Any pattern of nonparallel regression lines is an interaction. However, we focus on *simple linear interaction* in this chapter, which occurs when the effect of one regressor on Y is linearly related to the value of another regressor. This exists in Figure 13.1; the six regression lines depicted there can be found in Table 13.1. This table shows the six conditional effects of X_1 on Y for the six values of X_2. Observe that these six conditional effects have an exact linear relationship with X_2; each 1-unit increase in X_2 is associated with an increase of 0.1 in b_1.

We could formalize this linear relationship with the linear function $b_1 = -0.3 + 0.1X_2$, where -0.3 is the "intercept" and 0.1 is the "slope." Notice that this is an equation for a line. But it is a model not of Y but, rather, the conditional effect of X_1 on Y. Observe that when you set X_2 to

TABLE 13.1. The Six Regression Lines Depicted in Figure 13.1

X_2	b_0	b_1
0	10.0	−0.3
1	12.0	−0.2
2	14.0	−0.1
3	16.0	0.0
4	18.0	0.1
5	20.0	0.2

1, the function generates $b_1 = -0.3 + 0.1(1) = -0.2$, which is the conditional effect of X_1 when $X_2 = 1$ in Table 13.1 and the slope of the line relating X_1 to Y when $X_2 = 1$ in Figure 13.1. Other values of X_2 produce comparable conditional effects in Table 13.1 and slopes in Figure 13.1. So we have a linear model of the conditional effect of X_2 and hence *simple linear interaction*.

13.1.5 Representing Simple Linear Interaction with a Cross-Product

Simple linear interaction can be represented by a regression equation with two linear terms and a constructed *cross-product* term defined as the product of the two variables interacting. It looks like

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 \tag{13.1}$$

To see why, consider further Figure 13.1. We have just seen that the regression line when $X_2 = 0$ has slope −0.3 and the slope rises 0.1 for each 1-unit increase in X_2. In section 13.1.4 we said that we can represent this relationship between X_1's effect on Y and X_2 as $-0.3 + 0.1X_2$. We can also see that the Y-intercept of the regression line equals 10 when $X_2 = 0$ and increases by 2 (to 12, 14, etc.) for each 1-unit increase in X_2. In other words, the relation between the intercept and X_2 could be expressed as a function of the form $10 + 2X_2$. If we substitute these two functions into the simple linear regression equation $\hat{Y} = b_0 + b_1X_1$, this gives

$$\hat{Y} = b_0 + b_1X_1 = (10 + 2X_2) + (-0.3 + 0.1X_2)X_1$$

which can be written in equivalent form by multiplying X_1 through and removing parentheses as

$$\hat{Y} = 10 - 0.3X_1 + 2X_2 - 0.1X_1X_2$$

Thus, we see that a whole series of nonparallel regression lines can be represented by a single regression equation that includes the terms X_1, X_2, and X_1X_2.

We saw in Chapter 3 that a model without the X_1X_2 cross-product takes the form of parallel regression lines, which means X_1's effect on Y is constant across values of X_2. Such a model, $\hat{Y} = b_0 + b_1X_1 + b_2X_2$, generates parallel regression lines and can be thought of as a special case of equation 13.1, where b_3 is fixed to zero. That is, the parallel lines model, meaning no interaction, can be expressed as $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + 0X_1X_2$, whereas the nonparallel lines model expressed in equation 13.1 allows for interaction by allowing b_3 to be different from zero.

In equation 13.1, b_3 quantifies how much the conditional effect of X_1 changes as X_2 changes by 1 unit. As for any regression coefficient, a hypothesis test can be conducted on its value. No interaction in the population implies that ${}_Tb_3 = 0$. We can test the hypothesis of no interaction between X_1 and X_2 with a hypothesis test that ${}_Tb_3 = 0$ against the alternative hypothesis that ${}_Tb_3 \neq 0$. Any regression analysis program provides all the information needed to conduct this test, just as it does for all the other regression coefficients. Rejection of the null hypothesis means that X_1's effect on Y depends on X_2; that is, X_1 and X_2 interact, or X_2 moderates the effect of X_1 on Y. Alternatively, a confidence interval could be used to convey the two values between which ${}_Tb_3 = 0$ is likely to be. If zero is outside of the interval, this implies interaction.

13.1.6 The Symmetry of Interaction

Thus far we have discussed linear interaction by showing how in a model that includes X_1, X_2, and their cross-product X_1X_2, X_1's effect on Y depends linearly on X_2, meaning that X_1's effect on Y changes as X_2 changes. Interpreted that way, X_2 is the moderator and X_1 is the focal predictor. But we could have reversed the roles of X_1 and X_2 in the prior discussion and made X_1 the moderator and X_2 the focal predictor. That is, a parallel description of the interactive relationship between X_1 and X_2 tells us how the conditional effect of X_2 on Y changes as X_1 changes. That is, we could construe X_1 as a moderator of the effect of the focal predictor X_2 on Y.

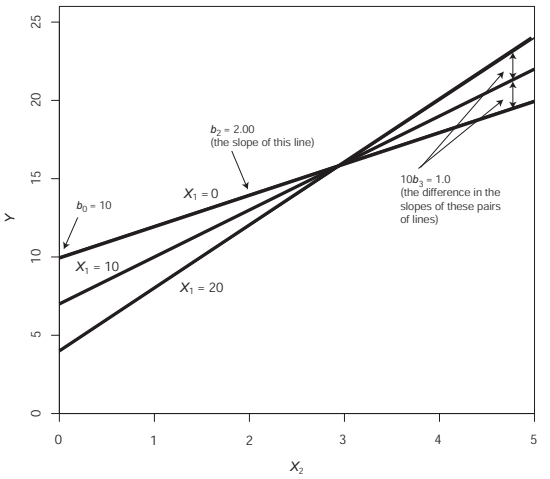


FIGURE 13.2. Another representation of simple linear interaction between X_1 and X_2.

Figure 13.2 is an alternative representation of the model $\hat{Y} = 10 - 0.3X_1 + 2.0X_2 + 0.1X_1X_2$. It looks very different than Figure 13.1, but in fact it represents the same model. In Figure 13.2 we have put X_2 on the horizontal axis and have shown different regression lines relating X_2 to Y for different values of X_1. Of course, this is only three of the many possible regression lines one could visualize.

The three regression lines depicted in Figure 13.2 are found in mathematical form $\hat{Y} = b_0 + b_2X_2$ in Table 13.2. Notice there that as X_1 rises by 10 units, the conditional effect of X_2 increases by 1 unit. Remembering that we chose X_1 values of 0, 10, and 20 arbitrarily, we could say more generally that as X_1 rises by 1 unit, the conditional effect of X_2 increases by 0.1 units. So we could represent the relation between X_2's conditional effect on Y and X_1 by the function $2.0 + 0.1X_1$. And observe that as X_1 increases by 10 units, the Y intercept decreases by 3 units. That is, as X_1 increases by 1 unit, the intercept *increases* by -0.3 units. In function form, the Y-intercept is related to X_1 as $10 - 0.3X_1$.

Using the same logic presented in section 13.1.5, we can substitute these two functions into the simple linear regression equation $\hat{Y} = b_0 + b_2X_2$ and this gives

$$\hat{Y} = b_0 + b_2X_2 = (10 - 0.3X_1) + (2.0 + 0.1X_1)X_2$$

TABLE 13.2. The Three Regression Lines Depicted in Figure 13.2

X_1	b_0	b_2
0	10.0	2.0
10	7.0	3.0
20	4.0	4.0

which can be written in equivalent form by multiplying X_2 through and removing parentheses as

$$\hat{Y} = 10 - 0.3X_1 + 2.0X_2 + 0.1X_1X_2$$

So, mathematically, the model is the same as when X_1 played the role of focal predictor and X_2 played the role of moderator. This is the *symmetry* property of interactions. Interaction between X_1 and X_2 means that the effect of X_1 on Y changes as X_2 changes *and* that the effect of X_2 on Y changes as X_1 changes. This change in the effect of one variable as the other changes is b_3, regardless of whether X_1 is the focal predictor and X_2 the moderator or the other way around. Regardless of how we construe the roles of X_1 and X_2, we can test a hypothesis of interaction using the estimate of b_3 and its test of significance in a regression analysis.

13.1.7 Interaction as a Warped Surface

In Chapter 3 we saw that an equation of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ can be represented either as a series of parallel lines or as a tilted plane in three-dimensional space. An equation with a cross-product term of the form

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

can be represented either as a series of *non*parallel lines, as in Figures 13.1 and 13.2, or as a *warped* surface in three-dimensional space. Figure 13.3 represents the model we've been discussing thus far: $\hat{Y} = 10.0 - 0.3X_1 + 2.0X_2 + 0.1X_1X_2$. The line in Figure 13.3 from $Y = 10$ to $Y = 4$ corresponds to the line labeled $X_2 = 0$ in Figure 13.1, and the line in Figure 13.3 from $Y = 20$ to $Y = 24$ corresponds to the line labeled $X_2 = 5$ in Figure 13.1. The other four lines from Figure 13.1 also appear in Figure 13.3. Any surface

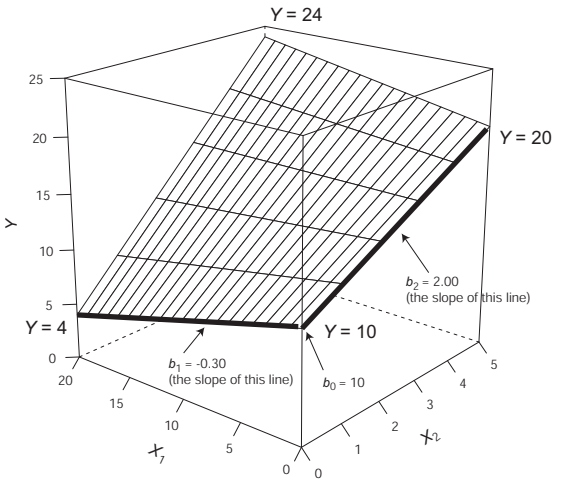


FIGURE 13.3. Three-dimensional representation of interaction.

representing a simple linear interaction looks like a bent piece of sheet metal, as in Figure 13.3.

13.1.8 Covariates in a Regression Model with an Interaction

Our discussion of linear regression with interactions generalizes to models with covariates. Covariates are added to a model with an interaction merely by including the covariates as additional regressors in the model, and the usual "holding the covariates constant" interpretation applies. As in an ordinary regression model with covariates, there is no practical limit to the number of covariates you could include. Only the sample size determines the limit. Because the cross-product term takes away 1 degree of freedom from $df_{residual}$, the maximum number of covariates you can include in a model with a cross-product term is $N - 4$.

13.1.9 The Meaning of the Regression Coefficients

In a model of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$, the following interpretations of the regression coefficients and regression constant apply regardless of how the variables are coded or scaled in the data.

- b_0 is the estimate of Y when both X_1 and X_2 equal zero (and, if the model includes covariates, when all covariates are equal to zero).

- b_1 is the estimated difference in Y between two cases that differ by 1 unit on X_1 but whose scores on X_2 equal zero (and are the same on any covariates). It is the conditional effect of X_1 when $X_2 = 0$.
- b_2 is the estimated difference in Y between two cases that differ by 1 unit on X_2 but whose scores on X_1 equal zero (and are the same on any covariates). It is the conditional effect of X_2 when $X_1 = 0$.
- b_3 is the estimated change in the conditional effect of X_1 as X_2 increases by 1 unit; alternatively, it is the estimated change in the conditional effect of X_2 as X_1 increases by 1 unit. When covariates are in the model, we impose the additional condition "holding the covariate(s) constant."

In this example, $b_1 = -0.3$ and is the slope of the line relating X_1 to Y when $X_2 = 0$, highlighted in Figures 13.1, 13.2, and 13.3. Here, $b_2 = 2.0$, which corresponds to the slope of the line relating X_2 to Y under the condition that $X_1 = 0$, as depicted in Figures 13.2 and 13.3. In section 14.4.4 we further emphasize the conditional nature of these effects, and that care must be taken when interpreting them so as to not misinterpret them as if they are "main effects" in an ANOVA sense, or "average" effects, which they very much are not.

The regression coefficient for the cross-product, $b_3 = 0.1$, requires more explanation and visualization. In this example, it is the difference in the conditional slope of the line relating X_1 to Y between two sets of cases that differ by 1 unit on X_2. This is easiest to see in Figure 13.1. It is the amount the slope for X_1 changes as you move up 1 point on the X_2 scale. Observe that the slope is increasing by 0.1 units as X_2 increases by 1 unit. By the symmetry of interactions, $b_3 = 0.1$ is also the difference in the conditional slope of the line relating X_2 to Y between two sets of cases that differ by 1 unit on X_1. In Figure 13.2, the slopes plotted are for values of X_1 that differ by 10 units. So the difference between consecutive conditional slopes for X_2 as you move 10 points up the X_1 scale is $10b_3 = 1.0$.

The regression constant, $b_0 = 10.0$ is the estimate of Y when both X_1 and X_2 are zero. As can be seen, the regression constant in this example appears in Figures 13.1, 13.2, and 13.3.

13.1.10 An Example with Estimation Using Statistical Software

In practice you will use a statistical package of some kind to estimate your regression models, and that includes models with a cross-product when

testing for interaction between regressors. All programs allow the analyst to construct a new variable defined as the product of X_1 and X_2 and then include it as a regressor along with X_1 and X_2. The regression coefficient and its test of significance is provided by the software, along with a confidence interval if desired.

We illustrate using the HOSPITAL data file available on this book's web page at *www.afhayes.com*. These data are fabricated but are motivated by a study of hospital workers described in Halbesleben (2010). Three hundred health care workers at a hospital were asked questions about their physical and mental work-related exhaustion, held in the variable `exhaust`, with higher scores reflecting greater feelings of exhaustion (X_1). Six months later, they were asked a series of questions to measure how frequently they use various methods of getting around safety protocols during their day-to-day work-related tasks with equipment and patients at the hospital (Y). This is held in a variable named `safety`. Also available in the data is the variable `tenure`, which quantifies how long the person has been working in the health care field (X_2). We refer to this in the discussion below simply as "job tenure" or "job experience."

For this example, we will test whether the relationship between exhaustion and use of safety protocol work-arounds interacts with job tenure (i.e., experience as a health care worker). To do so, we regress `safety` on `exhaust`, `tenure`, and their cross-product. To show that covariates are easily included, we will control for the sex of the employee (X_3), as well as age (X_4), by including them in the model as additional regressors. The data file does not include a variable that is the product of exhaust and tenure, so we construct it prior to running a regression command. For instance, in SPSS, the code would be

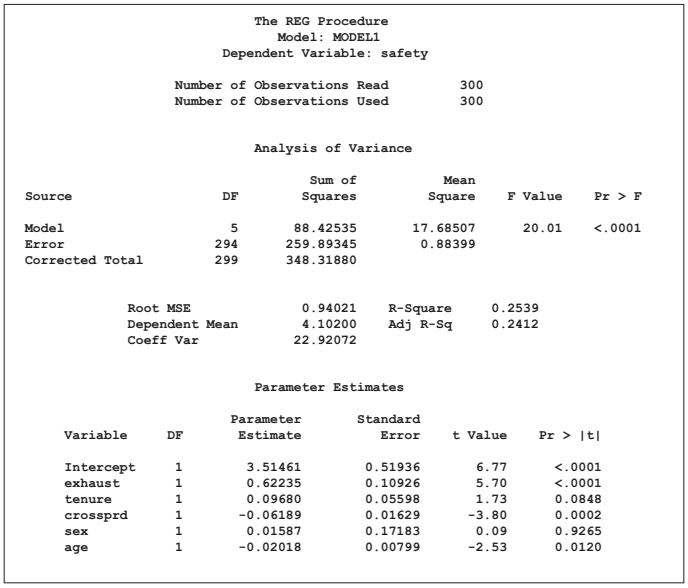
```
compute crossprd=exhaust*tenure.
regression/dep=safety/method=enter exhaust tenure crossprd sex age.
```

Comparable code in SAS is

```
data hospital;set hospital;crossprd=exhaust*tenure;run;
proc reg data=hospital;
model safety=exhaust tenure crossprd sex age;run;
```

and in STATA, try

```
gen crossprd=exhaust*tenure
regress safety exhaust tenure crossprd sex age
```



The REG Procedure
Model: MODEL1
Dependent Variable: safety

Number of Observations Read 300
Number of Observations Used 300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	88.42535	17.68507	20.01	<.0001
Error	294	259.89345	0.88399		
Corrected Total	299	348.31880			

Root MSE	0.94021	R-Square	0.2539
Dependent Mean	4.10200	Adj R-Sq	0.2412
Coeff Var	22.92072		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.51461	0.51936	6.77	<.0001
exhaust	1	0.62235	0.10926	5.70	<.0001
tenure	1	0.09680	0.05598	1.73	0.0848
crossprd	1	-0.06189	0.01629	-3.80	0.0002
sex	1	0.01587	0.17183	0.09	0.9265
age	1	-0.02018	0.00799	-2.53	0.0120

FIGURE 13.4. SAS output for a regression model containing a cross-product.

The output from SAS PROC REG can be found in Figure 13.4. The regression coefficient for the X_1X_2 cross-product is $-0.062, t(294) = -3.80, p = .0002$. Exhaustion and job tenure interact. Alternatively, we can say that job tenure moderates the relationship between exhaustion and use of safety protocol work-arounds.

As can be seen in the SAS output, the regression model is

$$\hat{Y} = 3.515 + 0.622X_1 + 0.097X_2 - 0.062X_1X_2 + 0.016X_3 - 0.020X_4 \quad (13.2)$$

Equation 13.2 can be written in equivalent form as

$$\hat{Y} = 3.515 + (0.622 - 0.062X_2)X_1 + 0.097X_2 + 0.016X_3 - 0.020X_4$$

which conveys the conditional effect of exhaustion (X_1) on the use of safety protocol work-arounds as a linear function of job tenure: $0.622 - 0.062X_2$.

Figure 13.5 visually represents the model. This model was generated by using equation 13.2 to produce $\hat{Y}$ for various values of X_1 and X_2, setting

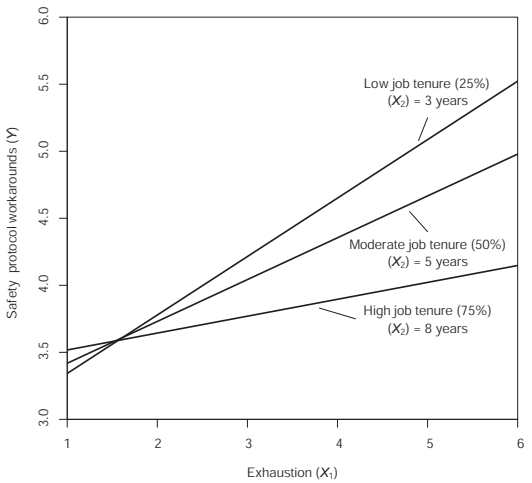


FIGURE 13.5. The interaction between exhaustion and job tenure.

the covariates X_3 and X_4 to their sample means.[1] Second, these values of $\hat{Y}$ as a function of X_1 and X_2 were plotted, connecting those representing the same value of X_2 with a line. The three values of X_2 used to produce this figure are the sample median job tenure or 50th percentile ($X_2 = 5$ years), the 25th percentile ($X_2 = 2$ years), and the 75th percentile ($X_2 = 8$ years). We might call these relatively moderate, relatively low, and relatively high in job tenure, respectively. Of course, these values are arbitrary and "relative" is sample specific, because they are defined in terms of the distribution in the sample. Other values of X_2 could be used, so long as they are within the range of the available data.

As is apparent in Figure 13.5, it seems that exhaustion has a bigger effect on the use of safety protocol work-arounds among those with less experience working on the job, as reflected in the steeper slope of the line when job tenure X_2 is larger. Among those with more experience, the slope while still positive is much less steep. So consider two health care workers who differ by 1 unit in their exhaustion. These findings say that these two people are estimated to differ more in their use of safety protocol work-arounds if they have less job experience relative to if they have more.

[1] Although it might seem strange or mathematically inappropriate to use the mean of a dichotomous predictor in these computations, doing so is mathematically legitimate, regardless of the two codes used to code the groups the dichotomous variable represents.

The slopes of the three lines in Figure 13.5 can be calculated from the linear function relating job tenure to the conditional effect of exhaustion on the use of safety protocol work-arounds. Recall that function is $0.622 - 0.062X_2$. Plugging values of 3, 5, and 8 into this function yield conditional effects of $0.622 - 0.062(3) = 0.436$, $0.622 - 0.062(5) = 0.312$, and $0.622 - 0.062(8) = 0.126$, which are the slopes of the lines in Figure 13.5.

In the earlier SPSS, SAS, and STATA codes, the product of the focal predictor and moderator was constructed prior to estimating the model. SAS has a feature in PROC GLM (not available in PROC REG) that eliminates the need to manually construct the product of the focal predictor and moderator in OLS regression. To include a cross-product involving exhaust and tenure, include **exhaust*tenure** in the PROC GLM model line, as in

```
proc glm data=hospital;
model safety=exhaust tenure exhaust*tenure sex age;run;
```

The RLM macro for SPSS and SAS described in Appendix A can also estimate a model with an interaction without requiring the analyst to first construct the product. For instance, in SPSS, the RLM command would be

```
rlm y=safety/x=sex age exhaust tenure/mod=1/ptiles=1/plot=1.
```

The **mod=1** option tells SPSS to estimate a model with a cross-product, with the cross-product being constructed from the last two variables in the **x=** list. The **ptiles=1** and **plot=1** components of the command are optional and produce the conditional effect of the focal predictor at the 25th, 50th, and 75th percentiles of the moderator, as well as a table values of $\hat{Y}$ to assist in the production of a plot such as Figure 13.5. See Appendix A for additional information about the use of the RLM macro.

13.2 Interaction Involving a Categorical Regressor

13.2.1 Interaction between a Dichotomous and a Numerical Regressor

If the relationship between X_1 and Y differs between two groups coded with X_2, such as if people assigned to a treatment group in an experiment are coded $X_2 = 1$ and those assigned to a control group are coded $X_2 = 0$, then one can think about the relationship between X_1 and Y as characterized with two regression lines, one for the treatment group, and another for the control group. In section 13.1.5 we saw how a whole series of nonparallel

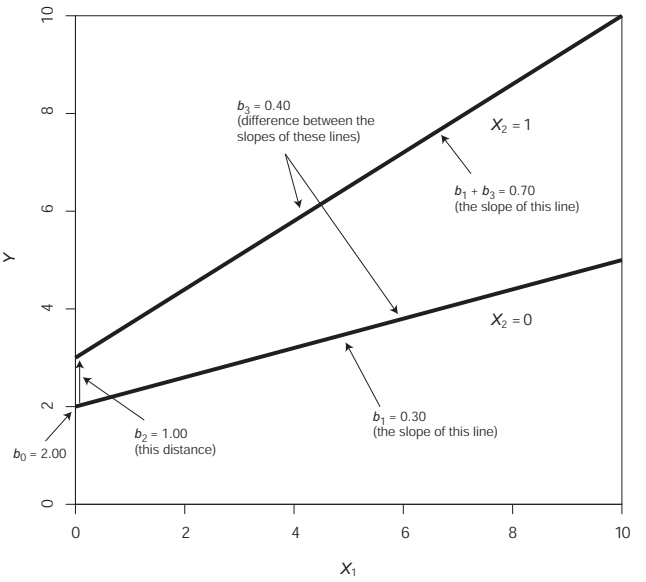


FIGURE 13.6. A plot depicting interaction between a dichotomous and a numerical variable.

regression lines can be represented by a single equation. Therefore, it should come as no surprise that we can use a single regression equation to represent two nonparallel regression lines. For instance, consider the two regression models below, one for people assigned to an experimental treatment group ($X_2 = 1$) and another for people assigned to a control group ($X_2 = 0$).

$$\hat{Y} = 3.0 + 0.7X_1 \text{ when } X_2 = 1 \text{ (treatment)}$$
$$\hat{Y} = 2.0 + 0.3X_1 \text{ when } X_2 = 0 \text{ (control)}$$

The regression lines for these equations appear in Figure 13.6. The difference between these two equations is

$$1.0 + 0.4X_1$$

Now consider the equation

$$\hat{Y} = 2.0 + 0.3X_1 + (1.0 + 0.4X_1)X_2 \tag{13.3}$$

where $1.0 + 0.4X_1$ is the difference between the two equations relating X_1 to Y when $X_2 = 0$ compared to when $X_2 = 1$. Observe that if we set X_2 to 0 in equation 13.3, which would be the case for those in the control condition, then

$$\hat{Y} = 2.0 + 0.3X_1 + (1.0 + 0.4X_1)(0)$$
$$\hat{Y} = 2.0 + 0.3X_1$$

which is the equation for the control group. But if we set X_2 to 1, then equation 13.3 becomes

$$\hat{Y} = 2.0 + 0.3X_1 + (1.0 + 0.4X_1)(1)$$
$$\hat{Y} = 3.0 + 0.7X_1$$

which is the equation for the treatment group. So equation 13.3 applies to both groups. By multiplying to remove parentheses, equation 13.3 becomes

$$\hat{Y} = 2.0 + 0.3X_1 + 1.0X_2 + 0.4X_1X_2$$

and we have expressed a single equation for the relationship between X_1 and Y that applies to both the treatment and the control groups.

So we can test for interaction between a numerical and a dichotomous regressor in exactly the same manner as when both regressors were numerical. We include the cross-product of the numerical and the dichotomous regressors in the model. A hypothesis test for the regression coefficient for the cross-product tests whether X_2 moderates X_1's effect on Y. The symmetry of interaction means that this also tests whether X_1 moderates the effect of X_2 on Y.

13.2.2 The Meaning of the Regression Coefficients

The interpretation of the regression coefficients presented in section 13.1.9 applies regardless of the scaling or coding of X_1 and X_2. In the equation for the previous example, $b_0 = 2.0$, $b_1 = 0.3$, $b_2 = 1.0$, and $b_3 = 0.4$. So $b_1 = 0.3$ is the conditional effect of X_1 when $X_2 = 0$. That is, two cases with $X_2 = 0$ but differ by 1 unit on X_1 are estimated to differ by 0.3 units on Y. As can be seen in Figure 13.6, this corresponds to the slope of the line relating X_1 to Y for those in the control group, because $X_2 = 0$ for those assigned to the control group.

The regression coefficient for X_2 is the conditional effect of X_2 when $X_1 = 0$. Two cases that differ by 1 unit on X_2 but that are 0 on X_1 are estimated to differ by $b_2 = 1.0$ units on Y. This is the estimated difference in Y between those in the treatment relative to the control group, conditioned on $X_1 = 0$. Thus, it is an estimated conditional mean difference, but under the condition that $X_2 = 0$. It is the distance between the points highlighted on Figure 13.6.

The regression coefficient for the X_1X_2 cross-product is $b_3 = 0.4$. This quantifies how the conditional effect of X_1 on Y changes as X_2 changes by 1 unit. A change of 1 unit on X_2 corresponds to the difference between the control and treatment conditions. In other words, $b_3 = 0.4$ is the difference between the slopes of regression lines linking X_1 to Y in the treatment group relative to the control group.

Knowing that in the control group the conditional effect of X_1 is $b_1 = 0.3$, and that as X_2 increases by 1 unit the conditional effect of X_1 changes by 0.4 units, we can calculate that the conditional effect of X_1 for those in the treatment group is $b_1 + b_3 = 0.3 + 0.4 = 0.7$. This is the slope of the line relating X_1 to Y in the treatment group (see Figure 13.6). In this example, b_3 represents the difference between these two slopes because the two groups are coded on X_2 such that they differ by 1 unit. If we used different codes, such as 0 and 2, then b_3 would be only one-half of the difference between the slopes. But the t- and p-values for b_3 would be the same.

The interpretation we gave to b_3 just now applies if we construe X_2 as the moderator and X_1 as the focal predictor. But remember that without changing the model whatsoever, we can flip the roles of focal predictor and moderator—the symmetry of interaction. In that case, b_3 quantifies how the conditional effect of X_2 on Y changes as X_1 changes by 1 unit. In this example, X_2 codes treatment or control and the two groups differ by 1 unit on X_2, so we can interpret b_3 as a measure of the amount the estimated difference in Y across the two groups changes as X_1 changes. So consider $X_1 = 3.0$. For the control group, the model generates

$$\hat{Y} = 2.0 + 0.3(3) + 1.0(0) + 0.4(3)(0) = 2.9$$

and for the treatment group the model generates

$$\hat{Y} = 2.0 + 0.3(3) + 1.0(1) + 0.4(3)(1) = 5.1$$

which is a difference of $5.1 - 2.9 = 2.2$ and the conditional effect of X_2 when $X_1 = 3.0$. This corresponds to the difference between points on the two

lines for the treatment and control groups when $X_1 = 3$. Now consider the same computations but when $X_1 = 4.0$, which produce

$$\hat{Y} = 2.0 + 0.3(4) + 1.0(0) + 0.4(4)(0) = 3.2$$

for the control group, and

$$\hat{Y} = 2.0 + 0.3(4) + 1.0(1) + 0.4(4)(1) = 5.8$$

for the treatment group, which is a difference of $5.7 - 3.2 = 2.6$ and also the conditional effect of X_2 when $X_1 = 4.0$. This corresponds to the difference between points on the two lines for the treatment and control groups when $X_1 = 4$.

Remember that interaction means that differences are different. When $X_1 = 4.0$, the difference in estimated Y between the groups is 2.6, and when $X_1 = 3$, the difference in estimated Y between the two groups is 2.2. The difference between these differences is $2.6 - 2.2 = 0.4 = b_3$. So b_3 quantifies a difference between differences. These computations can be repeated, changing the first value of X_1 but making sure the second value is 1 unit higher than the first value chosen. The difference between the differences in Y will always be $b_4 = 0.4$, regardless of the first value of X_1 chosen.

13.2.3 Interaction Involving a Multicategorical and a Numerical Regressor

We have modeled linear interaction between X_1 and X_2 by including the cross-product of X_1 and X_2 as a regressor in the model along with X_1 and X_2. This works when X_1 and X_2 are both numerical as well as when one is dichotomous. A similar procedure can be used when one of the variables is multicategorical. We saw in Chapters 9 and 10 that a multicategorical variable coding g groups can be represented with $g - 1$ regressors, such as a set of indicator codes, sequential codes, or some other system for coding the groups. Interaction between a multicategorical variable X_2 and numerical X_1 involves the construction of $g - 1$ cross-products involving X_1 and the $g - 1$ variables coding group, and including them in the regression model along with X_1 and the $g - 1$ group codes.

Consider a three-category variable X_2 represented with two indicator codes D_1 and D_2, such that group 1 is represented with $D_1 = 1$ and $D_2 = 0$, group 2 is coded with $D_1 = 0$ and $D_2 = 1$, and group 3 is the reference

group with $D_1 = D_2 = 0$. Interaction between X_1 and X_2 is estimated with the following model:

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2 + b_4X_1D_1 + b_5X_1D_2 \tag{13.4}$$

For instance, imagine an experiment with one control group (the reference group) and two treatment groups coded with two indicator codes D_1 and D_2. Suppose the resulting model that includes the interaction between X_1 and group is

$$\hat{Y} = 5.0 + 2.0X_1 - 1.0D_1 + 0.5D_2 + 1.0X_1D_1 - 1.5X_1D_2 \tag{13.5}$$

and thus $b_0 = 5.0, b_1 = 2.0, b_2 = -1.0, b_3 = 0.5, b_4 = 1.0, b_5 = -1.5$. This model is represented visually in Figure 13.7. As can be seen, it takes the form of three lines relating X_1 to Y, one for each group. This figure can be interpreted in two ways. It can be interpreted as three regression lines with different slopes (and intercepts), meaning that the relationship between X_1 and Y depends on group. It can also be interpreted as distances between points on the three lines conditioned on a specific value of X_1 that vary systematically across the the X_1 distribution. Either way, we can see that equation 13.5 represents three regression lines by plugging in values of D_1 and D_2 representing the groups and simplifying. For group 1

$$\begin{aligned}\hat{Y} &= 5.0 + 2.0X_1 - 1.0(1) + 0.5(0) + 1.0X_1(1) - 1.5X_1(0) \\ &= 4.0 + 3.0X_1\end{aligned}$$

Similarly, for group 2

$$\begin{aligned}\hat{Y} &= 5.0 + 2.0X_1 - 1.0(0) + 0.5(1) + 1.0X_1(0) - 1.5X_1(1) \\ &= 5.5 + 0.5X_1\end{aligned}$$

and for group 3

$$\begin{aligned}\hat{Y} &= 5.0 + 2.0X_1 - 1.0(0) + 0.5(0) + 1.0X_1(0) - 1.5X_1(0) \\ &= 5.0 + 2.0X_1\end{aligned}$$

These are the equations for the three lines in Figure 13.7. So equation 13.5 represents all three regression lines in one condensed form.

This model allows X_1's conditional effect on Y to differ across the three groups or the estimated difference on Y across the three groups to depend on X_1. In the former case, where X_2 is the moderator (represented with the

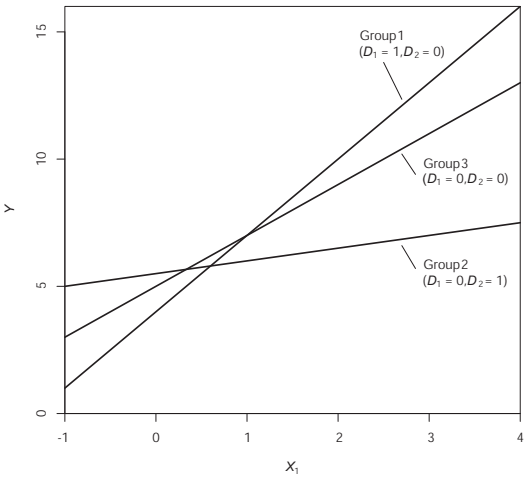


FIGURE 13.7. A plot depicting interaction between a multicategorical and a numerical variable.

set of regressors D_1 and D_2) and X_1 is the focal predictor, the conditional effect of X_1 on Y is most easily seen by reexpressing equation 13.4 as

$$\hat{Y} = b_0 + (b_1 + b_4D_1 + b_5D_2)X_1 + b_2D_1 + b_3D_2$$

which shows that X_1's effect on Y depends on the pattern of indicator codes representing group: $b_1 + b_4D_1 + b_5D_2$. Alternatively, equation 13.4 can be equivalently written as

$$\hat{Y} = b_0 + (b_2 + b_4X_1)D_1 + (b_3 + b_5X_1)D_2 + b_1X_1$$

which shows that the differences across the three groups depend on X_1. In this form, X_1 is the moderator and X_2 (represented with D_1 and D_2) is the focal predictor. Two functions characterize the difference across the groups on Y. The first relates the difference between group 1 and the reference group to X_1: $b_2 + b_4X_1$. The second relates the difference between group 2 and the reference group to X_1: $b_3 + b_5X_1$.

This system of modeling interaction between a multicategorical and a numerical regressor works for any number of groups. Simply include $g - 1$ variables required to code groups in the model along with X_1 and all $g - 1$ cross-products involving group codes and X_1. It can also be used for any

system of coding groups. We used indicator coding in the example above, but sequential, Helmert, or effect codes could be used instead.

13.2.4 Inference When Interaction Requires More Than One Regression Coefficient

In the first two examples in this chapter, interaction between X_1 and X_2 was carried in a single regression coefficient for the X_1X_2 cross-product and a hypothesis test undertaken using the t-statistic and p-value for that regression coefficient. But in the example with a multicategorical variable coding g groups, $g - 1$ cross-product terms were required to estimate interaction. A different approach to inference is required.

When the focal predictor or moderator is multicategorical, an inferential test of interaction relies on the strategy described in section 5.3.3 when we discussed inference about sets of variables. This strategy involves comparing the fit of two models, one that allows the focal predictor's effect to vary as a function of the moderator and another that fixes the focal predictor's effect to be independent of the moderator. If the first model fits better than the second, this implies interaction. But if there is no statistically significant difference in fit, parsimony tells us to prefer the simpler model that doesn't allow the focal predictor's effect to depend on the moderator.

To see how this works, consider the model we just introduced:

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2 + b_4X_1D_1 + b_5X_1D_2 \tag{13.6}$$

This is a flexible model in that it allows the effect of X_1 on Y to differ between the three groups coded with D_1 and D_2 (or the effect of group membership to vary as a function of X_1). It is the regression coefficients b_4 and b_5 that provide this flexibility, because they determine how much the focal predictor's effect varies with the moderator. We could remove that flexibility by forcing b_4 and b_5 to be zero, which would be equivalent to estimating

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2 \tag{13.7}$$

Descriptively speaking, these two models will fit differently, except in the rare case where b_4 and b_5 are both exactly zero. We know that R^2 for the model expressed by equation 13.6 will be larger than R^2 for the model expressed in equation 13.7, because equation 13.6 contains the same regressors as equation 13.7 plus a few more, and adding variables to a model almost always increases R^2 somewhat. Recognizing that the difference in R^2 between these two models is a squared semipartial multiple correlation,

we can test whether ${}_TSR^2$ is equal to zero using the F-test introduced in section 5.3.3. Under the null hypothesis of no interaction between X_1 and the multicategorical variable, this F-ratio is $F(g-1, df_{residual})$ distributed, where $df_{residual}$ is the residual degrees of freedom for the model that includes the $g-1$ cross-products (in this example, $g=3$). This is equivalent to testing the null hypothesis that all $g-1$ of the regression coefficients for the cross-products are equal to zero.

In fact, using the t-statistic and p-value for inference when only a single cross-product term is needed is a special case of this more general F-test. We discuss this point in section 14.4.5.

13.2.5 A Substantive Example

The POLITICS data file discussed earlier in the book includes a variable named `demoneg` that contains the survey respondents' evaluations of the Democrat running for President of the United States in that year's election. This evaluation is an aggregate of responses to questions asking how much certain traits such as "moral," "dishonest," and "a good leader" apply to the person. Higher scores on this variable reflect a more negative evaluation of the candidate. This will be the dependent variable in a model that includes the number of days per week the respondent reports engaging in political discussion with others (X_1:`pdiscuss`), the respondent's political party self-identification (X_2), and the interaction between party identification and political discussion frequency. Party identification is held in the variable named `party` and is coded 1 = Democrat, 2 = Republican, 3 = Independent. Because party identification is multicategorical, we represent it in the model with two regressors coding group. In this example, we use indicator coding, with Democrats coded $D_1 = 1$ and $D_2 = 0$ and Republicans coded $D_1 = 0$ and $D_2 = 1$. The Independents are set to 0 on both D_1 and D_2 and so function as the reference group. Thus, we estimate

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2 + b_4X_1D_1 + b_5X_1D_2$$

Because the interaction requires two regression coefficients to estimate it, one for each of the cross-products, we test the null hypothesis of no interaction by testing whether both ${}_Tb_4$ and ${}_Tb_5 = 0$. This is accomplished by examining whether the model above fits better than

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2$$

by examining the difference in R^2 between the two models, as discussed in section 13.2.4.

The SPSS code below constructs the indicator codes and estimates both models while also producing a test of the difference in fit of the two models:

```
compute d1=(party=1).
compute d2=(party=2).
compute crossp1=pdiscuss*d1.
compute crossp2=pdiscuss*d2.
regression/statistics defaults change/dep=demoneg/method=enter pdiscuss
    d1 d2/method=enter crossp1 crossp2.
```

The output generated by this code can be found in Figure 13.8. The best fitting regression model including the two cross-products is

$$\hat{Y} = 2.078 + 0.038X_1 - 0.049D_1 + 0.179D_2 - 0.071X_1D_1 + 0.027X_1D_2 \quad (13.8)$$

As can be seen in the output, adding the two cross-products significantly improves the fit of the model. The change in R^2 (which we also refer to as a squared semipartial multiple correlation) is 0.032, $F(2, 334) = 7.834, p < .001$. It seems that the relationship between political discussion frequency and evaluation of the candidate differs between the three groups. Alternatively, we can say that the difference between Democrats, Republicans, and Independents in their evaluation of the Democrat depends on frequency of political discussion.

Comparable code in SAS that conducts this analysis is

```
data politics;set politics;
d1=(party=1);d2=(party=2);crossp1=pdiscuss*d1;crossp2=pdiscuss*d2;run;
proc reg data=politics;
    model demoneg=d1 d2 pdiscuss crossp1 crossp2;
    test crossp1=0,crossp2=0;run;
```

and in STATA, try

```
gen d1=(party==1)
gen d2=(party==2)
gen crossp1=pdiscuss*d1
gen crossp2=pdiscuss*d2
regress demoneg d1 d2 pdiscuss crossp1 crossp2
test crossp1 crossp2
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.532[a]	.283	.276	.53659	.283	44.151	3	336	.000
2	.561[b]	.315	.305	.52600	.032	7.834	2	334	.000

a. Predictors: (Constant), d2, pdiscuss, d1

b. Predictors: (Constant), d2, pdiscuss, d1, crossp1, crossp2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	38.136	3	12.712	44.151	.000[b]
	Residual	96.743	336	.288		
	Total	134.880	339			
2	Regression	42.471	5	8.494	30.701	.000[c]
	Residual	92.409	334	.277		
	Total	134.880	339			

a. Dependent Variable: Democrat: Negative evaluation

b. Predictors: (Constant), d2, pdiscuss, d1

c. Predictors: (Constant), d2, pdiscuss, d1, crossp1, crossp2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.163	.089		24.274	.000
	pdiscuss	.018	.012	.072	1.552	.122
	d1	-.372	.087	-.291	-4.297	.000
	d2	.340	.087	.267	3.920	.000
2	(Constant)	2.078	.136		15.244	.000
	pdiscuss	.038	.027	.151	1.403	.162
	d1	-.049	.165	-.038	-.298	.766
	d2	.179	.172	.141	1.042	.298
	crossp1	-.071	.032	-.317	-2.206	.028
	crossp2	.027	.032	.130	.833	.405

a. Dependent Variable: Democrat: Negative evaluation

FIGURE 13.8. SPSS output from a model estimating an interaction between a multicategorical and a numerical regressor.

The RLM macro documented in Appendix A has an option for specifying an interaction between a multicategorical and a numerical or dichotomous regressor. It does all the computations itself, including the construction of the necessary cross-products and the test of interaction. For this analysis, the SPSS version of the RLM command is

```
rlm y=demoneg/x=pdiscuss party/catx=1/mod=1.
```

A visual depiction of the interaction represented by equation 13.8 can be found in Figure 13.9, generated by plugging in various values of political discussion frequency and indicator codes into the equation and then plotting the resulting $\hat{Y}$ values. In this example we had no covariates, but if we had, we could have set them to the sample mean to produce the $\hat{Y}$ values that go into the plot, as in the example in section 13.1.10.

Interpretation of the results using Figure 13.9 would depend on whether political discussion frequency is construed as the focal predictor or moderator. As focal predictor, attention would be directed toward the different slopes relating political discussion frequency to negativity of the evaluation of the Democrat. As can be seen, it appears that among Republicans, this relationship is positive with more discussion linked to a more negative evaluation, whereas with Democrats, the opposite appears to be the case. Among Independents, the relationship is in between, with a slight positive slope meaning more discussion relates to a more negative evaluation.

If political discussion frequency is construed as the moderator, then interpretation focuses on how the differences in estimated evaluation of the Democrat among people who differ in political party identification vary depending on political discussion frequency. In Figure 13.9, this leads to an examination of the distances between points on the lines between the three groups when political discussion frequency is held fixed and how these distances vary with discussion frequency.

Interpreted in this way, with political discussion frequency as the moderator, the most apparent result is that the differences in how the Democrat is perceived between people of different party identifications seem more pronounced (i.e., larger) among those who discuss politics more frequently, compared to those who discuss politics less frequently. If one assumes that people primarily talk about politics to like-minded others, we might interpret these findings to mean that political discussion frequency *polarizes* opinions. Alternatively, it could be that people who hold more extreme positions are more inclined to talk about politics with other people (like-minded or not).

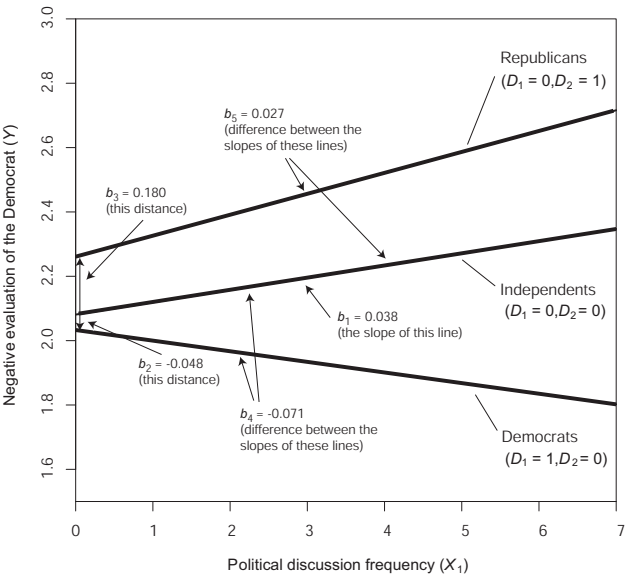


FIGURE 13.9. A plot depicting interaction between political discussion frequency and party identification.

13.2.6 Interpretation of the Regression Coefficients

Information from the regression model allows us to quantify the three slopes in Figure 13.9 and the difference between pairs of slopes. Recall from section 13.2.3 that we can express the model in equation 13.8 in the form

$$\hat{Y} = 2.078 + (0.038 - 0.071D_1 + 0.027D_2)X_1 - 0.049D_1 + 0.179D_2$$

which shows that the relationship between X_1 and Y varies with D_1 and D_2 (i.e., $b_1 + b_4D_1 + b_5D_2 = 0.038 - 0.071D_1 + 0.027D_2$), or, in other words, as a function of party identification, because D_1 and D_2 code a person's party self-identification. For the three combinations of D_1 and D_2 coding groups, we get

$$\begin{aligned}
\text{Democrats:} \quad & b_1 + b_4(1) + b_5(0) = 0.038 - 0.071(1) + 0.027(0) = -0.033 \\
\text{Republicans:} \quad & b_1 + b_4(0) + b_5(1) = 0.038 - 0.071(0) + 0.027(1) = 0.065 \\
\text{Independents:} \quad & b_1 + b_4(0) + b_5(0) = 0.038 - 0.071(0) + 0.027(0) = 0.038
\end{aligned}$$

for the three conditional effects of political discussion frequency X_1. Thus, the regression coefficient for political discussion frequency, $b_1 = 0.038$, is the conditional effect of X_1 on Y for Independents. That is the slope of the line relating X_1 to Y among Independents in Figure 13.9. But from the regression output in Figure 13.8, this is not statistically different from zero, $t(334) = 1.403, p = .162$. The slope of the line for Democrats is $b_1 + b_4 = 0.038 - 0.071 = -0.033$ and so $b_4 = -0.071$ is the difference between the slopes for Democrats relative to Independents. As can be seen in the output in Figure 13.8, this difference between these two slopes is statistically significant, $t(334) = -2.206, p = .028$. Similarly, the slope of the line for Republicans is $b_1 + b_5 = 0.038 + 0.027 = 0.065$ and so $b_5 = 0.027$ is the difference between the slopes for Republicans relative to Independents. But this difference is not statistically significant, $t(334) = 0.833, p = .405$.

The symmetry property of interactions tells us that we can interpret b_4 and b_5 in two ways. We just interpreted b_4 and b_5 as the difference between the slopes of the lines relating X_1 to Y. But we can also interpret b_4 and b_5 as how much differences between groups in how they perceive the Democrat change as X_1 changes by 1 unit. Recall from section 13.2.3 that we can express equation 13.8 in the form

$$\hat{Y} = 2.078 + 0.038X_1 + (-0.049 - 0.071X_1)D_1 + (0.179 + 0.027X_1)D_2$$

This means that the difference in how the Democrat is perceived between Independents and Democrats is $b_2 + b_4X_1 = -0.049 + -0.071X_1$, which is a function of political discussion frequency. For instance, when $X_1 = 1$, the difference is $-0.049 - 0.071(1) = -0.120$, meaning that Democrats who talk 1 day per week about politics perceive the Democrat 0.120 units less negatively than Independents who talk about politics 1 day a week. But when $X_1 = 2$ days per week, the difference is $-0.049 - 0.071(2) = -0.191$, meaning that such Democrats perceive the Democrat 0.190 units less negatively than such Independents. This change in the difference with a 1 unit change in X_1 is $b_4 = -0.071$, and it is invariant to where you start

on X_2. It is the amount the distance between the points on the Democrat and Independent lines in Figure 13.9 changes as X_1 increases by 1.

Using the same reasoning, the difference in how the Democrat is perceived between Independents and Republicans is $b_3 + b_5X_1 = 0.179 + 0.027X_1$. Regardless of the value of X_1 at which you start, as you increase X_1 by 1 unit (i.e., 1 day of political discussion), the difference in how Republicans and Independents perceive the Democrat changes by $b_5 = 0.027$ units. As can be seen in Figure 13.9, the gap between the Independent and Republican line is growing as X_1 increases by 1 unit. The gap is growing by 0.027 units per 1 unit change in X_1.

We have thus far neglected b_2 and b_3. The prior two paragraphs lead to the interpretation that these are the estimated difference in how negatively the Democrat is perceived between Independents and Democrats (b_2) and between Independents and Republicans (b_3) among people who report not discussing politics at all. That is, these conditional effects are conditioned on $X_1 = 0$. So among those who don't discuss politics at all, Democrats perceive the Democrat $b_2 = -0.049$ units differently than Independents. The negative sign means that Democrats perceive the Democrat *less* negatively than do Independents, as can be seen in Figure 13.9. Similarly, among those who don't discuss politics, Republicans perceive the Democrat $b_3 = 0.179$ units differently than Independents. The positive sign means that Republicans perceive the Democrat more negatively than do Independents (see Figure 13.9). But neither of these conditional effects is statistically significant.

13.3 Interaction between Two Categorical Regressors

Every example of interaction we have presented has included at least one numerical regressor serving the role of focal predictor or moderator. When both focal predictor and moderator are categorical, most researchers would conduct an ANOVA or ANCOVA. But AN(C)OVA is just a special case of the general linear model and can be conducted using a linear regression analysis program.

13.3.1 The 2 × 2 Design

Consider the 2 × 2 table of cell means in Table 13.3, which shows the mean of Y for males and females (X_2) randomly assigned to either a treatment or

a control group (X_1). The pattern of means in this table is consistent with interaction between X_1 and X_2, because the simple effect of experimental condition differs between men and women. Notice that for men, the mean difference between the treatment and control groups is $9 - 8 = 1$, but for women, the mean difference between the two groups is $5 - 7 = -2$. The difference between these differences is $1 - (-2) = 3$. Notice that this difference between differences is the same if we look at the simple effects of sex. For those assigned to the treatment group, the mean difference between men and women is $9 - 5 = 4$, but for those assigned to the control group, the mean difference between men and women is $8 - 7 = 1$. The difference between these differences is $4 - 1 = 3$.

When the difference between simple effects in a 2×2 design is not zero, this implies interaction. When this difference is zero, this means no interaction. Of course, we have not acknowledged sampling error in this discussion. A hypothesis test for this difference between differences takes the form of an F-test, as users of ANOVA are aware.

Suppose the treatment and control groups were coded $X_1 = 1$ and $X_2 = 0$, respectively. And suppose men were coded $X_2 = 1$ and females were coded $X_2 = 0$. In that case, the pattern of means in this table can be expressed with the linear model

$$\hat{Y} = 7 - 2X_1 + 1X_2 + 3X_1X_2 \tag{13.9}$$

That is, $b_0 = 7$, $b_1 = -2$, $b_2 = 1$, and $b_3 = 3$. You can plug values of 0 and 1 for X_1 and X_2 corresponding to the groups into equation 13.9 to verify that it generates the cell means. For instance, for males assigned to the control condition, $\hat{Y} = 7 - 2(0) + 1(1) + 3(0)(1) = 8$.

When the groups are coded this way, b_1 is the simple effect of experimental condition among women ($5 - 7 = -2$), b_2 is the simple effect of sex for those assigned to the control condition ($8 - 7 = 1$), b_3 is the difference between the simple effects (i.e., interaction; see the calculations above), and b_0 is the mean of Y for the females assigned to the control condition (see Figure 13.10, panel A). This is a *simple effects* parameterization of the 2×2 design, because b_1 and b_2 represent simple effects of X_1 and X_2, respectively. We have been calling these conditional effects, which is another name for a simple effect. The t- and p-values for these regression coefficients provide a test of the null hypothesis that the corresponding simple effect is zero. For the interaction, the square of its t value for b_3 will be equal to the F-ratio for the interaction from a 2×2 ANOVA, and its p-value would be the same.

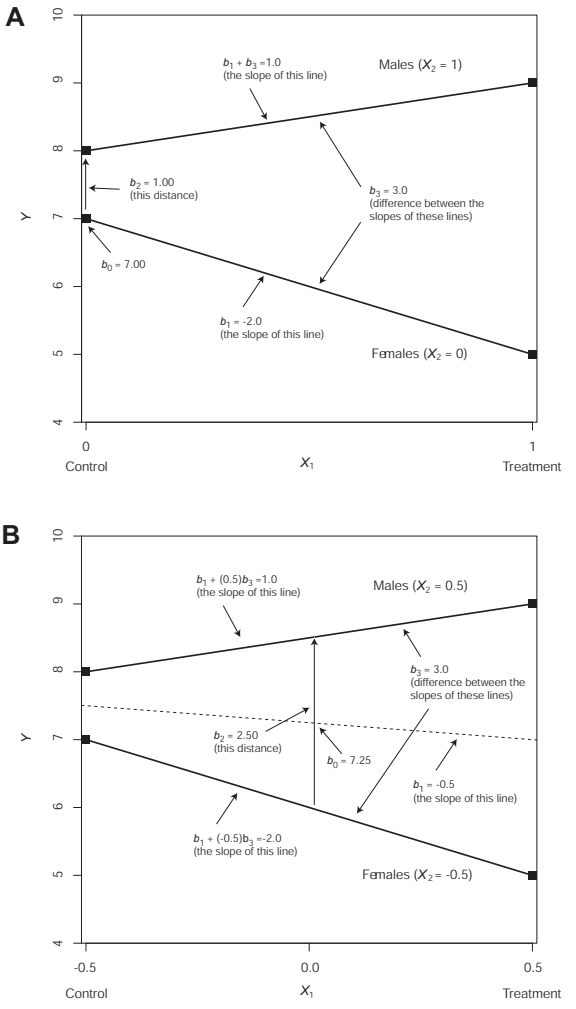


FIGURE 13.10. A visual representation of cell means and regression coefficients from a 2×2 design using simple effects (A) and main effects (B) parameterizations.

TABLE 13.3. A Table of Means from a 2 × 2 Design

	Sex (X_2)	
Condition (X_1)	Male	Female
Treatment	$\overline{Y} = 9.00$	$\overline{Y} = 5.00$
Control	$\overline{Y} = 8.00$	$\overline{Y} = 7.00$

An alternative parameterization is the *main effects* parameterization. In the lingo of ANOVA, a main effect is an *unweighted average simple effect*. In Table 13.3, the main effect of experimental condition is −0.5, calculated as the mean of the two simple effects of experimental condition: $[(9 - 8) + (5 - 7)]/2 = -0.5$. The main effect of sex is 2.5, calculated as the mean of the two simple effects of sex: $[(9 - 5) + (8 - 7)]/2 = 2.5$. To implement a main effects parameterization, code the treatment group $X_1 = 0.5$ and the control group $X_1 = -0.5$. Similarly, code males and females $X_2 = 0.5$ and $X_2 = -0.5$, respectively. In that case, the model that generates the group means is

$$\hat{Y} = 7.25 - 0.5X_1 + 2.5X_2 + 3X_1X_2 \tag{13.10}$$

and so $b_0 = 7.25$, $b_1 = -0.5$, $b_2 = 2.5$, and $b_3 = 3$. Plugging in values of −0.5 and 0.5 corresponding to the groups into equation 13.10 will reproduce the group means. For example, for males assigned to the control condition, $\hat{Y} = 7.25 - 0.5(-0.5) + 2.5(0.5) + 3(-0.5)(0.5) = 8$.

Observe that b_3 is the same compared to when the simple effects parameterization was used. Its *t*- and *p*-values would also be the same. But the other terms have changed. Now b_1 is the main effect of experimental condition and b_2 is the main effect of sex, as calculated earlier. Now b_0 is the unweighted mean of the four cell means: $(9 + 5 + 8 + 7)/4 = 7.25$ (see Figure 13.10, panel B). The squares of the *t*-values for these effects will correspond to the *F*-ratios for the corresponding effects from an ANOVA, and the *p*-values will be the same.

13.3.2 Interaction between a Dichotomous and a Multicategorical Regressor

So regression analysis can generate the results from a 2 × 2 ANOVA when the main effects parameterization is used, and it will generate the same test

of interaction and some of the tests of simple effects when the simple effects parameterization is used. But what about a design with more than two levels of a categorical variable?

If one of the variables is dichotomous but the other is multicategorical with g categories, the approach discussed in section 13.2.3 applies. If X_1 were a dichotomous variable in the example in that section, in some ways but not others, the approach described would would be equivalent to a 3×2 factorial ANOVA, with the similarity depending on how the groups are coded. The test for interaction between the dichotomous and the multicategorical variable would be the same as the F-test from the ANOVA. But the tests for the multicategorical and dichotomous regressors would not correspond to the "main effects" from the ANOVA unless the dichotomous variable was coded with two values equal but opposite in sign (e.g., -1 and 1 or -0.5 and 0.5) and the multicategorical variable was coded using the effect coding system described in section 10.1.3. The coding system for the categorical variables would not change the test of significance for the interaction. Its F-ratio and p-value would be the same regardless, and the increment in R due to the cross-product terms would be unaffected.

13.3.3 Interaction between Two Multicategorical Regressors

When both focal predictor and moderator are multicategorical with three or more categories, regression analysis can still duplicate ANOVA results, but most researchers will find this more tedious than just doing an ANOVA in the usual way. If the focal predictor has g_1 categories and moderator g_2 categories, then $(g_1 - 1)(g_2 - 1)$ cross-products are required to represent the $g_1 \times g_2$ interaction. This number can be large, and managing and interpreting all these regression coefficients, not to mention those quantifying simple or main effects, can be cumbersome. Yet if effect coding is used for both variables, ANOVA F-ratios for main and interaction effects can be generated by testing the contribution of sets of regression coefficients to explaining variance in Y using the hierarchical entry strategy described in section 5.3.3.

13.4 Chapter Summary

Linear interaction between a focal predictor X_1 and a moderator X_2 can be estimated and tested by including the cross-product X_1X_2 as a regressor along with X_1 and X_2. This works so long as neither X_1 nor X_2 is a multicategorical variable. In such a model, the regression coefficient for X_1

estimates the conditional effect of X_1 on Y when $X_2 = 0$, and the regression coefficient for X_2 quantifies the conditional effect of X_2 on Y when $X_1 = 0$. The regression coefficient for X_1X_2 quantifies the amount the conditional effect of X_1 on Y changes as X_2 changes by 1 unit. If this regression coefficient is statistically different from zero, we can say that X_1 and X_2 interact, or that X_2 moderates the effect of X_1 on Y.

The symmetry property of interaction tells us that we can flip the roles of focal predictor and moderator without changing the essence of the interaction or the regression coefficient for the X_1X_2 cross-product. If we think of X_2 as the focal predictor and X_1 as the moderator, then the regression coefficient for the product can be interpreted as how the conditional effect of X_2 changes as X_1 changes by 1 unit. This will be the same as the amount the conditional effect of X_1 changes as X_2 changes by 1 unit.

When either the focal predictor or moderator is multicategorical, interaction requires multiple cross-products in the model, and a test of interaction requires comparing the fit of two models, one with the cross-products and one without. Although estimation of the model is straightforward, interpretation is a bit more complex. But the same rules of interpretation apply, where the regression coefficients for the variables that define the interaction are conditional effects, and regression coefficients for cross-products quantify how the conditional effect of one variable changes as another variable changes by 1 unit. When the moderator is multicategorical with g groups, such a model will represent g regression lines relating the focal predictor to Y. When the focal predictor is multicategorical, the model can be used to quantify how the groups differ from each other on average at given values of the moderator, and how those differences change as the moderator changes.

When both X_1 and X_2 are categorical, regression analysis can be used to mimic factorial ANOVA. Interaction in a factorial ANOVA can be represented with a set of cross-product regressors and estimated using any regression analysis program. In simple designs involving only two variables with a few levels each, interpretation of the regression coefficients is fairly straightforward. But as the number of categories increases, so does the number of cross-products required, and interpretation of the effects can become quite complex.

14
Probing Interactions and Various Complexities

When two variables interact, one variable's effect on the dependent variable can be expressed as a function of the other variable involved in the interaction. We start this chapter by formalizing the functions linking one variable's effect to another in a linear interaction model. These functions can be used to generate estimates of the conditional effect of one variable at values of another, and inference can proceed with an estimate of the standard error of the conditional effect. We also discuss techniques for probing interactions based on estimates of conditional effects, including the Johnson–Neyman technique, which eliminates the need to choose a value of the moderator when probing an interaction. We end with a discussion of various complexities and controversies, including the difficulty of detecting interactions, distinguishing between interaction and curvilinearity, as well as models with multiple or higher-order interactions.

14.1 Conditional Effects as Functions

We have seen in Chapters 3, 13, and elsewhere in the book that in a model of the form

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

the effects of X_1 and X_2 are constant across values of the other variable. That is, the effect of a 1 unit change in X_1 on Y does not depend on X_2, just like the effect of a 1 unit change in X_2 on Y does not depend on X_1. When these effects are independent of the other variable, it is sensible to talk about *the* effect of X_1 or X_2 and draw inferences about its size.

But when a cross-product between X_1 and X_2 is included as a regressor in the model, as in

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 \tag{14.1}$$

then the effects of X_1 and X_2 can no longer be expressed by a single number. Each of their effects becomes a function of the other variable, and so it is not sensible to talk about the effect of X_1 on Y without conditioning that discussion on a value of X_2. Likewise, we can't talk about the effect of X_2 on Y without conditioning that discussion on X_1.

In this section, we formalize those functions in models that include a cross-product, depending on the combination of the focal predictor and moderator as continuous or categorical. These functions take the form of equations that can be used to generate an estimate of the conditional effect of X_1 on Y for a given value of X_2 and the conditional effect of X_2 on Y for a given value of X_1. Our discussion generalizes to models with covariates so long as those covariates are not allowed to interact with X_1 or X_2. In Chapter 13 we applied these functions, but in the context of specific examples. Our discussion here is more general and organizes all the functions into one place rather than spreading them throughout the entire chapter.

14.1.1 When the Interaction Involves Dichotomous or Numerical Variables

We can use the model in equation 14.1 when the focal predictor and moderator are dichotomous or numerical in any combination. That is, the mathematics to generate the function are the same for any of the four combinations of the dichotomous or numerical status of the focal predictor and moderator. A different function is required when either the focal predictor or moderator is multicategorical. We address this in section 14.1.2.

Notice in equation 14.1 that X_1 and X_2 both appear twice. We can rewrite it in a form where X_1 appears only once by grouping terms involving X_1 and then factoring out X_1 from the group that includes it. That is,

$$\begin{aligned}\hat{Y} &= b_0 + (b_1X_1 + b_3X_1X_2) + b_2X_2 \\ &= b_0 + (b_1 + b_3X_2)X_1 + b_2X_2\end{aligned}$$

In this form, we can see that X_1's effect on Y is a function of X_2. That function is $b_1 + b_3X_2$. So in a model of the form in equation 14.1,

$$\theta_{X_1} = b_1 + b_3X_2 \tag{14.2}$$

where θ_{X_1} is the *conditional effect* of X_1 on Y. So if we want an estimate of X_1's effect on Y when X_2 equals some value, we can generate it with an estimate of b_1 and b_3 and that value of X_2.

For instance, revisiting the example presented in section 13.1.10, we had $\hat{Y} = 3.515+0.622X_1+0.097X_2-0.062X_1X_2+0.016X_3-0.020X_4$ for the model of the use of safety protocol work-arounds with exhaustion as focal predictor X_1 and job tenure as moderator X_2, with sex and age as covariates (equation 13.2. Figure 13.5 depicts the model). Using equation 14.2 we estimate the conditional effect of exhaustion on use of safety protocol work-arounds for hospital workers with 5 years of job experience ($X_2 = 5$) as

$$\theta_{X_1} = b_1 + b_3X_2 = 0.622 - 0.062(5) = 0.312 \tag{14.3}$$

So when $X_2 = 5$, changing X_1 by 1 unit results in an increase in the estimate of Y by 0.312 units. This is the slope of the $X_2 = 5$ line in Figure 13.5.

If X_2, which is functioning as the moderator in the discussion above, is dichotomous, then equation 14.2 produces only two conditional effects of X_1, one for each of the groups defined by X_2. It makes no difference how X_2 is coded, for the estimation of the regression coefficients adjusts for the scaling of X_2 such that equation 14.2 generates the same values for θ_{X_1} regardless of how the two groups are coded. When X_2 is numerical, there may be many values of X_2, even an infinite number, depending on how finely measured X_2 is.

This same logic results in a function defining the conditional effect of X_2. Equation 14.1 can be rewritten by factoring X_2 out of terms involving X_2:

$$\begin{aligned}\hat{Y} &= b_0 + (b_2X_2 + b_3X_2X_1) + b_1X_1 \\ &= b_0 + (b_2 + b_3X_1)X_2 + b_1X_1\end{aligned}$$

which shows that X_2's effect on Y is a function X_1 of the form $b_2 + b_3X_1$. That is, the conditional effect of X_2 is

$$\theta_{X_2} = b_2 + b_3X_1$$

With estimates of b_2 and b_3, θ_{X_2} can be calculated for any value of X_1.

14.1.2 When the Interaction Involves a Multicategorical Variable

When the focal predictor or moderator is multicategorical with g categories, the regression model should include $g - 1$ variables coding group, as well as $g - 1$ cross-products involving those codes and the focal predictor (see section 13.2.3). In the discussion below, we assume that the moderator X_2 is multicategorical with three groups, but the procedure generalizes to $g > 3$ by including additional group codes and cross-products.

For $g = 3$ groups, the model to estimate interaction involving the multicategorical variable is

$$\hat{Y} = b_0 + b_1X_1 + b_2D_1 + b_3D_2 + b_4X_1D_1 + b_5X_1D_2 \tag{14.4}$$

where D_1 and D_2 are indicator codes, Helmert codes, sequential codes, effect codes, or codes derived from some other group coding system. The function defining the conditional effect of focal predictor X_1, which is dichotomous or numerical, is found by grouping terms involving X_1 and then factoring it out, as in

$$\begin{aligned}\hat{Y} &= b_0 + (b_1X_1 + b_4X_1D_1 + b_5X_1D_2) + b_2D_1 + b_3D_2 \\ &= b_0 + (b_1 + b_4D_1 + b_5D_2)X_1 + b_2D_1 + b_3D_2\end{aligned}$$

which shows that X_1's conditional effect on Y is a function of group, meaning a function of combinations of D_1 and D_2. That is,

$$\theta_{X_1} = b_1 + b_4D_1 + b_5D_2 \tag{14.5}$$

With estimates of b_1, b_4, and b_5, the conditional effect of X_1 can be derived for each of the three groups by plugging the values of D_1 and D_2 corresponding to the group in the coding system used. For example, revisiting the political discussion example from section 13.2.5, the model was $\hat{Y} = 2.078 + 0.038X_1 - 0.049D_1 + 0.179D_2 - 0.071X_1D_1 + 0.027X_1D_2$. In that analysis, indicator coding was used to code groups. For the group coded $D_1 = 0$ and $D_2 = 1$ (the Republicans), application of equation 14.5 yields $\theta_{X_1} = 0.038 - 0.071(0) + 0.027(1) = 0.065$ for the conditional effect of political discussion (X_1) for Republicans.

When the focal predictor is the multicategorical variable and X_1 is the moderator, then there are $g - 1$ conditional effects for a given value of X_1. Each of these conditional effects compare estimates of Y, conditioned on X_1, that reflect the group comparisons built into the coding system. For instance, if indicator coding is used, then the two conditional effects gauge

the difference between the group coded with D_i and the reference group conditioned on X_1 being some value.

By grouping terms into sets that share D_i in equation 14.4 and then factoring out D_i we get

$$\begin{aligned}\hat{Y} &= b_0 + b_1X_1 + (b_2D_1 + b_4X_1D_1) + (b_3D_2 + b_5X_1D_2) \\ &= b_0 + b_1X_1 + (b_2 + b_4X_1)D_1 + (b_3 + b_5X_1)D_2\end{aligned}$$

Recognizing that the categorical variable X_2 is represented with two variables, D_1 and D_2, we get two conditional effects of X_2, one for D_1 and one for D_2. Those functions are

$$\theta_{D_1} = b_2 + b_4X_1 \tag{14.6}$$

and

$$\theta_{D_2} = b_3 + b_5X_1 \tag{14.7}$$

both of which are functions of X_1. So using the model coefficients from the example above, when $X_1 = 2$, for instance, equations 14.6 and 14.7 yield $\theta_{D_1} = -0.049 - 0.071(2) = -0.191$ and $\theta_{D_2} = 0.179 + 0.027(2) = 0.233$. Thus, when $X_1 = 2$, the group coded with $D_1 = 1$ (Democrats) is estimated to differ by -0.191 units on Y compared to the reference group (Independents), and the group coded with $D_2 = 1$ (Republicans) is estimated to differ by 0.233 units on Y from the reference group of Independents.

14.2 Inference about a Conditional Effect

In section 14.1 we offered functions that relate the effect of X_1 on Y to X_2 and the effect of X_2 on Y to X_1 in a model that allows linear interaction between X_1 and X_2. Using these functions, one can generate a point estimate of the conditional effect of one variable conditioned on the other variable involved in the interaction. We next turn to inference about conditional effects.

14.2.1 When the Focal Predictor and Moderator Are Numerical or Dichotomous

Standard errors can be estimated for a conditional effect in order to generate a confidence interval for ${}_T\theta_{X_1}$ or ${}_T\theta_{X_2}$ or test a null hypothesis. Unless X_1 or

X_2 is multicategorical, in a linear interaction model of the form in equation 14.1, the standard errors for θ_{X_1} and θ_{X_2} are calculated as

$$SE(\theta_{X_1}) = \sqrt{SE(b_1)^2 + 2X_2 COV_{b_1 b_3} + X_2^2 SE(b_3)^2} \tag{14.8}$$

and

$$SE(\theta_{X_2}) = \sqrt{SE(b_2)^2 + 2X_1 COV_{b_2 b_3} + X_1^2 SE(b_3)^2}$$

These require estimates of the standard errors of b_1, b_2, b_3, and the value at which the effect is being conditioned. The standard errors are provided in the output of any OLS regression program. Also required is a covariance (COV) between regression coefficients. These usually are not found in default output in most programs but typically can be obtained by request.

Once a conditional effect and its standard error are calculated, a confidence interval can be constructed as the point estimate plus and minus the standard error times the critical value of t for a given degree of confidence (see Appendix C). For testing the null hypothesis that the conditional effect equals zero, the point estimate can be divided by the standard error and p-value for this ratio derived from the $t(df_{residual})$ distribution.

These computations are not difficult to do by hand, but we don't recommend attempting them, so we don't illustrate them here. It is too easy to mistakenly plug the wrong value into the wrong part of the formula, and even if you implement the formula correctly, rounding error is likely to creep into your computations unless you do them to many decimal places of accuracy even with a spreadsheet calculator of some kind.

We recommend letting a computer do the work for you using a method we call the *regression centering approach*. This method relies on the fact that in a model of the form $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$, b_1 estimates the conditional effect of X_1 on Y when $X_2 = 0$, and regression output includes its standard error, t-, and p-values, and even a confidence interval, if requested. What we seek is an estimate of the conditional effect of X_1 when X_2 is equal to some value λ of our choosing rather than zero. To estimate the conditional effect of X_1 when $X_2 = \lambda$, construct a new variable $X_2' = X_2 - \lambda$ and then estimate

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2' + b_3 X_1 X_2'$$

In this model, b_1 estimates the conditional effect of X_1 on Y when $X_2' = 0$. But notice that $X_2' = 0$ when $X_2 = \lambda$, so b_1 therefore estimates the conditional effect of X_1 when $X_2 = \lambda$. Importantly, $SE(b_1)$ is the estimated standard error

for this conditional effect and the same as you would get if you applied equation 14.8 to calculate the standard error. Obviously, this is easier, and the resulting standard error will be computed much more accurately than you will likely do by hand computation. Using this standard error, *t*- and *p*-values can be derived for null hypothesis testing, or a confidence interval constructed in the usual way.

This same logic could be used to estimate the conditional effect of X_2 on Y when X_1 equals some value ω along with its standard error and *p*-value. Or you could simultaneously center X_1 around some value of interest ω and X_2 around some value λ by constructing X_2' as discussed above and $X_1' = X_1 - \omega$ and then estimating $\hat{Y} = b_0 + b_1 X_1' + b_2 X_2' + b_3 X_1' X_2'$. That would generate the estimates, standard errors, *t*-, and *p*-values for the conditional effect of X_1 when $X_2 = \lambda$ and the conditional effect of X_2 when $X_1 = \omega$.

We illustrate this approach with the HOSPITAL data, estimating the conditional effect of exhaustion (X_1) on the use of safety protocol workarounds among people with 5 years of experience on the job (i.e., $X_2 = 5$), controlling for sex and age as in the earlier example. If we used X_1 and X_2 in the model along with X_1X_2, then the regression coefficient for X_1 along with its standard error would be conditioned on $X_2 = 0$. To condition it on $X_2 = 5$, we subtract 5 from all values of X_2 prior to constructing the product, and then we substitute the centered X_2 in the model for the original X_2. The SPSS code below accomplishes this.

```
compute tenurep=tenure-5.
compute crossprd=exhaust*tenurep.
regression/dep=safety/method=enter exhaust tenurep crossprd sex age.
```

Comparable code in SAS is

```
data hospital;set hospital;
tenurep=tenure-5;crossprd=exhaust*tenurep;run;
proc reg data=hospital;
model safety=exhaust tenurep crossprd sex age;run;
```

and in STATA, try

```
gen tenurep=tenure-5
gen crossprd=exhaust*tenure
regress safety exhaust tenurep crossprd sex age
```

STATA output can be found in Figure 14.1. Observe that the model is

$$\hat{Y} = 3.999 + 0.313X_1 + 0.097X_2 - 0.062X_1X_2 + 0.016X_4 - 0.020X_5$$

```
.
      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  5,   294) =   20.01
       Model |  88.4253474     5  17.6850695           Prob > F      =  0.0000
    Residual |  259.893453   294  .883991335           R-squared     =  0.2539
-------------+------------------------------           Adj R-squared =  0.2412
       Total |    348.3188   299  1.16494582           Root MSE      =  .94021

------------------------------------------------------------------------------
      safety |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exhaust |   .3129059   .0566463     5.52   0.000     .2014222    .4243895
     tenurep |   .0968027   .0559833     1.73   0.085     -.013376    .2069814
    crossprd |  -.0618893   .0162901    -3.80   0.000    -.0939493   -.0298292
         sex |   .0158722    .171829     0.09   0.926    -.3222987     .354043
         age |  -.0201783    .007985    -2.53   0.012    -.0358933   -.0044632
       _cons |   3.998619    .429063     9.32   0.000     3.154195    4.843043
------------------------------------------------------------------------------
```

FIGURE 14.1. STATA output from implementation of the regression centering approach.

The regression coefficient for exhaustion is $b_1 = 0.313, SE(b_1) = 0.057, t(294) = 5.524, p < .001$. So among people with 5 years of experience, the estimated relationship between exhaustion and the use of safety protocol work-arounds is positive and statistically significant. Observe that the regression coefficient for exhaustion in this model is θ_{X_1} and the same as we calculated using equation 14.3. But now we have a standard error, *t*- and *p*-values, and a confidence interval. Use of the regression centering method has avoided the need to apply formulas for θ_{X_1} and $SE(\theta_{X_1})$ using hand computations or other more tedious methods.

The regression coefficient for the X_1X_2 cross-product, its standard error, *t*- and *p*-values will not be affected by centering of X_1, X_2, or both in this fashion. They remain the same, as can be seen by comparing the outputs in Figures 14.1 and 13.4

The RLM macro offers a still easier approach to inference for conditional effects that doesn't even require centering of X_1 or X_2. The SPSS RLM command

```
rlm y=safety/x=sex age exhaust tenure/mod=1/modval=5.
```

tells SPSS to estimate the linear interaction model we've been discussing. The addition of **modval=5** to the command requests an estimate of the conditional effect of the focal predictor when the moderator is equal to 5. In the output, the conditional effect is provided along with a standard error, *t*- and *p*-values, and confidence interval. The SAS version of the RLM command is

```
%rlm (data=hospital,y=safety,x=sex age exhaust tenure,mod=1,modval=5);
```

When the moderator X_2 is dichotomous, the regression output will provide the conditional effect of X_1 for the group coded $X_2 = 0$ *if* you coded one of the groups zero. It may have occurred to you that if you want the conditional effect of X_1 for the other group, you could just rerun the analysis after recoding X_2 so that $X_2 = 0$ for that group. This will work, for it is equivalent to the regression centering approach. If your groups are coded 0 and 1 on X_2, then $X_2' = X_2 - 1$ yields $X_2' = 0$ when $X_2 = 1$.

14.2.2 When the Focal Predictor or Moderator Is Multicategorical

When X_1 or X_2 is multicategorical, the strategy for inference is changed slightly. In the case where the moderator X_2 is multicategorical, the simplest approach to inference about the conditional effect of X_1 is to estimate the model g times using an indicator coding system for coding the g groups on the moderator variable. When you run the regression analysis, the regression coefficient for X_1 will be θ_{X_1} for the reference group, and your output will provide a standard error, t-value, and p-value. Then redo the analysis, making a different group the reference group, and so forth. Repeat for a total of g times.

When the focal predictor is multicategorical but the moderator is dichotomous or numerical, there are two inferences you might want to make. One of these is an omnibus test of equality of the g estimates of Y conditioned on a particular value of X_2. For instance, from the political discussion example from section 13.2.5, you might want to know whether among people who discuss politics 2 days a week, there a difference between Democrats, Republicans, and Independents in how they perceive the Democrat running for President. Or you might want to do specific comparisons as reflected in the coding system for groups, but conditioned on a value of the moderator. For instance, do Democrats who discuss politics 2 days a week differ from Republicans who discuss politics 2 days a week in how they perceive the Democrat?

The omnibus test can be conducted using a variation of the hierarchical regression strategy described in section 5.3.3. This involves estimating two models. Before estimation of either model, X_2 is first centered around the value λ at which you want to condition the inference. That is, compute $X_2' = X_2 - \lambda$. The first model you estimate then includes X_2' and all the

cross-products involving D_i and X_2', but it *excludes* all D_i variables coding groups. So assuming three groups, this first model is

$$\hat{Y} = b_0 + b_3X_2' + b_4X_2'D_1 + b_5X_2'D_2 \tag{14.9}$$

The second model is the same as the first but adds the $g - 1$ D_i variables coding groups. So for two groups,

$$\hat{Y} = b_0 + b_1D_1 + b_2D_2 + b_3X_2' + b_4X_2'D_1 + b_5X_2'D_2 \tag{14.10}$$

Under the null hypothesis of no difference in Y between the three groups when $X_2 = \lambda$, the difference in R^2 between the models can be converted to an F-statistic in the same way described in section 5.3.3. A sufficiently small p-value leads to the inference that the groups differ on average in Y conditioned on $X_2 = \lambda$.

The logic of this method is apparent when you consider that equation 14.9 is identical to equation 14.10 when you impose the constraint that $b_1 = b_2 = 0$. But in the second model, b_1 and b_2 estimate differences between groups on Y when $X_2 = \lambda$. If both ${}_Tb_1$ and ${}_Tb_2$ are equal to zero, this implies that the groups don't differ on average on Y when $X_2 = \lambda$. So we are testing whether relaxing this assumption of equality of the conditional estimates of Y across the g groups (equation 14.10) yields a better-fitting model than when we assume they are the same (equation 14.9).

Equation 14.10 yields regression coefficients b_1 and b_2 that estimate the difference between pairs of group means or sets of means on Y conditioned on $X_2 = \lambda$, along with standard errors, t-values, and p-values. These can be used to test whether the estimates of Y for the groups coded by D_1 and D_2 differ from each other. For instance, if D_1 and D_2 are indicator codes, b_1 and b_2 and their tests of significance provide an inference as to whether the group coded by D_1 differs from the reference group on Y and whether the group coded by D_2 differs from the reference group on Y, conditioned on $X_2 = \lambda$.

Using computer software, implementation involves the combination of the regression centering approach described above with hierarchical entry of regressors. For instance, in SPSS we could test whether Republicans, Democrats, and Independents who talk about politics 2 days a week differ from each other in how they perceive the Democrat by executing the code below, assuming D_1 and D_2 are indicator codes that have already been constructed as described in section 13.2.5:

```
compute pdiscp=pdiscuss-2.
compute crosspd1=pdiscp*d1.
compute crosspd2=pdiscp*d2.
regression/statistics defaults change/dep=demoneg/method=enter pdiscp
    crosspd1 crosspd2/method=enter d1 d2.
```

The second model that allows the estimates of Y to differ between the three groups when $X_2 = 2$ fits better than the one that constrains them to be the same, $\Delta R^2 = .039, F(2, 334) = 9.413, p < .001$. So we can say that the three groups differ on average in how they perceive the Democrat, conditioned on talking about politics 2 days a week.

In SAS, the comparable commands are

```
data politics;set politics;
pdiscp=pdiscuss-2;crosspd1=pdiscp*d1;crosspd2=pdiscp*d2;run;
proc reg data=politics;
model demoneg=d1 d2 pdiscp crosspd1 crosspd2;test
crosspd1=0,crosspd2=0;run;
```

and in STATA, use

```
gen pdiscp=pdiscuss-2
gen crosspd1=pdiscp*d1
gen crosspd2=pdiscp*d2
regress demoneg d1 d2 pdiscp crosspd1 crosspd2
test crosspd1 crosspd2
```

The two regression coefficients for D_1 and D_2 are $b_1 = -0.190, t(334) = -1.658, p = .098$ and $b_2 = 0.233, t(334) = 1.942, p = .053$. So although we can say that Democrats, Independents, and Republicans who talk about politics 2 days a week differ on average in how they perceived the Democrat, we can't definitively say that the Democrats differ from Independents, or that Republicans differ from Independents. If we were to recode the groups, choosing Democrats as the reference, we'd find that they differ from Republicans. Democrats who talk about politics 2 days a week differ from Republicans who talk about politics 2 days a week, $b_1 = 0.424, t(334) = 4.338, p < .001$. More specifically, Republicans are estimated to perceive the Democrat 0.424 units more negatively than Democrats.

The RLM macro has a function built in to conduct this test without having to construct the indicator codes or center the moderator. The SPSS RLM command would be

```
**************************************************************************

Conditional Effect of Focal Predictor at Values of the Moderator Variable

Moderator value:
pdiscuss     2.0000

        Coeff          se           t           p       LLCI       ULCI
D1     -.1904       .1148    -1.6581       .0982     -.4162       .0355
D2      .2334       .1202     1.9415       .0530     -.0031       .4699

Test of equality of conditional means at this value of the moderator
    R2-chng           F         df1         df2           p
      .0386      9.4127      2.0000    334.0000       .0001

Estimated conditional means at this value of the moderator
      party        yhat
     1.0000      1.9630
     2.0000      2.3868
     3.0000      2.1534

**************************************************************************
```

FIGURE 14.2. SPSS RLM output.

```
rlm y=demoneg/x=party pdiscuss/mcfoc=1/modval=2.
```

and in SAS, use

```
%rlm (data=politics,y=demoneg,x=party pdiscuss,mcfoc=1,modval=2);
```

The relevant section of SPSS output can be found in Figure 14.2, where the statistics reported above can be found.

14.3 Probing an Interaction

We have seen how to test whether X_1's effect varies linearly with X_2 in a regression model and how to estimate the effect of X_1 when X_2 is set to a specific value. Researchers typically seek to better understand the nature of the dependency between X_2 and the effect of X_1 on Y as revealed by an interaction analysis by applying inferential statistical procedures such as described above, with the goal of making specific claims about the values of X_2 at which the focal predictor X_1 is related to Y and the values at which it is not. This exercise is often called *probing an interaction*.

14.3.1 Examining Conditional Effects at Various Values of the Moderator

With evidence that X_1 and X_2 interact, a common strategy for making sense of the conditional nature of the effect of focal predictor X_1 on Y is to estimate the conditional effect of X_1 at various values of moderator X_2 and test which of these conditional effects is different from zero using the procedure described in section 14.2. This method is known by various names, such as the *pick-a-point* approach, an *analysis of simple slopes*, or a *spotlight analysis* (Bauer & Curran, 2005; Spiller, Fitzsimons, Lynch, & McClelland, 2013).

In order to implement this approach, one must settle on values of X_2 at which to condition the estimate of X_1's effect. When X_2 is dichotomous or multicategorical, the choice is easy. You simply estimate X_1's effect conditioned on values of X_2 corresponding to the groups. These values manifest themselves in the form of the pattern of codes on D_1, D_2, and so forth.

But when X_2 is numerical, the choice is more difficult. Ideally, the values of a numerical X_2 that are chosen are meaningful in some way, either theoretically or practically. For instance, if your moderator X_2 is number of packs of cigarettes smoked a day, you might estimate the effect of X_1 on Y for people who don't smoke at all ($X_2 = 0$ packs), who smoke $X_2 = 1$ pack a day, and who smoke $X_2 = 2$ packs a day. But sometimes there is no basis for deciding what values of X_2 to choose. In these circumstances, a common convention is to use three values that operationalize "relatively low," "relatively moderate," and "relatively high" on X_2. Some possibilities include a standard deviation below the mean of X_2, the mean of X_2, and a standard deviation above the mean of X_2. An alternative strategy is the 25th, 50th, and 75th percentiles of the distribution of X_2. Each of these is just as arbitrary as the other. These relative values could be based on the distribution of X_2 in the sample, or some other distribution. For example, if X_2 contains scores on some standardized test, then you could use values from the distribution of published test norms.

We have already explained how to conduct an inference for a single conditional effect in section 14.2. Implementation of this method of probing interactions involves repeating this process as many times as desired for different values of X_2 in order to understand a bit about where in the distribution of X_2 the focal predictor X_1 has an effect and where it does not. Using the regression centering method, you would simply reconduct the analysis, say, three times, each time centering X_2 around a particular

```
**************************************************************************

Conditional Effect of Focal Predictor at Values of the Moderator Variable
     tenure     effect         se          t          p       LLCI       ULCI
     3.0000      .4367      .0715     6.1097      .0000      .2960      .5773
     5.0000      .3129      .0566     5.5239      .0000      .2014      .4244
     8.0000      .1272      .0659     1.9314      .0544     -.0024      .2569

Moderator values are 25th, 50th, and 75th percentiles of the moderator distribution

**************************************************************************
```

FIGURE 14.3. RLM output from the pick-a-point approach to probing an interaction.

moderator value, and interpreting the effect for X_1 as conditioned on X_2 being that value you centered it around.

The RLM macro described in Appendix A makes this much easier because it is automated, and it works for dichotomous, numerical, and multicategorical variables. When you specify a model with an interaction, RLM will automatically include in its output a section that provides the conditional effect of the focal predictor at various values of the moderator. If the moderator is multicategorical, it will provide estimates of the conditional effect of the focal predictor in each of the groups, along with standard errors, *t*- and *p*-values, and confidence intervals. If the moderator is numerical, it will implement the pick-a-point approach conditioned on values of the moderator that correspond to the sample mean and a standard deviation below and above the mean. An option also exists to condition the effect of the focal predictor on various percentiles of the distribution of the moderator.

The section of the output generated by the RLM command on page 390 for the safety protocol work-arounds study can be found in Figure 14.3. The **`ptiles=1`** option tells RLM to condition the moderator on the 25th, 50th, and 75th percentiles, and it provides estimates of the conditional effect of exhaustion (the focal predictor) on use of safety protocol workarounds at those values of job tenure (the moderator), along with standard errors, *t*- and *p*-values, and a 95% confidence interval. As can be seen, among those low in job tenure (the 25th percentile, $X_2 = 3$), the conditional effect of exhaustion is positive ($\theta_{X_1} = 0.437$) and statistically significant, $t(294) = 6.110, p < .0001$. It is also positive ($\theta_{X_1} = 0.313$) and statistically significant among those "moderate" in job tenure (at the 50th percentile, $X_2 = 5$), $t(294) = 5.524, p < .001$ Among those relatively high in job tenure (the 75th percentile, $X_2 = 8$), the conditional effect of exhaustion is positive ($\theta_{X_1} = 0.127$) but just misses statistical significance, $t(294) = 1.931, p = .054$.

These conditional effects correspond to the slopes of the lines in Figure 13.5.[1]

RLM implements a comparable method when X_1 is multicategorical. In that case, RLM conducts an F-test of equality of the estimated values of Y between the g groups conditioned on various values of a dichotomous or numerical X_2. It also provides $g - 1$ specific comparisons between groups, as determined by the coding system used.

14.3.2 The Johnson–Neyman Technique

The approach to probing an interaction discussed in section 14.3.1 suffers from the limitation that the data analyst must choose values of the moderator, and typically the choice is arbitrary when the moderator is numerical. Although conventions exist, such as described earlier, these conventions are also arbitrary. The choice of values of the moderator at which to condition the estimate of the focal predictor's effect can influence what an investigator reports as statistically significant and not. Without a rationale for preferring some values rather than others, different investigators who make different decisions could go away with different claims, even when analyzing the same data using the same model.

When the moderator is numerical, the Johnson–Neyman technique, called a *floodlight analysis* by Spiller et al. (2013), avoids this problem by analytically deriving the values of the moderator that represent "points of transition" between a statistically significant and nonsignficant conditional effect. These points of transition demarcate *regions of significance* of the focal predictor's effect. Whereas the pick-a-point approach involves the investigator choosing a value of the moderator and obtaining t- and p-values for θ_{X_1} at that value, the Johnson–Neyman technique asks what value of the moderator produces a p-value for θ_{X_1} that is exactly equal to the α-level chosen for the test, such as .05, or a confidence interval for the conditional effect that just touches zero. Originally introduced by Johnson and Neyman (1936) in the context of ANCOVA, the algebra for this derivation in regression analysis more generally is provided in

[1]Results like these raise an interesting conflict between conservative and parsimonious hypotheses. To say that exhaustion operates at low and medium job tenures but not at higher tenures is to imply an interaction between tenure and exhaustion. Thus, this model is less parsimonious but more conservative than a model that posits an effect of exhaustion at all levels of job tenure. There is no mechanical answer as to which of these models is preferable. Here we say merely that some scientists would consider it reasonable to assume an effect of exhaustion at all levels of job tenure, even if that effect was not significant at all levels, because that assumption produces a simpler model.

Bauer and Curran (2005), but is offered only for dichotomous or numerical focal predictors and requires that the moderator be numerical. See Bauer and Curran for the formulas, which we don't recommend attempting to implement by hand.

When the Johnson–Neyman technique is applied, two solutions will result, but one or both of these solutions may have no use to the researcher. For instance, one or both of the solutions may be beyond the range of measurement of the moderator and thus not be interpretable. Alternatively, one may be an imaginary number. When these are eliminated from the solution, the result is either zero, one, or two values of the moderator at which θ_{X_1} is just statistically significant at the α level of significance. Zero values means that θ_{X_1} is statistically significant for any value of the moderator in the range of the data or at no value in the range of the data. One value means that θ_{X_1} is statistically significant when the moderator is either above or below that value. Two values means that θ_{X_1} is statistically significant when the moderator is below the smallest value *and* above the largest value, or only *between* those two values.

The Johnson–Neyman technique is best left to a computer. No commercially available software that we are aware of provides this as an option when estimating a model with an interaction, but there are packages for R that will conduct it, as well as macros for SPSS and SAS such as MODPROBE (Hayes & Matthes, 2009), PROCESS (Hayes, 2013), and the RLM macro described in Appendix A. For the study of health care workers, adding **`jn=1`** while removing the **`ptile=1`** option from the RLM command on page 390 produces the section of output in Figure 14.4.

Toward the top, we can see that RLM has identified 7.9797 and 15.2716 as points on the continuum of job tenure at which the conditional effect of exhaustion on the use of safety protocol work-arounds is just statistically significant with a p-value of .05. At these two values, θ_{X_1} is 0.129 and −0.323, respectively, as can be seen in the table of values of θ_{X_1}, standard errors, t- and p-values, and confidence intervals in the rest of Figure 14.4. Looking at this output, we see that when job tenure is less than 7.9797, θ_{X_1} is statistically significant and positive. This is one of the regions of significance identified by the Johnson–Neyman technique. But between 7.9797 and 15.2716, θ_{X_1} is not statistically significant. Finally, above 15.2716, θ_{X_1} is statistically significant and negative—another region of significance.

Caution must be exercised when interpreting the results of the Johnson–Neyman technique to make sure you don't interpret certain regions of significance where there is very little data in that region. Toward the top

```
**************************************************************************

Moderator Value(s) Defining Nonsimultaneous Johnson-Neyman Significance Region(s)
      Value    % below    % above
     7.9696    74.3333    25.6667
    15.2716    98.3333     1.6667

Conditional Effect of Focal Predictor at Values of Moderator Variable
     tenure     effect         se          t          p       LLCI       ULCI
     1.0000      .5605      .0956     5.8652     .0000      .3724      .7485
     1.8500      .5079      .0846     6.0010     .0000      .3413      .6744
     2.7000      .4553      .0747     6.0963     .0000      .3083      .6022
     3.5500      .4026      .0661     6.0878     .0000      .2725      .5328
     4.4000      .3500      .0596     5.8701     .0000      .2327      .4674
     5.2500      .2974      .0559     5.3244     .0000      .1875      .4074
     6.1000      .2448      .0554     4.4195     .0000      .1358      .3539
     6.9500      .1922      .0583     3.2963     .0011      .0775      .3070
     7.8000      .1396      .0642     2.1763     .0303      .0134      .2659
     7.9696      .1291      .0656     1.9681     .0500      .0000      .2582
     8.6500      .0870      .0722     1.2049     .2292     -.0551      .2291
     9.5000      .0344      .0818      .4204     .6745     -.1266      .1955
    10.3500     -.0182      .0925     -.1967     .8442     -.2003      .1639
    11.2000     -.0708      .1040     -.6810     .4964     -.2754      .1338
    12.0500     -.1234      .1159    -1.0644     .2880     -.3516      .1048
    12.9000     -.1760      .1283    -1.3720     .1711     -.4285      .0765
    13.7500     -.2286      .1409    -1.6224     .1058     -.5060      .0487
    14.6000     -.2812      .1538    -1.8290     .0684     -.5838      .0214
    15.2716     -.3228      .1640    -1.9681     .0500     -.6456      .0000
    15.4500     -.3338      .1668    -2.0019     .0462     -.6620     -.0056
    16.3000     -.3864      .1799    -2.1483     .0325     -.7405     -.0324
    17.1500     -.4390      .1931    -2.2736     .0237     -.8191     -.0590
    18.0000     -.4917      .2064    -2.3819     .0179     -.8979     -.0854

Alpha level used for Johnson-Neyman method:
  .05

**************************************************************************
```

FIGURE 14.4. Johnson–Neyman output from the RLM macro.

of the output, notice that RLM says that 1.667% of the cases in the data are above 15.2716. In a sample of 300, this is only five cases. Most would be uncomfortable making a claim about the relationship between the focal predictor and the dependent variable in a region of the domain of the moderator where there are so few data. A more sensible interpretation of these results is that there is one region of significance defined as moderator values 7.9797 and less.

14.3.3 Testing versus Probing an Interaction

When probing an interaction using either of these methods, keep in mind that just because X_1 and X_2 interact, that doesn't mean that X_1's conditional effect on Y must be statistically different from zero for some value of X_2 and not others. It is possible that all conditional effects you estimate are

statistically different from zero. It is also possible for none of them to be different from zero, even though X_1 and X_2 interact. The result of a test of interaction says nothing about the pattern of significance or lack of significance of various conditional effects when the interaction is probed.

You may also find yourself tempted to probe a nonsignificant interaction. That is, you might observe after applying one of these methods that for some values of moderator X_2, the conditional effect of X_1 is different from zero, but it is not different from zero for some other values of X_2, yet X_1 and X_2 do not significantly interact. Be careful how you talk about such a pattern, taking care that you don't interpret it as if X_1's effect varies with X_2. A test of linear interaction tests whether X_1's effect, θ_{X_1}, varies linearly with X_2. If the answer to that question is no by a test of interaction, then you can think of X_1's effect as linearly independent of X_2, meaning that θ_{X_1} is not a linear function of X_2. Rather than interpreting a pattern of hypothesis tests for various conditional effects, it would be more sensible to reestimate the model without the X_1X_2 cross-product, which forces X_1's effect to be independent of X_2, and then interpret X_1's partial effect on Y, which is necessarily a constant.

14.3.4 Comparing Conditional Effects

When probing an interaction between X_1 and X_2 from a model that includes their cross-product as a regressor, we can estimate the conditional effect of X_1 at various values of X_2, as discussed earlier. One might think it would be worth testing whether those conditional effects differ significantly from each other. Although this seems reasonable, it can be shown that in a linear interaction model that includes X_1X_2 as a regressor, if X_1 and X_2 interact, then it follows that any two conditional effects of X_1 for different values of X_2 are statistically different, regardless of the two values of X_2 you choose. Conversely, if X_1 and X_2 do not interact, then any two conditional effects of X_1 for different values of X_2 are not statistically different from each other. This is true for both dichotomous and numerical X_2.

This is not intuitively obvious, but it can be shown that the ratio of the difference between two conditional effects of X_1 to the standard error of their difference is equal to the t-statistic for the regression coefficient for X_1X_2. So no further tests comparing conditional effects of X_1 conditioned on different values of X_2 are necessary once you have tested whether X_1 and X_2 interact.

14.4 Complications and Confusions in the Study of Interactions

14.4.1 The Difficulty of Detecting Interactions

Interactions are difficult to detect. It may be that substantial interaction effects just rarely exist in the real world. But this is unlikely, in our opinion, given that it seems improbable that any relationship you might find in a study is invariant across the seemingly unlimited number of possible moderator variables. More likely, such effects do exist as a matter of routine but probably are so small in magnitude that it takes large samples to detect them. That is, hypothesis tests for interaction are simply not very powerful in the kinds of sample sizes that are typical in research.

Given this, one probably has good reason to be initially skeptical of statistically significant interactions found in small samples. It may be that the interaction you have observed in a small sample reflects a real interaction that is large, but it may be just as possible that the interaction you are observing is the result of one or two cases in the data that are highly influential. It is a good idea to apply the kinds of methods discussed in Chapter 16 and rule out influential cases whenever you observe an interaction in a small sample before reporting it as real.

There are various reasons that interactions are hard to detect. In Chapter 17 we talk about how measurement error in regressors can lower power of hypothesis tests. As discussed by Busemeyer and Jones (1983), if X_1 and X_2 are measured with lots of random measurement error, their cross-product will contain still more random error, meaning regression coefficients for product terms may be highly unstable from sample to sample (and potentially biased).

McClelland and Judd (1993) describe and explain how interactions are harder to detect using observational data than when using experimental methods. Their argument rests on the fact that interaction between X_1 and X_2 is ultimately about partial association between X_1X_2 and Y when X_1 and X_2 are partialed out of both. The standard error for the regression coefficient for the cross-product is determined by the size of the residual variance of X_1X_2, and this tends to be smaller in observational studies than in experimental studies, where the investigator has more control over the joint distribution of X_1 and X_2 and therefore the variance of X_1X_2. We make a similar point in section 17.1.2 when we recommend designing studies so that the variance of independent variables is large because the standard error of a regressor's regression coefficient is inversely related to

TABLE 14.1. Curvilinearity Masquerading as Interaction

X_1	X_2	X_1X_2	Y
1	2	2	1
2	1	2	4
3	4	12	9
4	3	12	16
5	6	30	25
6	5	30	36
7	8	56	49
8	7	56	64

the variability of that regressor. X_1X_2 is simply a variable in a regression model, so our recommendation there applies to the analysis of interactions as well.

14.4.2 Confusing Interaction with Curvilinearity

Interaction and curvilinearity can at times be difficult to distinguish, and they can masquerade as each other. Consider, for instance, the data in Table 14.1. You can see for yourself that $Y = X_1^2$ exactly and so $R = 1$ if you were to estimate Y from X_1 and X_1^2. But X_1 and X_2 are highly correlated, and so the X_1X_2 cross-product is highly correlated with X_1^2. Regressing Y on X_1, X_2, and X_1X_2 yields

$$\hat{Y} = 0.5 + 4.5X_1 - 4.5X_2 + 1X_1X_2$$

and the regression coefficient for X_1X_2 is statistically significant, $t(4) = 3.578, p = .024$. Yet we know that Y was constructed without regard to X_2 or the cross-product. That is, there is no interaction, yet an analysis would suggest that X_2 moderates $X_1's$ effect on Y (or the other way around).

Another example with a dichotomous regressor makes the same point. In Figure 14.5 we see a scatterplot depicting the relationship between X_1 and Y among nine women ($X_2 = 0$) and nine men ($X_2 = 1$). If we ignore X_2 entirely and model Y as a parabolic function of X_1 we get the parabola shown, described by the function

$$\hat{Y} = 15.228 - 4.060X_1 + 0.340X^2$$

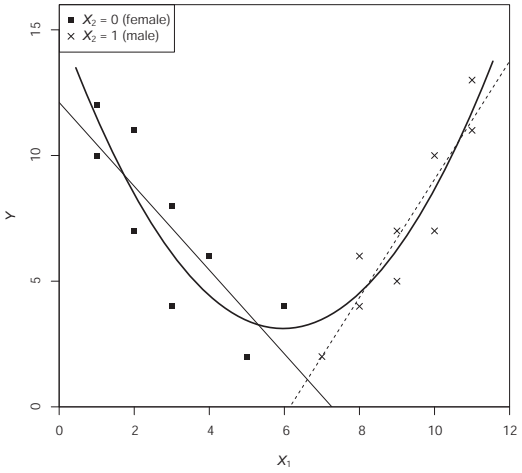


FIGURE 14.5. Curvilinearity and interaction can masquerade as each other.

The fit of this model is excellent; $R = 0.897$. Adding X_2 does not significantly improve the fit of the model, nor does adding both X_2 and X_1X_2. But if we remove X_1^2 and model Y from X_1, X_2, and their cross-product, the resulting model is

$$\hat{Y} = 12.111 - 1.667X_1 - 26.561X_2 + 4.017X_1X_2$$

which also fits very well ($R = 0.886$). Importantly, the regression coefficient for the cross-product is statistically significant. The regression lines relating X_1 and Y can be found in Figure 14.5. As can be seen, the regression weight predicting Y from X_1 does seem to differ dramatically between men and women. The point is that both of these are good accounts of the relationship between X_1 and Y. That relationship could be adequately described as either curvilinear or dependent on sex in this example.

Some argue that tests for interaction should not be accepted at face value without also examining for curvilinearity (either in a separate model or simultaneously) in order to reduce the likelihood of reporting interactions that could reflect nonlinearity rather than true interaction or to avoid misreporting the signs of interactions, such as when the relationship between focal predictor and Y increases with the moderator but the analysis finds the opposite (Ganzach, 1997; Lubinski & Humphreys, 1990). But product

terms and square terms are often highly correlated, so much so that when they are both in the model, standard errors are so large that no effects that would otherwise be found are statistically significant. Indeed, in the example just presented, estimating $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 + b_4X_1^2$ yields a model that fits very well ($R = 0.901$) but neither b_3 nor b_4 is statistically significant, even though both of the curvilinearity and interaction terms are statistically significant when estimated separately.

Lubinski and Humphreys (1990) suggest stepwise methods of variable selection, described in section 7.3.1, as a means of distinguishing between the two possible models, though these methods will not always select the correct model and their performance is dependent on many things (MacCallum & Mar, 1995). Regardless, it is important to recognize that when both nonlinearity and interaction are consistent with the data and also plausible conceptually or theoretically, the analysis may not be able to help you decide which is correct.

14.4.3 How the Scaling of *Y* Affects Interaction

We discussed transformation of variables in section 12.4. When sample sizes are fairly large, the results of a two-sample *t*-test comparing means are rarely much affected by transforming Y—for instance by replacing Y by $\log(Y)$ or e^Y. The same is true in regression with only linear terms. If X_j has a large partial relationship with Y, that relationship rarely vanishes if Y is transformed. But this is not true for interaction. Tranforming Y can greatly affect the size and significance of interactions. When a statistically significant interaction is found, it might vanish under some transformation of Y.

Consider the data in Table 14.2. If you were to regress Y on X_1, X_2, and their cross-product X_1X_2, you'd find exceptionally good fit ($R = 0.995$), with $\hat{Y} = -0.75 + 2.75X_1 + 8.00X_2 + 2.00X_1X_2$. The regression coefficient of 2.00 for the cross-product is statistically significant, $t(12) = 6.928, p < .001$. This regression coefficient means that the slope relating X_1 to Y differs by 2 between the two groups coded with X_1. That is, the slope linking X_1 to Y is steeper when $X_2 = 1$ compared to when $X_2 = 0$.

The fourth column in Table 14.2 contains a square root transformation of Y. Regressing $\sqrt{Y}$ on X_1, X_2, and X_1X_2 yields a perfectly fitting model, $R = 1$, and that perfectly fitting model is $\hat{Y} = 1.00 + 0.50X_1 + 2.00X_2 + 0.00X_1X_2$. The coefficient for the cross-product is exactly zero, meaning that the slopes of the lines relating X_1 to Y are exactly the same when $X_2 = 0$ and $X_2 = 1$. You can see this for yourself in the data without any regression analysis at

TABLE 14.2. A Data Set Illustrating the Effect of Transformations

X_1	X_2	Y	$\sqrt{Y}$
0	0	1.00	1.00
1	0	2.25	1.50
2	0	4.00	2.00
3	0	6.25	2.50
4	0	9.00	3.00
5	0	12.25	3.50
6	0	16.00	4.00
7	0	20.25	4.50
0	1	9.00	3.00
1	1	12.50	3.50
2	1	16.00	4.00
3	1	20.25	4.50
4	1	25.00	5.00
5	1	30.25	5.50
6	1	36.00	6.00
7	1	42.25	6.50

all. In both groups defined by X_2, as X_1 increases by 1 unit, $\sqrt{Y}$ increases by one-half a unit.

This effect of the scaling of Y on interaction applies as much to classical ANOVA as it does to regression analysis. It is not adequate to show that residuals are approximately normal distributions or equal in variance. These are tests of the sampling assumptions of linear models and have nothing to do with this problem, which can occur in very large samples or even if the entire population is available for study.

It is tempting to go away from this example with the message that one should not accept an interaction as real if it goes away with a transformation. However, if the original scaling of Y is highly meaningful and the transformation is arbitrary, the model with the original Y may sometimes be preferable even with the interaction term.

14.4.4 The Interpretation of Lower-Order Regression Coefficients When a Cross-Product Is Present

In a regression model that includes X_1, X_2, and their cross-product X_1X_2, the regression coefficients for X_1 and X_2 do not have the same interpretation as they do when X_1X_2 is excluded from the model. In a model of the form

$\hat{Y} = b_0 + b_1X_1 + b_2X_2$, b_1 and b_2 are *partial* effects. They represent how changes in one variable relate to changes in Y when the other variable is *fixed at any value*. This is what is meant by the term *holding constant*. But in a model of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$, b_1 and b_2 are *conditional* effects. They represent how changes in one variable relate to changes in Y when the other variable is *fixed at zero*. So b_1 quantifies the association between X_1 and Y when $X_2 = 0$, and b_2 quantifies the association between X_2 and Y when $X_1 = 0$.

To illustrate, remember that

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

can be written in two equivalent forms

$$\hat{Y} = b_0 + (b_1 + b_3X_2)X_1 + b_2X_2$$

and

$$\hat{Y} = b_0 + b_1X_1 + (b_2 + b_3X_1)X_2$$

In such a model, X_1's effect on Y is not constant but, rather, a *function* of X_2, as discussed in section 14.1. The effect of changing X_1 on Y will depend on the values of b_1, b_3, and X_2. But if $X_2 = 0$, then this function relating X_1 to Y reduces to a simpler form: $b_1 + b_3(0) = b_1$. So b_1 quantifies the relationship between X_1 and Y under the condition that $X_2 = 0$. By the same reasoning, b_2 quantifies the relationship between X_2 and Y under the condition that $X_1 = 0$, from $b_2 + b_3X_1$, the function relating X_1 to the effect of X_2. When X_1 is set to zero, this function reduces to $b_2 + b_3(0) = b_2$.

This is a very subtle point, but one that is very important point to understand. We have observed many instances in the literature of investigators interpreting b_1 and b_2 as "average" effects or "main effects" as in ANOVA, as the effect of X_1 and X_2 collapsing across the other variable. But that is not what b_1 and b_2 quantify. Indeed, b_1, b_2, or both, along with their tests of significance, may be meaningless if zero is meaningless in the measurement system or outside of the bounds of the measurement scale.

This confusion may stem from the fact that many investigators first learn about interactions in the context of ANOVA, where the rules of interpretation are different and don't generalize to linear models more broadly. There have been many articles written about this confusion (e.g., Friedrich, 1982; Hayes, Glynn, & Huge, 2012; Irwin & McClelland, 2001; Spiller et al., 2013), but the message has been slow to disseminate among users of regression analysis.

A simple solution to this potential for misinterpretation is to mean-center X_1 and/or X_2 prior to testing for interaction. A variable is mean-centered by subtracting its mean from all measurements on that variable. Doing so is not required, but it can aid interpretation of b_1 and b_2 and render them and their hypothesis tests meaningful. We discuss this further in section 14.4.5.

14.4.5 Some Myths about Testing Interaction

It is widely believed that one should never test interaction between X_1 and X_2 by including their cross-product without first mean-centering X_1 and X_2 prior to constructing their product and model estimation. That is, according to this myth, linear interaction should be tested by estimating

$$\hat{Y} = b_0 + b_1(X_1 - \overline{X}_1) + b_2(X_2 - \overline{X}_2) + b_3(X_1 - \overline{X}_1)(X_2 - \overline{X}_2) \qquad (14.11)$$

This practice is often justified by believers of this myth on the grounds that X_1X_2 is likely to be highly correlated with X_1, X_2, or both. By centering X_1 and X_2 around their means prior to constructing the product, the tolerance (see section 4.4.4) of the cross-product goes up, and this will lower the standard error of the regression coefficient for the product and produce a more powerful test of interaction. Thus, you often find people describing how they mean-centered X_1 and X_2 prior to producing the product "to avoid the problems produced by collinearity."

This myth has been widely debunked (see, e.g., Cronbach, 1987; Echambadi & Hess, 2007; Edwards, 2009; Friedrich, 1982; Hayes et al., 2012; Hayes, 2013; Irwin & McClelland, 2001; Kam & Franzese, 2007; Kromrey & Foster-Johnson, 1998; Sheih, 2011; Whisman & McClelland, 2005) but it is worth repeating the argument here in brief. Although it is true the tolerance of the product of mean-centered variables will be higher than the product of uncentered variables, this turns out not to matter, because the variance of the product also changes by mean-centering X_1 and X_2. The change in tolerance and the change in the variance of the cross-product completely offset each other in the standard error computations, resulting in no change in the standard error of the regression coefficient for the cross-product. So the standard error of the regression coefficient for the product is completely unaffected, as is the regression coefficient itself, and the t- and p-values as a result are the same whether you test interaction using centered or uncentered focal predictor and moderator. For a worked example of how these

offsetting changes to the tolerance and variance result in no change to the standard error, see Hayes (2013).

This said, it can be valuable for other reasons to mean-center X_1 and X_2 and estimate the model in equation 14.11 rather than one using the original metrics of X_1 and X_2. As discussed in section 14.4.4, b_1 and b_2 are conditioned on the other variable involved in the interaction being zero. But when you mean-center X_1 and X_2, zero on the mean-centered versions of X_1 and X_2 correspond to the sample means of X_1 and X_2. That means that b_1 in equation 14.11 estimates the effect of X_1 on Y among those average on X_2, and b_2 estimates the effect of X_2 on Y among those average on X_1. These will always be meaningful, as will their hypothesis tests and confidence intervals. So mean-centering can render meaningful those regression coefficients and tests that may not be meaningful if zero is not a meaningful value on the measurement scales.

Another myth you will frequently hear is that to properly test interaction between X_1 and X_2, you must build the regression model hierarchically by first estimating Y from X_1 and X_2 and then, in a second step, entering X_1X_2. If R^2 increases to a statistically significant degree when the cross-product is added, then this is evidence of interaction. In section 5.3.3 we discussed how to test whether adding variables to a model significantly improves the fit of the model. We did so assuming one was entering a set of variables, but a single variable can be thought of as a set of size one, so the method discussed there applies to this simpler version of hierarchical entry.

Although this hierarchical entry of regressors will work, it is required only if the interaction requires more than a single cross-product to represent it, such as in the examples in section 13.2.3 and 14.1.2. It is not required if only a single cross-product is needed to estimate the interaction. Hierarchical entry as discussed in section 5.3.3 will generate an F-ratio for the change in R^2 along with a p-value for testing the null hypothesis of no interaction. But this F-ratio is equal to the square of the t for the regression coefficient for the cross-product, and the two p-values (one for the F and one for the t) will be identical. Mathematically these are the same test. Furthermore, the change in R^2 when the cross-product term is added is equal to the squared semipartial correlation for X_1X_2 in the model that includes X_1, X_2, and their product. So if your regression program generates the semipartial correlation, you don't need to use hierarchical entry even to generate the change in R^2.

14.4.6 Interaction and Nonsignificant Linear Terms

You will often find when estimating a model with X_1X_2 as a regressor that the regression coefficient for the lower-order variables X_1, X_2, or both, are not statistically significant. In such a situation, you may be tempted to remove those nonsignificant variables from the model. Resist this temptation, for doing so will usually bias the test of interaction and also invalidate approaches to probing the interaction.

Often these coefficients and tests of significance are substantively meaningless. Remembering that the regression coefficients for X_1 and X_2 quantify the effect of one variable when the other is zero, it could be that b_1, b_2, or both, are estimating effects that are outside of the range of the observed data. Thus, the tests of significance can't even be interpreted in a meaningful way, and we shouldn't use such meaningless tests to guide our decisions about model construction.

Even if these regression coefficients and their tests are meaningful, the bias that can result is not worth the apparent parsimony that the exclusion of these variables buys. As a general rule, when a product of two regressors is in a model, keep the two variables that were used to construct the product in the model, regardless of the statistical significance of their regression coefficients. There are some circumstances where this rule can be violated (we did so in section 14.2.2, for example), but doing so in these circumstances is founded on relevant estimation principles. Lack of statistical significance is not a sufficiently principled basis for removing nonsignificant regressors that are used to form a cross-product.

14.4.7 Homogeneity of Regression in ANCOVA

In section 9.2, we showed how regression analysis can be used to conduct ANCOVA, which is a statistical technique employed when an investigator wants to compare the means of $g \geq 2$ groups while holding one or more covariates fixed. The literature on ANCOVA discusses an assumption one must make when doing so called *homogeneity of regression*. This assumption states that the relationship between the the covariate and Y is the same in all g groups. This assumption is represented in Figure 9.5, which visually represents the computation of adjusted means. Notice in this figure that the relationship between the covariate and Y is the same; the regression lines are parallel.

Having read this chapter, you can now see how the homogeneity of regression assumption is really an assumption of no interaction between

the covariate and group in the model of Y. You can also see how the difference between the groups on Y would vary with the covariate if the lines in Figure 9.5 were not parallel. This would make any test of differences between adjusted means hard to interpret. For this reason, it is worth testing this assumption prior to interpreting the results of an ANCOVA using the methods discussed in this chapter. If this assumption is not met (i.e., if the covariate interacts with group), then interpretation should focus on how the difference between the groups on Y depends on the covariate, using the methods discussed in section 14.3.

14.4.8 Multiple, Higher-Order, and Curvilinear Interactions

A regression model can include more than one interaction. For instance, you might propose that X_1's effect on Y varies linearly with X_2 and X_3 in an additive matter. Such a model would look like

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + b_5X_1X_3$$

and includes interactions involving X_1. In this model, the conditional effect of X_1 is defined by the function $b_1 + b_4X_2 + b_5X_3$; b_4 quantifies how the conditional effect of X_1 changes as X_2 changes by 1 unit and X_3 is held fixed, and b_5 quantifies how the conditional effect of X_1 changes as X_3 changes by 1 unit and X_2 is held fixed.

In models that include more than one interaction, it is likely that the cross-product terms in the model will be correlated. For instance, if your model includes three cross-products, X_1X_2, X_1X_3, and X_2X_3 representing three two-way interactions, substantial collinearity can result if any pairs of the three variables, X_1, X_2, and X_3, are highly correlated. As a result, they may be confused for each other if you test for each of these two-way interactions separately. This is unimportant if the focus is merely on dismissing interactions, as it sometimes is. But it should be considered if you want to demonstrate the existence of specific interactions.

A two-way interaction can be moderated; that is, it may be dependent on a third variable. Examine the means in Table 14.3 from a hypothetical experiment conducted in two different cities. In City A, the difference in the effect of experimental condition between men and women is $(9 - 8) - (5 - 7) = 3$, while in City B, the comparable value is $(6 - 4) - (3 - 6) = 5$. These values quantify the condition by sex interaction in each city, and they differ between the two cities. This illustrates *three-way* interaction: a condition $\times$ sex $\times$ city interaction. A two-way interaction means that the

TABLE 14.3. A Three-Way Interaction

	City A		City B	
	Male	Female	Male	Female
Treatment	$\overline{Y} = 9.00$	$\overline{Y} = 5.00$	$\overline{Y} = 6.00$	$\overline{Y} = 3.00$
Control	$\overline{Y} = 8.00$	$\overline{Y} = 7.00$	$\overline{Y} = 4.00$	$\overline{Y} = 6.00$

size of a conditional effect changes with another variable, while a three-way interaction means that the size of the two-way interaction changes with another variable. If City A is coded 1 and City B is coded 0, then the three-way interaction in the data in Table 14.3 is $3 - 5 = -2$.

In this example, all three variables are dichotomous. But as with two-way interaction, interaction can be defined when variables are dichotomous or continuous or numerical or multicategorical or any combination of these. Three-way interactions can be included in a model by forming all possible products involving the variables. For instance, a model with a three-way interaction between numerical or dichotomous variables X_1, X_2, and X_3 would look like

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + b_5X_1X_3 + b_6X_2X_3 + b_7X_1X_2X_3$$

If b_7 is statistically different from zero, this means that the two-way interaction between X_1 and X_2 varies with X_3. The symmetry property of interactions applies to higher-order interactions as well. So if b_7 is statistically different from zero, we can also say that the two-way interaction between X_1 and X_3 varies with X_2, or the two-way interaction between X_2 and X_3 varies with X_1.

Still higher-order interactions are defined similarly. A four-way interaction is defined as the change in a three-way interaction for each 1-unit change in a fourth variable, a five-way interaction is defined as the change in a four-way interaction for each 1-unit change in a fifth variable, and so on. Each is constructed by including all possible products of all the variables defining the highest-order interaction as regressors in the model, including all lower-order products and all the variables themselves.

Three-way interactions can be very hard to interpret, and interactions of order higher than three are nearly impossible to interpret. Remember that

a two-way interaction quantifies a difference between differences. A three-way interaction means the difference between differences that characterize a two-way interaction differs with another variable. Four-way interaction means that the difference between the difference between differences differs with a fourth variable, and a five-way interaction means that difference in the difference between the difference between differences differs with a fifth variable!

Although it is not uncommon for researchers conducting ANOVA to test four- or five-way interactions in complex experiments with many factors, this is generally not a good idea in our opinion unless you have a good reason for doing so, such as a strong theoretical orientation that predicts such an interaction. But even with a strong theory, tests on such higher-order interactions tend to be low in power, so a failure to find a significant four- or five-way interaction is not very informative about whether such an interaction actually does exist. But more important, few theories predict such a complex pattern of differences between differences between differences (between differences), and if you are theorizing a four- or five-way interaction, your troubles are going to be more in convincing critics that your theoretical orientation is parsimonious enough for anyone to take seriously than in demonstrating the existence of such an interaction. And you will find it next to impossible to convey your results to anyone in a manner that he or she will be able to keep straight, assuming you are even able to make sense of them yourself.

A curvilinear effect can be moderated. *Curvilinear interaction* is a change in the curvilinearity of one variable as another variable changes. For instance, if X_1 is represented by a parabola through a X_1^2 term in the regression, the coefficient of this square term measures the degree of curvilinearity, equaling zero if the best fit is provided by a straight line. If the optimum degree of curvature for X_1 changes linearly with another variable X_2, that would appear as an interaction between X_1^2 and X_2. Such a hypothesis could be tested by estimating a model that looks like

$$\hat{Y} = b_0 + b_1X_1 + b_2X_1^2 + b_3X_1X_2 + b_4X_1^2X_2$$

with b_4 and its test of significance carrying information about whether and by what amount the curvilinearity in the relationship between X_1 and Y varies with X_2.

As is probably apparent, these models are complex, and proper interpretation of models with multiple interactions, three-way or higher-order interactions, or curvilinear interactions requires care and an organized mind.

We do not discuss the estimation or interpretation of such models in this book. For guidance, see Aiken and West (1991) or Hayes (2013).

14.4.9 Artificial Categorization of Continua

In section 5.1.6 we cautioned against categorizing numerical variables prior to analysis. We repeat that warning here. Investigators often categorize continuous variables prior to testing interaction. Perhaps because many researchers are not familiar with the general regression strategy for testing interactions that we have covered in the last two chapters, they instead choose to split a continuous variable into low and high groups based on a mean or median split and then conduct a more familiar factorial ANOVA to test interaction. The reasons we offered in section 5.1.6 against artificial categorization apply here too. Artificial categorization prior to testing for interaction can either lower power to detect interactions that are real or produce false interactions that are not (Hayes, 2005; Humphreys, 1978; Maxwell & Delaney, 1993; Veiel, 1988). Given that it is no more difficult to test for interaction when focal predictor or moderator is continuous relative to when both are categorical, this practice is rarely justified.

14.5 Organizing Tests on Interaction

The number of possible interactions that can be constructed and estimated in a regression model can be very large. If there are 10 regressors that are numerical or dichotomous in any combination, there are $(10\times9)/2 = 45$ possible two-way interactions, meaning you would need a sample no smaller than 56 to estimate the model, and that wouldn't allow any inference because $df_{residual}$ would be zero if $N = 56$. Fit would be perfect and inference could not be undertaken unless you had a sample of larger than 56, and you'd likely be grossly overfitting the data unless you had a lot more than 56 cases. If you care to also estimate all possible three-way interactions, there are 120 of them, meaning you need a sample size no smaller than 176 just to calculate the effects. If interactions involve multicategorical variables, the needed sample size is even bigger, because interactions with multicategorical variables require more regression coefficients than interactions that involve dichotomous or numerical regressors.

Ignoring the needed sample size, which may not be problematic in many research situations, how would you go about testing so many possible interactions? Should you test them at all? Should you test them individually or as sets? How do you organize the sets? Do you group interactions by

variables involved (e.g., all interactions involving X_1) or by order of the interactions (e.g., all two-ways, all three-ways, and so forth)? These are important questions, and we begin to address them in this section.

Curiously, we often study interactions in order to dismiss them—to dismiss either their existence or their relevance, thus allowing the conclusion that the study's major conclusions apply equally to all cases. This fact can influence the analysis of interactions, as we shall see.

14.5.1 Three Approaches to Managing Complications

Complications always arise when analyzing data, and the number of possible interactions one could estimate may be considered a complication. In any statistical analysis, there are three general ways to handle complications, in order of increasing conservativeness:

1. Assume their absence—the standard method for some researchers as well as in introductory classes, where dealing with complications is beyond the scope of the class or perhaps the knowledge of the researcher.

2. Check for them, and assume their absence if they are not clearly present. If present, use an alternative statistical method that may be less powerful but is valid even in the presence of the complications.

3. Use the less powerful method even when no complications are detected, on the ground that the absence of complications is a null hypothesis that can never be proved.

The choice among these three methods arose in section 4.7.3. There we considered the possibility of simply failing to analyze covariates of doubtful importance (method 1 in the list above) or deleting from the model covariates found to have nonsignificant effects (method 2 in the list). We criticized both these approaches and recommended method 3, which in that case meant including covariates in the model even if they were both nonsignificant and of doubtful importance. But we shall see in section 14.5.2 that this conservative approach cannot be consistently applied when the complications are interactions.

If we consider all possible interactions in a model with k regressors, then there are even more interactions than implied in the earlier discussion. If we considered every possible multiplicative interaction involving k numerical regressors (i.e., all interactions up to order k), there would be $2^k - k - 1$

possible interactions. If $k = 10$, this amounts to over 1,000 interactions. You would need a sample of over 1,000 cases just to estimate such a model. And this ignores that nonmultiplicative interactions can be defined, though they are not discussed in this book. Therefore, we simply have to assume the absence of some interactions in most real-world research scenarios.

This problem is made more manageable by the fact that most tests for interactions yield nonsignificant results, and this is even more likely to be so for higher-order interactions. Thus, we typically ignore the highest-order interactions not because of their impossibility but because of their implausibility. Many investigators do not check for interactions at all—a practice we don't necessarily endorse. Most researchers do not check for three-way interactions unless their theory or hypothesis leads them to expect such an interaction. This seems more reasonable. Thus, of the three methods listed earlier, method 1, the least conservative, is actually the standard method for analyzing higher-order interactions. Simple interactions, such as two-way interactions, probably should be tested routinely, and certainly if your theory or hypothesis predicts it or you designed the study expecting it.

14.5.2 Broad versus Narrow Tests

We can distinguish between three types of tests for interaction:

- An overall test that includes in a single model all the interactions to be tested and that tests change in fit when all are dropped from the model.
- Variable-by-variable tests that perform one test for all the interactions involving a specific regressor—for instance, all interactions involving age.
- Simple interaction tests that test a specific interaction.

Tests on individual terms would typically be corrected for the number of tests performed—a number that may be large. When performing variable-by-variable tests, you should probably also correct for the number of tests performed (see Chapter 11), but that number will only be k. Each interaction is then counted twice; for instance, an age × income interaction is included in both the test of all interactions involving age and in the test of all interactions involving income. But this does not violate any assumptions.

For reasons that will soon become clear, we call the first of these approaches the broadest of the three, and the last the narrowest. One advan-

tage of breadth was first mentioned in section 4.7.2. Tests of vague null hypotheses are usually conducted with more power than tests of specific null hypotheses. Broad tests yield vague conclusions, and vague conclusions are reached more easily than specific conclusions. If several different interaction effects are largely independent and each has a small effect, a test of them as a set may be testing the significance of a larger effect, so that test may be more powerful than separate tests on separate interaction terms.

Another advantage of broader tests for interaction is that they are capable of detecting effects despite complementarity *among interactions*. For instance, suppose occupational prestige correlates highly with education, and suppose a dependent variable of self-confidence is determined among men largely by occupational prestige relative to education, whereas this effect does not operate among women. Then sex interacts strongly with a linear function of education and occupational prestige, even though the individual sex × education and sex × prestige interactions may both be small.

The disadvantages of broad tests of vague null hypotheses are important when the number of terms is large relative to sample size, and this is more likely to occur with interaction terms than with lower-order terms. The first disadvantage of broad tests occurs in its extreme form when there are many interaction terms, but only one has any effect on Y. Then a narrow set of tests is more likely to detect it. But the potential size of this advantage is limited by the number of interaction terms since the variance explained by a single interaction term cannot exceed the variance explained by the set of all interaction terms. On the other hand, under complementarity, the countervailing disadvantage of narrower tests can theoretically be enormous.

The second advantage of narrower tests relative to broad tests is that they use fewer degrees of freedom, leaving more for the residual. The smaller the sample, the more important this advantage. In an extreme case, an overall test can use up all the degrees of freedom and then some, so its power is necessarily zero.

These countervailing advantages make a simple choice among the three approaches very difficult. About all that can be said unambiguously is that the relative advantages of broader tests increase with sample size, so the broader tests are especially recommended when samples are large and the narrower tests when samples are small, with variable-by-variable tests perhaps the most appropriate for intermediate sample sizes. If the purpose of the analysis is to dismiss the interactions, then the most conservative

approach is to perform all three types of tests, with the hope that all yield nonsignificant results.

Of course, a failure to reject a null hypothesis does not prove the null hypothesis. A nonsignificant interaction may still exist in reality. So even if an interaction is nonexistent, a very conservative researcher might want to take into consideration the possibility that it exists by retaining it in the model. If the interaction involves only covariates, this is not a problem. But if the interaction involves independent variables, interpretation and discussion are harder, because you can't really talk about one variable's effect without conditioning it on the other variable it is (nonsignificantly) interacting with in the model. This invariably leads to awkward interpretations and writing that seem to convey that an interaction exists when you have not been able to establish that it does by the kind of evidence that scientists expect to see.

14.6 Chapter Summary

In a regression model that includes X_1, X_2, and X_1X_2, X_1's effect on Y can be expressed as a linear function of X_2, and X_2's effect can be expressed as a linear function of X_1. Using the function, an estimate of the conditional effect of one variable at a given value of the other can be calculated, and inference about its size undertaken with an estimate of the standard error. These computations are best left to a computer either through the regression centering strategy, which produces conditional effects and their standard errors without having to rely on hand computation, or relying on macros that others have produced, such as the RLM macro described in Appendix A.

It is common when evidence of interaction exists to *probe* the interaction by conducting inferences about various strategically or arbitrarily chosen conditional effects, in order to understand where in the distribution of the moderator the focal predictor is significantly related to the dependent variable. Most typically the investigator chooses a set of values of the moderator and conducts an inference about conditional effects at those values. But the arbitrariness of the selection of moderator values makes this strategy less attractive than the Johnson–Neyman technique, which algebraically derives "regions of significance"—the range or ranges of values in the domain of the moderator where the focal predictor is significantly related to the dependent variable and where it is not. This approach elimi-

nates the need to choose values of the moderator in advance when probing an interaction.

Entire books have been written about interactions in regression analysis, because it is a complex topic and it is easy to misanalyze your data or misinterpret your results if you don't know what you are doing. Interactions are generally somewhat difficult to detect, and when they are detected, they can be the result of nonlinearities rather than actual interaction. Interactions can also come and go with various transformations of Y. Perhaps the biggest difficulty that researchers seem to have is misinterpreting conditional effects that come out of a regression analysis with a cross-product as a predictor as if they are "main effects" in an ANOVA sense. There are also various myths circulating about how to properly test interaction, including that the variables involved in an interaction must first be mean-centered or that hierarchical entry is required to test interaction. We have debunked those myths in this chapter.

Even in a modest regression model with very few regressors, the number of potential interactions one could estimate can be large. Various strategies for managing the analysis of interactions can be employed, such as assuming their absence, testing for them as sets, testing them one variable at a time, or testing them individually. Most of these strategies involve multiple hypothesis tests, and researchers looking for interactions should be aware of the multiple test problem and the possibility of overfitting one's data when exploring in search of interactions.

15
Mediation and Path Analysis

There is always some kind of a process at work behind a relationship between two variables, whether psychological, sociological, cognitive, or biological in nature. Causal effects operate through *mechanisms*—a sequence of steps in which an independent variable causally influences a dependent variable by affecting an intermediary or *mediator* variable or variables, which then carry their own causal effect onto the dependent variable. In this chapter we overview *path analysis* as a means of statistically assessing *mediation*. We describe the algebra by which an independent variable's effect on a dependent variable can be broken into direct and indirect pathways of influence. After discussing inference about indirect effects and models with multiple mediators, we overview various complications as well as extensions to mediation analysis that can be undertaken using linear regression.

Good research often goes further than just establishing that there is some relationship between an independent and a dependent variable. As first discussed in section 6.2.1, even in experiments in which you can conclusively say that *X causes Y*, for all you know there may be some group of people for whom *X* does not cause *Y*, or where the effect of *X* on *Y* is even the opposite of what you observed in your study. Chapters 13 and 14 addressed how to examine the extent to which an independent variable's effect depends on another variable; this was interaction or moderation. These techniques can be used to help understand the contingencies or boundary conditions of a relationship or an effect.

Establishing that two variables are related also may not say much of anything about how that relationship comes to be, a point we also raised in section 6.2.1 and again in section 6.3.3. That is, what is the process or mechanism at work that leads to *X* being related to *Y*, or *X* causally influencing *Y*? For instance, knowing that a particular method of teaching increases student performance on a standardized math test leaves open questions as to how that effect operates. Does the method increase students'

motivation, how deeply they process information, or their ability to recall information or use it in new ways? True understanding of a phenomenon exists when we can say not only that X and Y are related, but also how that effect operates.

Establishing cause and effect is far more than just data analysis. In fact, by some arguments, statistics really doesn't have too much to say about cause–effect, as research design is ultimately key when it comes to the kinds of cause–effect claims one can make with data. But statistics can be used to quantify effects, rule out certain alternative interpretations for a relationship observed, and assess the role of sampling variability or "chance" in study findings. With the limitations of statistical analysis and the role it can play in answering cause–effect questions in mind (a point we again raise at the end of this chapter), here we address the topic of *mediation analysis* using linear regression. Mediation analysis is used to quantify *pathways of influence*, or the process or processes by which an independent variable can influence a dependent variable. Using the methods discussed in this chapter, you can ascertain the extent to which a variable can be said to be functioning as a mediator of the relationship between independent and dependent variable.

15.1 Path Analysis and Linear Regression

15.1.1 Direct, Indirect, and Total Effects

In a causal system, one variable may affect another *directly*, *indirectly* through other variables, or both directly and indirectly. For instance, exercise might affect weight loss in three ways. It could directly influence weight loss as calories are consumed during exercise. Or it could indirectly influence weight loss by curbing or stimulating appetite and therefore food consumption, or by raising metabolic rate, which increases how quickly calories are burned.

In a regression model estimating dependent variable Y from a set of two or more regressors, the regression coefficient for regressor X_j quantifies X_j's direct effect on Y, ignoring any indirect effects that may work through other variables. So when we regressed weight loss onto exercise, food intake, and metabolism in the example in section 3.2.4, we found a regression weight of 1.045 for exercise—its direct effect—meaning that when food intake and metabolism are held constant, each additional hour of weekly exercise translates into 1.045 units (104.5 grams) of average weekly weight loss. It ignores any effect that exercise might have on food consumption or

metabolism, both of which could affect weight loss as well. So if you wanted to know the amount that exercise affects weight loss in the aggregate, controlling for metabolism and food intake may not be a good idea.

Sometimes we are interested not in the direct or indirect effects but instead the *total effect* of some independent variable. As we discuss later, a variable's total effect is the sum of its direct and indirect effects. For instance, suppose a state's Department of Education is considering raising the requirements for promotion from eighth grade to ninth grade, in the hope that this will encourage students to study harder, and this will increase their average performance as measured by a nationally standardized achievement test. So they conduct an experiment in which students in some school districts have a higher requirement imposed and students from other districts have no such higher requirement, to see what affects this has on test performance across the state.

We could depict a causal system such as this one in the form of a *path diagram*, as in Figure 15.1, panel A. In this diagram, the arrows represent assumed causal affects, with the presumed direction of causal flow progressing in the direction of the arrow. So this diagram shows that higher requirements are assumed to affect study time, which in turn affects achievement as measured by performance on the test. This is the indirect effect of requirements on achievement through study time. Study time is called a *mediator* variable in this process, because it is the conduit through which requirements affect achievement. The indirect effect represents one mechanism by which changing requirements might influence achievement. But this model also includes a path from requirements directly to achievement—the direct effect of requirements. This direct effect means that raising requirements may influence achievement through some process other than study time. Because all effects must be mediated by *something*, a direct effect essentially represents another process or mechanism—an unspecified mediator—that is not formalized in the path diagram.

In this example, the school district would probably be most interested in assessing the total effect of a raised requirement on achievement. Regressing achievement on requirements (coded, for instance, 0 and 1 for lower and higher requirement school districts) while controlling for study time would not generate an estimate of the total effect of the requirements on achievement but, rather, just the direct effect, independent of its effects on study time. So we should not control for study time. Doing so is much like selecting a group of students who study the same amount but

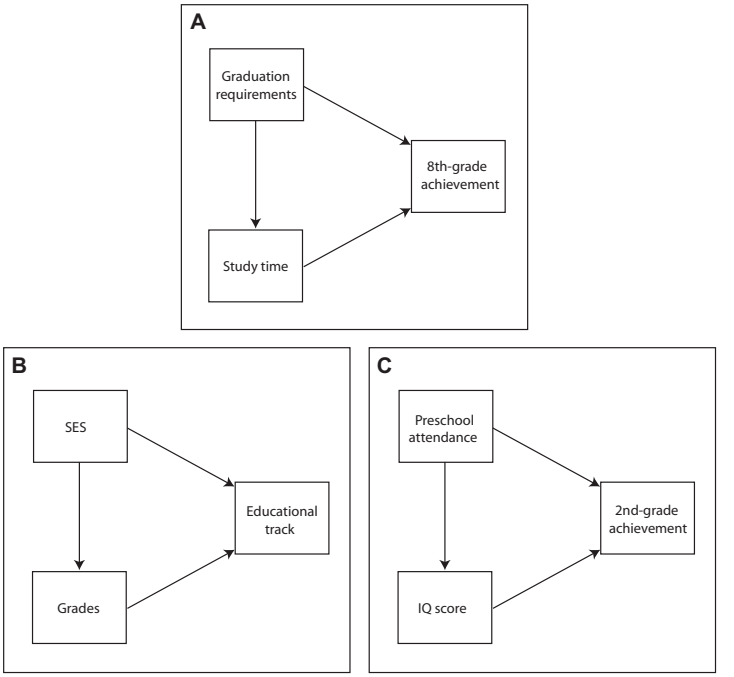


FIGURE 15.1. Three path diagrams depicting direct and indirect effects of graduation requirements on eighth-grade achievement (panel A), SES on educational track placement (panel B), and preschool attendance on second-grade achievement (panel C).

come from districts that differ in requirements. If the new requirements improve achievement only by making students study harder, we would expect to find no difference in average achievement between students coming from districts with different requirements when average study time is controlled. If we find an effect of requirements in this model, manifested by the regression coefficient for the dummy variable coding district type, this estimates the effect of requirements that operates through something other than study time—the direct effect. If we want to estimate the average difference in achievement between the two types of districts, we should do so without controlling for study time.

Both direct and indirect effects may provide useful information. If we control for study time, then the direct effect of requirements on achievement could tell us whether the change in requirements affects achievement by any mechanism *other than* by increasing study time. For example, the difference

in achievement between students who study the same amount but come from districts that differ in requirements may be attributable to something that the teachers are doing to raise achievement through some other process, such as the expectations the teachers have for good performance, which may increase motivation that manifests its effects through something other than time spent studying. Maybe the missing mediator is how efficiently the students study, rather than how much. Without a measure of efficiency of a child's study, this can't be tested. So the direct effect may be quantifying some other process at work that is not a part of the causal system being formally estimated.

In some situations, the direct effect or indirect effect (or both) may be of more interest than the total effect. For example, suppose we assume that SES affects grades in school, and that grades affect the school track that teachers recommend to a child's parents. Thus, we assume that SES affects track placement indirectly through grades, as depicted in Figure 15.1, panel B. Knowing whether this is true could be important theoretically and practically. But it would also be very interesting to know if SES also affects track placement directly. Are children from a middle-class background more likely to be placed on a high track than children from a working-class background who have earned the same grades? If so, what might this say about the system used to determine the educational track of a child? Might there be some kind of systematic bias that enhances opportunities for middle-class children and works against opportunities for working-class children?

Or consider the effect of preschool attendance on school achievement as measured at the end of second grade. Suppose an experiment had been done in which some children were randomly assigned to a preschool program, while others were assigned to a control group. There is then no requirement to control for factors like SES, because it can be assumed that those factors relate only randomly to the independent variable of preschool attendance. Controlling for it may increase the power of tests, but it isn't necessary. But some educators might argue that preschool attendance produces a temporary artifactual inflation of IQ test scores, that children with high IQ scores were given more attention by the first- and second-grade teachers, and that this mechanism produced the positive effect of preschool on achievement scores at the end of the second grade. In Figure 15.1, panel C, this hypothesis is represented by the indirect path from preschool to achievement through IQ scores. This is clearly important to assess. But other educators might consider this implausible, or they may

argue that it does not matter how preschool affects achievement so long as the effect does exist. The former educators would be most interested in the indirect and direct effect of preschool on achievement with IQ scores controlled; the latter group would be most interested in the total effect, with less concern about the direct and indirect effects.

15.1.2 The Regression Algebra of Path Analysis

The total, direct, and indirect effects in a path model or *mediation model*, as described in section 15.1.1, can be estimated with least squares regression analysis. Figure 15.2 represents a model with a single mediator, also called a *simple mediation model*, in generic form. Each of the causal arrows is labeled with a regression coefficient from one of the regression models in equations 15.1, 15.2, and 15.3.

The total effect of independent variable X_1 on dependent variable Y is estimated by regressing Y on X_1, as

$$\hat{Y} = b_0 + b_1 X_1 \tag{15.1}$$

Using regression analysis, we can break the total effect b_1 into two components—direct and indirect. The direct effect of X_1 on Y comes from a model of Y that includes X_1 and mediator variable X_2 as regressors

$$\hat{Y} = b_0 + b_2 X_1 + b_3 X_2 \tag{15.2}$$

where b_2 is the direct effect of X_1 on Y. The computation of the indirect effect of X_1 on Y through X_2 requires a third regression quantifying the effect of independent variable X_1 on mediator variable X_2:

$$\hat{X}_2 = b_0 + b_4 X_1 \tag{15.3}$$

Once b_4 is estimated, the indirect effect of X_1 on Y through X_2 can be calculated as the product of b_4 from equation 15.3 and b_3 from equation 15.2:

$$\text{Indirect effect of } X_1 \text{ on } Y = b_4 b_3 \tag{15.4}$$

With the direct and indirect effects defined, we can now express the total effect of X_1 as the sum of its direct and indirect effects. That is,

$$b_1 = b_2 + b_4 b_3 \tag{15.5}$$

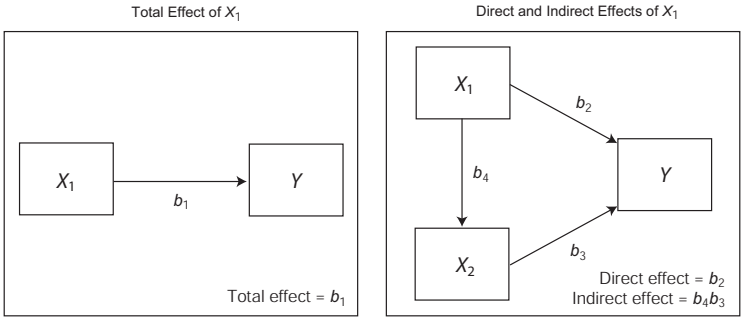


FIGURE 15.2. Path diagrams depicting the total, direct, and indirect effects of X_1 on Y, with the indirect effect operating through a single mediator X_2.

Equation 15.5 is always true whenever X_2 and Y are estimated using ordinary least squares regression. Because b_1 is equivalent to the sum of the direct and indirect effects, we could calculate b_1 directly using equation 15.1 or indirectly using equation 15.5 to get the total effect of X_1 on Y. However, only the direct calculation when conducted with regression software yields standard errors and statistical inferences for the total effect.

The direct, indirect, and total effects can be interpreted like regression coefficients, though the indirect effect is actually estimated as a product of two regression coefficients. The total effect b_1 estimates by how much two cases that differ by 1 unit on X_1 are estimated to differ on Y. The direct effect b_2 estimates by how much two cases that differ by 1 unit on X_1 but are equal on X_2 are estimated to differ on Y. The indirect effect b_4b_3 estimates the amount by which two cases that differ by 1 unit on X_1 are estimated to differ on Y through the sequence of causal steps in which X_1 affects X_2, which in turn carries its effect on to Y.

We illustrate these computations using the exercise and weight-loss data first introduced in Chapter 3. The model we estimate allows exercise frequency, X_1, to affect weight loss, Y, directly as well as indirectly through food intake, X_2. By this process, it is expected that when food intake is held constant, more exercise should translate into greater weight loss. This is the direct effect of exercise frequency on weight loss. But exercise could also increase appetite, manifested through greater food intake among those who exercise more. This increase in consumption of calories could decrease weight loss. The is the indirect effect of exercise frequency on weight loss through food intake.

Estimation of the linear regression models in equations 15.1, 15.2, and 15.3 yields

$$\hat{Y} = 4.000 + 1.750X_1$$

$$\hat{Y} = 6.000 + 2.000X_1 - 0.500X_2$$

$$\hat{X}_2 = 4.000 + 0.500X_1$$

and thus $b_1 = 1.750$, $b_2 = 2.000$, $b_3 = -0.500$, and $b_4 = 0.500$. The total effect of exercise frequency on weight loss is b_1. Each hour of exercise per week is associated with 1.750 units, or 175 grams, of weight loss per week. This total effect breaks into two components. The direct effect of exercise on weight loss is $b_2 = 2.000$, meaning that among people who consume the same number of calories above the minimum recommended, the person who exercises 1 additional hour per week more is estimated to lose 2.000 more units of weight, or 200 grams. The indirect effect of exercise frequency on weight loss through food consumption is $b_4 b_3 = 0.500(-0.500) = -0.250$. This is a *decrease* in weight loss as a result of exercise. This negative indirect effect results from greater food intake among those who exercise more ($b_4 = 0.500$ units of calories per hour of exercise), and each 1 unit of food intake results in a $b_4 = -0.500$ unit (50 gram) increase, meaning a *decrease*, in weight loss. Observe that as stated by equation 15.5, the direct and indirect effects do indeed add up to the total effect of exercise frequency on weight loss: $b_2 + b_3 b_4 = 2.000 + (-0.250) = 1.750 = b_1$.

We have estimated and expressed the direct, indirect, and total effects of X_1 in unstandardized form, meaning they are interpreted with respect to the original units of measurement of the variables. Standardized regression coefficients $\tilde{b}$ can be substituted into equations 15.1 through 15.5, and they still apply (but remember that we don't recommend expressing effects in standardized form when X_1 is a dichotomous variable; see section 5.1.5). In that case, the interpretations of these effects as discussed earlier generalize, except "1 unit" becomes "1 standard deviation."

15.1.3 Covariates

Our discussion thus far has assumed no covariates. But this is not an assumption of the path analysis algebra. We can add covariates to a path analysis represented by equations 15.1, 15.2, and 15.3 just by including them as regressors. However, equation 15.5 will be true only if the same covariates are included in all three of the these equations.

Suppose, for example, you wanted to include age (C_1) and weight in kilograms at the start of the study (C_2) as covariates in the weight-loss path analysis just presented. By including C_1 and C_2 in the models of Y and X_2, as in

$$\hat{Y} = b_0 + b_1X_1 + b_5C_1 + b_6C_2 \tag{15.6}$$

$$\hat{Y} = b_0 + b_2X_1 + b_3X_2 + b_7C_1 + b_8C_2 \tag{15.7}$$

$$\hat{X}_2 = b_0 + b_4X_1 + b_9C_1 + b_{10}C_2$$

then equation 15.5 still holds. All the interpretations of the total, direct, and indirect effects described in section 15.1.2 apply, but with the addition of "holding age and initial weight constant." But violating this rule (by, e.g., putting C_1 and C_2 only in the model of X_2 or putting C_1 in the model of Y and C_2 in the model of X_2) results in a total effect of X_1 that does not equal the sum of the direct and indirect effects. This should make sense, because you can't interpret the total, direct, and indirect effects as holding the covariate set constant if you haven't held the same covariates constant in all of the equations that are used to generate these effects.

15.1.4 Inference about the Total and Direct Effects

The total and direct effects of an independent variable on a dependent variable are quantified with regression coefficients b_1 and b_2 in equations 15.1 and 15.2 or, if covariates are included, equations 15.6 and 15.7. All regression programs will provide a standard error of these regression coefficients that can be used for testing a null hypothesis about these effects, or a confidence interval can be constructed in the usual way.

In the exercise and weight-loss example, the total effect of exercise is statistically different from zero, $t(8) = 4.850, p = .001$, as is the direct effect, $t(7) = 6.000, p < .001$. The degrees of freedom for these tests are the residual degrees of freedom for the corresponding model of the dependent variable from which the estimates are derived.

15.1.5 Inference about the Indirect Effect

When your hypothesis focuses on whether a variable is functioning as a mediator of the effect of the independent variable on the dependent variable, you need to be able to rule out chance as an explanation for the obtained indirect effect. A hypothesis-testing procedure can be used to test the null hypothesis that the true indirect effect, ${}_Tb_4{}_Tb_3$, equals zero, or you can construct a confidence interval for the indirect effect. Although

a hypothesis-testing framework is widely understood, confidence interval approaches with a confidence interval constructed in a specific way are more widely recommended, for reasons that will be made clear in this section.

With an estimate of the standard error of b_4b_3, we could proceed with inference in the usual way by computing a p-value for the ratio of the indirect effect to its standard error or by constructing a traditional confidence interval as the point estimate plus or minus about 2 standard errors. There are a few formulas for the standard error of the product of two statistically independent regression coefficients in the literature (b_4 and b_3 are statistically independent in this model). The simplest is

$$SE(b_4b_3) = \sqrt{b_4^2SE^2(b_3) + b_3^2SE^2(b_4)} \tag{15.8}$$

Equation 15.8 requires only b_3, b_4, and their standard errors, and these are available in any regression output. Sobel (1982) suggests using $Z = b_4b_3/SE(b_4b_3)$, with Z interpreted as a standard normal variable, and the p-value for testing the null hypothesis that ${}_Tb_{4T}b_3 = 0$ derived from a table of standard normal probabilities (see Appendix C) or a computer algorithm. In this example,

$$SE(b_4b_3) = \sqrt{(0.500^2)0.252^2 + (-0.500^2)0.433^2} = 0.251$$

and so $Z = -0.250/0.251 = -0.998$, with a two-tailed p-value of .318. We can say that the indirect effect is not statistically significant. A 95% confidence interval would be $b_4b_3 \pm 1.96SE(b_4b_3)$, which in this example is -0.742 to 0.242.

If you are going to use this *Sobel test*, we recommend doing the computations to many decimals places, as adding squares of small numbers can introduce lots of rounding error if the computations are done to only a few decimal places. But we don't recommend using this test anyway. The problem with the Sobel test is that the sampling distribution of the product of regression coefficients is not normal, or even symmetrical, so using the normal distribution for generating a p-value is not appropriate. Most experts in the statistical analysis of mediation discourage the use of this test in part for this reason.

Two better alternatives are the bootstrap confidence interval and the Monte Carlo confidence interval. These methods don't make any assumption about the shape of the sampling distribution of b_4b_3, but both require a computer, as they are computationally intensive and require many repet-

itive computations. We discuss the mechanics of bootstrapping in section 16.3.3. Suffice it to say now that in the bootstrap method, we construct many many estimates of the indirect effect by constructing a new data set with N cases from the existing data set by randomly sampling the N rows of the original data with replacement. We estimate the indirect effect in this new data in the same way we did in the original data, and then we repeat this process thousands of times to generate the *bootstrap distribution* of the indirect effect. Using the distribution of these thousands of *bootstrap estimates* of the indirect effect, we form a 95% confidence interval for the indirect effect as the values in the bootstrap distribution that define the 2.5th and 97.5th percentiles of the distribution.

A bootstrap confidence interval requires estimation of the mediation model thousands of times. A similar method that requires only one estimation of the model is the Monte Carlo confidence interval. For this method, you estimate b_3 and b_4 and their standard errors using regression analysis. Once these are calculated, they are used as inputs into an algorithm that generates a random draw from a normal distribution with a mean of b_4 and a standard deviation of $SE(b_4)$, which is then multiplied by a random draw from a normal distribution with a mean of b_3 and a standard deviation of $SE(b_3)$. Most statistical packages and computing languages have routines for generating random draws from various probability distributions. Like bootstrapping, this process of multiplying random draws from normal distributions is repeated thousands of times to produce a Monte Carlo distribution of the indirect effect. This Monte Carlo distribution is used to construct a 95% confidence interval for the indirect effect as the values in the Monte Carlo distribution that define the 2.5th and 97.5th percentiles of the distribution.

We discuss some programs you can use for generating bootstrap or Monte Carlo confidence intervals in section 15.1.6, using a more realistic example. Using some of these programs, a 95% bootstrap confidence interval for the indirect effect in the exercise and weight-loss example was −1.013 to 0.100. Using the Monte Carlo method, we got −0.892 to 0.174. Because the confidence interval includes zero, we cannot confidently rule out chance as the explanation for the obtained indirect effect. If the confidence interval did not include zero, then this would be consistent with mediation of the effect of exercise frequency on weight loss by food intake.

Research has shown that both of these confidence interval methods tend to perform pretty well and better than the Sobel test (see, e.g., Hayes & Scharkow, 2013; Preacher & Selig, 2012). More specifically, the Sobel

test tends to be lower in power than either the bootstrap or Monte Carlo confidence interval. Even if they were equal in power, the Sobel test relies on a patently false assumption that other methods avoid—that of normality of the sampling distribution of the product of regression coefficients. Why make an assumption you don't need to make when there are tests that are just as good that don't require that assumption?

15.1.6 Implementation in Statistical Software

Ordinarily, mediation analysis (or most any analysis for that matter) would not be undertaken with such a small sample, although there is nothing in the mathematics or statistical theory that would prevent you from doing so. Here we do a more realistic illustration using the HOSPITAL data file first introduced in Chapter 13, while showing how to generate an inference for the indirect effect. The data set, fabricated for this illustration but motivated by Halbesleben (2010), contains the responses of 300 health care employees at a hospital. At time 1, they were asked questions to measure their physical and emotional exhaustion (X_1:`exhaust`). Also available is a baseline measure of workplace injury (C_1:`injuryb`) that is an index based on the number and severity of injuries the health care worker had experienced since starting employment. Some months later, these same workers were asked how frequently they engage in various work-arounds to safety protocols, so as to avoid the time and hassle these safety measures require (X_2:`safety`), and their workplace injury frequency and severity was again quantified (Y:`injury`).

The path model is depicted in Figure 15.3. This model estimates the effects of physical and emotional exhaustion on later workplace injuries, directly as well as indirectly, through the use of safety protocol work-arounds. That is, according to this process, physical and emotional exhaustion may prompt workplace injuries because people who are exhausted are more likely to avoid the use of safety protocols, which in turn translates into a greater likelihood of injury. This is the indirect effect of exhaustion depicted in Figure 15.3. But exhaustion may influence workplace injuries through some other process not a part of this model (the direct effect). We use baseline workplace injuries as a covariate. Thus, the effect of exhaustion on later injuries and use of safety protocol work-arounds is assessed independent of how frequently the worker tended to get injured before the study started and measurements on X_1, X_2, and Y were obtained.

The total, direct, and indirect effects can be estimated using any statistics program capable of conducting regression analysis. We assume at this point

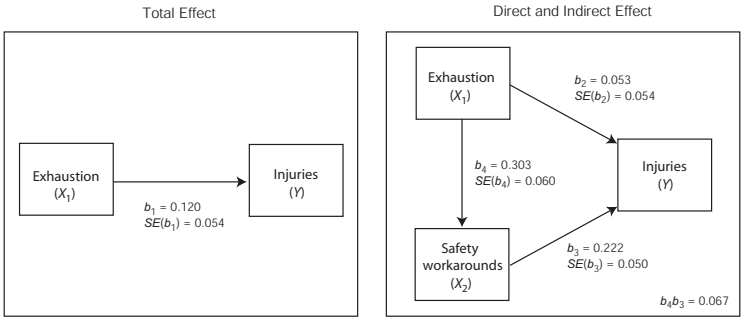


FIGURE 15.3. Path diagrams depicting the total, direct, and indirect effects of physical and emotional exhaustion on workplace injuries, with the indirect effect operating through the use of safety protocol work-arounds. Baseline workplace injury frequency, used as a covariate, is not depicted in the diagram.

in the book that you could estimate the corresponding equations of X_2 and Y using your chosen software. Doing so yields the following regression equations for the use of safety protocol work-arounds and later workplace injuries (see corresponding STATA output in Figure 15.4):

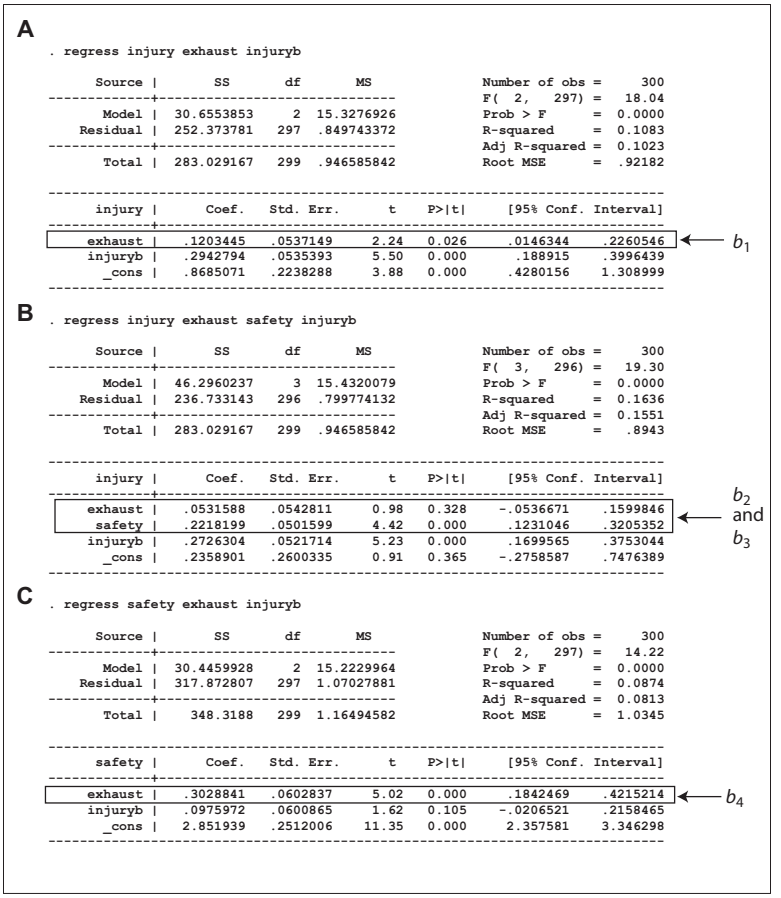
$$\hat{Y} = 0.869 + 0.120X_1 + 0.294C_1$$

$$\hat{Y} = 0.236 + 0.053X_1 + 0.222X_2 + 0.273C_1$$

$$\hat{X}_2 = 2.852 + 0.303X_1 + 0.098C_1$$

The relevant regression coefficients are superimposed on Figure 15.3. From these equations (also see Figure 15.4), they are $b_1 = 0.120$, $b_2 = 0.053$, $b_3 = 0.222$, and $b_4 = 0.303$. The regression analysis output will also include standard errors for the total and direct effects of exhaustion, along with t- and p-values for testing the null hypothesis that the effect equals zero. Confidence intervals can also be used for interval estimation.

The total, direct, and indirect effects of exhaustion on workplace injury are $b_1 = 0.120$, $b_2 = 0.053$, and $b_4b_3 = 0.303(0.222) = 0.067$, respectively. The total effect, with an estimated standard error of 0.054, is statistically significant, $t(297) = 2.240, p = .026$. Two hospital workers who are equal in initial workplace injuries but who differ by 1 unit in exhaustion are estimated to differ by 0.120 units in later workplace injuries, with the more exhausted person experiencing more injury. The direct effect is also positive, $b_2 = 0.053$, but with a standard error of 0.054, it is not statistically significant, $t(296) = 0.979, p = .328$.



A

```
. regress injury exhaust injuryb

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  2,   297) =   18.04
       Model |  30.6553853     2  15.3276926            Prob > F      =  0.0000
    Residual |  252.373781   297  .849743372           R-squared     =  0.1083
-------------+------------------------------           Adj R-squared =  0.1023
       Total |  283.029167   299  .946585842           Root MSE      =  .92182

------------------------------------------------------------------------------
      injury |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exhaust |   .1203445   .0537149     2.24   0.026     .0146344    .2260546   <-- b1
     injuryb |   .2942794   .0535393     5.50   0.000      .188915    .3996439
       _cons |   .8685071   .2238288     3.88   0.000     .4280156    1.308999
------------------------------------------------------------------------------
```

B

```
. regress injury exhaust safety injuryb

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  3,   296) =   19.30
       Model |  46.2960237     3  15.4320079            Prob > F      =  0.0000
    Residual |  236.733143   296  .799774132           R-squared     =  0.1636
-------------+------------------------------           Adj R-squared =  0.1551
       Total |  283.029167   299  .946585842           Root MSE      =   .8943

------------------------------------------------------------------------------
      injury |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exhaust |   .0531588   .0542811     0.98   0.328    -.0536671    .1599846   <-- b2
      safety |   .2218199   .0501599     4.42   0.000     .1231046    .3205352   <-- and b3
     injuryb |   .2726304   .0521714     5.23   0.000     .1699565    .3753044
       _cons |   .2358901   .2600335     0.91   0.365    -.2758587    .7476389
------------------------------------------------------------------------------
```

C

```
. regress safety exhaust injuryb

      Source |       SS       df       MS              Number of obs =     300
-------------+------------------------------           F(  2,   297) =   14.22
       Model |  30.4459928     2  15.2229964            Prob > F      =  0.0000
    Residual |  317.872807   297  1.07027881           R-squared     =  0.0874
-------------+------------------------------           Adj R-squared =  0.0813
       Total |   348.3188   299  1.16494582           Root MSE      =  1.0345

------------------------------------------------------------------------------
      safety |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exhaust |   .3028841   .0602837     5.02   0.000     .1842469    .4215214   <-- b4
     injuryb |   .0975972   .0600865     1.62   0.105    -.0206521    .2158465
       _cons |   2.851939   .2512006    11.35   0.000     2.357581    3.346298
------------------------------------------------------------------------------
```

FIGURE 15.4. STATA output from a regression-based path analysis of mediation of the effect of exhaustion on workplace injury through the use of safety protocol work-arounds.

```
************************ PROCESS Procedure for SPSS ************************

         Written by Andrew F. Hayes, Ph.D.       www.afhayes.com
    Documentation available in Hayes (2013). www.guilford.com/p/hayes3

**************************************************************************
Model = 4
    Y = injury
    X = exhaust
    M = safety

Statistical Controls:
CONTROL= injuryb

Sample size
        300

**************************************************************************
Outcome: safety

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2956      .0874     1.0703    14.2234     2.0000   297.0000      .0000

Model
             coeff         se          t          p       LLCI       ULCI
constant    2.8519      .2512    11.3532      .0000    2.3576    3.3463
exhaust      .3029      .0603     5.0243      .0000     .1842     .4215   <-- b4
injuryb      .0976      .0601     1.6243      .1054    -.0207     .2158

**************************************************************************
Outcome: injury

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .4044      .1636      .7998    19.2955     3.0000   296.0000      .0000

Model
             coeff         se          t          p       LLCI       ULCI
constant     .2359      .2600      .9072      .3651    -.2759     .7476
safety       .2218      .0502     4.4223      .0000     .1231     .3205   <-- b3 and
exhaust      .0532      .0543      .9793      .3282    -.0537     .1600   <-- b2
injuryb      .2726      .0522     5.2257      .0000     .1700     .3753

************************* TOTAL EFFECT MODEL ****************************
Outcome: injury

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .3291      .1083      .8497    18.0380     2.0000   297.0000      .0000

Model
             coeff         se          t          p       LLCI       ULCI
constant     .8685      .2238     3.8802      .0001     .4280     1.3090
exhaust      .1203      .0537     2.2404      .0258     .0146     .2261   <-- b1
injuryb      .2943      .0535     5.4965      .0000     .1889     .3996

***************** TOTAL, DIRECT, AND INDIRECT EFFECTS ********************

Total effect of X on Y
     Effect         SE          t          p       LLCI       ULCI
      .1203      .0537     2.2404      .0258     .0146     .2261   <-- b1

Direct effect of X on Y
     Effect         SE          t          p       LLCI       ULCI
      .0532      .0543      .9793      .3282    -.0537     .1600   <-- b2

Indirect effect of X on Y
           Effect    Boot SE   BootLLCI   BootULCI
safety      .0672      .0194      .0335      .1104   <-- b4b3

**************************************************************************
```

FIGURE 15.5. PROCESS for SPSS output from a simple mediation analysis estimating the direct and indirect effects of exhaustion on workplace injury, with the a 95% bootstrap confidence interval for the indirect effect through use of safety protocol work-arounds.

Most regression programs will not generate the indirect effect, although it can easily be constructed by multiplying b_4 and b_3 by hand. In this example, the indirect effect 0.067. This is due to the positive effect of exhaustion on the use of safety protocol work-arounds ($b_4 = 0.303$), which in turn is positively related to later injury ($b_3 = 0.222$). The result is a difference of 0.067 units in later workplace injury due to exhaustion indirectly through the use of safety protocol work-arounds. Notice that as promised, the indirect effect of exhaustion of 0.067 when added to the direct effect of exhaustion of 0.053 yields the total effect of exhaustion: $0.067 + 0.053 = 0.120$.

We have not yet discussed inference about the indirect effect. The methods discussed in section 15.1.5 are generally not implemented in regression programs, but they are available as special tools or freely available macros for SPSS, SAS, or R. For instance, the SOBEL or INDIRECT macros for SPSS and SAS (Preacher & Hayes, 2004, 2008) conduct simple mediation analysis and provide various inferential tests for the indirect effect that can be used for inference. Hayes (2013) provides code for construction of a Monte Carlo confidence interval in SPSS and SAS, and we provide code for STATA below. The Rmediate (Tofighi & MacKinnon, 2011) and MBESS (Kelley, 2007) packages are available for R users. Mediation analysis can also be programmed into a structural equation modeling program such as Mplus or AMOS, both of which have features for estimation of and inference about indirect effects, including bootstrap confidence intervals.

The PROCESS macro, freely available for SPSS and SAS and described in Hayes (2013), is a simple, widely used tool for mediation analysis, so we illustrate its use here for construction of an interval estimate of the indirect effect. PROCESS has various path analysis features that estimate all the regression coefficients in a mediation model and provide inferential tests for the direct, indirect, and total effects, including the Sobel test and a bootstrap or Monte Carlo confidence interval. PROCESS does not come with SPSS or SAS. It must be downloaded from *www.processmacro.org* and executed before SPSS or SAS will understand a PROCESS command. The features in PROCESS are documented in Hayes (2013), which also provides a more detailed discussion of mediation analysis than we provide in this book.

The output for this analysis from the SPSS version of PROCESS can be found in Figure 15.5. This output was generated with the command

```
process vars=exhaust safety injury injuryb/y=injury/x=exhaust/m=safety
   /model=4/total=1/boot=5000.
```

The equivalent command in the SAS version of PROCESS is

```
%process (data=hospital,vars=exhaust safety injury injuryb,y=injury,
    x=exhaust,m=safety,model=4,total=1,boot=5000);
```

As can be seen in Figure 15.5, the output contains all the regression coefficients, standard errors, and t- and p-values for each path in the causal system, as well as the covariates. At the very bottom of the output can be found the indirect effect, which is listed as 0.067 and is the same as we calculated earlier. The lower and upper bounds of a 95% bootstrap confidence interval based on 5,000 bootstrap samples are listed under "BootLLCI" and "BootULCI." The confidence interval is 0.034 to 0.110. As this is entirely above zero, we can conclude with 95% confidence that the indirect effect is positive. Statistically, this is evidence of mediation of the effect of exhaustion on workplace injury through the use of safety protocol workarounds. But as discussed in section 15.3.1, although this is consistent with mediation—a causal process—there is more to establishing cause–effect than just data analysis.

PROCESS is not available for STATA, but all the effects can be estimated using ordinary least squares regression in STATA with the **regress** command illustrated numerous times elsewhere in this book. Special STATA programming skills are required to generate a bootstrap confidence interval for the indirect effect. But a Monte Carlo confidence interval is fairly simple to generate in STATA once you have b_3 and b_4 and their standard errors from the output from the **regress** commands. The code below generates a Monte Carlo confidence interval for the indirect effect using 5,000 samples. In this code, b_4 and $SE(b_4)$ are in the second line of code and b_3 and $SE(b_3)$ are in the third line of code. The resulting output can be found in Figure 15.6, showing a 95% confidence interval for the indirect effect of 0.032 to 0.111. This is very similar to the confidence interval generated by bootstrapping and leads to the same conclusion about the indirect effect.

```
set obs 5000
gen b4 = (invnorm(uniform())*0.060)+0.303
gen b3 = (invnorm(uniform())*0.050)+0.222
gen b4b3 = b4*b3
centile b4b3, centile (2.5 97.5)
```

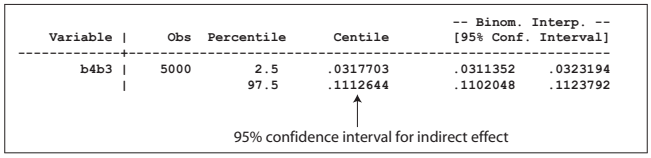


FIGURE 15.6. STATA output for a 95% Monte Carlo confidence interval for the indirect effect.

15.2 Multiple Mediator Models

The path analyses in sections 15.1.2 and 15.1.6 contained only one mediator. But more complex path models are possible that allow an independent variable to exert its effect on a dependent variable through more than one indirect pathway. For instance, in a *parallel multiple mediator model*, we have more than one mediator between independent and dependent variable, but those mediators are not connected to each other causally. Figure 15.7, panel B, depicts a parallel multiple mediator model with two mediators. In such a model, although the mediators might be correlated, no commitment is made that one causes the other. In the example depicted in Figure 15.7, panel B, exercise is modeled as affecting weight loss through three pathways. One pathway operates indirectly through food intake. A second pathway operates indirectly through metabolic rate. And the final pathway is direct, bypassing both food intake and metabolic rate.

But you may have some basis for believing that metabolic rate would influence food intake. In that case, you might prefer a *serial multiple mediator model*, such as in Figure 15.7, panel C. Now we have four pathways of influence, three indirect and one direct. One indirect effect operates only through food intake, one operates only only through metabolic rate, and one operates through metabolic rate and food intake in sequence or *serially*. The final pathway is the direct effect, bypassing both food intake and metabolic rate.

In both the parallel and the serial multiple mediator models, the total effect of the independent variable can be partitioned into direct and indirect components. These indirect and direct effects can be estimated using regression analysis.

15.2.1 Path Analysis for a Parallel Multiple Mediator Model

The total effect in a mediation model is not determined by how many mediators are placed in between the independent variable and the dependent

variable. If X_1 is the independent variable and Y is the dependent variable, the total effect of X_1 on Y is b_1 in a regression model estimating Y from X_1, as in equation 15.1. Covariates can be included if desired, as discussed in section 15.1.3. Using the EXERCISE data file, we know from the earlier analysis that the total effect of exercise frequency on weight loss is $b_1 = 1.75$ and is statistically different from zero. See Figure 15.7, panel A.

In a parallel multiple mediator model with k mediators, we can estimate the direct and indirect effects by regressing the dependent variable on the independent variable and all the mediators in one regression and then each of the mediators on the independent variable in k separate regressions with X_1 as the sole regressor. Covariates can be included in each of the equations, as discussed in section 15.1.3. In the example in Figure 15.7, panel B, that means estimating weight loss (Y) from exercise frequency (X_1), food intake (X_2), and metabolism (X_3):

$$\hat{Y} = b_0 + b_2X_1 + b_3X_2 + b_4X_3 \tag{15.9}$$

The regression coefficients using the EXERCISE data file are found in the path diagram in Figure 15.7, panel B. The direct effect of exercise frequency on weight loss is $b_2 = 1.046, SE = 0.422, t(6) = 2.476, p = .048$. This is statistically different from zero. Notice that this direct effect is different than the direct effect in section 15.1.2, because that simpler model did not include metabolism as a mediator, so it wasn't being statistically controlled when assessing exercise's direct effect on weight loss.

From the model of the dependent variable we also get the partial effects of the two mediators. In this example, $b_3 = -1.136$ and $b_4 = 0.634$. So holding exercise and metabolism constant, an additional 1 unit of food intake is related to a reduction in weight loss of 1.136 units (113.6 grams per week). But a 1-unit increase in metabolism, holding food intake and exercise constant, is associated with a 0.634-unit (63.4 grams per week) increase in weight loss.

The indirect effects of exercise frequency on weight loss require two regression coefficients each, one coming from the model of the dependent variable in equation 15.9, and the other coming from a regression estimating the mediator from the independent variable (and any covariates included in the model of the total effect). The models of the mediators from the independent variable are

$$\hat{X}_2 = b_0 + b_5X_1 \tag{15.10}$$

$$\hat{X}_3 = b_0 + b_6X_1 \tag{15.11}$$

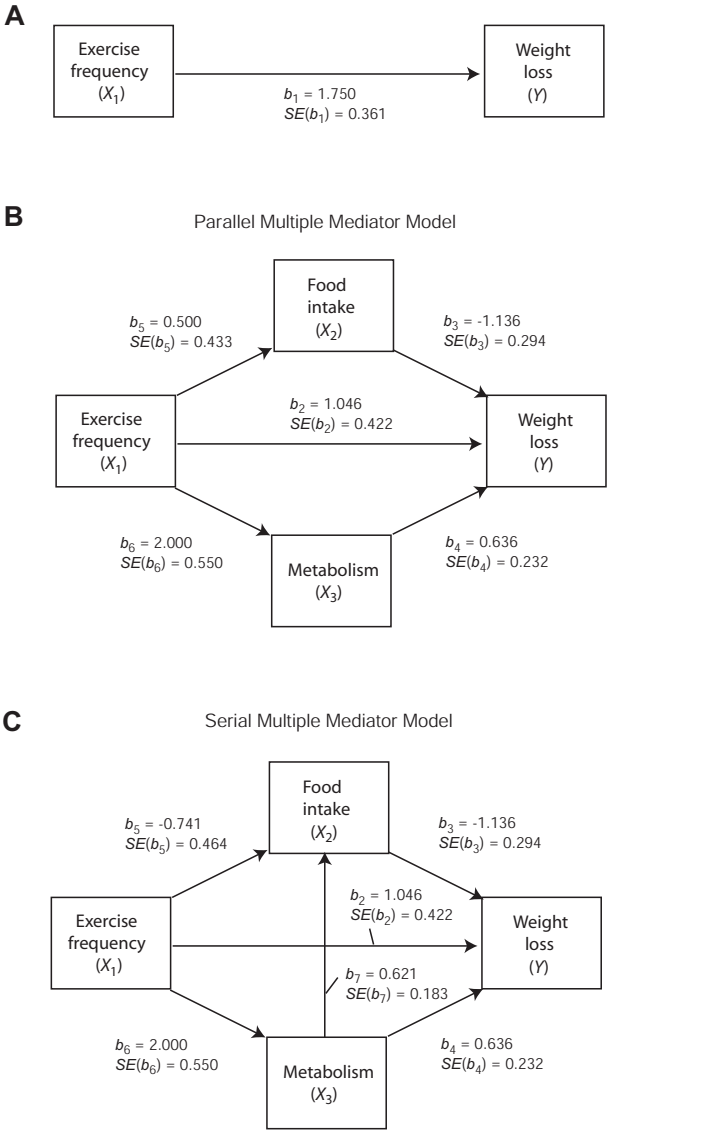


FIGURE 15.7. A parallel (panel B) and serial (panel C) multiple mediator model of the effects of exercise frequency on weight loss. The total effect is depicted in panel A.

In this example, $b_5 = 0.500$ and $b_6 = 2.000$, which are superimposed in the path diagram in Figure 15.7, panel B. So an additional hour of exercise translates into 0.500 units more food intake and 2.000 units higher metabolism. These estimates, combined with the estimates of the effects of the mediators on the dependent variable, give the indirect effects.

The indirect effect of X_1 on Y through X_2 is the product of the effect of X_1 on X_2 (b_5 from equation 15.10) and the effect of X_2 on Y, holding all else constant (b_3 from equation 15.9). That is, $b_5 b_3 = (0.500)(-1.136) = -0.568$. So 1 hour of additional exercise per week seems to reduce weight loss by 0.568 units (56.8 grams per week) indirectly through its effect on increasing food intake, which in turn lowers weight loss. The indirect effect of X_1 on Y through X_3 is calculated similarly as the product of the effect of X_1 on X_3 (b_6 from equation 15.11) and the effect of X_3 on Y, holding all else constant (b_4 from equation 15.9). In this case, $b_6 b_4 = (2.000)(0.636) = 1.273$. So an hour of additional exercise is related to an increase in weight loss by 1.273 units (127 grams per week) by increasing metabolism rate, which in turn is associated with greater weight loss.

Observe that, as promised, the total effect b_1 is the sum of the direct effect of X_1 and the two indirect effects of X_1 through X_2 and X_3. That is, $b_1 = b_2 + b_5 b_3 + b_6 b_4 = 1.046 + (-0.568) + 1.273 = 1.750$. So the 175 gram per week weight loss due to an additional 1 hour of exercise is due to its positive effects on increasing metabolism (the indirect effect through metabolism of 1.273), as well as some other process not a part of the model (the direct effect of 1.046). But some of that weight loss from exercise is counteracted by increased food consumption, which lowers weight loss (the indirect effect through food intake of −0.568).

We have already discussed inference about the direct and total effects. Inference for the indirect effect can be undertaken using any of the methods discussed in section 15.1.5. Using the PROCESS macro for SPSS and SAS, we generated 95% bootstrap confidence intervals for these indirect effects using 5,000 bootstrap samples and found evidence of mediation by metabolism (0.252 to 2.223) but not by food intake (−1.564 to 0.212). That is, we can confidently conclude that the indirect effect through metabolism is positive, but we can't say definitively that the indirect effect through food intake is different from zero.

15.2.2 Path Analysis for a Serial Multiple Mediator Model

The serial multiple mediator model differs from the parallel multiple mediator model only with the inclusion of a causal path between mediators.

In a serial multiple mediator model with only two mediators, as in Figure 15.7, panel C, the only difference relative to the parallel multiple mediator model is the inclusion of an effect of one mediator on the other. In this case, this model allows for a causal effect of X_3 on X_2. So the three equations required to estimate the direct and indirect effects are

$$\hat{Y} = b_0 + b_2X_1 + b_3X_2 + b_4X_3 \tag{15.12}$$

$$\hat{X}_2 = b_0 + b_5X_1 + b_7X_3 \tag{15.13}$$

$$\hat{X}_3 = b_0 + b_6X_1 \tag{15.14}$$

Notice that the equations for Y and X_3 are the same in this serial multiple mediator model relative to the parallel multiple mediator model. Because the direct effect of X_1 comes from the model of Y, it is unchanged by connecting mediators into a causal chain, as can be seen in Figure 15.7, panel C, which includes all of the regression coefficients for the model derived from equations 15.12, 15.13, and 15.14. And of course the total effect is not changed by configuring the causal connections between mediators, because the total effect is calculated without regard to what mediators are in the model or how they are interconnected.

The indirect effects of exercise frequency on weight loss, of which there are three now, are computed by multiplying the constituent components of each path in a chain linking X_1 to Y. Starting first with the indirect effect through metabolism only, this indirect effect is $b_6b_4 = 2.000(0.636) = 1.273$. Notice that this is the same as the corresponding indirect effect in the parallel multiple mediator model, because the equations that yield b_4 and b_6 are the same in the two models.

The indirect effect through food intake only is different from the parallel multiple mediator model, because in this serial mediation model the effect of exercise on food intake is estimated while controlling for metabolism. That effect is $b_5 = -0.741$, which is very different from the parallel multiple mediator model. When metabolism is held constant, more frequent exercise is associated with less food intake, not more, as in the parallel multiple mediator model. When this negative effect of −0.741 is multiplied by the negative partial effect of food intake on weight loss, $b_3 = -1.136$, the resulting indirect effect is $b_5b_3 = -0.741(-1.136) = 0.843$. So when the effect of metabolism on food intake is held constant, the indirect effect of exercise on weight loss through food intake is positive rather than negative, as it was in the parallel multiple mediator model.

The remaining indirect effect is the serial indirect effect that passes first to metabolism, then to food intake, and then to weight loss. This indirect effect is the product of the three constituent paths, $b_6b_7b_3 = 2.000(0.621)(-1.136) = -1.411$. Through this mechanism, exercise decreases weight loss in part by increasing metabolism, which then increases food intake, which results in less weight loss.

As in the simple and parallel multiple mediator models, the total effect is the sum of the direct and indirect effects. In this model, $b_1 = b_2 + b_6b_4 + b_5b_3 + b_6b_7b_3 = 1.046 + 1.273 + (-1.411) + 0.843 = 1.750$.

Inference about the indirect effect can be based on any of the methods already discussed. The PROCESS macro for SPSS and SAS can estimate this model and generate bootstrap confidence intervals for inference about indirect effects. Using PROCESS, 95% bootstrap confidence intervals (using 5,000 bootstrap samples) through metabolism only, food intake only, and both in serial, were 0.252 to 2.223, −0.058 to 1.628, and −2.540 to −0.315, respectively. So we can definitively claim that the indirect effect through metabolism alone is positive, and the indirect effect through metabolism and food intake in serial is negative. But we cannot definitively claim that the indirect effect through food intake alone is different from zero.

15.3 Extensions, Complications, and Miscellaneous Issues

15.3.1 Causality and Causal Order

Mediation is by definition a causal process, so it is impossible to talk about mediation in noncausal terms. Yet often our data collection efforts do not generate data or results that allow us to make unequivocal causal claims for at least some parts of a mediation system. We discussed some of the relative advantages and disadvantages of experimentation through random assignment compared to statistical control in Chapter 6. Random assignment is a nice design feature to have in a study, but often it isn't possible. Even when it is, there are limits to the claims one can make from randomized experiments. Statistical control can be used as a substitute for experimental control of covariates, but one never knows if one has controlled for the right covariates. This applies to mediation analysis as well. And even if one has a sense for what things should be controlled and what should not, the direction of causal order cannot always be established in nonexperimental studies.

In path analysis and mediation, things get even more complicated, because mediation involves a sequence of at least two causal relationships. Suppose you have randomly assigned participants in a study to levels of the independent variable, but the mediator and dependent variable are only measured. In the presence of good experimental design, random assignment to values of the independent variable allows for a causal inference about the effect of the independent variable on the mediator and the dependent variable, as well as the direction of causal order for those effects. But this does not allow you to conclude that the mediator affects the dependent variable. It could be that the dependent variable actually affects the mediator, meaning the dependent variable is actually the mediator, and your proposed mediator is actually the effect rather than intermediate between cause and effect. Although it is tempting to reconduct the analysis, flipping the role of the mediator and the dependent variable to see what happens, you will often find evidence of an indirect effect (and thus mediation) in the reconfigured model. Furthermore, there is no way to determine which configuration is correct in any absolute sense.

When the independent variable is not randomly assigned, you lose the ability to make unequivocal claims about the direction of causal order between the independent variable and the mediator and between the independent variable and the dependent variable as well. Now direction of causality and causal order can only be established through argument, logic, and theory, or the combination of these. Of course, this is true for any nonexperimental study, so this isn't a disadvantage of mediation analysis specifically.

Some argue that mediation analysis is inappropriate in such a design, and some go to such extremes as calling it "futile" (Maxwell, Cole, & Melissa, 2011). But we don't believe that mediation analysis is any more or less appropriate with nonexperimental data than it is with experimental data. The kinds of conclusions we can reach with any statistical analysis are always constrained in one way or another by the design of the study and the manner of data collection. Mediation analysis can be a useful approach to describing relationships and testing hypotheses, but connecting variables together in a theoretical causal system and then constructing measures of direct and indirect effects does not mean you can interpret the relationships in causal terms. The same kinds of design and interpretation considerations that complicate the interpretation of studies without random assignment discussed in Chapter 6 apply to mediation analysis. But so too do the limitations of random assignment.

In short, questions about cause and causal order require difficult judgments, and we may have seemed rather cavalier and loose in our use of causal language in the examples in this chapter. In practice, you may agonize for days over choices about how to properly order variables in your research in a sensible causal sequence. We do not intend to imply in our examples that the relationships are necessarily causal, or that it is easy or even possible to answer questions about cause–effect using mediation analysis. But mediation analysis can be used to estimate direct and indirect effects in a purely mathematical sense. Whether these conclusions really can be causal ones is not a judgment that statistical analysis can make for you.

15.3.2 The Causal Steps Approach

In sections 15.1 and 15.2 we described how to calculate an indirect effect and conduct an inference about whether the indirect effect is different from zero. If an indirect effect is statistically different from zero, then this supports a claim of mediation of the presumed causal variable on the proposed effect variable by the putative mediator. This approach is consistent with most modern perspectives on mediation analysis, but it differs from an approach that is now outdated but remains very popular and that you will still see being used, so it is worth discussing in brief.

Baron and Kenny (1986) popularized an approach to mediation analysis that never involves the computation of an indirect effect, nor does it involve any kind of statistical inference about the indirect effect. Their approach is sometimes called the *causal steps* approach to mediation analysis. It relies on a set of hypothesis tests about each path in the causal system and the pattern of statistical significance or lack of significance for the total and direct effects. Mediation is established by the causal steps or *Baron and Kenny* method only if the total effect of a variable is statistically significant and the paths that define an indirect effect (i.e., the effect of the independent variable on the mediator and the effect of the mediator on the dependent variable) are all different from zero by a hypothesis test or confidence interval. The effect of the independent variable on the dependent variable is said to be mediated *completely* by the mediator if all these conditions are met, and the direct effect of the independent variable on the dependent variable is not statistically significant. If all these conditions are met but the direct effect of the independent variable is statistically significant, then the effect is said to be *partially* mediated.

We mention this approach because it is still widely used, but we consider it to have historical relevance only. Most researchers who write and publish about mediation analysis don't recommend the use of the causal steps approach and instead advocate quantifying the indirect effect and conducting an inference about it rather than its constituent components. The arguments against the causal steps approach include that it is relatively lower in power, is inconsistent with the way science proceeds by quantifying effects of interest and conducting an inference about those quantities, and it relies on more inferential tests than is needed to test a mediation hypothesis. For a discussion of these arguments against the causal steps approach, as well as a skeptical view of the concepts of complete and partial mediation, see Hayes (2013).

15.3.3 Mediation of a Nonsignificant Total Effect

Mediation analysis can help to answer the question as to how an effect operates. Implied in this question is that there is an effect operating. That is, in practice, researchers often ask questions about mediation only when it has been established that there is some effect to be mediated. In statistical terms, this means that mediation analysis is only sensible if one has evidence of a total effect of the independent variable on the dependent variable. Absent evidence of such an effect (by a statistical significance or confidence interval standard), there is no effect to be mediated, and thus no point in conducting a mediation analysis.

But we disagree with this perspective and recommend not requiring evidence of a total effect of the independent variable on the dependent variable before estimating and testing indirect effects. The independent variable can causally affect the dependent variable even if they are not correlated. As Bollen (1989, p. 52) aptly states, "a lack of correlation does not disprove causation." Recall that the total effect of an independent variable is the sum of the direct and indirect effect(s) of that independent variable. There is no mathematical requirement that direct and indirect effects be of the same sign. If they differ in sign, then they may add to something very small, even zero, and not statistically significant by an inferential test. Yet the indirect effect may be statistically different from zero and perhaps even quite large.

As a concrete example, consider a study by Cole, Walter, and Bruch (2008). They found no statistically significant relationship between the extent to which teams at an automobile parts manufacturing facility engaged in dysfunctional behavior and the teams' performance as measured by

supervisor perceptions. Yet a mediation analysis was consistent with dysfunctional team behavior negatively affecting team performance indirectly through its effect on the negativity of the work climate (which translated to lower performance) as well as directly, with more dysfunctional team behavior *positively* affecting team performance when the negativity of the work climate was held constant. Both the indirect and direct effects of dysfunctional team behavior on performance were statistically significant, even though the total effect was not.

In complex models with more than one mediator between the independent and dependent variable, indirect effects through different mediators may be different in sign, and if the direct effect is weak, the result may be a total effect near zero and not statistically significant. For example, Pitts and Safer (2016) found no statistically significant relationship between combat experience and depression in a sample of U.S. Army medics. Yet combat experience had a positive indirect effect on depression through how threatened the medics felt during those experiences, and a negative indirect effect through the positivity of their view of their combat experience. Medics with more combat experience perceived greater threat during those experiences, which was positively related to depression (and hence a positive indirect effect). But they also had a more positive view of the combat experience, which was negatively related to depression (and hence a negative indirect effect). There was no direct effect of combat experience on depression, and when added to the two opposing indirect effects, the result was a small total effect that was not statistically significant.

The point is that in any causal systems, the sum of the direct and indirect effects of an independent variable—the total effect of that independent variable—is an aggregation of multiple pathways of influence that may work in opposing directions. Ignoring this and conditioning the hunt for an indirect effect on evidence of a statistically significant total effect means you will probably miss some interesting and perhaps even surprising, exciting, and theoretically important results.

15.3.4 Multicategorical Independent Variables

We have seen that the indirect effect of an independent variable on a dependent variable through a mediator can be calculated as a product of two regression coefficients, one quantifying the effect of the independent variable on the mediator, and the other quantifying the effect of the mediator on the dependent variable. But when the independent variable is multicategorical, there is no single regression coefficient quantifying the

effect of the independent variable on a mediator or the dependent variable. Recall from Chapters 9 and 10 that it takes $g - 1$ regression coefficients to represent a multicategorical variable's effect when that variable consists of g categories. The interpretation of these regression coefficients will depend on how the multicategorical variable is represented by the coding system used.

When the independent variable in a mediation analysis is a multicategorical variable, there is no single indirect, direct, or total effect of that variable but, rather, $g - 1$ *relative* indirect, direct, and total effects, terms introduced by Hayes and Preacher (2014). They discuss how to test for mediation when the independent variable is multicategorical. The logic is similar to what we described in this chapter and involves the product of regression coefficients relating the independent variable to the mediator and the mediator to the outcome. We refer interested readers to their article for a discussion of the mechanics of mediation analysis in this situation.

15.3.5 Fixing Direct Effects to Zero

If you are interested in mediation and have no basis for believing that an independent variable affects a dependent variable through any process other than the one through the proposed mediator, should you just fix the direct effect to zero? This would be accomplished by leaving the independent variable out of the equation for the dependent variable that includes the mediator or mediators (e.g., equations 15.2, 15.9, or 15.12).

For two reasons, we don't believe this is a good idea. First, remember that the total effect of the independent variable on the dependent variable is the sum of the direct and indirect effects. By fixing the direct effect of the independent variable to zero, you are forcing the indirect effect to equal the total effect. If the direct effect is actually equal to zero, then this is not a problem. But if you are wrong, then the indirect effect will be a biased estimator of the true indirect effect. It is better to let the data derive the direct effect rather than to fix it. Although the temptation to delete the direct effect if it is not statistically significant may be strong, resist the temptation to do so. A null hypothesis can never be proven true. Including a nonsignificant direct effect in a model is no different than leaving a nonsignificant covariate in the model, and we generally don't recommend removing covariates from a regression model just because they are not statistically significant (see section 4.7.3 and 17.1.3 for our discussion of nonsignificant or unnecessary covariates).

The second reason we discourage fixing a direct effect to zero relies on the fact that one interpretation of a direct effect is that it is the effect of the independent variable on the dependent variable that operates through some mediator that is not in the model. Most effects operate through multiple mechanisms simultaneously, and probably through mediators that are not in your model. Including the direct effect, even when it is not theorized or statistically significant, results in a more realistic model that doesn't force these unmodeled mechanisms to manifest themselves entirely in the other indirect effects you are estimating.

15.3.6 Nonlinear Effects

The path analysis algebra discussed in sections 15.1 and 15.2 is predicated on the assumption of linearity in the relationships in the causal system being modeled. We saw in Chapter 12 how regression analysis can be used to model curvilinear relationships by, for example, including the square of a regressor in a model. The estimation of direct and indirect effects can be more complicated than as presented here when relationships are modeled as curvilinear. An understanding of the algebra involved requires a little background in calculus, so we refer the interested reader to Hayes and Preacher (2010), who discuss mediation analysis involving nonlinear relationships. Their discussion is restricted to simple mediation models, though, in principle, the methods they discuss could be generalized to more complex models with more than one mediator.

15.3.7 Moderated Mediation

In Chapters 13 and 14 we showed how a linear regression analysis can be used to test for interaction between two regressors. Two variables X_1 and X_2 interact when X_1's effect on Y depends on X_2. By including the product of X_1 and X_2 as a regressor along with X_1 and X_2, X_1's effect on Y becomes a linear function of X_2.

Linear interaction can be combined with mediation analysis to yield a model of the effect of an independent variable on a dependent variable through one or more mediators that depend on a moderator variable. If the indirect effect of an independent variable depends on a moderator, then it is said that the mediation is moderated, called *moderated mediation*. For example, in the dysfunctional team behavior study mentioned in section 15.3.3, Cole et al. (2008) found that the indirect effect of dysfunctional team behavior on team performance through the negative tone of the work cli-

mate was larger among teams that were more expressive of their emotions. In other words, this mechanism responsible for the negative indirect effect was moderated by the emotional expressiveness of members of the team.

In this example, team expressiveness was a continuum, but the moderator could also be dichotomous, such as a person's biological sex, or something experimentally manipulated. For example, Witkiewitz and Bowen (2010) studied the effect of depression on substance use indirectly through the effect of increased craving, which in turn was positively related to later substance use. But this indirect effect did not exist among people who were given a kind of therapy that relied on techniques of meditation and mindfulness.

Hayes (2013) coined the term *conditional process analysis* to refer to a data-analytic strategy that focuses on quantifying and testing the contingencies of mechanisms, as in these examples. There are many ways that mediation and moderation analysis can be analytically combined, depending on where in the causal system the moderation is happening. Conditional process analysis can be done with linear regression analysis, although special tools or statistical programs (e.g., the PROCESS macro mentioned in section 15.1.6) are needed to conduct inferences about indirect effects and to test whether an indirect effect is moderated. For a discussion of moderated mediation or conditional process analysis and the regression mathematics that underlies it, see sources such as Edwards and Lambert (2007), Fairchild and MacKinnon (2009), Hayes (2013, 2015), and Preacher, Rucker, and Hayes (2007).

15.4 Chapter Summary

A relationship between two variables X and Y, whether causally established through experimentation or merely assumed to be causal, can be partitioned into two broad pathways of influence using path analysis. The indirect effect of X on Y quantifies the amount that changing X changes Y through a sequence of causal steps in which X causally influences a mediator variable, which in turn causally influences Y. An independent variable's direct effect quantifies the amount that changing X changes Y when all the mediator or mediators in the model are held constant. The direct and indirect effects sum to give the total effect of X on Y. Inferential tests about direct effects are available in almost any regression output. Inference about indirect effects can be undertaken in a number of ways, including the Sobel test or a Monte Carlo or bootstrap confidence interval. The latter two meth-

ods usually require special code or tools and are not generally available in most regression analysis routines found in commercial software.

Although mediation is a causal process, cause cannot be definitively established merely through data analysis. So, although regression analysis can be used to determine whether evidence is consistent with a mediation process, it cannot determine whether the relationships are actually causal ones. Mediation analysis can be extended beyond the simpler models discussed in this chapter, including models with multicategorical independent variables and models that don't assume linearity in the relationships between variables in the system. Mediation analysis can also be combined with interaction to model mechanisms that are contingent or otherwise dependent on other variables, also known as conditional process analysis or the analysis of moderated mediation.

16
Detecting and Managing Irregularities

This chapter addresses the topic of *regression diagnostics*. Diagnostic statistics are useful for identifying cases in an analysis that are "irregular" in some way. We introduce *leverage*, *distance*, and *influence* as measures of irregularity and discuss how irregular cases may distort a regression analysis and so are worth identifying prior to interpretation of the results. We describe how diagnostic statistics can be used for testing whether the assumptions of linear regression analysis are met, introduce some ways of dealing with assumption violations, and discuss how violations may affect the validity of the inferences one makes using regression analysis.

All too many investigators discover clerical errors in their data, such as inputting a person's age or a response to a question on a survey incorrectly, only after they have already spent hours on their data analysis and have perhaps reached conclusions that are hard to erase from their minds. Or after publication, a critic may point out that the researcher's main conclusion depended entirely on one research participant who was very unusual and perhaps should not have even been included. Some statistical techniques designed for avoiding mishaps like these are the topic of this chapter. We discuss some methods of detecting cases that are somehow "irregular," which we define later in a number of ways. We talk about what to do when they are detected and methods you might consider employing if you are worried about the effects such irregularities may have on the quality of the inferences you report. We provide only a rough overview of these topics, which can be quite complicated. A more extensive treatment of some of the topics we discuss and others we don't can be found in Berry (1993), Fox (1991), and Kaufman (2013), among others.

16.1 Regression Diagnostics

In the evolution of regression analysis, diagnostic statistics are relatively new, having been developed mostly since 1975 or so. These statistics have several purposes. First, they can help to detect clerical errors, such as inputting a person's height as 720 inches rather than 72, which can seriously distort an entire analysis if not caught. Second, they can detect violations of the secondary assumptions of homoscedasticity and normality. Third, they can be used to examine data that are suspect for some reason, such as questionnaire results from someone who appeared not to understand directions, to determine whether those data are irregular in some way.

Diagnostic statistics can also be used to identify cases whose presence in the analysis are greatly influencing the results. For this reason, they can easily be misused. For instance, using some of the statistics and methods in this chapter, a clinical psychologist could find that three people in a study, if deleted from the analysis, could improve the apparent effectiveness of a therapeutic method he or she developed. This discovery could lead the psychologist to look at the files of these patients and find some rationale for excluding them. But any tool can be misused, and diagnostic statistics are an important part of regression analysis. The best protection against misuse is to require authors to explain in detail the reasons for deleting any cases and the ways in which those deletions affected the major conclusions. Although you can be faulted for your decision to exclude cases, you can't be accused of misconduct or unethical behavior if you are open about what you have done.

Diagnostic statistics may also occasionally detect violation of the primary assumption of linearity. But intuition suggests that they would not be nearly as powerful for that purpose as the methods discussed in Chapter 12, and our own analyses confirm that conjecture. For example, in a small-scale simulation study, we found that a test on the regression coefficient for X^2 to detect curvilinearity correctly detected real nonlinearity 98% of the time, while an approach we describe in section 16.2.4 detected nonlinearity only 33% of the time.

One of the best ways of detecting irregularities is to search for cases that are "extreme" in one sense or another. Such cases are often called *outliers*, though we confine that term to a particular type of extreme case.

16.1.1 Shortcomings of Eyeballing the Data

When computers were in their infancy, one of the major arguments given against their use in data analysis was that computer analysis made it easier to overlook extreme or unusual cases, even when they reflect obvious errors such as adult human weights of 16 or 1,600 pounds. Most likely, this is a clerical error of some kind that should be fixed before data collection. One simple way of catching extreme cases such as this is to scan the data file with your eyes, just looking for things that seem amiss. This is easy to do if the data file is small, but with large data sets with many variables, such "eyeballing" of the data may miss important irregularities.

Statistical computer programs quickly met the objection mentioned above by making it easy to identify the highest and lowest score on every variable, so that such extreme cases can be called to the investigator's attention. We recommend that prior to conducting an analysis, you ask your computer program to print the smallest and largest values of every variable in the data. Doing this would condense information about extreme cases for all the variables into one small output and make it easy to detect problems, such as someone whose weight is 1,600 pounds or who is −4.5 years old. Such values in the data are likely to show up as the minimum or maximum value for the variable. If you see something like this, fix it or otherwise investigate the source of the problem. Maybe you or your research assistant simply mistyped a weight when entering the data. Or if the data were collected by a computer program, maybe there is a bug in the program that generates incorrect data in certain circumstances.

Today's computer programs allow us to go far beyond this basic step. Using statistics discussed in this chapter we can detect irregularities that could never be discovered by eyeballing the data or looking at maximums and minimums. For instance, suppose your data file contains information about employees at a particular company, and the records for one employee in your data include the following information:

- Present salary: $30,000
- Hours worked per week: 20
- Starting salary: $20,000
- Hours worked per week on starting: 40
- Number of years worked: 2

This case is very unusual, though it isn't obvious how unless you think carefully about it. Notice that the employee earns the equivalent of a full time employee (40 hours per week) who makes $60,000 per year. But only 2 years ago, when the person was working full time, he or she was making only $20,000, so the employee's salary is three times what it was only 2 years ago. Most people don't get such large raises so quickly. Such an unusual case may represent a clerical error or some other factor worth checking. But ignoring any one of the five entries in this person's file would make the case appear normal. For instance, if "number of years worked" were not shown, we might assume it was 10 or 20 instead of 2, and the case would appear normal. A similar argument can be made about any of the other entries. Only when all five entries are considered together is the case identified as unusual. But if these five entries were scattered among 20 or 30 other entries about the same employee, it is highly unlikely that eyeballing of the data matrix would reveal anything amiss. Nor would this case likely be brought to our attention if we look only at the minimum and maximum values across all the employees on all five of these variables. Some regression diagnostic statistics can easily detect such cases.

16.1.2 Types of Extreme Cases

A case can be extreme or otherwise noteworthy in three major ways, all of which can be quantified. A case has high *leverage* if its pattern of regressor scores (ignoring Y) puts it far from most or all other cases. Speaking a bit loosely, cases with the highest *distance* are those whose vertical distance from the regression surface is greatest. *Influence* measures how much a case's presence in the analysis actually moves the regression surface. As we see later, there is a sense in which influence is the product of leverage and distance, so high influence requires both high leverage and high distance.

The distinctions among distance, leverage, and influence are illustrated most easily in simple regression. Consider the data set in Figure 16.1. Suppose the sample contains only the 37 cases represented with a solid square. If you regressed Y on X for only these 37 cases, the resulting model would be $\hat{Y} = 4.0 + 0.0X$. Now suppose you added case A to the data, denoted with a hollow square in the figure, bringing the sample size to 38. This case is extreme in the distribution of X. But if case A is included in the analysis, the regression model is unaffected; it is denoted with the thin, solid black line, and its equation is identical: $\hat{Y} = 4.0 + 0.0X$. This case is high in leverage, low in distance, and low in influence.

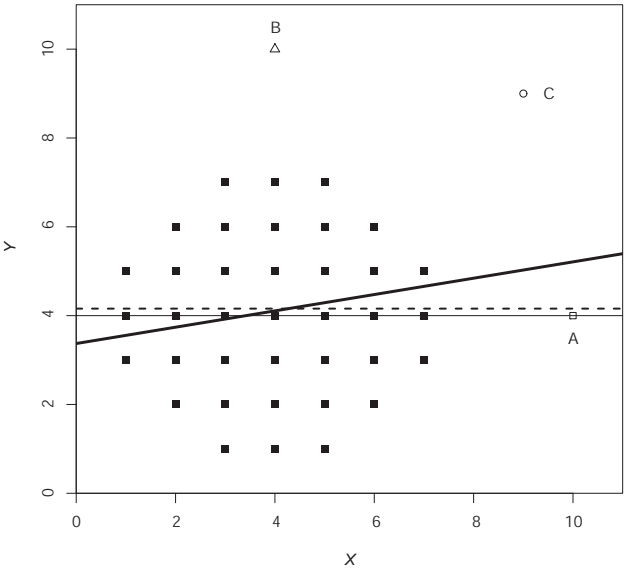


FIGURE 16.1. The influence of adding one case (represented with a hollow triangle, circle, or square) to a regression model containing 37 cases (represented with solid squares).

But now suppose you added only case B, denoted with the the hollow triangle. It is unusual on Y but quite ordinary on X. If it were included in the analysis, $\hat{Y} = 4.160 + 0.0X$, depicted with the dashed line. The regression constant has changed slightly, but the regression coefficient for X has not changed at all. This case is low in leverage, high in distance, and low in influence.

Finally, suppose you added only case C, denoted with the hollow circle. When it is included in the analysis, $\hat{Y} = 3.372 + 0.184X$, represented in the figure with the solid dashed line. Case C is high in leverage, high in distance, and high in influence. If leverage is potential to influence, then case C has realized that potential, whereas case A has not.

Users of regression analysis often focus on residuals when looking for extreme or influential cases, paying close attention to cases with large residuals (i.e., large distance). But this example shows that residuals are not necessarily the best way to identify influential cases, because cases that influence a regression analysis can "hide" by shrinking their own residual. Notice that case C pulls the regression line toward it, cutting its residual by

about 20%. Better would be some kind of a statistic that quantifies a case's residual relative to what it would be if the case weren't in the analysis. There is a measure of this, and we discuss it.

Complicating your learning of regression diagnostics, terminology is not standard, and there are many statistics in the literature that are not aptly named. For example, there is a measure called *Cook's distance* that we discuss later, but it is really a measure of influence rather than distance. Another statistic, called *Mahalanobis distance*, is a measure of leverage, because it quantifies how unusual a case's pattern of regressor scores is.

Cases with high leverage are here called *leverage points*. A case high in leverage has the potential to be influential, but it may not be. Cases high in distance are here called *outliers*, though other writers often use this term to describe any kind of extreme or unusual case. Cases high in influence are here called *influential cases* or *influential points*.

Leverage differs qualitatively from distance, in that cases extreme in distance can invalidate statistical inference in regression. But extreme leverage violates none of the standard assumptions of regression, because regression analysis makes no assumption about the distribution of regressors.[1] But high-leverage cases can affect power and precision of estimation. Consider a single dichotomous regressor such as sex. If a sample includes 90 men and 10 women, the difference between men and women on Y is going to be estimated with less precision than if the sample includes 50 men and 50 women. But the 10 women are going to be much higher in leverage than the 90 men.

16.1.3 Quantifying Leverage, Distance, and Influence

There is a variety of different ways that leverage, distance, and influence can be measured, depending on how you think about these concepts, and we talk about only some of them. They are all interrelated in one way or another.

Leverage. We start first with leverage, which we defined earlier as the atypicalness of a case's pattern of values on the regressors in the model. A case in a data set may have quite ordinary values on the individual regressors, but its combination of regressor values might be quite unusual. For instance, being 55 and being pregnant each are not particularly unusual

[1]There is a common misconception that regression analysis assumes normally distributed regressors. This is not true. We have seen that dichotomous variables can be used as regressors, and ANOVA is just a special case of regression analysis with dichotomous regressors. But dichotomous variables are by definition not normal.

if you were to randomly sample people from a broad population, but if pregnancy and age were both regressors in a regression model, then a 55-year-old pregnant woman would have high leverage since this combination would be very unusual in almost any sample (except perhaps a sample from a population of older pregnant women).

Consider a single variable X_1 used as a regressor in a simple regression model $\hat{Y} = b_0 + b_1 X_1$. Quantify the discrepancy between X_1 and $\overline{X}_1$ for case i in standard deviations of X_1, and then square this result. This is just the squared standardized value of X_1 for case i:

$$Z^2_{X_{1i}} = \left(\frac{X_{1i} - \overline{X}_1}{s_{X_1}} \right)^2 \tag{16.1}$$

The farther case i's X_1 value is from the mean of X_1, the larger is $Z^2_{X_{1i}}$. $Z^2_{X_{1i}}$ cannot be negative, and it will be zero only if $X_{1i} = \overline{X}_1$. $Z^2_{X_{1i}}$ is known as *the Mahalanobis distance* for case i, which we denote as MD_i, but it is really a measure of leverage rather than distance as we have defined the terms. Note that although we have introduced this statistic in the context of a simple regression model, Y is not used in its computation at all.

As defined in equation 16.1, we might call MD_i *univariate* Mahalonobis distance, because it is a measure of case i's atypicalness on a single variable. But Mahalanobis distance can be defined more generally in a multivariate form that considers a case's atypicalness on a set of regressors. Suppose we have a second variable X_2, and we want to calculate case i's atypicalness in its pattern of values on X_1 and X_2 considered jointly. You might think we could just calculate $Z^2_{X_{2i}}$ for X_2 in a comparable way and then add it to $Z^2_{X_{1i}}$ to get a multivariate Mahalanobis distance that considers both X_1 and X_2. The trouble with this reasoning is that if X_1 and X_2 are correlated, this sum would contain some redundancy. The stronger the correlation between X_1 and X_2, the more likely a case is to be atypical on both, and this sum would double-count part of the discrepancy. A preferred multivariate measure would quantify case i's atypicalness on X_2 accounting or adjusting for the correlation between X_1 and X_2.

Recall from section 2.4.2 that the residuals in a regression are uncorrelated with the regressor or regressors. Later, in section 3.2.2, we showed that if we regress X_2 on X_1, then the residuals from this regression, $X_{2.1}$, are uncorrelated with X_1, making $X_{2.1}$ a measure of X_2 that has been purified of its linear relationship with X_1. With these residuals calculated, we can then quantify how atypical case i is on $X_{2.1}$, the part of X_2 independent of

X_1, using the same logic as above. Case i's atypicalness on X_2 controlling for X_1 is

$$Z^2_{X_{2.1_i}} = \left(\frac{X_{2.1_i} - \overline{X}_{2.1}}{s_{X_{2.1}}} \right)^2 \tag{16.2}$$

but because $\overline{X}_{2.1} = 0$ (residuals always have a mean of zero), equation 16.2 simplifies slightly to

$$Z^2_{X_{2.1_i}} = \left(\frac{X_{2.1_i}}{s_{X_{2.1}}} \right)^2$$

The farther case i's $X_{2.1}$ value is from zero, regardless of sign, the larger is $Z^2_{X_{2.1_i}}$. Now we can add $Z^2_{X_{1i}}$ and $Z^2_{X_{2.1_i}}$ to get MD_i, the Mahalanobis distance for case i on the set of regressors X_1 and X_2. The larger MD_i, the more atypical is case i's pattern of values of X_1 and X_2.

You might wonder what would happen if we reversed the order of computations above, starting first with X_2 and then generating the residuals from regressing X_1 on X_2 to generate $X_{1.2}$. It turns out that this doesn't matter. MD_i will be the same. We can then further extend this logic to k regressors by adding successive values of Z_{ji} to those that come before, i.e., $Z^2_{X_{3.12_i}}$, $Z^2_{X_{4.123_i}}$, and so forth. The resulting MD_i calculated as the sum of all these k values of Z^2 will not be affected by the order of the partialing process.

MD_i will tend to be large for cases that are more distant from the center of a multivariate space defined by the joint distribution of the k regressors. But when statisticians use the term *leverage* in regression analysis, they are often not talking about MD_i but rather a different statistic h_i, which is often labeled case i's *hat value*. It is difficult to talk about the computation of h_i without using matrix algebra, so we refer interested readers to Appendix D where we provide the formula. It turns out that h_i is perfectly linearly correlated with MD_i. That is, their correlation across all N cases is exactly 1. So we know the case highest on MD_i is also highest on h_i, the case that is second highest on MD_i is second highest on h_i, and so forth. Unlike MD_i, which has no upper bound, h_i is always between $1/N$ and 1. Furthermore, $\overline{h} = (k + 1)/N$. From now on, whenever we make specific references to "leverage" in computations, we are referring to h and not MD. As we will see, h appears in the computation of many regression diagnostics, so it has more value in regression diagnostics analysis than MD.

In large data sets with more than a couple of regressors, there will often be one or two cases with large values of MD_i and h_i that stand out in the distribution relative to others. But in small samples, or when considering a

small number of regressors, it would not be uncommon to find several cases with large values. For instance, from the weight loss data set in Table 3.1, the largest MD_i calculated using exercise frequency and food intake is 2.500 and the largest h_i is 0.378. But four of the 10 cases have these largest values. So really they aren't atypical at all. Thus, these aren't perfect measures of atypicalness, but they generally are sensitive to the concept as most people would think about it.

Distance. Distance measures how far case i's Y value deviates from $\hat{Y}_i$. Cases with extreme distance are *outliers*. Such cases are more important than leverage cases because a sufficiently extreme outlier represents a violation of at least one of the standard assumptions of regression, while leverage points do not. An outlier may or may not have high leverage.

Outliers can be the result of clerical errors, so it is always worth checking that first. Assuming any outliers found are legitimate values, they may suggest revisions to the model are needed. For instance, if 80% of the cases in a sample were women and 20% were men, and most of the outliers were men, this might mean you need different models for men and women, and the predominance of women in the sample forces the model to fit the women's data. Thus, developing separate models for men and women, perhaps through the methods discussed in Chapters 13 and 14, may be appropriate. Or if there are too few men to develop a separate model for men, a large number of male outliers may suggest that they be excluded from the sample and that the conclusions of the model be applied only to women.

The most obvious measure of distance is a case's residual $e_i = Y_i - \hat{Y}_i$, but residuals can be refined. Cases with high leverage tend to pull the regression surface toward them more than other cases do, thereby shrinking their own residual. So residuals can be adjusted for the case's leverage. We can define a leverage corrected residual as $e_i/\sqrt{(1-h_i)}$. Leverage-corrected residuals are rarely actually used, but an interesting fact is that the square of a leverage-corrected residual equals the amount $SS_{residual}$ would drop if case i were excluded from the analysis.

The expected value of the squared leverage-corrected residual is ${}_T\text{Var}(Y.X)$, which is estimated by $MS_{residual}$. So leverage corrected residuals can be standardized by dividing them by $\sqrt{MS_{residual}}$, as

$$str_i = \frac{e_i}{\sqrt{(1-h_i)MS_{residual}}} \tag{16.3}$$

In large samples, residuals transformed by equation 16.3 are normally distributed with a mean of zero and a standard deviation of one. An additional transformation

$$tr_i = str_i \sqrt{\frac{df_{residual} - 1}{df_{residual} - str_i^2}}$$

results in residuals that are exactly t-distributed with $df_{residual} - 1$ degrees of freedom. We will refer to these as *t-residuals*.[2] Because t-residuals are exactly t-distributed, they are useful for testing some of the standard assumptions of regression, as discussed in section 16.2. The transformation of str_i to tr_i does not change the relative ordering of the cases on these measures of distance. Their rank correlation will be 1.

Earlier we said that cases with high leverage tend to pull the regression surface toward them more than cases with low leverage, thereby shrinking their own residuals. We also just said that tr_i quantifies distance for case i in reference to its $\hat{Y}_i$ when it is excluded from the analysis. It turns out h_i has a similar interpretation. Define e_i as case i's ordinary residual $Y_i - \hat{Y}_i$ and define ${}_{d}e_i$ as $Y_i - \hat{Y}_{i,\text{not}\,i}$, where $\hat{Y}_{i,\text{not}\,i}$ is defined as in section 7.2.3, as case i's estimate of Y derived from the model estimated without case i. It turns out that

$$h_i = \frac{{}_{d}e_i - e_i}{{}_{d}e_i}$$

In words, h_i equals the proportion by which case i lowers its own residual by pulling the regression surface (i.e., the model that produces $\hat{Y}$ for all cases) toward itself. Consider, for instance, a case with a residual of 6, meaning its Y is 6 points above its $\hat{Y}$. If that case's residual would be 8 points above its $\hat{Y}$ if it were excluded from the analysis, then that point's h_i is $(8 - 6)/8 = 0.25$ since inclusion of the point has pulled the regression surface 25% of the way toward the point. Thus, the highest possible value of h_i is 1. The lowest possible value is $1/N$; a case exactly at the mean on all regressors but above or below the mean on Y will not change any regression coefficients, but it will pull the entire regression surface up or down $1/N$th of the point's distance from the surface's previous location. This may be the simplest single definition of h_i, but it can't be considered the primary definition, because it obscures the important fact that h_i is computed without reference to Y. That is, h_i is determined entirely by the

[2]Terminology is inconsistent in the literature and computer software. What we call t-residuals other authors and some statistics programs call *studentized* residuals. A distinction is also made by some authors between *internally* studentized residuals and *externally* studentized residuals. In our notation, these are str_i and tr_i, respectively. SPSS produces something it calls *standardized* residuals, but these are something different still.

regressors, not by Y. It also reduces to an indeterminate form when ${}_d e_i = 0$, but h_i is just as precisely defined for such cases as for any other case. But h_i is not really a measure of distance, even though we have included this manner of defining h_i in this section.

Influence. The influence of a case is quantified by the extent to which its inclusion changes the regression solution or some aspect of it, such as the estimates it generates for Y. It is the cases that most change the regression surface by their inclusion in the analysis that we are most concerned about and wish to identify for further scrutiny. There are many ways one can measure influence. We restrict our discussion here to how the inclusion of case i changes $\hat{Y}$ for all cases or how b_j is changed by the inclusion of case i. But these aren't the only ways of quantifying influence; a case could have little influence on a regression coefficient or $\hat{Y}$, but its presence in a model could greatly influence R or $SE(b_j)$, for instance.

The standard measure of a case's influence on the regression surface was suggested by Cook (1977), and is here denoted $Cook_i$. This measure is inappropriately named *Cook's distance*; *Cook's influence* would be better. $Cook_i$ is proportional to the sum of squared changes in values of $\hat{Y}$ across all cases when case i is deleted from the analysis. To be precise, let d_{ij} denote the change in the value of case j's residual when the residuals are rederived after case i is deleted from the analysis. Then

$$Cook_i = \frac{\Sigma_{j=1}^{N} d_{ij}^2}{k \times MS_{residual}}$$

where k is the number of regressors. Thus, $Cook_i$ is a measure of the amount values of $\hat{Y}$ move when case i is deleted from the analysis. It can be thought of as the product of a particular measure of distance and a particular measure of leverage. The key formula is

$$Cook_i = str_i^2 \times \frac{h_i}{(1 - h_i)(k + 1)}$$

As discussed earlier, str_i ranks cases in the same order as tr_i, the best measure of distance from the regression surface. And all the rest of the right side is a measure of leverage in that it ranks cases in the same order as h_i.

Some have stated that $Cook_i$ is distributed as F with $k + 1$ and $N - k - 1$ degrees of freedom. But, in fact, the mean of an F distribution is always over 1, and values of $Cook_i$ are rarely found as high as 1. Also, the standard

assumptions of regression do not require any particular distribution for h_i, but h_i has a major effect on $Cook_i$, so no general rule can be stated for the distribution of $Cook_i$.

In multiple regression we can distinguish between *total influence* and *partial influence*. Whereas $Cook_i$ measures the influence of case i on the entire regression model, as manifested by what it generates for $\hat{Y}$ for each and every case, partial influence measures a case's influence on a specific regression coefficient b_j. If, say, 10 regressors include nine covariates and one independent variable X_1, then we may be more concerned about cases that substantially affect b_1 than about cases with high total influence. Thus, if your focus is on a specific regressor j, you may be particularly concerned about identifying cases that have a lot of influence on that specific b_j, but care little or not at all about how any case influences any of the other $k - 1$ regression coefficients or the regression constant.

A statistic called $dfbeta_i$ quantifies how much case i influences a specific regression coefficient. In a regression model with k regressors, there are $k + 1$ *dfbeta* values for each case, one for each regression coefficient and one for the constant. We will denote the $dfbeta_i$ for regressor j as $DB(b_j)_i$. It is defined as

$$DB(b_j)_i = b_j - b_{j,\text{not}\,i}$$

where $b_{j,\text{not}\,i}$ is b_j when case i is excluded from the analysis. For instance, if $b_j = 1$ but $b_{j,\text{not}\,i} = 0.25$, then $DB(b_j)_i = 0.75$, meaning that including case i in the analysis raises b_j by 0.75. Large values of $DB(b_j)_i$ relative to other cases suggests that case i is having a big effect on the estimate of the X_j's partial relationship with Y. It can be shown that

$$DB(b_j)_i = \frac{e_i ce_{ij}}{N(1 - h_i)\text{Var}(X_j)Tol_j}$$

where ce_{ij} is the residual for case i in the crosswise regression predicting X_j from the other regressors.

16.1.4 Using Diagnostic Statistics

The analysis of regression diagnostics is as much art as science. The ultimate objective is to flag any cases in the data that are unusual or extreme in some fashion for closer scrutiny. Some authors provide rules of thumb for deciding whether a certain diagnostic statistic is too large or offer ways of testing hypotheses about whether a certain diagnostic is larger than you would expect to observe by chance. These hypothesis tests and rules of

thumb make assumptions about the distribution of various diagnostics, but extreme cases may make those assumptions less tenable. So, with one exception mentioned here and later in section 16.2, we recommend instead a descriptive and holistic approach in which you look at the distribution of each of the diagnostics, notice those that really stand out as unusual relative to others, and see if there are some cases in the data that seem to consistently come to your attention using various diagnostics.

We illustrate this approach using the data set in Table 16.1. The 12 cases in the data represent two groups coded $X_1 = 0$ and $X_1 = 1$, such as an experimental and a control condition, along with two numerical variables, X_2 and X_3. The diagnostic statistics in Table 16.1 are generated from a regression estimating Y from X_1, X_2, and X_3.

As already mentioned, one of the first uses of diagnostic statistics is to identify clerical errors or other problems that may have occurred at the data entry or data generation stage of the research. We discussed the use of leverage for this purpose, as cases with an unusual pattern of scores on the regressors will often show up as high in leverage. A leverage measure such as h_i can be useful for the identification of such errors and supplement what can be learned by looking at the minimum and maximum values. We provided an example of how the minimum and maximum values may fail to detect a case with an unusual pattern of values in section 16.1.1.

The data in Table 16.1 provide another illustration of how simple eyeballing of the data or the use of maximum and minimum values can fail to uncover extreme cases. In these data, the third case is highest in leverage, with values of MD_i and h_i of 7.157 and 0.734, respectively. These values are no less than 75% larger than the corresponding statistics for the case with the next highest leverage. But only very careful examination shows that the value of 11 for X_2 is unusual not by itself but in relation to its X_3 value. Notice that all cases with relatively small values of X_3 also have values of X_2 that are relatively small, and that this is true regardless of whether X_1 is 0 or 1. But not so for case 3, which has quite a large value of X_2 even though this case's X_3 value is relatively small. So it doesn't fit the pattern of the association between X_2 and X_3. Yet examining case 3's values of X_1, X_2, and X_3 individually reveals nothing extreme or unordinary about this case, and 11 is neither the maximum nor the minimum value of X_2 in the data, so looking at the minimums and maximums would not flag this case as unusual. Measures of leverage have flagged this case as worthy of further attention. If these were your data, you might take a look at the data collection records to see whether X_2 was entered incorrectly for this case

TABLE 16.1. Various Measures of Leverage, Distance, and Influence

i	X_1	X_2	X_3	Y	$\hat{Y}$	e	${}_d e$	MD	h	str	tr	$Cook$	$DB(b_0)$	$DB(b_1)$	$DB(b_2)$	$DB(b_3)$
1	0	1	3	8	9.523	−1.523	−2.497	3.374	0.390	−0.932	−0.923	0.139	−0.942	0.610	0.121	−0.051
2	0	3	4	13	10.204	2.796	3.800	1.991	0.264	1.558	1.746	0.218	1.200	−0.790	−0.111	0.035
3	0	11	5	17	16.994	0.006	0.022	7.157	0.734	0.005	0.005	0.000	0.004	0.003	0.002	−0.002
4	0	7	8	7	8.857	−1.857	−2.377	1.488	0.219	−1.004	−1.005	0.071	−0.290	0.605	0.064	−0.085
5	0	9	8	10	10.893	−0.893	−1.096	1.117	0.185	−0.473	−0.449	0.013	−0.105	0.173	−0.011	0.000
6	0	12	12	10	8.528	1.472	2.349	3.192	0.374	0.889	0.876	0.118	−0.200	−0.594	−0.014	0.104
7	1	2	5	10	10.664	−0.664	−0.911	2.069	0.271	−0.372	−0.351	0.013	−0.164	−0.152	0.016	0.008
8	1	4	6	15	11.345	3.655	4.637	1.413	0.212	1.968	2.563	0.260	0.551	0.945	0.007	−0.091
9	1	3	4	11	13.037	−2.037	−3.048	2.733	0.332	−1.191	−1.228	0.176	−0.617	−0.839	−0.057	0.154
10	1	7	12	6	6.270	−0.270	−0.444	3.389	0.391	−0.165	−0.155	0.004	0.059	0.005	0.022	−0.033
11	1	9	10	11	11.016	−0.016	−0.020	1.395	0.210	−0.009	−0.008	0.000	0.002	−0.004	0.000	0.000
12	1	13	14	9	9.669	−0.669	−1.149	3.681	0.418	−0.419	−0.396	0.032	0.324	−0.182	−0.029	−0.018

or otherwise examine your measurement system to see if something went awry.

We might worry that case 3, because of its unusual pattern of values on the regressors, may distort the regression surface in some way. Diagnostic statistics can help identify whether this is so for case 3, or perhaps for some other case in the data. Starting first with distance, cases with a large discrepancy between Y and $\hat{Y}$ can suggest a violation of one of the assumptions of regression, such as normality or homoscedasticity. We recommend the use of the t-residual as the best measure of distance rather than relying on str_i or e_i. In section 16.2 we discuss a way of using the t-residuals for testing whether one of the assumptions of regression has been violated. For now, notice that case 3's t-residual is not particularly large in absolute value. We might be more concerned about case 8, with a t-residual of 2.563. You would expect only 1 in 29 cases in a regression analysis to have a t-residual this large or larger in absolute value if the assumptions of regression have been met. So in a sample of only 12 cases, this residual stands out as potentially unusual or uncommon to observe. But as will be seen in section 16.2.4, we would want to correct this probability for the fact that we have looked at 12 residuals rather than just 1 before claiming we have violated an assumption. This should remind you of the multiple test problem discussed in Chapter 11.

Remember that MD_i and h_i measures the atypicality of a case i's pattern of regressor values. Neither of these statistics is calculated in reference to Y. It could be that the large residual observed for case 8 reflects some kind of data entry error for Y. This would be worth checking. You could also calculate MD_i or h_i while treating Y as if it were a regressor. This could be accomplished by requesting your computer to produce one of these leverage measures when regressing some other variable in the data set on X_1, X_2, X_3, and Y. The dependent variable could even be a set of random numbers since the dependent variable is not used in the computation of leverage. When we did so, we found that case 8's leverage was not particularly large (though it was the second largest out of 12, it didn't stand out much from many of the other cases), thereby reducing our concern that its large t-residual is due to a clerical or computational error of some kind.

A case can be influential in that it changes $\hat{Y}$ a lot for all cases in the data, or it could be influential in its effect on one or more of the regression coefficients. The former is measured with $Cook_i$ and the latter with $DB(b_j)_i$. Observe that case 3, our case with the highest leverage, has a tiny $Cook$ value. Notice as well that the regression coefficients and regression constant, as

measured by the $DB(b_j)$ statistics, are barely affected at all by the inclusion of case 3. It has very little influence. The inclusion of case 8 (the case with the largest distance as well) has the biggest influence in shifting all cases' $\hat{Y}$ values around, because it has the largest value of *Cook*. Observe as well that it has the largest $DB(b_1)$ in absolute value. Its value of $DB(b_1) = 0.945$ means that b_1 is 0.945 larger than it would be if this case were excluded from the analysis. With the case included, $b_1 = 2.832$, which means that if this case were excluded, $b_1 = 1.887$. If X_1 coded a treatment or control condition, then including this case makes the adjusted mean difference in Y between the groups 0.945 units larger than it otherwise would be. But note that this value of $DB(b_1)$ is not particularly large relative to some of the other cases. Observe that cases 2 and 7 have values of $DB(b_1)$ that are not much smaller than 0.945 in absolute value. And whether case 8 is included or excluded does not influence whether we claim a statistically significant partial association between X_1 and Y in these data.

16.1.5 Generating Regression Diagnostics with Computer Software

Most good regression programs have options for saving and displaying various regression diagnostics for examination and analysis. Different programs use different labels in the code for generating the same statistic, so take a close look at your program's manual to make sure that you understand what is being generated.

The SPSS command below will generate all the regression diagnostics we have discussed in this chapter.

```
regression/dep=y/method=enter x1 x2 x3/
   save pred resid dresid sresid sdresid cook mahal leverage dfbeta.
```

The options following the **save** command produce, respectively, $\hat{Y}_i$, e_i, ${}_de_i$, str_i, tr_i, $Cook_i$, MD_i, $h_i - (1/N)$, and $DB(b_j)_i$. These diagnostics are inserted into the data file, though not in this order. Note that SPSS produces something called the "centered leverage" rather than h_i. To convert centered leverage to h_i, add $1/N$ to the centered leverage. SPSS labels some of these diagnostics differently than we have. For instance, what we are calling the t-residual, SPSS calls the "studentized deleted residual."

The SAS code below accomplishes something similar:

```
proc reg data=chap16;
    model y=x1 x2 x3/influence;
    output out=ch16diag p=pred r=resid student=str rstudent=t
    cookd=cook h=h;run;
proc print data=ch16diag;run;
```

This code produces a new file (named "ch16diag" in the code above) containing values for each case for all regressors and Y_i as well as $\hat{Y}_i$, e_i str_i, tr_i, $Cook_i$, and h_i, and prints these values on the screen. The influence option following the model command outputs (though does not save) $DB(b_j)_i$ values, though these are expressed in standardized form, meaning standard errors from the estimate of b_j. See the SAS documentation for guidance.

STATA can also generate diagnostic statistics from a regression analysis. For instance, the code below generates $\hat{Y}_i$, e_i, h_i, str_i, tr_i, and standardized $DB(b_j)_i$. The text prior to the comma provides a variable name for the diagnostics saved into the data file. The **list** command prints the diagnostics on the screen.

```
regress y x1 x2 x3
predict pred,xb
predict resid,residuals
predict h,hat
predict str,rstandard
predict tr,rstudent
predict dbb1,dfbeta(x1)
predict dbb2,dfbeta(x2)
predict dbb3,dfbeta(x3)
list pred resid h str t dbb1 dbb2 dbb3
```

The RLM macro described in Appendix A will produce all the diagnostics discussed in this chapter, except for the *dfbeta* values, by adding the **diagnose=1** option to the RLM command. The diagnose option also generates output showing the minimum and maximum values of the regressors and the outcome, $\hat{Y}$, and a few of these diagnostics. See the documentation in Appendix A.

16.2 Detecting Assumption Violations

In Chapter 4 we introduced the assumptions of linearity, normality, and homoscedasticity. In this section we describe some approaches to detecting

violations of these assumptions. These assumptions can be tested individually or they can be tested as a set, though testing them as a set provides only the vague conclusion that an assumption is violated without specifying which one.

16.2.1 Detecting Nonlinearity

Under the assumption of linearity, the expected value of the errors in estimation of Y for any combination of regressors is zero. Residuals can be used to determine whether the linearity assumption is violated, but none of the methods based on a residual analysis that you will find described here or in other books is likely to be as good at detecting nonlinearity as the methods discussed in Chapter 12.

In section 2.4.4 we provide an example of a nonlinear relationship, depicted in Figure 2.7 and replicated here in this section in Figure 16.2, panel A. The best-fitting line of the form $\hat{Y} = b_0 + b_1 X$ is found superimposed on the scatterplot. Notice that for both relatively large and relatively small values of X, the residuals are predominantly negative, but for moderate values of X, the residuals are predominantly positive. Figure 16.2, panel B, depicts the t-residuals generated from $\hat{Y} = 3.289 - 0.220X$, the best-fitting linear regression line, against X (the solid line in Figure 16.2, panel A). Notice the obvious pattern, with negative residuals for extreme values of X and positive residual in the middle of X. This kind of pattern, with residuals that are systematically positive or negative in certain ranges of the regressor, suggests that the relationship between X and Y is not well described as linear. Figure 16.2, panel C, is a comparable plot of t-residuals from the quadratic model $\hat{Y} = 1.254 + 1.587X - 0.359X^2$. The quadratic model itself is depicted with the dotted line in Figure 16.2, panel A. In the scatterplot of t-residuals against X, there appears to be no systematic tendency for residuals to be positive or negative in certain ranges of X, suggesting that any nonlinearity that does exist in the relationship between X and Y is well described by the quadratic model.

For models with more than one regressor, comparable plots of residuals, such as those in Figure 16.2, can be generated with $\hat{Y}$ on the X-axis. Alternatively, a residual scatterplot can be used to check for evidence of partial nonlinearity. For instance, if you are concerned that the partial relationship between X_1 and Y is nonlinear when you control for X_2, you can regress Y on X_1 and X_2, generate the residuals from this regression, and then plot the residuals against X_1, looking for evidence of nonlinearity in the plot.

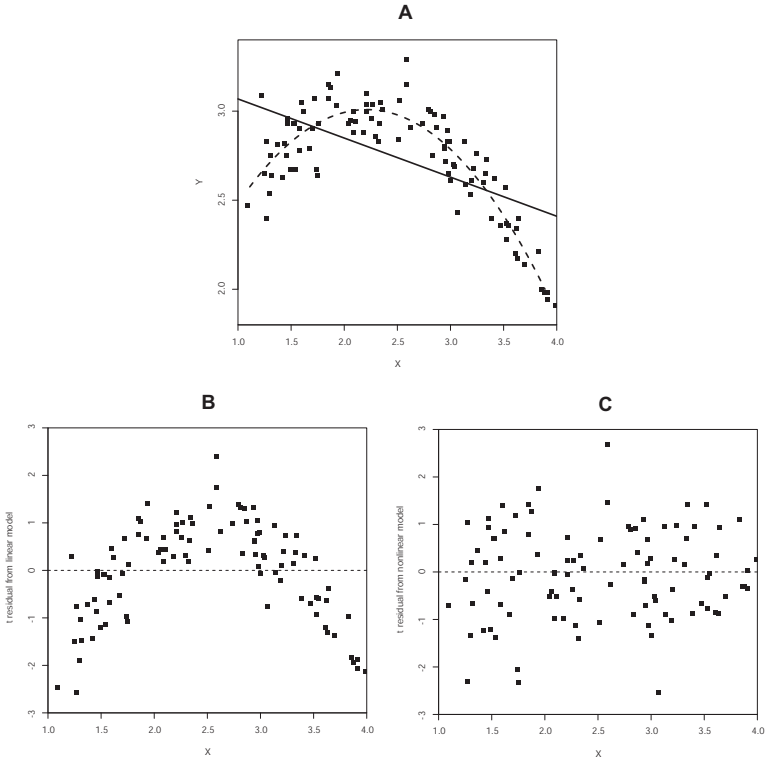


FIGURE 16.2. A nonlinear relationship (panel A) and the *t*-residuals from a model without (panel B) and with (panel C) the square of *X* as a regressor.

Intuition tells us that the conclusions we reach with an "eyeball" test of nonlinearity should be treated with a grain of salt. Looking at scatterplots such as these will tend to reveal only obvious nonlinearity, such as in this example. More subtle nonlinear relationships, such as a shallow curve, are not likely to be detected with the eye. There is also the possibility that your brain may detect a pattern in what is really just a random dispersion of the residuals in the plot. Systematic tests of nonlinearity described in Chapter 12 not only are superior, for they may detect nonlinearity we may not see, but also protect us from misinterpreting random variation as nonlinearity.

16.2.2 Detecting Non-Normality

Regression analysis assumes that the conditional distributions of Y are normal or, equivalently, that the errors in estimation of Y are normally distributed conditioned on the regressors. Some authors recommend constructing a histogram of the residuals (either e_i or tr_i) and eyeballing the histogram to see if you can detect evidence of non-normality. The two problems with this approach are just as described in the section on detecting nonlinearity—that we often see non-normality that really is just random variation, or we fail to see real non-normality when it exists. The eye is good at detecting only obvious non-normality, just as it is good at detecting only obvious nonlinearity. The second problem is that a histogram of the residuals reflects only the marginal distribution of the errors in estimation, ignoring the conditioning that is part of the assumption. The counter to this concern is that if the marginal distribution of the errors in estimation is non-normal, mostly likely so too is one or more of the conditional distributions.

There are formal tests of non-normality of the errors in estimation that one could apply. But they can detect non-normality that is trivial and not likely to affect the accuracy of the inferences one is making with a regression analysis. In Chapter 12 we discussed various transformations that can be used to reduce nonlinearity in relationships that also can have the effect of reducing non-normality in errors in estimation. But they carry with them the disadvantage that transformed metrics may be harder to interpret, and it can be perceived by potential critics as arbitrary and used in an attempt to make results cleaner than they actually are.

Our perspective is that unless you see clear evidence of fairly extreme non-normality in the residuals and have ruled out the existence of clerical errors and highly influential cases using the methods discussed in section 16.1.3, don't worry too much about all but extreme violations of normality. It turns out the normality assumption is one of the least important of the assumptions of regression for most of the widespread uses. You might also consider verifying that your results replicate when using one of the methods we discuss in section 16.3 that make weaker assumptions about the errors in estimation. But if the non-normality is inherent in the system of measurement of Y, such as the result of using a single-item ordinal response scale (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*) or small counts of things (e.g., how many televisions a person has), consider learning about one of the methods discussed in Chapter 18 designed for the modeling of

ordinal, discrete, or count outcomes, which are non-normal by definition or turn out to be so in most applications.

16.2.3 Detecting Heteroscedasticity

In most simple terms, homoscedasticity means that the conditional distributions of Y have equal variances. The assumption is most easily described in the context of simple regression and states that ${}_T\text{Var}(Y.X)$ is the same regardless of X. Because the conditional distribution of Y is centered around $\hat{Y}$, the assumption can also be expressed in terms of the variance of the errors in estimation ${}_T\text{Var}(e.X)$. Figure 16.3, panel A, depicts a sample of 500 cases from a population regression model $\hat{Y} = 5 + 0.25X$ with homoscedastic errors. As can be seen, there is no apparent pattern in the distribution of the residuals or, alternatively, the conditional distribution of Y given X. The residuals appear roughly equally dispersed around the regression line. It appears that the dispersion of Y given X is the same regardless of X.

In the description above, as well as what follows below, we can replace X with $\hat{Y}$, which, of course, is a linear combination of k values of X_j, the regressors in the model. That is, the assumption pertains to the conditional distribution of Y for the linear combination of k values of X_j that is $\hat{Y}$.

Violation of this assumption is known as *heteroscedasticity*. The most common type of heteroscedasticity occurs when ${}_T\text{Var}(Y.X)$, the true conditional variance of Y given X, is largest for the highest or lowest values of some regressor or combination of regressors, a situation we could call *ordinary* heteroscedasticity. Figure 16.3, panel B, depicts such a situation, where the variability of Y and therefore e_i is larger for higher values of X or $\hat{Y}$. Two alternative forms of heteroscedasticity are *butterfly* heteroscedasticity, as in Figure 16.3, panel C, and *inverse butterfly* heteroscedasticity, as in Figure 16.3, panel D. In butterfly heteroscedasticity, the conditional distribution of Y is larger at more extreme values of X or $\hat{Y}$, and in inverse butterfly heteroscedasticity, variability in Y is largest in the middle of the distribution of X or $\hat{Y}$.

In Figure 16.3 we place Y and X on the axes of the figures. But you could replace the Ys with residuals to produce partial scatterplots (see, e.g., Figures 3.10 and 3.12). When testing the significance of the regression coefficient for X_j or producing confidence intervals for ${}_Tb_j$, we would assume *partial homoscedasticity*, meaning that the variance of the errors in the estimation of Y when controlling for all regressors but X_j is uncorrelated with X_j when holding all other regressors constant.

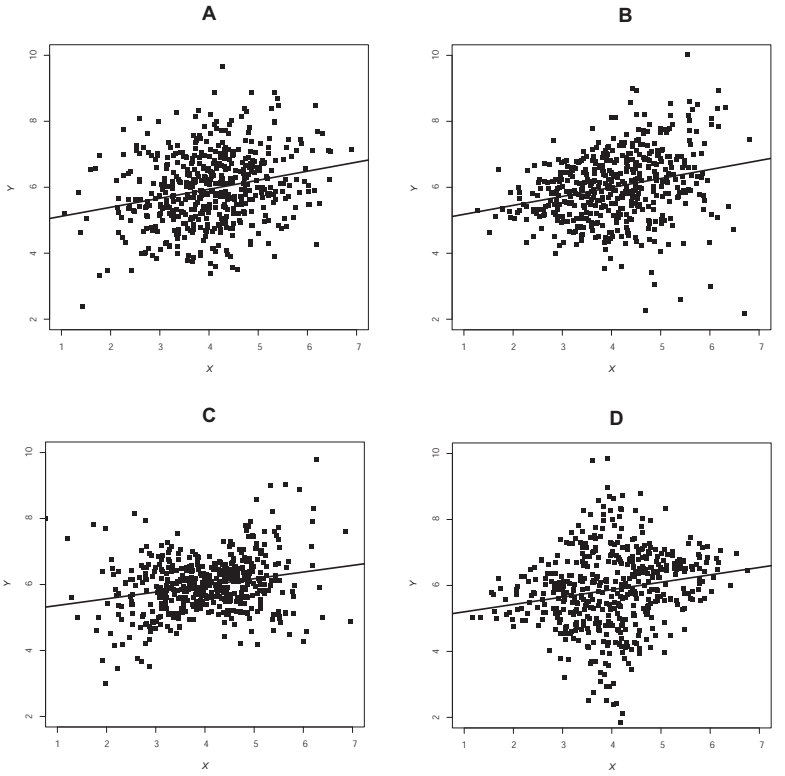


FIGURE 16.3. Scatterplots and the linear regression of Y on X reflecting homoscedasticity (panel A), ordinary heteroscedasticity (panel B), butterfly heteroscedasticity (panel C), and inverse butterfly heteroscedasticity (panel D).

Heteroscedasticity can occur in a number of ways. One way is the existence of an interaction involving one regressor and another variable that may or may not be one of the regressors in the model. For instance, we could imagine in Figure 16.3, panels B or C, drawing two lines relating X to Y, one for group A and another for group B, that differ in slope but are opposite in sign. By ignoring the existence of two subpopulations, each with a different relationship between X and Y, and estimating a single regression coefficient for X can produce a pattern of residuals or conditional distributions of Y that look like those in panel B or C. Including an interaction in the model can eliminate heteroscedasticity.

A second situation that can produce heteroscedasticity occurs when the population of interest is composed of two subpopulations, with one ranging more widely than the other on all the variables. For instance, consider natives born in a specific country and immigrants not born there. The immigrants will typically be from many nations and continents, while the natives by definition are from one. The immigrant subpopulation is likely to be more heterogeneous on many variables than the native subpopulation. When subpopulation A is more heterogeneous than subpopulation B on all variables in the analysis, then the extremes of each regressor will be dominated by group A, which also has a greater variance on Y than group B. This will produce butterfly heteroscedasticity. It can be shown that when a population consists of two equal-size bivariate normal populations, and ${}_TS_X$ and ${}_TS_Y$ are twice as large in one subpopulation as in the other but the simple regression coefficients relating Y to X are equal, ${}_TSE(b_1)$ is 17% larger than the value calculated from the regression formula ordinarily used to calculate $SE(b_1)$. This does not sound like much, but if the actual standard error is 17% larger than what your calculations show, then the probability of finding a significant association between X and Y when there is no real association is nearly twice as high as the α-level being used for the test.

A third situation that can produce heteroscedasticity occurs when Y is measured with less random error for certain cases than others that differ on the regressors. We discuss random measurement error in section 17.2. Suffice it to say now that random measurement error tends to increase a variable's variance relative to what it would be if that variable were measured without random error. If people who score higher (or lower) on X have their Y measured with more error, heteroscedasticity is the result.

Heteroscedasticity can also result when modeling discrete count outcomes using ordinary least squares regression. If your dependent variable were something like the number of times a person donated to political candidates in the last year, Y would be dominated by zeros and 1s, a few 2s, fewer 3s, and so forth. In a least squares linear model of a count Y such as this, the conditional variance of Y is typically positively correlated with its expected value. That means the variance of the errors in estimation will tend to be larger for people who are estimated by the model to donate more often.

Heteroscedasticity does not bias regression coefficients. Rather, heteroscedasticity exerts its influence on inference in regression analysis primarily through its effects on the estimates of the standard errors of the regression coefficients. Ordinary and butterfly heteroscedasticity tend to

result in estimates of standard errors that are too small. This produces confidence intervals that are too narrow and hypothesis tests for regression coefficients and ${}_TR$ that are invalid. Inverse butterfly heteroscedasticity tends to result in estimates of standard errors that are too large. This produces confidence intervals that are too wide and hypothesis tests that are lower in power than they otherwise would be if homoscedasticity were met.

When a regressor is dichotomous, we can talk about the conditional variance of Y in each of the two groups. Heteroscedasticity has its biggest effect on the standard error for the regression coefficient for a dichotomous regressor when the groups are different in size. When the smaller group is more variable on Y, the standard error for the dichotomous regressor tends to be underestimated, but when the smaller group is less variable on Y, the standard error tends to be overestimated.

Given that the quality of our inferences in regression analysis are dependent on the quality of our estimates of standard errors (since standard errors determine confidence interval width and p-values), it is worth testing for its existence so you can make an informed decision about how to proceed. There are many tests of heteroscedasticity that have been described in the regression analysis literature (e.g., Breusch & Pagan, 1979), and you may be familiar with some from the ANOVA literature, such as Levene's test. These generally require some belief about the nature of the heteroscedasticity (e.g., the variance in Y increases with X) or they make assumptions, putting you in the awkward predicament of wondering whether the assumptions of your test of assumptions are met.

Rather than describing these tests, of which there are several, we provide a fairly simple method of testing for ordinary and butterfly heteroscedasticity that can be conducted with any regression program that allows you to generate and save t-residuals, as most do. The test relies on the fact that under the standard assumptions of regression, $E(tr_i^2)$, the variance of the t-residuals, tr_i, is identical for all values on all regressors. So a significant association between tr_i^2 and any regressor or set of regressors is evidence for heteroscedasticity, and we can test for heteroscedasticity by testing the independence of tr_i^2 from the regressors.

The form of this test we advocate requires *normalizing* tr_i^2, which forces its distribution to one approximately normal in form. This process involves replacing the values of tr_i^2 with their rank position in the distribution, such that the smallest squared t-residual gets a value of 1, the next smallest a value of 2, and so forth, up to N. The rank order of ties can be determined

arbitrarily, or they can each be assigned the mean rank for which they are tied. Most statistical programs have a command for replacing scores with their rank position in the distribution.

With these ranks derived, divide them all by $N+1$ and then replace these with the value from the standard normal distribution that cuts off the lower $100\times[\text{rank}/(N+1)]\%$ of the normal distribution from the rest. These values can be found with the help of Appendix C, or they can be derived by your software. For example, if $N = 19$, then dividing the ranks 1 through 19 by 20 yields .05, .10, .15, and so forth, up to 0.95. From the standard normal distribution, these convert Z-scores of $-1.645, -1.282, -1.036, \ldots, 1.645$. These resulting normalized values or Z-scores are known as Van der Waerden scores.

These Z-scores are roughly normally distributed, and under the assumption of homoscedasticity they are independent of all the regressors. So to test for ordinary heteroscedasticity, we regress these Z-scores on all the regressors in the model and test the significance of the multiple correlation. If R in this regression is statistically significant, then the homoscedasticity assumption is violated. If you are particularly interested in certain regressors, you would look at the t-statistic for those regression coefficients in this regression with the Z scores as the dependent variable. A significant regression coefficient implies partial heteroscedasticity.

This approach only tests for ordinary heteroscedasticity. You could also test for butterfly or inverse butterfly heteroscedasticity by including the squares of numerical regressors in this model at the same time. A nonsignificant R would suggest no violation of the homoscedasticity assumption, whereas a significant R could mean either ordinary, butterfly, or inverse heteroscedasticity. But you could test collectively for any butterfly or inverse butterfly heteroscedasticity by testing all squared terms as a set using the method described in section 5.3.3, or you could test for heteroscedasticity due to a specific variable by testing the significance of the set defined as that variable's unsquared and squared terms. If a variable's squared term is nonsignificant, you could drop it and reestimate the model, examining the partial regression coefficient for that regressor as a test of ordinary partial heteroscedasticity, while allowing for butterfly or inverse butterfly heteroscedasticity involving other regressors that still have their squared terms in the regression.

We illustrate by testing for heteroscedasticity in the self-censorship analysis from section 10.2.4. Recall in that example we estimated a person's willingness to self-censor from his or her age and shyness. Age was a mul-

ticategorical variable with four ordinal age categories (Generation X, Generation Y, baby boomer, pre-baby boomer). In the code below we assume that three indicator variables coding age cohort are already constructed and held in variables *d1*, *d2*, and *d3*. The SPSS code below generates and normalizes the squared *t*-residuals and regresses these normalized residuals on age cohort, shyness, and the square of shyness.

```
regression/dep=wtsc/method=enter d1 d2 d3 shy/save sdresid.
compute trsq=sdr_1*sdr_1.
rank variables=trsq.
compute rtrsq=rtrsq/462.
compute z=idf.normal(rtrsq,0,1).
compute shysq=shy*shy.
regression/dep=z/method=enter d1 d2 d3 shy shysq.
```

In STATA, use

```
regress wtsc d1 d2 d3 shy
predict tr,rstudent
gen trsq=tr*tr
egen rtrsq=rank(trsq)
replace rtrsq=rtrsq/462
gen z=invnormal(rtrsq)
gen shysq=shy*shy
regress z d1 d2 d3 shy shysq
```

The SAS code below does the same analysis, assuming that the data file containing the regressors and the *t*-residuals are in a file named "ch16diag." SAS has a special procedure built into PROC RANK for generating Van der Waerden scores, which the code below utilizes.

```
data ch16diag;set ch16diag;trsq=t*t;shysq=shy*shy;run;
proc rank data=ch16diag normal=vw ties=mean out=ch16diag;
    var trsq;run;
proc reg data=ch16diag;
model trsq=d1 d2 d3 shy shysq;run;
```

From this analysis, $R = 0.135, F(5, 455) = 1.695, p = .134$, meaning we fail to reject the assumption of homoscedasticity. But this omnibus test doesn't preclude the possibility of partial heteroscedasticity. The regression coefficient for the square of shyness was not statistically significant, meaning no butterfly heteroscedasticity involving shyness. When the squared term

was removed and the model reestimated, none of the regression coefficients were statistically significant, nor was $R, F(4, 456) = 2.104, p = .079$. We also looked for evidence of partial heteroscedasticity involving age by adding the three indicator variables coding age to the model that already contained shyness (only the linear term). The increase in R was not statistically significant, $F(3, 456) = 1.744, p = .157$. Combined, these analyses support the conclusion that the homoscedasticity assumption is met.

In the next section we describe a test on the whole set of standard assumptions that can be performed by applying a Bonferroni correction to the p-value of the highest t-residual. But that test is not nearly as powerful at detecting heteroscedasticity as the test we just described. In 1,000 bivariate samples of size 50 from artificial populations with butterfly heteroscedasticity, the test we describe next failed to discover the problem in 371 samples, while the test described in this section failed in only six samples.

16.2.4 Testing Assumptions as a Set

In the prior pages we described some methods for examining the plausibility of the assumptions of linear regression analysis. We can conduct a more general test of the null hypothesis that none of the assumptions is violated against the alternative that at least one is violated. Perhaps the simplest method for detecting a violation of this set of assumptions relies on the distribution of t-residuals. As discussed in section 16.1.3, these follow an exact t-distribution under the standard assumptions of regression. Using the $t(df_{residual})$ distribution, one can derive a two-tailed p-value for tr_i.

The p-value for each t-residual is sometimes misinterpreted as testing the null hypothesis that case i falls on the true regression line. If that were so, then the proportion of significant t-residuals would approach 1 as N increases since almost no cases in fact fall exactly on the true regression line. But if the standard assumptions hold, we expect only 5% of the t-residuals to be significant at the .05 level, no matter how large the sample. The hypothesis tested using the p-value for each t-residual is actually that Y_i falls within a normal distribution of scores around the regression line. But because the number of residuals is N, a Bonferroni correction should be applied to the p-value for each tr_i to compensate for the fact that we are doing N hypothesis tests in search of something statistically significant. So the largest t-residual in absolute value is considered statistically significant only if its significance level is below some chosen α-level, such as 0.05, even after being multiplied by N. A statistically significant residual after

this Bonferroni correction suggests that at least one of the assumptions of regression is violated without specifying which one.

Most statistical packages have a command for generating the *p*-value for a *t*-statistic, so this test is fairly easy to implement once you have generated the *t*-residuals as discussed in section 16.1.5. Section 11.2.5 provides SPSS, SAS, and STATA code for generating *p* from *t*. This test is implemented in the RLM macro described in Appendix A. It provides output containing the largest *t*-residual and its Bonferroni-corrected *p*-value.

Another test of the standard assumptions of regression does not rely so heavily on individual *t*-residuals and may be considerably more powerful for detecting any violation that affects many residuals somewhat without affecting any single residual too greatly. In this test we pick some arbitrary probability, count the number of *t*-residuals that are statistically significant at this level (without a Bonferroni correction), and use the binomial distribution to test whether this number is greater than would be expected by chance. For instance, in a sample of 50 cases, by chance we would expect five *t*-residuals to be statistically significant with a *p*-value of no greater than .10. If we observe 11 such residuals, the binomial distribution tells us that that the probability of observing so many is only .0094; this indicates that at least one of the standard assumptions must be violated. The binomial test is not perfectly accurate for this use, for it assumes that the *N* residuals are statistically independent, and they are not quite independent. But in tests we have run, the error is small.

16.2.5 What about Nonindependence?

We have not yet addressed the assumption of independence. The assumption of independence pertains to the size of the errors in estimation—that there is no relationship between the error in the estimation of *Y* for case *i* and the error in estimation of *Y* for case *j*. This assumption can be harder to test than other assumptions and is probably routinely violated. Nonindependence can creep into a study in all kinds of ways if you aren't careful about your sampling, study design, and data collection procedure.

Nonindependence can have various effects on statistics from a regression analysis, but its effect on standard errors is one of the bigger concerns. Research shows that violation of the independence assumption can result in standard errors for regression coefficients that are too large or too small, but in most circumstances the result will be underestimation. As a result, confidence intervals will be too narrow and *p*-values inappropriately small when this assumption is violated.

To understand why, consider two studies identical in purpose but different in method. Suppose you are interested in comparing men and women in their attitudes toward a controversial social topic, such as gun control. You decide to ask 200 people their attitudes by randomly visiting houses in a city and asking the opinions of everyone at home at the time you visit. Because some houses have more than one person in the home, you won't need to visit 200 houses, but this doesn't change the fact that you will still end up talking to 200 people. Once you have talked to 200 people, you can then conduct a test comparing the attitudes of the men and women you ask.

Now consider a variation on this method, where you actually visit 200 houses because you decided to talk to only one of the people living at each house you visit. This may take more time than the variation of this study just described, but at the end you'll have 200 responses, and you can compare the responses of men and women, just as in the prior version.

In both variants of this study, $N = 200$. But the latter study contains more information about how men and women differ, because the responses of the men and women are more likely to be independent, with the caveat we describe later. Its *effective sample size* is 200 or nearly so, but the former study's effective sample size would be much smaller than 200. In the former study, people living together are likely to have similar attitudes, because we know that people influence each other, they selectively sort themselves into social groups based on similarity in beliefs, and they are more likely to be attracted to and partner with people who are like themselves. So if you were to regress a person's attitude about gun control on an indicator variable coding sex in order to test for sex differences, the errors in estimation of Y are not likely to be independent between people living in the same house. But this is not a problem in the second version of the study, because you have data from only one person in each house. The consequence is that we would expect the standard error for sex to be smaller in the first version of the study, because it is treating the 200 people as if they are providing independent information about variability between people in their attitudes. Although 200 people were asked about their attitudes, we don't have 200 independent measurements of those attitudes.

The problem with the analysis from the first version of the study is not easily fixed after the fact without relying on more complicated regression methods. Although you could include a set of indicator variables to code the house a person lives in, this would consume many degrees of freedom and could drastically lower the power of hypothesis tests.

Earlier we said that the version of the study based on only a single person interviewed at a selected house is more likely to produce independent responses than if everyone at the house were interviewed. This is true, but even then, nonindependence may exist. For instance, people living on the same street may know each other, talk to each other, and influence each other's attitudes. So two people living on the same street may give nonindependent responses even if they don't live in the same house. Or maybe people who are politically liberal are more likely to live on Equality Street, whereas politically conservative people are more likely to live on Liberty Street. Even if no one talks to his or her neighbors, the errors in estimation of a person's response may be related to errors in estimation for people living on the same street.

Or suppose you were to randomly call 500 people living throughout the United States to provide data on some variable of interest. This is common in survey research and public opinion polling. Such a sampling plan might seem like the epitome of a method that would satisfy the independence assumption. But people living in the same state or city might be more similar to each other on the variable you are measuring than people living in different states. Technically, this is a violation of independence, although researchers rarely do or even think much about it. And it is common in experimental research to collect data from people in groups. For instance, perhaps you are presenting stimuli to people on a computer screen, and to save time, you recruit five people at a time and sit them in front of different computers in the same room to collect data from them at the same time. But are their responses likely to be independent? Perhaps, but suppose that the dependent variable is affected by the temperature of the room. If the temperature of the room fluctuates from day to day or even hour to hour, this can produce nonindependence in the errors of estimation of Y between subsets of people in the room at the same time their data were collected. Obviously, if these people are allowed to interact during the study, this can also produce nonindependence, especially if they talk about the study itself, their responses to the questions, and so forth, as the data are being collected.

Although you may not be able to completely avoid or eliminate nonindependence, you can at least be conscious of its possible existence and try to reduce it through choices made about sampling and study design. After the fact, it is hard to eliminate unless you have a good idea of where it comes from. Of course, some methods you are already familiar with are designed with nonindependence in mind. An example is the paired-

samples *t*-test, which is designed for comparing the means of Y among people who are "matched" and hence nonindependent. There are some tests of independence that can be used for certain types of sampling and research designs, and there are special analytical methods that are well suited to modeling data that are likely to be nonindependent in some way, such as *multilevel modeling*. For a discussion of some of these methods and nonindependence more generally, as well as ways of quantifying nonindependence, see Griffin and Gonzales (1995), Grawitch and Munz (2004), Kenny and Judd (1986), Kenny, Mannetti, Pierro, Livi, and Kashy (2002), Luke (2004), O'Connor (2004), and Raudenbush and Bryk (2002).

16.3 Dealing with Irregularities

Neither heteroscedasticity nor non-normality affects the expected values of b_0, b_j, and $MS_{residual}$, so these statistics provide unbiased estimates of ${}_T b_0$, ${}_T b_j$, and ${}_T\text{Var}(Y.X)$ even in the presence of these conditions. But hypothesis tests and confidence intervals can be invalidated by violations of any of the standard assumptions. Thus, you typically should do something about cases suggesting violations of the standard assumptions.

But what do you do? There are many exceptions, but generally your four options are correction, transformation, elimination, and robustification. They are normally considered in that order. *Correction* refers simply to the correction of clerical errors. *Transformation* means applying a logarithmic or other transformation to a variable—either a regressor or the dependent variable—so that the case is no longer so extreme. *Elimination* means eliminating the case from the sample. *Robustification* means replacing the regression analysis by an alternate method less sensitive to extreme cases. Correction of clerical errors needs no discussion here, and transformations were discussed in Chapter 12. In the rest of this section, we discuss elimination and robustification.

When you eliminate a case simply because it is extreme in some sense, you are essentially adding a major qualification to your conclusions. You are admitting that the conclusions apply only to the subpopulation defined as the population of cases that exclude extreme cases like the one or ones you eliminated. At least four questions are left unanswered: (1) how the studied subpopulation differs from the rest of the population, (2) how large the included and excluded subpopulations are, (3) how the independent variables relate to the dependent variable in the excluded subpopulation, and (4) whether these relationships in the excluded subpopulation might

be so large as to make the relationships in the studied subpopulation irrelevant. Nevertheless, an extreme score may be the major available clue that a participant in a study did not understand the directions he or she was given in a survey or experiment, or that the experimental manipulation was done improperly for that one participant, or may in other ways provide a defensible reason for discarding the participant's data. Thus, elimination may be a reasonable choice. This is especially true if post hoc examination of the case reveals something odd about it—for instance, evidence that a person did not understand experimental directions. But for the reasons mentioned, elimination may sometimes be a reasonable choice even without such evidence.

There are two general types of robust approach. One set of approaches uses alternative methods for estimating the regression coefficients. The other uses ordinary formulas for the regression coefficients but some alternative method for calculating significance levels or estimates of standard errors. The former approaches essentially give less weight to outliers. This raises fundamental questions about the purpose of the regression. After all, down-weighting an outlier can lead to a regression solution that fails to represent adequately the fact that such outliers do occasionally occur. So we shall consider only the second approach, in which the investigator uses ordinary regression formulas to derive the best-fitting model, but employs an alternative method to find standard errors, confidence intervals, or *p*-values.

We consider four methods: heteroscedasticity-consistent standard errors, the jackknife, bootstrapping, and permutation tests. All of these are practical only with computers, but with the right software they take anywhere from a few seconds to a few minutes on an ordinary personal computer. None of these are panaceas for problems produced by various irregularities such as assumption violations. Even these methods are nonrobust in certain circumstances too numerous and complicated to outline here. Each has variants we do not describe to deal with some of the weaknesses of other variants. The point of our discussion below is to outline a bit about how these methods work, not to describe all the forms they take or offer recommendations as to the specific circumstances in which you might choose to use them. Each of these methods has been heavily studied. General overviews can be found in Edgington (1995), Efron and Tibshirani (1993), Good (2001), Lunneborg (2000), and Rodgers (1999).

16.3.1 Heteroscedasticity-Consistent Standard Errors

The formula for the standard error of a regression coefficient in section 4.4.3 that is implemented in most regression analysis programs assumes homogeneity in the variance of the errors in estimation. This assumption justifies the use of $MS_{residual}$ in the numerator of the formula as an estimate of the conditional variance of Y, which is assumed to be equal for all combinations of regressors.

There is a family of *heteroscedasticity-consistent* (HC) standard error estimators for the regression coefficients that do not require this assumption. They are known as *sandwich estimators* in the statistics literature, because their formulas in matrix algebra look like a sandwich, with the matrix of values on the regressors as the "bread" and the residuals, usually squared and possibly weighted in some fashion by each case's leverage, serving as the "meat." They are called HC estimators, because unlike the usual OLS standard error estimator, which is biased and does not converge with increased sample size to the proper value when the homoscedasticity-assumption is violated, the HC estimators approach the correct value with increasing sample size even in the presence of heteroscedasticity. In statistics, the converging of an estimator to its correct value with increasing sample size is a property called *consistency*.

Use of one of these standard errors does not require modifying the mathematics to estimate the regression coefficients. Rather, the usual standard error estimator is simply replaced with a HC standard error estimator. There are many forms HC estimators take, the earliest frequently attributed to White (1980) and often called the White or Huber–White estimator and denoted HC0. This early version has been improved into forms labeled HC1, HC2, HC3, and HC4. They defy nonmathematical description. We offer the formula for HC3 in matrix algebra form in Appendix D. Otherwise, see Cribaro-Neto (2004), Hayes and Cai (2007), and Long and Ervin (2000) for details about their computation and examples of application.

When heteroscedasticity is a concern, one of these estimators can provide more solid footing. But Long and Ervin (2000) make a case for the regular use of one of these standard errors even when the homoscedasticity assumption is met. This is because they tend to perform better when the homoscedasticity assumption is violated, regardless of the form heteroscedasticity takes, than the standard error estimator that assumes homoscedasticity. Research shows HC3 and HC4 tend to work best. Importantly, these standard error estimators work well even when the homoscedasticity assumption is reasonable. Given that these estimators are easy to compute

and are even available in some software packages (all these HC estimators are available in the RLM macro for SPSS and SAS described in Appendix A; STATA and SAS offer several of them as well), perhaps one day researchers will rid themselves of the homoscedasticity assumption and use one of these estimators for inference in regression analysis as a matter of routine.

16.3.2 The Jackknife

The jackknife, or *jackknifing*, was given its name by J. W. Tukey on the grounds that it may not be the very best tool for anything at all, but it's a serviceable tool in a great many situations. To jackknife a statistic or a test, divide the sample into g groups of equal size, where g is at least 10. In fact, in practice g is frequently set to N, so each "group" contains only one case. Then compute the statistic of interest after deleting group 1 from the sample; then add group 1 back in, delete group 2, recompute the statistic; then add group 2 back in, delete group 3, recompute the statistic; and continue in this manner through all g groups. At the end of this process, you will have g estimates of the statistic of interest. The standard deviation of these g estimates can be used to compute the standard error of the original statistic. Inference can then proceed in the usual way, by constructing a confidence interval using this jackknife estimate of the standard error. Or you could divide the observed statistic by this standard error and generate a p-value for testing the null hypothesis that the corresponding parameter equals zero using the normal distribution.

16.3.3 Bootstrapping

Like the jackknife method, the bootstrap method has been suggested for inference for virtually any statistic. It is based on a simple idea documented in Efron and Tibshirani (1993). If we make absolutely no assumptions about the nature of the population distributions of the variables measured, then the distribution of the measurements in the sample is in every respect the best estimate of the population distribution. That is, if our sample size is 50, then our best assumption-free estimate of the population distribution of the variables measured is that 1/50th of the cases are exactly like case 1, another 1/50th are exactly like case 2, and so on. We then draw, say, B independent random "bootstrap samples" of size 50 from this imaginary population, where B is some large number. This sampling of the original data is done with replacement, so that the bootstrap sample data set does not just reproduce the original data. We then compute the statistic(s) of

interest in each of these bootstrap samples, giving us B estimates of the corresponding parameter.

In one version of the bootstrap method, we then calculate the standard deviation of all B values of each statistic and use that as the estimated standard error of that statistic. As with the jackknife, the normal distribution is then ordinarily used to test hypotheses about the statistic or to find a confidence interval. B does not need to be particularly large when using bootstrapping in this way. Usually 100 or 200 bootstrap samples will do.

In the other version of the bootstrap we never compute a standard error but base our inferences on the number of bootstrap samples yielding statistics in various ranges. This requires a larger value of B—at least 1,000, but more is better. For example, if b_j is positive in the original sample, the proportion of the bootstrap estimates of b_j that yield negative values of b_j can serve as the significance level for testing the null hypothesis ${}_T b_j > 0$. Alternatively, a confidence interval for ${}_T b_j$ can be constructed by using the percentiles of the distribution of B values of b_j. For instance, for a 95% confidence interval, the lower and upper endpoints are defined as the 2.5th and 97.5th percentiles of the distribution of B bootstrap estimates of b_j.

16.3.4 Permutation Tests

Consider a simple correlation r_{XY} based on a sample of size N. Suppose we were to take the N measurements of Y and randomly match them with the N values of X and then recompute r_{XY}. Imagine doing this 999 times, so we have 1,000 values of r_{XY} including the original one. Suppose we find that the original correlation is the 28th-highest of all 1,000 values. We can then say that if these X scores had been matched randomly with these Ys, the probability is only 28/1,000, or .028, that the original correlation r_{XY} would have ranked so high. This value .028 is the one-tailed significance level p for the obtained correlation; it is a *permutation* or *randomization test* of random association. If we ignored the sign of r_{XY} both in the original data and in all 999 recomputed correlations, then a two tailed p-value is the proportion of the absolute values of the 1,000 correlations that are at least as large as the original absolute correlation.

In this example we held constant the order of measurements on X and randomly reassigned values of Y to those X values. In multiple regression we can hold constant the entire matrix of regressor scores, rescramble the order of the Ys many times, and recompute R and all values of b_j each time for construction of p-values using the same approach as in the simpler example.

Rescrambling the Ys themselves is actually not as powerful as an alternative method. To see why, suppose b_1 is high positive, and one person has extremely high measurements on X_1 and Y, but this person's measurement on Y is about what we would predict from his or her high measurement on X_1. This high Y will increase the variation across all the rescramblings of every b_j. This is as it should be for b_1, but it will also be true for every other b_j tested. So in testing the unique contribution of any regressor X_j, the most powerful procedure will generally be to use the portion of Y independent of all regressors except X_j. This means we should use a different column of residuals for each X_j, and still another column for testing R. Thus, we should altogether use $(k + 1)$ different columns of Y-residuals when constructing permutation tests for partial regression coefficients.

16.4 Inference without Random Sampling

In section 6.1.3 we mentioned briefly that valid statistical inferences may be drawn without random sampling, and even without either random sampling or random assignment. In an example presented there, we pondered a statistical test about the change from one decade to another in the proportion of female professors hired by a particular college. Or suppose a club of 50 local businesspeople contains 30 retailers and 20 others. If 25 of the retailers but only 10 of the others vote to change the bylaws, it is valid to perform a 2×2 test of independence in a cross-tabulation to test for a nonchance association between vote and type of business. But, again, there is no hint of either random assignment or random sampling from a broader population. When used in this way, tests of association test the null hypothesis of random association—the hypothesis that the association observed between two variables is caused solely by chance.

Both these examples could be instances of nonsampling, because there is no sampling at all. In the first example, we might study every professor ever hired by the college, or in the second example, the entire membership of the business club. But it is often difficult to distinguish between nonsampling and nonrandom sampling. For instance, in the second example we might think of the local club as a nonrandom sample of the population of members of other business groups in that city or in the nation. The distinction between nonsampling and nonrandom sampling is unnecessary, as well as ambiguous since the types of conclusions we can draw are much the same under both conditions. So the important distinction we must make is between the presence and absence of random sampling.

Nonrandom sampling and nonsampling are very common in both large-scale and small-scale research. On a small scale, suppose an experimenter posts an ad asking for volunteers to serve as participants in an experiment, and uses the first 20 people who sign up. Those participants are not a random sample from any broader population. But if the experimenter assigns the 20 subjects randomly to conditions, then the experiment has random assignment without random sampling. On a larger scale, many behavioral scientists study the entire population of interest: Analysts at the Educational Testing Service have data from all students who take College Board tests, workers at the American Association of Medical Colleges have data on every applicant to an American medical school, census analysts have data from virtually the entire U.S. population, and so on.

Frick (1998) and Mook (1987) discuss how it is inappropriate to put random sampling on a pedestal, thereby condemning all studies that fail to include it inferior in some way. But others have argued that studies that don't include a random sampling component are "pseudoscientific" (Potter, Cooper, & Dupagne, 1993). We agree with the former perspective. Random sampling certainly has a role to play in the kind of inferences we can make. But as Frick (1998) notes, we should distinguish between inferences about *process* and inferences about *populations*. Most researchers care about *process inference*: what is the process that generates the data and the obtained result? They often care less or not at all about *population inference*: does the result obtained reflect what would have been found if the entire population could have been included in the study? Of course, some people care very much about population inference. Public opinion pollsters who generate poll results you read about in the news are an example. Their business is founded on the importance of solid population inference. But most researchers have different research goals than the typical pollster has.

When a significant association between two variables is found under random sampling, it establishes both the *replicability* and *meaningfulness* of the association. We say an association is meaningful if valid hypothesis tests indicate that chance may be excluded from a list of the possible causes of the association. We say the association is replicable if we can have a certain confidence that a nonzero association will be observed again under specifiable conditions, such as drawing a large second sample from the same population. Finding a statistically significant association under nonrandom sampling establishes the association as meaningful, though not necessarily replicable. This at least allows us to speculate on the causes of

the association, as in the previous examples concerning the college's hiring practices or the business club.

When there is random *assignment* without random sampling, as in the example involving the signup sheet, we can go beyond such speculation. Then the existence of a causal relation can be demonstrated, though its generality or replicability is still unknown. In particular, if scores of a treatment group are significantly above those of a control group, then you have shown that the treatment increases at least some scores. This can be a finding of some interest if the dependent variable is a trait thought to be wholly beyond control, such as baldness—or if the independent variable is thought to be imperceptible, such as infrared light or messages flashed on a screen too fast to be seen consciously.

Conclusions of this sort can sometimes be generalized to a broader population, even without random sampling. This is possible if it is assumed that causation is unidirectional, meaning that exposure to the treatment condition rather than the control does not lower anyone's score on the dependent variable. Then, even without random sampling, we have shown that the treatment increases the population mean merely by demonstrating that it raises at least some scores in the population but doesn't lower any scores.

16.5 Keeping the Diagnostic Analysis Manageable

At the level we have now reached in regression analysis, it may be clear that statistical analysis is as much art as science, and not a set of mechanical do and do-not rules. But some general suggestions on the conduct of diagnostic analysis should be helpful.

We saw in Chapters 12, 13, and 14 that curvilinearity and interaction can distort analyses that ignore them, and the same is true of the various kinds of irregularities considered in this chapter. Thus, all these chapters concern potential complications. When should you check for them? You cannot do everything at once. There is no "right" order of checking for these complications, any more than there is a right order of checking for problems when you buy a used car. But there are three reasons for normally applying diagnostic methods before checking for unanticipated curvilinearity or interaction. First, diagnostic methods can uncover clerical errors, and such errors clearly should be detected as early as possible. Second, at least the basic diagnostic methods are easier, and it is always sensible to do easier things first. Third, experience suggests that diagnostic methods uncover

complications more often than do tests for curvilinearity and interaction, and you want to find any problems as soon as possible.

A diagnostic analysis always concerns a particular regression, so the first step in a diagnostic analysis is to identify the regression analysis you would conduct if there were no irregularities. The diagnostic analysis should focus on that regression.

The next step is to choose particular diagnostic methods and tests. We have described measures or tests for many types of irregularities involving leverage, distance, influence, partial influence, and several kinds of heteroscedasticity. And examination of partial influence and partial heteroscedasticity can be done for each regressor. Thus, the number of possible analyses may be large. You should not try to use every one of these tools in every possible analysis. Rather, you should focus on the three major goals of the diagnostic analysis: to check for clerical errors, to examine previously suspect cases, and to test the standard assumptions of regression. To check for clerical errors, check the cases with the highest scores on overall leverage, distance, and influence. Previously suspect cases should be checked primarily for excessive influence—either total or in part—for the most important regressors.

To test the standard assumptions of regression, nearly every analysis should include the Bonferroni-corrected test on the highest t-residual. In addition, tests for ordinary and butterfly heteroscedasticity described in section 16.2.3 should be routinely conducted. The exception might be if you choose to use a heteroscedasticity-consistent standard error estimator for inference, but even then, it isn't a bad idea to test for heteroscedasticity, because its detection could reveal things about the model that could be modified, such as including a missing interaction. And if the major focus of the analysis is on a single regression coefficient b_j, then pay special attention to things that might affect the quality of the estimate of b_j or inference about ${}_T b_j$. For any of these tests, absence of significance does not prove the assumptions hold, but at least violations of the assumptions have been given a chance to show themselves.

16.6 Chapter Summary

Regression diagnostics are used to detect unusual or irregular cases in a data set and to test the assumptions of regression. Before taking any regression analysis at face value, it is important to examine the data for any irregularities, such as impossibly large or small values of regressors

or the dependent variable or strange combinations of regressor values. Often these represent clerical errors or other data collection problems, and they should be fixed. But unusual cases may be hard to detect by merely eyeballing the data in search of something strange. Diagnostic statistics that measure leverage—the atypicality of a pattern of regressor values—can be helpful in this task.

If a case's value of Y is very far from $\hat{Y}$—the case's distance—this may represent a violation of one or more of the assumptions of regression analysis. The residuals, after a transformation, can be used to test whether the assumptions of normality or heteroscedasticity have been violated using one or more of the methods discussed in this chapter. Often, an assumption violation will have no deleterious effects on the quality of the resulting inference and conclusions reached, but you can never be sure, so it is worth looking for assumption violations so you can make an informed decision on what to do about it.

A case can also be highly influential, meaning that its presence in the analysis is having a large effect on the regression results. Measures of influence introduced in this chapter quantify the amount that the inclusion of a case affects the estimates of Y for all cases in data, as well as how a case changes the regression coefficients when it is included relative to when it is excluded from the analysis. These influence measures should be examined and appropriate action taken if a case appears to be distorting a regression analysis, especially if its inclusion seems to work in favor of a hypothesis you are advocating or claim is supported in the data. The decision to include or exclude a case from an analysis should not be taken lightly and needs to be justified. Most important is that you are open with consumers of your research about what you have done.

Assumption violations can affect the validity of the inferences reached with regression analysis or lower the power of hypothesis tests. It is worth examining how robust one's regression analysis is to assumption violations by employing an alternative method, such as bootstrapping or the use of heterscedasticity-consistent standard errors, to see if your conclusions change using one of these alternative methods. This should certainly be done when you have evidence that one or more of the assumptions has been violated, but even if you don't, evidence that an alternative method of inference does not change one's findings can be comforting to both yourself and consumers of your research.

17
Power, Measurement Error, and Various Miscellaneous Topics

This chapter touches on miscellaneous topics in regression analysis. We first address matters of statistical power and some study design considerations that can enhance the likelihood of finding effects that actually exist. We then touch on measurement error, both in terms of what it is and what effects it has on regression parameter estimates and hypothesis tests. We end by describing various problems that can occasionally arise in a regression analysis, such as missing data, rounding error, and overcontrol.

17.1 Power and Precision of Estimation

The power of a statistical test is the probability of obtaining a statistically significant effect if in fact an effect actually exists. We want power to be high when we conduct a hypothesis test. If the power to detect an effect is low, then a failure to reject the null hypothesis is uninformative about whether such an effect actually exists. On the other hand, when power is high, you can be more confident prior to conducting the analysis that if there is an effect to detect, there is a good chance you will find it with the hypothesis-testing procedure you are employing, and so the results of that analysis will be more meaningful and informative.

We can also think about power from the perspective of precision of estimation, although we wouldn't use the word *power* in that context. Results are more informative when interval estimates about a parameter are narrower. This should be obvious. It would be more informative and meaningful if I were to claim that the percentage of people who experience stress daily is between 55 and 75% than if I claimed it is somewhere between 35 and 95%. Whereas the former can be interpreted to mean that

the majority of people do experience stress daily, the latter means that the incidence of daily stress is somewhere between "the majority of people do not" and "most people do."

An important applied problem in data analysis is sample size selection. How large a sample is required in order to detect an effect of a certain size with a certain likelihood or probability? If your sample size is too small to detect an effect that may exist, there is little point to conducting the study with such a small sample. You need more data to detect the effect. On the other hand, scientists are usually operating under constraints in resources such as time and money. It isn't an efficient use of resources to collect more data than one needs to answer the question satisfactorily. So we want our samples to be big enough to detect effects that exist, but not so big that we are wasting resources. Of course, if the data are handed to you free and you've got lots of data, so much the better.

We don't dedicate space here to the specifics of sample size determination or power computations. Power and sample size selection is a complex problem with solutions that depend on so many things, and it often requires lots of educated guesswork in order to get good estimates of power and needed sample size. You will find entire books dedicated to the topic of power (Cohen, 1988) and journal articles about power and sample size in regression analysis (Algina & Moulder, 2001; Dupont & Plummer, 1998; Faul, Erdfelder, Buchner, & Lang, 2009; Gatsonis & Sampson, 1989). Also, freely available and commercially produced computer software exists to do power and sample size calculations (e.g., G*Power and PASS; Faul, Erdfelder, Lang, & Buchner, 2007). But here we talk about some principles in regression analysis as they relate to power and sample size selection to help guide your thinking.

17.1.1 Factors Determining Desirable Sample Size

A well-known but oversimplified rule of thumb for sample size selection is that the sample size in a regression analysis should be at least 10 times the number of regressors. Variations on this rule exist, but they all are based on the ratio of sample size to regressors. As discussed in section 7.2.2, this rule is quite satisfactory for prediction problems, when the focus is R or shrunken R.

But the rules for prediction and causal analysis are very different. When using regression for causal analysis, which typically focuses on regression coefficients or related measures of partial association, the necessary sample size depends heavily on the goals of the analysis. Specific conclusions (e.g.,

${}_Tb_j \neq 0$) generally require larger samples than vague ones (e.g., ${}_TR \neq 0$). Accurate estimates of effects as expressed in the form of narrow confidence intervals generally require larger samples than tests of null hypotheses that effects are zero. Tests of interaction generally require larger samples than tests for parameters in models without an interaction. Analyses with collinearity involving independent variables require larger samples than analyses with collinearity just among covariates or with no collinearity. At one extreme, the null hypothesis that ${}_TR = 0$ can often be tested powerfully with a few dozen cases. At the other extreme, hundreds or even thousands of cases might be needed to test for interaction effects, especially when imbedded in models with several interactions. If one wants a simple rule of thumb about sample size, we repeat what we articulated way back in section 4.7.3: larger is generally better.

17.1.2 Revisiting the Standard Error of a Regression Coefficient

When using linear regression for causal analysis, the focus is usually on measures of partial association for independent variables. Concerns about power are thus directed toward whether one has sufficient power to determine whether an independent variable is uniquely related to Y when other variables in the model are held constant. As discussed in section 4.5, if ${}_Tb_j = 0$, then ${}_Tpr_j$ and ${}_Tsr_j$ are also zero, and rejection of the null hypothesis that ${}_Tb_j = 0$ necessarily leads to a rejection of the null hypotheses that ${}_Tpr_j$ and ${}_Tsr_j = 0$. As the size of the p-value for a hypothesis test of a regression coefficient is determined in part by $SE(b_j)$, anything that affects $SE(b_j)$ will affect the power of a hypothesis test for all these measures of partial association. In section 4.4.3 we discussed things that affect $SE(b_j)$, but we repeat and expand on that discussion in this section in the context of statistical power by examining the factors that affect power of hypothesis tests for regression coefficients.

Equation 17.1 conveys the four components of $SE(b_j)$: $MS_{residual}$, N, $\text{Var}(X_j)$, and Tol_j. Equation 17.1 is simply a re-expression of equation 4.3 with the addition of a square root, which makes it the equation for the $SE(b_j)$ rather than its square.

$$SE(b_j) = \sqrt{\frac{MS_{residual}}{N \times \text{Var}(X_j) \times Tol_j}} \tag{17.1}$$

There is a positive relationship between a quantity expressed as a fraction and the size of the numerator. As the numerator grows, so does that

quantity, and as the numerator shrinks, so does the quantity. In equation 17.1, the only quantity in the numerator is $MS_{residual}$, which estimates the variance of the errors in estimation. Remember that $MS_{residual}$ is minimized by the least squares criterion. If $Y = \hat{Y}$ for every case in the data, then $MS_{residual} = 0$. As the discrepancies between Y and $\hat{Y}$ increase, $MS_{residual}$ increases. This means that the better the model fits the data (as manifested by a smaller $SS_{residual}$, a smaller $MS_{residual}$, and a larger R), the smaller the standard errors for *all* of the regression coefficients, and the larger the power of tests of the null hypothesis that ${}_Tb_j = {}_Tsr_j = {}_Tpr_j = 0$.

This means that anything you can do to reduce the size of $MS_{residual}$ while not affecting the other quantities in equation 17.1 will result in an increase in the power of hypothesis tests for both independent variables and covariates. For example, including another regressor in the model that is correlated with Y but uncorrelated with X_j will increase power of the hypothesis test that ${}_Tb_j = 0$ and reduce the width of interval estimates of ${}_Tb_j = 0$. We saw an example of that in section 6.3.1 when discussing the benefits of controlling for a pretest in an experiment rather than using a difference score. Hypothesis tests for regression coefficients and other measures of partial association are also conducted with greater power when Y is measured better, meaning with less random measurement error (i.e., higher reliability), because random measurement error increases $MS_{residual}$. We discuss measurement error in section 17.2.

There is a negative relationship between a quantity expressed as a fraction and the size of the denominator. As the denominator grows, then the quantity shrinks, and as the denominator shrinks, the quantity grows. In equation 17.1, there are three quantities in the denominator to address with respect to their effect on statistical power: N, $\text{Var}(X_j)$, and Tol_j. These are the sample size, the variance of regressor j, and regressor j's tolerance, respectively.

The presence of N in the denominator of equation 17.1 reflects what is already well known. All other things being equal, statistical power of hypothesis tests is larger in larger samples because increasing N shrinks $SE(b_j)$ for all regressors. Although hardly worth discussing further, it is worth emphasizing that almost anything that lowers power or increases the width of confidence intervals can always be offset simply by increasing the sample size. The exception would be something that makes a standard error zero or infinite, such as $MS_{residual} = 0$, or when any regressor's tolerance is zero.

$\text{Var}(X_j)$ quantifies variation in regressor j, and as it is in the denominator of equation 17.1, this means that if all other things in equation 17.1 are held fixed, power can be increased for a test on X_j by increasing variability in X_j. In the study design phase, this suggests that efforts should be taken to sample in such a fashion that variability in X_j is maximized. As discussed in section 2.3.1, restricting the range of measurement of X_j does not affect b_j, but it does affect its standard error and therefore statistical power; specifically, it raises its standard error, lowers statistical power, and widens confidence intervals for ${}_T b_j$.

The remaining quantity in equation 17.1 is Tol_j, regressor j's tolerance. Tolerance was discussed at length in section 4.4.4. Recall from that discussion that tolerance quantifies the uniqueness of information in regressor j relative to information contained in the other regressors: the proportion of variability in regressor X_j not explained by a linear combination of the other regressors. As the correlation between regressor j and the other regressors increases, the tolerance of regressor j decreases. As a result, $SE(b_j)$ increases, and the power of the hypothesis test for ${}_T b_j$ goes down. At its extreme, when regressor j can be perfectly predicted by the other regressors, $SE(b_j)$ is infinite and power for the hypothesis test of partial association between X_j and Y is zero (and, in fact, the regression computations can't even be done).

The effect of intercorrelation between predictors on the power of hypothesis tests of regression coefficients is perhaps the most vexing problem for investigators using regression analysis. Including a regressor in a model that is highly correlated with others will, all other things in equation 17.1 held fixed, lower the power of the hypothesis tests and increase the width of confidence intervals for the regression coefficients of those regressors it is highly correlated with through its effect on standard errors. But often things we want to control for in linear regression analysis are highly correlated with independent variables we most care about. Failing to control for them opens us up to criticism that effects for those independent variables we care most about are spurious. There is little that can be done about this other than being thoughtful when choosing what variables to include in a regression model.

"All other things in equation 17.1 held fixed" appears several times in our discussion above. It is easier to talk about how changing one term in equation 17.1 affects power under the condition that all other terms are fixed. But in reality, things that change one term may change another term in equation 17.1 in such a way that power is barely if at all affected.

For example, sampling so as to ensure large variance in X_j can increase power of hypothesis tests on ${}_Tb_j$, but it may also lower it by decreasing the tolerance of X_j, as increasing the range of X_j may result in the other regressors explaining more of the variance in Y than they otherwise would.

The fact that the effect of three of the four terms in equation 17.1 operate in ways that may work against each other, but that changing one may be a way of compensating for the effect of changing another, is one reason why we are skeptical of rules of thumb such as the ratio of sample size to regressors or whether variance inflation factors are too high. The negative effects on power of one change can be offset by strategic design choices that affect another factor that has positive effects on power.

The one term we exclude from this dance is N. Changing N generally will not affect any of the other terms unless the sample size is very small to start with. So we know that any negative effect of one term on power of hypothesis tests can always be offset by increasing the sample size. So sample size is always king when it comes to statistical power, and this is one reason why the best rule of thumb for sample size selection is simply that more is better.

We have emphasized how the power of hypothesis tests for measures of partial association between X_j and Y are affected by the four components that determine $SE(b_j)$. But these same things influence precision of estimation. If you seek more precise interval estimates (i.e., narrower confidence intervals), do those things discussed in this section that increase power. Conversely, those things that decrease power have the corresponding effect of increasing the width of confidence intervals, which means decreasing the precision of estimation.

17.1.3 On the Effect of Unnecessary Covariates

We have seen that including a covariate correlated with Y but uncorrelated with X_j can increase the power of the hypothesis test for ${}_Tb_j$ and, by extension, ${}_Tpr_j$ and ${}_Tsr_j$. But if that covariate is correlated with X_j, then the effect of its inclusion on power will depend on the balance of the effect of that covariate on $MS_{residual}$ and $1 - R_j^2$. Including it could raise power, it could have no effect on power, or it could even reduce power.

But what about including a covariate that is not actually necessary? By this, we mean a covariate that does not produce invalidity through overcontrol (discussed in section 17.3.4) but which need not be included because, unknown to the investigator, it actually has no effect on Y. In such a circumstance, the expected effect of including that covariate on

the power of the test for X_j's effect is equivalent to the effect of randomly discarding one case from the data or collecting data from one fewer research participants. It would slightly raise $MS_{residual}$ by subtracting one residual degree of freedom from the model, just as would randomly discarding a single case from the data.

17.2 Measurement Error

17.2.1 What Is Measurement Error?

Each variable in a regression analysis may be and often is measured with some error. Even something as simple as a person's height and weight are rarely measured perfectly, and variables such as attitudes, skills, and socioeconomic variables usually contain measurement error, sometimes in substantial amounts.

Measurement error is not the same as sampling error. *Sampling error* refers to the error in estimating means or other characteristics of the entire population, usually resulting from the process of randomly sampling from that population. *Measurement error* refers to the error in estimating individual characteristics or features of those being measured, such as a person's extraversion, level of education, or knowledge of the political process. If you were to take a test of intelligence offered by a commercial testing firm, you wouldn't care at all about the amount of sampling error that exists in the company's estimate of the average intelligence of people, but you would care a lot about how well your test score reflects your actual intelligence—the amount of measurement error that exists in the score you are given. In large samples, we may have very little sampling error in our estimates of parameters such as a population mean, but we may have lots of measurement error in the individual measurements used to construct that estimate.

In regression analysis, parameters such as ${}_{T}b_j$ are by definition unaffected by sampling error, but they may be distorted by measurement error. We could imagine a concept akin to the "true true regression weight," perhaps denoted ${}_{TT}b_j$, which is the parameter unaffected by both measurement and sampling error. We do not use this notation, but merely mention it to emphasize the point that the parameters that we have been discussing so far acknowledge one source of error but not the other.

Measurement error may be either random or nonrandom. If men tend to overreport their incomes on a questionnaire, whereas women underreport, the nonrandom measurement error can obviously distort conclusions about

the relative earnings of men and women. There is little we have to say in this section about nonrandom measurement error except to note its seriousness, so in our discussion we use the term *measurement error* to mean *random* measurement error.

Measurement error has many sources, depending on what is being measured and how it is being measured. The idea is that even though a person's "true" score on some variable doesn't change, repeated measurements of that variable taken from that person will not all be the same, due to a variety of forces operating on the person being measured and the measurement instrument or procedure being used. For instance, if you were to take a commercial test measuring your mathematical ability repeatedly, you might score 90 today, 95 tomorrow, and 86 the day after. Presumably your aptitude hasn't changed—your *true score* is constant at least over a short period—but your *observed score* varies. In this example, sources of random measurement error would be the specific questions you were asked one day rather than another, the amount of sleep you had the night before, the temperature in the room when you were taking the test, and so forth. All these things contribute to random measurement error, and they make any single measurement—the observed score—less trustworthy as an estimate of the true score.

A variable's *reliability* is defined as the proportion of its variability in observed scores that is attributable to variability in the true scores. The estimation of reliability is a complex matter. The various corrections for unreliability we present in this section assume you have a reasonable estimate of the reliability of the variable. For a discussion on the intricacies of the measurement process, including the computation of reliability, consult one of the many books available on psychometric theory, including Anastasi (1976), Nunnally (1978), or Traub (1994).

17.2.2 Measurement Error in Y

We have seen that in causal analysis, b_j estimates the direct effect of X_j on the dependent variable Y. Thus, it is both important and fortunate that ${}_T b_j$, the parameter estimated by b_j, is not affected at all by random measurement error in Y. To see why this is so, think first about a single mean. Even though random measurement error influences the measurements of individual cases in the data, it will not change the mean of the population, because if it is truly random, it will randomly raise half the scores and randomly lower half the scores by the same amount. Therefore, measurement error will not change a marginal (overall) mean and, by the same argument,

it will not change conditional means. But the population regression surface passes through those conditional means, so it too is unchanged. And the position of this surface is determined by ${}_Tb_0$ and the values of ${}_Tb_j$, so they must not change either. On the other hand, adding more and more measurement error to Y would eventually lower ${}_TR$, ${}_Tpr_j$, and ${}_Tsr_j$ to zero, so these values are clearly affected by measurement error.

With these conclusions concerning population values, we are now ready to consider the effects of measurement error on samples. We know that $E(b_j) = {}_Tb_j$; that is, b_j is an unbiased estimator of the true regression weight. Since measurement error does not change the true regression weight, b_j is still an unbiased estimator of the "true true" regression weight that would exist with no measurement or sampling error. By the same argument, b_0 is still an unbiased estimator of ${}_Tb_0$. But R and values of pr_j and sr_j, which are somewhat biased to begin with, are biased more by measurement error in Y.

Random measurement error in Y does increase the standard error of all b_js, and therefore lowers the statistical power of tests on the regression weights. But the tests are still valid.

17.2.3 Measurement Error in Independent Variables

When regression analysis is used for prediction (of the variety discussed in Chapter 7), random measurement error in regressors lowers R but does not invalidate whatever predictive ability a regression model offers. That is, when we are using regression for prediction, we need not assume absence of measurement error.

In causal analysis, measurement error in an independent variable X_j tends to lower b_j toward zero and also to raise $SE(b_j)$. Both of these effects result in a conservative bias, lowering the power of tests on b_j and increasing the width of confidence intervals. This means that a statistically significant value of b_j is not invalidated by measurement error in X_j since the effect of that error is to make it harder to reject the null hypothesis.

17.2.4 The Biggest Weakness of Regression: Measurement Error in Covariates

Measurement error in covariates has much more serious effects than measurement error in independent or dependent variables. Random measurement error in any regressor X_j changes the values of b, sr, and pr for all other regressors in the direction they would be moved if X_j were omitted

from the regression. We cannot usually tell this direction by inspection, so we often don't know whether measurement error in covariates has raised or lowered the values of b_j, sr_j, or pr_j for the independent variables that we care most about.

To see how this can work, consider the formula for b_1 with two regressors provided in section 3.4.5. Suppose that without measurement error the three correlations among the variables are $r_{YX_1} = 0.3$, $r_{YX_2} = 0.4$, and $r_{YX_1X_2} = 0.5$, and all standard deviations are 1. Then

$$\begin{aligned} b_1 &= \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \times \frac{s_Y}{s_{X_1}} \\ &= \frac{0.3 - 0.4 \times 0.5}{1 - 0.5^2} \times \frac{1}{1} = 0.133 \end{aligned}$$

Now suppose that measurement error in X_2 lowers both correlations involving X_2 to half their correct value, so that we observe $r_{YX_2} = 0.2$ and $r_{X_1X_2} = 0.25$. Measurement error in X_2, of course, will not affect s_Y, s_1, or r_{YX_1}. Therefore, we calculate

$$b_1 = \frac{0.3 - 0.2 \times 0.25}{1 - 0.25^2} \times \frac{1}{1} = 0.267$$

Measurement error in X_2 has in this case doubled b_1. In more complex situations, we cannot easily predict either the size or the direction of the change in any b_j produced by measurement error in the other regressors.

17.2.5 Summary: The Effects of Measurement Error

We can list three possible effects of measurement error, in order of seriousness:

1. Least serious is to leave values of b_j as unbiased estimates of ${}_Tb_j$ but to raise their standard errors $SE(b_j)$. This widens confidence intervals and lowers the power of tests, but does not affect the validity of the tests.

2. More serious is to introduce conservative bias into estimates of ${}_Tb_j$ while still leaving the tests on ${}_Tb_j$ as valid tests.

3. Most serious is to introduce unknown bias into estimates of ${}_Tb_j$; this simply invalidates tests on ${}_Tb_j$.

Measurement error in Y has the first of these three effects, error in an independent variable has the second, and error in covariates has the third. These points were all developed in the preceding sections.

The effect of measurement error in covariates is certainly one of the biggest weaknesses in regression analysis. But this should come as no surprise, as a major purpose of regression analysis is to control for covariates, and it cannot control for covariates well if the covariates are not measured well. At its extreme, if a covariate is measured with total inaccuracy, so that it is all just random error, then that covariate is not being controlled at all. Thus, the result of measurement error is that we have less control than we wish. In the next section, we discuss some partial solutions for the problems created by random measurement error.

But before doing so, it is worth putting the seriousness of this problem in context. Remember that we include covariates in a model because we believe that the association between an independent variable X_j and dependent variable Y is being distorted by the covariate. But as discussed in section 6.1.1, there is an infinite number of possible covariates, and one never knows whether the correct one has been controlled. If it has not, then the hypothesis test on X_j is invalid because b_j is not estimating X_j's actual effect. Yet users of regression analysis seem to accept never knowing for certain whether the proper covariates are in the model. Not including a covariate is like including that covariate but measuring it so badly that its reliability is zero. So measurement error in covariates is not a bigger problem than not controlling for the right covariates. Indeed, so long as the covariate is measured with nonzero reliability, including this imperfectly measured covariate in the model is better than not including it at all.

Furthermore, although it is easy to show the effects of measurement error in simple models, its effect in more complex models with many regressors is hard to predict with many covariates that are imperfectly measured. It is conceivable that the various biases cancel each other out to some extent. Even if there is bias resulting from measurement error in covariates, remember that measurement error in X_j will shrink b_j toward zero, and this can offset bias due to measurement error in covariates. Finally, when Y is also measured with error, $SE(b_j)$ increases, making it hard to find effects. So if the null hypothesis is true, measurement error in Y is going to make it harder to find real effects. Tests are more conservative when Y is measured with error, thereby reducing the likelihood of misreporting effects as real that are solely attributable to bias created by measurement error in covariates.

Although regression has many weaknesses, it is not always apparent how those weaknesses will influence one's conclusions, if at all. Furthermore, all methods have their various weaknesses. Linear regression is still a useful analytical tool, and we think its weaknesses are far offset by its tremendous usefulness and versatility. It is not likely that science is going to ban linear regression analysis because of its weaknesses. So it is worth understanding for its own sake, and because it serves as a foundation for more advanced methods that you will also find uses for in your research life that may not suffer as much from these weaknesses.

17.2.6 Managing Measurement Error

Measurement error is one of the most important limitations of linear models. There are methods for accounting for its influence, and we discuss a few of those methods in this section. All of these methods require some kind of estimate of the reliability of measurement of a variable. Reliability in theory is between 0 and 1, where 0 means that the observed measurements are all random error, and 1 means that there is no random measurement error at all. The proper estimation of reliability is not a simple task. There are many ways of estimating reliability, and they generally produce different estimates depending on the assumptions they make and how they are calculated. All we say here is that a serious attempt to estimate reliability, with awareness of these complexities, is likely to produce better estimates than using values of 1 for reliability that are implicitly assumed when we simply ignore the problem. In what follows we assume that we have appropriate estimates of the reliability of the variables that we need. We shall denote the reliability of Y by rel_Y and the reliability of X_j as rel_j.

To correct for measurement error in Y, we can divide adjusted R or values of pr_j or sr_j by $\sqrt{rel_Y}$. This gives estimates of the corrected values of ${}_T R$, ${}_T pr_j$, and ${}_T sr_j$—values that would exist if there were no measurement error in Y. But values of b_j are already unbiased estimators of ${}_T b_j$, so they should not be altered. And it can be shown that confidence intervals and hypothesis tests on R, pr_j, sr_j, b_j, or any other statistics need not and should not be corrected for unreliability in Y. This unreliability genuinely raises the standard errors of all those statistics, and those increased standard errors should be taken into account when computing hypothesis tests or confidence intervals.

To correct an estimate of ${}_T b_j$ for measurement error in the corresponding independent variable X_j, we can use the formula

$$\text{Corrected estimate of } {}_T b_j = \frac{b_j}{1 - \left(\frac{1 - rel_j}{N \times Tol_j/(N-k)}\right)} \tag{17.2}$$

This complex expression is derived as follows. Combining equation 2.5 with the ideas in section 3.2.2 gives the result

$${}_T b_j = \frac{{}_T\text{Cov}(Y, \text{unique portion of} X_j)}{{}_T\text{Var}(\text{unique portion of} X_j)}$$

It can be shown that the covariance between any two variables is not affected by measurement error in either one, so that the numerator of this fraction is unaffected by measurement error. But the denominator is increased by measurement error. If we arbitrarily assume ${}_T\text{Var}(X_j) = 1$ when measurement error exists, then removing measurement error would lower that variance by $(1 - rel_j)$. But this reduction comes entirely out of ${}_T\text{Var}(\text{unique portion of} X_j)$, whose value before the reduction was ${}_T Tol_j$. Therefore, the ratio of the two values of ${}_T\text{Var}(\text{unique portion of} X_j)$—after and before the removal of measurement error—is $[{}_T Tol_j - (1 - rel_j)]/{}_T Tol_j$. Dividing ${}_T b_j$ by this ratio corrects ${}_T b_j$ for measurement error. But when we apply the conclusions of section 4.3.1 to the crosswise regression predicting X_j from the other $k - 1$ regressors, we find that an unbiased estimator of ${}_T Tol_j$ is $N \times Tol_j/(N - k)$. Substituting this value for ${}_T Tol_j$ in the previous ratio, and continuing in the natural way, gives equation 17.2.

To correct for measurement error in covariates, one can estimate the size but not the standard errors or significance of partial regression coefficients by multiplying the variances of all covariates by their reliabilities while leaving all covariances unchanged, and by then deriving the regression formulas in the usual way from the modified variance–covariance matrix. The standard errors and statistical significance of partial regression coefficients can be estimated with the method provided by Fuller and Hidiroglou (1978).

All of these methods of managing measurement error are applied to the components that are used to construct regression statistics or to the regression statistics themselves generated by a regression program. An alternative approach is the use of a structural equation modeling program such as LISREL, Mplus, or Amos. This is an entirely different approach to estimating linear models that can disentangle sampling variance and mea-

surement error if certain measurement requirements are met. Structural equation modeling is beyond the scope of this book, but all researchers who rely on regression methods should eventually develop some familiarity with this type of modeling. Some resources on the topic include Byrne (2012), Bollen (1989), and Kline (2015).

17.3 An Assortment of Problems

In this section we describe various problems that arise in regression analyses. Some are less serious than you might imagine. Others are serious if unattended to but have solutions that are often simple and acceptable. Still others require supplementary analyses, such as checks for nonlinearities or outliers (extreme measurements) that are described in other chapters. Still other problems are best handled by more advanced or specialized statistical methods detailed in other books. The purpose of this section is to convey a general idea of the seriousness of various problems and the complexities of their solutions without examining the more complex solutions in detail. Some of these solutions we discuss elsewhere in the book.

17.3.1 Violations of the Basic Assumptions

The central assumptions of regression were mentioned in Chapter 4: linearity, normality, homoscedasticity, and random sampling. Chapter 12 described methods for detecting nonlinearity and for transforming variables to make them satisfy the requirement of linearity. The assumption of normality is relatively unimportant, especially in large samples. Chapter 16 describes procedures for detecting and correcting for deviations from homoscedasticity. And section 16.4 shows that useful conclusions can sometimes be drawn even in the absence of random sampling. Thus, we are not completely lost even when all four of the standard sampling assumptions of regression are violated.

17.3.2 Collinearity

Collinearity has been discussed in various places in the book, including sections 3.4.1, 4.4.4, 4.7.1, and 5.3.3. But one major problem left unsolved is the problem of identifying collinear sets. When the number of regressors, k, is large, there will frequently be several regressors with nonsignificant partial relations to Y. But removing all these variables from the regression may lower R significantly. This suggests strongly that the regressors contain one

or more collinear sets. But which variables are in those sets, and which are not? Knowing the answer allows us to be far more specific in our conclusions. For instance, if income, education, and occupational status form a collinear set that relates significantly to Y, you may be able to conclude that Y is affected by SES. But if you merely know that these three variables are among 10 heterogeneous variables whose removal significantly lowers R, you cannot draw such a specific conclusion.

We can often tentatively identify collinear sets merely by inspecting the matrix of correlations among the regressors or even looking at the variable names. Variables that are highly correlated may be measuring similar things, and they may even have similar names. But more elaborate methods are available for more difficult problems.

Factor analysis is a method capable of discovering sets of regressors highly correlated with each other because they are measuring something similar. Factor analysis is too complex to explain here; whole books on factor analysis are available (e.g., Gorsuch, 1983; Kim & Mueller, 1978; Kline, 1994; Thompson, 2004). When variables are highly correlated because they are measuring something in common, factor analysis can help the data analyst to identify one or more things that highly correlated variables have in common. This can help to guide changes in measurement decisions, such as finding a way to aggregate variables that are measuring something in common into a single measure. This often eliminates the collinearity.

All subsets regression is another approach that can be useful. We described all subsets regression in section 7.3.2, although we did not recommend its use for the prediction problems considered there. As described in that section, all subsets regression can efficiently find R for every possible subset of regressors. It was invented for the purpose of finding subsets of *few* regressors yielding *high* values of R. But suppose we turn it around, using it to find subsets of *many* regressors yielding values of R well *below* the R found from the entire set. Such a situation suggests that the *excluded* subset of variables is highly important. For instance, if a subset of seven regressors out of 10 yields an R far below that found from the entire set of 10, it means that the excluded three regressors are highly important when considered as a set. Running an all subsets regression generates all possible values of R from combinations of the regressors, and we can then scan the output to find low values of R associated with a large number of regressors. Then the excluded variables are important as a set even though none may be significant individually. Because all subsets regression examines every

possible set of regressors, this method identifies without fail those sets of regressors whose deletion most lowers R.

This method, however, does have a weakness related to regression to the mean. When we scan many sets of regressors to identify the most important sets, the sets so identified may not actually be as important as they appear in the sample. Methods of dealing with multiple hypothesis tests, discussed in Chapter 11, can deal with this problem.

17.3.3 Singularity

Singularity exists if one regressor has a perfect crosswise multiple correlation of 1.0 when predicted from the other regressors. If that happens, some regression programs will refuse to run even though the data would answer the researcher's questions if it were treated correctly. Other programs will always run but will give answers that are often incorrect for causal analysis. We tell how to avoid both these mishaps after describing some basic points about singularity.

Singularity would occur if an educational researcher had measures of verbal and mathematical skill and studied a model that included both of these as regressors along with their sum. Or a sociologist might include as regressors measures of educational attainment, income, and occupational prestige, plus an overall measure of SES that is a weighted sum of those three variables. We saw in section 9.1.3 that singularity can also be produced by including a dummy variable for every category of a multicategorical variable. Other instances of singularity might involve more complex relationships.

If singularity exists in a set of regressors, there may still be regressors not involved in the singularity. For instance, in any of the examples above there may also be regressors of age and sex that are not involved in the singularity. To deal with singularities we need to know something about which variables are involved. We define a *minimal singular set* as the set of variables actually involved in a singularity. Every variable in such a set is perfectly predictable from the others in the set. In the educational example just mentioned, we could write Sum = Verbal + Math, or Verbal = Sum − Math, or Math = Sum − Verbal. Each of these equations perfectly predicts scores on one of the variables from the other two. Similar equations can be written for the other examples of singularity given above or for any other instance of singularity. A two-variable minimal singular set can exist only if the two variables correlate +1 or −1 with each other, but more complex rules are needed to identify the variables in larger sets.

A set of regressors may contain two or more minimal singular sets, and relations among them may be complex. There may be two nonoverlapping minimal singular sets, as when one set contains X_2, X_4, and X_6, while another contains X_3, X_5, and X_8. Or the two sets may partially overlap, as when one contains X_1, X_2, X_3, and X_4, while another contains X_1, X_2, X_5, and X_6. A single set of variables can even have singularities in every possible subset of three or more. For instance, if X_3, X_4, and X_5 can all be predicted perfectly from different linear functions of X_1 and X_2, then any of those five variables can be predicted perfectly from any two others. If you have studied permutations and combinations, you will know that could be thought of as 10 different minimal singular sets since 5!/(2!3!) is the number of combinations of five things taken three at a time, and that equals 10.

If singularity exists within a set of regressors, there is no unique solution to the regression. To see why, consider the simple case in which $X_1 = X_2 + X_3$. Suppose we form some weighted composite of those three variables. Suppose we then increase the weights of X_2 and X_3 by 1 each and lower the weight of X_1 by 1. Those increases for X_2 and X_3 are equivalent to increasing the weight of X_1 by 1. Therefore, the decrease for X_1 will cancel out the two increases, and the composite will be unchanged. In fact, there are infinitely many ways we could change all three of the weights while leaving the composite unchanged. Therefore, when we try to predict Y from X_1, X_2, and X_3, there is no one best composite, because there are infinitely many different composites that give the same estimates of Y.

Regression programs deal with this in either of two ways. If singularity is detected, some regression programs will stop without printing any results, while others will delete as many regressors from the model as are necessary to remove the singularities, using whatever algorithm or arbitrary rule to make the choice that it is programmed to use. We will call the automatic removal approach ARRES for "Arbitrary Removal of Regressors to Eliminate Singularity," and we will call the former option non-ARRES. ARRES may be perfectly satisfactory for pure prediction, but it can lead to serious errors in causal modeling. An uninformed user studying causation would likely accept ARRES output as the best possible output, and would take the absence of output from a non-ARRES program to mean that no answers are possible. Both of these are serious errors. The proper approach is simpler for ARRES programs than for non-ARRES, so we'll start with that. For simplicity we'll assume here that we're working with just one regression, not a path analysis model involving both direct and indirect effects.

To see why ARRES output should not simply be accepted, imagine that we have three regressors—an independent variable and two covariates. Suppose no two of the regressors correlate perfectly with each other, but the three regressors form a minimal singular set. ARRES may drop out one of the covariates, thus eliminating the singularity, and the program will print results for the independent variable and the remaining covariate. These results will imply to the user that the singularity has been "taken care of." But if an independent variable is in a minimal singular set, there is actually no way to assess the effect of that variable on Y. Recall that any regressor's effect on Y is assessed by studying the relation between Y and the component of that regressor independent of all covariates, and in this case there is no such component. So the proper conclusion is that in this data set, we cannot estimate the independent variable's effect on Y.

But this problem arose only because the singular set included an independent variable as well as covariates. If a minimal singular set includes only covariates, and you drop only the variables that need to be dropped to remove the singularity, you will be retaining all the nonredundant variance in the covariates. Thus, dropping variables from a singular set produces no problems so long as all the variables in the set are covariates. You will get the same value of R and the same regression weights for the independent variables, regardless of which covariates the ARRES program drops.

Some ARRES programs will drop regressors from minimal singular sets in the reverse order that the user entered them into the model, so variables entered last are dropped first. We'll call this an "ordered" ARRES program. You may be able to determine whether your ARRES program is ordered by reading its manual, though this may not be documented. You could also just play around with your program to figure out whether it drops variables in any particular order when a singularity exists. If you have an ordered program, you can enter all the covariates first and all the independent variables last. Then any independent variable in a minimal singular set with covariates will be eliminated from the analysis. That's the proper procedure for such variables, because your data allow you to reach no conclusions about that variable's effect on Y. Any removal of covariates will do no damage because, as already mentioned, you will still be retaining all nonredundant variance in covariates. Thus, ordered ARRES programs will give proper estimates of the effects of the independent variables on Y, provided you enter all covariates into the model before all independent variables.

You can still get the output you need with a nonordered ARRES program, though it takes extra steps. For each independent variable, run a crosswise regression predicting it from all the covariates. Any independent variable in a minimal singular set with covariates will show a crosswise multiple correlation of 1. You must abandon any such independent variable until you have new data, for reasons just described. With these variables removed, you can run the original model. All variables automatically removed in that final run will be covariates, and for reasons also just described, you will still get proper estimates of the effects of the independent variables on Y.

We describe next what we believe to be the simplest way to proceed with a non-ARRES program. Remember that a non-ARRES program won't run when there is a singularity. Temporarily ignoring the independent variables, our first goal is to find the largest possible nonredundant set of covariates. To do this, first predict Y from just the covariates. If this regression runs, you know there is no singularity among covariates and you have accomplished that first goal. If that regression fails to run, set aside Y from your thinking for now and start a process in which you predict one covariate from some or all of the others. If any regression runs and yields $R < 1$, you know there are no singularities in any of the variables involved in that regression. If a regression runs but yields $R = 1$, you know there are no singularities among the regressors in that run because the regression ran, but there is a singularity involving the predicted variable because R was 1. If any variable in that regression was not involved in the singularity, its regression weight will be exactly 0, because perfect prediction was achieved without it. Those with nonzero weights were all involved. This allows you to identify one minimal singular set, though others may also exist. Some programs may not print the values of t for the nonzero regression weights in this regression, because those values of t will be infinite, but that lack of t-values doesn't matter. If the regression fails to run, you have learned that there is singularity within the regressors in that run. By experimenting with various runs of this sort, you can typically find fairly quickly all the minimal singular sets among the covariates. That information will allow you to identify the largest possible nonredundant set of covariates.

After accomplishing this goal, use this reduced set of covariates to predict each of the independent variables separately. These regressions will all run because you have removed any singularity within their regressors. If any of these regressions yields $R = 1$, it means that this independent

variable has no variance independent of the covariates, and in this data set it is not possible to study that variable's effect on Y. That variable must therefore be set aside.

Finally, predict Y from the remaining independent variables and the reduced set of covariates just described. If this regression fails to run, it must be because of singularities among independent variables, since you have eliminated all singularities involving covariates. You can apply some of the procedures above to find these singularities. Ultimately you will eliminate all singularities and a regression will run, predicting Y from some or all of the independent variables and some or all of the covariates, though not all of both those sets. This regression will give you the information you need.

17.3.4 Specification Error and Overcontrol

Specification errors are errors in assumptions about which variables affect which others. These assumptions are essential for causal analysis, when the focus is on regression coefficients. They are irrelevant for prediction-related problems, such as those addressed in Chapter 7, when the focus is on the multiple correlation. The basic specification errors are *undercontrol* and *overcontrol*.

Undercontrol is essentially the problem we have been considering since the beginning of this book: failure to control relevant variables. In any sort of causal analysis, we must assume that the regression includes all variables affecting Y, or that those variables excluded that affect Y are uncorrelated with those in the regression. This is simply saying we must control for all necessary covariates.

We must also assume that Y does not affect any of the regressors. Inclusion in the regression of a variable affected by Y is *overcontrol*. One part of this requirement is obvious and another part is not. It is obvious that if we want to interpret an association between X_j and Y as being due to the effect of X_j on Y, then we must assume that the association is not produced by the effect of Y on X_j. Thus, when interpreting any given partial regression weight b_j, we must assume that Y does not affect that particular regressor.

But the problem of overcontrol has a second, less obvious aspect. Even when interpreting a single partial regression weight b_j, we must assume that Y does not affect any of the *other* regressors. Therefore, unnecessary inclusion of extra covariates can raise the problem of overcontrol.

The problem of overcontrol can be be explained by an example. Suppose we are interested in the effect of academic aptitude on study time. It might

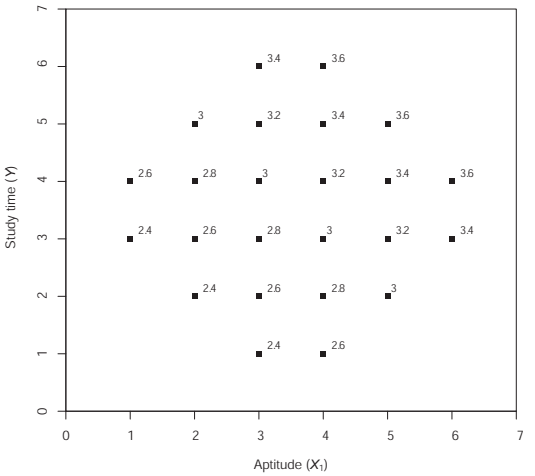


FIGURE 17.1. A graphical illustration of overcontrol.

be argued that naturally brighter students foresee professional careers that require studious preparation, so those students study harder than others. Contrariwise, it might be argued that bright students find that they can obtain satisfactory grades without much study, so that they study less than others. Both of these might be true, but it is still meaningful to ask which effect is stronger. We can examine this question by seeing whether the correlation between an aptitude measure X_1 and study time Y is positive or negative.

Suppose that, in fact, aptitude does not affect study time, so that ${}_T r_{YX_1} = 0$. In a sample of 24 students, we might observe a scatterplot that looks something like that in Figure 17.1. For the moment ignore the numbers in scatterplot; just interpret the dots. The scatterplot clearly shows a correlation of zero between aptitude and study time; that is, $r_{YX_1} = 0$.

Maybe this is just one of several related questions we are researching, and in the questionnaire measuring aptitude and study time we have also measured each student's GPA, so this is in our data file as an additional variable. Because GPA is in our data anyway, we decide to go ahead and control for it statistically; we include it in the regression of Y as X_2.

Those numbers in the body of the scatterplot are these GPAs, X_2. We see that they are arranged much as we might expect; students high on both

aptitude and study time have the highest GPAs, those low on both have the lowest GPAs, and the intermediate GPAs are obtained by students high on aptitude but low on study time, or high on study time but low on aptitude, or intermediate on both.

The simplest way to see the effect of controlling for GPA is to consider a simpler method of control: Select a subgroup of people who all have the same GPA and observe the conditional correlation (the correlation in that subgroup) between aptitude and study time. For instance, select all the students with a GPA of 3.0. A glance at the scatterplot shows this conditional correlation to be highly negative. This negative correlation is what we would expect. Speaking loosely, it says that there are three kinds of students getting intermediate GPAs of 3.0: those low on aptitude but high on study time, those high on study time and low in aptitude, and those intermediate on both. All other students have GPAs above or below 3.0. But notice that the same is true regardless of GPA. Holding constant GPA, those relatively lower in study time tend to be relatively higher in aptitude. Thus, we find a negative partial relationship between aptitude and study time when GPA is controlled, even though we know that there is no true relationship between aptitude and study time.

Figure 17.2 shows a path diagram depicting causal associations among these variables. Our dependent variable of study time affects the "covariate" of GPA, so controlling GPA is a specification error. One way to think about the problem of overcontrol is to say that if X_j affects Y and Y affects another regressor X_k, then by controlling X_k we are removing part of the effect of X_j, and we do not want to do that.

The major factor leading a data analyst to overcontrol is that the very word *control* sounds so good that controlling more variables would always seem to be better. It is so easy to statistically control for many variables in a regression analysis that one might simply control for as many things as are available in the data. But, in fact, the decision to control or not control each variable should be based on careful thought.

Some authors use the term *overcontrol* to refer to the small loss of statistical power that results from the inclusion of each unnecessary covariate. We prefer the meaning given in this section. Since *under*control refers to a potentially drastic loss of validity that can result from failure to control for even one covariate, it is reasonable to use *overcontrol* to refer to a potentially drastic loss of validity resulting from incorrectly controlling even one extra variable, rather than to refer to a *small* loss of power from such control.

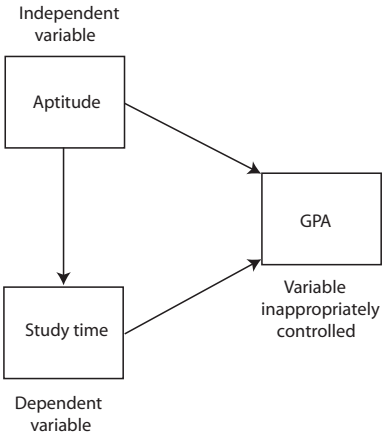


FIGURE 17.2. Controlling for variables caused by the dependent variable results in *overcontrol*.

There is a third form of overcontrol that results not when Y affects a covariate in a regression but, rather, when X_j affects the covariate, although it is appropriate to call this overcontrol a "problem" only in certain circumstances. Suppose your interest was in knowing whether aptitude affects GPA, and you decide to control for study time. If aptitude affects study time, and study time affects GPA (as depicted in Figure 17.2), then by controlling for study time you are removing a part of the process by which aptitude affects GPA. The partial regression coefficient for aptitude in this regression is estimating the *direct effect* of aptitude on GPA. If you are interested in the direct effect, a concept we introduced in Chapter 15, that is fine, and this is not overcontrol. Quite the contrary, this kind of control is necessary to estimate the direct effect. But if you are interested in the *total effect* of aptitude (i.e., the effect operating both directly on the dependent variable and indirectly through a mediator), you'd be underestimating its effect by controlling for the mediator. This would be a form of overcontrol.

17.3.5 Noninterval Scaling

Regression might seem to rely heavily on the assumption of equal-interval scaling. Yet many of the scales researchers use are merely ordinal in nature. An example would be responses to a 5-point attitude scale (e.g., level of agreement with a statement), or a person's evaluation of the quality of food at a restaurant (*poor*, *fair*, *average*, *good*, *great*). It is common in

the sciences for measures of personality, attitudes, and other things to be constructed from an aggregation of more than one ordinal variable. A common form of aggregation would be the use of the unweighted mean to a set of questions with an ordinal response format. Or different items may be given different weights prior to averaging them. Such aggregations are typically treated like interval scales and used in regression analysis without a second thought. We do not take a stand on the legitimacy of doing so, noting only that it is somewhat controversial.

When an ordinal variable is the dependent variable and has only a few values observed or possible, there are methods closely related to linear regression analysis that are designed to appropriately handle such ordinal, discrete variables. The linear regression model based on the least squares criterion can give approximations, sometimes reasonably good ones, to what more sophisticated methods can give. But given that methods designed for ordinal dependent variables are well established in the literature and implemented in most good statistics packages, learning them is a good investment of time, but only after you have a good working understanding of the material in this book. We briefly discuss some of these methods in Chapter 18.

An ordinal variable with only a few measurement values can be used as a regressor in a linear regression analysis if we think of it as a categorical variable. Each scale point on the ordinal variable can be thought of as a category rather than a quantity, and the method introduced in Chapter 9 can be used. And we saw in Chapter 10 some coding systems for categorical variables produce regression coefficients that contain information about how Y changes as the measurements move up or down the ordinal scale.

Otherwise, the term *equal-interval scaling* is used in a number of ways. A critic may say that a variable lacks equal-interval scaling, because a scatterplot depicting its relationship with another variable exhibits some kind of nonlinearity. Or the fact that the distribution is non-normal might be cited. Or it may be argued that a variable's units are not equal to each other in importance. We can deal with all of these issues, but the solutions have little to do with each other.

Chapter 12 describes tests for nonlinearity in a regression analysis and transformations of regressors that can restore linearity. Section 16.2.2 discusses the detection of non-normality. Differences in importance of units must be dealt with on a case-by-case basis. Suppose a high jumper is considering a new training regimen that may raise the best jumps from 6.5 to 7 feet but which may just as well produce minor strains that will lower

performance to 6 feet; the jumper simply has to make a subjective judgment about the relative importance of the two differences: between 6 and 6.5 feet and between 6.5 and 7 feet. A gymnast rated on a subjective 10-point scale must make exactly the same kind of judgment. Demonstrating a very risky skill that improves a rating from 9 to 10 may or may not be worth the risk relative to a somewhat less risky skill that improves a rating from 8 to 9. It is totally irrelevant that the units of the high-jumper's scale are equal-interval in a simple physical sense while the units on the gymnast's scale are not. Similarly, conclusions based on regression must often be supplemented by subjective judgments about the importance of effects predicted by the regression, but the questions are the same whether or not the variables are equal-interval in some mechanical sense.

17.3.6 Missing Data

The term *missing data* has more than one use. For example, we might worry about some conditions in an experiment "missing data" when the number of cases in the groups is not the same. But this is not a problem, and the methods discussed in Chapters 9 and 10 for categorical variables make no assumptions about equality of the sizes of the groups.

In regression analysis, the term refers to the situation in which there are cases with data available on some regressors but missing on others. For instance, a respondent to a survey may answer all the questions except for those on income and marital status. Or participants in an experiment may not respond to all the stimuli, or perhaps procedural or technical errors produce some cases that are missing data on key variables.

We shall consider a few approaches to the problem of missing data, but only briefly to give you an idea of the options. There are entire books dedicated to missing data (e.g., Allison, 2001; Enders, 2010; Graham, 2012; Little & Rubin, 2002) that discuss many of the available methods and their relative advantages and disadvantages. Good journal-article-length treatment can be found by Graham (2009) and Schafer and Graham (2002). This is an important topic, and since all researchers eventually confront the problem, we recommend developing some familiarity with these methods at some point. We consider three methods here: pairwise deletion, listwise deletion, and a few forms of imputation.

Pairwise deletion and listwise deletion are the simplest and best-known approaches to the problem. In *pairwise deletion*, each covariance (the foundation of regression computations) is calculated using those cases for which data are available for the two variables involved. Thus, in principle, every

covariance in the regression computations could be computed on a different subset of cases. When a regression analysis is done using covariances computed in this way, it can yield results that could never occur in a single sample. Pairwise deletion generally can't be recommended for linear models.

By far the most common approach to missing data is *listwise deletion*: deleting from the analysis any case for which any data are missing on any of the variables in the analysis. Listwise deletion will distort estimated means unless the data meet a rather strict assumption: that data are missing solely due to chance. But we are usually more interested in regression coefficients than in means. For estimating regression coefficients, the validity of listwise deletion depends upon a much less strict assumption we call the *noncontribution of missingness*. Little and Rubin (2002) call this the assumption of random missingness, but this name doesn't seem sufficiently informative. The assumption is that missingness makes no contribution to the prediction of Y independent of the regressors. To understand this assumption more precisely, imagine for every regressor X_j a variable F_{X_j}, where F stands for "filled in"; that is, F_{X_j} is what X_j would be if all the missing data on X_j were magically and accurately filled in, as if they weren't missing at all. Imagine also a dummy variable $Miss_{X_j}$ which is 1 on cases for which X_j is missing, and 0 otherwise. There are also variables F_Y and $Miss_Y$ for Y. Then the crucial assumption of listwise deletion is that the *Miss* variables make no contribution to the prediction of F_Y independent of the F_{X_j} variables.

Why is this assumption necessary? Recall that controlling for any regressor amounts to estimating the effects of other regressors in a subpopulation of cases that all have the same values on the regressor being controlled. One way of accomplishing that end would be an extreme form of exclusion of cases—restricting the sample to a subsample of cases with the same value on the regressor being controlled. Listwise deletion effectively controls the *Miss* variables by this method. It deletes from the sample any case scoring 1 on any such variable, leaving only the cases scoring 0 on all. But if a variable makes no independent contribution to Y, then controlling for it does not change the slopes of other variables in the regression. Therefore, controlling the *Miss* variables by listwise deletion does not affect the slopes of the other regressors.

Listwise deletion also leaves the analysis undistorted if the *Miss* variables are uncorrelated with the F_j variables, but this condition is too rare to be of much interest. On the other hand, it is often reasonable to assume

the noncontribution of missingness. For instance, in a questionnaire study, suppose that the *Miss* variables are determined primarily by secretiveness. Then there may be a great many regression problems in which it is reasonable to assume secretiveness has no relation to Y independent of the regressors. Often secretiveness or missingness will be determined by the regressors themselves. For instance, people with exceptionally high or low incomes might be especially reluctant to report them. If income is one of the regressors in the analysis, then it may be quite plausible to assume the noncontribution of missingness.

Rather than discarding cases that are missing data, the missing data can be *imputed* through a variety of methods. The goal of imputation is to replace cases that are missing with reasonable guesses or estimates as to what the missing data would be if they were not missing. One simple procedure, *mean imputation*, is to replace cases that are missing on a variable with the mean of that variable as computed from the cases not missing. *Regression imputation* involves replacing data for missing cases with estimates derived from a regression model estimating that variable from other variables in the data using the cases that are not missing to build the model. *Hotdeck imputation* replaces the missing data on a variable for each case with the values on those variables that are observed in cases that are similar to those missing (so-called "donor" cases), with "similar" being defined in various ways.

Though fairly simple to implement, these imputation methods are difficult to recommend except when the amount of missing data is fairly small (no more than 5 to 10% of cases) and even then arguments can be made for avoiding them. Most important, standard errors tend to be underestimated when imputed data are treated as if they are real, producing confidence intervals that are inappropriately narrow and hypothesis tests that are invalid. *Multiple imputation* gets around this by constructing many imputed data sets, each of which is analyzed. The various estimates (e.g., regression coefficients and standard errors) from each analysis are then fed into an algorithm that generates a proper point estimate and standard error that can be legitimately used for inference.

There are a few methods that don't involve any kind of formal imputation but, instead, rely on estimates of means, variances, and covariances using information that is available from the cases not missing. These methods include the *EM algorithm* and *Full Information Maximum Likelihood*. These are complex methods, but good books on missing data discuss their mechanics and implementation in various software packages.

17.3.7 Rounding Error

Many modern statistical programs carry their computations out to many digits of accuracy. Yet surprisingly, rounding error can still occasionally creep into computations. To illustrate this for yourself, trying running the code below in SPSS.

```
data list free/d1.
begin data.
0
end data.
compute d1=10000000000000.5-10000000000000.4.
compute d2=(.6+.1)*10.
compute d3=(.7+.1)*10.
format all (F16.15).
execute.
```

This code does some fairly rudimentary computations. The first computation subtracts 10,000,000,000,000.4 from 10,000,000,000,000.5 and displays the result in a variable in the data file named *d1*. The correct answer, of course, is 0.1. But SPSS shows the answer as .099609375. The second computation results in 7.000, which is what SPSS does generate for *d2*. The third computation is almost identical to the second and should generate 8.000, yet SPSS shows the answer is 7.9999999999. These apparent computational errors are not unique to SPSS.

Most computing programs, including SAS, STATA, Excel, and others, are susceptible to varying degrees to problems traceable to rounding error due to the way that computers represent and store numbers. The problem is usually exacerbated when doing mathematical operations with numbers that are very large or small and very similar to each other, such as under conditions of near singularity in regression analysis. They can also be a problem with variables with small standard deviations or ranges relative to the means. For this reason, extremely low tolerances in regression analysis or when using a variable whose range is small relative to its mean should prompt some consideration of the problem of rounding error.

You can test your preferred program's vulnerability to rounding error due to the second circumstance—small variability relative to the mean—by asking it to construct the standard deviation of three numbers. Start with the three numbers (10,000,000,001), (10,000,000,002), (10,000,000,003). Each of these numbers contains nine zeros between the first and last digits. If your program uses $N - 1$ in standard deviation computations, it should

TABLE 17.1. Another Data Set Useful for Detecting Rounding Error in Regression

X_1	X_2	X_3	Y
1	1	1	1
−1	−1	−1	−1
d	0	0	0
0	d	0	0
0	0	d	0

display 1 as the standard deviation. Now add the same number of zeros between the first and last digits of all three numbers. With between nine and 15 zeros, SPSS displayed 1 as the standard deviation. But with 16 zeros, it gave a standard deviation of 1.1547, and with 17 zeros it gave a standard deviation of 0.

Here is another way of testing your program's vulnerability. Create a data set such as that in Table 17.1, with the number of cases N equal to $k+2$, where k is the number of regressors. In this table, $k = 3$. Set all variables for all cases to 0, with the following exceptions. Set all variables including Y for case 1 to 1, and set all variables including Y to −1 for case 2. For the remaining cases, set regressor $i - 2$ to $+d$, where i is the case number.

Using this data file, regress Y on the k regressors. Rounding error in most computations tends to increase as d approaches 0 since correlations among regressors all approach 1, as does R. Some exact values using data set up in this manner can be found in Table 17.2. Looking at your regression program's output, you can check how close your regression program comes to generating these exact values. Try it for different values of d, making d increasingly close to 0. You should not use only powers of .1 (e.g., .01, .001, .0001, and so on), since some programs work better on those numbers than on others.

Although no computer program can give accurate answers to all problems, it seems reasonable to expect a program to issue warnings when rounding error is likely. Yet no commercial statistical packages do so. This doesn't mean that you shouldn't trust the output your regression analysis program generates. Most of the time things will be fine. But be aware that it can happen even on well-established and widely respected and used programs.

TABLE 17.2. Formulas That Give Exact Answers Using the Data in Table 17.1

$SS_{residual} = 2d^2/(Nk + d^2)$
$R^2 = 1 - (d^2/(Nk + d^2))$
Overall $F = N/d^2$
$b_1 = b_2 = \ldots = b_k = N/(Nk + d^2)$
$t_1 = t_2 = \ldots = t_k = N/\sqrt{2N(k-1) + 3d^2}$
$r_{X_iX_j} = (2N - d^2)/(2N + (N-1)d^2)$
$Tol_j = 2d^2N(N^2 - 2N + d^2)/((2N + d^2N - d^2)(2N^2 - 6N + 3d^2))$

17.4 Chapter Summary

The power of a test is the probability of it rejecting a false null hypothesis. In regression analysis, the null hypotheses that investigators test usually pertain to measures of multivariate association, such as ${}_TR$ and ${}_TSR$, or the partial relationship between a single regressor and Y, such as ${}_Tb_j$, ${}_Tpr_j$, or ${}_Tsr_j$. Deriving the power of a hypothesis test or determining the sample size needed to detect an effect with sufficiently high probability can be a complex task best left to computers. But understanding the factors that affect the standard error of a regression coefficient facilitates study design and planning choices that increase the likelihood of correctly rejecting false null hypotheses and generating estimates that are more precise.

The presence of random measurement error is the norm rather than the exception. Although almost anything can be measured, measuring things well requires careful thought. Even well-established methods of measuring personality, attitudes, opinions, and other things that scientists routinely study yield data that contain random measurement error. Most of our treatment of linear regression, up to this chapter, has ignored this. The seriousness of the effects measurement error has on estimation and inference in regression analysis depends on whether the measurement error is in the independent variable, the dependent variable, or covariates. Least serious is measurement error in the dependent variable, and most serious is measurement error in covariates. Although random measurement error is nearly unavoidable, there is no reason to panic. In complex models it is

difficult to determine how disruptive, if at all, measurement error is to the validity of one's conclusions. Although it is best to manage measurement error by minimizing it at the measurement phase, there are some statistical techniques one can employ to compensate for its effects. Better still is the use of structural equation modeling, a topic that is beyond the scope of this book.

At some point you will encounter various problems we address in this chapter, including the use of regressors that are very highly correlated or that end up producing singularities, thereby preventing the estimation of the desired model. Although it is tempting to control for any variable you have available in the data, in this chapter we discussed why you should give careful thought to the selection of covariates, so as to reduce or avoid the problem of overcontrol. Missing data is a common problem when conducting any kind of data analysis, and there are many methods available, some better than others, for dealing with it when it is pervasive. And even in these days of low-cost, fast computing technologies, rounding errors can still creep into computations.

18
Logistic Regression and Other Linear Models

In this closing chapter we discuss the application of principles of linear regression to the analysis of a dichotomous dependent variable. We discuss the modeling of the probability of an event as the *odds* or *logit* of the event and show how to calculate estimates of probabilities from a linear combination of regressors. We describe the interpretation of logistic regression coefficients and various inferences, such as testing or estimating the fit of the model. We close the book with a brief overview of some other analytical methods based on the linear model and provide references to sources for more information about these various extensions of the linear model.

18.1 Logistic Regression

In every analysis we have discussed thus far, the dependent variable was assumed to be or was treated as continuous in nature. But not all dependent variables are continuous or even numerical. Sometimes we are interested in modeling a dichotomous dependent variable, meaning that it can take on only one of two possible values. For example, a person may or may not be successful at a task, lose his or her job, or have a child before getting married. A person may get a question right or wrong, or respond one way as opposed to another way. A person may or may not die in a given period, may or may not have a heart attack before age 50, or may or may not completely recover from a surgical procedure or traumatic experience. A company may or may not make a profit during a fiscal year, or a neuron may or may not fire in a certain set of biological circumstances. Many phenomena of interest to scientists are not continuous in nature.

When analyzing dichotomous dependent variables, special modifications to the procedures described so far must be made. In this section we discuss the application of the linear model to the analysis or prediction of a dichotomous dependent variable from one or more regressors. The procedure described here is known as *logistic regression* and is very closely related to multiple regression. With a good understanding of multiple regression, you will find that it does not take long to become comfortable with logistic regression.

That said, our intention is not to make you an expert on logistic regression after reading 19 pages. As we note in section 18.1.8, book-length treatments exist on the topic. Our goal here is only to give you a brief introduction so you can see the similarities and the differences between logistic regression and ordinary linear regression and to prepare you for more thorough and comprehensive study on your own.

18.1.1 Measuring a Model's Fit to Data

One of the fundamental concepts of statistical inference is the consistency between a model and a set of data. Even the most basic inferential tests can be thought of as a comparison between a model and the data. For example, for the independent samples t-test comparing two means, we set up a null hypothesis that the population mean difference equals zero. When data are collected, we reject the null hypothesis if the data are not consistent with it. The p-value from the t-test is essentially a measure of consistency between the data and the null hypothesis. We decide that fit is inadequate if p is less than some predetermined α-level, the level of significance for the test.

Applied to regression analysis, we can think of the task a regression program faces as one of maximizing the fit of the model to the data. The consistency between the model and the data is measured by $SS_{residual}$, with a value of zero indicating perfect fit. Any change to a model, such as by adding a regressor, is evaluated by its ability to increase the consistency between the model and the data as measured by $SS_{residual}$.

When we apply this concept to a dichotomous dependent variable, our first step is to develop a means of measuring the consistency or lack of consistency between the model and the data. Then, just as we try to maximize consistency in linear regression by trying to minimize $SS_{residual}$, we can fit a model to a dichotomous dependent variable by maximizing the chosen measure of consistency.

In most general terms, we can think of the fit of the model in terms of how likely it would be to observe the data actually observed if the model

is correct. In the case of a dichotomous dependent variable, we can ask the probability of the observed data if a certain model is correct. This is the model's *likelihood*. For instance, suppose that a particular model or theory states that success at a task is determined primarily by practice at the task rather than by just listening to lecturers talk about how to do it. Suppose we applied this model to three people who vary in practice and exposure to lectures on the topic, and this model predicts that the probabilities that person A, B, and C will succeed at the task they are given are 0.6, 0.7, and 0.2, respectively. Now suppose that person A and person B actually do succeed, and person C fails. If we think of success as a variable Y coded 1 for success and 0 for failure, then our three measurements of Y are 1, 1, and 0. We can now ask, if the model is correct (i.e., the model that generates these three probabilities), what is the probability of observing these three values of Y for these three people?

If we assume that the performances of these three people are independent, then we can apply the multiplicative law of probabilities to figure out this probability. According to the model, the probability of success for person A was 0.6, the probability of success for person B was 0.7, and the probability of failure for person C was 0.8 (remember the model asserts the probability of person C's success is 0.2, so the probability of his or her failure is 0.8). The multiplicative law of independent events say that the probability of all three of these things happening is the product of their individual probabilities. So the probability of this set of results if the model is correct is $0.6 \times 0.7 \times 0.8 = 0.336$. This is the model's likelihood.

But suppose a second model asserts that exposure to classroom lectures is more important than practice, and when this model is applied to these three people who vary in practice and exposure to lectures, it implies the probabilities of success for person A, B, and C are 0.6, 0.8, and 0.1, respectively. Using the logic above, the probability of observing Ys of 1, 1, and 0 for these three people is $0.6 \times 0.8 \times 0.9 = 0.432$. Since the observed outcomes have a higher likelihood under this model than the first, we can say that overall, the second model is more consistent with the data than the first.

We can formalize the likelihood in the form of a *likelihood function*. Let Y_i be whether (1) or not (0) the event happens for case i, and let PE_i be the estimated probability of the event for case i from some model. Then case i's contribution to the likelihood or fit of the model is

$$Fit_i = Y_i \times PE_i + (1 - Y_i) \times (1 - PE_i) \tag{18.1}$$

If the event does not happen for person i, then we have $Y_i = 0$, so that the first term in equation 18.1 equals zero and Fit_i reduces to $1 - PE_i$. If the event does happen for person i, then we have $Y_i = 1$, so that the second term equals zero and Fit_i reduces to PE_i. But PE_i and $1 - PE_i$ are the probabilities the model calculates for the event happening and not happening, respectively, for case i. Thus, whether the event happens or not, Fit_i equals the probability the model has assigned to the outcome that is ultimately observed. Therefore, the model's likelihood or consistency with the data equals the product of the values of Fit_i for all cases in the data. We can express this as

$$\text{Likelihood} = \Pi_{i=1}^{N} Fit_i \tag{18.2}$$

where Π is the multiplication operation (as opposed to Σ, which denotes addition).

Values of the likelihood can be extremely small, especially in large samples. For instance, if Fit_i were 0.9 for each of 1,000 cases, the likelihood would equal $.9^{1000} = 1.75 \times 10^{-46}$. So we usually report the natural logarithms of likelihood values. But because these are always negative (since a likelihood must be less than 1, and the natural logarithm of a number less than 1 is negative), the value usually reported by a computer doing logistic regression is either $-ln$(likelihood) and denoted $-LL$, or twice this value (and denoted $-2LL$), for which we would say "negative log likelihood" or "negative two log likelihood." Formally,

$$-LL = -\Sigma_{i=1}^{N} ln(Fit_i) \tag{18.3}$$

and $-2LL$ is twice this. These values cannot be negative (which is nonintuitive given that we call it the "negative log likelihood"). Generally, they are positive and measure *lack of fit* between data and model; the smaller the value of $-LL$ or $-2LL$, the better the model fits the data. A value of zero for $-LL$ or $-2LL$ means that model perfectly fits the data. Thus, $-LL$ or $-2LL$ is like $SS_{residual}$ in that it's never negative, zero implies perfect fit, and smaller positive values imply better fit. Also, like $SS_{residual}$, these values typically increase with N.

18.1.2 Odds and Logits

Derivation of the likelihood requires a value of PE_i estimated for each case in the data. We can think of a probability as a kind of mean: a mean of zeros and ones. For instance, if I have 10 males in my sample, six of

whom have a college degree, then I might estimate the probability that a male will get a college degree is 0.6. This is the mean of six ones and four zeros. Given that a regression equation is a model of conditional means, you might think we could just regress a dichotomous Y with values 1 and 0 on a set of regressors to generate an estimated probability of the event as a function of the regressors. However, there are many reasons why we should not do this. We focus on only one of these reasons, and that has to do with the fact that a probability has to be between zero and one. A weighted linear sum of regressors could generate a probability greater than 1, or less than 0. For instance, suppose we found that each 1-unit increase in age increases the probability of some event by 0.05. Now suppose that the probability of the event for a 40-year-old is 0.7. By this model, that means that the probability for a 50-year-old would be $0.7 + (10 \times .05) = 1.2$, and the probability for a 20-year-old would be $0.7 + (-20 \times 0.05) = -0.3$. Neither of these "probabilities" can be such, as a probability must be between 0 and 1.

Consider a different version of this problem. Suppose we have two dichotomous regressors, sex and training at some task. So we have men and women in the sample who either have training or do not. Suppose we find that among untrained people, 50% of men and 70% of women succeed at the task; this is a difference of 20 percentage points. Now suppose that the effect of sex on success does not depend on training. It very well could be that the training increases success, such that perhaps 90% of trained men succeed. But if sex differences in success do not depend on training, then we might think therefore that 110% of women should succeed, which is also an increase of 20 percentage points. But this can't happen.

To model probabilities, we want to convert probabilities into something that has no upper or lower bound, so that estimates from a linear model are not larger or smaller than what is possible. We do this by modeling not the probability of an event, but the *log odds* or *logit* of the event (pronounced "low-jit"). Odds are similar to probabilities; odds simply repackage a probability into a different metric. If the probability of an event is 0.5, then this means that half of the time we expect the event to occur and half of the time we expect it not to occur. So the odds of it happening are 1 to 1, or 1. If the probability of an event occurring is 0.25, the odds of it occuring are 1 to 3 or 0.333. And if the probability if 0.75, then the odds are 3 to 1, or 3. Odds are related to probabilities by the function

$$Odds_i = \frac{PE_i}{1 - PE_i} \tag{18.4}$$

Whereas a probability is bound between 0 and 1, odds are bound between 0 and positive infinity. So this doesn't completely solve the problem. But if we take the logarithm of an odds, then we have a number that can be anywhere between plus and minus infinity. We call this log odds a logit. It is defined as

$$\text{logit}(PE_i) = ln\left(\frac{PE_i}{1 - PE_i}\right)$$

By modeling the log odds of an event—logit(PE_i)—we don't have to worry about the model generating an estimate that is impossible. Importantly, once we have estimated logit(PE_i), we can convert it to a probability with the function

$$PE_i = \frac{e^{\text{logit}(PE_i)}}{1 + e^{\text{logit}(PE_i)}} \tag{18.5}$$

Recall from section 12.4.1 that e is approximately 2.718828.

Now reconsider the problem with the trained and untrained men and women. The success rates of men and women who are untrained are 0.5 and 0.7, respectively, in probability terms. These probabilities correspond to 0 and .847 on the logit scale, which is a difference of 0.847. Now if trained men had a probability of success of 0.90, this is a logit of 2.197. If sex difference doesn't depend on training, then on the logit scale, we'd expect the logit for trained women to be 2.197 + 0.847 = 3.044. Using equation 18.5, this corresponds to a probability of success of 0.955, which satisfies the requirement that a probability must be between 0 and 1.

There is no proof that the logit scale is in any sense the best possible scale for modeling probabilities. But it turns out to be useful, and it tends to give reasonable results. Furthermore, there is a well-developed literature in linear models for modeling logits and testing hypotheses about how variables are related to the likelihood or probability of events occurring. So it has become very popular.

18.1.3 The Logistic Regression Equation

In an ordinary linear model, the equation linking the regressors to the dependent variable is

$$\hat{Y} = b_0 + \Sigma_{j=1}^{k} b_j X_j$$

and the computer finds the values of b_0 and k values of b_j that maximize the accuracy of estimations of Y by minimizing $SS_{residual}$. If we assume that logit(PE) is linearly related to regressors, which turns out to be a fairly

reasonable assumption much of the time, then we can express those logits as a linear function

$$\text{logit}(PE) = b_0 + \Sigma_{j=1}^{k} b_j X_j \tag{18.6}$$

In equation 18.5, we expressed PE as a function of logit(PE), so substituting equation 18.6 into equation 18.5 gives

$$PE = \frac{e^{b_0 + \Sigma_{j=1}^{k} b_j X_j}}{1 + e^{b_0 + \Sigma_{j=1}^{k} b_j X_j}}$$

For any set of specified values of b_0 and k values of b_j, we can calculate PE for each case and then use equation 18.3 to calculate the model's fit to the data. Those values that minimize $-LL$ (or twice its value, $-2LL$) are the values most consistent with the observed data. A logistic regression program will find those values of b_0 and b_j as well as their standard errors.

18.1.4 An Example with a Single Regressor

To illustrate logistic regression, we use the data set (hypothetical) in Table 18.1 from a sample of 24 international companies studied by an organizational psychologist. The data file is available at this book's web page at *www.afhayes.com* and is named LEADER. A variable in the data file named `profit` is set to 1 if the company made a profit last year; otherwise, `profit` is set to zero. And each company was classified as large (`size` = 1) or small (`size` = 0). The leadership ability of the chief executive officer (CEO, and named `ceo` in the data) is also available from judgments provided by the board of directors, where a higher rating corresponds to higher perceived leadership ability.

A Single Dichotomous Regressor. Consider first whether there is a difference between large and small companies in the probability of making a profit. In these data, 90% of the large companies made a profit last year, whereas only 14.3% of the small companies made a profit. If you think of the "event" as having made a profit, then the estimated probability of this event for a large company is 0.900, and the probability of this event for a small company is 0.143.

A logistic regression analysis with profit status as the dependent variable and size of the company as the sole independent variable yields a model of the logit or log odds of making a profit from company size. Doing so yields

$$\text{Estimated logit} = -1.792 + 3.989X_1$$

TABLE 18.1. Company Profit Status, Size, and Leadership Ability of the CEO

Company	Profit Y	Company Size X_1	CEO rating X_2
1	1	1	3.4
2	0	0	2.7
3	0	0	3.2
4	1	1	2.7
5	1	1	3.4
6	1	1	3.8
7	1	1	4.2
8	1	1	3.4
9	0	1	3.7
10	1	1	4.1
11	1	1	4.5
12	1	0	4.3
13	1	0	4.8
14	0	0	3.2
15	0	0	2.6
16	0	0	2.9
17	0	0	3.4
18	0	0	3.1
19	0	0	2.4
20	0	0	2.7
21	1	1	2.1
22	0	0	4.3
23	0	0	2.3
24	0	0	3.2

The regression coefficient for company size is $b_1 = 3.989$, which means that the log odds of making a profit increase as X_1 increases by 1 unit. This 1-unit increase is the difference between the small ($X_1 = 0$) and large ($X_1 = 1$) companies. It seems the log odds of making a profit are higher for larger than smaller companies.

Applying this model to the two company sizes gives

$$\text{Small companies:} \quad -1.792 + 3.989(0) = -1.792$$
$$\text{Large companies:} \quad -1.792 + 3.989(1) = 2.197$$

as the estimated logit or log odds of making a profit for the two types of companies. But the logit is a strange metric to use for discussing these

TABLE 18.2. Estimated Logit, Odds, and Probability of Making a Profit

CEO rating	Logit	Odds	*PE*
1.0	−3.572	0.028	0.027
1.5	−2.853	0.058	0.055
2.0	−2.133	0.118	0.106
2.5	−1.414	0.243	0.196
3.0	−0.694	0.500	0.333
3.5	0.026	1.026	0.506
4.0	0.745	2.106	0.678
4.5	1.465	4.325	0.812
5.0	2.184	8.882	0.899

results. We can undo the log of the odds by raising e to the power of the log odds, resulting in the *odds* of making a profit. Doing so yields

$$\text{Small companies:} \quad e^{-1.792+3.989(0)} = e^{-1.792} = 0.167$$
$$\text{Large companies:} \quad e^{-1.792+3.989(1)} = e^{2.197} = 9.000$$

as the odds for small and large companies. Alternatively, we could apply equation 18.5 to get the estimated probabilities of making a profit. Doing so generates

$$\text{Small companies:} \quad \frac{e^{-1.792+3.989(0)}}{1+e^{-1.792+3.989(0)}} = 0.143$$
$$\text{Large companies:} \quad \frac{e^{-1.792+3.989(1)}}{1+e^{-1.792+3.989(1)}} = 0.900$$

as the estimated probabilities, which are exactly what we calculated earlier. So the logistic regression model has regenerated the observed probabilities.

A Single Numerical Regressor. No change to the mechanics is required for a numerical regressor. Suppose we wanted to know whether a company is more likely to make a profit if it is run by a CEO perceived as having more leadership skills. Regressing the dependent variable of profit status on the leadership ability of the CEO (X_2) using logistic regression yields

$$\text{Estimated logit} = -5.011 + 1.439X_2$$

The regression coefficient for leadership of the CEO is $b_1 = 1.439$, which means that the log odds of making a profit increase as X_2 increases. It

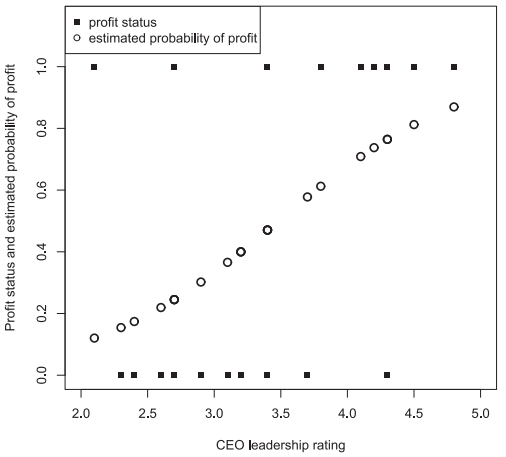


FIGURE 18.1. Profit status and estimated probability of making a profit as a function of CEO leadership ability.

seems the log odds of making a profit are higher in companies run by a CEO perceived to be higher in leadership abilities.

Leadership ability is a continuous scale. This model generates many logits, one for each possible value of X_2. Table 18.2 contains the estimated logits, odds, and probabilities of making a profit for different CEO ratings. Figure 18.1 is a scatterplot of profit status against leadership rating, as well as a plot of the estimated probabilities of making a profit applying equation 18.5. It is apparent that the likelihood of making a profit (whether measured in terms of log odds, odds, or probabilities) goes up as the leadership ability of the CEO increases.

18.1.5 Interpretation of and Inference about the Regression Coefficients

In a logistic regression analysis, the logit is modeled as a linear function of regressors. In models with only a single regressor, b_0 is the estimated logit when the regressor is equal to zero, and b_1 is the estimated difference in the logit between two cases that differ by 1 unit on the regressor. In that sense, logistic regression coefficients are similar to ordinary regression

coefficients. As the logit is positively related to the probability of the event, a positive logistic regression coefficient for a regressor means that the probability of the event increases with increases in the regressor, whereas a negative logistic regression coefficient means that the probability of the event decreases as the regressor increases.

A logistic regression program will produce a standard error of the logistic regression coefficient, as well as some kind of statistic and associated *p*-value for testing the null hypothesis of no association between the probability of the event and the regressor. Some programs display this as a *t*-statistic, others as a *Z*-statistic, both formed as ratios of the logistic regression coefficient to its standard error. Still others report a *Wald* statistic, which is just the square of this ratio. A confidence interval can be constructed in the usual way as the point estimate plus or minus about two standard errors, but this confidence interval usually doesn't have any meaningful interpretation, because the logit scale is not substantively meaningful.

A transformation of a logistic regression coefficient is often applied prior to interpretation. By raising e to the power of a logistic regression coefficient, the result is an *odds ratio*. An odds ratio is just that—a ratio of odds. To understand an odds ratio, consider the odds of making a profit for small and large companies. We saw in section 18.1.4 that the odds of a profit if the company is small are estimated as 0.167. But for large companies, the estimated odds are 9.000. This ratio is 9.000/0.167, which is about 54 if you account for rounding error in doing computations here to only three decimal places. Recall that the regression coefficient for company size was 3.898. And notice that $e^{3.898} = 54$. This is the factor change in the odds of the event as the regressor increases by 1 unit. When $X_1 = 0$ (small company), the odds of a profit is 0.167. When X_1 increases by 1 unit ($X_1 = 1$ for large companies), the odds go up to 9.000. This is a multiplication of the odds by 54.

Applying this to the example model of profit from CEO leadership ability, $b_1 = 1.439$, and $e^{1.439} = 4.2$. So we can say that an increase in 1 scale point in the leadership ability of the CEO is associated with about a 4.2 factor increase in the odds of making a profit. Observe that this is true in Table 18.2. For instance, for a CEO with a 2.0 rating, the estimated odds of a profit is 0.118. But for a CEO with a 3.0 rating, the estimated odds is 0.500. The ratio of these odds is 4.2.

The odds ratio can be interpreted in this manner regardless of what value in the distribution of the regressor serves as the baseline. For instance,

notice in Table 18.2 that for a CEO with 4.0 rating, the odds of a profit are about 4.2 times the odds of a CEO with a 3.0 rating making a profit. But this ratio does not apply to probabilities. The constraints on a probability being between zero and one mean that you can't apply ratio approaches such as this to evaluating the relative sizes of probabilities as the regressor changes.

Confidence intervals can be generated in an odds ratio metric as well. This is accomplished by raising e to the power of the upper and lower bounds of the confidence interval for the logit. The result will be a confidence interval for the true odds ratio.

What would happen if the logistic regression coefficient for a regressor were zero? This would mean that there is no relationship between the regressor and whether or not the event occurs or how likely the event is to occur. If you raise e to the zero power, you get 1.00. That means that a 1-unit increase in the regressor changes the odds of the event by a factor of 1. But a factor change of 1 is not a change at all. If we multiply a number by 1, we just get that number.

An odds ratio has a lower bound of 0 but no upper bound, but a logistic regression coefficient can be negative. If a logistic regression coefficient is negative, then the odds ratio is less than 1. This means that the odds of the event is decreasing as the regressor increases, as is the probability. For example, an odds ratio of 0.8 means that the odds of the event *increases* by a factor of 0.8. But when you multiply a number by something between 0 and less than 1, you get a number smaller than the original number. So an odds ratio of 0.80 translates into a *decrease* in the odds as the regressor increases by 1 unit.

18.1.6 Multiple Logistic Regression and Implementation in Computing Software

Logistic regression routines are available in most all software that conducts ordinary regression analysis. We illustrate code to conduct logistic regression in SPSS, SAS, and STATA in the context of a *multiple* logistic regression model, meaning a model with more than one regressor. Suppose, for instance, we wanted to model the probability of a company making a profit from both its size and the perceived leadership ability of the CEO. A logistic regression command in SPSS to conduct such an analysis looks very much like an ordinary regression command:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	19.413	2	.000
	Block	19.413	2	.000
	Model	19.413	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	13.691[a]	.555	.741

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	size	4.624	1.718	7.244	1	.007	101.872
	ceo	1.790	.962	3.459	1	.063	5.989
	Constant	-8.149	3.799	4.600	1	.032	.000

a. Variable(s) entered on step 1: size, ceo.

FIGURE 18.2. SPSS output from a multiple logistic regression analysis.

```
logistic regression profit/method=enter size ceo.
```

The equivalent command in SAS is

```
proc logistic data=leader descending;
   model profit=size ceo/rsquare;
run;
```

and in STATA, use

```
logit profit size ceo
```

Figure 18.2 contains an abbreviated version of the output from SPSS. Before discussing this model, it is worth pointing out a feature of SAS that can be confusing if you aren't aware of it. In SPSS and STATA, the numerically largest value or code in the data for the dependent variable will be chosen as the "event." In this case, with no profit coded 0 and profit coded 1, making a profit is the event. But SAS defaults to the *lowest* value as the event. This means that the logistic regression coefficients and regression constant will have the opposite signs in SAS than what SPSS and STATA give, *unless* you specify **descending** as an option, as in the command above. This tells SAS to treat the numerically highest code on the dependent variable as the event, rather than the other way around.

Under the output column labeled "B" can be found the logistic regression coefficients and constant. As can be seen, the model is

$$\text{Estimated logit} = -8.149 + 4.624X_1 + 1.790X_2$$

Using this model, the estimated odds and probabilities of a company making a profit can be derived for a company of a given size (X_1) whose CEO is given a certain leadership rating by the board (X_2). For example, for a large company with a CEO rated 2.0 on leadership, the estimated odds of making a profit are

$$e^{-8.149+4.624(1)+1.790(2)} = e^{0.055} = 1.057$$

and the estimated probability is

$$\frac{e^{0.055}}{1 + e^{0.055}} = 0.514$$

The logistic regression coefficients for size and sex are both positive and statistically significant. These are *partial* logistic regression coefficients. Each variable in the model serves as a statistical control when assessing the partial association between X_j and the probability of event. So we can say that *holding the perceived leadership ability of the CEO constant*, or *controlling for CEO leadership ability*, large companies are more likely to make a profit than small companies, because b_1 is positive and statistically different from zero. Further, we can say that the odds of a large company making a profit are $e^{b_1} = e^{4.624} = 101.900$ times the odds of a small company making a profit (and note that this odds ratio is printed in the SPSS output under the column labeled "Exp(B)" and is computed with a bit more precision than our hand computation). But we can't say the same thing in probability terms, because the difference in the probabilities of making a profit for two companies different in size but run by a leader with the same perceived leadership ability will depend on the specific value of leadership chosen when making the comparison. The estimated probability of the large company will always be larger than the estimated probability of the smaller company, but we can't say just how much larger without choosing a value of perceived leadership ability.

We can also say that holding company size constant, companies run by CEOs perceived as better leaders are more likely to make a profit. More specifically, we can say that holding company size constant, two companies run by CEOs who differ by 1 unit in perceived leadership abil-

ity are estimated to differ in the odds of making a profit by a factor of $e^{b_2} = e^{1.790} = 5.989$. That is, the company run by the CEO 1 unit higher in perceived leadership ability is estimated to be almost six times more likely to make a profit. But as in the prior paragraph, in this statement, *likely* is a reference to the relative *odds* and not probabilities. We can't make such a statement about probabilities. To talk about such a difference in probability terms, you have to pick a company size and two values of CEO rating and do the computations to get the estimated probabilities for these specific cases.

18.1.7 Measuring and Testing the Fit of the Model

In section 18.1.1 we saw that when using the multiplicative law of probabilities, we can quantify the fit of a set of estimated probabilities to the observed data, and therefore the fit of a model that generates those probabilities. Logistic regression generates an estimated probability of the event, and the data include whether or not the event occurred for each case. So a simple measure of fit is the Pearson correlation between the observed values and the estimated probabilities. For a perfectly fitting model, the correlation between PE and Y would be 1. In the two-regressor model of the probability of a company making a profit in section 18.1.6, this correlation is 0.826.

Although this seems intuitively reasonable, most regression programs generate measures of fit that rely on the likelihood of the data. These measures all rely on a comparison between the likelihood for the observed model and a model that contains no regressors—the so-called *constant-only model*. To understand this, imagine in this example that you didn't know a company's size or the perceived leadership of its CEO. In that case, if you wanted to estimate any randomly selected company's probability of making a profit, your best guess would be the proportion of companies in the data available that made a profit. In this case, from the data in Table 18.1, you can see that 11 of the 24 companies made a profit. So $11/24 = 0.458$ is a reasonable estimate of the probability of making a profit if you had no other information available to you.

We could translate that probability into an odds using equation 18.4, which would result in an odds of making a profit of 0.846. Translated yet again into a logit, which is the natural logarithm of the odds, we have −0.167. Thus, the constant only model of the probability of making a profit would be

$$\text{Estimated logit} = -0.167$$

This simple model has a likelihood, which can be generated using equation 18.2 and then converted into $-LL$ with equation 18.3 or to $-2LL$ by multiplying $-LL$ by two. In this example, $-LL = 16.552$ and $-2LL = 33.104$. To ease the discussion that follows, rather than referring to both values, we will use the $-2LL$ version and denote it $-2LL_c$, where the "c" subscript denotes "constant-only." So $-2LL_c = 33.104$.

Now consider the model of the probability of making a profit that includes the size of the company and the CEO's leadership rating. In this example $-2LL = 13.691$. In SPSS, $-2LL$ for the model is found in the section of output under the heading "-2 Log Likelihood." As can be seen, SPSS reports 13.691. We will denote this $-2LL_m$, so $-2LL_m = 13.691$. Most programs will display one of these statistics for the model being estimated. Some will display both $-2LL_m$ and $-2LL_c$ in the same output.

There are two things that we can do with this information, keeping in mind that for a perfectly fitting model, $-LL = -2LL = 0$. Since a better-fitting model has $-2LL$ closer to zero, and we know that a model that excludes all regressors has a $-2LL = -2LL_c$, we can ask how close $-2LL_m$ is to 0 relative to $-2LL_c$. If you think of $-2LL_c$ as a distance from zero, this question can be phrased as how much of the distance between fit of the constant-only model and a hypothetical perfectly fitting model has the observed model "traveled?" This ratio is called the McFadden R^2. It is defined as

$$\text{McFadden } R^2 = \frac{-2LL_c - (-2LL_m)}{-2LL_c}$$

which in this example is 0.586. This is sometimes called a "pseudo-R^2" because it is often interpreted like a squared multiple correlation between the observed probabilities and the actual values of Y (which are 0 and 1 in the data), but it isn't quite the same thing as a squared multiple correlation in reality. In fact, the square of the correlation between PE_i and Y_i in this example is $0.846^2 = 0.682$.

Attempts have been made to improve the McFadden R^2. One of these is the Cox and Snell R^2, defined as

$$\text{Cox and Snell } R^2 = 1 - e^{-[-2LL_c-(-2LL_m)]/n}$$

which in this example is 0.555. But a true R^2 can be 1, yet it can be shown that the Cox and Snell measure can't quite achieve 1 in some situations. A

correction for this is the Nagelkerke R^2, which is the Cox and Snell measure divided by its maximum possible value:

$$\text{Nagelkerke } R^2 = \frac{\text{Cox and Snell } R^2}{1 - e^{-(-2LL_c)/n}}$$

which in this example is 0.741.

These three measures of fit can differ quite dramatically, as can be seen in this example. Complicating matters, there are many other measures of fit that will produce still different values than these. Different logistic regression programs will display none, some, or all of these measures, and they will not always label them the same. As can be seen in Figure 18.2, SPSS provides the Cox and Snell and Nagelkerke R^2 measures but not the McFadden R^2.

A second use of $-2LL_c$ and $-2LL_m$ is for testing a hypothesis about the fit of a model. We saw in section 4.3.2 in ordinary regression that if there is no relationship between any of the regressors and Y in the population, this implies that all true regression coefficients for the regressors are equal to zero or, alternatively, that ${}_TR = 0$. An F-test was described in that section for testing this hypothesis.

A comparable test exists in logistic regression. Under the null hypothesis that all of the true logistic regression coefficients equal zero (meaning the observed model fits no better than the constant-only model), $-2LL_c - (-2LL_m)$ follows a χ^2 distribution on k degrees of freedom, where k is the number of regressors. This is called a *likelihood ratio test*. Most logistic regression programs will display this test. In SPSS, it can be found in the section titled "Omnibus Tests of Model Coefficients" in the row labeled "Model." In the two-regressor example we've been discussing, $\chi^2(2) = 19.413, p < .001$, where $\chi^2 = -2LL_c - (-2LL_m)$. We can reject the null hypothesis. There is a relationship between the probability of making a profit and either company size, leadership of the CEO, or both. That is, this model fits better than you would expect "just by chance."

In section 5.3.3 we discussed a test for sets of regressors. That test was used to determine whether adding one or more variables to a regression model improves its fit. In logistic regression, the likelihood ratio test just described can also be used to test whether adding additional regressors to an existing model improves its fit. Suppose you are estimating a dichotomous variable from three regressors X_1, X_2, and X_3. Call the $-2LL$ from that model $-2LL_A$, for "model A." You want to know if adding X_4, X_5, and X_6 to model A significantly improves the fit of the model. Call this

model B, which has $-2LL_B$. In this example, $-2LL_A - (-2LL_B)$ follows χ^2 distribution on three degrees of freedom if the variables added to model A to produce model B don't improve the fit. This is equivalent to a test of the null hypothesis that the true logistic regression coefficients for all the variables added to model A to produce model B are equal to zero. More generally, if you add k regressors to model A to produce model B, then under the null hypothesis of no partial relationship between any of the k regression coefficients added and the outcome when holding the variables in model A constant, the difference in the $-2LL$ is distributed as χ^2 on k degrees of freedom. Most logistic regression programs can conduct this test and generate a p-value, or it can be done manually by estimating the two models and comparing the difference in the $-2LL$ values to entries in a table of critical values of $\chi^2(k)$, such as in Appendix C.

18.1.8 Further Extensions

In this book we have covered many topics in linear regression analysis across 17 chapters, and we could write equivalent chapters on these various topics for logistic regression. Many of the same methods described in this book have equivalent versions in logistic regression. For example, multicategorical variables can be included as regressors in a logistic regression model using any of the coding schemes discussed in Chapters 9 and 10. Nonlinear models such as those described in Chapter 12 can be fit to dichotomous dependent variables using logistic regression, as can models with interactions using the techniques in Chapters 13 and 14. And there are many diagnostic statistics in logistic regression similar to those discussed in Chapter 16 that can be used for finding irregular cases, quantifying influence, and testing model assumptions. There are entire books dedicated to logistic regression that discuss these topics, and with the background you now have in linear regression and this brief introduction to logistic regression, you should be prepared to tackle any of these topics, covered in such sources as Allison (2012), Hosmer, Lemeshow, and Sturdivant (2013), Jaccard (2001), Long (1997), Menard (2002), and Pampel (2000).

18.1.9 Discriminant Function Analysis

Discriminant function analysis is an older statistical technique that has largely been replaced by logistic regression, which has fewer requirements and assumptions. But you may need to read research that used discriminant function analysis, so a few words about it seem appropriate here.

Discriminant function analysis is used when the dependent variable Y is dichotomous, but it assumes all regressors are continuous. Technically, it assumes that within each group of cases defined by a common value of Y, the distribution of all the regressors is multivariate normal, though the tests are reasonably accurate even if that assumption is not exactly true. It assumes that the regressor scores have the same pattern of standard deviations and correlations in the two Y groups. Discriminant function analysis can be used to make many of the same kinds of inferences as can be done with logistic regression, such as calculating for each person the probability of the event. It turns out that the significance tests on partial and total relationships in discriminant function analysis are exactly the same as those found when an ordinary regression program (not logistic regression) is applied to a dichotomous Y. Discriminant function analysis computes a linear function of the regressors called a *discriminant function* and derives a curvilinear relationship between that function and the probability of the event. That discriminant function turns out to correlate perfectly with the $\hat{Y}$ values you would find from applying ordinary regression analysis using the dichotomous Y as the dependent variable. Those $\hat{Y}$ values cannot be interpreted as probabilities since they may be below 0 or above 1. But they nevertheless correlate perfectly with the discriminant function, which can be combined with a curvilinear formula to find probabilities.

18.1.10 Using OLS Regression with a Dichotomous *Y*

With a dichotomous variable Y, the conditional distributions of Y cannot be normally distributed, because Y can have only two values. Thus, if we used ordinary regression to estimate Y from one or more other variables, the assumption of normality would be badly violated. This raises questions about the accuracy of any statistical inferences you can make using linear regression analysis with a dichotomous dependent variable, such as significance tests on individual regressors or on the regression model as a whole. However, we mentioned in section 4.1.2 that the central limit theorem says that the larger the sample size the less important it is that conditional distributions be normal. This suggests that with large sample sizes, these inferences might be valid. Also, suppose a researcher's primary interest is in the relationship between Y and one continuous independent variable X_1, and the other variables in the model are covariates. In that case, the one inference of most interest concerns the partial relationship between Y and X_1. But if we are interested in the partial relationship between two variables, the test is the same test regardless of which one is the depen-

dent variable. If we instead thought of X_1 as the dependent variable and Y as a regressor, we would be studying the partial relationship between a continuous dependent variable and a dichotomous independent variable, and we know that regression includes no requirement that independent variables be normally distributed. Thus, our test of interest might be reasonably valid even with a fairly small sample. And even if both X_1 and Y are dichotomous, the ordinary regression test is pretty accurate with larger samples. The same is true for the test on the regression as a whole. Thus, even when Y is dichotomous, the most common inferences in ordinary regression are often reasonably valid.

Suppose you have used ordinary regression analysis many times, and you are thinking of conducting a logistic regression analysis for the first time. The preceding argument suggests that applying your ordinary regression analysis to your data can give you a pretty good idea of the significance of the regression as a whole and of individual partial relationships of interest. If those relationships turn out to be nonsignificant and with large p-values, then maybe you don't need to bother learning logistic regression analysis right now. But if they are significant, you may want to learn about and run the logistic regression analysis to confirm it.

18.2 Other Linear Modeling Methods

Although linear regression analysis is a very versatile "data analytic system" (Cohen, 1968), it can't be used for every data problem you confront. Yet the idea of modeling a dependent variable as a weighted sum of regressors underlies may other statistical methods, and so the researcher familiar with regression analysis can often use these other methods with only a little bit of extra study. In this last section of the book, we offer brief descriptions of some of these methods, along with some references you can explore to further educate yourself.

18.2.1 Ordered Logistic and Probit Regression

Ordinary regression analysis is typically the method of choice when the dependent variable is continuous or at least numerical, with many possible values, and the measurement scale is interval in nature. We saw in section 18.1 that when the dependent variable is dichotomous and so has only two possible values, logistic regression is appropriate. Between these two extremes are discrete, ordinal dependent variables. Researchers often ask participants to respond to a single question used as a dependent variable

that has only a few response options and for which the response options scale a quantity at only the ordinal level. Examples include level of agreement (e.g., *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*), evaluation (e.g., *poor*, *fair*, *good*, *excellent*), or frequency (e.g., *never*, *a few times*, *usually*, *always*). Or a clinical psychologist might evaluate whether a patient has improved since the start of therapy, not changed, or worsened. Although it is not difficult to find people analyzing dependent variables such as these with ordinary regression, and there is some debate about and research on its appropriateness (see, e.g., Noreen, 1988; Taylor, West, & Aiken, 2006), methods are available designed for dependent variables like these.

There are extensions of logistic regression that can be used for discrete ordinal dependent variables. They take many forms and collectively go by the name *ordinal logistic regression*, but all rely on the idea of modeling the odds of the dependent variable taking one value rather than some other value that is ordinally higher on the scale. There are many ways this could be done. For instance, with a 3-point ordinal scale, you could simultaneously model from one or more regressors the odds of giving ordinal response 1 relative to response 2 or 3, as well as the odds of giving ordinal response 2 relative to 3. Or you could model the odds of giving response 1 rather than 2 and also the odds of giving response 2 rather than response 3. By making certain assumptions, you can use a single logistic regression coefficient to estimate how the odds of making an ordinally higher response relative to a lower response as a regressor increases.

Ordinal regression doesn't make any assumptions about the distribution of the ordinal dependent variable. If you are willing to assume that the ordinal response categories reflect an underlying continuum that is normally distributed, then *probit regression* can be used. If you think about it, many ordinal scales are ordinal not because the underlying variable is ordinal but because the researcher has forced a continuum into a set of ordinal categories. For instance, if I ask you to evaluate the quality of a restaurant using a 4-point ordinal scale, you are forced to translate your feelings into one of the four response options. But people's beliefs probably don't follow a discrete distribution but, rather, an underlying continuum. It is only the measurement process that has made the response ordinal.

Once you feel comfortable with logistic regression, you should have no trouble expanding your knowledge of linear modeling methods to ordinal logistic or probit regression. There are several good books that cover these and related methods, including Borooah (2001), Long (1997), and O'Connell (2005). A single and very readable chapter on ordinal logistic regression

can be found in Orme and Combs-Orme (2009), and some journal articles include Ananth and Kleinbaum (1997) and Valenta, Pitha, and Poledne (2006).

18.2.2 Poisson Regression and Related Models of Count Outcomes

A variable is a count if the smallest possible value is zero and only integers can be observed. Examples include the number of times a person reports having donated to a political campaign in the last year, how many computers a person owns, or how many symptoms of a particular psychological disorder a person experiences. When a count variable is the dependent variable in an analysis, sometimes it is acceptable to use regular regression analysis, but in some circumstances ordinary regression analysis is not appropriate. For example, consider our measure of political knowledge used in the analysis reported in Chapter 12. Strictly speaking, that is a count variable, because it was formed by adding up the number of questions that a person was able to answer correctly. Yet most researchers wouldn't be too concerned about using ordinary regression for this, because the number of things being counted is fairly large (there were 22 questions on the knowledge test) and relatively few of the observations were on the very bottom or top end of the scale (the mean was 11.3, and 80% of the measurements were between 6 and 17).

But often it is the case that most of the observations are at the very bottom or the very top of the distribution, and only a few of the possible values are observed. For example, if you asked people how many car accidents in the last 2 years they have been in, most people would say zero, some would say one, very few would say two or three, and you aren't likely to find anyone who has been in four or more car accidents in the last 2 years unless he or she is in the business of crash-testing cars. A variable like the number of children a person has would probably be similar, with many zeros, ones, and twos, some threes, fewer fours, and relatively few fives and higher. On the other end of the spectrum, if you asked a child during the school year how many days in the last week he or she went to school, almost all would say 5 days, a few might say 4, fewer would say 3, and almost no children would report not having gone to school at all that week.

Modeling dependent variables that are counts such as these, where there are relatively few values observed and most cluster toward the bottom or top end of the distribution, ordinary regression analysis is not the best,

and sometimes not even a particularly good analytical choice. Statistical methods based on the linear model are designed especially for modeling discrete variables that are counts such as this. Examples include *Poisson regression* and *negative binomial regression*, two methods that are very closely related. Long (1997) offers a few chapters on this topic; an entire book on the regression of count outcomes is available by Cameron and Trivedi (2013); and see Coxe, West, and Aiken (2009) and Gardner, Mulvey, and Shaw (1995) for readable journal-article-length treatments. A very readable book chapter on these methods can be found in Orme and Combs-Orme (2009).

18.2.3 Time Series Analysis

Time series analysis is superior to regression analysis for some inferential purposes when the units of analysis are measured over many sequential time periods, such as weeks, months, or years. The example on changes in the population of the United States in section 2.4.4 illustrates this kind of data. Time series data are particularly challenging, because they usually violate the assumption of independence when ordinary regression analysis methods are used. Typically, there is a component of a measurement at time t that can be predicted from values of the dependent variable taken at time $t - 1$, $t - 2$, and so forth, and this makes it challenging to get good estimates of standard errors of regression coefficients.

The most popular time series method is based on something called the *autoregressive integrated moving average* model, or *ARIMA* for short. ARIMA modeling is quite common in economics and other fields in which data are collected at regular intervals over extended periods of time, such as economic indicators, company performance measured quarterly, and so forth. But you can find examples of time series analysis in other fields such as public health, criminology, and virtually anywhere else that relies on statistics collected over time. Although ARIMA modeling is the standard for modeling time series data, regression methods can be used with some forms of time series data as well. There are many good books on the topic of time series analysis from different approaches (e.g., Box, Jenkins, & Reinsel, 2008; Enders, 2009; Milhoj, 2013; Ostrom, 1990; Yaffee & McGee, 2000).

18.2.4 Survival Analysis

Survival analysis, also called *event history analysis*, is often used when the dependent variable is the time or date at which an event occurs. An example

would be a person's death, when he or she is first married, how long after completion of therapy a person relapses, or something less dramatic, such as how long it takes a person to complete a task. One of the problems with analyzing time to event is that sometimes the event doesn't occur. For instance, your dependent variable might be how many minutes it takes a person to solve a moderately complex mathematics problem. But some people may not actually finish the problem. In such cases, what should be used as the value for the dependent variable? If you truncate it at some value such as the amount of time elapsed before the participant was not allowed to continue trying, then no distinction is being made between those who actually do finish it in that time, those who might have eventually finished it if given more time, and those who would never complete it regardless of the amount of time they were given. This complicates time to event analysis using ordinary regression methods.

Unlike in regression analysis, survival analysis allows for the possibility that the event doesn't occur for a particular case in the data set. For such cases, there is no time of event, but they can still be used in the analysis. Their dependent variable is called *censored*. For some kinds of problems involving time, logistic regression can be used, such as when time is measured discretely (e.g., whether or not the event occurred in one of a few predetermined time periods). Otherwise, more specialized procedures such as *Cox regression* are required. See Allison (2010, 2014), Box-Steffensmeier and Jones (2004), and Hosmer, Lemeshow, and May (2008) for book-length introductions to these methods, or individual chapters in Singer and Willett (2003). A four-journal article series on survival analysis can be found in the *British Journal of Cancer* (Bradburn, Clark, Love, & Altman, 2003a, 2003b; Clark, Bradburn, Love, & Altman, 2003a, 2003b).

18.2.5 Structural Equation Modeling

Linear regression analysis can be thought of as a special form of *structural equation modeling* (SEM), also known as *covariance structure modeling*. SEM is more of a multifunctional tool, like a Swiss army knife, than a single distinct method. Methods that can be thought of as forms of SEM include simple path analysis, as discussed in Chapter 15, confirmatory factor analysis, latent growth modeling, latent class analysis, and many others. An understanding of regression analysis is fundamental to mastering its more general version as SEM.

Three features of SEM are worth noting that make it distinct from regression and other methods discussed above. First, in ordinary regression

analysis, there is only a single dependent variable being modeled. But in SEM, you can model many dependent variables at once, and some of those dependent variables can simultaneously be functioning as independent variables or covariates in the same analysis. This allows for sets of relationships to be modeled all at once to better understand complex processes, including those that evolve over time.

Second, recall from section 17.2 that in ordinary regression analysis, measurement error in independent and dependent variables can lower power, and measurement error in covariates can bias the estimation of regression coefficients. But SEM offers a means of dealing with problems due to measurement error when variables are modeled as latent rather than observed. In latent variable SEM, a variable can be represented by a set of *indicators* that is modeled as caused by the latent variable. Examples of indicators would be responses to specific questions on a test or attitude survey. Rather than using a sum or average of responses to a set of questions, the covariation between the indicators is modeled as being caused by a latent variable not actually observed in the data. The latent variables can then be tied together in a linear model to examine how variables relate to one another.

Third, whereas ordinary regression defines the "best" linear model as the one that minimizes $SS_{residual}$, in SEM, you can choose from many definitions of best, some of which are better suited to certain kinds of problems than others. Unlike in ordinary regression analysis (and like logistic regression, though we didn't discuss this earlier), an SEM program may not return the best solution for a given definition of best. That is because software that does SEM attempts to find the solution iteratively, by proposing one solution, quantifying its fit, and then modifying the solution repeatedly until further modification no longer improves fit by more than some specified amount. But when such an iterative procedure is used, there is no guarantee that the solution it prints is the best available.

Bollen (1989) is the classic but technical treatment of SEM. There are many other books that are more applied in orientation (e.g., Kline, 2015; Raykov & Marcoulides, 2006). Others worth exploring are dedicated to the implementation of SEM in specific software programs such as Mplus and AMOS (see, e.g., Byrne, 2009, 2012; Geiser, 2012).

18.2.6 Multilevel Modeling

It is common in some research for data to be collected from people or other units that are "nested" in some fashion under different higher-level

research units. For example, employees at a company are often organized into certain functional areas, such as manufacturing, marketing, human resources, and so forth. And school-age children reside in classrooms, and classrooms reside in schools, which reside in districts. Or when you measure somebody repeatedly over time on one or more variables, each measurement at a given time point is nested under the person providing the data.

Multilevel linear models are designed to properly handle problems that are produced by nonindependence when such nesting exists in one's data and there is the possibility that values on the dependent variable may differ systematically across higher-level organizational units. Multilevel modeling can also be used to answer questions about how variables at one level (e.g., the number of children in a classroom) influence the effects of variables at other levels (e.g., how the sex of the child in that classroom relates to his or her performance). This is because in multilevel analysis, the data analyst is not forced to fix one variable's effect to be the same across all higher-level units.

To illustrate this idea, suppose a cognitive psychologist is interested in examining how people come to understand certain concepts like "chair." The researcher might show a person 50 photographs of pieces of furniture and ask the person to classify it as a chair or not, or to evaluate how "chair-like" the piece is. Each piece of furniture may be quantified on a variety of variables, such as size, number of legs, squareness, puffiness, and so forth. In this example, the furniture pieces are *level-1 units*, and these features of the furniture are *level-1* regressors. And the dependent variable of judgment of chair-likeness is also a level-1 variable. *Level-2* regressors might include the age of the participant, his or her sex, and perhaps his or her score on a cognitive abilities test. Using multilevel modeling, we could simultaneously examine the relationship between the dependent variable and both level-1 and level-2 regressors, while also allowing the values of level-2 regressors to influence the effect of level-1 regressors on the dependent variable.

If there were 20 participants in this study, each of which gives 50 responses (one for each piece of furniture), we can't just treat these like 1,000 independent data points. Doing so would ignore the fact that people probably differ tremendously from each other in how "chair-like" they perceive furniture as being. This variation needs to be accounted for when modeling responses to the dependent variable from the regressors such as size, number of legs, sex of the participant, his or her cognitive ability, and so forth.

Or it could be that the relationship between a feature, such as number of legs, is related to a judgment such as chair-likeness differently for different types of people. Maybe the relationship between the regressor of number of legs and the dependent variable is larger among people who are younger. This might remind you of the concept of interaction discussed in Chapters 13 and 14. If age (a level-2 variable) affected the relationship between number of legs and judgments of chair-likeness (both level-1 variables), in multilevel modeling this is called a *cross-level interaction*.

One important feature of multilevel modeling is the fact that the data analyst has the flexibility to decide whether the effects of level-1 variables on the dependent variable are fixed to be the same across level-2 research units (called a *fixed effect*), or to vary randomly (called a *random effect*). Alternatively, the analyst could allow the effect of level-1 variables to depend on level-2 variables, as in the example just given.

Multilevel modeling procedures are built into many statistical packages such as SPSS, SAS, and STATA. There are also programs dedicated to multilevel modeling such as HLM. And some programs often used for SEM, such as Mplus, can also estimate multilevel models; some forms of multilevel analysis can be set up as a structural equation model. Multilevel analysis is, like SEM, a very versatile analytical tool, and you can spend years studying it and only begin to thoroughly understand it. Good books on the topic include the classic but more technical one by Raudenbush and Bryk (2002). Others that are more applied in orientation include Bickel (2007), Hox, Moerbeek, and Schoot (2002), and Luke (2004). Some broad conceptual journal articles to get you started on multilevel modeling using SPSS and SAS are available in Hayes (2006) and Singer (1998).

18.2.7 Other Resources

Learning new methods sometimes is made easier when it is structured in the form of a class or seminar. Most universities offer courses in many of the topics described here. If you are no longer a student or you don't have access to such courses at your own university or one nearby, there is an entire industry dedicated to giving training to researchers in short workshops or seminars lasting anywhere from a few hours to a few days to a week. Providers come and go, but as of the writing of this book, you will find a variety of courses on these and other topics offered by Statistical Horizons (*www.statisticalhorizons.com*), StatsCamp (*www.statscamp.org*), and the Global School in Empirical Research Methods (*www.gserm.ch*), among many others.

18.3 Chapter Summary

Linear modeling is far more versatile than what has been covered in the first 17 chapters of this book. In this last chapter we briefly introduced the application of the principles of linear modeling to the analysis of a dichotomous dependent variable using logistic regression analysis. In logistic regression, the focus is on estimating the relationship between one or more regressors and the likelihood of an event occurring. Although we often think about "likelihood" in probability terms, logistic regression instead models the log odds or logit of the event as a linear function of regressors. Once the model of the logit is constructed, the model can be used to produce estimates of the probability of event for any combination of regressor scores. Using logistic regression, it is possible to assess the relationship between one regressor while holding others constant, or to test the contribution of a set of regressors to predicting the probability of the event when variables in another set are held constant.

Logistic regression is only one variant on linear modeling. Linear models have many uses, and many statistical procedures that go by different names and are used for various purposes are based on the linear model. With an understanding of the fundamentals of linear regression analysis covered in this book, you are now in a position to tackle some of its extensions, including logistic and ordered logistic regression, survival analysis, probit regression, time series analysis, Poisson and negative binomial regression, multilevel modeling, and covariance structure modeling, among many other methods.

APPENDICES

Appendix A
The RLM Macro for SPSS and SAS

RLM is a macro that conducts ordinary least squares regression analysis. It is not intended to replace SPSS's REGRESSION module or PROC REG or PROC GLM in SAS; RLM is actually quite limited relative to what is built into SPSS or SAS. But RLM has a few regression analysis features you will not find in SPSS and SAS already, including some routines for probing interactions, coding categorical variables, spline regression, dominance analysis, testing assumptions, and inference without assuming homoscedasticity. It makes simple some tasks that are quite complicated or tedious in SPSS REGRESSION or SAS PROC REG and PROC GLM. It should be used as a supplement to rather than a replacement for what comes with SPSS and SAS.

This appendix describes how to install and execute RLM, how to set up an RLM command, and it documents its many features. As RLM is modified and features are added, supplementary documentation will be released at *www.afhayes.com*. This documentation focuses on the SPSS version of RLM. All features and functions described below are available in the SAS version as well and work as described here, with minor modifications to the syntax. A section devoted to SAS (see page 599), describes some of the differences in syntax structure for the SAS version of RLM compared to what is described below.

Preparing for Use

RLM can be used as either a command-driven macro or installed as a custom dialog for setting up the model using SPSS's point-and-click user interface. When executed as a macro, the RLM.sps file (available from *www.afhayes.com*) should first be opened as a syntax file. Once it has been opened, execute the entire file exactly as is. ***Do not modify the code at***

all. Once the RLM.sps program has been executed, it can be closed and the RLM command is available for use in any SPSS program. Running RLM.sps activates the macro, and it will remain active so long as SPSS remains open. The RLM file must be loaded and reexecuted each time SPSS is opened in order to use the features of the RLM command. See the "Examples" section starting on page 584 for how to set up an RLM command in a syntax window.

To install RLM as a custom dialog into the SPSS menus, execute RLM.spd (available from *www.afhayes.com*) by double-clicking it on the desktop (SPSS 23 or earlier) or opening and installing it from within SPSS under the Utilities (SPSS 23 or earlier) or Extensions → Utilities (SPSS 24 and later) menu. Administrative access to the machine on which RLM is being installed is required when using a Windows operating system, and you must execute SPSS as an administrator. Once successfully installed, RLM will appear as a new menu item in SPSS nested under Analyze → Regression. If you do not have administrative access, contact your local information technology specialist for assistance in setting up administrative access to the machine on which you wish to install RLM. Some options available in the macro cannot be accessed through the dialog box. Installing the dialog box does not automatically run RLM.sps. If you wish to use the command syntax to execute an RLM command rather than the dialog box, you still need to first load and execute RLM.sps.

Syntax Structure

RLM along with a variable following **y=** and at least one following **x=** are required. Commands in brackets are optional. Brackets, parentheses, and asterisks should not be included in the RLM command. "**" Denotes the default argument when the option is omitted.

```
rlm y=yvar/x=xvarlist [/conf=ci(95**)]
                      [/stand=(0**)(1)]
                      [/covcoeff=(0**)(1)]
                      [/mod=(0**)(1)]
                      [/mcx=(1)(2)(3)(4)(5)(6)]
                      [/mcmod=(1)(2)(3)(4)(5)(6)]
                      [/mcfoc=(1)(2)(3)(4)(5)(6)]
                      [/center=(0**)(1)]
                      [/ptiles=(0**)(1)]
                      [/modval=mval]
```

```
[/jn=(0**)(1)]
[/plot=(0**)(1)]
[/subsets=(0**)(1)]
[/dominate=(0**)(1)]
[/diagnose=(0**)(1)]
[/crossv=(0**)(1)]
[/spline=joint1,joint2,...]
[/contrast=weight1,weight2,...]
[/hc=(0)(1)(2)(3)(4)]
[/settest=nvar(0**)]
[/decimals=dec(F10.4**)].
```

Model Specification

An RLM command has only two *required* arguments:

- A single quantitative outcome variable *yvar* listed in the **y=** specification (i.e., **y=***yvar*), where *yvar* is the name of the variable in the data functioning as the dependent variable Y in the model.
- At least one regressor *xvarlist* listed in the **x=** specification (i.e., **x=***xvarlist*), where *xvarlist* is the name of one or more variables in the data file. With the exception of the last or second to last variable, all variables in *xvarlist* must be dichotomous or a quantitative variable with interval-level scaling properties. For multicategorical variables, see the section below.

With these minimum specifications, RLM will conduct an ordinary least squares regression estimating *yvar* from the variables in *xvarlist* and generate output such as the model R, adjusted R, standard error of estimate, regression coefficients with standard errors, t- and p-values, and confidence intervals, regression ANOVA summary table, and the zero order, partial, and semipartial correlations. Various output and test options can be specified but are not required. Most options are toggled on or off with a 0 (off) or 1 (on) as the argument for the option. Some options require an argument other than 0 or 1, as described in various places in this documentation.

Although RLM has a number of error-trapping routines built in, it will not catch all errors produced by improper formatting of an RLM command, improper listing of variables and variable names, and so forth. Any errors it has trapped will be displayed in an errors section of the RLM output.

Errors it has not successfully trapped will appear as a long list of execution errors that will be largely unintelligible.

RLM also has no features to detect singularities in the data matrix. Singularities will generally appear in the output as a matrix inversion error. When such an error appears, do not interpret any output. Some ways of detecting singularities in a data matrix are are discussed in section 17.3.3.

Examples

```
rlm y=newlaws/x=media age alcuse neuro discuss/settest=3/dominate=1
    /subsets=1/conf=90.
```

- Regresses *newlaws* on *media*, *age*, *alcuse*, *neuro*, and *discuss*.
- Conducts a test that the regression coefficients for *alcuse*, *neuro*, and *discuss* are all zero. (**settest=3**)
- Conducts a dominance analysis. (**dominate=1**)
- Generates the multiple correlation for all subset models containing at least one of the regressors. (**subset=1**)
- Generates 90% confidence intervals for all regression coefficients rather than 95% intervals. (**conf=90**)

```
rlm y=votes/x=donate winner partyid/mcx=2/hc=3/covcoeff=1/stand=1.
```

- Regresses *votes* on *donate*, *winner*, and *partyid*.
- Specifies *partyid* as a multicategorical variable and uses sequential coding of groups. (**mcx=2**)
- Employs the HC3 standard error estimator of the regression coefficients for inference. (**hc=3**)
- Prints the variance–covariance matrix of the regression parameter estimates. (**covcoeff=1**)
- Provides the standardized regression coefficients. (**stand=1**)

```
rlm y=mathprob/x=gender explms treat/mod=1/jn=1.
```

- Regresses *mathprob* on *gender*, *explms*, and *treat* and the product of *explms* and *treat*. (**mod=1**)
- Implements the Johnson–Neyman technique for finding regions of significance of the effect of *explms* on *mathprob* conditioned on *treat*. (**jn=1**)

```
rlm y=know/x=educ attn sex age elab/mcfoc=1/decimals=F12.6/modval=4
   /plot=1/hc=4.
```

- Regresses *know* on *educ*, *attn*, *sex*, *age*, *elab* and the product of *age* and *elab*.
- Specifies that *age* is a multicategorical variable and uses indicator coding to represent groups. (**mcfoc=1**)
- Generates a test of the effect of *age* on *know* when *elab* is equal to 4. (**modval=4**)
- Allocates 12 characters to display of numbers and displays six decimals places of accuracy after the decimal. (**decimals=F12.6**)
- Produces a table of estimated values of *know* for various combinations of *age* and *elab*. (**plot=1**)
- Employs the HC4 standard error estimator of the regression coefficients for inference. (**hc=4**)

```
rlm y=happy/x=commit close desire/spline=3,6,12.
```

- Conducts a spline regression analysis, estimating *happy* from *commit*, *close*, and *desire*.
- The spline segments are based on values of *desire*, with joints at the points 3, 6, and 12 on the measurement scale. (**spline=3,6,12**)

```
rlm y=turnout/x=frame euskept peffic risk/mod=1/center=1/ptiles=1.
```

- Regresses *turnout* on *frame*, *euskept*, *peffic*, *risk*, and the product of *peffic* and *risk*. (**mod=1**)
- *peffic* and *risk* are mean-centered prior to analysis. (**center=1**)
- Generates the conditional effect of *peffic* on *turnout* at values of *risk* corresponding to the 25th, 50th, and 75th percentiles of the distribution of *risk*. (**ptiles=1**)

```
rlm y=jobsat/x=calling livecall/mcmod=3/center=1/plot=1.
```

- Regresses *jobsat* on *calling*, and *livecall* and a set of variables to estimate the interaction between *calling*, and *livecall*.
- Specifies *livecall* as a multicategorical moderator variable and employs Helmert coding of groups. (**mcmod=3**)
- Produces a table of estimated values of *jobsat* for various combinations of *calling* and *livecall*. (**plot=1**)
- Mean-centers *calling* prior to analysis. (**center=1**)

Standardized Regression Coefficients

Standardized regression coefficients are available as optional output in RLM by specifying **stand=1** in the RLM command line. By default, standardized regression coefficients are not displayed. When requested, they will be found in the last column of the section of output showing the simple, semipartial, and partial correlations, under the column heading "Stand."

Covariance Matrix of Regression Coefficients

RLM will display the variance–covariance matrix of the regression coefficients by specifying **covcoeff=1** in the RLM command line.

Level of Confidence for Confidence Intervals

By default confidence intervals are set to 95%. The level of confidence can be changed with the **conf** option, setting the *ci* argument to the desired confidence between 50 and 99.999. For instance, for 90% confidence intervals, add **conf=90** to the RLM command line.

Regression Diagnostics

RLM will print and save various statistics for screening the data for various irregularities and testing assumptions by specifying **`diagnose=1`** in the RLM command line. When this option is specified, the output will include a table containing the minimum and maximum values of all variables in the model, the minimum and maximum predicted value ($\hat{Y}$), residual, and *t*-residual, as well as the smallest Bonferroni-corrected *p*-value for the largest *t*-residual, with a case number identifier. RLM will also produce a new data set containing all the variables in the model as well as each case's predicted value $\hat{Y}$, residual (*e*), deleted residual ($_{d}e$), Studentized residual (*str*), *t*-residual (*tri*), Mahalanobis distance (*MD*), leverage (*h*), and Cook's distance (*Cook*). This can be used to examine or test various model assumptions and identify unusual cases. For a discussion of these statistics and various methods, see Chapter 16.

Shrunken *R*

RLM can produce three estimates of shrunken *R*. These include one based on the Browne formula and the two leave-one-out estimates described in section 7.2.2. To obtain these estimates, request them with the **`crossv`** option, setting its toggle to 1 (i.e., **`crossv=1`**).

Heteroscedasticity-Consistent Standard Errors

By default, RLM uses the estimator for the standard errors of the regression coefficients described in section 4.4.3 (equation 4.3) that assumes homoscedasticity of the errors in estimation. RLM can also generate standard errors using the HC0, HC1, HC2, HC3, or HC4 heteroscedasticity-consistent standard error estimators described in Cribaro-Neto (2004), Hayes and Cai (2007), and Long and Ervin (2000). A heteroscedasticity-consistent standard error estimator is requested by setting the argument for the **`hc`** option to 0, 1, 2, 3, or 4 (e.g., **`hc=3`** generates the HC3 estimator). Any computation that relies on the variance–covariance matrix of the regression coefficients (e.g., regression coefficient standard errors) will automatically employ the HC estimator when this option is requested, including the the Johnson–Neyman method, tests of conditional effects in moderation analysis, test of the significance of *R* and change in R^2, and linear combinations of regression coefficients. See section 16.3.1.

All Subsets Regression

The **subsets** option in RLM conducts all subsets regression, discussed in sections 7.3.2 and 17.3.2. When this option's toggle is set to 1 (i.e., **subsets=1**), RLM generates output containing R for all possible models containing at least one regressor. The output takes the form of a table with the variable names at the top and models occupying the rows. The table entries for each row contain zeros and ones under the variable name. A 1 in the column designates that the variable in that column is included in the model. R is in the last column.

The number of possible models explodes as the number of regressors increases, and computing time and memory requirements increase accordingly. For this reason, all subsets regression is available only for models that include 15 or fewer regressors. All subsets regression is not available for models that specify moderation using the **mod**, **mcmod**, **mcfoc** options or models with a variable specified as multicategorical using the **mcx** option.

Dominance Analysis

Dominance analysis is a method for determining the relative importance of regressors in a model. Output from a dominance analysis is requested by specifying **dominate=1** in the RLM command line. The RLM macro will display a dominance table, discussed in section 8.4. The entries in the dominance table are the proportion of the possible subset models in which the variable in the row contributes more to prediction accuracy than the variable in the column. The diagonals of the dominance table are zero, and the cells symmetrically located around the diagonal usually sum to one.

Dominance analysis is not available for models with interactions or multicategorical variables. Thus, the **dominate** option is ignored when used in conjunction with the **mod**, **mcmod**, **mcfoc** or **mcx** options. Dominance analysis requires a lot of computations that require time and memory. Consequently, dominance analysis is available only for models with 15 or fewer regressors.

Spline Regression

RLM can conduct spline regression, discussed in section 12.3, wherein separate linear models relating one variable to the outcome are estimated between joints defined by user-specified values on the measurement scale

of the variable defining the splines. Spline regression is conducted by using the **spline** option, followed by a list of joint values separated by commas. The joint values should be values on the measurement scale of the variable listed *last* in ***xvarlist***. For example, if ***age*** were listed last, then the option **spline=30,40,50** would specify splines for the ***age*** variable, with the joints defined at ages 30, 40, and 50. Up to 10 joints may be specified when using the **spline** option. Joint locations must be listed in ascending order of value, with no ties, and all spline segments must contain at least two cases.

The variable listed last in ***xvarlist*** cannot be multicategorical, and so the **spline** option is incompatible with the **mcx** option. Spline regression is also not available in a model with an interaction specified in the model using the **mod**, **mcfoc**, or **mcmod** option or for use in conjunction with dominance analysis, all subsets regression, or when using the **settest** option.

The features of the **spline** option cannot be accessed through the RLM dialog box.

Multicategorical Regressors

Multicategorical regressors can be included in an RLM command using any system for coding groups such as those described in Chapters 9 and 10. RLM has an option for automatically generating $g - 1$ variables representing one (and only one) multicategorical regressor coding g groups with one of six coding systems. Such a multicategorical regressor must be listed *last* in the list of variables in ***xvarlist***. The coding system to be employed is specified as the argument in the **mcx** option. The six options (with argument in parentheses) available are indicator coding (1), sequential coding (2), Helmert coding (3), effect coding (4), weighted Helmert coding (5), and weighted effect coding (6). These coding systems are described in Chapters 9 and 10. When the **mcx** option is employed, RLM will display a matrix of the $g - 1$ codes used for each of the g groups at the top of the output.

A variable specified as multicategorical cannot contain more than 10 categories. RLM discerns the number of categories using the number of unique numerical codes in the variable specified as multicategorical. Dominance analysis and all subsets regression are not available when a multicategorical variable is specified using the **mcx** option.

The examples below assume that the variable called "cond" represents four groups coded in the data with the numbers 1, 3, 4, and 6.

Indicator Coding

When the argument for **mcx** is set to 1, $g - 1$ indicator codes are used to represent groups with the largest numerical code treated as the reference category. Indicator coding is discussed in Chapter 9. The $g - 1$ indicator codes will correspond to groups as coded in the multicategorical variable in ascending sequential order. For example, **mcx=1** implements the coding system below:

cond	D_1	D_2	D_3
1	1	0	0
3	0	1	0
4	0	0	1
6	0	0	0

There is no option for changing which group is treated as the reference. If you want to designate a different group as the reference group, recode the multicategorical variable prior to using RLM so that the reference group you desire is coded with the numerically largest code.

Sequential Coding

When the argument for **mcx** is set to 2, $g - 1$ sequential codes represent groups. Sequential coding allows for the comparison of group j to the group one ordinal position higher on the categorical variable. Sequential coding is discussed in section 10.1.1. RLM will assume that the ascending ordinality of the multicategorical variable corresponds to the ascending sequence of arbitrary numerical codes in the multicategorical variable. For example, **mcx=2** implements the coding system below:

cond	D_1	D_2	D_3
1	0	0	0
3	1	0	0
4	1	1	0
6	1	1	1

Helmert Coding

When the argument for **mcx** is set to 3, Helmert coding is used. Helmert coding allows for the comparison of group j to all groups ordinally higher on the categorical variable. Helmert coding is discussed in section 10.1.2. RLM will assume that the ascending ordinality of the multicategorical variable corresponds to the ascending sequence of arbitrary numerical codes in the multicategorical variable. Helmert coding is also useful for setting up certain orthogonal contrasts for a nominal multicategorical variable. For example, **mcx=3** implements the coding system below:

cond	D_1	D_2	D_3
1	−3/4	0	0
3	1/4	−2/3	0
4	1/4	1/3	−1/2
6	1/4	1/3	1/2

Effect Coding

When the argument for **mcx** is set to 4, effect coding is used, with the group with the largest numerical code left out of the coding scheme. Effect coding is discussed in section 10.1.3. The indicator variables correspond to the groups in ascending sequential order in the coding of the multicategorical variable. For example, **mcx=4** implements the coding system below:

cond	D_1	D_2	D_3
1	1	0	0
3	0	1	0
4	0	0	1
6	−1	−1	−1

Weighted Helmert Coding

When the argument for **mcx** is set to 5, weighted Helmert coding is used, with the weights determined by the sizes of the groups. Weighted Helmert coding is discussed in section 10.3.2. RLM will assume that the ascending

ordinality of the multicategorical variable corresponds to the ascending sequence of arbitrary numerical codes in the multicategorical variable. For example, assuming the sample sizes for the groups coded 1, 3, 4, and 6 are 20, 40, 30, and 10, respectively, **mcx=5** implements the coding system below:

cond	D_1	D_2	D_3
1	−0.750	−0.125	−0.125
3	0.250	−0.625	−0.125
4	0.250	0.375	−0.375
6	0.250	0.375	0.625

Weighted Effect Coding

When the argument for **mcx** is set to 6, weighted effect coding is used, with group with the largest numerical code left out of the coding scheme. The weights are determined by the sizes of the groups coded with the multicategorical variable. Weighted effect coding is discussed in section 10.3.1. For instance, if the sample sizes for the groups coded 1, 3, 4, and 6 in the variable "cond" are 20, 40, 30, and 10, respectively, **mcx=6** implements the coding system below:

cond	D_1	D_2	D_3
1	1	0	0
3	0	1	0
4	0	0	1
6	−2	−4	−3

Linear Interaction

RLM can assist in the estimation of models that include linear moderation of one variable's effect on *yvar* by another variable in the model through a set of options for specifying the interaction, probing it, and visualizing it. There are numerous ways an interaction can be specified. All these methods assume that the last two variables in *xvarlist* are the focal predictor and moderator, respectively. When listing variables in the RLM command after

x=, the moderator variable should be listed *last*, and the focal predictor should be listed *second to last*. The moderator is the variable that you believe is influencing or is otherwise related to the size of the focal predictor's effect on *yvar*.

To specify a linear interaction between the focal predictor and moderator, use the **mod** option, setting its argument to 1 (i.e., **mod=1**). Except as discussed in the section of this documentation titled "Multicategorical Variables and Interaction," doing so will include the product of the focal predictor and the moderator as an additional regressor in the model. The user does not need to construct the product of these two variables in the data, as RLM does all the necessary computations internally. It will not add it to the data file. If the focal predictor or moderator is a multicategorical variable rather than a dichotomous or continuous dimension, see the "Multicategorical Variables and Interaction" section in this documentation.

Specifying an interaction generates additional output, including the difference in R^2 for the model with and without the product term. As discussed in section 14.4.5, this test is mathematically equivalent to the *t*-test for the the regression coefficient for the product. Additional output provided by RLM will depend on other options specified for probing or visualizing the interaction.

Dominance analysis and all subsets regression are not available when specifying a model that includes an interaction. There are no options implemented in RLM for interactions higher than second order.

Probing Interactions and Generating Conditional Effects

In any model that specifies an interaction, RLM can produce estimates of the conditional effect of the focal predictor at various values of the moderator—so-called "simple slopes." By default, when a moderator is dichotomous, conditional effects of the focal predictor at the two values of the moderator are generated. But when a moderator is quantitative, conditional effects of the focal predictor are estimated by default at the sample mean of the moderator, as well as plus and minus one standard deviation from the moderator mean.

Three alternatives for probing interactions are available in RLM. For a quantitative moderator, the **ptiles** option generates conditional effects at the 25th, 50th, and 75th percentiles of the distribution of the moderator. This option is available by setting the argument in the **ptiles** option to 1

(i.e., **ptiles=1**). Unlike when the mean and ± one standard deviation from the mean is used, these percentile values are guaranteed to be within the range of the observed moderator variable data.

The second alternative is to request the conditional effect of interest at a specific, single value of the moderator. This is accomplished through the use of the **modval** option, setting the corresponding argument *mval* to the value of the moderator at which you'd like the estimate of the conditional effect of the focal predictor. For example, to generate an estimate of the conditional effect of the focal predictor on *yvar* when the moderator is equal to 3, append **modval=3** to the RLM command. This option for probing an interaction is not available in the custom dialog version of RLM. Only one value of *mval* can be specified in a single run of RLM.

The third alternative implemented in RLM is the Johnson–Neyman technique, requested by setting the argument in the **jn** option to 1 (i.e., **jn=1**). This approach identifies the value(s) on the moderator variable continuum at which point (or points) the effect of the focal predictor on *yvar* transitions between statistically significant and not, using the α-level of significance as the criterion. By default, $\alpha = 0.05$. This can be changed using the **conf** option, setting the desired confidence to $100(1 - \alpha)$. For example, for $\alpha = 0.01$, specify **conf=99**. In addition to identifying points of transition, RLM produces a table to aid in the identification of the regions of significance as well as information about the percentage of cases in the data above ("% Above") and below ("% Below") the points of transition in significance this procedure identifies. See section 14.3.2 for a discussion of the Johnson–Neyman technique.

Multicategorical Variables and Interaction

By default, RLM assumes that the focal predictor and moderator are dichotomous and/or continuous when an interaction is specified using the **mod** option described earlier. If the moderator is a multicategorical variable coding membership in one of g groups, $g > 2$, it should be specified as such using the **mcx** command, described in section earlier titled "Multicategorical Regressors." The combination of the **mod=1** along with **mcx** estimates a model that includes $g - 1$ product terms between the regressors coding group and the focal predictor, along with a test of interaction using the difference in R^2 as the test statistic. RLM will also generate estimates of the effect of the focal predictor on *yvar* in each of the g groups, along with statistics for inference.

An equivalent approach is the use of the **mcmod** option, which stands for "multicategorical moderator." The same six possible methods for coding groups can be used with the **mcmod** option as are available for the **mcx** option. A number 1 through 6 should be provided as the argument for **mcmod**. For example, **mcmod=2** tells RLM that the moderator variable is multicategorical and to use sequential coding of groups. When the **mcmod** option is used, it is not necessary to use the **mod** option, as the specification of **mcmod** in the RLM command implies an interaction between the two variables listed last in the **x=** list.

If the focal predictor is multicategorical, it should be specified as such using the **mcfoc** option, which stands for "multicategorical focal predictor." Like the **mcx** and **mcmod** options, six coding options are available and the option desired in the form of a number 1 through 6 should be provided. For example, **mcfoc=3** specifies that the focal predictor is multicategorical and tells RLM to use Helmert coding of the *k* groups. The use of the **mcfoc** implies a model with an interaction, so use of the **mod** option in the RLM command line is not required.

When the focal predictor is specified as multicategorical, in addition to a test of interaction using the difference in R^2 as the test statistic, RLM provides a test of differences between the estimated group means of *yvar* at various values of the moderator and focal predictor. When the moderator is dichotomous, two such tests are provided, one for each value of the moderator. When the moderator is a continuum, these tests are conducted for moderator values corresponding to the mean, a standard deviation below the mean, and a standard deviation above the mean, unless the **ptiles** or **modval** options are used to override this default.

The Johnson–Neyman method for probing an interaction is not available when the focal predictor is specified as a multicategorical variable. Thus, the **jn** option is ignored when used in conjunction with **mcfoc**.

The **mcmod** and **mcfoc** options cannot be used simultaneously in an RLM command. A model with a multicategorical focal predictor and moderator is an ANOVA or ANCOVA model. Use the GLM or UNIANOVA routines in SPSS to estimate such a model.

When an interaction involves a multicategorical focal predictor or moderator, the **center** option centers only the variable involved in the interaction that is not multicategorical.

Mean-Centering in Models with an Interaction

In models that include parameters for estimating interaction effects, the user has the option of requesting RLM to mean-center variables used in the construction of products of regressors prior to model estimation by setting the argument in the ***center*** option to 1 (i.e., **center=1**). All output will be based on the focal predictor and moderator in the mean-centered metric (e.g., the regression coefficient for the focal predictor will be conditioned at the mean of the moderator, and the conditional effects of the focal predictor on *yvar* at values of the moderator will be based on values of the moderator after mean-centering).

By default, variables used to form products are not mean-centered. When mean-centering is requested, arguments of options used for estimating conditional effects at specific values of the moderator(s) should be values based on a mean-centered metric. For example, the RLM command

```
RLM y=smoking/x=anxiety surgery addict/mod=1/modval=1.5.
```

will produce the conditional effect of `surgery` on `smoking` when `addict` = 1.5, whereas the RLM command

```
RLM y=smoking/x=anxiety surgery addict/mod=1/modval=1.5/center=1.
```

produces the conditional effect of `surgery` on `smoking` when `addict` is 1.5 measurement units above the sample mean of `addict`.

In models with a multicategorical variable involved in the interaction specified with the **mcfoc** or **mcmod** options, centering is not undertaken for the multicategorical variable.

Visualizing Interactions

To help visualize an interaction, the **plot** option generates a table of estimated values of *yvar* from the model using various values of the focal predictor and moderator. This table is generated by setting the argument in the **plot** option to 1 (i.e., **plot=1**). Any covariates in the model are set to their sample mean when deriving the estimated values in the table generated. In the table, the estimated value of *yvar* is listed as "yhat." The table can be entered into the user's preferred graphing program for visualization.

The **plot** option also generates the estimated standard error of the estimated values of *yvar* from the model. These are listed in the table under the column labeled "se(yhat)."

Inference for Sets of Regressors

RLM provides a test that all of the regression coefficients for a subset of the regressors in the model are zero. By adding **settest=*nvars*** to the RLM command, a test of the null hypothesis that the regression coefficients for the last ***nvars*** regressors in the **x=** list are equal to zero is conducted. For example, **settest=3** conducts a test that the regression coefficients for the last three regressors are equal to zero. This test take the form of an *F*-ratio and is equivalent to a test of the null hypothesis that the last ***nvars*** variables in the **x=** list do not improve the fit of the model as measured by the increase in the squared multiple correlation when those variables are added. For a discussion of this test, see section 5.3.3.

If the last variable in the **x=** list is specified as multicategorical using the **mcx** option, then the set includes all of the variables coding group (i.e., if the multicategorical variable codes g groups, then all $g-1$ variables coding groups are included in the set). Thus, the numerator degrees of freedom for the *F*-ratio, which usually is equal to ***nvars***, will be larger when **settest** is used in conjunction with **mcx**.

When the **settest** option is used in conjunction with the **hc** option, the *F*-test for the set of regressors does not assume homoscedasticity and instead relies on a heteroscedasticity-consistent variance–covariance matrix for the regression coefficient estimates. For details on the computation of the *F*-ratio in this case, see Hayes and Cai (2007).

The **setttest** option is implemented in the RLM dialog box in a section of the box labeled "Variables in subset test." The number selected here chooses the last m regressors in the regressors box for inclusion in the test, where m is a number between one and nine. For instance, by selecting "2," the last two variables in the "Regressors" box are included in the set, and RLM produces a test of the null hypothesis that adding these two variables to the model does not significantly improve its fit as quantified by the increase in the multiple correlation that results when these two variables are added. As discussed in section 5.3.3, this is equivalent to the null hypothesis that the regression coefficients for both of these regressors are equal to zero.

This test is not available in models that include an interaction.

Inference for Linear Combinations of Regression Coefficients

A regression model contains a regression constant and k regressors. RLM can construct a weighted linear combination of these $k + 1$ estimates of regression parameters, the standard error of the linear combination, and a confidence interval for inference about its value. In addition, RLM provides a test of the null hypothesis that the linear combination equals zero. The linear combination is specified with the **contrast** option along with a set of $k + 1$ weights for the regression coefficient and each of the regression coefficients. For instance, from the example in section 4.6,

```
rlm y=wtloss/x=exercise food/contrast=1,3,7.
```

constructs estimated weight loss ($\hat{Y}$) for someone who exercises 3 days per week and with a weekly food intake of 7 units. In addition to the estimate, RLM produces $SE(\hat{Y})$, a confidence interval, and the two-tailed p-value for testing the null hypothesis that the estimate equals zero.

When using the **contrast** option, the weights for each regression coefficient must be in the same order as the regression coefficients appear in the RLM output from top to bottom, with the weight for the regression constant always listed first. The number of weights provided must be $k + 1$. Some RLM options (**mod**, **mcmod**, and **mcfoc**) generate variables in the model that do not appear in the RLM command line. Prior to specifying the weights, run your RLM command without the **contrast** option so you will know how many weights are needed and in what order they should be listed.

The contrast option can also be used to conduct an inference for the difference between two regression coefficients in the model. For example,

```
rlm y=pknow/x=npnews natnews sex age/contrast=0,1,-1,0,0.
```

regresses *pknow* on *npnews*, *natnews*, *sex*, and *age* while generating a test of equality of the regression coefficients for *npnews* and *natnews*, along with the standard error of the difference and a confidence interval.

The features of the **contrast** option cannot be accessed through the RLM dialog box.

Decimal Place Precision in Output

RLM generates numerical output to four decimals places of resolution by default. This can be changed with the ***dec*** argument when using the

decimals option. This argument is set to `F10.4` by default, meaning numbers in the output will contain up to 10 characters, with four of these to the right of the decimal. In this argument, **Fa.b** sets the number of characters allocated to numbers to ***a*** and the number of decimal places to display to the right of the decimal point to ***b***. For example, **decimals=F12.6** specifies 12 characters with six to the right of the decimal place. In the ***dec*** argument, **a** should be larger than **b**.

Missing Data

A case will be deleted from the analysis if user- or system-missing on any of the variables in *yvar* or *xvars*. There are no features in RLM for imputing data or using more sophisticated missing data routines such as multiple imputation or the EM algorithm.

Installation, Execution, and Syntax Modifications for SAS Users

The SAS version of RLM functions similarly to the SPSS version, and most of the instructions described in this appendix apply to the SAS version, with only the minor modifications described below. Like the SPSS version, the SAS version is a program file (RLM.sas), which when executed creates a new command that SAS understands called **%rlm**. Once RLM.sas is executed (without changing the file whatsoever), then the **%rlm** command is available for use and the program can be closed. Once you close SAS, you have to define the **%rlm** command by executing RLM.sas again. **RLM for SAS requires the PROC IML module**. To determine whether you have the PROC IML module installed, run the following commands in SAS:

```
proc iml;
print "PROC IML is installed";
quit;
```

When this code is executed, check the log for any errors, as well as your output window for the text "PROC IML is installed." Any errors in the log or a failure to see this text suggests that PROC IML is not installed on your version of SAS.

The syntax structure for RLM for SAS is almost identical to the SPSS version. There are a few important exceptions:

- The command name is **%rlm** rather than **rlm**.
- All parts of the command between **%rlm** and the ending of the command must be in parentheses. SAS commands ordinarily must end in a semicolon (**;**). Though an **%rlm** command can end with a semicolon, it is not required.
- The data file being analyzed must be specified in the command as **data=***file*, where ***file*** is the name of a SAS data file.
- Options and specifications must be delimited with a comma (**,**) rather than a slash (**/**).
- When requesting regression diagnostics using the **diagnose** option, specify the name of the file to store the diagnostic statistics following the equals sign, as in **diagnose=***filename*, where ***filename*** is any valid SAS datafile name.
- Joint values and weights should not be separated by commas when using the **spline** and **contrast** options, respectively.
- When specifying the number of decimal places in output using the **decimals** option, the "F" should be left off the ***dec*** argument. For example, to set 12 characters for numbers, with six after the decimal, use **decimals=12.6**.

For example, suppose the data corresponding to the example on page 585 were stored in a SAS work file named POLITICS. The SAS version of the RLM command corresponding to this example would be

```
%rlm (data=politics,y=know,x=educ attn sex age elab,mcfoc=1,
    decimals=12.6,modval=4,plot=1,hc=3);
```

Notes

- In the SPSS version of RLM, variable names are restricted to eight characters or fewer. If any of your variable names are longer than eight characters, shorten them before using RLM.
- Do not use STRING formatted variables in any of your models. Doing so will produce errors. All variables should be NUMERIC format.
- RLM ignores the variable "Measure" settings in SPSS set in the Variable View pane (i.e., nominal, ordinal, scale).

- The RLM procedure code cannot be imbedded in a syntax file with an `INCLUDE` command in SPSS, but it can be called with an `INSERT` command. This eliminates the need to manually load and run RLM.sps prior to execution of a set of commands that calls the RLM macro. See the *Command Syntax Reference* available through the Help menu in SPSS for details on the use of the `INSERT` command.

Appendix B
Linear Regression Analysis Using R

R is a programming language that is growing in popularity among scientists. The R user's attraction to R is attributable to its price (downloadable for free from *www.r-project.org*), flexibility, and power. It is also known for its steep learning curve. In this appendix, we illustrate some R commands for linear regression analysis using only the functions that come with R when it is downloaded. There are many freely available "packages" for R that make some of the things discussed here easier or that produce output in a more user-friendly format. This appendix is not intended as a stand-alone guide to the use of R. Many books on the use of R as a programming language, graphics production system, and data analysis tool are available, and the Internet is filled with advice from users assisting other users.

The command below reads the text version of the EXERCISE file, first used in Chapter 3 and available on this book's web page at *www.afhayes.com*. The code specifies that the file is stored on the computer's C drive in a folder named "data."

```
health<-read.table(file="C:\\data\\exercise.txt",header=TRUE)
```

Once executed, the result is a "data frame" named `health`. The **`header=TRUE`** line tells R that the variable names are at the top of the data file. To make sure the data were read correctly, we can print it to the screen by typing the data frame object name

```
health
```

which results in the following appearing on the output screen after execution:

```
   id exercise food metab sex wtloss
1   1        0    2    15   0      6
2   2        0    4    14   0      2
3   3        0    6    19   0      4
4   4        2    2    15   1      8
5   5        2    4    21   1      9
6   6        2    6    23   0      8
7   7        2    8    21   1      5
8   8        4    4    22   1     11
9   9        4    6    24   0     13
10 10        4    8    26   0      9
```

In section 3.2.1 we conducted a regression analysis estimating weight loss from exercise frequency and food intake. A linear regression with *wtloss* as the dependent variable and *exercise* and *food* as regressors is conducted with the **lm** command. We tell R that the results will be stored in an object named model1 and that the variables should be pulled from the health data frame by appending health$ to the beginning of variable names.

```
model1<-lm(health$wtloss~health$exercise+health$food)
```

Note that variable names are case sensitive in R, so make sure you follow case consistently. A variable named *wtloss* is different than one named *WTLOSS* which is different from one named *Wtloss*.

The **summary** command shows the results of the regression analysis held in the model1 object:

```
summary(model1)
```

```
Call:
lm(formula = health$wtloss ~ health$exercise + health$food)

Residuals:
   Min     1Q Median     3Q    Max
    -2     -1      0      1      2

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.0000     1.2749   4.706 0.002193 **
```

```
health$exercise   2.0000     0.3333   6.000 0.000542 ***
health$food      -0.5000     0.2520  -1.984 0.087623 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.512 on 7 degrees of freedom
Multiple R-squared:  0.8376,    Adjusted R-squared:  0.7912
F-statistic: 18.05 on 2 and 7 DF,  p-value: 0.001727
```

Standardized regression coefficients aren't produced by R unless you ask for them. Use **scale** to generate standardized regression coefficients.

```
lm(scale(health$wtloss)scale~(health$exercise)+scale(health$food))
```

```
Call:
lm(formula = scale(health$wtloss) ~ scale(health$exercise) +
    scale(health$food))

Coefficients:
     (Intercept)  scale(health$exercise)   scale(health$food)
       8.240e-17               9.872e-01           -3.265e-01
```

The standardized regression coefficients are in scientific notation in this output.

It can be a nuisance to have to tell R the data frame that the variables are held in for every command, and when you do, the data frame appears in all output appended to all variable names, which can make for messy output. When executing a set of commands applied to the same data frame, you can eliminate this requirement and pretty up the output with the command

```
attach(health)
```

Having executed the **attach** command, we no longer have to specify the data frame source for the variables in future commands (which we will not in all that follows). Now the regression command can be rewritten and executed as

```
model1<-lm(wtloss~exercise+food)
```

A table containing the sums of squares and other statistics found in a regression ANOVA summary table for the results in the `model1` object is

generated with the **anova** command:

```
anova(model1)
```

which generates as output

```
Analysis of Variance Table

Response: wtloss
          Df Sum Sq Mean Sq F value   Pr(>F)
exercise   1   73.5  73.500 32.1562 0.000758 ***
food       1    9.0   9.000  3.9375 0.087623 .
Residuals  7   16.0   2.286
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Confidence intervals for the regression coefficients stored in the `model1` object are produced with the **confint** command. The level of confidence desired can be specified, as in

```
confint(model1,level=0.95)
```

```
                2.5 %     97.5 %
(Intercept)  2.985316 9.01468433
exercise     1.211792 2.78820808
food        -1.095829 0.09582931
```

Residuals can be extracted from the `model1` object and saved as a new variable. Listing a variable by itself prints its contents to the screen:

```
resid<-residuals(model1)
resid
```

```
 1  2  3  4  5  6  7  8  9 10
 1 -2  1 -1  1  1 -1 -1  2 -1
```

The numbers 1 through 10 identify the case numbers. The estimates of Y from the model (i.e., $\hat{Y}$) can be extracted and printed similarly:

```
yhat<-fitted(model1)
yhat
```

```
 1  2  3  4  5  6  7  8  9 10
 5  4  3  9  8  7  6 12 11 10
```

The residuals and predicted values can be appended to the `health` data and then printed to the screen with

```
health<-data.frame(health,resid,yhat)
health
```

```
   id exercise food metab  sex wtloss resid yhat
1   1        0    2    15    0      6     1    5
2   2        0    4    14    0      2    -2    4
3   3        0    6    19    0      4     1    3
4   4        2    2    15    1      8    -1    9
5   5        2    4    21    1      9     1    8
6   6        2    6    23    0      8     1    7
7   7        2    8    21    1      5    -1    6
8   8        4    4    22    1     11    -1   12
9   9        4    6    24    0     13     2   11
10 10        4    8    26    0      9    -1   10
```

Various functions are available for regression diagnostics. For example, leverage and *t*-residuals can be constructed using

```
hat<-hatvalues(model1)
tres<-rstudent(model1)
```

and then appended to the `health` data frame:

```
health<-data.frame(health,hat,tres)
health
```

```
   id exercise food metab sex wtloss resid yhat      hat      tres
1   1        0    2    15   0      6     1    5 0.377778  0.818520
2   2        0    4    14   0      2    -2    4 0.266667 -1.761661
3   3        0    6    19   0      4     1    3 0.377778  0.818520
4   4        2    2    15   1      8    -1    9 0.350000 -0.798936
5   5        2    4    21   1      9     1    8 0.127778  0.680531
6   6        2    6    23   0      8     1    7 0.127778  0.680531
7   7        2    8    21   1      5    -1    6 0.350000 -0.798936
8   8        4    4    22   1     11    -1   12 0.377778 -0.818520
```

```
9   9        4    6    24   0      13     2   11 0.266667  1.761661
10 10        4    8    26   0       9    -1   10 0.377778 -0.818520
```

Partial and semipartial correlations can be calculated from residuals by correlating them, as discussed in section 3.3. It is not necessary to construct a new variable containing the fitted values to calculate the residuals, as the fitted values can be constructed by imbedding an **lm** command inside a mathematical operation. For instance, to construct the residuals from a model estimating weight loss from exercise, try

```
residy<-wtloss-fitted(lm(wtloss~exercise))
```

The residuals estimating food intake from exercise frequency would be constructed similarly:

```
residx<-food-fitted(lm(food~exercise))
```

Pearson's coefficient of correlation is produced with the **cor** command. For instance, the partial correlation between weight loss and food intake controlling for exercise frequency is the Pearson correlation between `residy` and `residx`, generated with

```
cor(residy,residx)
```

```
[1] -0.6
```

and the semipartial correlation is Pearson's correlation between `wtloss` and `residx`:

```
cor(wtloss,residx)
```

```
[1] -0.3022756
```

Mathematical transformations to variables are easily undertaken. For instance, if you wanted to substract 2 from all values in the variable *food*, use

```
food<-food-2
```

though this would be dangerous, because it would replace all values of *food* in the data with their new values. It might be safer to create a

new variable *foodt*, as in

```
foodt<-food-2
```

Or suppose you wanted to test for linear interaction between food intake and exercise using the methods discussed in Chapter 13. A new variable that is the product of food intake and exercise can be created using

```
foodexer<-food*exercise
```

and now this product *foodexer* included as an additional predictor of weight loss in a regression model along with food intake and exercise frequency.

In section 5.3.3 we discussed tests of multivariate partial association. In the example in that discussion, we examined whether factors in one's control (food intake and exercise) explain variation in weight loss when controlling for factors not in one's control (sex and metabolism). The test is undertaken by examining the change in R^2 when food intake and exercise are added to a model of weight loss that already includes metabolism and sex. In R, this is accomplished by estimating the two models and storing their results in separate objects, named `model2` and `model3` in the code below:

```
model2<-lm(wtloss~metab+sex)
model3<-lm(wtloss~metab+sex+exercise+food)
```

The two models are then compared with an F test with the **anova** command, listing the object containing the results from the model with fewer regressors first:

```
anova(model2,model3)
```

```
Analysis of Variance Table

Model 1: wtloss ~ metab + sex
Model 2: wtloss ~ exercise + food + metab + sex
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      7 51.766
2      5  6.797  2    44.969 16.539 0.006248 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

R is widely respected for its graphics and data visualization capabilities. A simple scatterplot depicting the relationship between weight loss (on the *Y*-axis) and food intake (on the *X*-axis) is produced by the **plot** option.

```
plot(y=wtloss,x=food)
```

This very minimalist scatterplot produced by the above code can be embellished in many ways. For example, the code below adds options for labeling the axes, specifies the range of values to place on the axes, superimposes a regression line on the plot, and changes the symbol used in the plot.

```
plot(y=wtloss,x=food,xlim=c(0,10),ylim=c(0,14),xlab="Food
   intake",ylab="Weight loss",pch=15)
abline(lm(wtloss~food),lwd=2)
```

With practice and the assistance of a good book or a knowledgeable friend to guide you, you can produce very nice publication-quality graphics in R, with little if any extra editing outside of the R program. Indeed, the majority of the figures in this book were created entirely with R. For example, Figure 3.1 was created from the exercise data file using the R code below:

```
mark<-c(15,15,15,21,21,21,21,17,17,17)
plot(y=wtloss,x=food,xaxs="i",yaxs="i",pch=mark,xlim=c(0,10),
   ylim=c(0,15),xlab=expression('Food intake (X'[2]*')'),
   ylab="Pounds lost (Y)",cex=1.2)
legend.txt<-c(as.expression(bquote(X[1]~"= 0 hours of exercise
   per week ")),as.expression(bquote(X[1]~"= 2 hours of
   exercise per week ")),as.expression(bquote(X[1]~"= 4 hours
   of exercise per week ")))
legend("topleft",legend=legend.txt,bty="o",cex=0.8,
   pch=c(15,21,17))
```

Appendix C
Statistical Tables

TABLE C.1. Right-Tail Normal Probabilities

Table entries are the proportion of the area in the normal distribution to the right of the *Z*-score computed as the sum of the bold entries to the top and left. For instance, the second entry in the final column shows that the right-tailed area is .4247 when $Z = 0.19$. Table entries were generated with the IDF.NORMAL function in SPSS.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

TABLE C.2. Critical Values of t

Table entries are t-values that cut off the upper (one-tailed) or extreme (two-tailed) p proportion of the $t(df)$ distribution from the rest. Table entries were generated with the IDF.T function in SPSS.

	One-tailed p-value									
	.10	**.05**	**.025**	**.01**	**.005**	**.0025**	**.001**	**.0005**	**.00025**	**.0001**
	Two-tailed p-value									
df	**.20**	**.10**	**.05**	**.02**	**.01**	**.05**	**.002**	**.001**	**.0005**	**.0002**
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587	5.049	5.694
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	4.417	4.880
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850	4.146	4.539
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725	3.996	4.352
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646	3.902	4.234
35	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591	3.836	4.153
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551	3.788	4.094
45	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520	3.752	4.049
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496	3.723	4.014
55	1.297	1.673	2.004	2.396	2.668	2.925	3.245	3.476	3.700	3.986
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460	3.681	3.962
65	1.295	1.669	1.997	2.385	2.654	2.906	3.220	3.447	3.665	3.942
70	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435	3.651	3.926
75	1.293	1.665	1.992	2.377	2.643	2.892	3.202	3.425	3.639	3.911
80	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416	3.629	3.899
85	1.292	1.663	1.988	2.371	2.635	2.882	3.189	3.409	3.620	3.888
90	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402	3.612	3.878
95	1.291	1.661	1.985	2.366	2.629	2.874	3.178	3.396	3.605	3.869
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390	3.598	3.862
110	1.289	1.659	1.982	2.361	2.621	2.865	3.166	3.381	3.587	3.848
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373	3.578	3.837
130	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367	3.571	3.828
140	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361	3.564	3.820
150	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357	3.558	3.813
200	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340	3.539	3.789
300	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323	3.519	3.765
400	1.284	1.649	1.966	2.336	2.588	2.823	3.111	3.315	3.510	3.754
500	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310	3.504	3.747
1000	1.282	1.646	1.962	2.330	2.581	2.813	3.098	3.300	3.492	3.733

TABLE C.3. Critical Values of F

Table entries are F values that cut off the upper p proportion of the $F(df_1, df_2)$ distribution from the rest. Table entries were generated with the IDF.F function in SPSS.

	***p*-value**						
df_2	**.100**	**.050**	**.025**	**.010**	**.005**	**.0025**	**.001**
				$df_1 = 1$			
10	3.285	4.965	6.937	10.044	12.826	16.036	21.040
15	3.073	4.543	6.200	8.683	10.798	13.133	16.587
20	2.975	4.351	5.871	8.096	9.944	11.940	14.819
30	2.881	4.171	5.568	7.562	9.180	10.889	13.293
40	2.835	4.085	5.424	7.314	8.828	10.411	12.609
50	2.809	4.034	5.340	7.171	8.626	10.138	12.222
75	2.774	3.968	5.232	6.985	8.366	9.789	11.731
100	2.756	3.936	5.179	6.895	8.241	9.621	11.495
200	2.731	3.888	5.100	6.763	8.057	9.377	11.155
500	2.716	3.860	5.054	6.686	7.950	9.234	10.957
1000	2.711	3.851	5.039	6.660	7.915	9.187	10.892
				$df_1 = 2$			
10	2.924	4.103	5.456	7.559	9.427	11.572	14.905
15	2.695	3.682	4.765	6.359	7.701	9.173	11.339
20	2.589	3.493	4.461	5.849	6.986	8.206	9.953
30	2.489	3.316	4.182	5.390	6.355	7.365	8.773
40	2.440	3.232	4.051	5.179	6.066	6.986	8.251
50	2.412	3.183	3.975	5.057	5.902	6.770	7.956
75	2.375	3.119	3.876	4.900	5.691	6.497	7.585
100	2.356	3.087	3.828	4.824	5.589	6.365	7.408
200	2.329	3.041	3.758	4.713	5.441	6.175	7.152
500	2.313	3.014	3.716	4.648	5.355	6.064	7.004
1000	2.308	3.005	3.703	4.626	5.326	6.028	6.956
				$df_1 = 3$			
10	2.728	3.708	4.826	6.552	8.081	9.833	12.553
15	2.490	3.287	4.153	5.417	6.476	7.634	9.335
20	2.380	3.098	3.859	4.938	5.818	6.757	8.098
30	2.276	2.922	3.589	4.510	5.239	5.999	7.054
40	2.226	2.839	3.463	4.313	4.976	5.659	6.595
50	2.197	2.790	3.390	4.199	4.826	5.466	6.336
75	2.158	2.727	3.296	4.054	4.635	5.222	6.011
100	2.139	2.696	3.250	3.984	4.542	5.105	5.857
200	2.111	2.650	3.182	3.881	4.408	4.936	5.634
500	2.095	2.623	3.142	3.821	4.330	4.838	5.506
1000	2.089	2.614	3.129	3.801	4.305	4.805	5.464

TABLE C.3. Critical Values of F (continued)

	p-value						
df_2	**.100**	**.050**	**.025**	**.010**	**.005**	**.0025**	**.001**
				$df_1 = 4$			
10	2.605	3.478	4.468	5.994	7.343	8.888	11.283
15	2.361	3.056	3.804	4.893	5.803	6.796	8.253
20	2.249	2.866	3.515	4.431	5.174	5.967	7.096
30	2.142	2.690	3.250	4.018	4.623	5.253	6.125
40	2.091	2.606	3.126	3.828	4.374	4.934	5.698
50	2.061	2.557	3.054	3.720	4.232	4.753	5.459
75	2.021	2.494	2.962	3.580	4.050	4.525	5.159
100	2.002	2.463	2.917	3.513	3.963	4.415	5.017
200	1.973	2.417	2.850	3.414	3.837	4.257	4.812
500	1.956	2.390	2.811	3.357	3.763	4.166	4.693
1000	1.950	2.381	2.799	3.338	3.739	4.136	4.655
				$df_1 = 5$			
10	2.522	3.326	4.236	5.636	6.872	8.288	10.481
15	2.273	2.901	3.576	4.556	5.372	6.263	7.567
20	2.158	2.711	3.289	4.103	4.762	5.463	6.461
30	2.049	2.534	3.026	3.699	4.228	4.776	5.534
40	1.997	2.449	2.904	3.514	3.986	4.470	5.128
50	1.966	2.400	2.833	3.408	3.849	4.297	4.901
75	1.926	2.337	2.741	3.272	3.674	4.078	4.617
100	1.906	2.305	2.696	3.206	3.589	3.973	4.482
200	1.876	2.259	2.630	3.110	3.467	3.822	4.287
500	1.859	2.232	2.592	3.054	3.396	3.734	4.176
1000	1.853	2.223	2.579	3.036	3.373	3.706	4.139
				$df_1 = 6$			
10	2.461	3.217	4.072	5.386	6.545	7.871	9.926
15	2.208	2.790	3.415	4.318	5.071	5.891	7.092
20	2.091	2.599	3.128	3.871	4.472	5.111	6.019
30	1.980	2.421	2.867	3.473	3.949	4.442	5.122
40	1.927	2.336	2.744	3.291	3.713	4.144	4.731
50	1.895	2.286	2.674	3.186	3.579	3.976	4.512
75	1.854	2.222	2.582	3.052	3.407	3.763	4.237
100	1.834	2.191	2.537	2.988	3.325	3.662	4.107
200	1.804	2.144	2.472	2.893	3.206	3.515	3.920
500	1.786	2.117	2.434	2.838	3.137	3.430	3.813
1000	1.780	2.108	2.421	2.820	3.114	3.402	3.778

TABLE C.3. Critical Values of F (continued)

	p-value						
df_2	**.100**	**.050**	**.025**	**.010**	**.005**	**.0025**	**.001**
				$df_1 = 7$			
10	2.414	3.135	3.950	5.200	6.302	7.564	9.517
15	2.158	2.707	3.293	4.142	4.847	5.616	6.741
20	2.040	2.514	3.007	3.699	4.257	4.850	5.692
30	1.927	2.334	2.746	3.304	3.742	4.194	4.817
40	1.873	2.249	2.624	3.124	3.509	3.902	4.436
50	1.840	2.199	2.553	3.020	3.376	3.737	4.222
75	1.798	2.134	2.461	2.887	3.208	3.529	3.955
100	1.778	2.103	2.417	2.823	3.127	3.429	3.829
200	1.747	2.056	2.351	2.730	3.010	3.286	3.647
500	1.729	2.028	2.313	2.675	2.941	3.203	3.542
1000	1.723	2.019	2.300	2.657	2.919	3.176	3.508
				$df_1 = 8$			
10	2.377	3.072	3.855	5.057	6.116	7.328	9.204
15	2.119	2.641	3.199	4.004	4.674	5.404	6.471
20	1.999	2.447	2.913	3.564	4.090	4.648	5.440
30	1.884	2.266	2.651	3.173	3.580	4.001	4.581
40	1.829	2.180	2.529	2.993	3.350	3.713	4.207
50	1.796	2.130	2.458	2.890	3.219	3.551	3.998
75	1.754	2.064	2.366	2.758	3.052	3.346	3.736
100	1.732	2.032	2.321	2.694	2.972	3.248	3.612
200	1.701	1.985	2.256	2.601	2.856	3.107	3.434
500	1.683	1.957	2.217	2.547	2.789	3.025	3.332
1000	1.676	1.948	2.204	2.529	2.766	2.998	3.299
				$df_1 = 9$			
10	2.347	3.020	3.779	4.942	5.968	7.140	8.956
15	2.086	2.588	3.123	3.895	4.536	5.235	6.256
20	1.965	2.393	2.837	3.457	3.956	4.487	5.239
30	1.849	2.211	2.575	3.067	3.450	3.847	4.393
40	1.793	2.124	2.452	2.888	3.222	3.563	4.024
50	1.760	2.073	2.381	2.785	3.092	3.402	3.818
75	1.716	2.007	2.289	2.653	2.927	3.199	3.561
100	1.695	1.975	2.244	2.590	2.847	3.103	3.439
200	1.663	1.927	2.178	2.497	2.732	2.963	3.264
500	1.644	1.899	2.139	2.443	2.665	2.882	3.163
1000	1.638	1.889	2.126	2.425	2.643	2.855	3.130

TABLE C.3. Critical Values of F (continued)

	p-value						
df_2	**.100**	**.050**	**.025**	**.010**	**.005**	**.0025**	**.001**
				$df_1 = 10$			
10	2.323	2.978	3.717	4.849	5.847	6.987	8.754
15	2.059	2.544	3.060	3.805	4.424	5.097	6.081
20	1.937	2.348	2.774	3.368	3.847	4.355	5.075
30	1.819	2.165	2.511	2.979	3.344	3.720	4.239
40	1.763	2.077	2.388	2.801	3.117	3.438	3.874
50	1.729	2.026	2.317	2.698	2.988	3.279	3.671
75	1.685	1.959	2.224	2.567	2.823	3.078	3.416
100	1.663	1.927	2.179	2.503	2.744	2.982	3.296
200	1.631	1.878	2.113	2.411	2.629	2.844	3.123
500	1.612	1.850	2.074	2.356	2.562	2.764	3.023
1000	1.605	1.840	2.061	2.339	2.541	2.737	2.991
				$df_1 = 11$			
10	2.302	2.943	3.665	4.772	5.746	6.861	8.586
15	2.037	2.507	3.008	3.730	4.329	4.982	5.935
20	1.913	2.310	2.721	3.294	3.756	4.245	4.939
30	1.794	2.126	2.458	2.906	3.255	3.615	4.110
40	1.737	2.038	2.334	2.727	3.028	3.335	3.749
50	1.703	1.986	2.263	2.625	2.900	3.177	3.548
75	1.658	1.919	2.170	2.494	2.736	2.977	3.295
100	1.636	1.886	2.124	2.430	2.657	2.881	3.176
200	1.603	1.837	2.058	2.338	2.543	2.744	3.005
500	1.583	1.808	2.019	2.283	2.476	2.664	2.906
1000	1.577	1.798	2.006	2.265	2.454	2.638	2.874
				$df_1 = 12$			
10	2.284	2.913	3.621	4.706	5.661	6.754	8.445
15	2.017	2.475	2.963	3.666	4.250	4.884	5.812
20	1.892	2.278	2.676	3.231	3.678	4.151	4.823
30	1.773	2.092	2.412	2.843	3.179	3.525	4.001
40	1.715	2.003	2.288	2.665	2.953	3.246	3.642
50	1.680	1.952	2.216	2.562	2.825	3.089	3.443
75	1.635	1.884	2.123	2.431	2.661	2.890	3.192
100	1.612	1.850	2.077	2.368	2.583	2.795	3.074
200	1.579	1.801	2.010	2.275	2.468	2.658	2.904
500	1.559	1.772	1.971	2.220	2.402	2.578	2.806
1000	1.552	1.762	1.958	2.203	2.380	2.552	2.774

TABLE C.4. Critical Values of Chi-Square

Table entries are χ^2 values that cut off the upper p proportion of the $\chi^2(df)$ distribution from the rest. Table entries were generated with the IDF.CHISQ function in SPSS.

	***p*-value**						
df	**.100**	**.050**	**.025**	**.010**	**.005**	**.0025**	**.001**
1	2.706	3.841	5.024	6.635	7.879	9.141	10.828
2	4.605	5.991	7.378	9.210	10.597	11.983	13.816
3	6.251	7.815	9.348	11.345	12.838	14.320	16.266
4	7.779	9.488	11.143	13.277	14.860	16.424	18.467
5	9.236	11.070	12.833	15.086	16.750	18.386	20.515
6	10.645	12.592	14.449	16.812	18.548	20.249	22.458
7	12.017	14.067	16.013	18.475	20.278	22.040	24.322
8	13.362	15.507	17.535	20.090	21.955	23.774	26.124
9	14.684	16.919	19.023	21.666	23.589	25.462	27.877
10	15.987	18.307	20.483	23.209	25.188	27.112	29.588
11	17.275	19.675	21.920	24.725	26.757	28.729	31.264
12	18.549	21.026	23.337	26.217	28.300	30.318	32.909
13	19.812	22.362	24.736	27.688	29.819	31.883	34.528
14	21.064	23.685	26.119	29.141	31.319	33.426	36.123
15	22.307	24.996	27.488	30.578	32.801	34.950	37.697
16	23.542	26.296	28.845	32.000	34.267	36.456	39.252
17	24.769	27.587	30.191	33.409	35.718	37.946	40.790
18	25.989	28.869	31.526	34.805	37.156	39.422	42.312
19	27.204	30.144	32.852	36.191	38.582	40.885	43.820
20	28.412	31.410	34.170	37.566	39.997	42.336	45.315
22	30.813	33.924	36.781	40.289	42.796	45.204	48.268
24	33.196	36.415	39.364	42.980	45.559	48.034	51.179
26	35.563	38.885	41.923	45.642	48.290	50.829	54.052
28	37.916	41.337	44.461	48.278	50.993	53.594	56.892
30	40.256	43.773	46.979	50.892	53.672	56.332	59.703
32	42.585	46.194	49.480	53.486	56.328	59.046	62.487
34	44.903	48.602	51.966	56.061	58.964	61.738	65.247
36	47.212	50.998	54.437	58.619	61.581	64.410	67.985
38	49.513	53.384	56.896	61.162	64.181	67.063	70.703
40	51.805	55.758	59.342	63.691	66.766	69.699	73.402
45	57.505	61.656	65.410	69.957	73.166	76.223	80.077
50	63.167	67.505	71.420	76.154	79.490	82.664	86.661
55	68.796	73.311	77.380	82.292	85.749	89.035	93.168
60	74.397	79.082	83.298	88.379	91.952	95.344	99.607
65	79.973	84.821	89.177	94.422	98.105	101.600	105.988
70	85.527	90.531	95.023	100.425	104.215	107.808	112.317
75	91.061	96.217	100.839	106.393	110.286	113.974	118.599
80	96.578	101.879	106.629	112.329	116.321	120.102	124.839
85	102.079	107.522	112.393	118.236	122.325	126.195	131.041
90	107.565	113.145	118.136	124.116	128.299	132.256	137.208
95	113.038	118.752	123.858	129.973	134.247	138.288	143.344
100	118.498	124.342	129.561	135.807	140.169	144.293	149.449

TABLE C.5. Fisher's *r*-to-*Z* Transformation

Table entries are the Fisher *Z*-values corresponding to the correlation that is the sum of the bold values in the row and column in which that *Z*-value resides. For instance, the last entry on this page shows that $Z = 0.863$ when $r = 0.698$.

	.000	**.002**	**.004**	**.006**	**.008**	**.010**	**.012**	**.014**	**.016**	**.018**
.00	0.000	0.002	0.004	0.006	0.008	0.010	0.012	0.014	0.016	0.018
.02	0.020	0.022	0.024	0.026	0.028	0.030	0.032	0.034	0.036	0.038
.04	0.040	0.042	0.044	0.046	0.048	0.050	0.052	0.054	0.056	0.058
.06	0.060	0.062	0.064	0.066	0.068	0.070	0.072	0.074	0.076	0.078
.08	0.080	0.082	0.084	0.086	0.088	0.090	0.092	0.094	0.096	0.098
.10	0.100	0.102	0.104	0.106	0.108	0.110	0.112	0.114	0.117	0.119
.12	0.121	0.123	0.125	0.127	0.129	0.131	0.133	0.135	0.137	0.139
.14	0.141	0.143	0.145	0.147	0.149	0.151	0.153	0.155	0.157	0.159
.16	0.161	0.163	0.165	0.168	0.170	0.172	0.174	0.176	0.178	0.180
.18	0.182	0.184	0.186	0.188	0.190	0.192	0.194	0.196	0.199	0.201
.20	0.203	0.205	0.207	0.209	0.211	0.213	0.215	0.217	0.219	0.222
.22	0.224	0.226	0.228	0.230	0.232	0.234	0.236	0.238	0.241	0.243
.24	0.245	0.247	0.249	0.251	0.253	0.255	0.258	0.260	0.262	0.264
.26	0.266	0.268	0.270	0.273	0.275	0.277	0.279	0.281	0.283	0.286
.28	0.288	0.290	0.292	0.294	0.296	0.299	0.301	0.303	0.305	0.307
.30	0.310	0.312	0.314	0.316	0.318	0.321	0.323	0.325	0.327	0.329
.32	0.332	0.334	0.336	0.338	0.341	0.343	0.345	0.347	0.350	0.352
.34	0.354	0.356	0.359	0.361	0.363	0.365	0.368	0.370	0.372	0.375
.36	0.377	0.379	0.381	0.384	0.386	0.388	0.391	0.393	0.395	0.398
.38	0.400	0.402	0.405	0.407	0.409	0.412	0.414	0.417	0.419	0.421
.40	0.424	0.426	0.428	0.431	0.433	0.436	0.438	0.440	0.443	0.445
.42	0.448	0.450	0.453	0.455	0.457	0.460	0.462	0.465	0.467	0.470
.44	0.472	0.475	0.477	0.480	0.482	0.485	0.487	0.490	0.492	0.495
.46	0.497	0.500	0.502	0.505	0.508	0.510	0.513	0.515	0.518	0.520
.48	0.523	0.526	0.528	0.531	0.533	0.536	0.539	0.541	0.544	0.547
.50	0.549	0.552	0.555	0.557	0.560	0.563	0.565	0.568	0.571	0.574
.52	0.576	0.579	0.582	0.585	0.587	0.590	0.593	0.596	0.599	0.601
.54	0.604	0.607	0.610	0.613	0.616	0.618	0.621	0.624	0.627	0.630
.56	0.633	0.636	0.639	0.642	0.645	0.648	0.650	0.653	0.656	0.659
.58	0.662	0.665	0.669	0.672	0.675	0.678	0.681	0.684	0.687	0.690
.60	0.693	0.696	0.699	0.703	0.706	0.709	0.712	0.715	0.719	0.722
.62	0.725	0.728	0.732	0.735	0.738	0.741	0.745	0.748	0.751	0.755
.64	0.758	0.762	0.765	0.768	0.772	0.775	0.779	0.782	0.786	0.789
.66	0.793	0.796	0.800	0.804	0.807	0.811	0.814	0.818	0.822	0.825
.68	0.829	0.833	0.837	0.840	0.844	0.848	0.852	0.856	0.860	0.863

TABLE C.5. Fisher's *r*-to-*Z* Transformation (continued)

	.000	.002	.004	.006	.008	.010	.012	.014	.016	.018
.70	0.867	0.871	0.875	0.879	0.883	0.887	0.891	0.895	0.899	0.904
.72	0.908	0.912	0.916	0.920	0.924	0.929	0.933	0.937	0.942	0.946
.74	0.950	0.955	0.959	0.964	0.968	0.973	0.978	0.982	0.987	0.991
.76	0.996	1.001	1.006	1.011	1.015	1.020	1.025	1.030	1.035	1.040
.78	1.045	1.050	1.056	1.061	1.066	1.071	1.077	1.082	1.088	1.093
.80	1.099	1.104	1.110	1.116	1.121	1.127	1.133	1.139	1.145	1.151
.82	1.157	1.163	1.169	1.175	1.182	1.188	1.195	1.201	1.208	1.214
.84	1.221	1.228	1.235	1.242	1.249	1.256	1.263	1.271	1.278	1.286
.86	1.293	1.301	1.309	1.317	1.325	1.333	1.341	1.350	1.358	1.367
.88	1.376	1.385	1.394	1.403	1.412	1.422	1.432	1.442	1.452	1.462
.90	1.472	1.483	1.494	1.505	1.516	1.528	1.539	1.551	1.564	1.576
.92	1.589	1.602	1.616	1.630	1.644	1.658	1.673	1.689	1.705	1.721
.94	1.738	1.756	1.774	1.792	1.812	1.832	1.853	1.874	1.897	1.921
.96	1.946	1.972	2.000	2.029	2.060	2.092	2.127	2.165	2.205	2.249
.98	2.298	2.351	2.410	2.477	2.555	2.647	2.759	2.903	3.106	3.453

Appendix D
The Matrix Algebra of Linear Regression Analysis

Advanced readers of this book may be interested in the matrix algebra underlying linear regression analysis. This appendix describes how the regression coefficients, standard errors, and various other statistics in regression analysis are calculated in matrix algebra form.

Let $\mathbf{X}$ be a $N \times (k+1)$ data matrix containing the values of the k regressors in the columns for cases in the rows, including a column of 1's for the regression constant, and let $\mathbf{y}$ be a $N \times 1$ column vector containing values of the dependent variable. For the exercise data in Table 5.1, with weight loss as the dependent variable,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 2 & 15 & 0 \\ 1 & 0 & 4 & 14 & 0 \\ 1 & 0 & 6 & 19 & 0 \\ 1 & 2 & 2 & 15 & 1 \\ 1 & 2 & 4 & 21 & 1 \\ 1 & 2 & 6 & 23 & 0 \\ 1 & 2 & 8 & 21 & 1 \\ 1 & 4 & 4 & 22 & 1 \\ 1 & 4 & 6 & 24 & 0 \\ 1 & 4 & 8 & 26 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 6 \\ 2 \\ 4 \\ 8 \\ 9 \\ 8 \\ 5 \\ 11 \\ 13 \\ 9 \end{bmatrix}$$

where the regressor data are in the order exercise, food intake, metabolism, and sex in the columns of $\mathbf{X}$ after the first column of 1's representing the regressioning constant. The least squares constant and regression coefficients, including the regression constant, are produced with the matrix expression

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

which results in a $(k + 1) \times 1$ vector $\mathbf{b}$

$$\mathbf{b} = \begin{bmatrix} -0.967 \\ 1.151 \\ -1.133 \\ 0.600 \\ -0.404 \end{bmatrix}$$

containing the regression constant as the first entry and the regression coefficients for exercise, food intake, metabolism, and sex in that order.

Estimates of Y from the model are calculated as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

which yields

$$\hat{\mathbf{y}} = \begin{bmatrix} 5.761 \\ 2.895 \\ 3.627 \\ 7.659 \\ 8.991 \\ 8.327 \\ 4.458 \\ 11.892 \\ 11.229 \\ 10.161 \end{bmatrix}$$

With $\hat{\mathbf{y}}$ calculated, residuals can be calculated as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 6 \\ 2 \\ 4 \\ 8 \\ 9 \\ 8 \\ 5 \\ 11 \\ 13 \\ 9 \end{bmatrix} - \begin{bmatrix} 5.761 \\ 2.895 \\ 3.627 \\ 7.659 \\ 8.991 \\ 8.327 \\ 4.458 \\ 11.892 \\ 11.229 \\ 10.161 \end{bmatrix} = \begin{bmatrix} 0.239 \\ -0.895 \\ 0.373 \\ 0.341 \\ 0.009 \\ -0.327 \\ 0.542 \\ -0.892 \\ 1.771 \\ -1.161 \end{bmatrix}$$

and the residual sum of squares as

$$SS_{residual} = \mathbf{e'e}$$

which produces 6.797. Dividing $SS_{residual}$ by the residual degrees of freedom $N - k - 1$ results in $MS_{residual}$, which for this model is 6.797/5 = 1.359.

The expression

$$\mathbf{\Sigma} = MS_{residual}\mathbf{(X'X)}^{-1}$$

is the $(k+1) \times (k+1)$ variance–covariance matrix $\mathbf{\Sigma}$ of the regression constant and regression coefficients

$$\mathbf{\Sigma} = \begin{bmatrix} 11.938 & 1.401 & 0.600 & -0.855 & -1.251 \\ 1.401 & 0.257 & 0.075 & -0.111 & -0.197 \\ 0.600 & 0.075 & 0.100 & -0.062 & -0.006 \\ -0.855 & -0.111 & -0.062 & 0.068 & 0.069 \\ -1.251 & -0.197 & -0.006 & 0.069 & 0.755 \end{bmatrix}$$

The square root of the diagonal elements of $\mathbf{\Sigma}$ is a $k+1$ vector of standard errors of the regression constant and the k regression coefficients:

$$\begin{bmatrix} 3.455 \\ 0.507 \\ 0.316 \\ 0.261 \\ 0.869 \end{bmatrix}$$

The "hat matrix" $\mathbf{H}$ places a significant role in regression algebra. It is defined as:

$$\mathbf{H} = \mathbf{X(X'X)}^{-1}\mathbf{X'}$$

and is an $N \times N$ matrix

$$\hat{\mathbf{H}} = \begin{bmatrix} 0.476 & 0.247 & 0.197 & \cdots & -0.001 & 0.031 & -0.109 \\ 0.247 & 0.544 & 0.170 & \cdots & -0.128 & 0.064 & 0.026 \\ 0.197 & 0.170 & 0.424 & & & -0.074 & 0.039 \\ 0.147 & 0.243 & -0.155 & & & 0.047 & -0.104 \\ \vdots & & & & & & \\ -0.272 & 0.128 & 0.185 & & & -0.110 & 0.118 \\ -0.001 & -0.128 & -0.174 & & & 0.183 & 0.096 \\ 0.031 & 0.064 & -0.074 & \cdots & 0.183 & 0.445 & 0.392 \\ -0.109 & 0.026 & 0.039 & \cdots & 0.096 & 0.392 & 0.467 \end{bmatrix}$$

The H matrix is referred to as a hat matrix because it puts hats on **y**. That is,

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \begin{bmatrix} 5.761 \\ 2.895 \\ 3.627 \\ 7.659 \\ 8.991 \\ 8.327 \\ 4.458 \\ 11.892 \\ 11.229 \\ 10.161 \end{bmatrix}$$

and so **H** can be used to generate estimates of Y without formally calculating the regression coefficients **b**. The diagonal of **H** is an $N \times 1$ vector **h** of leverage values h_i, one for each case in the data:

$$\mathbf{h} = \begin{bmatrix} .476 \\ .544 \\ .424 \\ .548 \\ .580 \\ .320 \\ .780 \\ .417 \\ .445 \\ .467 \end{bmatrix}$$

These leverage values can be used in a number of ways. For instance, the squared standard errors in the diagonal of $\boldsymbol{\Sigma}$ are "best" only when homoscedasticity is met. An alternative standard error estimator that doesn't assume homoscedasticity is calculated by using each case's h_i to adjust the residuals to construct a "heteroscedasticity-consistent" variance–covariance matrix. The HC3 estimator mentioned in section 16.3.1 and implemented in the RLM macro documented in Appendix A is

$$\boldsymbol{\Sigma}_{\text{HC3}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag\left[\frac{e_i^2}{(1-h_i)^2}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

which in these data results in

$$\boldsymbol{\Sigma}_{HC3} = \begin{bmatrix} 15.584 & 1.735 & .930 & -1.069 & -2.492 \\ 1.735 & .435 & .067 & -.119 & -.718 \\ .930 & .067 & .237 & -.109 & .221 \\ -1.069 & -.119 & -.109 & .085 & .082 \\ -2.492 & -.718 & .221 & .082 & 2.343 \end{bmatrix}$$

The square root of the diagonal elements of $\boldsymbol{\Sigma}_{HC3}$ are the HC3 standard errors of the regression coefficients in $\mathbf{b}$:

$$\begin{bmatrix} 3.948 \\ .659 \\ .487 \\ .292 \\ 1.531 \end{bmatrix}$$

More matrix expressions for various regression-analysis and statistics-related computation can be found in Searle (1982).

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129–333.

Ahlberg, J. H., Nilson, E. N., & Walsh, J. L. (1967). *The theory of splines and their application*. New York, NY: Academic Press.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage Publications.

Algina, J., & Moulder, B. C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement*, *61*, 633–649.

Allison, P. D. (1990). Change scores as dependent variables. *Sociological Methodology*, *20*, 93–114.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed.). Cary, NC: SAS Institute.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application* (2nd ed.). Cary, NC: SAS Institute.

Allison, P. D. (2014). *Event history and survival analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*, 1323–1333.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York, NY: Macmillan.

Azen, R., & Budescu, D. V. (2003). Dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*, 129–148.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, *57*, 289–300.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery

rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60–83.

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage Publications.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York, NY: Guilford Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Borooah, V. K. (2001). *Logic and probit: Ordered and multinomial models*. Thousand Oaks, CA: Sage Publications.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*, *26*, 211–252.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). New York, NY: John Wiley & Sons.

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event history modeling: A guide for social scientists*. New York, NY: Cambridge University Press.

Bradburn, M., Clark, T. G., Love, S. B., & Altman, D. G. (2003a). Survival analysis part III: Multivariate data analysis—choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*, 605–611.

Bradburn, M., Clark, T. G., Love, S. B., & Altman, D. G. (2003b). Survival analysis part II: Multivariate data analysis—an introduction to concepts and methods. *British Journal of Cancer*, *89*, 431–436.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*, 1287–1294.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, *28*, 79–87.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542–551.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*, 549–562.

Byrne, B. M. (2009). *Structural equation modeling with AMOS* (2nd ed.). New York, NY: Routledge.

Byrne, B. M. (2012). *Structural equation modeling with Mplus*. New York, NY: Routledge.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). New York, NY: Cambridge University Press.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003a). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*, 232–238.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003b). Survival analysis part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, *89*, 781–786.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. B. (1968). Multiple regression as a general data analytic system. *Psychological Bulletin*, *70*, 426–443.

Cohen, J. B. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 240–253.

Cole, M. S., Walter, F., & Bruch, H. (2008). Affective mechanisms linking dysfunctional behavior to performance in work teams: A moderated mediation study. *Journal of Applied Psychology*, *93*, 945–958.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*, 15–18.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*, *91*, 121–136.

Cribaro-Neto, F. (2004). Asymptotic inference under heteroscedasticity of unknown form. *Computational Statistics and Data Analysis*, *45*, 215–233.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin*, *102*, 414–417.

Cronbach, L. J., & Gleser, C. G. (1965). *Psychological tests and personnel decisions* (2nd ed.). Champaign, IL: University of Illinois Press.

Dana, J., & Thomas, R. (2006). In defense of clinical judgment . . . and mechanical prediction. *Journal of Behavioral Decision Making*, *19*, 413–428.

D'Andrade, R., & Dart, J. (1990). The interpretation of r versus r^2 or why percent of variance accounted for is a poor measure of effect size. *Journal of Quantitative Anthropology*, *2*, 47–59.

Darlington, R. B. (1978). Reduced variance regression. *Psychological Bulletin*, *85*, 1238–1255.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dawes, R. M. (1988). *Rationale choice in an uncertain world*. San Diego, CA: Harcourt Brace Jovanovich.

de Cani, J. S. (1984). Balancing Type I risk and loss of power in ordered Bonferroni procedures. *Journal of Educational Psychology*, *76*, 1036–1037.

Derksen, S., & Keselman, H. J. (1992). Backward, forward, and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265–282.

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, *19*, 589–601.

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated regression models. *Marketing Science*, *26*, 438–445.

Edgington, E. S. (1995). *Randomization tests*. New York, NY: Dekker.

Edmonds, R. (1986). Characteristics of effective schools. In U. Neisser (Ed.), *The school achievement of minority children* (pp. 93–104). Hillsdale, NJ: Lawrence Erlbaum Associates.

Edwards, J. R. (2009). Seven deadly myths of testing moderation in organizational research. In C. E. Lance & R. J. Vanderberg (Eds.), *Statistical and methodological*

myths and urban legends (pp. 143–164). New York, NY: Routledge.
Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*, 1–22.
Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.
Enders, C. K. (2010). *Applied missing data analysis*. New York: NY: Guilford Press.
Enders, W. (2009). *Applied econometric time series* (3rd ed.). New York: NY: John Wiley & Sons.
Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, *10*, 87–99.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analysis. *Behavior Research Methods*, *41*, 1149–1160.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.
Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, *41*, 84–86.
Fox, J. (1991). *Regression diagnostics*. Thousand Oaks, CA: Sage Publications.
Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, and Computers*, *30*, 527–535.
Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, *26*, 797–833.
Fuller, W. A., & Hidiroglou, M. A. (1978). Regression estimation after correction for attenuation. *Journal of the American Statistical Association*, *73*, 99–104.
Ganzach, Y. (1997). Misleading interactions and curvilinear terms. *Psychological Methods*, *2*, 235–247.
Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analysis of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*, 392–404.
Gatsonis, C., & Sampson, A. R. (1989). Multiple regression: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516–524.
Geiser, C. (2012). *Data analysis with Mplus*. New York: NY: Guilford Press.
Good, P. I. (2001). *Resampling methods: A practical guide to data analysis* (2nd ed.). Boston, MA: Birkhauser.
Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.
Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.
Grawitch, M. J., & Munz, D. C. (2004). Are your data nonindependent?: A practical guide to evaluating nonindependence and within-group agreement. *Understanding Statistics*, *3*, 231–257.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, *26*, 499–510.

Greville, T. N. E. (1969). *Theory and applications of spline functions*. New York, NY: Academic Press.

Griffin, D., & Gonzales, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*, *118*, 430–439.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.

Halbesleben, J. R. B. (2010). The role of exhaustion and workaround in predicting occupational injuries: A cross-lagged panel study of health care professionals. *Journal of Occupational Health Psychology*, *15*, 1–16.

Hamilton, D. (1987). Sometimes $r^2 > r^2_{yx_1} + r^2_{yx_2}$: Correlated variables are not always redundant. *American Statistician*, *41*, 129–132.

Hayes, A. F. (2005). *Statistical methods for communication science*. New York, NY: Routledge.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*, 385–410.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*, 1–22.

Hayes, A. F., & Cai, L. (2007). Using heteroscedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Hayes, A. F., Glynn, C. J., & Huge, M. E. (2012). Cautions regarding the interpretation of regression coefficients and hypothesis tests in linear models with interactions. *Communication Methods and Measures*, *6*, 1–11.

Hayes, A. F., Glynn, C. J., & Shanahan, J. E. (2005). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research*, *17*, 298–323.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*, 924–936.

Hayes, A. F., & Preacher, K. J. (2010). Estimating and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, *45*, 627–660.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 451–460.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, *24*, 1918–1927.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandanivian Journal of Statistics*, *6*, 65–70.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time to event data* (2nd ed.). New York, NY: John Wiley & Sons.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). New York, NY: John Wiley & Sons.

Hox, J. J., Moerbeek, M., & Schoot, R. v. (2002). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Humphreys, L. G. (1978). Research on individual differences requires correlational analysis, not ANOVA. *Intelligence*, *2*, 1–5.

Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, *38*, 100–109.

Irwin, J. R., & McClelland, G. H. (2002). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, *40*, 366–371.

Jaccard, J. (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage Publications.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Kam, C. D., & Franzese, R. J. (2007). *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor, MI: University of Michigan.

Kaufman, R. L. (2013). *Heteroskedasticity in regression*. Thousand Oaks, CA: Sage Publications.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*, 979–984.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*, 422–431.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*, 126–137.

Kim, J., & Mueller, C. M. (1978). *Introduction to factor analysis: What it is and how to do it*. Thousand Oaks, CA: Sage Publications.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*, 296–310.

Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement*, *58*, 42–68.

Kuss, O. (2013). The danger of dichotomizing continuous variables: A visualization. *Teaching of Statistics*, *35*, 78–79.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, *7*, 409-443.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*, 385–393.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Lunneborg, C. E. (2000). *Data analysis by resampling*. Pacific Grove, CA: Duxbury.

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, *118*, 405–421.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

Marsh, L. C., & Cormier, D. R. (2002). *Spline regression models*. Thousand Oaks, CA: Sage University Press.

Maxwell, S. E., Cole, D. A., & Melissa, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, *46*, 816–841.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.

Meehl, P. E. (1957). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Milhoj, A. (2013). *Practical time series analysis*. Cary, NC: SAS Institute.

Mook, D. G. (1987). In defense of external invalidity. *American Psychologist*, *38*, 379–387.

Noreen, E. (1988). An empirical comparison of probit and OLS regression hypothesis tests. *Journal of Accounting Research*, *26*, 119–133.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

O'Connell, A. (2005). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage Publications.

O'Connor, B. P. (2004). SPSS and SAS programs for addressing interdependence and basic levels-of-analysis issues in psychological data. *Behavior Research Methods, Instruments, and Computers*, *36*, 17–28.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211.

Orme, J. G., & Combs-Orme, T. (2009). *Multiple regression with discrete dependent variables*. Oxford, UK: Oxford University Press.

Ostrom, C. W. (1990). *Time series analysis: Regression techniques* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, *97*, 307–315.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage Publications.

Pitts, B. L., & Safer, M. A. (2016). Retrospective appraisals mediate the effects of combat experiences on PTS and depression symptoms in U.S. Army medics. *Journal of Traumatic Stress*, *29*, 65–71.

Potter, W. J., Cooper, R., & Dupagne, M. (1993). The three paradigms of mass media research in mainstream communication journals. *Communication Theory*, *3*, 317–335.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*, 717–731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*, 77–98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). New York: NY: Psychology Press.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441–456.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166–169.

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*, *76*, 1028–1034.

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, *25*, 666–678.

Ryan, T. A. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. *Psychological Bulletin*, *57*, 318–328.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York, NY: John Wiley & Sons.

Shaw, M., Mitchell, R., & Dorling, D. (2000). Time for a smoke?: One cigarette reduces your life by 11 minutes. *British Medical Journal*, *320*, 53.

Sheih, G. (2011). Clarifying the role of mean centering in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, *64*, 462–477.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*, 323–355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290–312). San Francisco, CA: Jossey-Bass.

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, *50*, 277–288.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coursely categorized. *Educational and Psychological Measurement*, *66*, 228–239.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.

Tofighi, D., & MacKinnon, D. P. (2011). Rmediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, *43*, 692–700.

Traub, R. E. (1994). *Reliability for the social sciences*. Thousand Oaks, CA: Sage Publications.

Tu, Y.-K., Gunnel, D., & Gilthorpe, M. S. (2008). Simpon's paradox, Lord's paradox, and suppression effects are the same phenomenon—the reversal paradox. *Emerging Themes in Epidemiology*, *5*, 1–9.

Valenta, Z., Pitha, J., & Poledne, R. (2006). Proportional odds logistic regression—effective means of dealing with limited uncertainty in dichotomizing clinical outcomes. *Statistics in Medicine*, *25*, 4227–4234.

van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonstandomized studies: The difference. *Multivariate Behavioral Research*, *48*, 895–922.

Veiel, H. O. F. (1988). Base-rates, cut-points and interaction effects: The problem with dichotomized continuous variables. *Psychological Medicine*, *18*, 703–710.

Whisman, M. A., & McClelland, G. H. (2005). Designing, testing, and interpreting interactions and moderator effects in family research. *Journal of Family Psychology*, *19*, 111–120.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Wiggins, J. F. (1973). *Personality and prediction*. Reading, MA: Addison-Wesley.

Witkiewitz, K., & Bowen, S. (2010). Depression, craving, and substance use following a randomized trial of mindfulness-based relapse prevention therapy.

Journal of Consulting and Clinical Psychology, *78*, 362–374.

Yaffee, R. A., & McGee, M. (2000). *An introduction to time series analysis and forecasting: With applications of SAS and SPSS*. New York, NY: Academic Press.

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education*, *69*, 203–224.

Author Index

Note: *t* following a reference indicates a table, *f* following a reference indicates a figure, and *n* following a reference indicates a note.

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

V

W

Y

Z

Subject Index

Note: *t* following a reference indicates a table, *f* following a reference indicates a figure, and *n* following a reference indicates a note.

B

C

D

E

F

G

H

I

J

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z

About the Authors

Richard B. Darlington, PhD, is Emeritus Professor of Psychology at Cornell University. He is a Fellow of the American Association for the Advancement of Science and has published extensively on regression and related methods, the cultural bias of mental tests, the long-term effects of preschool programs, and, most recently, the neuroscience of brain development and evolution.

Andrew F. Hayes, PhD, is Professor of Quantitative Psychology at The Ohio State University. His research and writing on data analysis have been published widely, and he is the author of *Introduction to Mediation, Moderation, and Conditional Process Analysis* and *Statistical Methods for Communication Science*. Dr. Hayes teaches data analysis, primarily at the graduate level, and frequently conducts workshops on statistical analysis throughout the world. His website is *www.afhayes.com*.