

Data Analysis

A Model Comparison Approach to
Regression, ANOVA, and Beyond

Third Edition

**Charles M. Judd,
Gary H. McClelland, and
Carey S. Ryan**



Data Analysis

Third Edition

Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond is an integrated treatment of data analysis for the social and behavioral sciences. It covers all of the statistical models normally used in such analyses, such as multiple regression and analysis of variance, but it does so in an integrated manner that relies on the comparison of models of data estimated under the rubric of the general linear model.

Data Analysis also describes how the model comparison approach and uniform framework can be applied to models that include product predictors (i.e., interactions and nonlinear effects) and to observations that are nonindependent. Indeed, the analysis of nonindependent observations is treated in some detail, including models of non-independent data with continuously varying predictors as well as standard repeated measures analysis of variance. This approach also provides an integrated introduction to multilevel or hierarchical linear models and logistic regression. Finally, *Data Analysis* provides guidance for the treatment of outliers and other problematic aspects of data analysis. It is intended for advanced undergraduate and graduate level courses in data analysis and offers an integrated approach that is very accessible and easy to teach.

Highlights of the third edition include:

- a new chapter on logistic regression;
- expanded treatment of mixed models for data with multiple random factors;
- updated examples;
- an enhanced website with PowerPoint presentations and other tools that demonstrate the concepts in the book; exercises for each chapter that highlight research findings from the literature; data sets, R code, and SAS output for all analyses; additional examples and problem sets; and test questions.

Charles “Chick” M. Judd is Professor of Distinction in the College of Arts and Sciences at the University of Colorado at Boulder. His research focuses on social cognition and attitudes, intergroup relations and stereotypes, judgment and decision making, and behavioral science research methods and data analysis.

Gary H. McClelland is Professor of Psychology at the University of Colorado at Boulder. A Faculty Fellow at the Institute of Cognitive Science, his research interests include judgment and decision making, psychological models of economic behavior, statistics & data analysis, and measurement and scaling.

Carey S. Ryan is a Professor in the Department of Psychology at the University of Nebraska at Omaha. She has research interests in stereotyping and prejudice, group processes, and program evaluation.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Data Analysis

A Model Comparison Approach to
Regression, ANOVA, and Beyond

Third Edition

Charles M. Judd
University of Colorado

Gary H. McClelland
University of Colorado

Carey S. Ryan
University of Nebraska at Omaha

Third edition published 2017
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2017 Taylor & Francis

The right of Charles M. Judd, Gary H. McClelland, and Carey S. Ryan to be identified as authors of this work has been asserted by them in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

First edition published by Harcourt Brace Jovanovich 1989

Second edition published by Routledge 2009

Library of Congress Cataloging in Publication Data

Names: Judd, Charles M., author. | McClelland, Gary H., 1947– author. | Ryan, Carey S., author.

Title: Data analysis : a model comparison approach to regression, ANOVA, and beyond / Charles M. Judd, University of Colorado, Gary H. McClelland, University of Colorado, Carey S. Ryan, University of Nebraska at Omaha.

Description: Third Edition. | New York : Routledge, 2017. | Revised edition of Data analysis, 2007. | Includes bibliographical references and indexes.

Identifiers: LCCN 2016055541 | ISBN 9781138819825 (hardback : alk. paper) | ISBN 9781138819832 (pbk. : alk. paper) | ISBN 9781315744131 (ebook)

Subjects: LCSH: Mathematical statistics.

Classification: LCC QA276 .J83 2017 | DDC 519.5/36—dc23

LC record available at <https://lcn.loc.gov/2016055541>

ISBN: 978-1-138-81982-5 (hbk)

ISBN: 978-1-138-81983-2 (pbk)

ISBN: 978-1-315-74413-1 (ebk)

Typeset in Times New Roman
by Florence Production Ltd, Stoodleigh, Devon, UK

Visit the companion website: www.dataanalysisbook.com

Contents

<i>Preface</i>	vii
1 Introduction to Data Analysis	1
2 Simple Models: Definitions of Error and Parameter Estimates	10
3 Simple Models: Models of Error and Sampling Distributions	25
4 Simple Models: Statistical Inferences about Parameter Values	43
5 Simple Regression: Estimating Models with a Single Continuous Predictor	72
6 Multiple Regression: Models with Multiple Continuous Predictors	103
7 Moderated and Nonlinear Regression Models	135
8 One-Way ANOVA: Models with a Single Categorical Predictor	168
9 Factorial ANOVA: Models with Multiple Categorical Predictors and Product Terms	205
10 ANCOVA: Models with Continuous and Categorical Predictors	229
11 Repeated-Measures ANOVA: Models with Nonindependent Errors	260
12 Incorporating Continuous Predictors with Nonindependent Data: Towards Mixed Models	292
13 Outliers and Ill-Mannered Error	314
14 Logistic Regression: Dependent Categorical Variables	339
<i>References</i>	354
<i>Appendix</i>	357
<i>Author Index</i>	362
<i>Subject Index</i>	363



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Preface

We are stunned that twenty-seven years after publishing the first edition of our text, we find ourselves preparing this third edition. The first edition resulted from our teaching a graduate level course in data analysis in the 1980s at the University of Colorado (Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. Orlando, FL: Harcourt Brace Jovanovich). The second edition was published in 2008, with the invaluable assistance of Carey Ryan (Judd, C. M., McClelland, G. H., & Ryan, C. S. (2008). *Data analysis: A model comparison approach*. New York, NY: Routledge). Now we once again find ourselves happy to see the continuing interest in our approach and our book, as revealed by the continuing course adoptions and the enthusiasm of our publisher.

There is much that is new and different in this third edition. At a later point in this Preface, we make clear the significant changes that we have undertaken. At the same time, our basic model comparison approach and our way of thinking about data have not changed at all and remain the core of the book. Accordingly, we reproduce below parts of the Preface of the second edition, making clear our assumptions and our approach from the very beginning.

GOALS AND ASSUMPTIONS

Statistics courses, textbooks, and software are usually organized in the same way that a cookbook is organized. Typically, various recipes are given in different chapters for different kinds of research designs or data structures. In fact, numerous statistics books include a chart at the beginning, pointing the reader to various chapters and ostensibly different statistical procedures, depending on what their data look like and the kinds of questions that they wish to answer. As a result, social and behavioral scientists, as consumers of statistics, typically organize their statistical knowledge in much the same way: “With this sort of data, I know to do this test. With that sort of data, I know to do that test.”

This book has been written under the assumption that this sort of organization for statistical knowledge has distinct costs. To extend the cookbook analogy, cooks who rely on cookbooks do not know how to proceed when they wish to prepare a dish for which no recipe has been included. When students are confronted with data that do not fit nicely into one of the categories for which they have learned a statistical procedure, frustration and error frequently occur. The student may proceed to use the wrong test, making inappropriate assumptions about the structure of the data in order to use a well-

learned statistical procedure. We have encountered students who have bludgeoned their data to fit an orthogonal analysis of variance with equal sample sizes because it is the only test they know. We have also heard some of our colleagues moan that a given data analysis problem would be simple if only they had a particular specialized computer program that was available at their previous university.

A totally different sort of organization forms the basis of this book. Our focus is on how statistical procedures can be seen as tools for building and testing models of data. A consistent framework is used throughout the book to develop a few powerful techniques for model building and statistical inference. This framework subsumes all of the different statistical recipes that are normally found in the cookbooks, but it does so in a totally integrated manner. As we build and test models, we develop a consistent vocabulary to refer to our models and inference procedures. At each step, however, we clearly show the reader how our vocabulary and models can be translated back into the recipes of the old cookbooks.

We are convinced that this integrated framework represents a better way to learn and think about data analysis. Students who understand data analysis through our framework will be able to ask the questions of their data that they want to ask instead of the questions that the designer of some textbook or of some computer statistical package assumed they would want to ask. In the end, students will use theory and their own intelligence to guide their data analysis instead of turning over the analysis to a cookbook set of procedures or to a computer program. Intelligent data analysis is our goal. Rather than applying recipes that have been learned by rote, we wish to enable students to write their own recipes as they do data analysis.

A few words are in order concerning our assumptions about the reader and the structure of the book. We assume that the reader is a student or professional researcher in the behavioral or social sciences, business, education, or a related field who wants to analyze data to answer significant theoretical or practical questions in a substantive discipline. We assume that the reader does not want to become a statistician but is more interested in the substantive discipline, with data analysis being a tool rather than an end in itself.

We assume that the reader is reasonably facile with basic algebraic manipulations because almost all of our models for data are expressed algebraically. But we do not assume any other mathematical training such as calculus or matrix algebra. A very important assumption is that the reader has access to a good multiple regression program in a statistical software package. But, importantly, we do not care which one.

The book is appropriate for semester or year-long statistics and data analysis courses emphasizing multiple regression and/or analysis-of-variance and taught at the upper undergraduate or graduate levels. Previous editions have been successfully used with students from psychology, sociology, anthropology, political science, linguistics, cognitive science, neuroscience, biology, geology, geography, museum sciences, applied mathematics, marketing, management science, and organizational behavior.

Our assumptions about the reader and our goal of training data analysts instead of statisticians have prompted several decisions about the structure of the book:

1. We present only enough mathematical derivations to provide conceptual clarity. We will generally assume that the mathematical statisticians have done their job correctly, and so we will use many of their results without proving them ourselves. At the

same time, however, we cannot abandon all mathematical details because it is extremely important that the data analyst be able to recognize when the data analysis and particularly the model of the data are inappropriate or in need of modification. The choice of which derivations to include is therefore guided by the goal of training educated data analysts and not mathematical statisticians.

2. We let the computer do the computational work. Most statistics cookbooks present many different formulas for each statistic. The different formulas facilitate hand calculation for different organizations of the raw data. We assume that the reader is interested in the science and not the calculations and will use the most efficient means to do the calculations—the computer. Ironically, the old hand-calculation formulas can be disastrous when implemented in computer programs because of problems of rounding errors and other computer idiosyncrasies. Neither the computational formulas that are best for computers nor the hand-calculation formulas are very conceptual. Hence, we avoid both computer- and hand-oriented formulas. We present instead formulas that emphasize the concepts but may not be useful for direct computations either by hand or by computer.
3. We try as much as possible to work from the general to the specific so that an integrated framework emerges. Many statistics books begin with details and simple statistics and then slowly build up to more general models, changing concepts and notation frequently along the way. Although we begin with simple models of data and work up, we do so within the context of a consistent overall framework, and we present the simple models using the same concepts and notation that we use with the more complex models presented later. Most importantly, we use the same inferential statistics throughout.
4. We do not try to cover all of statistics. There are many statistical procedures that we have left out either because they are infrequently used or because the same thing can be accomplished with a more general model that we do cover. We provide the data analyst with a limited stock of models and statistical procedures, but the models we do provide are quite general and powerful. The goal is to learn a few powerful techniques very thoroughly.
5. Our framework for data analysis is consistent with what has been termed the regression approach or the general linear model. Historically, the regression approach has often been contrasted, erroneously, with an analysis-of-variance approach. And, in this tradition, regression has been more often associated with surveys, quasi-experiments, and nonexperimental data, while analysis of variance has more frequently been used for randomized, laboratory experiments. These historical associations caused many people to have a false belief that there is a fundamental difference between the two approaches. We adopt the more general regression approach and show how all of analysis of variance can be accomplished within this framework. This has the important advantage of reducing the number of specialized statistical techniques that must be learned for particular data analysis problems. More importantly, we show how the regression approach provides more control in the analysis of experimental data so that we can ask specific, theoretically motivated questions of our data—questions that often differ from the standard questions that the procedures of a traditional analysis of variance presume we want to ask.
6. We had a difficult decision to make about the vocabulary to be used in this book. On the one hand, we are convinced that the traditional names for the various

statistical tests perpetuate false distinctions. For that reason, we would have preferred to abandon them in favor of a shorter list of names from our integrated framework. However, to do so would have been irresponsible because many readers of this book will have had an introductory course using the traditional names, will have to talk to colleagues who have not read this book, and will have to read journal reports of data analysis that perpetuate the traditional names. We therefore do both: Consistent terminology and notation are used throughout the book to provide an integrated framework, but we also provide translations, usually at the beginning and end of each chapter, between our terms and their traditional counterparts.

HOW THE BOOK IS ORGANIZED AND WHAT IS NEW THIS TIME AROUND

The preface or introduction in most statistics textbooks describes alternative orders in which the chapters may be studied, which is precisely what we think is wrong with most statistics textbooks. When each chapter or section is self-contained, students are often befuddled by the abrupt change of topics across chapter boundaries. To avoid these abrupt changes, we have organized the chapters in a specific sequence and used a unifying framework for data analysis throughout. Each chapter generally focuses on a single concept, but the understanding of that concept will depend on the concepts developed in the previous chapters. Thus, we do not think the chapters should be read in any order other than the one in which we present them. This means, of course, that our book cannot be used as a cookbook. But then, we have made clear that we think the cookbook organization of statistical knowledge is misguided.

While the reader who uses our book to look up a specific test may be frustrated by the sequential organization of our book, we are convinced that this organization represents a better way to learn statistics. It substantially reduces the number of separate concepts that must be mastered and allows progressive accumulation of integrated knowledge about data analysis rather than disconnected amassing of discrete recipes.

Chapter 1 presents an overview of data analysis through the concepts of data, model, and error and integrates them in the equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

The concept of proportional reduction in error (PRE) is introduced as a tool for sorting our way through alternative models for a given set of data. Chapter 2 considers alternative definitions of error and shows the implications for the model for each of the alternative definitions. In the course of examining the alternative definitions of error, we develop the usual descriptive statistics for location and spread. However, we prefer to view these descriptive statistics in terms of models of data. Chapter 3 introduces a small set of reasonable assumptions about error and uses those assumptions, in conjunction with the concept of sampling distributions of statistics, to choose from among the alternative definitions of error offered in Chapter 2. The sum of squared errors (SSE) is chosen to be our standard definition of error. This chapter also introduces the concept of sampling distributions of statistics. New to this edition, this chapter also contains an intuitive explanation of maximum likelihood estimation. Chapter 4 uses the material of

Chapter 3 to make inferences about models and their parameters. The inferential test developed in this chapter is equivalent to a test usually identified as the one-sample t -test, but we present it in a more general form that we are able to use in all later chapters. All the necessary concepts for statistical inference used throughout the book are contained in Chapters 2, 3, and 4, although the models that are explored in these chapters are very simple.

The remainder of the book is mostly about building models that vary in their complexity and appropriateness for different kinds of data. Chapter 5 is about models that make conditional predictions of data based on a single continuous variable. Such models are usually referred to as simple regression. This chapter uses the inferential techniques developed in Chapter 4 to ask questions about the simple regression models and their parameters. Chapter 6 extends the concepts of Chapter 5 to models making conditional predictions of data based upon multiple continuous variables. Elsewhere, the material in this chapter is often referred to as multiple regression. An important concept in this chapter is redundancy among predictors. Chapter 7 introduces the important technique of including multiplicative products of continuous predictor variables in models. We then use that technique to develop curvilinear models and to consider the interaction between two or more predictor variables. Chapters 5–7 contain all the essential information for what is generally known as regression analysis.

Chapters 8–11 consider models that use categorical variables to make conditional predictions of data. Such models are often known as the analysis of variance. The key link between this material and the previous chapters is provided by Chapter 8, which introduces contrast codes. The use of contrast codes allows all the model-building and testing techniques of regression analysis (Chapters 5–7) to be applied without modification to models with categorical predictor variables. Chapter 8 develops these important contrast codes for the case of single categorical predictor variables, often known as one-way analysis of variance. Chapter 9 shows how models using products, developed in Chapter 7, can be applied to the contrast codes of Chapter 8 to produce models of increasing complexity when there are multiple categorical predictor variables. These models are often referred to as analysis of variance for factorial designs. Chapter 10 makes the easy generalization to models that include both continuous and categorical predictor variables; these models are sometimes known as analysis of covariance and, more recently, mediational analyses. Also included here are models that include products of continuous and categorical variables, models known elsewhere as moderated regression models.

An important assumption underlying statistical inference with all the models is that each data observation is independent from all other observations. This assumption is unlikely to be true when more than one observation comes from each unit of analysis (e.g., when each person or group provides more than one data observation). Chapters 11 and 12 present an approach for building models in such situations that is entirely consistent with the material in previous chapters. This problem is known in psychology as within-subject or repeated measures analysis of variance, in contrast to the models in Chapters 8–10 that apply to between-subject analysis of variance. More generally, this problem is referred to as nonindependent or correlated errors. Chapter 11 presents the basics of our approach for analyzing data with correlated errors. This approach consists of developing a separate set of contrast codes for the multiple observations from each unit of analysis, using those codes to construct new data variables, and then using

the machinery of previous chapters to model those new data variables. Chapter 12 extends that approach to what might be called within-subject regression by including continuous predictor variables that can vary either between or within the nonindependent observations in a dataset. Included here is an introduction to what are called linear mixed models, that allow random error variance to accrue from multiple sources rather than a single one. By extension we provide an introduction to what are called multilevel or hierarchical models as well as models for data with crossed random factors. This material is new to this edition.

Chapter 13 considers the problems of outliers and other violations of the assumptions about error made in Chapter 3. We use the concepts developed in previous chapters to develop an outlier model that provides a principled method for detecting outliers and mitigating their effects on the model for the remainder of the data. This chapter also considers techniques for the detection and treatment of violations of assumptions underlying the procedures for model building and statistical inference presented in the previous chapters. The effect of all the remediation techniques presented is to transform data or otherwise restructure the problem so that all the machinery previously developed may be appropriately applied.

New to this edition is Chapter 14 which deals with logistic regression, permitting the analysis of dichotomous rather than continuous outcome or dependent variables. Although the inferential machinery for these models is rather different, the basic model comparison approach that we have developed has a nice parallel in logistic models.

Also new to this edition are exercises that supplement each chapter. These exercises and their answers can be found at our supplementary website, www.dataanalysisbook.com/.

Our enhanced website, www.dataanalysisbook.com/ also contains additional materials for instructors and students including PowerPoint presentations and other tools that demonstrate the concepts in the book, datasets, R code and SAS output for all analyses, additional examples and problem sets, and test questions.

In preparing this revision we realize the huge debt that we owe all of our students over the years. We thank them for the many ways in which they have made us think harder and deeper as teachers and data analysts.

Introduction to Data Analysis

1

This book is about data analysis. In the social and behavioral sciences we often collect batches of data that we hope will answer questions, test hypotheses, or disprove theories. To do so we must analyze our data. In this chapter, we present an overview of what data analysis means. This overview is intentionally abstract with few details so that the “big picture” will emerge. Data analysis is remarkably simple when viewed from this perspective, and understanding the big picture will make it much easier to comprehend the details that come later.

OVERVIEW OF DATA ANALYSIS

The process of data analysis is represented by the following simple equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

DATA represents the basic scores or observations, usually but not always numerical, that we want to analyze. MODEL is a more compact description or representation of the data. Our data are usually bulky and of a form that is hard to communicate to others. The compact description provided by the model is much easier to communicate, say, in a journal article, and is much easier to think about when trying to understand phenomena, to build theories, and to make predictions. To be a representation of the data, all the models we consider will make a specific prediction for each observation or element in DATA. Models range from the simple (making the same prediction for every observation in DATA) to the complex (making differential predictions conditional on other known attributes of each observation). To be less abstract, let us consider an example. Suppose our data were, for each state in the United States, the percentage of households that had internet access in the year 2013; these data are listed in Figure 1.1. A simple model would predict the same percentage for each state. A more complex model might adjust the prediction for each state according to the age, educational level, and income of the state’s population, as well as whether the population is primarily urban or rural. The amount by which we adjust the prediction for a particular attribute (e.g., educational level) is an unknown *parameter* that must be estimated from the data.

The last part of our basic equation is ERROR, which is simply the amount by which the model fails to represent the data accurately. It is an index of the degree to which the model mispredicts the data observations. We often refer to error as the *residual*—the part that is left over after we have used the model to predict or describe the data. In other words:

$$\text{ERROR} = \text{DATA} - \text{MODEL}$$

FIGURE 1.1 Percentage of households that had internet access in the year 2013 by US state

<i>i</i>	<i>US State</i>	<i>Percentage</i>	<i>i</i>	<i>US State</i>	<i>Percentage</i>
1	AK	79.0	26	MT	72.1
2	AL	63.5	27	NC	70.8
3	AR	60.9	28	ND	72.5
4	AZ	73.9	29	NE	72.9
5	CA	77.9	30	NH	80.9
6	CO	79.4	31	NJ	79.1
7	CT	77.5	32	NM	64.4
8	DE	74.5	33	NV	75.6
9	FL	74.3	34	NY	75.3
10	GA	72.2	35	OH	71.2
11	HI	78.6	36	OK	66.7
12	IA	72.2	37	OR	77.5
13	ID	73.2	38	PA	72.4
14	IL	74.0	39	RI	76.5
15	IN	69.7	40	SC	66.6
16	KS	73.0	41	SD	71.1
17	KY	68.5	42	TN	67.0
18	LA	64.8	43	TX	71.8
19	MA	79.6	44	UT	79.6
20	MD	78.9	45	VA	75.8
21	ME	72.9	46	VT	75.3
22	MI	70.7	47	WA	78.9
23	MN	76.5	48	WI	73.0
24	MO	69.8	49	WV	64.9
25	MS	57.4	50	WY	75.5

The goal of data analysis is then clear: We want to build the model to be a good representation of the data by making the error as small as possible. In the unlikely extreme case when $\text{ERROR} = 0$, DATA would be perfectly represented by MODEL .

How do we reduce the error and improve our models? One way is to improve the quality of the data so that the original observations contain less error. This involves better research designs, better data collection procedures, more reliable instruments, etc. We do not say much about such issues in this book, but instead leave those problems to texts and courses in experimental design and research methods. Those problems tend to be much more discipline specific than the general problems of data analysis and so are best left to the separate disciplines. Excellent sources that cover such issues are Campbell and Stanley (1963), Cook and Campbell (1979), Judd and Kenny (1981a), Maruyama and Ryan (2014), Reis and Judd (2014), Rosenthal and Rosnow (2008), and Shadish, Cook, and Campbell (2002). Although we often note some implications of data analysis procedures for the wise design of research, we in general assume that the data analyst is confronted with the problem of building the best model for data that have already been collected.

The method available to the data analyst for reducing error and improving models is straightforward and, in the abstract, the same across disciplines. Error can almost always be reduced (never increased) by making the model's predictions conditional on additional information about each observation. This is equivalent to adding parameters to the model and using data to build the best estimates of those parameters. The meaning of "best

estimate” is clear: we want to set the parameters of the model to whatever values will make the error the smallest. The estimation of parameters is sometimes referred to as “fitting” the model to the data. Our ideal data analyst has a limited variety of basic models. It is unlikely that any of these models will provide a good fit “off the rack”; instead, the basic model will need to be fitted or tailored to the particular size and bulges of a given data customer. In this chapter, we are purposely vague about how the error is actually measured and about how parameters are actually estimated to make the error as small as possible because that would get us into details to which we devote whole chapters later. But for now the process in the abstract ought to be clear: add parameters to the model and estimate those parameters so that the model will provide a good fit to the data by making the error as small as possible.

To be a bit less abstract, let us again consider the example of internet access by state. An extremely simple model would be to predict a priori (that is, without first examining the data) that in each state the percentage of households that has internet access is 75. This qualifies as a model according to our definition, because it makes a prediction for each of the 50 states. But in this model there are no parameters to be estimated from the data to provide a good fit by making the error as small as possible. No matter what the data, our model predicts 75. We will introduce some notation so that we have a standard way of talking about the particulars of DATA, MODEL, and ERROR. Let Y_i represent the i th observation in the data; in this example Y_i is simply the percentage of households that have internet access for the i th state. Then our basic equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

for this extremely simple model becomes:

$$Y_i = 75 + \text{ERROR}$$

We can undoubtedly improve our model and reduce the error by using a model that is still simple but has one parameter: predict that the percentage is the same in all states, but leave the predicted value as an unspecified parameter to be estimated from the data. For example, the average of all 50 percentages might provide a suitable estimate. We will let β_0 represent the unknown value that is to be estimated so that our slightly more complex, but still simple, model becomes:

$$Y_i = \beta_0 + \text{ERROR}$$

It is important to realize that we can never know β_0 for certain; we can only estimate it.

We can make our model yet more complex and reduce the error further by adding more parameters to make *conditional predictions*. For example, innovations reputedly are adopted on the east and west coasts before the middle of the country. We could implement that in a model that starts with a basic percentage of internet use (β_0) for all states, which is adjusted upward by a certain amount (β_1) if the state is in the Eastern or Pacific time zones and reduced by that same amount if the state is in the Central or Mountain time zones. More formally, our basic equation now has a more complex representation, namely:

$$Y_i = \beta_0 + \beta_1 + \text{ERROR} \text{ if the state is in the Eastern or Pacific time zones}$$

$$Y_i = \beta_0 - \beta_1 + \text{ERROR} \text{ if the state is in the Central or Mountain time zones}$$

In other words, our model and its prediction would be conditional on the time zone in which the state is located.

Another slightly more complex model would make predictions conditional on a continuous, rather than a categorical, predictor. For example, we might make predictions conditional on the proportion of college graduates in a state, presuming that college graduates are more likely to be internet users. We again start with a basic percentage of internet users (β_0) for all states, which is adjusted upward by a certain amount (β_1) for each percentage point a state's proportion of college graduates is above the national average and reduced by the same amount for each percentage point a state's proportion of college graduates is below the national average. More formally, letting X_i represent the amount a state's proportion of college graduates is above or below the national average:

$$Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$$

In words, the percentage of college graduates is the condition in this model on which we base our differential or conditional prediction of a state's internet use.

We can continue making our model yet more complex by adding parameters to make similar adjustments for income, urban versus rural population, etc. By so doing we will be adding still more implicit hypotheses to the model.

It might appear that the best strategy for the data analyst would be to add as many parameters as possible, but this is not the case. The number of observations in DATA imposes an inherent limit on the number of parameters that may be added to MODEL. At the extreme, we could have separate parameters in our model for each observation and then estimate the value of each such parameter to be identical to the value of its corresponding DATA observation. For example, our prediction might contain statements such as, *if the state is Kentucky, then estimate its parameter to be 68.5*, which is the percentage of households in Kentucky that have internet access. That procedure would clearly reduce the error to zero and provide a perfect fit. But such a model would be uninteresting because it would simply be a duplicate of data and would provide no new insights, no bases for testing our theories, and no ability to make predictions in slightly different circumstances. A paramount goal of science is to provide simple, parsimonious explanations for phenomena. A model with a separate parameter for each observation is certainly not parsimonious. Our ideal model, then, is a compact description of the data and has many fewer parameters than the number of observations in data.

We now have an obvious conflict. The goal of reducing the error and providing the best description of DATA leads us to add parameters to the model. On the other hand, the goal of parsimony and the desire for a compact, simple model lead us to remove parameters from the model. The job of the data analyst is to find the proper balance between these two conflicting objectives. Thus, the ultimate goal is to find the smallest, simplest model that provides an adequate description of the data so that the error is not too large ("too large" will be defined later). In still other words, the data analyst must answer the question of whether it is worthwhile to add yet more parameters to a model.

Returning to the example of internet access, we will want to ask whether the extra complexity of making predictions conditional on time zone, educational level, income, urban versus rural population, etc. is worth the trouble. By so doing, we will simultaneously be asking whether the hypotheses implicit in the more complex models are true. For example, if we decide that conditioning our prediction of internet access on

the percentage of college graduates is not worthwhile, then we will have effectively rejected the hypothesis that college education is related to higher internet access. This is the essence of testing hypotheses.

Although we are still being vague about how to measure the error, we can be more precise about what we mean by: “Are more parameters worthwhile?” We will call the model without the additional parameters the *compact model* and will refer to it as Model C. The alternative, *augmented model*, Model A, includes all the parameters, if any, of Model C plus some additional parameters. The additional parameters of Model A may reduce the error or leave it unchanged; there is no way the additional parameters can increase the error. So it must be that:

$$\text{ERROR}(A) \leq \text{ERROR}(C)$$

where $\text{ERROR}(A)$ and $\text{ERROR}(C)$ are the amounts of error when using Models A and C, respectively. The question of whether it is worthwhile to add the extra complexity of Model A now reduces to the question of whether the difference between $\text{ERROR}(C)$ and $\text{ERROR}(A)$ is big enough to worry about. It is difficult to decide based on the absolute magnitude of the errors. We will therefore usually make relative comparisons. One way to do that is to calculate the *proportional reduction in error* (PRE), which represents the proportion of Model C’s error that is reduced or eliminated when we replace it with the more complex Model A. Formally:

$$\text{PRE} = \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

The numerator is simply the difference between the two errors (the amount of error reduced) and the denominator is the amount of error for the compact model with which we started. An equivalent expression is:

$$\text{PRE} = 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

If the additional parameters do no good, then $\text{ERROR}(A)$ will equal $\text{ERROR}(C)$, so $\text{PRE} = 0$. If Model A provides a perfect fit, then $\text{ERROR}(A) = 0$ and (assuming Model C does not also provide a perfect fit) $\text{PRE} = 1$. Clearly, values of PRE will be between 0 and 1. The larger the value of PRE, the more it will be worth the cost of increased complexity to add the extra parameters to the model. The smaller the value of PRE, the more we will want to stick with the simpler, more parsimonious compact model.

For example, ignoring for the moment how we calculate the error, assume that total $\text{ERROR} = 50$ for the simple model that says that internet access is the same in all states and that $\text{ERROR} = 30$ for the model with the additional parameter for the percentage of college graduates. Then, $\text{ERROR}(C) = 50$, $\text{ERROR}(A) = 30$, and:

$$\text{PRE} = 1 - \frac{30}{50} = .40$$

That is, increasing the complexity of the model by considering educational level would reduce the error by 40%.

Let us review where we are. We have transformed the original problem of the conflicting goals for the model (parsimony and accurate representation of data) into a

consideration of the size of PRE for comparing Model C and Model A. Unfortunately, we have still not solved the problem of the conflicting goals for the model, because now we must decide whether a given PRE (e.g., the 40% reduction in the previous example) is big enough to warrant the additional parameter(s). The transformation of the original problem has moved us closer to a solution, however, for now we have a PRE index that will be used no matter how we finally decide to measure the error. More importantly, PRE has a simple, intuitive meaning that provides a useful description of the amount of improvement provided by Model A over Model C.

Deciding whether a PRE of, say, 40% is really worthwhile involves inferential statistics. An understanding of inferential statistics must await the details of measuring the error, sampling distributions, and other topics that are developed in Chapters 2, 3, and 4. We can, however, now specify two considerations that will be important in inferential statistics. First, we would be much more impressed with a PRE of, say, 40% if it were obtained with the addition of only one parameter instead of with four or five parameters. Hence, our inferential statistics will need to consider the number of extra parameters added to Model C to create Model A. *PRE per parameter added* will be a useful index. Second, we noted that n , the number of observations in DATA, serves as an upper limit to the number of parameters that could be added to the model. We will be more impressed with a given PRE as the difference between the number of parameters that were added and the number of parameters that could have been added becomes greater. Hence, our inferential statistics will consider how many parameters *could have been added* to Model C to create Model A but were not. In other words, we will be more impressed with a PRE of 40% if the number of observations greatly exceeds the number of parameters used in Model A than if the number of observations is only slightly larger than the number of parameters.

The use of PRE to compare compact and augmented models is the key to asking questions of our data. For each question we want to ask of DATA, we will find appropriate Models C and A and compare them by using PRE. For example, if we want to know whether educational level is useful for predicting the percentage of households that have internet access, we would compare a Model C that does not include a parameter for educational level to a Model A that includes all the parameters of Model C plus an additional parameter for educational level. If Model C is a simple model (i.e., a single-parameter model that makes a constant prediction for all observations), then we are asking whether educational level *by itself* is a useful predictor of internet access. If there are other parameters in Model C, then we are asking whether educational level is a useful predictor of internet access *over and above* the other parameters. (We discuss this at length in Chapter 6.) As another example, if we want to ask whether several factors, such as time zone, educational level, urban versus rural population, and income, are simultaneously useful in predicting internet access, we would use PRE to compare a Model C that did not have parameters for any of those factors to a Model A that did have those parameters in addition to those in Model C.

In the usual language for statistical inference, Model C corresponds to the *null hypothesis* and Model A corresponds to the alternative hypothesis. More precisely, the null hypothesis is that all the parameters included in Model A but not in Model C are zero (hence, the name “null”) or equivalently that there is no difference in error between Models A and C. If we reject Model C in favor of Model A, then we reject the null hypothesis in favor of the alternative hypothesis that is implied by the difference between

Models C and A. That is, we conclude that it is unreasonable to presume that all the extra parameter values in Model A are zero. We discuss this fully in Chapter 4.

NOTATION

To facilitate seeing interrelationships between procedures, we use consistent notation throughout. Y_i represents the i th observation from DATA. The first observation is numbered 1 and the last is n , for a total of n observations in DATA. \hat{Y}_i represents the Model's prediction of the i th observation. Y will always represent the variable that we are trying to predict with our model. Other variables giving information about each observation on which we might base conditional predictions will be represented by X . In other words, X will always be used to represent the predictor(s) of Y . So, X_{ij} represents the value of the j th predictor variable for the i th observation. For example, in the internet access example the 50 percentages would be represented as, Y_1, Y_2, \dots, Y_{50} , with $n = 50$. X_{i1} could be 1 if the state is in the Eastern or Pacific time zones, and -1 if the state is in the Central or Mountain time zones. In this example, we could use X_{i2} to represent the proportion of college graduates, in which case the value of X_{i2} would be the proportion of college graduates for the i th state.

For model parameters we will use $\beta_0, \beta_1, \dots, \beta_j, \dots, \beta_{p-1}$, for a total of p parameters. Even if we were to know the values of these parameters exactly, we would not expect the model to predict the data exactly. Instead, we expect that some random error will cause the model to predict less than perfectly. We let ε_i represent the unknown amount by which we expect the model to mispredict Y_i . Thus, for the simple model, the basic equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

can be expressed in terms of the true parameter β_0 and the error ε_i as:

$$Y_i = \beta_0 + \varepsilon_i$$

We can never know the true β parameters (or ε) exactly. Instead, we will have estimates of β that we will calculate from the data. These estimates will be labeled, $b_0, b_1, \dots, b_j, \dots, b_{p-1}$, respectively. We use \hat{Y}_i to represent the prediction for the i th observation based on the calculated b values. We then let e_i represent the amount by which the predicted value or \hat{Y}_i mispredicts Y_i ; that is:

$$e_i = Y_i - \hat{Y}_i$$

The Greek letters β and ε represent the true but unknowable parameters and the Roman letters b and e represent estimates of those parameters calculated from DATA. For the simple model, the model part of the basic data analysis equation is:

$$\text{MODEL: } \hat{Y}_i = b_0$$

and we can express that basic equation in terms of our parameter estimates as either:

$$Y_i = \hat{Y}_i + e_i$$

or

$$Y_i = b_0 + e_i$$

Either of the two previous equations are estimates for the basic equation expressed in terms of the unknown parameters as:

$$Y_i = \beta_0 + \varepsilon_i$$

The quantity $\beta_j X_{ij}$ tells us how much we should adjust the basic prediction for the i th observation based on the j th predictor variable. For example, if X_{ij} equals the proportion of college graduates in a state, then β_j specifies how much to adjust, upward or downward, depending on the sign of β_j , the internet access prediction for particular states. A more complicated model involving more parameters can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_j X_{ij} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

\hat{Y}_i , the MODEL portion of the data analysis equation, is then represented in terms of the parameter estimates as:

$$\text{MODEL: } \hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_j X_{ij} + \dots + b_{p-1} X_{i,p-1}$$

The equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

can again be expressed in two ways:

$$Y_i = \hat{Y}_i + e_i$$

or

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_j X_{ij} + \dots + b_{p-1} X_{i,p-1} + e_i$$

Note that when β values are used on the right side of the equation the appropriate symbol for the error is always ε_i and when b values (i.e., estimates of β s) are used on the right side of the equation the appropriate symbol for the error is always e_i . The reason is that in the first instance the error ε_i is unknown, while in the second instance an estimated value e_i can actually be calculated once \hat{Y}_i is calculated.

We will have to develop a few special symbols here and there, but in general the above notation is all that is required for all the models we consider in this book.

SUMMARY

The basic equation for data analysis is:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

The data analyst using this equation must resolve two conflicting goals: (a) to add parameters to MODEL so that it is an increasingly better representation of DATA with correspondingly smaller ERROR, and (b) to remove parameters from MODEL so that it will be a simple, parsimonious representation of DATA. Resolving this conflict is equivalent to asking whether the additional parameters are worth it. We use PRE, the index of the proportional reduction in error, to answer this question by comparing appropriately chosen Models C and A. In the traditional language of statistical inference, this is equivalent to comparing a null hypothesis and an alternative hypothesis. The next

several chapters provide the information necessary to judge when PRE is large enough to warrant rejecting Model C in favor of Model A.

We use consistent notation throughout the book to specify our statistical models. Y represents the variable that we are trying to predict (“y-hat,” that is, \hat{Y} , represents a predicted value of Y) and X represents a predictor variable. Lower-case Greek letters represent true, but unknown, characteristics of the population and Roman letters represent the estimates of those characteristics. Thus, β represents a true, but unknown, population parameter and b represents an estimate of that parameter, which is calculated from DATA. Similarly, ε_i represents the true, but unknown, error of prediction and e_i represents an estimate of that error, which is calculated from DATA. The same statistical model might therefore be expressed using unknown population parameters, for example, $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$, or using symbols that represent estimates that are calculated from DATA, for example, $Y_i = b_0 + b_1 X_{i1} + e_i$. Finally, we may also express the model in terms of predicted values, for example, $\hat{Y}_i = b_0 + b_1 X_{i1}$.

2

Simple Models

Definitions of Error and Parameter Estimates

In this chapter we consider the very simplest models—models with one or even no parameters. These simple models make the same prediction for all data observations; there are no differential predictions conditioned on whatever else we might know about each observation. Such a simple model may not seem very realistic or useful. However, this simple model provides a useful baseline against which we can compare more complicated models, and it will turn out to be more useful than it might appear at first. For many of the questions we will want to ask about our data, the appropriate Model C will be the simple model with no parameter or one parameter; this will be compared to the more complex Model A. Also, the simple model provides a useful first-cut description of data.

OVERVIEW OF THE SIMPLE MODEL

Formally, the simplest model is:

$$Y_i = B_0 + \varepsilon_i$$

where B_0 is some specified value not based on this particular batch of data (i.e., it is a specific a priori numeric value), and ε_i is the true error or the amount by which Y_i differs from B_0 . This simple model in which no parameters are estimated from the data is not frequently used in the social sciences because we seldom have theories sufficiently powerful to make an explicit prediction for a parameter. Such models are much more common in fields such as biology. For example, a medical study measuring human body temperature might reasonably consider the model: except for error, all temperatures are 37°C. In this case, $B_0 = 37$ so the formal model would be:

$$Y_i = 37 + \varepsilon_i$$

Although we will sometimes want to consider a specified or hypothesized value of B_0 in order to ask an explicit question about data, it is much more common to consider the equation:

$$Y_i = \beta_0 + \varepsilon_i$$

where β_0 is a true parameter that is estimated from the data. Continuing with the medical example, suppose that the body temperatures were all from people who had taken a certain drug. We might suspect that, except for error, they all have the same body temperature, but it is not the usual body temperature of 37°C. We use β_0 to represent whatever the

body temperature might be for those who have taken the drug. It is important to realize that β_0 is unknowable; we can only estimate it from the data. In terms of these true values, ε_i is the amount by which Y_i differs from β_0 if we were ever to know β_0 exactly. We use b_0 to indicate the estimate of β_0 that we derive from the data. Then the predicted value for the i th observation is:

$$\hat{Y}_i = b_0$$

and

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

becomes

$$Y_i = b_0 + e_i$$

where e_i is the amount by which our prediction misses the actual observation. Thus, e_i is the estimate of ε_i . The goal of tailoring the model to provide the best fit to the data is equivalent to making the errors:

$$e_i = Y_i - b_0$$

as small as possible. We have only one parameter, so this means that we want to find the estimate b_0 for that one parameter β_0 that will minimize the errors. However, we are really interested not in each e_i but in some aggregation of all the individual e_i values. There are many different ways to perform this aggregation. In this chapter, we consider some of the different ways of aggregating the separate e_i into a summary measure of the error. Then we show how each choice of a summary measure of the error leads to a different method of calculating b_0 to estimate β_0 so as to provide the best fit of the data to the model. Finally, we consider expressions that describe the “typical” error.

Measures of location or *measures of central tendency* are the traditional names for the parameter estimates of β_0 developed from the different definitions of error. These names are appropriate because the parameter estimate in the simple model tells us about the location of a typical observation or about the center of a batch of data. Specific instances include the mode, median, and mean. *Measures of variability* or *measures of spread* are the traditional names for the expressions for typical errors. These names are appropriate because expressions for typical errors tell us how variable the observations are in a batch of data or, equivalently, how far the data spread out from the center. Specific instances include the median absolute deviation and standard deviation. Together, measures of central tendency and spread are known as *descriptive statistics*. However, this suggests a false distinction between these statistics and those to come later. We want to emphasize that the parameter estimates for β_0 in the simple model are no more nor less descriptive than the parameter estimates we will develop for more complicated models. Models and their parameter estimates always provide descriptions of data. Hence, we will generally avoid the phrase “descriptive statistics” and just refer to parameter estimates. The reader should be aware, however, that when other textbooks refer to descriptive statistics they are generally referring to the material in this chapter.

CONCEPTUAL EXAMPLE

Before considering simple models and measures of error more formally, we consider some conceptual examples that will help to build useful intuitions so that the subsequent mathematical formulas will be less abstract. Suppose that the data consisted of the five observations 1, 3, 5, 9, 14, representing the number of books read over the summer by each of five elementary school students. These observations are plotted in Figure 2.1.

The simple model makes the same prediction for all five observations. The horizontal line represents the value of that constant prediction. The vertical lines drawn from each observation to the prediction line represent the amount by which the prediction misses the actual data value. In other words, the length of the line is e_i . One way to find the best value for \hat{Y} and, equivalently, the best estimate b_0 for β_0 is to adjust the \hat{Y} line up or down so that the length of the lines is a minimum. (Note that we have dropped the subscript i from \hat{Y}_i because all the predictions are the same for the simple model.) In other words, we can use trial and error to find the best estimate. For example, we might want to try 7 as our estimate. The data, the prediction line for $\hat{Y} = b_0 = 7$, and the errors are graphed in Figure 2.2. Note that the five line lengths are now 6, 4, 2, 2, 7, with a sum of 21.

For an estimate of 5, the line lengths were 4, 2, 0, 4, 9, with a sum of 19. The estimate $b_0 = 5$ thus produces less total error than $b_0 = 7$, so we can eliminate 7, in favor of 5, as an estimate if our goal is to minimize the total error. We can continue to try other estimates until we find the best one. Figure 2.3 shows the sum of the line lengths for different choices of b_0 between 0 and 10. The sum of the line lengths reaches a minimum of 19 when $b_0 = 5$, so that is our best estimate of β_0 . We would get more total error, a larger sum of line lengths, if we used a value of b_0 that was either lower or higher than 5. Hence, $b_0 = 5$ is the optimum estimate. Note that 5 is the middle of our five observations; the middle observation in a batch of data that have been sorted from smallest to largest is often called the *median*.

It is interesting to ask how we would have to adjust the estimate b_0 if one of the observations were dramatically changed. For example, what if the 14 were replaced by 140 so that the five observations were 1, 3, 5, 9, and 140? Before reading on, test your intuitions by guessing what the new value for b_0 will be. Figure 2.4 shows the sum of the line lengths for different possible values of b_0 . Although all of the sums are much larger than before, the minimum of 145 still occurs when $b_0 = 5$! The middle or median

FIGURE 2.1 Error as the sum of line lengths (estimate is $\hat{Y} = b_0 = 5$)

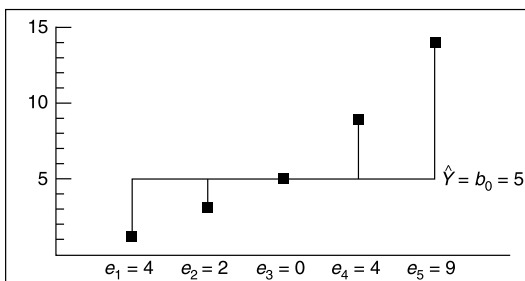


FIGURE 2.2 Error as the sum of line lengths (estimate is $\hat{Y} = b_0 = 7$)

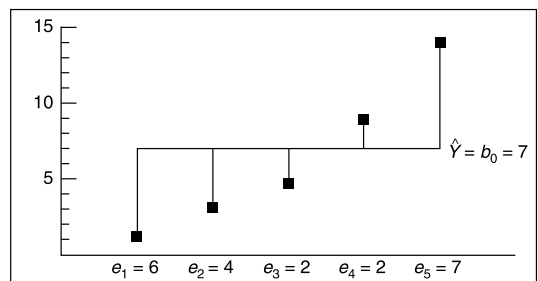
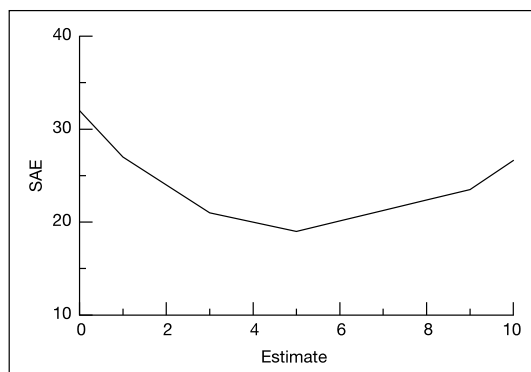
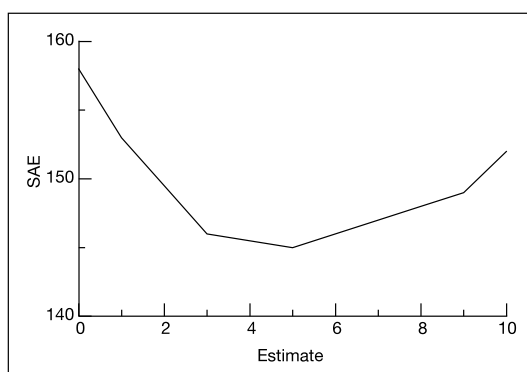


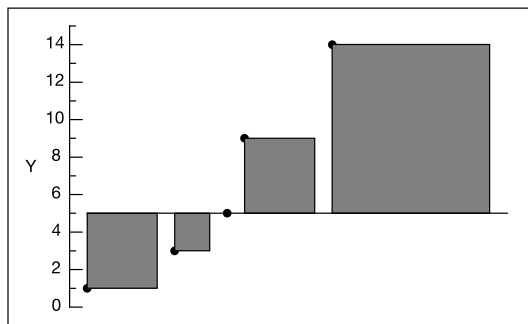
FIGURE 2.3 Sum of absolute error (SAE) as a function of \hat{Y}_i **FIGURE 2.4** Sum of absolute error (SAE) as a function of \hat{Y}_i with extreme observation (140)

observation is still the best estimate even though one of the observations has been increased by a factor of 10.

Above we used the sum of the error line lengths as an aggregate summary index of error. This simple sum may not always be reasonable. For example, the simple sum implies that several small errors (e.g., four errors of length 1) are equivalent to one large error (e.g., one error of length 4). Instead, we may want to charge a higher penalty in the error index for big errors so that an error of 4 counts more than four errors of 1. One way to accomplish this is to square the line lengths before summing them. For example, $4^2 = 16$ adds a lot more to the error sum than $1^2 + 1^2 + 1^2 + 1^2 + 4$. Figure 2.5 depicts the original set of five observations with this new definition of error; each e_i is now represented by a square, the length of whose side is determined by the distance between the observation and the horizontal line representing the constant prediction \hat{Y}_i of the simple model. The aggregate error is simply the sum of those squares.

Again, we can use brute force to find a value of \hat{Y}_i and b_0 that will make that sum of squares as small as possible. Let us consider the possible estimates of 5 and 7, which we evaluated when we were using the sum of the line lengths as the error measure. For $b_0 = 5$, the five areas of the squares are:

$$4^2 = 16, \quad 2^2 = 4, \quad 0^2 = 0, \quad 4^2 = 16, \quad 9^2 = 81$$

FIGURE 2.5 Error as the sum of squares (estimate is $\hat{Y}_i = b_0 = 5$)

and the sum of those squares is 117. For $b_0 = 7$, the five areas are 36, 16, 4, 4, 49, and the sum of squares is 109. So $b_0 = 5$, which was the best estimate when using line lengths, is no longer the best estimate when we use squares, because $b_0 = 7$ produces a smaller sum of squares or smaller error. Figure 2.6 shows the sum of squares for different possible values of b_0 between 0 and 10. The best value for b_0 is about 6.4 with a minimum sum of squares of about 107. The estimates of 5 and 7 are not bad—sums of squares of 117 and 109, respectively—but clearly inferior to the optimum estimate of 6.4. Although not obvious, we will prove later that the best estimate when using squared errors is simply the arithmetic average or *mean* of the observations. For the five observations:

$$\frac{1 + 3 + 5 + 9 + 14}{5} = \frac{32}{5} = 6.4$$

which produces b_0 , the best estimate of β_0 .

It is interesting to ask again what would happen to the estimate b_0 if one of the observations were dramatically changed: say, the 14 were replaced by 140. Before reading on, again check your intuition by guessing the new estimate b_0 . Figure 2.7 shows the sum of squares for the revised set of observations. The minimum sum of squares no longer occurs when $b_0 = 6.4$; instead, the minimum now occurs when b_0 is a whopping 31.6, which is again the average of the five observations. But note that although that estimate is the best, it is not very good, with a total sum of squares of about 14,723.

Before formalizing these examples in equations, it is useful to summarize the concepts introduced. First, the best estimate b_0 for a simple, one-parameter model is the constant prediction \hat{Y} , which minimizes the sum of the errors. Although we will generally have better ways to estimate the parameter than brute force, it is important to realize that the best estimate could be found by trial and error until we could no longer make the error any smaller. In fact, computer programs sometimes use precisely this strategy. Second, the choice of a method for summarizing or expressing the error—the lengths of the error lines or their squares in the above examples—affects the best estimate b_0 . We will soon see that there are many other plausible choices for error terms. Third, if total error is the sum of the line lengths, then the median of the observations provides the best estimate b_0 , where the median is simply the middle observation. Fourth, if total error is the sum of the squared line lengths, then the best estimate b_0 is the arithmetic

FIGURE 2.6 Sum of squared errors (SSE) as a function of \hat{Y}_i

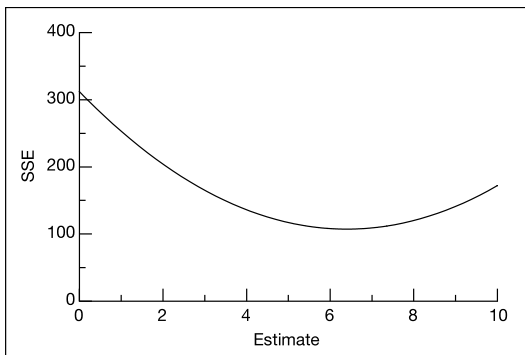
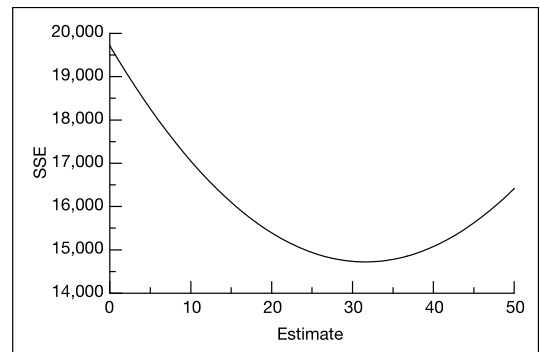


FIGURE 2.7 Sum of squared errors (SSE) as a function of \hat{Y}_i with extreme observation (140)



average or mean of the observations. Fifth, the median does not change when an extreme observation is made more extreme, but the mean can change dramatically. Sixth, many times in this book we will encounter the phrase “sum of squares”; it is useful to realize that it can indeed be represented geometrically as a literal summation of squares.

FORMALITIES FOR SIMPLE MODELS

As always, we begin with the basic equation for data analysis:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

For simple models, MODEL states that all observations in DATA are essentially the same, so:

$$Y_i = \beta_0 + \varepsilon_i$$

where β_0 represents the true, but unknown, parameter and ε_i represents the individual error disturbances around the unknown parameter. Then b_0 is the estimate of that parameter based on the data at hand. So the actual model used for predicting each Y_i becomes:

$$\text{MODEL: } \hat{Y}_i = b_0$$

The basic data analysis equation can then be written as:

$$Y_i = \hat{Y}_i + e_i \quad \text{or} \quad Y_i = b_0 + e_i$$

The error or residual associated with each observation is then simply the difference between the data and the model prediction, or:

$$e_i = Y_i - \hat{Y}_i$$

Our problem is how to select a single value for b_0 to represent all of the data. Clearly, we want b_0 to be in the “center” of the data so that it will be more or less close to all of the observations. Hence, estimates of b_0 are often called *measures of central tendency*. But we need to be much more precise about defining the center and what we mean by “close.” As always, the key is to make e_i as small as possible. Instead of looking at each individual e_i , we need to consider ways of aggregating the separate e_i values into a summary measure of the total error. Once we have done that, it should be a simple procedure to choose a b_0 that will make the summary measure of error as small as possible. We now turn to a consideration of possible summary measures.

Count of errors (CE)

One possibility is simply to count the number of times Y_i does not equal \hat{Y}_i . This ignores the size of the individual errors and only counts whether an error occurred. Formally:

$$\text{ERROR} = \sum_{i=1}^n I(e_i) = \sum_{i=1}^n I(Y_i - \hat{Y}_i) = \sum_{i=1}^n I(Y_i - b_0)$$

where $I(e_i) = 1$ if $e_i \neq 0$ and $I(e_i) = 0$ if $e_i = 0$. Functions such as $I()$ are often called *indicator* functions and are simply a fancy way of representing whether or not something is to be included in an equation.

Sum of errors (SE)

In order not to lose information about the size of the e_i , we might add all the e_i values so that:

$$\text{ERROR} = \sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n (Y_i - b_0)$$

But this is clearly unsatisfactory because it allows positive and negative errors to cancel one another. For example, suppose that b_0 underestimated one observation by 1000 and overestimated another observation by the same amount so that $e_1 = 1000$ and $e_2 = -1000$. Adding those two errors would produce zero, incorrectly implying that there was no error. We have no reason to be more interested in overestimates than underestimates, so one solution is to ignore the positive and negative signs of the e_i .

Sum of absolute errors (SAE)

One way to remove the signs is to sum the absolute values of the errors:

$$\text{ERROR} = \sum_{i=1}^n |e_i| = \sum_{i=1}^n |Y_i - \hat{Y}_i| = \sum_{i=1}^n |Y_i - b_0|$$

The sum of absolute errors is the formal equivalent of summing the line lengths in Figure 2.1. As noted in the conceptual example above, it may not always be desirable to count one big error (e.g., an error of 4) as being the same as the equivalent amount of small errors (e.g., four errors of 1). The conceptual example therefore suggests the next measure of error.

Sum of squared errors (SSE)

Another way to remove the signs from the e_i is to square each one before summing them:

$$\text{ERROR} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0)^2$$

The sum of squared errors is the formal equivalent of adding up the error squares in Figure 2.5. Besides removing the signs, the squaring has the additional effect of making large errors more important.

Weighted sum of squared errors (WSSE)

So far, all of the suggested error measures have given equal weight to each observation. For a variety of reasons we may want to give more weight to some observations and less weight to others when calculating the aggregate error. For example, we may have reason to believe that certain observations are questionable or suspect because they were collected with less precision or less reliability than the other observations. Or we may not want to count an error of 10 as being the same when $\hat{Y}_i = 1000$ as when $\hat{Y}_i = 5$. In the former instance, the error of 10 amounts to only a 1% error, while in the latter

instance it is an error of 200%. Or, finally, we might be suspicious of a couple of extreme observations just because they are “outliers” with respect to other observations. Whatever our reasons, it is easy to incorporate a weight w_i for each observation into the formal definition:

$$\text{ERROR} = \sum_{i=1}^n w_i e_i^2 = \sum_{i=1}^n w_i (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n w_i (Y_i - b_0)^2$$

The weights w_i might be assigned a priori based on judgments of data quality or some formal index of each observation’s reliability. One possibility is to weight all observations equally, in which case $w_i = 1$ for all i . This weighted sum of squared errors becomes simply the sum of squared errors above.

Statisticians have created some very clever ways of defining weights to solve a variety of complicated problems. We will encounter examples of those weights later in the context of specific problems. For now, just be aware that the use of weights gives us a great deal of flexibility in defining the aggregate measure of error.

ESTIMATORS OF β_0

As demonstrated by the conceptual examples presented earlier in this chapter, the choice of a method for aggregating the e_i influences the estimate b_0 . It should not be surprising, therefore, that for each definition of aggregate error presented above there is a different way of calculating b_0 from the data. We could use the brute-force method for each definition of error by trying different values of b_0 until we found the one that gave the minimum value for error. However, it turns out that for each of the above definitions of error we can define a way of calculating b_0 from the data so that b_0 is guaranteed to produce the minimum possible value for error. While we will almost always use the calculation method, it is important to remember that the definition of the best estimate of β_0 is the value of b_0 that produces the least error. We list in Figure 2.8 the definition of b_0 for each definition of error considered so far.

Proof that the Mean Minimizes SSE

In this section, we present a formal proof that the mean \bar{Y} does indeed produce the smallest possible value of SSE. Although we generally avoid proofs, we think it is important to understand that the choice of the mean as an estimator is not arbitrary; instead, the choice of SSE as an aggregate measure of ERROR also dictates the choice of the mean as the best estimator. Similar proofs can be given to show that the indicated estimator minimizes the corresponding definition of error in the list in Figure 2.8.

FIGURE 2.8 Estimators for each definition of error

<i>Error definition</i>	<i>b_0 estimator of β_0</i>
Count of errors	Mode = most frequent value of Y_i
Sum of absolute errors	Median = middle observation of all the Y_i
Sum of squared errors	Mean = average of all the Y_i
Weighted sum of squared errors	Weighted mean = weighted average of all the Y_i

Let us begin by assuming that \hat{Y} , the best estimator for reducing the sum of squared errors, is something other than

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}.$$

(Note that here \hat{Y} does not have an i subscript because we make the same prediction for all observations in the simple model.) Our strategy is to proceed with that assumption until we reach a point where we can see that it is a bad assumption. At that point we will know that the only reasonable assumption is $\hat{Y} = \bar{Y}$.

The sum of squared errors is:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y})^2$$

Obviously, $\bar{Y} - \bar{Y} = 0$; we can add zero within the parentheses without changing the sum of squared errors. That is:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y})^2$$

Rearranging the terms slightly, we get:

$$\text{SSE} = \sum_{i=1}^n [(Y_i - \bar{Y}) + (\bar{Y} - \hat{Y})]^2$$

Squaring gives:

$$\text{SSE} = \sum_{i=1}^n [(Y_i - \bar{Y})^2 + 2(Y_i - \bar{Y})(\bar{Y} - \hat{Y}) + (\bar{Y} - \hat{Y})^2]$$

Breaking the sums apart yields:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n [2(Y_i - \bar{Y})(\bar{Y} - \hat{Y})] + \sum_{i=1}^n (\bar{Y} - \hat{Y})^2$$

The last term contains no subscripts so the summation is equivalent to adding up the same quantity n times; hence, the summation sign can be replaced by multiplication by n . Similarly, the quantities without subscripts in the middle term can be taken outside the summation sign to give:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \hat{Y}) \sum_{i=1}^n (Y_i - \bar{Y}) + n(\bar{Y} - \hat{Y})^2$$

Now let us concentrate on the middle term. Note that:

$$\begin{aligned}
 \sum_{i=1}^n (Y_i - \bar{Y}) &= \sum_{i=1}^n Y_i - n\bar{Y} \\
 &= \sum_{i=1}^n Y_i - n \left(\frac{\sum_{i=1}^n Y_i}{n} \right) \\
 &= \sum_{i=1}^n Y_i - \sum_{i=1}^n Y_i \\
 &= 0
 \end{aligned}$$

Hence, the middle term of SSE includes a multiplication by zero, which eliminates that term. We are left with:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\bar{Y} - \hat{Y})^2$$

We want SSE to be as small as possible. We have no freedom in the first term: Y_i and \bar{Y} are whatever they happen to be. We do, however, have freedom to choose \hat{Y} in the second term to make SSE as small as possible. Clearly, $n(\bar{Y} - \hat{Y})^2$ is positive, so it is making SSE larger. But if we let $\hat{Y} = \bar{Y}$, then $n(\bar{Y} - \hat{Y})^2 = 0$ and we are no longer adding anything extra to SSE. For any estimate of \hat{Y} other than \bar{Y} , we will be making SSE larger. Hence, SSE is as small as possible when $\hat{Y} = \bar{Y}$, and the minimum is:

$$\text{SSE} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Any other choice for \hat{Y} would produce a larger SSE; thus, the mean is the best estimator for reducing SSE in the simple model.

Describing Error

If the goal is simply to describe a batch of data, then we can apply the simple model and use one or all of the measures of central tendency—mode, median, and mean—as descriptive statistics. When doing so, it is also useful to present a description of the typical error. Reporting the total error (e.g., SAE or SSE) is not desirable because the total depends so heavily on the number of observations. For example, aggregate error based on 50 observations is likely to be larger than aggregate error based on 15 observations, even if the typical errors in the former case are smaller. For each measure of central tendency there is a corresponding customary index of the typical error. We consider each below.

Modal error

When we use the count of the errors (CE) as the aggregate index of error, there can only be two values for e_i : either $e_i = 0$ when $Y_i = \hat{Y}$ or $e_i = 1$ when $Y_i \neq \hat{Y}$. The typical error is simply the more frequent or modal error. Modal error is seldom used, but we have presented it for completeness.

Median absolute deviation

When we use the sum of absolute errors (SAE) as the aggregate index of error, it is customary to use the median absolute error or deviation from the prediction to represent the typical error. To find the median absolute deviation, simply sort the $|e_i|$ into ascending order and find the middle one.

Standard deviation

When we use the sum of squared errors (SSE) as the aggregate index of error, the index is somewhat more complex. We will be making extensive use of SSE throughout this book. To avoid having to introduce more general formulas later, we present the more general formula now and then show how it applies in the case of simple models. In a general model with p parameters, those p parameters have been used to reduce the error. In principle, the maximum number of parameters we could have is n —one parameter for each observation—in which case the error would equal zero. Thus, there are $n - p$ potential parameters remaining that could be used to reduce the remaining error. A useful index of error is then the remaining error per remaining potential parameter. This index has the name *mean squared error* (MSE) and is given by:

$$\text{MSE} = \frac{\text{SSE}}{n - p} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - p}$$

For the simple model considered in this chapter there is only the one parameter β_0 to be estimated, so $p = 1$ and the estimate of Y_i is $b_0 = \bar{Y}$, the mean value. For the simple model, MSE has the special name *variance* and is commonly represented by s^2 , that is:

$$\text{Variance} = s^2 = \text{MSE} = \frac{\text{SSE}}{n - 1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

MSE represents the typical *squared* error; to express the typical error in the units in which the original data were recorded, it is useful to take the square root of MSE, which is often referred to, especially on computer printouts, as *ROOT MSE*. For the simple model, the square root of the variance or the MSE has the special name *standard deviation* and is given by:

$$\text{Standard deviation} = s = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

Another index sometimes used when SSE is used as the aggregate index of error is the *coefficient of variation*. It is common for the size of the standard deviation to be proportional to the size of the mean. For example, if $\bar{Y} = 10,000$, we would expect the typical error or standard deviation to be much larger than when $\bar{Y} = 10$. Although this need not be the case, it usually is true. To remove the effect of the overall magnitude of the data from the description of the error, the coefficient of variation expresses the size of the standard deviation as a proportion of the mean, that is:

$$\text{Coefficient of variation} = \text{CV} = \frac{s}{\bar{Y}}$$

An example

We will use the percentage of households that had internet access in the year 2013 by US state, which are listed in Figure 1.1, as an example to illustrate the simple model and the descriptors of central tendency and error. To facilitate finding the mode and the median, we have rearranged the data of Figure 1.1 in Figure 2.9 in order of increasing percentages. Our goal is to fit to these data the simple model that has just one parameter. Thus, the basic data analysis equation is:

$$Y_i = \beta_0 + \varepsilon_i$$

and we want to fit the model to the data by finding the estimate b_0 for β_0 that minimizes error—the e_i in the equation:

$$Y_i = b_0 + e_i$$

How we find the estimate b_0 depends on which definition of aggregate error we adopt.

If CE is adopted as the criterion, then the best estimate for β_0 is the mode. To find the mode, we simply observe which percentage is the most frequent. For these data, there are eight values that occur twice (i.e., 72.2, 72.9, 73.0, 75.3, 76.5, 77.5, 78.9, and 79.6) and none that occurs more than twice. Thus, there are really eight modes. It is frequently the case with continuous variables that there is no mode, or at least not a single mode, so the mode is usually not as useful as either the median or the mean for such data. If we were to round the data to the nearest whole number, there would be only two modes of 73 and 79. Using a mode of 73 (or 79) to predict the rounded data, the prediction would be accurate 6 times and incorrect 44 times.

If SAE is adopted as the criterion, then the best estimate for β_0 is the median. There are 50 observations, so there are two middle values—the 25th and 26th. (If there is an odd number of observations then there will be only one middle observation.) The 25th and 26th values are both 73.0; so, in this case, the best estimate is 73.0. (If the two middle observations were different, they could be averaged to produce the single estimate of the median.) The middle columns of Figure 2.10 present the prediction \hat{Y}_i , the error $e_i = Y_i - \hat{Y}_i$, and the absolute error based on the median as the estimate b_0 . In this case, total error = 200.3. Any other estimate for β_0 would produce a larger value for SAE. Although it is not obvious from Figure 2.10, the 25th and 26th largest absolute errors are 3.20 and 3.30, so the median absolute deviation or MAD = 3.25.

If SSE is adopted as the criterion, then the best estimate for β_0 is the mean. The average of the 50 observations gives 72.806 as the estimate b_0 . The last set of columns

FIGURE 2.9 Percentage of households that had internet access in 2013 by US state (sorted by percentage)

<i>i</i>	State	Percentage	Rank	<i>i</i>	State	Percentage	Rank
24	MS	57.4	1	49	WI	73.0	26
4	AR	60.9	2	12	ID	73.2	27
1	AL	63.5	3	3	AZ	73.9	28
31	NM	64.4	4	13	IL	74.0	29
18	LA	64.8	5	9	FL	74.3	30
48	WV	64.9	6	8	DE	74.5	31
40	SC	66.6	7	32	NY	75.3	32
36	OK	66.7	8	45	VT	75.3	33
42	TN	67.0	9	50	WY	75.5	34
17	KY	68.5	10	28	NV	75.6	35
14	IN	69.7	11	46	VA	75.8	36
25	MO	69.8	12	23	MN	76.5	37
22	MI	70.7	13	39	RI	76.5	38
33	NC	70.8	14	7	CT	77.5	39
41	SD	71.1	15	37	OR	77.5	40
35	OH	71.2	16	5	CA	77.9	41
43	TX	71.8	17	11	HI	78.6	42
26	MT	72.1	18	20	MD	78.9	43
10	GA	72.2	19	47	WA	78.9	44
15	IA	72.2	20	2	AK	79.0	45
38	PA	72.4	21	30	NJ	79.1	46
34	ND	72.5	22	6	CO	79.4	47
19	ME	72.9	23	21	MA	79.6	48
27	NE	72.9	24	44	UT	79.6	49
16	KS	73.0	25	29	NH	80.9	50

in Figure 2.10 gives the values of \hat{Y} (or b in the case of the simple model), $e = Y - \hat{Y}$, and e_i^2 . Note that the sum of the errors equals zero exactly and, necessarily, that the sum of the data observations equals the sum of the predictions. This is characteristic of predictions based on minimizing the SSE. The actual SSE equals 1355.028. Again, any other estimate for β_0 would produce a larger SSE. The variance or, more generally, the MSE equals:

$$s^2 = \frac{\text{SSE}}{n - 1} = \frac{1355.028}{49} = 27.654$$

and the standard deviation or root-mean-squared error equals:

$$s = \sqrt{\text{MSE}} = \sqrt{27.654} = 5.259$$

Finally, CV is given by:

$$\frac{s}{\bar{Y}} = \frac{5.259}{72.806} = .072$$

Note that in this particular example the three estimates of β_0 (using the three definitions of error) were very similar: 73 (the first mode), 73.0, and 72.806. This is often the case for “well-behaved” data, but there is no guarantee that data will be well

FIGURE 2.10 Predictions and errors using the median and mean to estimate b_0 in the simple model

i	US state	Percentage	Median			Mean		
			\hat{Y}_i	e_i	$ e_i $	\hat{Y}_i	e_i	e_i^2
1	AK	79.0	73	6.0	6.0	72.806	6.194	38.366
2	AL	63.5	73	-9.5	9.5	72.806	-9.306	86.602
3	AR	60.9	73	-12.1	12.1	72.806	-11.906	141.753
4	AZ	73.9	73	0.9	0.9	72.806	1.094	1.197
5	CA	77.9	73	4.9	4.9	72.806	5.094	25.949
6	CO	79.4	73	6.4	6.4	72.806	6.594	43.481
7	CT	77.5	73	4.5	4.5	72.806	4.694	22.034
8	DE	74.5	73	1.5	1.5	72.806	1.694	2.870
9	FL	74.3	73	1.3	1.3	72.806	1.494	2.232
10	GA	72.2	73	-0.8	0.8	72.806	-0.606	0.367
11	HI	78.6	73	5.6	5.6	72.806	5.794	33.570
12	IA	72.2	73	-0.8	0.8	72.806	-0.606	0.367
13	ID	73.2	73	0.2	0.2	72.806	0.394	0.155
14	IL	74.0	73	1.0	1.0	72.806	1.194	1.426
15	IN	69.7	73	-3.3	3.3	72.806	-3.106	9.647
16	KS	73.0	73	0.0	0.0	72.806	0.194	0.038
17	KY	68.5	73	-4.5	4.5	72.806	-4.306	18.542
18	LA	64.8	73	-8.2	8.2	72.806	-8.006	64.096
19	MA	79.6	73	6.6	6.6	72.806	6.794	46.158
20	MD	78.9	73	5.9	5.9	72.806	6.094	37.137
21	ME	72.9	73	-0.1	0.1	72.806	0.094	0.009
22	MI	70.7	73	-2.3	2.3	72.806	-2.106	4.435
23	MN	76.5	73	3.5	3.5	72.806	3.694	13.646
24	MO	69.8	73	-3.2	3.2	72.806	-3.006	9.036
25	MS	57.4	73	-15.6	15.6	72.806	-15.406	237.345
26	MT	72.1	73	-0.9	0.9	72.806	-0.706	0.498
27	NC	70.8	73	-2.2	2.2	72.806	-2.006	4.024
28	ND	72.5	73	-0.5	0.5	72.806	-0.306	0.094
29	NE	72.9	73	-0.1	0.1	72.806	0.094	0.009
30	NH	80.9	73	7.9	7.9	72.806	8.094	65.513
31	NJ	79.1	73	6.1	6.1	72.806	6.294	39.614
32	NM	64.4	73	-8.6	8.6	72.806	-8.406	70.661
33	NV	75.6	73	2.6	2.6	72.806	2.794	7.806
34	NY	75.3	73	2.3	2.3	72.806	2.494	6.220
35	OH	71.2	73	-1.8	1.8	72.806	-1.606	2.579
36	OK	66.7	73	-6.3	6.3	72.806	-6.106	37.283
37	OR	77.5	73	4.5	4.5	72.806	4.694	22.034
38	PA	72.4	73	-0.6	0.6	72.806	-0.406	0.165
39	RI	76.5	73	3.5	3.5	72.806	3.694	13.646
40	SC	66.6	73	-6.4	6.4	72.806	-6.206	38.514
41	SD	71.1	73	-1.9	1.9	72.806	-1.706	2.910
42	TN	67.0	73	-6.0	6.0	72.806	-5.806	33.710
43	TX	71.8	73	-1.2	1.2	72.806	-1.006	1.012
44	UT	79.6	73	6.6	6.6	72.806	6.794	46.158
45	VA	75.8	73	2.8	2.8	72.806	2.994	8.964
46	VT	75.3	73	2.3	2.3	72.806	2.494	6.220
47	WA	78.9	73	5.9	5.9	72.806	6.094	37.137
48	WI	73.0	73	0.0	0.0	72.806	0.194	0.038
49	WV	64.9	73	-8.1	8.1	72.806	-7.906	62.505
50	WY	75.5	73	2.5	2.5	72.806	2.694	7.258
Sum		3640.3	3650	-9.7	200.3	3640.3	.000	1355.03

behaved and that the three estimates will be similar. Later we will see that a major discrepancy between the three estimates, especially between the median and the mean, should alert us to special problems in the analysis of such data. Note also in this example that the median absolute deviation and the standard deviation produce different estimates for the typical error—3.25 and 5.26, respectively. This is not surprising given the different definitions of error used.

SUMMARY

In terms of the basic data analysis equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

the simple model with one parameter is expressed as:

$$Y_i = \beta_0 + \varepsilon_i$$

Fitting the model to the data consists of finding the estimator b_0 of β_0 that makes the errors e_i as small as possible in the equation:

$$Y_i = b_0 + e_i$$

To fit the simple model to data, we must first define how the individual error terms e_i are to be aggregated into a summary index of error. Once we have chosen an aggregate index of error, we can find, by trial and error if necessary, the best estimate b_0 of β_0 that minimizes error. Important definitions of aggregate error are: (a) the count of errors, (b) the sum of absolute errors, and (c) the sum of squared errors. For each of these definitions, there is a different, well-defined best estimator for β_0 that can be found by calculation rather than by trial and error. These estimators are, respectively: (a) the mode—the most frequent value of Y_i , (b) the median—the middle value of all the Y_i , and (c) the mean—the arithmetic average of all the Y_i . Also, for each of these three definitions of aggregate error there is an expression for representing the “typical” value for e_i . These expressions are, respectively: (a) the modal error, (b) the median absolute deviation, and (c) the standard deviation. Collectively, these best estimators and these expressions for the typical error are known as *descriptive statistics* because they provide a first-cut description of a batch of data. We prefer to view them simply as estimators for the simple model using different definitions of error.

In Chapter 3, we will choose one of the three definitions of error to be our standard definition. We will make this choice on the basis of reasoned principles. However, the estimates and aggregate indices for the sum of squared errors are more easily obtained from the standard computer statistical systems than are those for other definitions of aggregate error. The reader should therefore anticipate that we will choose the sum of squared errors to be our standard definition of aggregate error.

Simple Models

Models of Error and Sampling Distributions

In Chapter 2 we considered alternative aggregate summary measures for error and saw how different choices of an aggregate error index led us to different estimates for the single parameter of the simple model. Now we need to make a choice of which aggregate index of error we will generally use. To do so, we must state our assumptions about the behavior of the errors ε_i . We need to have a model for error. In this chapter, we consider a reasonable model of error that is often appropriate and show how that model of error directs us to a certain aggregate index. In later chapters, we show how to check the appropriateness of the chosen model for error and then consider what remedies to apply in our data analysis when the model for error is inappropriate.

In this chapter, our focus is on the error term ε_i , which should be included in the full statement of our models. For the simple model, the complete statement is:

$$\text{MODEL: } Y_i = \beta_0 + \varepsilon_i$$

This model says that were it not for random perturbations represented by the ε_i , all the Y_i would equal β_0 exactly. What are these random perturbations? They are anything that causes the observed datum Y_i to deviate from the true value β_0 . For the internet access example, the delivery or installation of equipment may have been unusually slow in some states; other states may have had either unusually good or unusually poor record-keeping procedures; still others may have had unusually high or low numbers of individuals whose work involved internet use. Any of these or countless other factors may have caused the observed percentages of households with internet access to deviate from the single value β_0 predicted by the simple model.

You might object that a factor such as internet use at work ought to be included in the model of household internet access rather than in the error. Indeed, internet use at work and the other factors listed above as causes of error may be important predictors that ought to be included in the model. However, if we are using a simple model with only one parameter, then we must include all those factors in the error because those factors cause the observed Y_i to deviate from the single value β_0 predicted by the model. Later we will consider more complex models that will allow us to move such factors from the error to the model. To the extent that moving a factor from the error to the model reduces the error, we will have improved our model and its predictions. How to do that must await subsequent chapters; for now our problem is what assumptions we can reasonably make about the nature of the error.

SAMPLING DISTRIBUTIONS

As was shown in Chapter 2, for each aggregate index of error there is an associated estimator of β_0 in the simple model. For example, for the sum of absolute errors the median is the best estimator, and for the sum of squared errors the mean is the best estimator. Part of our strategy for selecting an aggregate index of error will be to examine the performance of various estimators under some reasonable assumptions about the behavior of the errors ε_i . Below we develop a metaphorical representation of the simple model, which will help our intuitions about error and provide a context for examining the relative performance of the median and the mean.

Metaphorical Representation of Error

Suppose that Nature has a bag containing a large number of tickets and on each ticket is written a particular value for ε_i . Figure 3.1 is a list of 100 values for ε_i that might be on the tickets.

To determine the value for a particular observation Y_i , Nature reaches into her bag of tickets and *randomly* draws one ticket; she then adds the number on the ticket, the value of the error ε_i , to the true value β_0 . Let us assume, for example, that $\beta_0 = 50$ so that the simple model becomes:

$$\text{MODEL: } Y_i = 50 + \varepsilon_i$$

Suppose the following 20 values were on tickets randomly drawn from a bag containing the tickets of Figure 3.1:

-9	28	-17	22	35
-12	-23	1	10	11
6	-24	-19	-26	-31
-4	8	20	-14	-2

Then the first datum would equal $50 + (-9) = 41$ because -9 is on the first error ticket drawn. The second datum would equal $50 + 28 = 78$ and so on. The 20 observed values would be:

41	78	33	72	85
38	27	51	60	61
56	26	31	24	19
46	58	70	36	48

Note that for the moment we are ignoring the issue of how it is determined which numbers are to be written on the tickets representing the errors. This is a very important issue to which we return later. Had a different set of error tickets been drawn we would have obtained a different set of data.

Statisticians often use the word *sample* to refer to a batch of data that could have been different if, in terms of our metaphor, a different set of error tickets had been drawn. Thus, the 20 data observations above represent a sample of the possible data values that could have been obtained. The mean and median for this particular sample

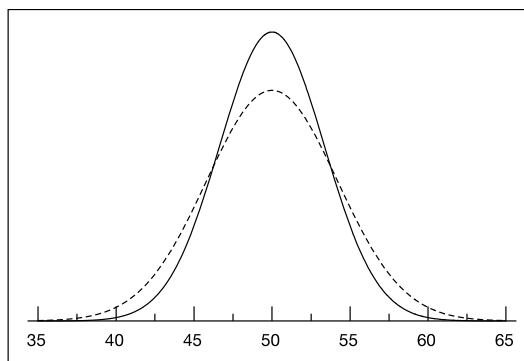
FIGURE 3.1 100 hypothetical values of ε_i in Nature's bag of tickets

-11	6	1	7	14	0	-1	9	1	2
32	15	12	-5	4	7	-6	-22	1	-2
20	11	29	-14	-8	-13	-7	-25	-3	-4
-19	8	-4	0	-15	17	-23	-8	6	17
3	-5	-15	11	2	-14	-20	4	31	24
2	-28	-31	12	-5	10	29	-10	19	-5
-6	5	0	-4	-5	-2	-12	13	4	-23
25	-19	-11	-22	3	4	3	28	10	-26
-9	35	7	-10	22	-17	10	5	-15	20
-5	-24	-4	-2	12	-26	10	-7	-11	-24

of 20 observations are 48 and 47, respectively. Note that neither the mean nor the median exactly equals the actual value of β_0 specified in the model as 50 for this example. Had a different sample been drawn, the mean and median would likely have been different. For some samples the mean and median would be closer to the true value, and for others they would be farther away. If the mean and median are to be good estimates of β_0 , we would hope that on average they would be close to β_0 . For this example we can perform a simulation experiment to determine likely values for the mean and median for samples of size 20. On each simulation round we randomly select another sample of 20 error tickets, add 50 to each one, and then calculate the mean and median of the resulting sample. If we perform many, many simulation rounds, generating many, many different samples and computing the mean and median in each, then we can treat each mean or median as a datum itself. We can then construct a frequency polygon for these data. Such frequency distributions are known as *sampling distributions* because they show the distribution of a particular statistic based on many, many samples.

Properties of Estimators

Sampling distributions provide all the information necessary for us to evaluate the relative performance of various estimators. Below we examine the performance of the median and mean with respect to three desirable properties of an estimator: unbiased, efficient, and consistent. Each concept is illustrated and defined in terms of Figure 3.2,

FIGURE 3.2 Sampling distributions for the mean (solid line) and median (dashed line) for $n = 20$ 

which depicts the sampling distributions for both the mean and the median based on an infinite number of samples of size 20 from the bag of error tickets in Figure 3.1.

Unbiasedness

Reassuringly, both distributions have peaks at 50, indicating that 50 is the most likely value for both the mean and the median. If the peak and the average value from the sampling distribution for an estimator or statistic equal the specified value for the parameter (50 in this case), then the estimator is said to be *unbiased*.

Efficiency

For these sampling distributions, there are very few instances for which the mean or the median is very far away from 50; the bulk of the observed values are between about 45 and 55. Note, however, that the mean appears to be a slightly better estimator than the median because there are somewhat fewer values of the median near 50 and there are somewhat more values of the median in the “tails” of the distribution far away from 50. For this example the mean is said to be more efficient than the median because with the same number of observations the mean provides more precision.

Consistency

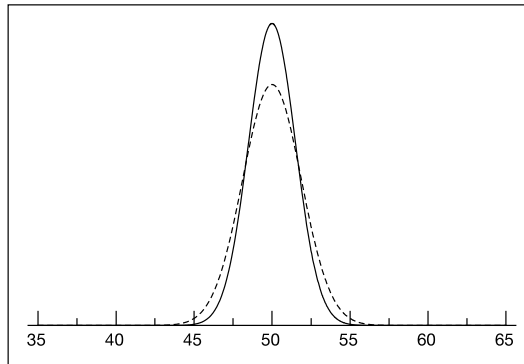
It is interesting to ask what the sampling distributions would be like if the sample sizes were larger. For example, what if the mean and median were based on samples of size 100 instead of samples of size 20? Figure 3.3 shows the sampling distributions for the mean and the median based on an infinite number of samples of size 100.

Compared with Figure 3.2 it is clear that the sampling distributions based on samples of size 100 are much narrower than those based on samples of size 20. The narrowness indicates that a greater proportion of the observed values of the mean and median were near 50 and that even fewer observed values were far away from 50. If the sampling distribution of an estimator or statistic becomes narrower as the sample size increases, then it is said to be a *consistent* estimator. Note also that again the mean is slightly more efficient than the median because its sampling distribution is a little narrower.

NORMAL DISTRIBUTION OF ERRORS

Unbiasedness, consistency, and efficiency are obviously desirable attributes for an estimator of β_0 in the simple model. Both the mean and the median and many other estimators are unbiased and consistent, so those properties offer no basis for choosing an estimator and its corresponding index of aggregate error. The efficiency of estimators does differ, and in the above example the mean was slightly more efficient than the median. Hence, our example suggests that choosing the sum of squared errors as our aggregate index of error and using its associated estimator (i.e., the mean) would serve us well. But that conclusion depends on the particular distribution of error tickets in Figure 3.1. Had there been some other distribution of error tickets in Nature’s bag, we might have found the median to be more efficient. Thus, we must make an assumption

FIGURE 3.3 Sampling distributions for the mean (solid line) and median (dashed line) for $n = 100$

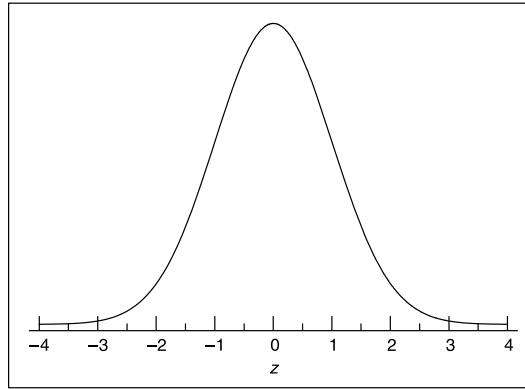
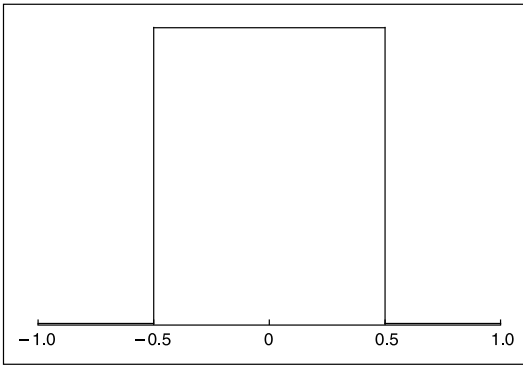
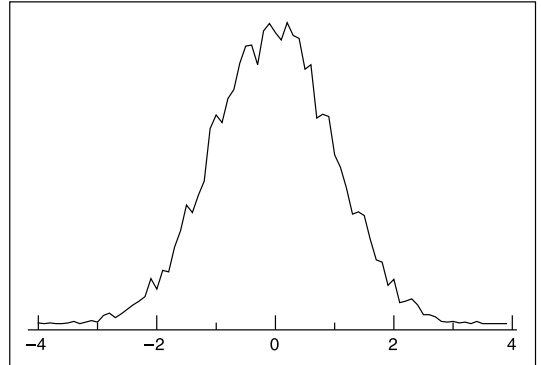


or state our model about the contents of Nature's bag of error tickets. Equivalently, we will be stating our assumption about the distribution of the error tickets.

We will assume that the errors have a *normal distribution*. That is, we assume that a frequency distribution for the error tickets would have the general shape of the normal curve depicted in Figure 3.4. This is an idealized representation that would only be obtained from a bag containing an infinite number of tickets. The 100 error tickets of Figure 3.1 were, in effect, randomly selected from a bag containing an infinite number of normally distributed errors.

There are many other possible distributions of errors that we might have assumed, but there are good reasons for assuming the normal distribution. First, there is an abundance of empirical data indicating that the distribution of errors is often approximately normal. That is, in fact, how it acquired the name "normal" distribution. Second, as we shall see in Chapter 13, it is often possible to transform a non-normal distribution of errors into a nearly normal distribution. Third, and most importantly, there is a reasonable model of the errors that causes us to expect them to be distributed normally. We turn now to a description of that model.

Above, we mentioned a number of possible components of the errors for the internet access data. It is unlikely that in any particular case only a single error component is perturbing the observed value away from the parameter value specified in the simple model. For example, it might be that internet access was altered because of an unusual installation process *and* unusual record keeping *and* publicity about internet security problems. Some of the factors will push the observed value above the model value, and others will push it below the model value. The observed difference between the observed and model values, $Y_i - \hat{Y}_i$, will actually be determined by the *sum* of the individual error components. The central limit theorem, a well-known theorem in mathematical statistics, shows that the distribution of the sum or average of a number of components will approximate to a normal distribution no matter what the original distribution of the individual components. Thus, if we assume that the actual error is the sum of a number of individual errors, then it is reasonable to expect that the distribution of the errors will approximate to a normal distribution. The central limit theorem also shows that the greater the number of components in the sum or average, the better the approximation to a normal distribution.

FIGURE 3.4 Normal probability curve**FIGURE 3.5** Uniform distribution**FIGURE 3.6** Frequency distribution for the sum of 12 uniform errors

The central limit theorem is of such importance that a demonstration of the concept is worthwhile. Suppose that the distribution of the individual error components is such that any value between -0.5 and $+0.5$ is just as likely as any other. This distribution is known as the *uniform distribution* because each possible value has the same probability. As Figure 3.5 illustrates, the uniform distribution is very different from a normal distribution.

Suppose that each value on an error ticket is the sum of 12 small individual error components randomly selected from the uniform distribution. Figure 3.6 shows the frequency distribution for 10,000 such error tickets. Clearly, even though the distribution for the individual components was not normal, the distribution for the sum closely approximates to a normal distribution. Thus, if it is reasonable to assume that the observed error is actually the sum of a lot of small random error components, then it is reasonable to assume that the distribution of the errors is approximately normal.

Formal Specification of the Normal Distribution

Fortunately, there are precise mathematical statements of the normal distribution, so we do not have to depend on simulations of drawing tickets from a bag, as we used

in our metaphor above. The curve for the relative frequency of the normal distribution represented in Figure 3.4 is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

This equation has two parameters: μ , the expected value or mean of the distribution, and σ^2 , the variance of the distribution. Different values of μ simply shift the curve of Figure 3.4 back and forth along the horizontal axis. Figure 3.7 shows two normal curves that have the same variance ($\sigma^2 = 1$) but different means ($\mu_1 = 0$, $\mu_2 = 3$). The variance parameter σ^2 determines the spread of the normal curve. The larger σ^2 is, the wider the distribution. Figure 3.8 shows three normal curves that each have a mean of $\mu = 0$ but increasing variances of 0.25, 1, and 4.

All normal distributions have the same shape, so it is common to use the distribution with $\mu = 0$ and $\sigma^2 = 1$ as a standard reference distribution. Figure 3.9 gives the probability that a value x randomly sampled from the normal distribution will be less than the tabled values of z ranging between -3 and $+3$. For example, the probability for $z = 1.0$ equals .841, which indicates that 84.1% of the time a randomly sampled value from this distribution will be less than 1.0. Note that the probability for $z = 0$, the mean for this particular distribution, equals .5, indicating that we would expect 50% of the observations to be below zero and 50% of the observations to be above zero. Tables of the normal distribution are available in many statistical books, and most computer statistical packages provide functions for easily calculating the required probabilities.

To find the expected proportion of observations between any two particular values, we simply subtract their associated probabilities. For example, the proportion of observations between $z = 0$ and $z = 1$ is $.841 - .500 = .341$. Figure 3.10 shows the expected proportion of observations between various landmarks.

We will not be using the z table of Figure 3.9 very much; instead we will be using similar tables for some special distributions. For most purposes, it is sufficient to remember that approximately 68% of the observations from a normal distribution fall between -1 and $+1$, approximately 95% fall between -2 and $+2$, and almost all (approximately 99.7%) fall between -3 and $+3$. The tails of the normal distribution

FIGURE 3.7 Normal probability curves with same variance and different means

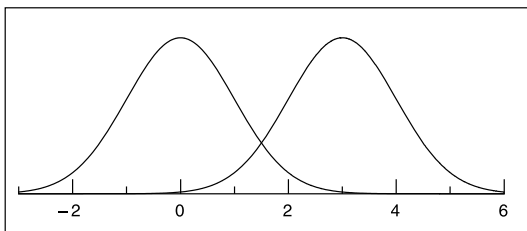


FIGURE 3.8 Normal probability curves with same mean but different variances

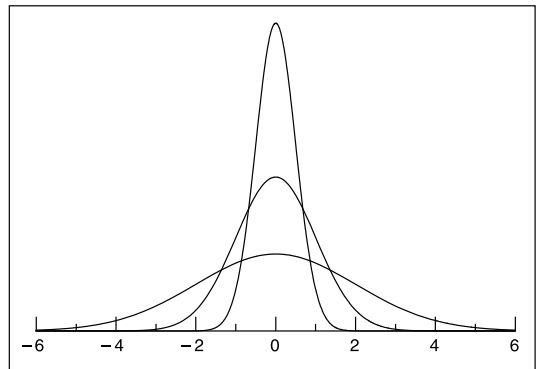
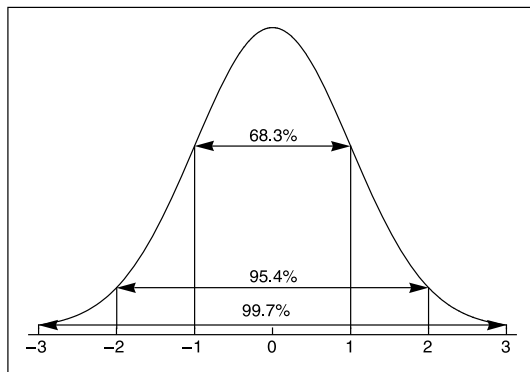


FIGURE 3.9 Cumulative probabilities for the normal distribution for $\mu = 0$ and $\sigma^2 = 1$

z	$p(x < z)$	z	$p(x < z)$
-3.0	.001	0.0	.500
-2.9	.002	0.1	.540
-2.8	.003	0.2	.579
-2.7	.003	0.3	.618
-2.6	.005	0.4	.655
-2.5	.006	0.5	.691
-2.4	.008	0.6	.726
-2.3	.011	0.7	.758
-2.2	.014	0.8	.788
-2.1	.018	0.9	.816
-2.0	.023	1.0	.841
-1.9	.029	1.1	.864
-1.8	.036	1.2	.885
-1.7	.045	1.3	.903
-1.6	.055	1.4	.919
-1.5	.067	1.5	.933
-1.4	.081	1.6	.945
-1.3	.097	1.7	.955
-1.2	.115	1.8	.964
-1.1	.136	1.9	.971
-1.0	.159	2.0	.977
-0.9	.184	2.1	.982
-0.8	.212	2.2	.986
-0.7	.242	2.3	.989
-0.6	.274	2.4	.992
-0.5	.309	2.5	.994
-0.4	.345	2.6	.995
-0.3	.382	2.7	.997
-0.2	.421	2.8	.997
-0.1	.460	2.9	.998
0.0	.500	3.0	.999

FIGURE 3.10 Standard normal probability curve ($\mu = 0, \sigma^2 = 1$)



theoretically extend to $+\infty$ and $-\infty$; however, observations more extreme than 3 are infrequent, and observations more extreme than 4 are very rare.

Normal distributions with means other than 0 and/or variances other than 1 are easily compared with the reference distribution in Figure 3.9. Suppose that Y is an observation from a normal distribution with mean μ and variance σ^2 . Subtracting μ from Y will shift the distribution along the horizontal axis, so the mean for the constructed variable $Y - \mu$ equals 0. Although it is not obvious, it can be shown that dividing $Y - \mu$ by σ will change the spread of the distribution so that it coincides with our reference distribution. (Note that we divide by σ , not σ^2 .) The division by σ rescales the original variable so that it is measured in units of the standard deviation. Thus if Y is from a normal distribution with mean μ and variance σ^2 , then z , defined as:

$$z = \frac{Y - \mu}{\sigma}$$

has a normal distribution with mean 0 and variance 1. Values calculated according to this formula are known as *z scores* or, because of their use with the standard reference distribution, *standard normal deviates*. “Deviates” refers to the size of the deviation from the mean, measured in units of the standard deviation. For example, intelligence test scores (IQ scores) are scaled so that $\mu = 100$ and $\sigma = 15$. To find the proportion of individuals we would expect to have IQ scores that are less than or equal to 127, given that IQ scores are normally distributed, we would first compute the corresponding z score. In this case:

$$z = \frac{127 - 100}{15} = 1.8$$

That is, an IQ score of 127 is 1.8 units (i.e., 1.8 standard deviations) above the mean (of 100). The associated cumulative probability in Figure 3.9 for $z = 1.8$ is .964. So, we would expect about 96% of the population to have IQ scores lower than 127.

We can also convert z scores back to the original scale. The Y score that corresponds to a particular z is given by:

$$Y = \mu + z\sigma$$

In the same IQ example, if we wanted to know what IQ score corresponded to $z = 1.6$ (i.e., a score 1.6 standard deviations above the mean), we would calculate:

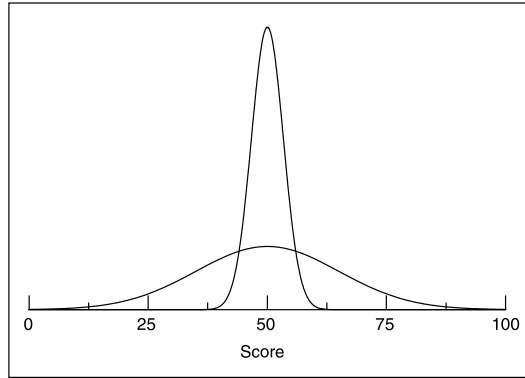
$$100 + 1.6(15) = 124$$

Sampling Distributions Revisited

Now that we have made an explicit assumption about the distribution of the errors, we can reconsider the sampling distributions for estimators such as the mean and median. Let us again return to our example of the simple model in which $\beta_0 = 50$. That is, our model is:

$$Y_i = 50 + \varepsilon_i$$

FIGURE 3.11 Normal distribution of observations and sampling distribution of the mean for $n = 20$



We now make the explicit assumption that the ε_i are randomly sampled from a normal distribution with $\mu = 0$ and some constant but unknown variance σ^2 . Adding 50 to each error value simply translates the normal distribution along the horizontal axis, so the Y_i data values will also have a normal distribution with mean $\mu = 50$ and variance σ^2 . In order to be able to draw a picture of this distribution, we have assumed that $\sigma^2 = 225$. Then the distribution of the Y_i values would have the theoretical curve depicted in Figure 3.11 as the broad curve.

If we were to calculate the mean from samples of size n , the theoretical sampling distribution of such means would also have a normal distribution. The mean of the sampling distribution of means equals the mean of the original theoretical distribution, which in this case is 50. Mathematical statisticians have shown that the variance of this sampling distribution would be σ^2/n if the error tickets are drawn from a normal distribution. In our example, if the sample size were 20, then the variance of the sampling distribution for the means would equal $225/20 = 11.25$. If we were to collect many, many samples of size 20 and calculate the mean for each sample, the sampling distribution of those means would approximate the theoretical sampling distribution also depicted in Figure 3.11 as the peaked curve. Note that, as was the case when we simulated the sampling distribution, the spread for the mean is very much smaller than the original spread. If we were to collect larger samples, the sampling distribution for the mean would be even narrower because we would be dividing σ^2 by a larger value in order to determine its variance.

Mathematical statisticians have also shown that the variance for the sampling distribution of the median based on samples of size n is approximately equal to:

$$\frac{\pi\sigma^2}{2n}$$

if the error tickets are drawn from a normal distribution. This is the same as the variance for the mean except that it is multiplied by $\pi/2$ (which equals approximately 1.57), so the variance for the median must be larger. That is, the median is less efficient when error values have a normal distribution. In particular, the variance for the sampling distribution of the mean is about $2/\pi = 64\%$ smaller than the variance for

the sampling distribution of the median. Another way to say this is that for a sample of size 100 using the sample median has about the same efficiency as using the mean with a sample of size 64. As our earlier simulation suggested, but did not prove, the mean is clearly preferable to the median if the distribution of errors is normal. However, for some non-normal distributions of errors the median can actually be more efficient than the mean.

CHOICE OF AN ESTIMATOR AND INDEX OF ERROR

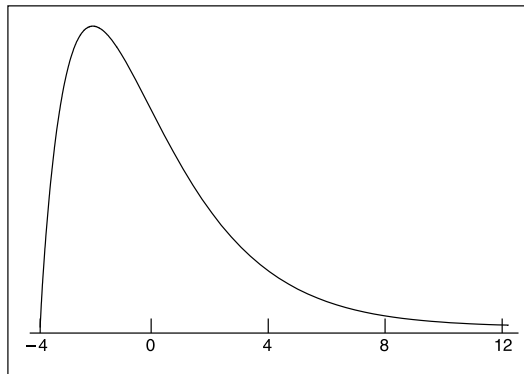
Now that we have made an explicit assumption about the distribution of the errors, we can choose the estimator and its corresponding definition of aggregate error. We want an estimator that is unbiased, efficient, and consistent.

It can be shown mathematically that if the distribution of errors is normal, then the mean is a more efficient estimator than any other unbiased estimator. That is, the mean is more efficient than the median, the mode, and any other statistic whose sampling distribution would be centered at the parameter value specified in the simple model. The mean is also a consistent estimator. Hence, we will choose it as the best estimator. We also will adopt its corresponding definition of aggregate error, which is the sum of squared errors (SSE).

There are other good reasons for selecting SSE and the mean besides the theoretical reasons above based on the assumption that the errors ε_i have a normal distribution. One very important reason is practicality. By far the greatest number of statistical procedures have been developed for SSE. Consequently, there are very good computer procedures for accomplishing the statistical tests based on SSE. There are some very practical reasons for the ubiquity of statistical procedures based on SSE. It is much easier to work with squares in mathematical proofs than with absolute values. Calculating an average is also generally easier than ordering all of the data to find the middle one. More importantly, for complex models with many parameters, the only way to estimate the best values for the parameters using most other aggregate indices of error, such as the sum of absolute errors (SAE), is brute force or iterative search algorithms. In contrast, when using SSE, we will almost always be able to use explicit formulas for calculating the best parameter estimates. Even worse, the parameters based on SAE are not necessarily unique. That is, sometimes many different values for the parameters all produce the same minimum value of SAE. This is never a problem with SSE.

Other Distributions of Errors

It is important to remember that our choice of the mean and SSE is highly dependent on our assumption that the errors are normally distributed. If the errors have some other distribution, then choosing the mean and SSE might not be appropriate. Figure 3.12 shows another possible distribution for the errors. Asymmetric distributions such as this one are *skewed* distributions. Note that in this case very large positive errors are more likely than very large negative errors. There are many types of data that can produce errors with a skewed distribution. For example, in any study measuring reaction times the smallest scores are obviously constrained to be above zero, but there is no constraint, other than the experimenter's patience, on how large the scores can be. This will tend

FIGURE 3.12 Example of a skewed distribution

to produce a scattering of observations very far above the mean and many observations just below the mean, resulting in a skewed distribution of the errors as in Figure 3.12. In Chapter 13, we discuss procedures for detecting and transforming skewed and other non-normal distributions of the errors.

One of the most troubling possible error distributions is one that has the general shape of the normal distribution except that it has “thicker” tails. That is, extreme, especially very extreme, errors are more likely to occur than would be expected from the normal distribution. These extreme errors produce extreme observations that can have a very deleterious effect on the mean. As we demonstrated in Chapter 2, one wild observation can cause a dramatic change in the mean. The problem is that SSE pays a lot of attention to extreme observations in order to reduce the really big errors, so estimates get pulled a lot in the direction of the extreme observations, so much so that the mean may not be a very good estimate of central tendency when there are a few extreme observations. The median is much more resistant to the effects of these wild observations or outliers, but we have to give up a lot of efficiency to get this robustness. We will stick with the mean and SSE, but we will have to remember that they are not resistant to outliers. To make sure that a few extreme observations are not seriously distorting our results, we will have to check carefully for outliers each time we use statistical procedures based on SSE. A number of useful follow-up procedures have been developed precisely for this purpose, and we will examine them in Chapter 13.

Other Criteria for Estimators

We have adopted unbiasedness, efficiency, and consistency as reasonable criteria for an estimator. These criteria have been adopted by many statisticians. However, not all statisticians agree on the desirability of these criteria or, especially, on their relative desirability. For example, some statisticians argue that we should be willing to accept a little bit of bias in order to achieve greater efficiency. For some of the complex models that we consider later there are biased estimators that may have greater efficiency than the unbiased estimators we present. Other statisticians argue that the mean and SSE are so susceptible to outlier problems that estimators that are somewhat less efficient but more robust should be preferred. In general, our approach is to stick with the mean and SSE while always checking assiduously for potential outliers that may be distorting the results.

Other Assumptions about Errors

In the preceding discussion of our model of errors and its metaphorical representation of Nature drawing error tickets from a bag, we made several implicit assumptions. We need to make these assumptions explicit so that we will understand what assumptions have been made and we can be alert to potential violations of these assumptions. Below we consider in turn three standard assumptions: that the ε_i are independent of one another, that they are identically distributed, and that their mean is zero. These assumptions will be especially important in considering the sampling distributions for statistics developed later, particularly in Chapter 4.

Independence

The formal assumption is that each error is independent of the value of any other error. In terms of our metaphor, this means that each error ticket is selected randomly from the bag and then replaced in the bag before the next ticket is drawn. A violation of independence occurs when the value on one error ticket allows us to make a reasonable guess about the value on another error ticket. In terms of the metaphor, it would be a violation of independence if error tickets came in pairs attached to one another so that if we drew one we would automatically draw the other. As a practical example, suppose that our data consisted of internet usage rates for each US state for each of several years. If a given state's internet usage rate were above the mean or median in one year, then it would likely be above the mean or median in other years. Thus, when we use the simple model, all the errors for that state are likely to be positive and probably of about the same size. Thus, knowing the error for that state in one year allows us to make a reasonable guess about the errors for other years. As a result, the errors would not be independent. This is an example of *time-series* data—the same units are observed or measured across a number of different time periods.

Time-series data seldom have errors that satisfy independence. As a similar example, suppose that we have two attitude scores from each subject: one score measured before the subject reads a persuasive communication and one measured after the subject reads the communication. This is known as a *within-subject* design. Again, it is probable that someone who is below the mean or median on the first measure (the “pretest”) will also be below the mean or median on the second measure (“the post-test”). Thus, if the simple model were used, the errors for each subject would likely be either both positive or both negative, so independence would once again be violated. As a final example, consider a survey that asks husbands and wives what proportion of various household tasks they perform. For a given couple, if the husband, say, claims to perform a small proportion of a particular task, then it is likely that the wife would claim to perform a large proportion of that particular task. If we tried to use a single simple model for all the data, it is likely that independence would be violated within couples. Knowing that the error for one spouse is negative (i.e., proportion claimed below the mean or median) would lead us to expect that the error for the other spouse would be positive (i.e., proportion claimed above the mean or median). Note that this dependence, unlike that in the other examples, is a *negative dependence* because the signs of the errors for each spouse are likely to be opposite.

Unlike the assumption about the normal distribution of errors, which is very robust, the assumption that the errors are independent of each other is not at all robust. That is,

our statistical inference will be seriously affected if the independence assumption is incorrect. Thus, we must be especially alert for possible dependencies among the errors. If the independence assumption is violated, then all is not lost. In Chapters 11 and 12 we show how to analyze such data so that, after suitable transformations and rearrangements, independence is satisfied.

Identically distributed

Another very important assumption is that all the errors are sampled from the *same* distribution. In terms of our metaphor, this assumption would be violated if Nature had *two* bags of error tickets: one bag containing error values of a reasonable size and another bag containing error tickets with much larger values. If Nature draws mostly from the first bag, then most of the resulting data values will be of a reasonable size. But if Nature occasionally draws an error ticket from the second bag, we will observe several “wild” data values that are very different from the other observations. Such data values are known as *outliers*. In other words, a violation of this assumption occurs if most of the errors are sampled from a normal distribution with variance σ^2 and a few errors are sampled from a normal distribution with a variance very different from σ^2 . Although it would technically be a problem if the second variance were much smaller than the first one, the greater problem occurs when the second variance is very much larger than the first.

A slightly different type of violation of this assumption occurs, for example, in experiments when the manipulation itself causes the error variance to change. In general, there is a problem whenever the size of the variance depends on the mean. In this context, the assumption of identical distributions is equivalent to assuming that the variances are equal, so it is sometimes called the *homogeneity of variance* assumption or sometimes *homoscedasticity*. Its violation is *heterogeneity of variance* or *heteroscedasticity*.

The identical distribution assumption is fairly robust. The variances must be very different before there are important problems for the data analysis. The most serious problems arise from outliers—a few observations that have errors sampled from a distribution with a very much greater variance. In Chapter 13, we consider procedures for detecting outliers and consider what to do about them. We also develop techniques for detecting heterogeneity of variance and the use of transformations to reduce the heterogeneity to acceptable limits.

Errors are unbiased

The final important assumption is that the errors are unbiased or, equivalently, that the expected value or mean of the distribution from which the errors are sampled equals zero. That is, our assumption is that, on average, the errors do not change the model parameters. If, instead, the errors were not random but systematic, such that errors averaged, say, below zero, then the systematic bias would appear in the model rather than in the error. Consider, for example, internet access data. Suppose that internet access was not properly documented for some households, for example, for those that had access through other households’ wireless internet connections. Then any estimate of the true internet access rate would necessarily be an underestimate. In effect, the systematic underreporting of error would be represented in the model and not in the error. As data

analysts we cannot do much to detect a systematic bias in the error by examining the data. Hence, it is incumbent on the collector of the data to ensure that the errors are indeed random with mean zero.

Depending on the substantive domain, ensuring that the errors are random involves aspects of experimental design, construction of survey questionnaires, interviewing procedures, instrument reliability, etc. We do not address such issues because they are highly domain-specific. However, if those issues have not been addressed in the actual collection of the data, then the use of the statistical procedures described in this book may be very misleading. As data analysts we offer no remedies for the problems of systematic error. Hence, we will always assume that errors are sampled from a distribution that has a mean of zero.

Estimator for σ^2

A final issue we need to consider is an estimator of σ^2 . We know that the mean is an estimator of the true but unknown population mean μ or β_0 . It turns out that the mean squared error (MSE) is an unbiased estimator of the true but unknown population error variance σ^2 . MSE is an index of the extent to which data deviate from the values predicted by the model. For the simple model that predicts the mean ($Y_i = b_0 + e_i$), MSE has the special name *variance* and special symbol s^2 . Thus, for the simple model that predicts the mean:

$$\text{MSE} = \text{Variance} = s^2 = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$$

which estimates the population error variance σ^2 . For more complex models with p parameters:

$$\text{MSE} = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-p}$$

which estimates σ^2 .

Note that MSE is an index of the *squared* deviations of the data from the values that are predicted by the model. A more easily interpreted index is the square root of MSE or root-mean-squared error (RMSE). RMSE estimates the population error standard deviation σ . For the simple model that predicts the mean ($Y_i = b_0 + e_i$), RMSE has the special name *standard deviation* and the special symbol s . Thus, the mean \bar{Y} , the variance s^2 , and the standard deviation s calculated from a sample of data provide estimates of the parameters μ , σ^2 , and σ that characterize a normal distribution.

Maximum Likelihood Estimation

Having laid out our assumptions concerning the distribution of errors, including that they are independent and normally distributed, it becomes possible to talk about another way of estimating parameters in models. Earlier we saw that if we minimize the sum of squared errors in estimating β_0 , then our best estimate equals the sample mean. Thus, the mean is the best *least-squares* estimate in the sense that it minimizes the sum of

squares. It turns out that for this simple model it is also what is called the *maximum likelihood* estimate, that is, the value of the estimate that maximizes the joint likelihood of all the data in the sample, given an underlying normal distribution with some variance of errors, σ^2 .

To unpack this phrase, imagine that we assume that the data in a sample do in fact come from an underlying normal distribution. Given some true value of μ and σ for that underlying distribution, we can calculate the likelihood of each data point in that distribution. This likelihood is given by the density function for the normal distribution given earlier:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Assuming all the data points in a sample come from the same distribution and are independent of each other, then one can calculate the *joint* likelihood of all the data points by taking the product of all of the likelihoods of each of the individual data points. A *maximum likelihood* estimate is that estimate of β_0 or μ that maximizes that joint likelihood, given some value of σ . And it turns out that, assuming the data are in fact normally distributed, that best estimate is \bar{Y} , the sample mean.

To illustrate this, consider the sample of 20 data points that Mother Nature sampled where the true value of β_0 was 50 and where each data point had an error ticket tacked onto it. Those 20 data points are reproduced below:

41	78	33	72	85
38	27	51	60	61
56	26	31	24	19
46	58	70	36	48

Let us assume that the standard deviation of the error tickets in Mother Nature's bag is 15. Then it becomes possible to calculate the likelihood of each individual data point under the normal distribution with different estimates of β_0 or μ .

For instance, if we started by estimating μ as 55, then the likelihood of each of the individual data points (given the normal distribution, μ estimated as 55, and $\sigma=15$) is

.0172	.0082	.0091	.0140	.0036
.0140	.0047	.0257	.0252	.0246
.0265	.0041	.0074	.0031	.0015
.0222	.0261	.0161	.0119	.0239

Each of these likelihoods is calculated using the above function to solve for $f(x)$. And the joint likelihood of all 20 data values is 6.695×10^{-40} (the product of all 20 individual likelihoods), an incredibly small number!

We could then try other estimates of μ and recalculate the likelihoods of each of the individual data points and their joint likelihood. What we would like to do is find the estimate of β_0 or μ that maximizes the joint likelihood. And it turns out that the best estimate is the sample mean, which for this sample of 20 data points is 48. The joint likelihood of the 20 data points given the estimate of 48 is 5.909×10^{-39} which is still

a very small number. But the point is that this joint likelihood is larger than what we got when we used the estimate of 55 and in fact it is larger than any other estimate we might use.

In this simple model, the maximum likelihood estimate, since it is \bar{Y} , is the same as the least-squares estimate. And in most of the models we consider in this book, this will continue to be the case. That is, in general, the parameter estimate that minimizes SSE in a sample of data will also be the maximum likelihood estimate. However, at latter points in the book, when we consider models for data that are dichotomous (Chapter 14) or those for data with multiple sources of error (Chapter 12), then we will need to rely on maximum likelihood estimates rather than least-squares estimates.

SUMMARY

As a result of random perturbations in data, sample statistics, such as the mean and the median, are unlikely to equal the true population parameters that they estimate; in other words, different samples will usually have different means (and medians) even when the samples are from the same population.

Sampling distributions show the distributions of statistics based on a very large number of samples of a particular size. So, for example, we could draw a thousand samples from the same population, calculate the mean of each sample, and then graph the set of 1000 means. The mean of this sampling distribution would equal the mean of the original theoretical distribution, but the sampling distribution would be narrower. Sampling distributions become narrower as the size of the samples from which the statistic is calculated becomes larger. The variance of the sampling distribution of the mean is equal to σ^2/n .

A comparison of the sampling distributions for the mean and the median indicates that both are unbiased and consistent estimators of the true population mean and median. However, the mean is generally more efficient than the median. For this reason, as well as practical considerations, we use the mean to estimate β_0 in the simple model and its associated aggregate index of error (the sum of squared errors) to estimate σ^2 . However, the conclusion that the mean is more efficient than the median is based on the assumption that errors are normally distributed. If the distribution of errors is skewed or contains unusual values, the mean may not be a good estimate of central tendency.

All normal distributions have the same shape and are thus often described with reference to a standard normal distribution in which $\mu = 0$ and $\sigma^2 = 1$. Normal distributions with other means and/or variances can be compared with the standard normal distribution by computing z scores or standard normal deviates. A z score reflects an observation's distance from the mean in standard deviation units. Approximately 68% of the observations in a normal distribution fall between z scores of -1 and $+1$, approximately 95% fall between z scores of -2 and $+2$, and nearly all (about 99.7%) fall between z scores of -3 and $+3$.

In addition to being normally distributed, errors are assumed to be independent of one another, identically distributed, and unbiased (i.e., have a mean of zero). Data analysts cannot determine whether errors are unbiased. Determining whether the normality, independence, and identical distribution assumptions have been violated is discussed in subsequent chapters, as are potential remedies.

Just as the mean is an estimator of the true but unknown population mean μ or β_0 , the mean squared error (MSE) is an unbiased estimator of the true but unknown population error variance σ^2 and the RMSE or square root of MSE is an unbiased estimator of the population error standard deviation σ . In the case of the simple model that predicts the mean, the MSE is known as the *variance* s^2 and the RMSE is known as the *standard deviation* s .

Finally, the mean is not only the best least-squares estimate of β_0 in the simple model, it is also the maximum likelihood estimate of β_0 if the data are normally distributed. Indeed, in general, the parameter estimates that minimize SSE will be the same as the maximum likelihood estimates. However, for dichotomous data or for data with multiple sources of error, we will use maximum likelihood estimates rather than least-squares estimates.

Simple Models

Statistical Inferences about Parameter Values

In Chapter 2 we considered various alternatives for how we should measure aggregate error in the equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

Although the sums of both the absolute errors and the squared errors seem to be reasonable alternatives, we decided in Chapter 3, for reasons of efficiency, practicality, and tradition, to define total error as the sum of the squared errors. In the simplest models, where we are estimating the single parameter β_0 , the choice of the sum of squared errors as the definition of error implies that the best estimate is the sample mean (as we proved in Chapter 2).

In this chapter, we develop procedures for asking questions or testing hypotheses about simple models. Defining and answering interesting questions is the purpose of data analysis. We first consider the logic of answering questions about data for the case of the simplest models because it is easy to focus on the logic when the models are simple and because the logic generalizes easily to more complex models. The specific statistical test presented in this chapter is equivalent to the “one-sample t -test.” We do not derive this test in terms of the t statistic; we prefer instead to construct this test using concepts and procedures that are identical to those required for the more complex models we will consider later.

The generic problem is that we have a batch of data for which we have calculated b_0 , the mean, as an estimate of β_0 . Our question is whether β_0 is equal to some specific value. For example, we might want to know whether the body temperatures of a group of patients administered a therapeutic drug differed from the normal body temperature of 37°C. We will let B_0 equal the value specified in our question. The statement:

$$\beta_0 = B_0$$

represents our *hypothesis* about the true value of β_0 . Such statements are often called *null hypotheses*. The calculated value of b_0 will almost never exactly equal B_0 , the hypothesized value of β_0 . That is, the compact model:

$$\text{MODEL C: } Y_i = B_0 + \varepsilon_i$$

in which no parameters are estimated will almost always produce a larger error than the augmented model:

$$\text{MODEL A: } Y_i = \beta_0 + \varepsilon_i$$

in which β_0 is estimated by b_0 , the mean of the batch of data. We will calculate PRE, the proportional reduction in error index developed in Chapter 1, to see how much better the predictions of Model A are than those of Model C. The original question then becomes not whether Model A is better than Model C, but whether Model A is better *enough* than Model C that we should reject the hypothesis that $\beta_0 = B_0$. Deciding what value of PRE is “better enough” is the essence of statistical inference and is the focus of this chapter.

To be less abstract, we will consider a detailed example. Suppose that 20 tickets were available for a lottery that had a single prize of \$1000. How much would individuals be willing to pay for a 1 in 20 or 5% chance of winning the \$1000 prize? The expected value of a ticket in this particular lottery would be \$1000/20 tickets or \$50. One might hypothesize that people would focus on the magnitude of the prize (i.e., the \$1000 payoff) and thus be willing to pay more than the expected value of a ticket (i.e., \$50). But one might also hypothesize that people would focus on the likelihood of losing whatever amount they paid and thus be generally willing to pay less than the \$50 expected value of a ticket. Formally, we are comparing these two models:

MODEL A: Bid = $\beta_0 + \varepsilon_i$
 MODEL C: Bid = $50 + \varepsilon_i$

Suppose that 20 individuals participated in our hypothetical lottery by submitting the following bids to buy a ticket:

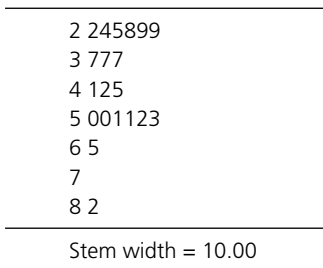
41	50	51	28	29
24	82	37	42	37
45	50	37	22	52
25	53	29	65	51

These data are displayed in Figure 4.1 in what is called a Stem and Leaf plot, where the left column of numbers indicates the left-most digit of each bid (the values of ten) and the numbers to the right of this column indicate the second digit of each bid (i.e., there were six bids in the twenties: 22, 24, 25, 28, 29, 29).

The average bid is \$42.5, which is obviously less than the expected value of \$50. However, we want to know whether these bids are really different from \$50 or whether their mean is below \$50 simply as a result of random variation in the data. In other words, even if the true state of the world were such that people generally are willing to

pay the expected value of \$50, it is unlikely that the average bid in our sample would equal \$50 *exactly*. So, we need to determine whether the average bid of 42.5 that we obtained is different *enough* from 50 that we would be willing to conclude that, on average, people are willing to pay less than the expected value. In this example, β_0 represents the *true* typical amount that people would be willing to pay for a lottery ticket. We do not know, nor can we ever know, exactly what this *true* value is. The *hypothesized* value for β_0 is B_0 , and in this example it equals 50,

FIGURE 4.1 Stem and Leaf plot for the 20 lottery bids



the expected value of a ticket in our lottery. Note that \$50 was not estimated from our data. Rather, \$50 is an a priori hypothesized value: it is a specific value that was determined before looking at the data. The *estimated* value for β_0 is b_0 , and in this case it equals 42.5, the mean or average of the 20 bids, because the mean minimizes the sum of squared errors. In other words, for the compact model (which represents the null hypothesis), the prediction is given by:

$$\text{MODEL C: } \hat{Y}_i = 50$$

and for the augmented model (in which β_0 is estimated from the data), the prediction is given by:

$$\text{MODEL A: } \hat{Y}_i = 42.5$$

So our question is whether the predictions of Model A are better enough to infer that Model C is unreasonable.

To answer our question, we want to calculate PRE. To do so, we first need to calculate the error for each model. The necessary calculations are displayed in Figure 4.2. For the compact model, $\hat{Y}_{iC} = 50$, so the sum of squared errors from the compact model, $\text{SSE}(C)$, is given as:

$$\text{SSE}(C) = \sum_{i=1}^n (Y_i - \hat{Y}_{iC})^2 = \sum_{i=1}^{20} (Y_i - 50)^2 = 5392$$

The squared error using the compact model for each bidder is listed in the third column of Figure 4.2, along with the sum of 5392 for $\text{SSE}(C)$. For the augmented model, $\hat{Y}_{iA} = 42.5$, so:

$$\text{SSE}(A) = \sum_{i=1}^n (Y_i - \hat{Y}_{iA})^2 = \sum_{i=1}^{20} (Y_i - 42.5)^2 = 4267$$

The squared error using the augmented model for each bidder is listed in the fourth column, along with its sum of 4267 for $\text{SSE}(A)$. Then the proportional reduction in error using Model A instead of Model C is given by:

$$\frac{\text{SSE}(C) - \text{SSE}(A)}{\text{SSE}(C)} = \frac{5392 - 4267}{5392} = .209$$

That is, Model A using b_0 , the *estimated* value of β_0 , has 20.9% less error than Model C using B_0 , the *hypothesized* value of β_0 . Later in this chapter, we will determine whether 20.9% less error is enough to warrant rejecting Model C (\$50) in favor of Model A.

We note in passing that for Model A one observation (bidder 7) is responsible for a substantial proportion of the total SSE. Although the presentation of formal procedures for investigating outliers must wait until Chapter 13, large errors associated with a few observations should make us suspect the presence of outliers. Remember that SSE and its associated estimators, such as the mean, are not resistant to outliers.

FIGURE 4.2 Lottery bids and error calculations for 20 bidders in a hypothetical lottery

<i>Bidder number</i>	<i>Bid</i> Y_i	<i>Squared errors</i>	
		<i>Compact</i> $(Y_i - B_0)^2$	<i>Augmented</i> $(Y_i - b_0)^2$
1	41	81	2.25
2	50	0	56.25
3	51	1	72.25
4	28	484	210.25
5	29	441	182.25
6	24	676	342.25
7	82	1024	1560.25
8	37	169	30.25
9	42	64	0.25
10	37	169	30.25
11	45	25	6.25
12	50	0	56.25
13	37	169	30.25
14	22	784	420.25
15	52	4	90.25
16	25	625	306.25
17	53	9	110.25
18	29	441	182.25
19	65	225	506.25
20	51	1	72.25
Sum	850.00	5392.00	4267.00
Mean	42.50		

DECOMPOSITION OF SSE

At this point it is useful to introduce a table that summarizes our analysis so far. The sum of squares reduced (SSR) is defined as:

$$\text{SSR} = \text{SSE}(C) - \text{SSE}(A)$$

and represents the amount of error that is reduced by using Model A instead of Model C. Then it is obvious that:

$$\text{SSE}(C) = \text{SSR} + \text{SSE}(A)$$

In other words, the original error $\text{SSE}(C)$ can be decomposed into two components: (a) the reduction in error due to Model A (i.e., SSR) and (b) the error remaining from Model A (i.e., $\text{SSE}(A)$). It is common to summarize the results of an analysis in a table having separate rows for SSR , $\text{SSE}(A)$, and $\text{SSE}(C)$. Figure 4.3 provides the generic layout for such tables, which are referred to as analysis of variance or ANOVA tables because they analyze (i.e., separate or partition) the original variance or error into component parts. Figure 4.4 presents the ANOVA summary table for our example. Note that the SS (sum of squares) for the total line, which represents $\text{SSE}(C)$, is indeed the sum of SSR and

FIGURE 4.3 Generic ANOVA layout

Source	SS	PRE
Reduction using Model A	SSR	SSR/SSE(C)
Error for Model A	SSE(A)	
Total	SSE(C)	

FIGURE 4.4 ANOVA summary table for lottery example

Source	SS	PRE
Reduction (using β_0)	1125	.209
Error (using β_0)	4267	
Total (error using B_0)	5392	

SSE(A); for our example, $5392 = 1125 + 4267$. PRE is readily obtained from the SS column using the formula:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(C)}}$$

We will use these tables, which give the decomposition of the sums of squares, as the basic summary for all our statistical tests. Later, we will add several other useful columns to such tables.

SSR is easily understood and often easily calculated as the difference between SSE(C) and SSE(A). However, there is another representation for SSR, which provides additional insight for the comparison of Models C and A. It can be shown (we essentially did it in Chapter 2 for the case of the simple model; the more general proof does not provide useful insights so we omit it) that:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_{iC} - \hat{Y}_{iA})^2 \quad (4.1)$$

where \hat{Y}_{iC} and \hat{Y}_{iA} are, respectively, the predictions for the i th observation using Model C and Model A. This formula will be useful later for calculating certain SSRs that are not automatically provided by typical computer programs. More important are the insights it provides. For a fixed SSE(C), the larger SSR is, the larger PRE is, and the larger the improvement provided by using Model A instead of Model C. This formula shows that SSR is small when Models C and A generate similar predictions for each observation. In the extreme case when Models C and A are identical (i.e., they produce the same predictions), then $\text{SSR} = 0$ and $\text{PRE} = 0$. Conversely, SSR will be large to the extent that Models C and A generate *different* predictions. Thus, SSR is a direct measure of the difference between Models C and A and $\text{PRE} = \text{SSR}/\text{SSE(C)}$ is a proportional measure of that difference.

Equation 4.1 is useful for calculating the SSR for the simple models considered in this chapter. Although we generally avoid multiple computational formulas, we present this one because many computer programs do not conveniently provide the necessary information for computing PRE in our terms. We will illustrate the use of Equation 4.1 by computing the SSR for the lottery example. The value predicted by MODEL A is $\hat{Y}_{iA} = \bar{Y} = 42.5$ and the value predicted by Model C is the hypothesized value $\hat{Y}_{iC} = B_0 = 50$. So, according to Equation 4.1:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_{iC} - \hat{Y}_{iA})^2 = \sum_{i=1}^{20} (50 - 42.5)^2 = \sum_{i=1}^{20} (7.5)^2 = 20(56.25) = 1125$$

That is, SSR equals the constant $7.5^2 = 56.25$ summed 20 times. Thus, $SSR = 20(56.25) = 1125$, which is the same value that we obtained by calculating SSR directly as $SSE(C) - SSE(A)$ in Figure 4.3. Thus, for simple models, by comparing a Model A that estimates one parameter with a Model C that estimates none, the following formula is often handy:

$$SSR = \sum_{i=1}^n (B_0 - \bar{Y})^2 = n(B_0 - \bar{Y})^2 \quad (4.2)$$

We will have many occasions to use Equation 4.2.

SAMPLING DISTRIBUTION OF PRE

It might seem that a difference in parameter estimates of $50 - 42.5 = 7.5$ and $PRE = 20.9\%$ are “large enough” to infer that Model C is unreasonable relative to Model A. Unfortunately, such a conclusion may not be warranted statistically, and it is important to understand why. To gain this understanding, we need to focus on the error term ε_i that is included in the full statement of Model C:

$$\text{MODEL C: } Y_i = B_0 + \varepsilon_i$$

As noted before, this model says that were it not for random perturbations represented by ε_i all the Y_i values would equal B_0 exactly. In Chapter 3 we made the assumption that the ε_i values are all sampled randomly and independently from a normal distribution with mean 0 and variance σ^2 . We also saw in Chapter 3 that the exact value for the mean calculated from a sample of size n would depend on the particular sample of errors. Sometimes the calculated mean would be above the true value B_0 , and sometimes it would be below. That is, there would be a sampling distribution for the mean. If Model C were correct, sometimes the sample mean would be somewhat higher than 50 and other times it would be somewhat lower, but it would seldom equal 50 exactly. For example, we would most likely have obtained a different mean if the bids had been gathered before lunch or the day before or a day later, because the pattern of random perturbations would have been different.

Similar to the sampling distribution for the mean, there is also a sampling distribution for PRE. Just as b_0 , calculated from the data, is the estimate of the unknown true parameter β_0 , so too PRE, calculated from the data, is the estimate of the unknown true proportional reduction in error η^2 . For the moment, let us assume that Model C is correct (i.e., that $\beta_0 = B_0$) and consider what the sampling distribution for PRE would be. In other words, we begin by assuming that the null hypothesis is true. In these terms, $\eta^2 = 0$ is equivalent to Model C being correct and to Model A making absolutely no improvement relative to Model C. We saw in Chapter 3 that b_0 has a sampling distribution, so even if Model C is correct we would not expect our estimate b_0 to equal B_0 exactly. But we know that b_0 produces the smallest possible sum of squared errors, so $SSE(A)$, using b_0 , must always be at least a little smaller than $SSE(C)$, using B_0 . For example, in the lottery data the mean will seldom equal 50 exactly, even if the true parameter value were 50, and thus the SSE calculated using the sample mean will always be a little less than $SSE(C)$.

(see the proof in Chapter 2). Hence, even if the true proportional reduction in error $\eta^2 = 0$ (as it must when Model C is correct), the calculated PRE will always be at least a little greater than zero (never less than zero). PRE is therefore a biased estimator of η^2 because PRE will always overestimate the true value of η^2 . We will return to this issue of the bias in PRE later. For now, the important point is that we should not expect the calculated PRE to equal zero even when Model A makes no improvement on Model C. Thus, we cannot base our decision about the validity of Model C simply on whether or not $\text{PRE} = 0$.

If we cannot use $\text{PRE} \neq 0$ as a criterion for rejecting Model C, then we need to consider the sampling distribution of PRE to determine whether the calculated value of PRE is a *likely* value, *assuming that Model C is correct*. If the calculated value of PRE is a likely value, then we ought not to reject Model C and its equivalent null hypothesis that $\beta_0 = B_0$. On the other hand, if the calculated value of PRE is an unlikely value when Model C is assumed to be true, then we ought to reject Model C in favor of Model A. In terms of our example, we need to consider the sampling distribution of PRE to determine for this case whether $\text{PRE} = .209$ is a likely value, assuming that Model C is correct (i.e., that $\beta_0 = 50$). If it is a likely value, then we ought not to reject Model C and its equivalent hypothesis that $\beta_0 = B_0 = 50$; there would be no evidence that the lottery bids are different from what would be expected if the null hypothesis were true. If $\text{PRE} = .209$ is an unexpected value, then we ought to reject Model C and its hypothesis in favor of Model A and its estimate that $\beta_0 = b_0 = 42.5$; in other words, we would conclude that the lottery bids were significantly lower than the expected value of \$50 for a ticket.

We could develop the sampling distribution for PRE in this example using the same simulation strategy we used in the previous chapter. That is, we could put error tickets of the appropriate size into a bag and then do many simulation rounds, in each of which we would randomly select 20 error tickets, add 50 to each, and calculate PRE for that sample. The only problem with this strategy is that we do not know what size error tickets to place in the bag to be sampled. In other words, we do not know the variance σ^2 of the normal distribution of errors. However, as was noted in the previous chapter, the mean squared error provides an estimate of σ^2 . In particular, for Model A with one parameter the estimate is:

$$s^2 = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} = \frac{4267}{19} = 224.58$$

We could therefore conduct the simulation by sampling error values from a normal distribution with mean 0 and variance 224.58. The 100 error tickets in Figure 3.1 were sampled from such a normal distribution, so they could be used as the bag of error tickets for the simulation.

As an example of a simulation round, suppose that the following 20 error tickets were drawn from Figure 3.1:

17	31	-2	-6	-17
-4	28	29	17	1
-12	-6	3	-25	2
24	15	-20	4	-3

These error terms when added to the value of $B_0 = 50$ of Model C yield the following 20 scores:

67	81	48	44	33
46	78	79	67	51
38	44	53	25	52
74	65	30	54	47

The mean of the resulting 20 scores is 53.8 and $SSE(C)$, using 50 as the model prediction for all the observations, and $SSE(A)$, using 53.8 as the model prediction, are easily calculated to be 5554 and 5265.2, respectively. Thus:

$$PRE = \frac{5554 - 5265.2}{5554} = .052$$

Then a new simulation round with a new sample of error values would produce a different mean and PRE. These simulation rounds could conceptually be repeated until there were enough PRE values to make a sampling distribution for PRE.

Alas, the simulation strategy outlined above for generating the sampling distribution of PRE will not work because $s^2 = 224.58$ is only an *estimate* of the true variance of the errors σ^2 . Just as it is unlikely that $\bar{Y} = b_0$ (the calculated mean for a sample) will equal B_0 exactly, it is also unlikely that the calculated variance s^2 will equal σ^2 exactly. In other words, we are uncertain about the exact size of the error tickets that should be placed in the bag.

We could conduct a more complex sampling simulation to account for our uncertainty about the size of the error tickets. However, it would be tedious if we had to do a new simulation round for each data analysis. Fortunately, this is not necessary because mathematical statisticians have specified the sampling distribution for PRE based on the assumptions we made in Chapter 3 about the behavior of the errors ε_i . Even though we will not actually do simulations for generating a sampling distribution, it is important to remember that the mathematical formula for that distribution is derived from the assumption that the error values are randomly sampled from a distribution with mean 0 and variance σ^2 and that the sampling could be represented by drawing error tickets from a bag.

Figure 4.5 provides a tabular description of the sampling distribution of PRE for the particular case of samples of size 20, again assuming the validity of Model C. That is, if Model C were correct (in our case, $\beta_0 = 50$), and if we compared Model C ($\hat{Y} = 50$) with Model A ($\hat{Y} = b_0 = \bar{Y}$) from samples, then Figure 4.5 presents the proportional frequency and cumulative proportion for each range of PRE. The proportional frequencies are plotted in Figure 4.6. As we would expect, the sampling distribution in Figure 4.6 shows that values of PRE near zero are the most likely. It also shows that values of PRE greater than .2 are infrequent.

The cumulative proportions are generally more useful than the proportions for individual ranges. The cumulative proportion is simply the total proportion for the range of PREs from zero to the value of interest. For example, to find the cumulative proportion for the range from 0 to .03, we simply add the proportions for the three component ranges: 0 to .01, .01 to .02, and .02 to .03. For this range, $.334 + .125 + .088 = .547$. That is, 54.7% of the simulated PRE values are less than or equal to .03. The cumulative

FIGURE 4.5 Tabular description of the sampling distribution of PRE for testing the simple model with 20 observations

<i>PRE range</i>	<i>Proportion</i>	<i>Cumulative proportion</i>
.00-.01	.334	.334
.01-.02	.125	.459
.02-.03	.088	.547
.03-.04	.068	.615
.04-.05	.055	.670
.05-.06	.045	.716
.06-.07	.038	.754
.07-.08	.032	.786
.08-.09	.028	.814
.09-.10	.024	.838
.10-.11	.021	.858
.11-.12	.018	.876
.12-.13	.016	.892
.13-.14	.014	.905
.14-.15	.012	.917
.15-.16	.010	.928
.16-.17	.009	.937
.17-.18	.008	.945
.18-.19	.007	.952
.19-.20	.006	.958
.20-.21	.005	.963
.21-.22	.005	.968
.22-.23	.004	.972
.23-.24	.004	.976
.24-.25	.003	.979
.25-.26	.003	.982
.26-.27	.002	.984
.27-.28	.002	.986
.28-.29	.002	.988
.29-.30	.002	.990
.30-.31	.001	.991
.31-.32	.001	.993
.32-.33	.001	.994
.33-.34	.001	.995
.34-.35	.001	.995
.35-.36	.001	.996
.36-.37	.001	.997
.37-.38	.001	.997
.38-.39	.000	.998
.39-.40	.000	.998

FIGURE 4.6 Sampling distribution of PRE for testing the simple model with 20 observations

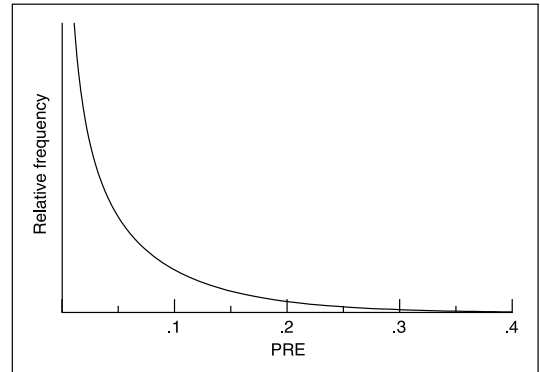
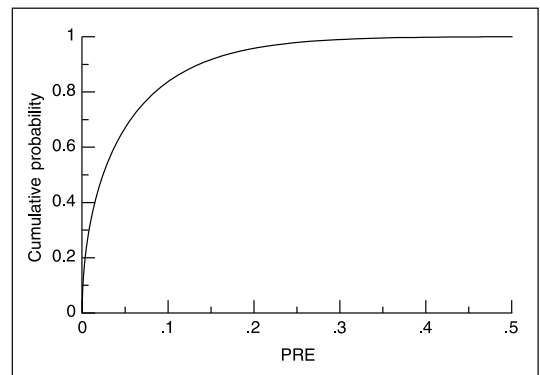


FIGURE 4.7 Cumulative proportions for $n - PA = 19$ ($PA =$ number of parameters for Model A)



proportions are displayed in the last column of Figure 4.5 and graphed in Figure 4.7. We can see, for example, from both the table and the graph of the cumulative proportions that PRE is less than .1 about 84% of the time.

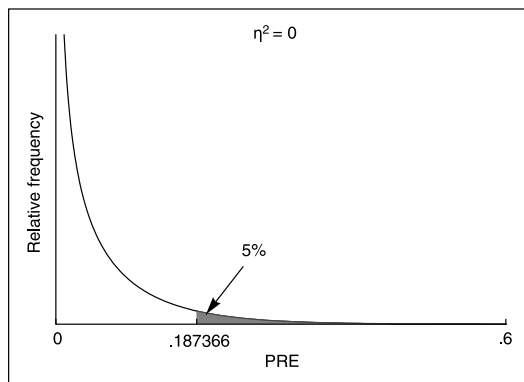
We can now ask whether a value for PRE of .209 is likely if Model C is correct. From the table in Figure 4.5 we see that approximately 96% of the time PRE would be less than .209. Or, in other words, a PRE as large as .209 would be obtained only about

4% of the time if Model C were correct. We can finally answer our question. It is unlikely (less than 5% chance) that we would have obtained a PRE this large had Model C been correct. We can therefore reasonably reject Model C in favor of Model A. This is equivalent to rejecting the null hypothesis that $\beta_0 = B_0 = 50$. Substantively, the data indicate that participants were willing to pay less than the expected value of the lottery tickets.

CRITICAL VALUES

From the mathematical equations describing the sampling distribution for PRE, we can determine for our example data that if Model C were correct (and thus $\eta^2 = 0$) we would expect 95% of the simulated values of PRE to be below the precise value of .187. This sampling distribution is plotted in Figure 4.8. In the social sciences it is customary to consider a value of PRE to be surprising if it occurs by chance less than 5% of the time when the null hypothesis is true. Thus, .187 is the *critical value* for this example; any value of $PRE > .187$ causes us to reject Model C. Using the equations, for any number of observations we can calculate the value of PRE for which we would expect 95% of the simulated PRE values to be below if Model C were correct. Figure 4.9 gives the 95% (and 99%) critical values for selected numbers of observations.

FIGURE 4.8 Distribution of PRE for $n - PA = 19$, assuming that $\eta^2 = 0$ (PA = number of parameters for Model A)



STATISTIC F

Figures of the critical values for PRE are rare in statistics books. Much more common, for largely historical reasons, are tables of F , a statistic closely related to PRE. As we shall see below, F is a simple function of PRE, so if we know PRE, the number of observations, and the number of parameters in Models C and A, then we also know F and vice versa. By re-expressing PRE, F also provides additional insights about the proportion of error reduced. We therefore turn to the motivation for calculating F and then consider its sampling distribution.

FIGURE 4.9 Critical values (95% and 99%) for PRE and F for testing models that differ by one parameter (PA = number of parameters for Model A)

$n - PA$	95%		99%	
	PRE	F	PRE	F
1	.994	161.45	1.000	4052.18
2	.903	18.51	.980	98.50
3	.771	10.13	.919	34.12
4	.658	7.71	.841	21.20
5	.569	6.61	.765	16.26
6	.499	5.99	.696	13.75
7	.444	5.59	.636	12.25
8	.399	5.32	.585	11.26
9	.362	5.12	.540	10.56
10	.332	4.97	.501	10.04
11	.306	4.84	.467	9.65
12	.283	4.75	.437	9.33
13	.264	4.67	.411	9.07
14	.247	4.60	.388	8.86
15	.232	4.54	.367	8.68
16	.219	4.49	.348	8.53
17	.208	4.45	.331	8.40
18	.197	4.41	.315	8.29
19	.187	4.38	.301	8.19
20	.179	4.35	.288	8.10
22	.164	4.30	.265	7.95
24	.151	4.26	.246	7.82
26	.140	4.23	.229	7.72
28	.130	4.20	.214	7.64
30	.122	4.17	.201	7.56
35	.105	4.12	.175	7.42
40	.093	4.09	.155	7.31
45	.083	4.06	.138	7.23
50	.075	4.03	.125	7.17
55	.068	4.02	.115	7.12
60	.063	4.00	.106	7.08
80	.047	3.96	.080	6.96
100	.038	3.94	.065	6.90
150	.025	3.90	.043	6.81
200	.019	3.89	.033	6.76
500	.008	3.86	.013	6.69
∞		3.84		6.63

The two reasons for calculating F are (a) to examine the proportional reduction in error *per additional parameter* added to the model and (b) to compare the proportion of error that was reduced (PRE) with the proportion of error that remains ($1 - \text{PRE}$). In the context of the simple models that we are considering in this chapter, PRE is obtained by the addition of only one parameter. But later we will want to consider the improvement produced by models that add more than one parameter. To avoid having to present different formulas as the models become more complex, we will present the general definition of F here. The key idea is that a given PRE, let us say .35, is more impressive

when obtained by the addition of a single parameter than when it is obtained by the addition of several parameters. So, we want to consider *PRE per parameter*. That is, we divide PRE by the number of additional parameters used in Model A that are not used in Model C. We will let PA and PC represent the number of parameters for Model A and Model C, respectively. Then, the number of additional parameters is simply PA – PC. Hence, F is based on the quantity:

$$\frac{\text{PRE}}{\text{PA} - \text{PC}}$$

which is simply the proportional reduction in error per additional parameter. For the simple models of this chapter, there are no parameters to be estimated for Model C and only one for Model A, so PC = 0, PA = 1, and PA – PC = 1.

Similarly, we need to consider the remaining proportion of the error, $1 - \text{PRE}$, in terms of the number of additional parameters that *could* be added to reduce it. As noted in Chapter 1, the most parameters we can have is one for each observation Y_i . If there are n observations and we have already used PA parameters in Model A, then at most we could add $n - \text{PA}$ parameters to some more complex model. So:

$$\frac{1 - \text{PRE}}{n - \text{PA}}$$

is the proportion of remaining error per parameter that *could* be added to the model. In other words, this is the average remaining error per additional parameter. If we added a parameter to the model at random, even a parameter that was not really useful, we would expect at least some reduction in error. The proportion of remaining error per parameter or the average remaining error tells us the value of PRE to expect for a worthless parameter. If the parameter or parameters added to the model are genuinely useful, then the PRE per parameter that we actually obtain ought to be substantially larger than the expected PRE per parameter for a useless, randomly selected parameter. An easy way to compare PRE per parameter obtained with the expected PRE per parameter is to compute the ratio of the two quantities; this gives the definition of F as:

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} \quad (4.3)$$

We can think of the numerator of F as indicating the average proportional reduction in error per parameter added, and the denominator as the average proportional reduction in error that could be obtained by adding all possible remaining parameters. For Model A to be significantly better than Model C, we want the average error reduction for the parameters added to be much greater than the average error reduction we could get by adding the remainder of the possible parameters. Hence, if F is about 1, then we are doing no better than we could expect on average, so values of F near 1 suggest that we should not reject the simpler Model C. Values of F much larger than 1 imply that the average PRE per parameter added in Model A is much greater than the average that could be obtained by adding still more parameters. In that case, we would want to

reject Model C (and its implicit hypothesis) in favor of Model A. For the example of Figure 4.2 where $PRE = .209$ and $n = 20$:

$$F = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} = \frac{.209/(1 - 0)}{.791/(20 - 1)} = \frac{.209}{.0416} = 5.02$$

In other words, we obtained a 20.9% reduction in error per parameter by using Model A, and the further reduction we could get by adding all the additional parameters is only 4.16% per parameter. Their ratio of 5.02 suggests that adding to Model C the one specific additional parameter β_0 , which is estimated by the mean, yields a substantially better (about five times better) PRE than we could expect by randomly adding a parameter from the remaining ones. In other words, the increased complexity of Model A is probably worth it. But again we need to consider the sampling distribution of F to determine whether a value of 5.02 is indeed “surprising.”

Again, mathematical equations exist for describing the sampling distribution of F , given the assumptions about error discussed in Chapter 3. If we assume that the errors ε_i are independently, identically, and normally distributed, then F has what is known as an *F distribution*. The 95% and 99% critical values for F for testing simple models are listed in Figure 4.9 next to their corresponding values of PRE. F and PRE are redundant in the sense that one exceeds its critical value if and only if the other one exceeds its critical value. For the example, $PRE = .209$ exceeds its critical value of .187, and so necessarily the corresponding $F = 5.02$ exceeds its critical value of 4.38. Thus, either PRE or F leads us to reject Model C and its implicit hypothesis that $\beta_0 = B_0 = 50$. Note that for most reasonable numbers of observations the 95% critical value for F is about 4. If we ignore the fractional part of F , then a useful rule of thumb that reduces the need to consult statistical tables frequently is to reject Model C in favor of Model A whenever F is greater than 5. If F is between 4 and 5, then you will probably have to look it up in the table, and if it is below 4, then there is no hope unless the number of observations is extremely large. Critical values of PRE and F for testing the more complex Model A that differs from Model C by more than one parameter are listed in the Appendix as a function of $PA - PC$, often called the “numerator degrees of freedom” because that term appears in the numerator of the formula for F (Equation 4.3), and $n - PA$, often called the “denominator degrees of freedom” because it appears in the denominator of the formula for F .

It is useful to add the degrees of freedom (df) and F to the basic summary table we began earlier. Figure 4.10 presents such a table for our example. It is our policy to avoid multiple computational formulas for the same quantity and instead to present only one conceptual formula. However, we must break that policy for F in this instance because F is traditionally calculated by an equivalent but different formula based on Figure 4.10. Figures constructed using the alternative formula for F are ubiquitous, so the reader has no choice but to learn this alternative in addition to the conceptual formula for F presented above. The alternative formula for F is:

$$F = \frac{SSR/(PA - PC)}{SSE(A)/(n - PA)} = \frac{MSR}{MSE}$$

For our example, this yields:

$$F = \frac{SSR/(PA - PC)}{SSE(A)/(n - PA)} = \frac{1125/(1 - 0)}{4267/(20 - 1)} = \frac{1125}{224.58} = 5.01$$

This agrees with our previous calculation except for a small rounding error. MSR represents the *mean squares reduced*, and MSE represents the *mean square error*. To facilitate this calculation, we usually add an “MS” column to the summary table. The final column, labeled “*p*,” gives the probability of obtaining a PRE and *F* that is that large or larger if $\eta^2 = 0$. In this case, PRE and *F* exceed the 95% critical values so the probability of getting a PRE or *F* that is that large or larger if Model C were correct is less than .05. With the additional columns, Figure 4.11 provides a detailed summary of our analysis of the lottery bids.

FIGURE 4.10 Analysis of variance summary table: decomposition of sums of squares

Source	SS	df	MS	F	PRE	<i>p</i>
Reduce, Model A	SSR	PA - PC	MSR = $\frac{SSR}{PA - PC}$	$\frac{MSR}{MSE}$	$\frac{SSR}{SSE(C)}$	
Error for Model A	SSE(A)	n - PA	MSE = $\frac{SSE(A)}{n - PA}$			
Total	SSE(C)	n - PC				

FIGURE 4.11 ANOVA summary table for lottery example

Source	SS	df	MS	F	PRE	<i>p</i>
Reduce, Model A	1125	1	1125.00	5.01	.209	<.05
Error for Model A	4267	19	224.58			
Total	5392	20				

STATISTICAL DECISIONS

We have now defined the essence of statistical inference: if PRE and *F* exceed their respective critical values then the simpler Model C is rejected in favor of the more complex Model A. We now have a rule for resolving the inherent tension in data analysis between reducing the error as much as possible and keeping the model of the data as parsimonious as possible. It is important to recognize, however, that statistical inference is probabilistic and therefore not infallible. That is, if Model C is actually the correct model, then 5% of the time we will obtain values of PRE and *F* that exceed their 95% critical values. Our rule is to reject Model C if those statistics exceed their 95% critical values, so in such instances we will have made a mistake in rejecting Model C. There is no way to avoid making occasional mistakes of that type. By adopting a 95% critical value, we are implicitly accepting that for those data for which Model C is correct we are willing to risk a 5% chance of incorrectly rejecting it. Mistakes of this type are known as *Type I* errors. The choice of 5% as an acceptable rate of Type I errors is inherently

arbitrary. If we want to be more cautious we could choose a rate of 1%, or if we are willing to risk more Type I errors then we might choose a rate of 10%.

We also can commit a *Type II* error. A Type II error occurs when Model C is incorrect but the obtained values of PRE and F still do not exceed their critical values. Thus, a Type II error occurs when we fail to reject Model C when we should. That is, Model C is incorrect and Model A is significantly better, but we are unlucky in terms of the error tickets drawn and miss seeing the difference. Figure 4.12 summarizes the statistical decision that confronts us and defines the ways in which both the right and wrong decisions can be made. Statistical inference can be viewed as a game with Nature. Nature determines whether Model C is correct or incorrect. The goal of the data analyst is to “guess” which is the case. The data analyst uses the data to make an informed guess. Specifically, if PRE and F exceed their critical values, then the decision is to “reject Model C”; otherwise the decision is “do not reject Model C.” If Nature has determined that Model C is correct, then, using a 95% critical value, we will decide correctly 95% of the time and incorrectly (i.e., make a Type I error) 5% of the time. The chance of making a Type I error is often labeled α and referred to as the *significance level*.

On the other hand, if Nature has determined that Model C is incorrect, then we will decide correctly the proportion of times that PRE and F exceed their critical values, and we will decide incorrectly (i.e., make a Type II error) the proportion of times that PRE and F fall below their critical values. The chance of making a Type II error is, unfortunately, often labeled β , which should not be confused with β_0 , β_1 , etc., which we use to represent parameters in a model. The proportion of times the correct decision is made when Nature has determined that Model C is incorrect, $1 - \beta$, is often referred to as the *power* of a statistical test.

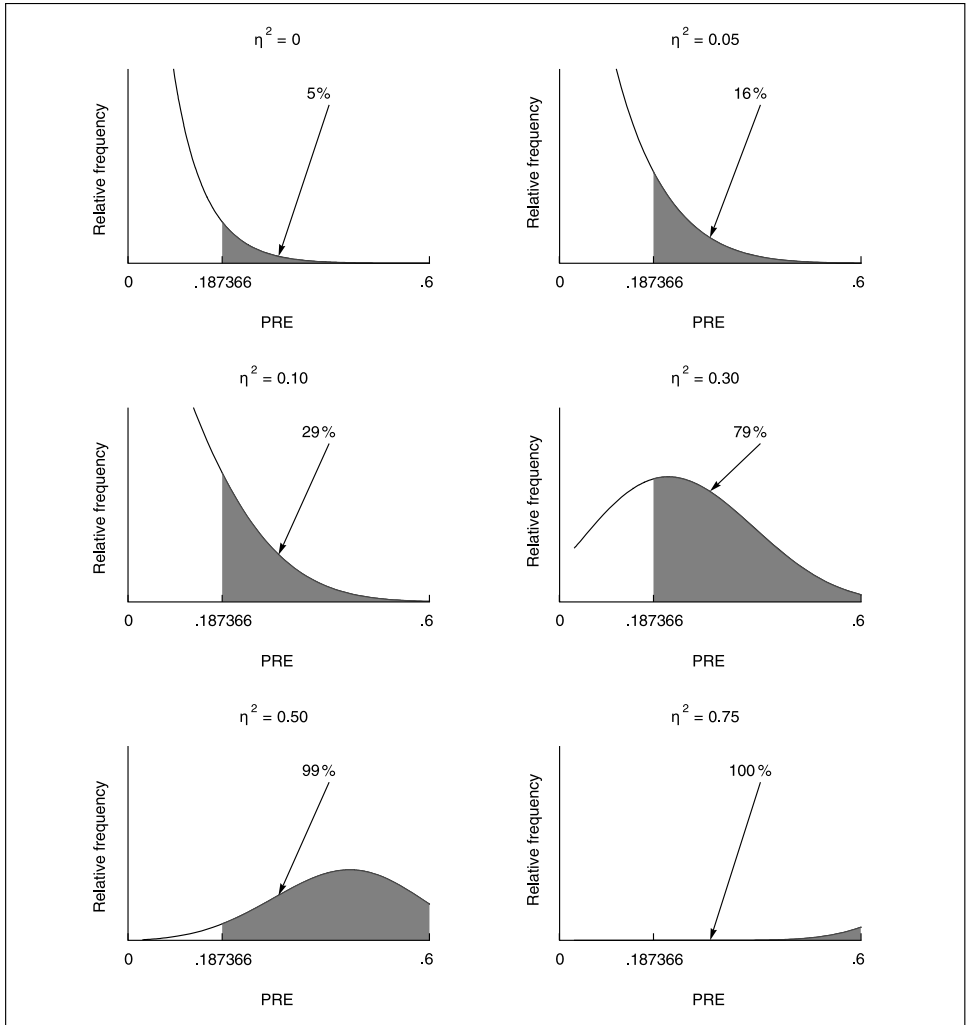
FIGURE 4.12 The statistical decision and the two types of errors

Statistical decision	True state of Nature	
	Model C correct	Model C incorrect
“Reject Model C”	Type I error	Correct decision
“Do not reject Model C”	Correct decision	Type II error

ESTIMATING STATISTICAL POWER

To determine the chances of a Type II error or to determine the power, we need to know the sampling distribution for PRE and F , assuming that Model C is incorrect. We cannot determine such a sampling distribution in general because to say that Model C is incorrect is to say only that the true proportional reduction in error η^2 is greater than zero. However, using the equations provided by mathematical statisticians, we can easily derive the sampling distributions for the specific values of η^2 we might want to consider. For our example in which $PA - PC = 1$ and $n - PA = 19$, Figure 4.13 displays plots of the sampling distribution for PRE assuming progressively greater true values of PRE, that is, η^2 . Note that if $\eta^2 = .1$, the probability of obtaining a PRE greater than the critical value and thus rejecting Model C is 29% as compared with only 5% if Model C is correct and $\eta^2 = 0$. Figure 4.14 displays the cumulative probability distributions for other

FIGURE 4.13 Distributions of PRE for $PA - PC = 1$ and $n - PA = 19$, assuming various values for η^2



observed values of PRE if we were to assume that η^2 were equal to 0, .05, .1, .3, .5, and .75, but again only for the very particular conditions of our example: $PA - PC = 1$ and $n - PA = 19$. Different sampling distributions would be obtained for other combinations of $PA - PC$ and $n - PA$. The column for $\eta^2 = 0$ corresponds exactly to the cumulative probability distribution of Figure 4.5. Each entry in Figure 4.14 is the probability that PRE calculated from the data would be less than the value of PRE specified for that row. For example, the value of .08 in the row for $PRE = .30$ and the column for $\eta^2 = .50$ means that if η^2 (i.e., the true PRE) were really .5, then the probability of obtaining a PRE (calculated from the data) of .3 or lower is 8%.

We can use the cumulative probability distributions of PRE for different values of η^2 to perform “what if” analyses. For example, we can ask, “*what* would the chances of making a Type II error be *if* $\eta^2 = .3$?” To answer this question, we first must decide what chance of a Type I error we are willing to risk. If we adopt the customary value

FIGURE 4.14 Cumulative sampling distributions of PRE for various η^2 when $PA - PC = 1$ and $n - PA = 19$

PRE	True PRE, η^2					
	0	.05	.1	.3	.5	.75
.00	.00	.00	.00	.00	.00	.00
.05	.67	.47	.31	.03	.00	.00
.10	.84	.65	.49	.08	.00	.00
.15	.92	.78	.63	.16	.01	.00
.20	.96	.86	.74	.25	.02	.00
.25	.98	.92	.83	.37	.04	.00
.30	.99	.95	.89	.49	.08	.00
.35	1.00	.97	.94	.61	.14	.00
.40	1.00	.99	.96	.72	.23	.00
.45	1.00	.99	.98	.81	.35	.00
.50	1.00	1.00	.99	.88	.48	.00
.55	1.00	1.00	1.00	.93	.62	.01
.60	1.00	1.00	1.00	.97	.75	.04
.65	1.00	1.00	1.00	.99	.86	.11
.70	1.00	1.00	1.00	1.00	.94	.25
.75	1.00	1.00	1.00	1.00	.98	.47
.80	1.00	1.00	1.00	1.00	.99	.73
.85	1.00	1.00	1.00	1.00	1.00	.92
.90	1.00	1.00	1.00	1.00	1.00	.99
.95	1.00	1.00	1.00	1.00	1.00	1.00

in the social sciences of $\alpha = .05$, then, as before, we select the critical value of PRE so that the calculated value of PRE has only a .05 probability of exceeding that critical value if $\eta^2 = 0$. In this case, we see from Figure 4.14 that the probability of obtaining a PRE less than or equal to .20 equals .96, so only 4% of the observed values of PRE should be greater than .20 if $\eta^2 = 0$. Remember that $\eta^2 = 0$ implies that Model C and Model A are identical, so our decision rule will be to reject Model C in favor of Model A if $PRE > .20$. Now we can use the column for $\eta^2 = .3$ to answer our “what if” question. The entry in that column for the row for $PRE = .20$ reveals that the probability that the calculated PRE will be less than .20 is .25. That is, even if $\eta^2 = .3$ (i.e., there is a real difference between Model C and Model A), there is still a 25% chance that we will obtain a value of PRE below the critical value and hence will not reject Model C. In other words, the probability of making a Type II error is .25. Conversely, $1 - .25 = .75$ is the probability of obtaining $PRE > .20$ and hence rejecting Model C in favor of Model A if $\eta^2 = .3$. In other words, the power (probability of not making a Type II error) is .75.

We can easily do the “what if” analysis for other values of η^2 . For example, if $\eta^2 = .05$, which implies a small difference between Models C and A, then power equals $1 - .86 = .14$. That is, the chances of obtaining a PRE large enough to reject Model C in favor of Model A would be only 14% for this small difference between the two models. On the other hand, if $\eta^2 = .75$, which implies a very large difference between Models C and A, then power equals $1 - .00 = 1$. That is, for this large difference we would be virtually certain to obtain a PRE large enough to reject Model C in favor of Model A.

FIGURE 4.15 Power table for $\alpha = .05$ when PC = 0 and PA = 1

n	Critical values		Prob(PRE > critical value)							
			True PRE, η^2							
	F	PRE	0	.01	.03	.05	.075	.1	.2	.3
2	161.45	.994	.05	.05	.05	.05	.05	.06	.06	.07
3	18.51	.903	.05	.05	.05	.06	.06	.07	.08	.11
4	10.13	.771	.05	.05	.06	.06	.07	.08	.11	.15
5	7.71	.658	.05	.05	.06	.07	.08	.09	.14	.21
6	6.61	.569	.05	.05	.06	.07	.09	.10	.17	.26
7	5.99	.499	.05	.06	.07	.08	.10	.12	.20	.31
8	5.59	.444	.05	.06	.07	.09	.11	.13	.23	.36
9	5.32	.399	.05	.06	.08	.09	.12	.14	.26	.41
10	5.12	.362	.05	.06	.08	.10	.13	.16	.29	.46
11	4.96	.332	.05	.06	.08	.11	.14	.17	.32	.50
12	4.84	.306	.05	.06	.09	.11	.15	.18	.35	.54
13	4.75	.283	.05	.06	.09	.12	.16	.20	.38	.58
14	4.67	.264	.05	.06	.09	.13	.17	.21	.41	.62
15	4.60	.247	.05	.07	.10	.13	.18	.23	.44	.66
16	4.54	.232	.05	.07	.10	.14	.19	.24	.46	.69
17	4.49	.219	.05	.07	.10	.14	.20	.25	.49	.72
18	4.45	.208	.05	.07	.11	.15	.21	.27	.52	.74
19	4.41	.197	.05	.07	.11	.16	.22	.28	.54	.77
20	4.38	.187	.05	.07	.12	.16	.23	.29	.56	.79
22	4.32	.171	.05	.07	.12	.18	.25	.32	.61	.83
24	4.28	.157	.05	.08	.13	.19	.27	.35	.65	.87
26	4.24	.145	.05	.08	.14	.20	.29	.37	.69	.89
28	4.21	.135	.05	.08	.15	.22	.31	.40	.72	.92
30	4.18	.126	.05	.08	.15	.23	.33	.42	.75	.93
35	4.13	.108	.05	.09	.17	.26	.37	.48	.82	.96
40	4.09	.095	.05	.10	.19	.29	.42	.54	.87	.98
45	4.06	.085	.05	.10	.21	.32	.46	.59	.91	.99
50	4.04	.076	.05	.11	.23	.36	.51	.64	.93	**
55	4.02	.069	.05	.11	.25	.39	.55	.68	.95	**
60	4.00	.064	.05	.12	.27	.42	.58	.72	.97	**
80	3.96	.048	.05	.14	.34	.53	.71	.84	.99	**
100	3.94	.038	.05	.17	.41	.62	.81	.91	**	**
150	3.90	.026	.05	.23	.57	.80	.93	.98	**	**
200	3.89	.019	.05	.29	.70	.90	.98	**	**	**
500	3.86	.008	.05	.61	.98	**	**	**	**	**

** Power > .995

The cumulative sampling distributions are unwieldy, and only a few of the numbers are actually needed for the “what if” analyses, so a “power table” that only gives the power probabilities for specified levels of η^2 is more useful. Figure 4.15 gives the power probabilities for selected values of η^2 and n when PC = 0, PA = 1, and $\alpha = .05$. We can use this table to do the same “what if” analyses that we did above, as well as many others. For our example problem we simply use the row for $n = 20$: for $\eta^2 = .05$, power = .16; and for $\eta^2 = .3$, power = .79. (The small differences from the power calculations above are due to using the more precise critical value of .187 for PRE instead of .20.)

The power table allows us to ask another kind of “what if” question: “What would the power be if the sample size were increased?” For example, how much would the power increase if the lottery bids were evaluated with 30 bidders instead of 20? If $\eta^2 = .05$, then power = .23, which is better than .16 for 20 bidders, but still not very good. If $\eta^2 = .3$, then power = .93, which is much higher than .79 for 20 bidders and gives an excellent chance of rejecting Model C in favor of Model A. Note that for $n = 50$ we are virtually certain of rejecting Model C whenever the true PRE η^2 is equal to or greater than .3.

Too many researchers fail to ask “what if” power questions before they collect their data. The consequence is often a study that has virtually no chance of rejecting Model C even if the idea that motivated the study is correct. With power tables such as Figure 4.15 (most software programs, such as R and SAS, have easily accessible procedures for calculating power), asking “what if” power questions is so easy that there is no excuse for not asking those questions before collecting data. A natural question is how high should statistical power be? Cohen (1977) suggested that power should be at least .8. However, the ultimate decision is how much the researcher is willing to accept the risk of not finding a significant result even when the ideas motivating the study are correct.

Now that we know how to answer easily “what if” power questions, we need to know what values of true PRE or η^2 are appropriate for those “what if” questions. There are three ways to obtain an appropriate value for η^2 to use in the power analysis: (a) obtain values of PRE from similar research; (b) use Cohen’s (1977) suggested values for “small,” “medium,” and “large” effects; and (c) compute expectations for PRE from guesses about the parameter values. We consider each in turn.

First, with sufficient experience in a research domain, researchers often know what values of PRE are important or meaningful in that domain. Those values of PRE from experience can be used directly in the power table. For example, if, based on past experience, we thought that important effects (such as the effect of the lottery bids) produced PREs greater than or equal to .1, then we could use the $\eta^2 = .1$ column of the power table. If we wanted to ensure that power $> .8$, then going down the column we find that the first power $> .8$ requires a sample size between 60 and 80, probably about 73, which is a much greater number of participants than we included in our test of whether people were willing to pay significantly less than the expected value of a lottery ticket.

In using the results of past studies to select an appropriate η^2 for the “what if” power analysis, we must remember that calculated values of PRE are biased because on average they overestimate η^2 . The following simple formula can be used to remove the bias from PRE:

$$\text{Unbiased estimate of } \eta^2 = 1 - (1 - \text{PRE}) \left[\frac{n - \text{PC}}{n - \text{PA}} \right]$$

For our example in which we calculated $\text{PRE} = .209$, the unbiased estimate of η^2 equals:

$$1 - (1 - .209) \left[\frac{20 - 0}{20 - 1} \right] = 1 - .791 \left[\frac{20}{19} \right] = .167$$

Thus, although the value of PRE calculated from the data is .209, for planning further research our best unbiased guess for the true value of η^2 is only .167. The correction for bias has more of an effect for small values of $n - \text{PA}$ than for large values. In essence,

the adjustment corrects for the ability of least squares to capitalize on chance for small sample sizes. These unbiased estimates are thus sometimes referred to as “adjusted” values.

A second and related method for finding appropriate values of η^2 is to use the values suggested by Cohen (1977) as “small” ($\eta^2 = .02$), “medium” ($\eta^2 = .13$), and “large” ($\eta^2 = .26$). Our power table does not have columns for these specific values of η^2 , but .03, .1, and .3 could be used instead. Although these suggested values for small, medium, and large effects are inherently arbitrary, they do represent experience across a wide range of social science disciplines. If you have sufficient experience in a research domain to consider these suggested values unreasonable, then simply use those values that are reasonable based upon your experience. The goal of a power analysis conducted before the collection of data is not an exact calculation of the statistical power but an indication of whether there is much hope for detecting the effect you want to find with the sample size you have planned. If a study would not have much chance of distinguishing between Model C and Model A for a large effect ($\eta^2 = .26$ or .3), then there is little if any reason for conducting the study.

As an example of this approach, let us estimate the power for detecting small, medium, and large effects for the lottery bids using our sample of 20 bidders. Using the row of the power table for $n = 20$ and the columns for $\eta^2 = .03$, .1, and .3, we find that the respective powers are .12, .29, and .79. In other words, we would not have much chance of detecting small and medium effects but a decent chance of detecting a large effect. If we wanted to be able to detect medium effects, then we would need to increase the number of participants in our study.

A third approach to finding an appropriate value of η^2 to use in “what if” power analyses involves guesses about the parameter values and variance. To have reasonable expectations about the parameter values and variance generally requires as much or more experience in a research domain as is necessary to know typical values of PRE. Thus, this third approach is generally less useful than the first two. We present this approach in order to be complete and because the derivation of this approach provides further useful insights about the meaning of PRE and η^2 . Also, this approach requires describing in detail the data that one expects to obtain, and such an exercise can often be useful for identifying flawed research designs.

We begin with our familiar definition of PRE:

$$\text{PRE} = \frac{\text{SSE}(C) - \text{SSE}(A)}{\text{SSE}(C)} = \frac{\text{SSR}}{\text{SSE}(C)}$$

We have noted before that $\text{SSE}(C) = \text{SSE}(A) + \text{SSR}$ (i.e., the error for the compact model includes all the error of the augmented model plus the error that was reduced by the addition of the extra parameters in the augmented model). Hence, substituting for $\text{SSE}(C)$ yields:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE}(A) + \text{SSR}} = \frac{1}{\text{SSE}(A)/\text{SSR} + 1}$$

To obtain a definition of the true proportional reduction in error η^2 , we simply estimate $\text{SSE}(A)$ and SSR , using not the data but the parameter values of B_0 and β_0 that are of interest.

For example, we noted above that:

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_{iC} - \hat{Y}_{iA})^2$$

If we thought that the effect of the lottery bids would be to decrease the mean bid from 50 to 40, then $\hat{Y}_{iC} = B_0 = 50$ represents the null hypothesis and $\hat{Y}_{iA} = \beta_0 = 40$ represents an alternative hypothesis that we want to test. For that situation we would *expect*:

$$\text{SSR} = \sum_{i=1}^{20} (50 - 40)^2 = \sum_{i=1}^{20} 10^2 = \sum_{i=1}^{20} 100 = 2000$$

In other words, the SSR that we expect is simply 100 added up 20 times (once for each bidder). We saw in Chapter 2 that $\text{SSE}/(n - \text{PA})$ was an estimate of the variance σ^2 . If we use our expected value of $\beta_0 = 40$ to calculate $\text{SSE}(A)$, then we are not using data to estimate any parameters, so $\text{PA} = 0$. Hence, $\text{SSE}(A)/n = \sigma^2$, so the value of $\text{SSE}(A)$ that we *expect* to obtain is:

$$\text{SSE}(A) = n\sigma^2$$

Thus, if we have a reasonable guess or expectation for the variance, then we can easily calculate the value of $\text{SSE}(A)$ that we would expect. Having good intuitions about what variance to expect is usually as difficult or more difficult than knowing what PRE to expect. Both are based on previous experience in a research domain. Good guesses for the variance often depend on previous experience with the particular measure for Y . For our example, suppose that past data for the lottery lead us to expect that σ^2 is about 400; then, we would expect:

$$\text{SSE}(A) = n\sigma^2 = (20)400 = 8000$$

We now can return to our formula for PRE to calculate the value that we expect for the true proportional reduction in error η^2 (given our guesses for B_0 , β_0 , and σ^2).

$$\text{Expected } \eta^2 = \frac{1}{\text{SSE}(A)/\text{SSR} + 1} = \frac{1}{8000/2000 + 1} = \frac{1}{4 + 1} = .2$$

In other words, $\eta^2 = .2$ corresponds to our guesses about B_0 , β_0 , and σ^2 . We now can use the power tables to find the power we would have for comparing Model C, which predicts that the mean bid will be 50, against Model A, which predicts that the mean bid will be 40, when the variance is about 400. Using Figure 4.15, we find that for 20 observations the probability of rejecting Model C (i.e., deciding that the lottery bids were less than the expected value of a ticket) is only about .56 even if we think that it will on average be \$10 less than the expected value of \$50. This means that the researcher has only a little more than a 50/50 chance of *deciding* that lottery bids are lower than the expected value of a ticket even when they really are lower by \$10. Using the power table, we can see that testing our hypothesis with twice the number of persons would increase the power substantially to .87. In this case, it would seem advisable to test our hypothesis with a larger sample size rather than with a sample size that offered little hope of finding the effect. We can similarly calculate the power for other “what if” values of B_0 , β_0 , and σ^2 that we might want to consider.

IMPROVING POWER

The relatively low power—the high probability of making a Type II error—for the apparently reasonable evaluation of the hypothesis we have been considering may be startling. Unfortunately, low power is a problem that plagues data analysis far more frequently than is commonly realized. Low power creates serious difficulties. For example, consider the plight of researchers trying to evaluate the effectiveness of an innovative educational curriculum or a new therapy for a serious illness. If the power of detecting the effect is only about 50/50, there is a fairly high risk of concluding that the new curriculum or therapy is not effective even when it actually is. When one considers the time and money that may be invested in research—not to mention the potential of the findings to benefit people and advance science—it generally makes little sense to design and conduct studies that have little chance of finding effects even when the effects really exist. In our particular example, the obtained value of PRE allowed us to reject the hypothesis that $\beta_0 = 50$; we were either lucky, or the true value of β_0 was considerably less than the alternative value of 50 that we considered above. In general, however, we want to increase the power of the statistical inference. There are three basic strategies for improving power: reduce error, increase α , and/or increase the number of observations. We consider each in turn.

Reducing Error

One way to reduce error is to control as many of the possible random perturbations as possible. In our lottery example, one might reduce error and obtain more power by providing clear instructions to participants, making sure participants were well rested, eliminating distractions in the bidding environment, and using a more reliable bidding procedure. In other words, error is reduced by obtaining data of better quality. In the equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

the model will account for a higher proportion of the data if the data are of higher quality and hence have less error. Less error allows us to obtain a more powerful look at our data. Although reducing error by such means may be the most effective method for improving power, the techniques for doing so are usually domain-specific and outside the scope of this book.

Another way to reduce error is to improve the quality of the model. Again in the equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

error will be smaller for data of fixed quality if the model can be improved to account for more of the data. How to use models more complex than the simple models we have been considering in these beginning chapters is the subject of the remainder of the book, so we cannot give too many details here. The general idea is to build models that make predictions conditional on additional information we have about the observations. In the lottery example we might know, say, which bidders had participated in lotteries before and which ones, if any, had ever won a lottery. If having participated in a lottery

before makes a difference in the amount an individual is willing to pay for a lottery ticket, then we can make different predictions conditional on whether (or perhaps how often) the bidder had participated in lotteries previously. By doing so we will have, in essence, removed what was formerly a random perturbation—previous participation in lotteries—from the error and included it in the model. Again the reduced error will give us a more powerful look at our data. In later chapters we explicitly consider the addition of parameters to the model for the purpose of improving power.

Increasing α

A different way to improve power is to increase α , the probability of a Type I error. The probabilities of Type I and II errors are linked in that if we choose a critical value that increases (decreases) α then we simultaneously and unavoidably decrease (increase) the probability of a Type II error. For our lottery data with $n = 20$, Figure 4.16 shows power as a function of η^2 and α . As α increases from .001 to .25 the critical values for F and PRE decrease. It obviously becomes easier for the values of F and PRE calculated from the data to beat these critical values, so the power increases as α increases. For example, if we do a “what if” power analysis with $\eta^2 = .2$, then the power at $\alpha = .05$ is .56, but if we increase α to .1, then power increases to .70.

Editors of scientific journals are wary of Type I errors and will seldom accept the use of $\alpha > .05$ for statistical inference. However, there are many practical data analysis problems when a higher α is justified to increase power. Characteristics of such data analyses are (a) that increasing power in any other way is infeasible, (b) that rejection of Model C if it is indeed false would have important practical consequences, and (c) that the costs associated with a Type I error are not great. Consider, for example, the statistical decision problem faced by a researcher testing the effectiveness of an innovative new curriculum. It may be difficult for the researcher to control any other sources of error to increase power. It would certainly be important to identify a curriculum that could improve student learning. The consequence of a Type I error would probably be further trial use of the curriculum in several classrooms the following year, which may not involve a significant cost. Thus, the researcher might well adopt a higher α to choose the critical values for PRE and F for her statistical inference.

Conversely, note that reducing α also reduces power. In the lottery example, lowering α to .01 would increase our protection against the possibility of a Type I error but would

FIGURE 4.16 Power for PA = 1, PC = 0, and $n = 20$ for various levels of α

α	Critical values		<i>Prob(PRE > critical value)</i>							
	F	PRE	True PRE, η^2							
			0	.01	.03	.05	.075	.1	.2	.3
.001	15.08	.443	.00	.00	.00	.01	.01	.02	.09	.22
.005	10.07	.346	.01	.01	.02	.03	.05	.07	.21	.43
.01	8.18	.301	.01	.02	.03	.05	.08	.11	.30	.54
.025	5.92	.238	.03	.04	.07	.10	.15	.20	.44	.69
.05	4.38	.187	.05	.07	.12	.16	.23	.29	.56	.79
.1	2.99	.136	.10	.13	.20	.26	.34	.42	.70	.88
.25	1.41	.069	.25	.29	.38	.46	.55	.63	.85	.96

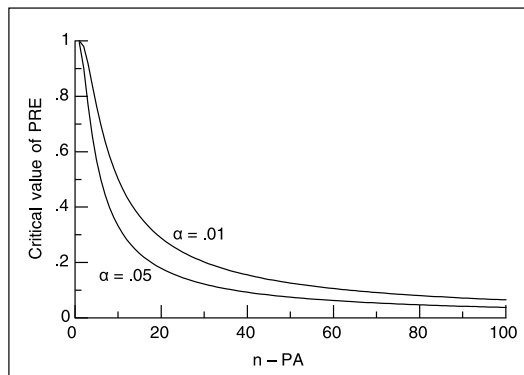
reduce power to less than .30, a level for which it would not likely be worth conducting the study. If the costs of a Type I error are very high—for example, if reliance on the research findings involved extensive teacher retraining in the case of an innovative new curriculum—then the use of restrictive values of α may be appropriate. However, in some cases, such as when the sample size is very small, as it is for our lottery study, reducing α would reduce the power so much that it would no longer be worthwhile to do the study—the chances of seeing anything are too low with such a low-powered “microscope.”

Increasing n

Probably the most common technique for increasing the power of statistical tests is to increase the number of observations. Figure 4.17 is a graphical display of the 95% ($\alpha = .05$) and 99% ($\alpha = .01$) critical values of Figure 4.9 as a function of the number of observations. The value of PRE required to reject Model C drops dramatically as the number of observations increases. For example, had there been 51 participants instead of 20 in our lottery study, a PRE of only .075 would have been required ($\alpha = .05$) to reject the Model C that assumed $B_0 = 50$ instead of the PRE of .187 required in our example. The drop in the critical values for F and PRE needed to reject Model C corresponds to an increase in power, as we have noted several times. For example, for $\eta^2 = .2$ as a “what if” value of the true proportional reduction in error, the power for 20 observations is .54 (see Figure 4.15), but with 51 observations the power increases to .93.

There are two reasons for not routinely using a large number of observations. First, it may be infeasible, due to cost or other data collection constraints, to obtain more observations. Second, power might be so high that some statistically significant results may be misleading. By turning up the power on our metaphorical statistical microscope to extraordinary levels, we might detect flaws in Model C that are statistically reliable but are trivial substantively. For example, with 120 observations any PRE greater than .032 is cause to reject Model C in favor of Model A. However, the 3.2% reduction in error means that Model A may be a trivial improvement over Model C. For this reason, one should always report not just the statistical inference about whether Model C is or is not rejected but also the obtained values for PRE and F so that the reader can evaluate

FIGURE 4.17 Critical values of PRE as a function of the number of observations



the magnitude by which Model A improves on Model C. All else being equal, more statistical power is always better; we just need to be careful in interpreting the results as substantively important when power is very high.

It is also important not to dismiss a small but reliable PRE just because it is small. Rejecting Model C in favor of Model A may be important theoretically even if Model A provides only a slight improvement. Whether a given PRE is substantively interesting will depend on theory and prior research experience in the particular domain.

CONFIDENCE INTERVALS

Confidence intervals provide an alternative way for considering statistical inference. Although, as we shall see later, confidence intervals are exactly equivalent to statistical inference as described above, they reorganize the information in a way that can give useful insights about the data and our model.

A *confidence interval* simply consists of all those possible values of a parameter that, when used as a hypothesis for Model C, would not cause us to reject Model C. For example, the mean for the 20 lottery bids is 42.5 and estimates the parameter β_0 in Model A. We have already determined that 50 is not in the confidence interval for β_0 because, when we used 50 as the prediction for Model C, we obtained an unlikely PRE value that would occur less than 5% of the time. Any value greater than 50 would produce an even larger PRE, so none of those values is included in the confidence interval. Conceptually, we could find the boundary of the confidence interval by trying increasingly higher values (i.e., higher than b_0) for B_0 in Model C until we found a value for B_0 that produced a PRE that would not cause us to reject Model C. For example, if we tried $B_0 = 45$, the respective sums of squared errors for Models C and A would be 4392 and 4267, yielding $\text{PRE} = .028$, which is below the critical value for $\alpha = .05$; hence 45 is in the confidence interval for β_0 , and the boundary must be somewhere between 45 and 50. We would also need to search for the lower boundary below the estimated value of 42.5. We can avoid this iterative search because it can be shown that the boundaries are given by:

$$b_0 \pm \sqrt{\frac{F_{\text{crit};1, n-1; \alpha} \text{MSE}}{n}} \quad (4.4)$$

For the simple model of the lottery bids, $\text{MSE} = s^2 = \text{SSE}/(n-1) = 4267/19 = 224.58$. $F_{\text{crit};1; n-1; \alpha}$ is the critical value at level α for F with 1 degree of freedom for the numerator (i.e., the difference in the number of parameters between the two models) and $n-1$ degrees of freedom for the denominator (i.e., the number of observations minus the number of parameters in Model A). For $\alpha = .05$, the critical value of $F_{1,19;.05} = 4.38$. For these data, $b_0 = \text{the mean} = 42.5$, so the boundaries of the confidence interval are given by:

$$42.5 \pm \sqrt{\frac{(4.38)224.58}{20}} \quad \text{or} \quad 42.5 \pm 7.01$$

Thus, the lower boundary is $42.5 - 7.01 = 35.49$ and the upper boundary is $42.5 + 7.01 = 49.51$. We are therefore 95% ($1 - \alpha$) confident that the true value for β_0 is in the

interval [35.49, 49.51]. Any B_0 not in this interval that we might try for Model C would, with the present data, produce values of PRE and F that would lead us to reject Model C. Any B_0 in this interval would produce values of PRE and F below the critical values, so we would not reject Model C.

When estimating statistical power we described how to do an a priori power analysis. Those “what if” power analyses are necessarily inexact. Confidence intervals are useful for describing post hoc the actual statistical power achieved in terms of the precision of the parameter estimates. Wide confidence intervals represent low statistical power and narrow intervals represent high statistical power.

EQUIVALENCE TO THE t -TEST

In this optional section we demonstrate the equivalence between the statistical test for the simple model developed in this chapter and the traditional *one-sample t-test* presented in most statistics textbooks. We do so to allow readers with exposure to traditional textbooks to make the comparison between approaches.

The one-sample t -test answers whether the mean of a set of observations equals a particular value specified by the null hypothesis. The formula for the one-sample t -test is:

$$t_{n-1} = \frac{\sqrt{n}(\bar{Y} - B_0)}{s}$$

where n is the number of observations, \bar{Y} is the calculated mean, B_0 is the value specified by the null hypothesis, and s is the standard deviation of the set of observations. With the appropriate assumptions about Y —the same assumptions that we made about the distribution and independence of the ε_i —calculated values of t can be compared to critical values of *Student’s t-distribution*. Tables of this distribution are available in many statistics textbooks. However, separate tables are not really needed because squaring t with $n - 1$ degrees of freedom yields an F with 1 and $n - 1$ degrees of freedom. Thus, the F tables in the Appendix may readily be used.

To show the equivalence between F as presented in this chapter and the usual t -test, we begin with the definition of F for the simple model; that is:

$$F_{1,n-1} = \frac{\text{PRE}/1}{(1 - \text{PRE})/(n - 1)}$$

We know that $\text{PRE} = \text{SSR}/\text{SSE}(C)$, and it is easy to show that:

$$1 - \text{PRE} = 1 - \frac{\text{SSR}}{\text{SSE}(C)} = \frac{\text{SSE}(C) - \text{SSR}}{\text{SSE}(C)} = \frac{\text{SSE}(A)}{\text{SSE}(C)}$$

Substituting these values into the definition for F yields:

$$F_{1,n-1} = \frac{\text{SSR}/\text{SSE}(C)}{[\text{SSE}(A)/\text{SSE}(C)]/(n - 1)} = \frac{\text{SSR}}{\text{SSE}(A)/(n - 1)}$$

But from Equation 4.2 we know that for the simple model SSR can be replaced with $n(B_0 - \bar{Y})^2$, and from Chapter 2 we know that $SSE(A)/(n - 1)$ is s^2 , the variance of the set of observations. Substituting these values yields:

$$F_{1,n-1} = \frac{n(B_0 - \bar{Y})^2}{s^2}$$

Taking the square root of this last equation gives the final result of:

$$\sqrt{F_{1,n-1}} = t_{n-1} = \frac{\sqrt{n}(\bar{Y} - B_0)}{s}$$

We provide the above derivation not to present yet another computational formula but to show that our model comparison approach to statistical inference is statistically identical to the traditional approach. The use of PRE and F for comparing models is nothing more than a repackaging of the traditional approach. This repackaging has the important consequence of making it easy to generalize to more complicated models and data analysis questions. We will use PRE and F just as we did in this chapter for statistical inference throughout the remainder of the book. In contrast, the traditional t -test does not generalize nearly so easily. Also, even though the t -test must produce exactly the same conclusion with respect to the null hypothesis, it does not automatically provide a measure of the magnitude of the result. In our model comparison approach, PRE automatically provides a useful measure of the magnitude.

AN EXAMPLE

In this section we illustrate the techniques of this chapter using the internet access data that were presented in Figure 1.1. Suppose that marketing researchers had projected that 75% of households would have internet access by the year 2013. Was the marketing researchers' projection overly optimistic? The question is equivalent to comparing the following two models:

$$\text{MODEL A: } Y_i = \beta_0 + \varepsilon_i$$

$$\text{MODEL C: } Y_i = 75 + \varepsilon_i$$

We know that the mean is $\bar{Y} = 72.806$ (see Chapter 2), so the estimated Model A is:

$$\hat{Y}_i = 72.806$$

For the statistical inference we need to calculate PRE and F from $SSE(A)$ and $SSE(C)$. We also know that the variance or the MSE is 27.654 (see Chapter 2). Since $MSE = SSE(A)/(n - 1)$, we can easily obtain $SSE(A)$ by multiplying MSE by $n - 1$; thus, $SSE(A) = 27.654(49) = 1355.046$; this value (within rounding error) is also given in Figure 2.10. We can compute SSR using:

$$SSR = n(B_0 - \bar{Y})^2 = 50(75 - 72.806)^2 = 240.682$$

Then it is easy to get $SSE(C)$ from:

$$SSE(C) = SSE(A) + SSR = 1355.046 + 240.682 = 1595.728$$

The computations of PRE and F are then easy:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE}(C)} = \frac{240.682}{1595.728} = .151$$

and

$$F_{1,49} = \frac{\text{PRE}/1}{(1 - \text{PRE})/(n - 1)} = \frac{.151}{.849/49} = 8.71$$

From the tables in the Appendix, the critical values for PRE and F (for $\alpha = .05$) are, respectively, about .075 and 4.03. The obtained values clearly exceed the critical values, so we can reject Model C in favor of Model A. Thus, the 15.1% reduction in error obtained by using the estimate $b_0 = 72.806$ instead of the null hypothesis value of $B_0 = 75$ is statistically significant. We can therefore conclude that the percentage of households that had internet access was significantly lower than the marketing researchers' projection. We might summarize our results for a journal article as follows:

On average, across states the percentage of households that had internet access in the year 2013 ($M = 72.806$) was significantly lower than the projected value of 75%, $\text{PRE} = .151$, $F(1, 49) = 8.71$, $p < .05$.

From the above it is also easy to calculate the 95% confidence interval for β_0 , the true average percentage of households that had internet access across states. Substituting the appropriate values into Equation 4.4 yields:

$$72.806 \pm \sqrt{\frac{4.03(27.654)}{50}} \quad \text{or} \quad 72.806 \pm 1.493$$

which gives an interval of [71.313, 74.299]. Using this interval, we can easily ask other questions. For example, had the marketing researchers projected that $B_0 = 73$, we would not conclude that the actual percentage of households with internet access was significantly less than the projection, because 73 is included in the 95% confidence interval.

SUMMARY

In Chapter 1 we noted that the equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

implies an inherent tension in data analysis between reducing error as much as possible and keeping the model as simple or parsimonious as possible. Whenever we consider adding an additional parameter to the model so that it will fit the data better and thereby reduce error, we must ask whether the additional complexity of the model is worth it. In this chapter we have developed inferential machinery for answering whether the additional complexity is worth it.

To decide whether the benefits of the additional parameters in Model A outweigh the benefits of the parsimony and simplicity of Model C, we first calculate $\text{SSE}(A)$ and

SSE(C), respectively, the sum of squared errors for the augmented model (which incorporates the additional parameters) and the compact model (which does not include those parameters). The sum of squares reduced, SSR, is simply the difference between them:

$$\text{SSR} = \text{SSE(C)} - \text{SSE(A)}$$

Then we calculate the proportional reduction in error attributable to the additional parameters, which is given by:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{\text{SSR}}{\text{SSE(C)}}$$

Another related statistic is the ratio of the proportional reduction in error per parameter added to the potential proportional reduction in error per remaining unused parameter, which is given by:

$$F = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})}$$

We then compare the calculated values of PRE and F to the distribution of values we would expect *if* Model C, the compact model, *were true*. If the calculated values of PRE and F would have been unlikely if Model C were true, then we reject Model C and conclude that the extra complexity of Model A is worth it. On the other hand, if the calculated values are ones that might reasonably have been obtained if Model C were true, then we do not reject Model C and without further evidence we would not accept the additional complexity of Model A.

This inferential machinery is merely a guide for decision making and is not infallible. There are two kinds of errors that we can make. A Type I error occurs when Model C is in fact correct but by chance we happen to get unusual values of PRE and F and so reject Model C. The probability of a Type I error is α and defines how unusual PRE and F have to be before we reject Model C. A Type II error occurs when Model C is in fact false or inferior to Model A but by chance we happen to get values of PRE and F that are not unusual and so fail to reject Model C. We generally select α , the probability of a Type I error, and try to reduce the chances of a Type II error by collecting better data with less error and by increasing the number of observations. Reducing the chances of a Type II error is referred to as increasing the statistical power of an inference.

We developed this inferential machinery in the context of asking a question for the simple model. However, *exactly* the same procedure will work for all the more complex models we consider in subsequent chapters. In this chapter, we have learned all we need to know as data analysts about statistical inference. The remainder of our task is to learn how to build more complex and interesting models of our data.

5

Simple Regression

Estimating Models with a Single Continuous Predictor

We have used the simple single-parameter model to illustrate the use of models, the notion of error, and inference procedures to be used in comparing augmented and compact models. We have focused on this single-parameter model in so much detail because the estimation and inference procedures that we developed within this very simple context generalize to much more complicated models. That is, regardless of the complexity of a model, estimation from here on will be done by minimizing the sum of squared errors, just as we did in the single-parameter case, and inference will be done by comparing augmented and compact models using PRE and F . So the detail on single-parameter models has been necessitated by our desire to present in a simple context all of the statistical tools that we will use in much more complex situations.

However, as we noted, single-parameter models are only infrequently of substantive or theoretical interest. In many ways, the example from the last chapter, where we wanted to test the hypothesis that people were willing to pay the expected value of \$50, is unusual in the behavioral sciences. More frequently such a priori values do not exist, and instead we may be asking whether the mean in one group of respondents (e.g., those who were trained in the meaning of expected values) differs from the mean in another group of respondents (e.g., those who received no such training). Or, returning to the data on internet access, while it is certainly possible that we would be interested in testing whether some a priori percentage (e.g., 75%) is a good estimate of mean internet access, it is much more likely that we would be interested in examining the determinants or correlates of internet access rates. In other words, our interest is more likely to center on attempts to explain the internet access data than on tests of alternative values for the mean access rate.

To examine these types of substantive issues, we need to consider models having more than a single parameter. Initially, we will consider only two-parameter models, taking the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The exact definition of the terms in this two-parameter model will be detailed below. For the present we simply note that we are making predictions of Y_i conditional upon some other variable X_i , since the model's predictions from this two-parameter model change as X_i changes, assuming that β_1 takes on some value other than zero.

Actually, there are two variations on this two-parameter model, each of which is illustrated by one of the two examples we have just discussed. In the first example, concerning whether training in the meaning of expected values influences how much one is willing to pay for a lottery ticket, we want to examine whether those who receive

such training are willing to pay more or less than those who do not. This amounts to asking whether we need different predictions from the model for the two groups of respondents or whether a single prediction suffices regardless of whether training was received or not. In other words, we want to compare a model in which predictions are made conditional on knowing whether or not a given respondent received the training with one where the identical prediction is made for every respondent. For this comparison, the augmented model is a two-parameter model, defining X_i in such a way that it identifies whether or not the respondent received training. We might, for instance, define X_i as follows:

$$X_i = -1, \quad \text{if a respondent did not receive training}$$

$$X_i = +1, \quad \text{if a respondent did receive training}$$

If the estimated value of β_1 in the above two-parameter model is something other than zero, the model then gives different predicted values for participants in the two groups. For example, if b_0 (the estimated value of β_0) equals 46 and b_1 (the estimated value of β_1) equals 3, then the prediction for the respondents without training is:

$$\hat{Y} = b_0 + b_1 X_i = 46 + 3(-1) = 46 - 3 = 43$$

and the prediction for the respondents who receive training is:

$$\hat{Y} = b_0 + b_1 X_i = 46 + 3(1) = 46 + 3 = 49$$

Notice that there are only two possible values for X_i in this example, and hence only two predicted values. Respondents either receive training or not, and our decision about the numerical values used to represent this training is arbitrary. For instance, had we defined X_i differently, giving respondents with training a value of 2 on X_i and those without training a value of 4, the two-parameter model would still generate different predictions for the two groups of students, assuming the estimated value of β_1 does not equal zero.

Now consider the second example. Suppose we wanted to explain variation in the internet access data and we suspected that average educational level of residents in the US states, measured as the percentage of residents with a college degree, might be a reasonable candidate for an explanatory factor. In other words, we thought that internet access rates might be higher in states where more people had graduated from college. So we might use the two-parameter model to make predictions for states conditional on college graduation rates, defining this as X_i . This variable has many possible values, and it would be unlikely that any two states would have exactly the same graduation rate and, hence, the exact same values on X_i . Therefore, instead of making two different predictions as in the lottery example, our two-parameter model now is likely to make different predictions for each state, since each state is likely to have a unique value of X_i . Another difference between this two-parameter model and the lottery example is that here values on X_i are less arbitrary than they were in the lottery example. Each state has an actual college graduation rate that can be compared with other states and the information about such state-to-state differences needs to be represented in the values we assign to X_i for each state.

The difference between these two examples lies in the nature of the units of measurement of the predictor variable, the variable upon which the predictions are

conditional. In the lottery example, no training versus training is a categorical variable, in the sense that all participants are in one group or the other. This means that while X_i needs to code the distinction between the two training conditions, it does not matter which group of respondents is given the higher value on X_i nor does it matter which two values are used. On the other hand, college graduation rate is what we call a continuous predictor variable, in the sense that different states have different values and the differences among these values are meaningful.

While all of the procedures for building models, estimating parameters, and comparing augmented and compact models can be used regardless of whether the predictor variable (or variables) is categorical or continuous, it is conceptually useful to treat models with categorical predictor variables separately from models whose predictors are assumed to be continuously measured. Initially, we will consider models that contain only predictors presumed to be continuously measured.

In the current chapter, we treat two-parameter models having a single, continuously measured predictor variable. Then, in Chapters 6 and 7, we consider models with multiple continuously measured predictors. In traditional statistical jargon these three chapters (Chapters 5, 6, and 7) deal with simple and multiple regression, including polynomial and moderated regression models. Then in Chapters 8 and 9 we turn our attention to models having categorical predictors. Again, in traditional statistical jargon, these chapters deal with analysis of variance. Finally, in Chapter 10 we consider models in which some predictors are categorical variables and some are continuous variables. Such models are traditionally referred to as analysis of covariance models. Our approach, however, to each type of model, regardless of the chapter, will be uniform. Rather than describing seemingly different statistical techniques for multiple regression, analysis of variance, and analysis of covariance, we will estimate parameters just as we have done in the simple single-parameter case, and we will test hypotheses by comparing augmented and compact models. So, while our treatment of categorical predictors is located in different chapters from our treatment of continuous predictors, the same procedures will be used throughout.

DEFINING A LINEAR TWO-PARAMETER MODEL

We now confine our attention to two-parameter models with a single continuous predictor variable. As an example, we will use the data contained in Figure 5.1 to ask whether differences between US states in their internet access rates are predictable from differences in their college graduation rates. As we speculated above, it seems reasonable that internet access may be higher in states where the population is relatively better educated.

Figure 5.2 is a scatterplot of the tabular data from Figure 5.1. The vertical axis represents internet access rates, and the horizontal axis represents college graduation rates. Each point in this plot represents one of the 50 US states. The question that we would like to ask is whether we can use graduation rates to predict internet access. Or, expressed differently, do our predictions of internet access improve by making those predictions conditional on knowledge of graduation rates?

We will use a simple linear model to generate such conditional predictions. As already discussed, this model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is a state's internet access rate and X_i is its college graduation rate. Returning to our generic equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

we see that the model in this two-parameter equation is represented by

$$\beta_0 + \beta_1 X_i$$

In terms of estimated parameter values, the predictions made by this model for each state's access rate are given by:

$$\hat{Y}_i = b_0 + b_1 X_i$$

Because ERROR equals DATA minus MODEL, the residuals in this two-parameter model can be expressed as follows:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (b_0 + b_1 X_i) \end{aligned}$$

Let us examine each of the parameter estimates in this model and see what each is telling us. First, consider b_0 . In the single-parameter model, we saw that b_0 equaled the mean value of Y_i , assuming that we define error as the sum of squared errors. Another way of saying the same thing is that in the single-parameter model b_0 is our predicted value for each state. However, in this two-parameter model we wish to take further information into account in making each state's prediction. We are making each state's prediction conditional on its college graduation rate. Therefore, b_0 is not the predicted value for each state, because the predictions vary as a function of graduation rates:

$$\hat{Y}_i = b_0 + b_1 X_i$$

There is one case, however, when this model predicts an internet access rate equal to b_0 . This is clearly when X_i equals zero, for then:

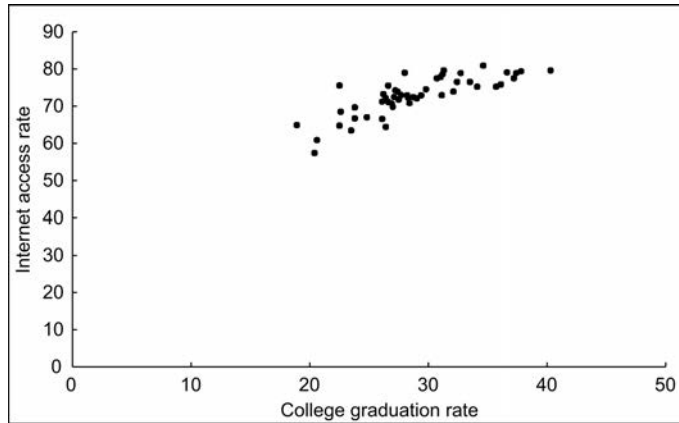
$$\hat{Y}_i = b_0 + b_1(0) = b_0$$

This, then, provides the interpretation for the parameter estimate b_0 in this two-parameter model: b_0 is our prediction of Y_i when X_i equals zero. As we will see for our example, this prediction may not be very useful because the data from which we estimate this parameter may not include data points having values of X_i near zero.

FIGURE 5.1 Internet access rates and college graduation rates

<i>US state</i>	<i>Internet access rate</i>	<i>College graduation rate</i>
AK	79.0	28.0
AL	63.5	23.5
AR	60.9	20.6
AZ	73.9	27.4
CA	77.9	31.0
CO	79.4	37.8
CT	77.5	37.2
DE	74.5	29.8
FL	74.3	27.2
GA	72.2	28.3
HI	78.6	31.2
IA	72.2	26.4
ID	73.2	26.2
IL	74.0	32.1
IN	69.7	23.8
KS	73.0	31.1
KY	68.5	22.6
LA	64.8	22.5
MA	79.6	40.3
MD	78.9	37.4
ME	72.9	28.2
MI	70.7	26.9
MN	76.5	33.5
MO	69.8	27.0
MS	57.4	20.4
MT	72.1	29.0
NC	70.8	28.4
ND	72.5	27.1
NE	72.9	29.4
NH	80.9	34.6
NJ	79.1	36.6
NM	64.4	26.4
NV	75.6	22.5
NY	75.3	34.1
OH	71.2	26.1
OK	66.7	23.8
OR	77.5	30.7
PA	72.4	28.7
RI	76.5	32.4
SC	66.6	26.1
SD	71.1	26.6
TN	67.0	24.8
TX	71.8	27.5
UT	79.6	31.3
VA	75.8	36.1
VT	75.3	35.7
WA	78.9	32.7
WI	73.0	27.7
WV	64.9	18.9
WY	75.5	26.6

FIGURE 5.2 Scatterplot of internet access rates and college graduation rates for each of the 50 US states



The second parameter estimate in the model, b_1 , tells us how our predictions change as X_i changes. Suppose we had two observations differing in their values on X_i by one unit, with X_i for the first observation being one unit larger than X_i for the second. According to the model, our predictions for the two data points would differ by b_1 since:

$$\begin{aligned}\hat{Y}_1 - \hat{Y}_2 &= (b_0 + b_1X_1) - (b_0 + b_1X_2) \\ &= b_1X_1 - b_1X_2 \\ &= b_1(X_1 - X_2) \\ &= b_1\end{aligned}$$

So, b_1 tells us by how much our predictions of Y_i change as X_i increases by one unit. Notice that in this derivation we did not specify what the actual values of X_1 and X_2 were. We only specified that they were one unit apart from each other. Hence, this implies that b_1 in this two-parameter model is constant, regardless of the level of X_i . This is what was meant by the definition of this sort of two-parameter model as a linear model. As X_i changes by some set amount, our predictions of Y_i change by a constant amount, regardless of the value of X_i .

To review, b_0 and b_1 tell us very different things: b_0 is a predicted value (a \hat{Y}_i) at a particular value of X_i , namely when X_i equals zero; b_1 is not a predicted value, rather it is the difference between two predicted values as we move from a smaller X_i to one that is one unit larger.

Let us look at this two-parameter model graphically for the example in which internet access is predicted from college graduation rates. Figure 5.3 presents the graph of the model set against the data. All of the predictions \hat{Y}_i lie on the line defined by the model. Errors of prediction, e_i , as in the single-parameter model, are defined as vertical distances between the line and an actual observation. That is, an error or residual is the difference between Y_i and \hat{Y}_i . b_0 is the value of \hat{Y}_i when X_i equals zero; it is frequently called the intercept because it is the value on the vertical axis of the graph where the prediction function crosses or “intercepts” it. b_1 is the difference in \hat{Y}_i for each unit increase in X_i . We can think of it as the slope of the line, since algebraically it is the difference in predicted values between any two points on the line per their difference in X_i values: *rise over run*:

$$b_1 = \frac{(\hat{Y}_2 - \hat{Y}_1)}{(X_2 - X_1)}$$

where the subscripts designate any two points on the line. Notice that the slope can take on any positive or negative value. If the slope is positive, it means that the model predicts higher values of Y_i as X_i increases. If the slope is negative, the model predicts lower values of Y_i as X_i increases.

ESTIMATING A LINEAR TWO-PARAMETER MODEL

Given some sample of data, how do we estimate the parameters of this sort of model? We want to use our sample of data to generate values of b_0 and b_1 in the equation:

$$\hat{Y}_i = b_0 + b_1 X_i$$

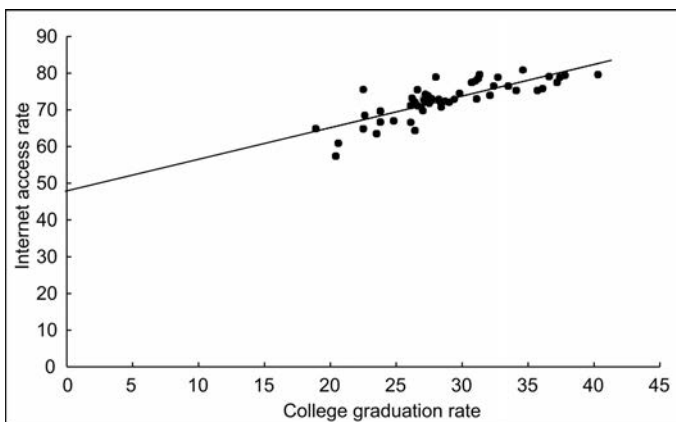
that are good estimates of the true (but unknown) parameters β_0 and β_1 . To do this, we decided in Chapter 3 that for the single-parameter model we would derive estimates that minimize the sum of squared errors. This preference was due to the fact that if the errors are normally distributed, least-squares parameter estimates are unbiased, consistent, and relatively efficient. This continues to be the case in the context of the present two-parameter model and will continue to be the case in more complicated models with many parameters, which we consider in subsequent chapters. For now, we want to derive estimated values of β_0 and β_1 that minimize $\Sigma(Y_i - \hat{Y}_i)^2$. The resulting *least-squares* parameter estimates are given as:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

The derivation of these estimates is given in Box 5.1.

FIGURE 5.3 Scatterplot with two-parameter model predicting internet access from college graduation rates



Box 5.1 Algebraic Derivation of Least-Squares Estimates of β_0 and β_1

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2$$

given that $\hat{Y}_i = b_0 + b_1 X_i$. We now add $(\bar{Y} - \bar{Y})$ and $b_1(\bar{X} - \bar{X})$ inside the parentheses to this expression for SSE. Since both of these expressions equal zero, we have not changed the equality. Thus:

$$\text{SSE} = \sum (Y_i - \bar{Y} + \bar{Y} - b_0 - b_1 X_i + b_1 \bar{X} - b_1 \bar{X})^2$$

Grouping terms yields the equivalent expression:

$$\text{SSE} = \sum [(Y_i - \bar{Y}) + (\bar{Y} - b_0 - b_1 \bar{X}) - b_1(X_i - \bar{X})]^2$$

If we square the term in brackets and distribute the summation sign, this gives the equivalent expression:

$$\begin{aligned} \text{SSE} &= \sum (Y_i - \bar{Y})^2 + 2(\bar{Y} - b_0 - b_1 \bar{X}) \sum (Y_i - \bar{Y}) \\ &\quad - 2b_1 \sum (Y_i - \bar{Y})(X_i - \bar{X}) + n(\bar{Y} - b_0 - b_1 \bar{X})^2 \\ &\quad - 2b_1(\bar{Y} - b_0 - b_1 \bar{X}) \sum (X_i - \bar{X}) + b_1^2 \sum (X_i - \bar{X})^2 \end{aligned}$$

Since both $\sum (Y_i - \bar{Y})$ and $\sum (X_i - \bar{X})$ equal zero, this expression reduces to:

$$\begin{aligned} \text{SSE} &= \sum (Y_i - \bar{Y})^2 - 2b_1 \sum (Y_i - \bar{Y})(X_i - \bar{X}) + n(\bar{Y} - b_0 - b_1 \bar{X})^2 \\ &\quad + b_1^2 \sum (X_i - \bar{X})^2 \end{aligned}$$

Since the third term in this expression, $n(\bar{Y} - b_0 - b_1 \bar{X})^2$, is necessarily positive, to minimize SSE we would like to set it equal to zero. Therefore, we wish values of b_0 and b_1 such that:

$$n(\bar{Y} - b_0 - b_1 \bar{X}) = 0$$

Dividing both sides of this equality by n gives us:

$$\bar{Y} - b_0 - b_1 \bar{X} = 0$$

or, equivalently:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

We have now reduced our expression for SSE, assuming the desire to minimize it, to:

$$\begin{aligned} \text{SSE} &= \sum (Y_i - \bar{Y})^2 - 2b_1 \sum (Y_i - \bar{Y})(X_i - \bar{X}) + b_1^2 \sum (X_i - \bar{X})^2 \\ &= \sum (Y_i - \bar{Y})^2 + \sum (X_i - \bar{X})^2 \left[b_1^2 - 2b_1 \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right] \end{aligned}$$

Let us now add to and subtract from this expression the quantity:

$$\sum (X_i - \bar{X})^2 \left(\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \right)^2$$

Thus:

$$\begin{aligned} \text{SSE} &= \Sigma(Y_i - \bar{Y})^2 \\ &+ \Sigma(X_i - \bar{X})^2 \left[b_1^2 - 2b_1 \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} + \left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \right)^2 \right] \\ &- \Sigma(X_i - \bar{X})^2 \left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \right)^2 \end{aligned}$$

Rearranging terms and taking the square root of the term in brackets gives us:

$$\begin{aligned} \text{SSE} &= \Sigma(Y_i - \bar{Y})^2 - \Sigma(X_i - \bar{X})^2 \left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \right)^2 \\ &+ \Sigma(X_i - \bar{X})^2 \left[b_1 - \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \right]^2 \end{aligned}$$

The last term in this expression for SSE is necessarily positive. Therefore, to minimize SSE we want this last term to equal zero. This occurs if:

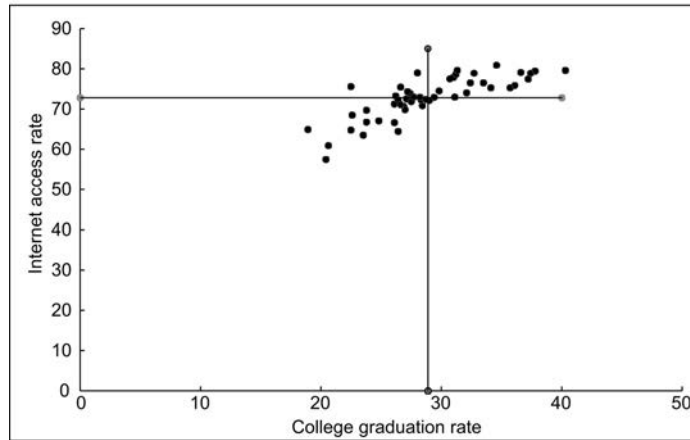
$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

We can think about the formula for the estimated slope in a couple of different ways. One way is to divide both the numerator and denominator of the formula by $n - 1$:

$$\begin{aligned} b_1 &= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/(n - 1)}{\Sigma(X_i - \bar{X})^2/(n - 1)} \\ &= \frac{s_{XY}}{s_X^2} \end{aligned}$$

In this last expression, the numerator of the slope, s_{XY} , is known as the covariance of X and Y , and the denominator is, of course, the variance of X .

Examining the crossproduct of $(X_i - \bar{X})(Y_i - \bar{Y})$ for any given observation helps to conceptualize the meaning of the covariance between two variables. To aid with this conceptualization, in Figure 5.4 we have added a horizontal line at the mean of Y and a vertical line at the mean of X to the scatterplot of the data. Any given observation will have a positive value for its crossproduct if it lies either in the upper right quadrant of the scatterplot or the lower left quadrant of the scatterplot, where the quadrants are defined by the intersecting lines at the two means. Positive values of the crossproduct thus come from observations that are either above the mean on both variables or below the mean on both. On the other hand, observations with negative values for their crossproducts will lie in the other two quadrants, having values that are below the mean on one variable but above the mean on the other. The covariance is (roughly) the average of all these

FIGURE 5.4 Scatterplot with horizontal line at \bar{Y} and a vertical line at \bar{X} 

individual crossproducts. Given that the denominator of the slope is always positive in value, the sign of the slope is determined by the sign of the covariance. This means that the slope will be positive if most of the observations in a scatterplot are either above the means on both variables or below the means on both variables. A negative slope happens when most of the observations fall into the other two quadrants: above the mean on one variable but below it on the other. A slope near zero would occur when the observations are randomly distributed throughout all four quadrants.

The other way to think conceptually about the meaning of the slope involves applying a bit of algebra to the original formula for the slope given above, yielding:

$$b_1 = \sum w_i \left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \right], \text{ where } w_i = \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

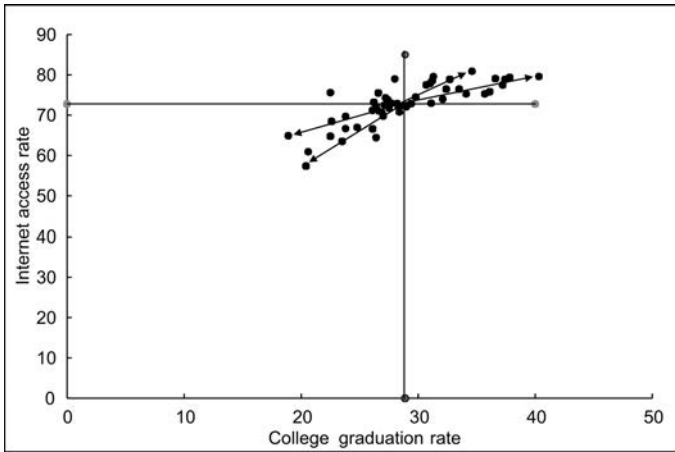
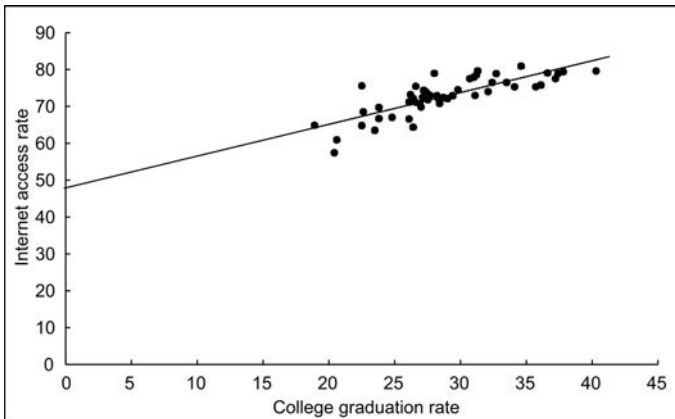
The term in the brackets in this equation can be thought of as the slope suggested by each individual observation—it is the rise over run of a line that goes between the joint mean of the two variables and a particular observation. In Figure 5.5 we have added a few of these individual “slopes” for a few observations. The w_i can be thought of as a weight assigned to each observation, where the weight represents the proportion of the total sum of squares of X that is attributable to the particular observation. In essence, we can think of each observation as having a slope that it “prefers” (between the joint mean and itself), that gets a certain weight or vote in determining the value of the slope for all the observations. The more extreme the observation is on X , the greater the vote.

Using the above formulas to calculate both parameter estimates, b_0 and b_1 , or more efficiently using a regression routine in one of the various statistical software packages, we can calculate the estimated intercept and slope for the model where we regress internet access rates on college graduation rates. (Notice here that in regression terminology one regresses Y on X , not the other way around.) The resulting estimated two-parameter model for these data is:

$$\hat{Y}_i = 47.912 + 0.862X_i$$

This prediction function is graphed in Figure 5.6 as a straight line on the scatterplot we saw before.

FIGURE 5.5 Scatterplot with individual slopes

FIGURE 5.6 Scatterplot with two-parameter model: $\hat{Y}_i = 47.912 + 0.862X_i$ 

(This figure is identical to Figure 5.3 where we plotted the slope and intercept prior to indicating how they were estimated.)

Let us interpret each of the regression coefficients, b_0 and b_1 , in this equation. The first equals 47.912. This is the value predicted by the model for a state's internet access rate if none of the population in the state had graduated from college. While this intercept is the best unbiased estimate of this prediction based on a linear model of these data, it is clearly a relatively meaningless value, because no state in the data had a college graduation rate anywhere near zero.

The value of the slope, 0.862, tells us that if we found two states differing in college graduation rates by 1%, our model predicts that the internet access rate would be .862% higher in the better educated state.

In Figure 5.7 we present for each state its Y_i , X_i , \hat{Y}_i , residual, and squared residual. The sum of these squared residuals, $\sum(Y_i - \hat{Y}_i)^2$, across all 50 states is also given. Having used the least-squares criterion guarantees that no other values of b_0 and b_1 would give us a smaller sum of squared residuals for these data.

FIGURE 5.7 Predicted values and residuals for internet access data by state

<i>US state</i>	Y_i	X_i	\hat{Y}_i	e_i	e_i^2
AL	63.5	23.5	68.169	-4.669	21.780
AK	79.0	28.0	72.048	6.952	48.330
AZ	73.9	27.4	71.531	2.369	5.612
AR	60.9	20.6	65.669	-4.769	22.743
CA	77.9	31.0	74.634	3.266	10.667
CO	79.4	37.8	80.496	-1.096	1.201
CT	77.5	37.2	79.978	-2.478	6.140
DE	74.5	29.8	73.600	0.900	0.810
FL	74.3	27.2	71.358	2.942	8.655
GA	72.2	28.3	72.307	-0.107	0.011
HI	78.6	31.2	74.806	3.794	14.394
ID	73.2	26.2	70.496	2.704	7.312
IL	74.0	32.1	75.582	-1.582	2.503
IN	69.7	23.8	68.428	1.272	1.618
IA	72.2	26.4	70.669	1.531	2.344
KS	73.0	31.1	74.720	-1.720	2.958
KY	68.5	22.6	67.393	1.107	1.225
LA	64.8	22.5	67.307	-2.507	6.285
ME	72.9	28.2	72.220	0.680	0.462
MD	78.9	37.4	80.151	-1.251	1.565
MA	79.6	40.3	82.651	-3.051	9.309
MI	70.7	26.9	71.100	-0.400	0.160
MN	76.5	33.5	76.789	-0.289	0.084
MS	57.4	20.4	65.497	-8.097	65.561
MO	69.8	27.0	71.186	-1.386	1.921
MT	72.1	29.0	72.910	-0.810	0.656
NE	72.9	29.4	73.255	-0.355	0.126
NV	75.6	22.5	67.307	8.293	68.774
NH	80.9	34.6	77.737	3.163	10.005
NJ	79.1	36.6	79.461	-0.361	0.130
NM	64.4	26.4	70.669	-6.269	39.300
NY	75.3	34.1	77.306	-2.006	4.024
NC	70.8	28.4	72.393	-1.593	2.538
ND	72.5	27.1	71.272	1.228	1.508
OH	71.2	26.1	70.410	0.790	0.624
OK	66.7	23.8	68.428	-1.728	2.986
OR	77.5	30.7	74.375	3.125	9.766
PA	72.4	28.7	72.651	-0.251	0.063
RI	76.5	32.4	75.841	0.659	0.434
SC	71.1	26.1	70.410	-3.810	14.516
SD	67.0	26.6	70.841	0.259	0.063
TN	71.8	24.8	69.290	-2.290	5.244
TX	79.6	27.5	71.617	0.183	0.033
UT	75.3	31.3	74.893	4.707	22.156
VT	75.8	35.7	78.685	-3.385	11.458
VA	78.9	36.1	79.030	-3.230	10.433
WA	64.9	32.7	76.099	2.801	7.846
WV	73.0	18.9	64.204	0.696	0.484
WI	75.5	27.7	71.789	1.211	1.467
WY	75.5	26.6	70.841	4.659	21.706

SSE = 480.003

We can divide the sum of squared errors by the remaining degrees of freedom for error, $n - p$ (which in this case equals $n - 2$), to calculate the mean square error:

$$\text{MSE} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{480.003}{48} = 10.00$$

Just as b_0 and b_1 are unbiased estimates of β_0 and β_1 under the least-squares criterion, so also the mean square error is an unbiased estimate of the variance of ε_i . It estimates how variable the errors of prediction are at each level of X_i . As we will discuss, it is assumed that the variance of these errors is constant across all values of X_i . The square root of this mean square error is known as the *standard error of prediction*.

We will do one more example using two other variables from the states dataset. For this example, we are going to examine whether a state's population density (measured in 2010 as hundreds of people per square mile) can be used to predict the automobile fatality rate in the state (measured in 2010 as the number of fatalities per 100 million vehicle miles traveled). One certainly might expect more automobile accidents in states that are more densely populated, but it is less clear what one might expect in terms of fatalities from such accidents. On the one hand, if the accident rate is higher in more densely populated states, one might also predict a higher fatality rate. On the other hand, in more densely populated states, perhaps accidents are less likely to result in fatalities since more of the accidents are likely to be simply fender-benders rather than more serious high-speed collisions.

The parameter estimates from the regression model make clear how these variables are related:

$$\hat{Y}_i = 1.28 - 0.05X_i$$

Let us interpret both parameter estimates in this model. Doing so will make clear that their interpretation depends on the metric in which the two variables are measured. First, the intercept, 1.28, represents the predicted number of fatalities (per 100 million vehicle miles driven) if a state's population density were zero. Of course this number is not very informative, since no state has a population density that is zero. Yet, it is the best linear prediction from these data, albeit well outside of the range of actual values of density found in the data. The slope, -0.05 , is negative, meaning that in more densely populated states the fatality rates are lower. The exact interpretation is that for every increase in population density of 100 people per square mile (the measurement metric of X_i) we predict a decrease in the fatality rate of .05 per 100 million vehicle miles driven (the measurement metric of Y_i).

AN ALTERNATIVE SPECIFICATION

It will prove useful at later points to be able to specify regression models in which the predictor variables have been put into “mean-deviation” form or “centered.” What this means is that for every observation we have taken the value of the variable and subtracted from it the variable's mean value. Thus, if the predictor variable is X_i , the mean-deviated or centered variable is $(X_i - \bar{X})$. This centered variable will necessarily have a mean of zero, i.e., $(\bar{X} - \bar{X}) = 0$. We will then regress Y_i on $(X_i - \bar{X})$ rather than on X_i :

$$Y_i = b'_0 + b'_1 (X_i - \bar{X})$$

The question is how these new parameter estimates, when the predictor is centered, differ from the parameter estimates that result from the estimation we have considered to this point, with an uncentered predictor:

$$\hat{Y}_i = b_0 + b_1 X_i$$

To answer this question, we can examine the formulas for the parameter estimates that we gave earlier, but this time with a centered X_i :

$$b'_1 = \frac{\sum((X_i - \bar{X}) - (\bar{X} - \bar{X}))(Y_i - \bar{Y})}{\sum((X_i - \bar{X}) - (\bar{X} - \bar{X}))^2}$$

$$b'_0 = \bar{Y} - b_1(\bar{X} - \bar{X})$$

Since $(\bar{X} - \bar{X}) = 0$, it follows that

$$b'_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = b_1$$

$$b'_0 = \bar{Y}$$

In other words, centering the predictor has no effect upon the slope, i.e., $b'_1 = b_1$, but it does change the intercept. The intercept with the predictor in mean-deviated or centered form will be the mean of Y_i .

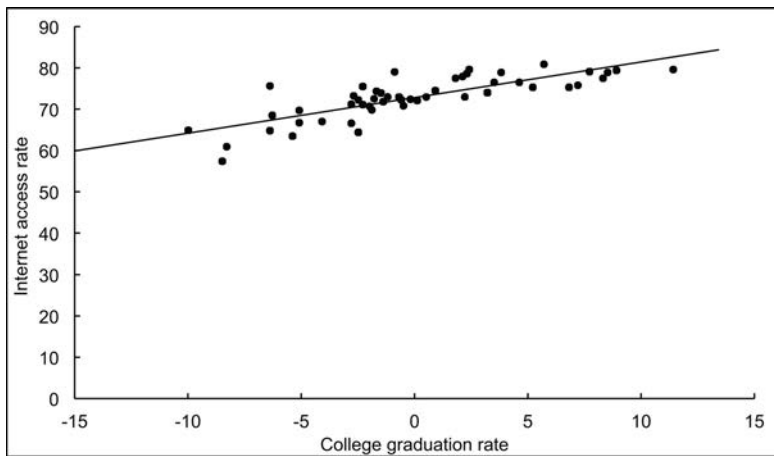
Conceptually, if we think graphically about the scatterplot of data and the superimposed regression line, by centering all we are really doing is changing the scale of the X_i axis in the scatterplot, redefining the value of X_i so that its mean equals zero. Such a scatterplot with the X_i axis centered is given in Figure 5.8. As this makes clear we have not changed the observations in the scatterplot at all; we have simply shifted the origin point in the plot. In a fundamental sense, our prediction function has not changed at all; the same line, having the same slope, minimizes the squared errors of prediction. The only change derives from the change in the location of the zero point on the horizontal axis. That zero point is now at the mean of X_i and, accordingly, the value of the intercept (i.e., the value of Y_i where the prediction function crosses the vertical axis) changes. It is now \bar{Y} . Obviously, this means that the regression line inevitably goes through the point defined by the joint means of the two variables.

In the case of the estimated model predicting internet access rates from college graduation rates, when the latter variable is centered, the resulting least-squares parameter estimates are:

$$\hat{Y}_i = 72.806 + 0.862(X_i - \bar{X})$$

The slope is unchanged by the centering of X_i but the intercept has changed. It no longer equals 47.912, rather it equals \bar{Y} , which is 72.806. One can still use the conventional interpretation for the intercept: It remains the predicted value when the predictor equals zero, i.e., when $(X_i - \bar{X})$ equals zero. And of course $(X_i - \bar{X})$ equals zero when $X_i = \bar{X}$.

Because all that has changed with centering the predictor is the zero point on the horizontal axis of the scatterplot, we are still dealing fundamentally with the same

FIGURE 5.8 Scatterplot and prediction function with centered X 

data and the same regression line, making the same predictions for each observation. Unsurprisingly, then, centering the predictor leaves the mean squared error of the model unchanged. In a deep sense the regression results are unchanged by this transformation.

STATISTICAL INFERENCE IN TWO-PARAMETER MODELS

Now that the basics of estimation and interpretation in simple regression models are clear, we turn to the issue of statistical inference, asking questions about parameter values in such models. Our approach to statistical inference in the case of two-parameter models will be identical to the approach we adopted in the single-parameter case. That is, we will compare an augmented model (in which both parameters are estimated from the data) to a compact one (in which one or more parameters are fixed at a priori values). We will calculate the sum of squared errors associated with both the augmented and compact models; from these we will then compute PRE, the proportional reduction in error as we go from the compact to the augmented model. This PRE, and its associated F statistic, can then be compared to their critical values, making the assumptions of normality, constant variance, and independence of residuals. Such a comparison permits a test of the null hypothesis that the values of the parameters fixed in the compact model are in fact the true unknown parameter values. Put the other way around, we are testing whether the estimated parameter values in the augmented model depart significantly from their a priori values specified in the compact model.

Given that the augmented model for such comparisons is now the two-parameter simple regression model, there are alternative compact models with which it may be compared. On the one hand, one may be interested in asking questions about the slope, comparing the augmented model to a compact one in which the slope parameter has been set to some a priori value. On the other hand, there may arise occasions when one is interested in testing a null hypothesis about the intercept in this two-parameter model, comparing it to a compact one in which the intercept has been fixed at some a priori value. We consider each one in turn.

Inferences about β_1

To ask statistical inference questions about the slope is to ask about associations: as the predictor variable (X_i) increases, what is our conclusion about whether the dependent variable (Y_i) increases or decreases, and at what rate. Our augmented two-parameter model for such questions is the one we have been using throughout this chapter:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The compact one, in its generic form, with which we will be making comparisons, is:

$$\text{MODEL C: } Y_i = \beta_0 + B_1 X_i + \varepsilon_i$$

where B_1 is some a priori value to which the slope parameter has been set. As before, the null hypothesis that is to be tested by the comparison between these two models is:

$$H_0 : \beta_1 = B_1$$

In words, the inferential test is whether the slope differs significantly from the a priori value.

By far the most common form this comparison takes is when the a priori value of the slope in the compact model equals zero. In this case the compact model is:

$$\text{MODEL C: } Y_i = \beta_0 + 0X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \varepsilon_i$$

and the null hypothesis is:

$$H_0 : \beta_1 = 0$$

The question being asked is whether X_i is a useful predictor of Y_i . If it is a useful predictor, then the predicted values of Y_i should change, either increasing or decreasing, as X_i changes. If we do no better with the augmented model (in which the slope is estimated from the data) than with the compact one (where the slope is constrained to equal zero) then it implies that the two variables may be unrelated. We do just as well making a constant prediction of all Y_i values regardless of an observation's X_i value as we do making conditional predictions.

For this comparison, the compact model is what was considered as the augmented model in the previous chapter. This will frequently be the case throughout the remainder of the book as we consider more complex models and additional inferential questions: What is for one question the augmented model becomes the compact model for a different question. So, in Chapter 4, we tested null hypotheses about a constant predicted value for every observation. Now we are testing whether we need to make conditional predictions and the compact model is one in which we estimate from the data a constant predicted value for each observation. That constant predicted value, when we estimate b_0 in Model C from the data, will not be the same as the intercept in Model A, when we estimate that model in our data. The best least-squares estimate of β_0 in Model C will be the mean of Y_i , just as it was (when we treated it as the augmented model) in Chapter 4. In the estimated augmented model (with two parameters), however, the estimate of β_0 will in general not be the mean of Y_i (unless of course the predictor has been centered). This makes clear another important point that will remain true as we consider more complex models in later chapters: the best estimate for a given parameter

in general depends on what other parameters are estimated in the model. In the compact single-parameter model that we are considering, the best estimate of β_0 will not in general be the same as the best estimate of β_0 in the two-parameter augmented model.

To ask whether the slope equals zero, and thus whether X_i is a useful predictor of Y_i , is the most common inferential question that one might ask about the slope. But it is certainly not the only question one might ask. Other Model Cs, with other values of B_1 , and thus other null hypotheses, might occasionally also be of interest. For instance, there are occasions when we are interested in testing the null hypothesis that:

$$H_0 : \beta_1 = 1$$

In this case, the augmented model remains the same two-parameter simple regression model, but Model C becomes:

$$\text{MODEL C: } Y_i = \beta_0 + 1X_i + \varepsilon_i$$

By casting all statistical inference questions in the form of Model A/Model C comparisons, testing such a null hypothesis becomes entirely feasible, even if standard statistical texts and software programs do not routinely provide such tests.

Testing the null hypothesis that $\beta_1 = 0$

In the context of our model that predicted states' internet access rates from their college graduation rates, let us test the first of the above null hypotheses, asking whether predictions of internet access are improved by making them conditional on college graduation rates, compared to a compact model that sets the slope at zero and thus makes a constant prediction for internet access for all states.

In terms of parameters, the models to be compared are:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

These are estimated as:

$$\text{MODEL A: } \hat{Y}_i = 47.0912 + 0.862X_i$$

$$\text{MODEL C: } \hat{Y}_i = 72.806$$

In Figures 5.9 and 5.10, we present the results of these two models for each state. In the first two columns of values in Figure 5.9 both Y_i (internet access rate) and X_i (college graduation rate) values for each state are given. Then the next three columns provide the results of Model C. The first of these provides the predicted values, \hat{Y}_{iC} . These are necessarily the same value for every state, since Model C predicts simply the mean internet access rate for each. Next, for each state we give its error, e_i , and its squared error, e_i^2 . Across all 50 states, the sum of squared errors for Model C equals 1355.028. This model has a single estimated parameter, hence $n - PC$ equals 49 (with 50 states). Thus, the mean square error for Model C is:

$$1355.028/49 = 27.654$$

Given that this is the simplest single-parameter model, as defined in the previous chapter, this mean square error is also called the variance of Y_i .

FIGURE 5.9 Model comparison for internet access data

<i>US state</i>	<i>Internet access rate</i>	<i>College graduation rate</i>	\hat{Y}_{IC}	e_{IC}	e_{IC}^2	\hat{Y}_{IA}	e_{IA}	e_{IA}^2	$(\hat{Y}_{IA} - \hat{Y}_{IC})^2$
AL	63.5	23.5	72.806	-9.306	86.602	68.169	-4.669	21.780	21.502
AK	79.0	28.0	72.806	6.194	38.366	72.048	6.952	48.330	0.575
AZ	73.9	27.4	72.806	1.094	1.197	71.531	2.369	5.612	1.626
AR	60.9	20.6	72.806	-11.906	141.753	65.669	-4.769	22.743	50.937
CA	77.9	31.0	72.806	5.094	25.949	74.634	3.266	10.667	3.342
CO	79.4	37.8	72.806	6.594	43.481	80.496	-1.096	1.201	59.136
CT	77.5	37.2	72.806	4.694	22.034	79.978	-2.478	6.140	51.438
DE	74.5	29.8	72.806	1.694	2.870	73.600	0.900	0.810	0.630
FL	74.3	27.2	72.806	1.494	2.232	71.358	2.942	8.655	2.097
GA	72.2	28.3	72.806	-0.606	0.367	72.307	-0.107	0.011	0.249
HI	78.6	31.2	72.806	5.794	33.570	74.806	3.794	14.394	4.000
ID	73.2	26.2	72.806	0.394	0.155	70.496	2.704	7.312	5.336
IL	74.0	32.1	72.806	1.194	1.426	75.582	-1.582	2.503	7.706
IN	69.7	23.8	72.806	-3.106	9.647	68.428	1.272	1.618	19.167
IA	72.2	26.4	72.806	-0.606	0.367	70.669	1.531	2.344	4.567
KS	73.0	31.1	72.806	0.194	0.038	74.720	-1.720	2.958	3.663
KY	68.5	22.6	72.806	-4.306	18.542	67.393	1.107	1.225	29.301
LA	64.8	22.5	72.806	-8.006	64.096	67.307	-2.507	6.285	30.239
ME	72.9	28.2	72.806	0.094	0.009	72.220	0.680	0.462	0.343
MD	78.9	37.4	72.806	6.094	37.137	80.151	-1.251	1.565	53.949
MA	79.6	40.3	72.806	6.794	46.158	82.651	-3.051	9.309	96.924
MI	70.7	26.9	72.806	-2.106	4.435	71.100	-0.400	0.160	2.910
MN	76.5	33.5	72.806	3.694	13.646	76.789	-0.289	0.084	15.864
MS	57.4	20.4	72.806	-15.406	237.345	65.497	-8.097	65.561	53.422
MO	69.8	27.0	72.806	-3.006	9.036	71.186	-1.386	1.921	2.624
MT	72.1	29.0	72.806	-0.706	0.498	72.910	-0.810	0.656	0.011
NE	72.9	29.4	72.806	0.094	0.009	73.255	-0.355	0.126	0.202
NV	75.6	22.5	72.806	2.794	7.806	67.307	8.293	68.774	30.239
NH	80.9	34.6	72.806	8.094	65.513	77.737	3.163	10.005	24.315
NJ	79.1	36.6	72.806	6.294	39.614	79.461	-0.361	0.130	44.289
NM	64.4	26.4	72.806	-8.406	70.661	70.669	-6.269	39.300	4.567
NY	75.3	34.1	72.806	2.494	6.220	77.306	-2.006	4.024	20.250
NC	70.8	28.4	72.806	-2.006	4.024	72.393	-1.593	2.538	0.171
ND	72.5	27.1	72.806	-0.306	0.094	71.272	1.228	1.508	2.353
OH	71.2	26.1	72.806	-1.606	2.579	70.410	0.790	0.624	5.741
OK	66.7	23.8	72.806	-6.106	37.283	68.428	-1.728	2.986	19.167
OR	77.5	30.7	72.806	4.694	22.034	74.375	3.125	9.766	2.462
PA	72.4	28.7	72.806	-0.406	0.165	72.651	-0.251	0.063	0.024
RI	76.5	32.4	72.806	3.694	13.646	75.841	0.659	0.434	9.211
SC	66.6	26.1	72.806	-6.206	38.514	70.410	-3.810	14.516	5.741
SD	66.6	26.6	72.806	-1.706	2.910	70.841	0.259	0.063	3.861
TN	71.1	24.8	72.806	-5.806	33.710	69.290	-2.290	5.244	12.362
TX	67.0	27.5	72.806	-1.006	1.012	71.617	0.183	0.033	1.414
UT	71.8	31.3	72.806	6.794	46.158	74.893	4.707	22.156	4.356
VT	79.6	35.7	72.806	2.494	6.220	78.685	-3.385	11.458	34.563
VA	75.3	36.1	72.806	2.994	8.964	79.030	-3.230	10.433	38.738
WA	75.8	32.7	72.806	6.094	37.137	76.099	2.801	7.846	10.844
WV	78.9	18.9	72.806	-7.906	62.505	64.204	0.696	0.484	73.994
WI	64.9	27.7	72.806	0.194	0.038	71.789	1.211	1.467	1.034
WY	75.5	26.6	72.806	2.694	7.258	70.841	4.659	21.706	3.861

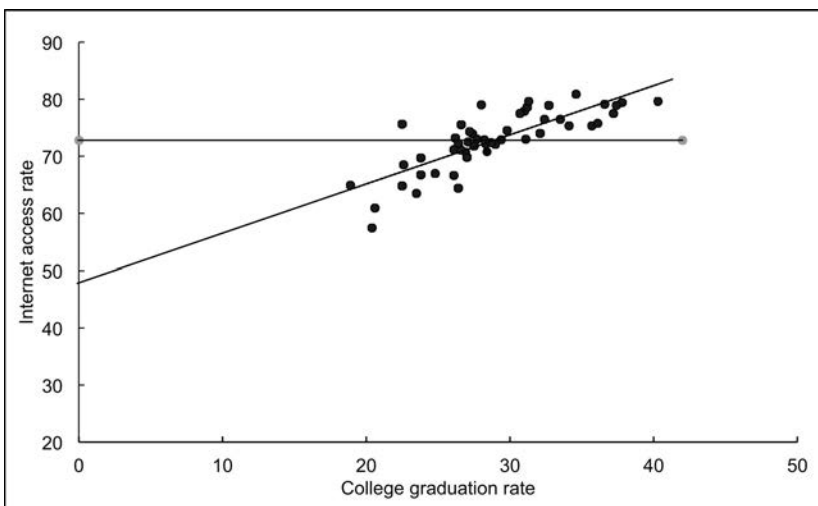
The next three columns in Figure 5.9 give the parallel results for Model A, first the predicted values (\hat{Y}_{iA}) and then errors of prediction and squared errors. Notice in this case that the predicted values now are not constant; rather, they are a linear function of the value of X_i for each state. And, again, the sum of the squared errors across the 50 states is SSE for Model A, equaling 480.00. In Model A we have estimated two parameters, accordingly PA equals 2, $n - PA$ equals 48, and the mean square error for this model is 10.00.

Model A, since it uses an additional parameter to model the data, necessarily does better than Model C in terms of sums of squared errors. That is not to say, however, that in every state the prediction made by Model A is better than that made by Model C. Examine Kansas (KS) for instance. Its internet access rate is 73.0%. Model C predicts it to be the mean of 72.806, which is a difference of 0.194. Model A, on the other hand, making a prediction conditional on Kansas' college graduate rate, predicts it to be 74.720, missing by 1.720. For this particular state, the Model A prediction is not as accurate as the Model C prediction. Yet, on average across states the squared errors of prediction must be at least as small from Model A as they are from Model C, simply because an additional parameter has been estimated.

Figure 5.10 presents the results graphically. The horizontal line superimposed on the data points is the prediction function of Model C. The other line is the prediction function of Model A. The inferential question that we now want to ask is whether the reduction in the SSEs when we replace Model C with Model A has been worth the reduction in the error degrees of freedom due to estimating an additional parameter. Graphically, the question is: Do we do sufficiently better with the sloped line to merit the added complexity over and above the horizontal prediction function? To answer this, we compute PRE, the proportional reduction in error as we move from Model C to Model A:

$$\text{PRE} = \frac{\text{SSE}(C) - \text{SSE}(A)}{\text{SSE}(C)} = \frac{1355.028 - 480.003}{1355.028} = .6458$$

FIGURE 5.10 Model C and Model A



Thus, when we make conditional predictions in this example, the total errors of prediction are reduced by more than 64% compared to simply predicting the mean level of internet access for each state.

The numerator of PRE is the sum of squares reduced (SSR) and can be calculated directly by taking the squared difference between the predicted value for Model A and that for Model C, and summing these across the 50 states:

$$\text{SSR} = \text{SSE}(C) - \text{SSE}(A) = \sum (\hat{Y}_{iA} - \hat{Y}_{iC})^2$$

In the final column of Figure 5.9, we present these squared differences in predicted values for every state. The sum of the numbers in this final column necessarily equals the difference in the sum of squared errors between the two models, i.e., $1355.028 - 480.003$ or 875.025 .

We can also compute the F statistic associated with the comparison between these two models. Below we do this with the equivalent expressions for F , either in terms of PRE or in terms of the sums of squares of the two models:

$$\begin{aligned} F &= \frac{\text{PRE}/(\text{PA}-\text{PC})}{(1-\text{PRE})/(n-\text{PA})} = \frac{.6458/1}{(1-.6458)/48} = 87.50 \\ &= \frac{\text{SSR}/(\text{PA}-\text{PC})}{\text{SSE}(A)/(n-\text{PA})} = \frac{875.025/1}{480.003/48} = 87.50 \end{aligned}$$

Using the appropriate tables in the Appendix, we compare the computed values of either PRE or F with their critical values, given that the assumptions of normality, constant variance, and independence of errors are met. The critical values at $\alpha = .05$ (with 1 and 48 degrees of freedom) are approximately .08 for PRE and 4.04 for F . Clearly, our computed values exceed these, and hence we can reject Model C in favor of the two-parameter Model A that makes conditional predictions. Our conclusion is that college graduation rates are a useful predictor of internet access rates in these data. Further, once we have rejected the null hypothesis of no relationship between the two variables, we can make a conclusion about the direction of the relationship between them, based on the sign of the estimated slope: In states where the percentage of people who graduated from college is higher, there are higher rates of internet access.

Figure 5.11 summarizes the results of the statistical analysis in an ANOVA table of the same type as we developed in Chapter 4. The first row provides information about the reduction in error achieved by including X_i as a predictor in the model (i.e., by using Model A instead of Model C). The entry in the SS column for that row is the SSR computed earlier. Associated with this SSR is a single degree of freedom, $\text{PA} - \text{PC} = 1$, for this comparison. The next row provides information about the error remaining in Model A, and its associated degrees of freedom ($n - \text{PA}$). The final row provides similar information for Model C. Calculating $\text{MS} = \text{SS}/\text{df}$ provides the basis for calculating F according to the sum of squares formula presented above. In the p column is indicated the fact that the computed PRE and F exceed the critical value with $\alpha = .05$. Finally, the value of PRE is given, computed as the SSR divided by SSE for Model C.

One final comment is appropriate about the statistics we have just computed. Throughout this book, we will use PRE to refer to the proportional reduction in error when replacing any compact model with an augmented one. In this sense, PRE is a very

general statistic that is informative regardless of the specifics of the models that are being compared. However, in the history of inferential statistics, a variety of specific names for PRE have been developed for particular model comparisons. For the case at hand, testing whether the slope in a single predictor model is different from zero, the square root of PRE is known as the Pearson correlation coefficient, commonly referred to simply as the correlation or r . Thus, the inferential test we have just conducted is equivalent to a test of whether the true correlation between two variables differs from zero. If PRE and F do not surpass their critical values, then it means that the true value of PRE (η^2) may equal zero. In fact, since the F statistic we have computed has only a single degree of freedom in its numerator, its square root is a t statistic with $n - PA$ ($n - 2$ in this case) degrees of freedom. It is simple algebra to show that the square root of the F formula we have given is the same as the t that is typically given in traditional statistics texts for testing whether a correlation between two variables is significantly different from zero:

$$t_{n-2} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

As a result, our conclusion in this case that Model A does significantly better than Model C is equivalent to concluding that the two variables are related, that β_1 differs from zero, that the true correlation differs from zero, and that the true PRE (η^2) differs from zero.

Testing null hypotheses for other values of β_1

There may be occasions when theory makes strong predictions about particular values of slopes other than zero. For instance, if two variables are measured in the same metric, then we might be interested in whether a one-unit difference on one variable is associated with a one-unit difference on the other. To illustrate, we continue to use the internet access/college graduation example. Since they are both measured in the same metric (percentage of the state who either has internet access or has attended college), we will ask the question of whether a 1% increase in college graduation is associated with a 1% increase in internet usage on average across the states. Obviously, we do not have any strong theory that would make this prediction; we use it for illustrative purposes only. The model comparison for this question is:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + 1X_i + \varepsilon_i$$

And the null hypothesis is that $\beta_1 = 1$.

FIGURE 5.11 ANOVA source table test of $H_0: \beta_1 = 0$ in simple regression context

Source	SS	df	MS	F	p	PRE
Reduction (using $b_1 = 0.862$)	875.025	1	875.025	87.50	< .001	.646
Error (using $b_1 = 0.862$)	480.003	48	10.000			
Total error (using $b_1 = 0$)	1355.028	49	27.654			

Unfortunately few statistical programs readily permit the estimation of parameters in a model in which some parameters are fixed at values other than zero. In this case, however, we can easily estimate β_0 in Model C by noting that if we subtract X_i from both sides of Model C we obtain:

$$Y_i - X_i = \beta_0 + \varepsilon_i$$

so the mean of the constructed variable $Y_i - X_i = 43.922$ estimates β_0 in Model C. The estimated models are then:

$$\text{MODEL A: } Y_i = 47.912 + 0.862X_i + e_i$$

$$\text{MODEL C: } Y_i = 43.922 + 1X_i + e_i$$

The SSE for Model C equals 502.486, again with $n - PC$ equal to 49. Model A makes better predictions, since as we saw above its sum of squared errors equals 480.003. SSR equals 22.483, with PRE and F computed as:

$$\text{PRE} = \frac{22.483}{502.486} = .045$$

$$F_{1,48} = \frac{22.483/1}{502.486/48} = 2.15$$

These values do not beat their respective critical values. Thus we cannot reject the null hypothesis—there is no evidence to conclude that the true slope is different from 1.00.

Confidence intervals of β_1

Recall from Chapter 4 that the confidence interval defines the range of values for the parameter for which we would fail to reject the null hypothesis. In other words, if we tested a null hypothesis that the parameter equals a value that lies within the confidence interval, we would fail to reject that null hypothesis. If the null hypothesis to be tested specifies that the parameter equals a value that lies outside of the confidence interval, we would reject that null hypothesis. In this sense, the confidence interval is entirely redundant with inferential tests about the value of a parameter.

In the case of β_1 in the two-parameter model we are considering, its confidence interval is given as:

$$b_1 \pm \sqrt{\frac{F_{\text{crit}; 1, n-2; \alpha} \text{MSE}}{(n-1)s_X^2}}$$

where $F_{\text{crit } 1, n-2; \alpha}$ is the critical value of F at level α , with degrees of freedom of PA – PC = 1 and $n - PA = n - 2$. MSE is the mean square error from the augmented model, again based on $n - 2$ degrees of freedom. And s^2 is the variance of the predictor variable.

For the internet access/college graduation example, the critical F with 1 and 48 degrees of freedom, using $\alpha = .05$, equals 4.04, MSE from Model A equals 10.00, and the variance of the predictor variable (college graduation rates) equals 24.04. Accordingly, the confidence interval for β_1 is:

$$0.862 \pm \sqrt{\frac{4.04(10.00)}{49(24.04)}} = 0.862 \pm 0.185$$

or

$$0.677 \leq \beta_1 \leq 1.047$$

Based on the present data, we can thus say we are confident that the true value for the slope in this two-parameter regression model, predicting internet access rates from college graduation rates, lies somewhere between 0.677 and 1.047. Notice that zero lies outside of this interval and that 1.00 lies within it. Both of these are consistent with the results already reported for our two null hypotheses: the null hypothesis that the parameter equals zero was rejected; the null hypothesis that it equaled 1.00 was not.

Although we might lament the fact that many statistical packages do not permit the estimation of models in which parameter values are fixed at values other than zero, the confidence interval permits a general approach for testing any null hypothesis about the slope. If the value that is specified by the null hypothesis lies within the interval, it would not be rejected. All other null hypotheses would be.

The formula for the confidence interval provides some insights into the factors that influence statistical power—the probability of rejecting the null hypothesis when it is false—and how to improve it. In general, the narrower the confidence interval, the more precision we have in estimating a parameter and statistical power means that we have greater precision, i.e., narrower confidence intervals. According to the formula, what are the factors that affect the width of the interval, and therefore power?

First, the critical value of F affects its width. If we use a smaller α level, thus reducing Type I errors, the critical value of F increases, thereby widening the confidence interval and resulting in less power and greater probability of Type II errors. This is the inherent tradeoff between Type I and Type II statistical errors.

Second, the width of the confidence interval is affected by the mean square error from the augmented model. As the variability of errors of prediction is reduced, the confidence interval becomes narrower. Thus, whatever we can do to reduce error, such as improving the quality of measurement of Y_i , will increase power.

Third, as n increases, all else being equal, power increases. This is reflected by the fact that $n - 1$ appears in the denominator of the confidence interval.

And, finally, the variance of the predictor variable X_i appears in the denominator. As X_i becomes more variable, the interval narrows and power improves. Given some predictor variable, we will examine its effects as a predictor with more precision (assuming a linear model) if we make sure that we sample widely across its values.

Power analysis in tests of simple regression

In Chapter 4 we performed “what if” power analyses for the simple model making a constant prediction for each observation. We can use exactly the same process to ask “what if” power analysis questions for simple regression models using one predictor variable. We perform “what if” power analyses for particular values of the true proportional reduction in error, η^2 , which may be of interest, in exactly the same way as before. For example, for the internet access data, we might want to know the power of detecting a relationship between it and some variable—detecting that the slope for a predictor is different from zero—when in fact we think that $\eta^2 = .20$. To do this, we can use a software program, such as R or SAS, to calculate power as we mentioned in the last chapter. For this Model A/Model C comparison, $PA - PC = 1$ and $n - PA = 48$.

If $\eta^2 = .20$ then the power calculator in SAS informs us that the power of our test is roughly .92. That is, if we were to do a study with an n of 50 and expect to find a relationship between a predictor and Y_i with a true PRE of .20, we would have roughly a 92% chance of correctly rejecting the null hypothesis.

Given that we now know the procedure for asking questions to determine the power with which we can assess whether one variable is significantly related to another in a simple regression model, we need to know what values of η^2 are appropriate and expected for such comparisons. As before, prior experience in a research domain may provide values of PRE that can be used directly in the power table. For the simple regression case, we might have estimates of PRE available from previous research, typically reported as the correlation between two variables, rather than as PRE itself. In this case, we need to square the correlation coefficient to obtain the estimate of PRE, since $\text{PRE} = r^2$ for this question. As before, we would want to convert past empirical values of PRE to unbiased estimates of η^2 , using the same adjustment formula as before:

$$\text{Unbiased estimate of } \eta^2 = 1 - \frac{(1 - \text{PRE})(n - \text{PC})}{n - \text{PA}}$$

To illustrate, suppose prior research has reported a correlation between two variables of .33 based on a sample size of 30. We intend to do a study, examining the relationship between the same two variables, but we intend to use an n of 50. What we would like to know is the power of our planned study, if in fact the two variables are related as strongly as they were reported to be in the prior research. To do this, we first convert the previously reported correlation to PRE, by squaring it: $.33^2 = .11$. We then convert this PRE to an unbiased estimate of η^2 using the above formula and the sample size from the prior study that reported the .33 correlation:

$$\text{Unbiased estimate of } \eta^2 = 1 - \frac{(1 - .11)(29)}{28} \approx .078$$

We then use SAS to estimate our power in the new study we plan to conduct. With an n of 50, $n - \text{PA}$ for our study will be 48, and with an anticipated η^2 of .078, our power is approximately .51. Given this result, we may want to think further about our anticipated study. It might be worthwhile, for instance, to recruit a larger sample to increase power.

If we do not have relevant prior experience for estimating η^2 , we can use the values suggested in Chapter 4 for “small” ($\eta^2 = .03$), “medium” ($\eta^2 = .10$), and “large” ($\eta^2 = .30$) effects. From these and the anticipated n for a new study, we can get the approximate power.

A third and final approach for finding an appropriate value of η^2 for “what if” power analyses for the simple regression model involves guesses about the parameter values and variances, just as in Chapter 4. Again, to have reasonable expectations about the parameter values and variances generally requires as much or more experience in a research domain as is necessary to know typical values of PRE. We present it, however, for completeness.

The formula that relates true values of the parameters to η^2 , the true value of PRE, is:

$$\eta^2 = \beta_1^2 \frac{\sigma_X^2}{\sigma_Y^2}$$

where β_1 is the true parameter value for the slope, σ_x^2 is the true variance of the predictor variable, and σ_y^2 is the true variance of the dependent variable. In other words, given some alternative hypothesis that specifies what we think is the correct value for the slope, β_1 , and given that we want to determine the power of our test of the null hypothesis that β_1 equals zero, we can calculate the corresponding η^2 using the above expression, assuming we have estimates of the variances of both X_i and Y_i . We can then take that value, and the anticipated $n - 2$ for the study in the planning, and estimate power.

Inferences about β_0

Our discussion so far has concentrated exclusively on inferences about the slope in the two-parameter simple regression model. But one may certainly also compare this augmented two-parameter model with a compact one that fixes the intercept, rather than the slope, at some a priori value:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = B_0 + \beta_1 X_i + \varepsilon_i$$

where B_0 represents an a priori value for the intercept. The null hypothesis tested by such a comparison would then be:

$$H_0 : \beta_0 = B_0$$

Although perhaps not of frequent theoretical interest in the behavioral sciences, one particular form of this compact model involves what is known as “regression through the origin” in which the intercept is fixed at zero:

$$\text{MODEL C: } Y_i = 0 + \beta_1 X_i + \varepsilon_i$$

Recall that the intercept is defined as the value predicted by the model when X_i equals zero. Thus, regression through the origin means that the line goes through the $(0, 0)$ point on the scatterplot.

Importantly, null hypothesis tests about the intercept are of a fundamentally different nature than those about the slope. When we are making inferences about the slope, we are asking about the rate of change in predicted values. And when we test that the slope is zero, we are asking whether there is any change in the predicted values when the predictor varies, i.e., whether X_i and Y_i are related. On the other hand, inferences about the intercept are inferences about predicted values, not inferences about changes in predicted values. Specifically, we are asking whether the predicted value of Y_i when X_i equals zero differs significantly from some a priori value, be it zero (in regression through the origin) or some other value.

In many cases, there is no intrinsic interest in the predicted value of Y_i when X_i equals zero. This may be because the value of zero lies outside of the range of the X_i values represented in the dataset, as in the data that we have been using where no state has a college graduation rate that is near zero. Even if zero is a value within the range of X_i in the dataset, it may not be a point that has much theoretical interest. However, with simple transformations of X_i , such as centering it or deviating it from its mean (discussed earlier), the zero value becomes of considerably greater potential interest. Earlier in this chapter we saw that when X_i is centered the intercept in the estimated

model will equal the mean of Y_i (i.e., \bar{Y}). Accordingly, with a centered predictor, the following Model A/Model C comparison is equivalent to asking questions about the mean of Y_i :

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

$$\text{MODEL C: } Y_i = B_0 + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

with the following null hypothesis:

$$H_0: \beta_0 = B_0 \text{ or } \mu_Y = B_0$$

where μ_Y is the true mean of Y_i .

Let us illustrate such a test with the internet access data that we have used throughout this chapter, asking the same question that we did at the end of the last chapter—whether the projection of a mean rate of 75% was overly optimistic for the year 2013. But this time, we will ask the question in the context of a simple regression model that makes conditional predictions of internet access based on states' college graduation rates. We will then compare our test in this model with our results from Chapter 4 to examine how the present test differs from that used there.

Estimating a model in which internet access rates are regressed on college graduation rates, with the latter variable in its centered or mean-deviated form, gives the following estimates:

$$\text{MODEL A: } \hat{Y}_i = 72.806 + 0.862(X_i - \bar{X})$$

with a sum of squared errors of 480.003. As explained earlier, the intercept in this model is now the mean value for the internet access variable, while the slope has not changed compared to the model in which X_i was not centered. Additionally, in a deep sense, this model is identical to the one with X_i uncentered, in that it makes the same predictions and therefore necessarily has the same sum of squared errors.

To test the null hypothesis that the true mean of Y_i equals 75, we want to compare this model to a compact one in which the intercept has been fixed at 75:

$$\text{MODEL C: } \hat{Y}_i = 75 + b_1(X_i - \bar{X})$$

This comparison permits a test of the null hypothesis:

$$H_0: \beta_0 = 75 \text{ or } \mu_Y = 75$$

As we said previously, many computer packages for data analysis do not readily permit the estimation of models in which parameters have been fixed at values other than zero. In this case, with a centered predictor, it can be shown that the best least-squares estimate for β_1 does not change even if we fix the intercept at some a priori value.¹ Accordingly, Model C is estimated as:

$$\text{MODEL C: } \hat{Y}_i = 75 + 0.862(X_i - \bar{X})$$

The sum of squared errors from this model can be directly computed across the observations. More simply, it can be computed by first calculating the sum of squares reduced (SSR) as we move from Model C to Model A.

Recall that $SSR = \Sigma(Y_{iA} - Y_{iC})^2$, accordingly:

$$SSR = \Sigma((72.806 + 0.862(X_i - \bar{X})) - (75 + 0.862(X_i - \bar{X})))^2$$

Because both predicted values have the same slope for the centered predictor, this reduces to:

$$SSR = \Sigma((72.806) - (75))^2 = 50(-2.194)^2 = 240.682$$

Accordingly, the sum of squared errors for the compact model is:

$$SSE(C) = SSE(A) + SSR = 480.003 + 240.682 = 720.685$$

Now that we have the sums of squared errors, we can calculate PRE and F for the comparison and the test of the null hypothesis:

$$PRE = \frac{SSR}{SSE(C)} = \frac{240.682}{720.685} = .334$$

$$F = \frac{.334/1}{(1 - .334)/48} = \frac{240.682/1}{480.003/48} = 24.07$$

All of this is summarized in the ANOVA source table of Figure 5.12.

Let us now compare these results with the test of the same null hypothesis reported at the end of Chapter 4, in the context of the simplest single-parameter model. There, estimated Models A and C were:

$$\text{MODEL A: } \hat{Y}_i = 72.806$$

$$\text{MODEL C: } \hat{Y}_i = 75$$

The sum of squared errors associated with Model A was 1355.046, that is, the total sum of squares of Y_i around its mean. While the SSR was found to be:

$$SSR = \Sigma((72.806) - (75))^2 = 50(-2.194)^2 = 240.682$$

Thus, the results of this test of the same null hypothesis yielded the ANOVA source table in Figure 5.13.

Although both tests resulted in the rejection of the null hypothesis, clearly the two approaches differ substantially in terms of the obtained PRE and F . Those values, in the context of the two-parameter simple regression model, are more than twice what they were for the same test in the context of the single-parameter model of Chapter 4. And this difference is attributable entirely to the difference in the sum of squared errors for Model A (and the concomitant difference in $n - PA$). The sum of squared errors for Model A in the context of the simple regression model equals 480.003, with $n - PA$ equal to 48. The same values in the context of the single-parameter model of Chapter 4 are $SSE = 1355.046$ and $n - PA = 49$. As a result, the MS error values (denominators of F) are markedly different: 10.00 (for the simple regression model) versus 27.654 (for the Chapter 4 single-parameter model). Importantly, the numerator of F is the same in both tests, with SSR equal to 240.682 and $PA - PC = 1$.

The difference in the sums of squared errors for Model A in the two source tables is, not surprisingly, the sum of squares that is attributable to the predictor variable

FIGURE 5.12 ANOVA source table test of $H_0: \beta_0 = \mu_Y = 75$ in simple regression context

Source	SS	df	MS	F	ρ	PRE
Reduction (using $b_0 = 72.806$)	240.682	1	240.682	24.07	< .001	.334
Error (using $b_0 = 72.806$)	480.003	48	10.000			
Total error (using $b_0 = 75$)	720.685	49	14.708			

FIGURE 5.13 ANOVA source table test of $H_0: \beta_0 = \mu_Y = 75$ in single-parameter model (from Chapter 4)

Source	SS	df	MS	F	ρ	PRE
Reduction (using $b_0 = 72.806$)	240.682	1	240.682	8.71	< .01	.151
Error (using $b_0 = 72.806$)	1355.046	49	27.654			
Total error (using $b_0 = 75$)	1595.728	50	31.915			

(college graduation rates), on the basis of which conditional predictions are made in the simple regression model. That is, earlier, we reported that the SSR attributable to a Model A that made conditional predictions of internet access, using college graduation rates as a predictor, compared to a Model C that predicted the mean value of Y_i , was 875.025. This is exactly the difference between each Model A used in these two tests. Model A for the Chapter 4 single-parameter version of the test makes unconditional predictions of Y_i ; Model A for the simple regression version of the test makes predictions of Y_i conditional on (centered) X_i . Making these conditional predictions means that Model A in the simple regression context has one fewer degrees of freedom for error ($n - PA = 48$) than Model A in the Chapter 4 version of the test (where $n - PA = 49$). But the loss in degrees of freedom has been more than compensated for by the substantial difference between the sums of squared errors of the two Model As. As a result, the test in the context of the conditional simple regression model has substantially more statistical power than the same test conducted in the context of the single-parameter model of Chapter 4.

In Chapter 4 we mentioned that the test we reported is known as the single-sample t -test. The advantage of our model comparison approach is that we have been able to generalize this test to cases where conditional predictions are made by the models that are compared. In the jargon traditionally used in the statistical inference literature, we have just conducted a single-sample t -test while controlling for a “covariate.”

A further advantage of our approach is that it permits us to conduct inferential tests about predicted values other than the mean. Suppose, for instance, that we had some reason to want to ask about internet usage rates in states where the college graduation rate was 70%. Rather than centering the predictor around its mean value, one could deviate the predictor from the value of 70. Then one could estimate a Model A in which the predictor variable was this deviated variable:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1(X_i - 70) + \varepsilon_i$$

In a deep sense we would be dealing with the same conditional model; we have simply moved the zero point on the x -axis to what was the value of 70. Accordingly, the slope

remains the same, while the estimated intercept in the model would be the predicted value of Y_i when X_i equals 70. This model might then be compared to a Model C that uses the same deviated predictor but fixes the intercept at some a priori value of interest.²

Confidence interval for the intercept

The confidence interval for the intercept represents the range of values for the intercept that would not be rejected by an inferential statistical test. In simple regression models, with a single predictor variable, the confidence interval for the intercept is:

$$b_0 \pm \sqrt{\frac{F_{\text{crit}; 1, n-2; \alpha} \text{MSE}(\sum X_i^2)}{n \sum (X_i - \bar{X})^2}}$$

This is the confidence interval for the intercept regardless of whether the predictor has been deviated from some value or not. If it has been deviated, then of course the X_i terms in the confidence intervals are the new values following deviation.

When using a centered predictor, deviated from its mean value, then \bar{X} for the centered predictor equals zero, and the above formula for the confidence interval reduces to:

$$b_0 \pm \sqrt{\frac{F_{\text{crit}; 1, n-2; \alpha} \text{MSE}}{n}}$$

which is the formula that we gave in Chapter 4 for the confidence interval for β_0 in the single-parameter model (except there the critical F value had 1 and $n - 1$ degrees of freedom). Of course, with a predictor variable in the model that is a useful predictor of Y_i , the MSE in the numerator of this confidence interval should be considerably smaller than the MSE in the numerator of the interval used in Chapter 4, without a predictor. This difference reflects the increase in power when conducting inferential tests about the mean in the context of a useful predictor compared to the same test in the single-parameter context of Chapter 4.

To illustrate, with the centered predictor from the simple regression model used in this chapter, the confidence interval for β_0 (equivalently for μ_y) equals:

$$72.806 \pm \sqrt{\frac{4.04(10.000)}{50}}$$

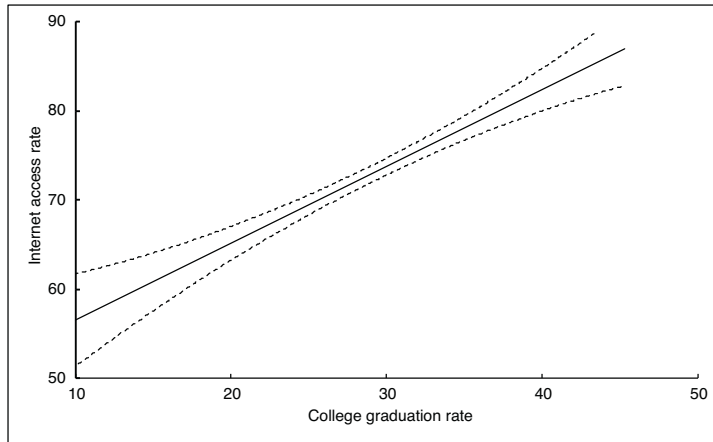
$$71.907 \leq \beta_0 \leq 73.705$$

On the other hand, in the context of the single-parameter model of Chapter 4, the same confidence interval equals:

$$72.806 \pm \sqrt{\frac{4.03(27.654)}{50}}$$

$$71.313 \leq \beta_0 \leq 74.299$$

Clearly, the confidence interval for the mean of Y_i is smaller in the context of the simple regression model, reflecting the substantial increase in power resulting from the inclusion of the predictor variable (with its associated reduction in errors of prediction).

FIGURE 5.14 Confidence limits of predicted values

Using the general formula for the confidence interval for β_0 in this two-parameter model given above:

$$b_0 \pm \sqrt{\frac{F_{\text{crit}; 1, n-2; \alpha} \text{MSE}(\sum X_i^2)}{n \sum (X_i - \bar{X})^2}}$$

We can generate confidence intervals for predicted values at all possible levels of the predictor variable (by deviating the predictor from each of those levels and then calculating the resulting confidence intervals for the varying intercepts). In Figure 5.14 we have graphed the 95% confidence limits for the predicted values of internet access at values of the predictor (college graduation rate) ranging from 10% to about 45%. The middle straight line in the graph represents the predicted values and the curved lines above and below it represent the upper and lower confidence limits. What this figure makes clear is that the confidence interval is narrowest near the mean value of the predictor, where the predicted value is the mean of the dependent variable, and it becomes wider as we depart in either direction from that mean. Thus, we have greater precision in inferring predicted values near the joint mean of the bivariate distribution than when we move further away along the horizontal axis.

Two-Parameter Model Comparisons

To conclude this chapter on two-parameter simple regression models, we should note that MODEL A/MODEL C comparisons are now also possible involving $PA - PC > 1$. For instance, suppose we wanted to simultaneously ask about fixed a priori values both for the intercept and for the slope:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = B_0 + B_1 X_i + \varepsilon_i$$

where both B_0 and B_1 are fixed a priori values (zero or any other value). The resulting null hypothesis from this comparison has two different components to it:

$$H_0 : \beta_0 = B_0; \beta_1 = B_1$$

To illustrate a situation where this sort of model comparison might be of interest, suppose we had data on the heights of each member of father–son pairs and we wanted to examine how these two height measures were related. We might estimate a Model A predicting each son’s height from his father’s height, estimating both the slope and intercept as in Model A above. This model might then be meaningfully compared with a Model C in which the intercept was fixed at 0 and the slope at 1, yielding the following model comparison and null hypothesis:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = X_i + \varepsilon_i$$

$$H_0 : \beta_0 = 0; \beta_1 = 1$$

Such a comparison asks whether sons and fathers are perfect resemblances of each other in terms of height. Except for error, Model C assumes that sons’ heights equal their fathers’ heights.

There is nothing statistically wrong with such a two-parameter model comparison and such a compound null hypothesis. The resulting PRE and F would be computed in the same ways as always, albeit with $PA - PC = 2$ degrees of freedom for the numerator. Since there is nothing to estimate in Model C, the predicted values and sum of squared errors from this model could be easily obtained.

The problem comes in interpreting the results. If Model C is rejected in favor of Model A (i.e., if the null hypothesis is rejected) we will not be able to be confident about *why* it was rejected. Maybe it is the case that the a priori value in Model C for the intercept is wrong. Maybe it is the case that the a priori value in Model C for the slope is wrong. Maybe both a priori values are wrong. All we can say is that Model A is preferred over Model C, but we will not know why.

It is for this reason that we prefer model comparisons where $PA - PC = 1$, where the numerator of the F statistic has only a single degree of freedom. There is nothing wrong with statistical tests involving more than one degree of freedom in the numerator of F , and we will occasionally discuss them and even recommend them. In general, however, we will refrain from such unfocused model comparisons. We simply note that in the context of the present two-parameter simple regression model they become possible for the first time.

SUMMARY

In this chapter we considered models with a single predictor, measured more or less continuously. We started by considering the definitions of both the intercept and slope in such models, with the former being a particular predicted value (when the predictor equals zero) and the latter being the unit difference between predicted values. We then provided formulas for estimating these parameters and used these to illustrate alternative ways of thinking about what a slope estimate means. Finally, we considered

models in which the predictor is centered or put into mean-deviation form. In such cases, the slope of the predictor is unchanged while the intercept equals the mean of the data variable, Y .

The second half of the chapter was devoted to model comparisons, treating the two-parameter, single-predictor model as Model A and comparing it to an alternative Model C, testing inferences about the slope and the intercept (and both simultaneously). Inferential tests of the slope most frequently make comparisons with a Model C that fixes the slope at zero, thus testing the null hypothesis that the predictor is not a useful one, or equivalently that the predictor and Y are unrelated to each other. There are occasions, however, when other null hypotheses about the slope are of interest and we illustrated these. Inferences about the intercept are most frequently of interest when the predictor has been centered, thus permitting inferences about the mean of Y . Testing null hypotheses about the mean of Y will be more powerful in the presence of a predictor when in fact that predictor is a useful predictor; that is, it explains a significant amount of variation in Y . This was illustrated by making comparisons with the simplest model comparisons of Chapter 4.

Notes

- 1 It is only in the case of a centered predictor in simple regression models that the estimate of the slope will remain constant regardless of whether the intercept is estimated or fixed at various a priori values. Unless predictors are centered around their mean, this will generally not be the case.
- 2 Importantly, the estimated slope in such a Model C will differ from the estimated slope in Model A. As we mentioned previously, only when the predictor variable is centered around its mean will the estimated slope remain the same regardless of whether the intercept is estimated or fixed. For all other cases, the slope estimate will vary.

Multiple Regression

Models with Multiple Continuous Predictors

In the previous chapters we considered the simple model that makes the same prediction for all the observations in the data and also the simple regression model. We now increase complexity by considering models that make predictions conditional on multiple predictor variables. For example, in our discussion of the rates of internet use, we suggested several plausible hypotheses about variables that might be predictive of each state's internet use—proportion of college graduates, urban versus rural, and population density. The simple regression model limited us to testing one predictor variable at a time. To test our different hypotheses using the simple regression model, we would have to first test the model with one predictor, say, proportion of college graduates, and then test another one-variable model with, say, population density as the predictor. It is obviously more efficient to test more than one predictor simultaneously. More importantly, understanding differences in the states' internet use will likely require more than one predictor. For example, we might need to base our predictions on *both* the proportion of college graduates and the population density. The solution we consider in this chapter is *multiple regression*.

Multiple regression is very similar to simple regression. There are, however, a few important differences. We begin by briefly considering these similarities and differences with respect to the model, the new problem of redundancy among the predictors, and statistical inference. Then we will consider an extended example designed to illustrate the interpretation of the parameters in the multiple regression model.

MULTIPLE REGRESSION MODEL

We want to consider models of the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (6.1)$$

where, as before, Y_i represents an observed data value and ε_i represents the error disturbance. X_{ij} represents the value of the i th observation on the j th predictor variable and β_j is the *partial regression coefficient* representing the weight we should give to X_{ij} in making our predictions conditional on X_{ij} . In other words, the partial regression coefficients are the degree to which we adjust our prediction \hat{Y}_i by each observation's X_{ij} value. In the multiple regression model the coefficients are called *partial regression coefficients* because, as we shall see later, the value of, say, β_1 may well depend on whether, say, the predictor X_{i2} and its parameter β_2 are included in the model. To remind us that the meaning of β_j is conditional on the other predictor variables included in the

equation, we sometimes use the notation $\beta_{j.123\dots p-1}$. The letter before the dot in the subscript represents the variable with which the regression coefficient β is associated and the numbers after the dot represent the other variables that are simultaneously included in the model equation. But this notation is cumbersome, so we, and certainly most computer programs, more often just use β_j with the understanding that its meaning depends on all the other parameters and predictors included in the model.

We often refer to the model of Equation 6.1 as the multiple regression model or sometimes as the general linear model. Linear means that the separate components, after first being weighted by their respective β_j , are simply added together. The general linear model is indeed very general and constitutes an extremely important tool for data analysis. As we shall see in subsequent chapters, even the “linear” part is not much of a restriction because many apparently nonlinear models can be represented in terms of Equation 6.1 by a suitable choice of the predictor variables. It is also clear that the models of the previous chapters are simply special cases of the multiple regression model. In fact, a great many statistical procedures can be represented in terms of Equation 6.1 with clever construction of the X_{ij} predictor variables. The following chapters are devoted to consideration of a number of interesting and very important special cases, such as analysis of variance, and a number of special problems in the context of this model, such as violation of the assumption that the ε_i error terms are independent. In this chapter, we present the basics of the general model—basics that we will use again and again.

Redundancy

The power and generality of the multiple regression model do not come without a cost; that cost is a definite increase in complexity and the introduction of some special problems that do not arise in the case of the simple models considered previously. One such problem is redundancy among the predictors. Suppose, for example, that we were attempting to predict the weights of a sample of elementary school children with two predictor variables: height measured in inches and height measured in centimeters. Our two predictor variables are obviously completely redundant—either would do as well as the other for predicting weight, and if we had already used one as a predictor there would clearly be no benefit to adding information from the other predictor.

Redundancy among the predictor variables is seldom so extreme as measuring height in inches and centimeters. A more typical example would be predicting the weights of the elementary school children using height and age as predictors. Height and age are obviously related—knowing a child’s age we can make a reasonable guess as to his or her height—but they are not completely redundant, because our guess of height from age would not be perfect. This means that height and age share in their ability to reduce error in the predictions of weight. If we first used height as a predictor, we would be improving our predictions (and reducing our error) in some of the same ways we would have if we had first used age as a predictor. Sorting out the effects of predictor variables and their relative importance is clearly going to be difficult when there is redundancy among our predictors. In general, there is redundancy when it is possible to predict, at least somewhat, one or more of the X_{ij} with some combination of the other predictor variables. In our analysis of multiple regression models we will have to be alert for redundancy and be very careful about the model interpretations we make in such cases. We will consider special techniques for doing so.

Statistical inference

Although we will have more choices of models to compare and test with multiple predictor variables, the multiple regression model poses no new problems for statistical inference. To make statistical inferences we use what is now our standard operating procedure for any statistical model. We first fit our model to the data by estimating parameters so as to minimize error. Next we calculate PRE for our augmented model (Model A) relative to some suitably chosen compact model (Model C) defined by the null hypothesis that we wish to test. Then we calculate F , an index of how much PRE we get per extra parameter. We then compare these calculated statistics (PRE and F) with their critical values, values from their sampling distributions that would surprise us if Model C (and the null hypotheses) were true. If the obtained values of PRE and F are surprising, then we reject Model C in favor of Model A. On the other hand, if the obtained values of PRE and F are not surprising, then we do not reject Model C. We consider, in turn, each of these steps for the multiple regression model, but only the details of estimating parameters are really new. The statistical inference process itself is *exactly* the same as before.

ESTIMATING PARAMETERS IN MULTIPLE REGRESSION

At the conceptual level, the problem of estimating parameters for the multiple regression model is exactly the same as for previous models that we have considered. We want to find the least-squares estimators of $\beta_0, \beta_1, \dots, \beta_{p-1}$ that will minimize SSE, the sum of squared errors. That is, we want to find the estimates b_0, b_1, \dots, b_{p-1} in the estimated model:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + \dots + b_{p-1} X_{i,p-1}$$

so that:

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2$$

is as small as possible. In that way we will ensure that the model fits the data as closely as possible so as to make the error (as measured by SSE) as small as possible.

If there is no redundancy among the predictors, then the same procedures we used in Chapter 6 to estimate slopes in the simple regression model will produce the appropriate estimates for the parameters in the multiple regression model. That is, the best estimate for each slope is:

$$b_j = \frac{\sum (X_{ij} - \bar{X}_j) (Y_i - \bar{Y})}{\sum (X_{ij} - \bar{X}_j)^2}$$

and the best estimate of the intercept is:

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 - \dots - b_{p-1} \bar{X}_{p-1}$$

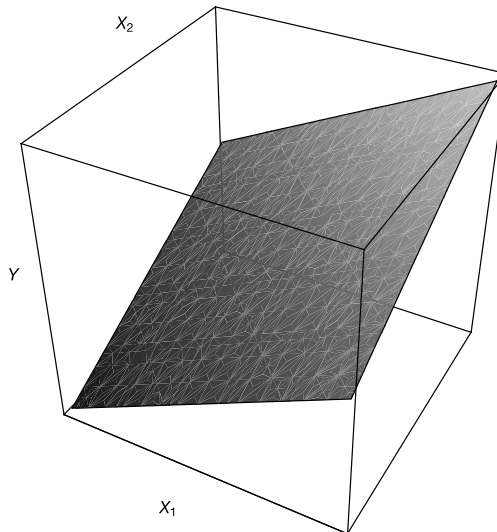
These formulas apply whenever there is no redundancy among the predictors. To say there is no redundancy among the predictors is equivalent to saying that the correlation between every pair of predictors equals zero. Unless the predictor variables are especially constructed to meet this condition (as in some experimental designs), at least some

redundancy among the predictors will always be present. In the presence of redundancy, deriving the least-squares parameter estimates is considerably more complicated. We can no longer use the procedures outlined in Chapter 5 and given by the above formula for b_j to estimate those parameters. There are well-defined algorithms for finding the unique least-squares estimates when there is redundancy among the predictors. However, these algorithms are extremely tedious on hand calculators for even just a few predictors and are, on the whole, neither very conceptual nor helpful in aiding our understanding of multiple regression. We therefore will leave the calculation of parameter estimates to the computer. All standard computer statistical packages have multiple regression procedures that can estimate the parameters of any of the models we consider in this book.

Before turning over all of our calculations to the computer, however, it is important that we have a firm understanding of the underlying concepts. For the simple single-parameter model we found the single value b_0 (a point, geometrically) that minimized SSE. For simple regression we found two values b_0 and b_1 that defined a line: b_0 is the intercept of that line, and b_1 is its slope. For a multiple regression model with two predictors the three estimates b_0 , b_1 , and b_2 define a plane, as depicted in Figure 6.1. The two predictor variables define two of the axes in the three-dimensional space (the two horizontal ones), and the data variable provides the vertical one. Error is represented geometrically as the vertical distance between each observation and the plane that represents the model.

In simple regression we conceptually moved the line around, changing its slope and its intercept, until we found the position of the line that minimized total error. Similarly, for the multiple regression model we move the plane around until we find the location that minimizes total error. (Actually we want to minimize squared error, but we have not complicated the picture in Figure 6.1 by adding the appropriate squares; conceptually the idea is the same.) The intersection of the plane with the y -axis defines b_0 , the slope of the plane with respect to X_{i1} defines b_1 , and the slope with respect to X_{i2} defines b_2 .

FIGURE 6.1 Plane defined by multiple regression model with two predictors



Conceptually the complicated computer algorithms simply find the location of the plane that minimizes the sum of squared errors, where an error is defined as the vertical distance from each observation to the model plane. For more than two predictors the model is equivalent to a hyperplane in spaces with four or more dimensions. Although it is impossible to draw a picture of such hyperplanes, the concept is exactly the same. The best location for the hyperplane minimizes the sum of squared errors, and the hyperplane's intercept and slopes define the parameter estimates.

INTERPRETING PARAMETERS IN MULTIPLE REGRESSION

We will use a detailed examination of the data in Figure 6.2 to develop our understanding of the meaning of the partial regression coefficients in a multiple regression model. This dataset, which appears in many SAS examples (SAS Institute, 2008) and in other examples found on the web, contains the weight, height, age, and sex of 19 middle school or junior high students. It is particularly useful for developing our understanding of partial regression coefficients because we have good intuitions about the relationships among these variables. Our goal will be to develop a model of the variable weight. We expect boys to weigh more than girls unless the girls are a lot older and taller than the boys. We expect older students to be heavier, except if the younger students happen to be exceptionally tall. We will see that such intuitions underlie the interpretation of multiple regression models.

We will begin by considering age and height as predictors of weight. That is, our model is:

$$Wt_i = \beta_0 + \beta_1 Age_i + \beta_2 H_i + \varepsilon_i$$

FIGURE 6.2 Sex, age, height, and weight of 19 middle school students

<i>Name</i>	<i>Sex</i>	<i>Age</i>	<i>Height (in.)</i>	<i>Weight (lb)</i>
Alfred	M	14	69.0	112.5
Antonia	F	13	56.5	84.0
Barbara	F	13	65.3	98.0
Camella	F	14	62.8	102.5
Henry	M	14	63.5	102.5
Jamal	M	12	57.3	83.0
Jane	F	12	59.8	84.5
Janet	F	15	62.5	112.5
Jayden	M	13	62.5	84.0
John	M	12	59.0	99.5
Joyce	F	11	51.3	50.5
Judy	F	14	64.3	90.0
Louise	F	12	56.3	77.0
Mary	F	15	66.5	112.0
Philip	M	16	72.0	150.0
Robert	M	12	64.8	128.0
Sequan	M	15	67.0	133.0
Thomas	M	11	57.5	85.0
William	M	15	66.5	112.0

where β_1 is the amount we adjust each person's weight estimate up or down depending on his or her age and β_2 is the amount we adjust each person's weight estimate up or down depending on his or her height. Figure 6.3 depicts data for the weights, heights, and ages for these 19 students as a three-dimensional scatterplot. As landmarks, the graph denotes the points for Philip (72 inches tall, age 16, and 150 pounds) and Joyce (51.3 inches tall, age 11, and 50.5 pounds). It is easy to see that the points in Figure 6.3 are not randomly scattered but instead tend to fall along a plane. More importantly, it

FIGURE 6.3 Three-dimensional scatterplot of the relationship of age and height to weight

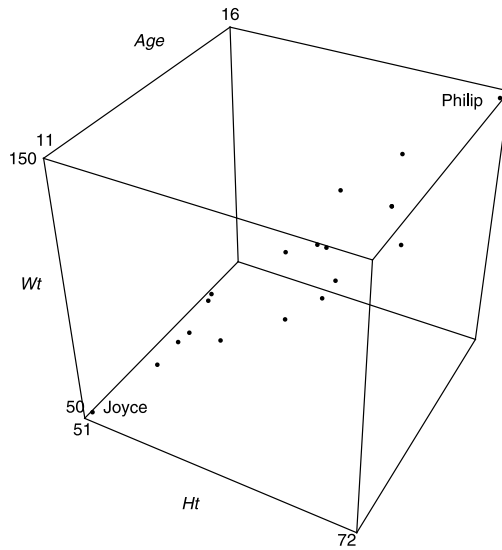
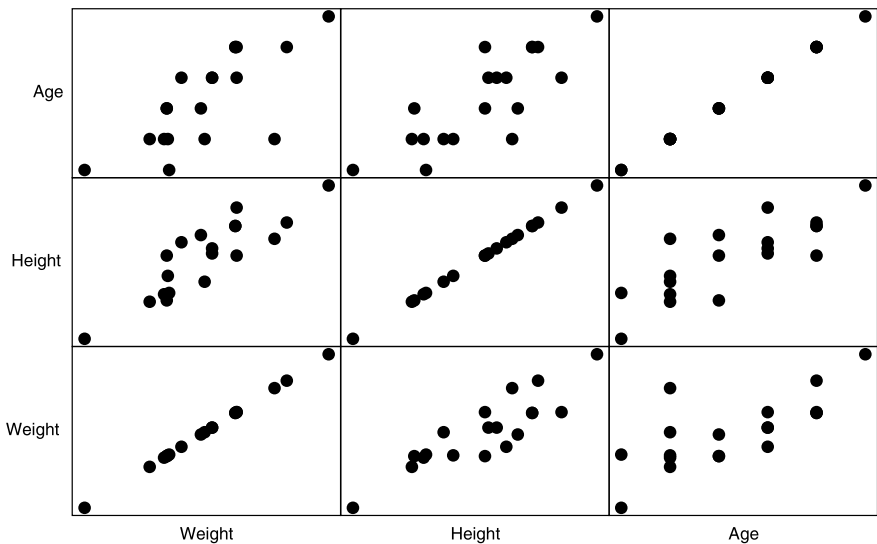


FIGURE 6.4 Matrix of two-dimensional scatterplot of the relationships among age, height, and weight



is easy to see that older, taller students tend to be heavier. Another common way to display data for multiple predictors, especially when more than two predictors precludes viewing the n -dimensional scatterplot, is a matrix of two-way scatterplots as depicted in Figure 6.4. In the last row we can see that taller students tend to be heavier and also that older students tend to be heavier. In the right-most scatterplot in the second row we can see that older students also tend to be taller; that is, height and age are redundant predictors. Because of the redundancy we cannot use the simple formulas. Any multiple regression program provides these estimates of the model parameters:

$$\hat{W}t_i = -141.22 + 1.28Age_i + 3.60Ht_i$$

But what exactly do these partial regression coefficients mean? We will use a series of simple models and regressions to help us understand the meaning of these coefficients.

A Model of Weight

We begin by considering a simple model for the weights of these students. That is:

$$Wt_i = \beta_0 + \varepsilon_i$$

Of course, the best estimate for this simple model is the mean, so:

$$\hat{W}t_i = \bar{W}t = 100.03$$

The weights of the 19 students, arrayed along the x-axis in the same order as they are listed in Figure 6.2, are displayed in the left panel of Figure 6.5. The horizontal line at 100.03 represents the simple model for these data. The points for Philip, the heaviest student, and Joyce, the lightest student, are again labeled as landmarks. The errors or residuals are easily computed as:

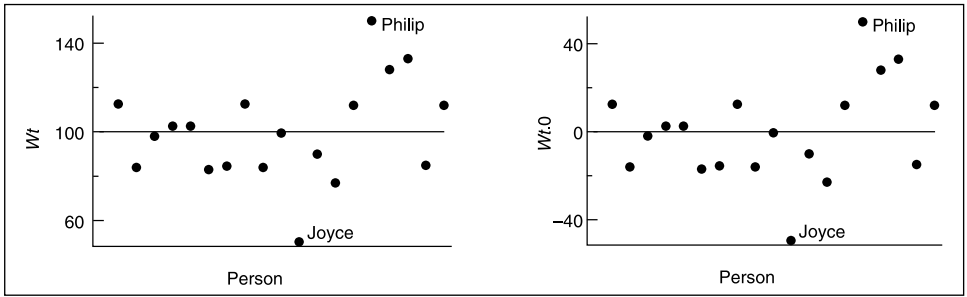
$$e_i = Wt_i - 100.03$$

For example, the errors for Philip and Joyce are, respectively, $151 - 100.03 = +50.97$ and $50.5 - 100.03 = -49.53$. We will be examining the errors for a number of different models for weight in this and the following sections so we introduce the notation $Wt.0$ to represent the errors from the simple model for weight. You can read “ $Wt.0$ ” as “the part of weight that remains after using a model with β_0 .” The right panel of Figure 6.5 displays these values of $Wt.0$. A positive value (i.e., those points above the horizontal line at zero) of $Wt.0$ indicates that the student is heavier than the average student, and a negative value (i.e., those points below the line) indicates that the student is lighter than the average student.

A Model of $Wt.0$ using Age

The values of $Wt.0$ in the right panel of Figure 6.5 represent when a student’s weight is unexpectedly or surprisingly heavy relative to our simple model of weight. If we are to improve on the simple model, then we need a variable that will help us predict when a student’s weight is unusually heavy or light amongst this group of students. The variable Age is an obvious candidate. For reasons of consistent and subsequent interpretation we will construct $Age.0$ in the same way we constructed $Wt.0$. That is, a simple model of age is:

FIGURE 6.5 A simple model for weight



$$Age_i = \beta_0 + \varepsilon_i$$

and the best estimate of the model is the mean, or:

$$\hat{Age}_i = \bar{Age} = 13.32$$

Then we can compute the errors or residuals as:

$$e_i = Age.0_i = Age_i - 13.32$$

For example, $Age.0$ for Philip and Joyce, respectively, are $16 - 13.32 = 2.68$ and $11 - 13.32 = -2.32$. The value of $Age.0$ for Philip means that he is 2.68 years older than the typical student in this group and the negative value for Joyce means that she is 2.32 years younger than the typical student.

It is now natural to ask whether a student who is older than average also tends to be heavier than average. That is equivalent to a model using $Age.0$ to predict $Wt.0$, or:

$$Wt.0_i = \beta_1 Age.0_i + \varepsilon_i$$

Note that we have omitted the intercept β_0 in this model. We know that the means of both $Wt.0$ and $Age.0$ must be zero. Given that the regression line must go through the point defined by the two means, the intercept is necessarily zero. Hence, there is no reason to include the intercept in the model. Figure 6.6 presents all the values of Wt , $Wt.0$, Age , and $Age.0$ for the 19 students. Our question is then whether a student being older than average predicts being heavier than average? In other words, does $Age.0$ predict $Wt.0$? Figure 6.7 provides the scatterplot for these two variables. Clearly, as $Age.0$ increases, so does $Wt.0$. We can use the formula for the slope from Chapter 5 or a simple regression program to estimate the model (we urge the reader to do the computations for this and subsequent simple regressions in this section using a hand calculator or using a computer program). The resulting model is:

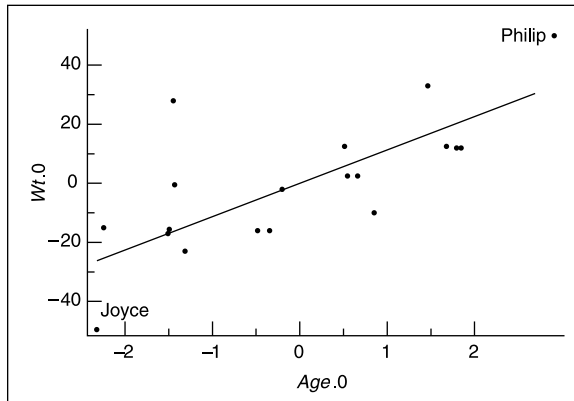
$$\hat{Wt.0} = 11.31 Age.0$$

In other words, for each year older (or younger) than average, we expect the student to weigh 11.31 pounds more (or less).

For someone like Philip who is 16, or 2.68 years older than the average student, we expect him to weigh, according to the model, $11.31(2.68) = 30.31$ pounds more than the average student. He in fact weighs 49.97 pounds more than the average student. Hence, the error after using both an intercept in the model and age as a predictor is $Wt.0, Age = 49.97 - 30.31 = 19.66$ pounds. Another way to say this is that Philip, “adjusting for”

FIGURE 6.6 Values of Wt , $Wt.0$, Age , and $Age.0$

Name	Wt	$Wt.0$	Age	$Age.0$
Alfred	112.5	12.47	14	0.68
Antonia	84.0	-16.03	13	-0.32
Barbara	98.0	-2.03	13	-0.32
Camella	102.5	2.47	14	0.68
Henry	102.5	2.47	14	0.68
Jamal	83.0	-17.03	12	-1.32
Jane	84.5	-15.53	12	-1.32
Janet	112.5	12.47	15	1.68
Jayden	84.0	-16.03	13	-0.32
John	99.5	-0.53	12	-1.32
Joyce	50.5	-49.53	11	-2.32
Judy	90.0	-10.03	14	0.68
Louise	77.0	-23.03	12	-1.32
Mary	112.0	11.97	15	1.68
Philip	150.0	49.97	16	2.68
Robert	128.0	27.97	12	-1.32
Sequan	133.0	32.97	15	1.68
Thomas	85.0	-15.03	11	-2.32
William	112.0	11.97	15	1.68
Mean	100.03	0	13.32	0

FIGURE 6.7 Relationship between $Age.0$ and $Wt.0$ 

or “controlling for” his age, is 19.66 pounds heavier than expected. Similarly, Joyce, who is 2.32 years younger than the average student, is expected to weigh $11.31(-2.32) = -26.24$ pounds lighter than the average student. However, she is in fact 49.53 pounds lighter so she weighs $Wt.0, Age = -49.53 - (-26.24) = -23.29$ pounds less than expected for her age. The values of $Wt.0, Age$ for all 19 students are displayed in Figure 6.8 (due to rounding errors in the hand calculations, the values are slightly different from those computed above). Note that although most of the errors become smaller (i.e., $Wt.0, Age < Wt.0$), the errors for a few students become larger. For example, although John’s weight is close to the mean weight (i.e., his $Wt.0 = -0.53$), he is 14.35 pounds heavier than we would expect based on his relatively young age of 12 (i.e., his $Wt.0, Age = 14.35$).

FIGURE 6.8 SSE for simple and conditional models

<i>Name</i>	<i>Wt</i>	<i>Wt.0</i>	<i>Wt.0,Age</i>	<i>Ht.0,Age</i>
Alfred	112.5	12.47	4.74	4.76
Antonia	84.0	-16.03	-12.46	-4.96
Barbara	98.0	-2.03	1.54	3.84
Camella	102.5	2.47	-5.26	-1.44
Henry	102.5	2.47	-5.26	-0.74
Jamal	83.0	-17.03	-2.15	-1.37
Jane	84.5	-15.53	-0.65	-1.13
Janet	112.5	12.47	-6.56	-4.53
Jayden	84.0	-16.03	-12.46	1.04
John	99.5	-0.53	14.35	0.33
Joyce	50.5	-49.53	-23.35	-4.58
Judy	90.0	-10.03	-17.76	0.06
Louise	77.0	-23.03	-8.15	-2.37
Mary	112.0	11.97	-7.06	-0.53
Philip	150.0	49.97	19.63	2.18
Robert	128.0	27.97	42.85	6.13
Sequan	133.0	32.97	13.94	-0.03
Thomas	85.0	-15.03	11.15	1.62
William	112.0	11.97	-7.06	-0.53
Mean	100.03	0	0	0
SS		9335.74	4211.25	
PRE			0.55	

On the whole, the errors are smaller so the sum of squared errors decreases by making predictions conditional on age. In Figure 6.8, the sum of squares for *Wt.0* (i.e., the SSE for a simple model making unconditional predictions) is 9335.74, but decreases to 4211.25 for *Wt.0,Age* (i.e., the SSE for a model making predictions conditional on age); the proportional reduction in error is 0.55. In other words, $PRE = .55$ for this model comparison:

$$\text{MODEL A: } Wt_i = \beta_0 + \beta_1 Age + \varepsilon_i$$

$$\text{MODEL C: } Wt_i = \beta_0 + \varepsilon_i$$

Verifying this value for PRE using a simple regression program is a useful exercise for the reader.

A Model of *Wt.0,Age* using *Ht.0,Age*

Why is, for example, Philip heavier than expected and Joyce lighter than expected for their age? Perhaps Philip is unusually tall and Joyce is unusually short for their age. We might be tempted simply to add Height to the model. However, our intuitions and visual inspection of the height by age scatterplot suggests that height and age are partially redundant. That is, height and age share some of the same information that might be used to predict weight. The only useful part of the height information is the part that is not redundant with age. We can easily find that part by using age to predict height in this model:

$$Ht = \beta_0 + \beta_1 Age + \varepsilon_i$$

and computing the residuals:

$$e_i = Ht_{i,0,Age} = Ht_i - \hat{H}t_i$$

A simple regression reveals that:

$$\hat{H}t_i = 25.22 + 2.79Age_i$$

In other words, we predict that these students grow about 2.79 inches each year. In particular, the model predicts Philip's height as $25.22 + 2.79(16) = 69.9$ inches, but he is actually 72 inches tall. Hence, he is indeed unusually tall for his age by $Ht_{i,0,Age} = 72 - 69.9 = 2.1$ inches. Because he is taller for his age, we might want to consider a model in which we adjust his predicted weight upward. Joyce, based on her age of 12, is expected to be $25.22 + 2.79(11) = 55.9$ inches tall. However, she is only 51.3 inches tall, or $Ht_{i,0,Age} = 51.3 - 55.9 = -4.6$ inches shorter than we expected for her age. And so perhaps her weight prediction should be adjusted downward. In general, we want to ask whether students who are taller for their age (or when "adjusting for" or "controlling for" age) tend to be heavier than expected for their age. In other words, we are asking whether there is a relationship between $Ht_{i,0,Age}$ (the part of height that is unrelated to age) and $Wt_{i,0,Age}$ (the part of weight that is unrelated to age). The values for $Ht_{i,0,Age}$ for all 19 students are listed in Figure 6.8. We want to consider the model:

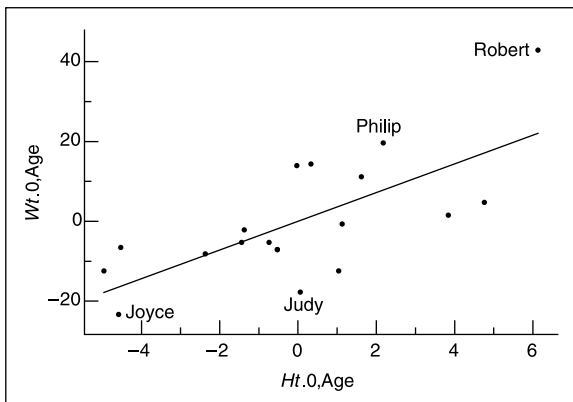
$$Wt_{i,0,Age} = \beta_1 Ht_{i,0,Age} + \varepsilon_i$$

Figure 6.9 shows the data to answer this question in a scatterplot. Indeed, those students who are taller than expected for their age ($Ht_{i,0,Age}$ is high) tend to be heavier than expected for their age ($Wt_{i,0,Age}$ is also high). The estimated model is:

$$\hat{W}t_{i,0,Age} = 3.6Ht_{i,0,Age}$$

As with any slope, this means that as $Ht_{i,0,Age}$ increases by one unit, the prediction of $Wt_{i,0,Age}$ increases by 3.6 pounds. But let us be more precise, even though it is a bit wordy: For each inch students are taller than expected for their age, we expect them to weigh an additional 3.6 pounds more than expected for their age. Another way to say this is that after adjusting height and weight for age, each additional inch predicts an increase of 3.6 pounds. Note that for those students for whom their heights were exactly

FIGURE 6.9 Relationship between $Ht_{i,0,Age}$ and $Wt_{i,0,Age}$



what we expected based on their age (i.e., $Ht.0, Age = 0$), we would expect them to weigh exactly what we expected based on their age (i.e., $\hat{Wt}.0, Age = 0$). Philip is 2.1 inches taller than expected for his age so we expect him to weigh $3.6(2.1) = 7.6$ pounds more than expected for his age, but he actually weighs 19.6 pounds more than expected for his age. Or, Philip weighs $Wt.0, Age, Ht = 19.6 - 7.6 = 12$ pounds more than we would expect based on a model using both age and height. Similarly, Joyce, 4.6 inches shorter than expected for her age, weighs $Wt.0, Age, Ht = -23 - 3.6(-4.6) = -23 - (-16.6) = -6.4$ pounds less than we would expect based on her age and height. When we just considered weight, Philip and Joyce were by far the heaviest and lightest students, respectively; naively we may have even thought that they were unhealthily overweight and underweight because they were so far from the average weight of these students. However, after accounting for (or adjusting for) their age and height, both Philip and particularly Joyce are close to the average weight expected. On the other hand, Robert and Judy, both marked in Figure 6.9, neither of whom appeared extreme when we only considered average weight, are the furthest from the weights we would expect given their heights and ages. Note that Robert is unusually tall for his age of 12, but he is still far heavier (by almost 21 pounds) than we would expect because of his extra height. Judy is almost exactly the height we would expect for her age of 14, but she is about 16 pounds lighter than we would expect for her combination of age and height. In other words, it is only after adjusting our weight predictions for age and height that we are able to see who is truly overweight (far above the regression line in Figure 6.9) or underweight (far below the regression line).

For each of the 19 students, Figure 6.10 lists their weights and their residuals from the various models we have considered so far. As a touchstone, let us consider the row

FIGURE 6.10 Students' weights and residuals for each of three models

Name	Wt	Wt.0	Wt.0, Age	Wt.0, Age, Ht
Alfred	112.5	12.47	4.74	-12.38
Antonia	84.0	-16.03	-12.46	5.38
Barbara	98.0	-2.03	1.54	-12.27
Camella	102.5	2.47	-5.26	-0.08
Henry	102.5	2.47	-5.26	-2.60
Jamal	83.0	-17.03	-2.15	2.78
Jane	84.5	-15.53	-0.65	-4.71
Janet	112.5	12.47	-6.56	9.73
Jayden	84.0	-16.03	-12.46	-16.20
John	99.5	-0.53	14.35	13.16
Joyce	50.5	-49.53	-23.35	-6.88
Judy	90.0	-10.03	-17.76	-17.98
Louise	77.0	-23.03	-8.15	0.37
Mary	112.0	11.97	-7.06	-5.15
Philip	150.0	49.97	19.63	11.79
Robert	128.0	27.97	42.85	20.80
Sequan	133.0	32.97	13.94	14.05
Thomas	85.0	-15.03	11.15	5.32
William	112.0	11.97	-7.06	-5.15
Mean	100.03	0	0	0
SS		9335.74	4211.25	2121.04
PRE			0.55	0.50

of values for Philip. His weight of 150 pounds is about 50 ($Wt.0$) pounds heavier than average, about 19.6 ($Wt.0, Age$) pounds heavier than average for students of his age, and only about 12 ($Wt.0, Age, Ht$) pounds heavier than we would expect for someone of his age and height. Thus, the deviation or error between Philip's actual weight (the data) and our model of weight steadily decreases from 50 to 20 to 12. On the whole, although there are exceptions (such as Judy), the errors decrease as more predictors are included in the model (i.e., $Wt.0, Age < Wt.0, Age, Ht$). Most importantly, the sum of squared errors steadily decreases. As we saw before, 55% of the squared error for the simple model is reduced when we add age. Now, 50% of the squared error for the model with age is reduced when we add the part of height not redundant with age. Overall, relative to the simple model of weight, using both age and height reduces error by 77%:

$$PRE = \frac{9335.74 - 2121.04}{9335.74} = .77$$

of the original squared error. We will return to these values of PRE for different model comparisons when we consider statistical inference for multiple regression models later in this chapter.

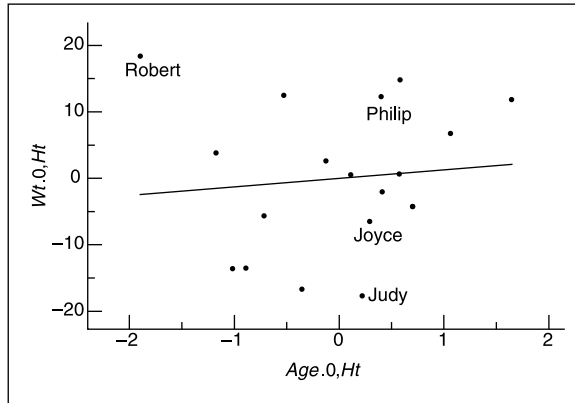
Conclusion: Interpretation of Partial Regression Coefficients

The goal of this series of simple regressions using residuals is to better understand regression coefficients. Now that we have a good understanding of the data from this detailed examination, it is time to return to the original goal. Earlier we noted that the estimate of the model from multiple regression is:

$$\hat{W}t_i = -141.22 + 1.28Age_i + 3.60Ht_i$$

The coefficient of 3.6 for Ht is *exactly* the same as the slope in Figure 6.9 where it is the coefficient predicting $Wt.0, Age$ from $Ht.0, Age$. Therefore, the meaning of the 3.6 in the multiple regression equation is exactly the same as above: for each inch taller students are for their age, we expect them to be 3.6 pounds heavier for their age. Or we might say, after adjusting both height and weight for age, for each inch taller, a student is expected to weigh 3.6 pounds more. In other words, the coefficient of 3.6 in the multiple regression model describes the relationship between that part of height that is not redundant or shared with age and that part of weight that is not redundant or shared with age. It is common in reports of multiple regression to write something simple like: "When controlling for age, weight increases 3.6 pounds for each additional inch of height." However, it is important to remember the more precise meaning of the regression coefficient that we developed in the previous section and to remember that it is the slope in the scatterplot of two residuals in Figure 6.9. Statisticians and statistics programs often refer to plots like Figure 6.9 as *partial regression plots* because of their basis for defining the partial regression slope or coefficient. The wise data analyst will routinely examine partial regression plots, whose many benefits include (a) a visual representation of the regression coefficient and (b) visual identification of any unusual patterns or observations (such as Robert and Judy).

To interpret the coefficient of 1.28 for age in the multiple regression equation, we do not need to do the complete series of simple regressions parallel to those that we did

FIGURE 6.11 Partial regression plot between $Age.0,Ht$ and $Wt.0,Ht$ 

above. Instead, we can reverse the role of height and age in the above statements interpreting the coefficient for height and we can examine the computer-generated partial regression plot in Figure 6.11 relating $Age.0,Ht$ to $Wt.0,Ht$. Thus, for each year older students are than we would guess from their heights, we expect them to be 1.28 pounds heavier for their height. Or we might say, after adjusting both age and weight for height, for each year older, a student is expected to weigh 1.28 pounds more. Still other language would be: After statistically equating two students on height or “holding height constant,” we expect the student older by one year to weigh only 1.28 pounds more than the younger student. In that context, 1.28 pounds does not seem like much; indeed, the slope in Figure 6.11 appears as if it might not differ significantly from a slope of zero (we will check this later).

Note that, as must be the case, Robert and Judy have the largest residuals (i.e., greatest distance from the regression lines in both of the partial regression plots) because they are unexpectedly heavy and light, respectively, given their heights and ages. And again, Philip and Joyce are closer to their expected weights (i.e., closer to the regression lines) based on their heights and ages. In other words, $Wt.0,Age,Ht = Wt.0,Ht,Age$. Although overall predictions of weight have improved, considerable error remains. We might now ask whether there is a variable that could predict when students are unusually heavy (or light) for their heights and ages. Looking at the names above the regression line (i.e., unusually heavy for their height and ages) and those below (i.e., unusually light for their height and ages) suggests an obvious variable to consider.

STATISTICAL INFERENCE IN MULTIPLE REGRESSION

Now that we understand the meaning of the partial regression coefficients, the remaining issue is how to determine whether those coefficients are significantly different from zero and to ask other model comparison questions involving models with multiple predictors. The general strategy for doing statistical inference or asking questions about data is exactly the same as before. In particular, the calculations of PRE and F pose no new problems as they are defined exactly as before. The only change is that the extra parameters in the multiple regression equation give us lots of freedom in defining Model

A and Model C. As we shall see, different choices allow us to ask different questions and test different hypotheses. The only difficulty then is selecting the appropriate compact and augmented models to test a specific hypothesis. Once we have the appropriate models, estimation of the parameters (using a computer statistical package) and calculations of PRE and F are straightforward. Each question will have its own PRE and F .

For any set of data there will be many possible questions. There are several generic types of questions that occur frequently in multiple regression models. We consider each generic type separately.

Testing an Overall Model

Sometimes we want to ask whether our predictors *as a group* are any better than the simple model, which uses no predictors and predicts the mean for every observation. For example, college admissions offices have the practical and interesting question of whether student information such as high school percentile rank and Scholastic Aptitude Tests (SAT), both verbal and quantitative, predict performance in college. To answer that question, we could compare a model that made predictions of first-year cumulative grade point average conditional on high school rank, SAT verbal, and SAT quantitative to a simple model that made a single unconditional prediction of grade point average using the mean.

In general, we are comparing the following Models C and A:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

To be a test of the overall model, Model C must be the simple model with only one parameter. The null hypothesis being tested—that Model C is sufficient—is equivalent to saying that it is not useful to make the predictions conditional on all of the $p - 1$ predictor variables as a group. Formally:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

Model C has one parameter and Model A has p parameters—one for each of the $p - 1$ predictors plus one parameter for the intercept (β_0). The appropriate sums of squared errors are given by:

$$\text{SSE(C)} = \Sigma(Y_i - \bar{Y})^2 \quad \text{and} \quad \text{SSE(A)} = \Sigma(Y_i - \hat{Y}_{Ai})^2$$

where:

$$\hat{Y}_{Ai} = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_{p-1} X_{i,p-1}$$

And the difference between these two, SSR, is given by:

$$\text{SSR} = \text{SSE(C)} - \text{SSE(A)} = \Sigma(\hat{Y}_{Ai} - \bar{Y})^2 = \Sigma(\hat{Y}_{Ai} - \bar{Y})^2$$

The PRE for testing the overall model relative to the simple mean model (Model A versus Model C above) has the special name “coefficient of multiple determination” and is usually represented by the symbol R^2 . All multiple regression computer programs, which estimate the parameters, will also provide R^2 directly so it is seldom necessary

to calculate PRE by hand for the overall model. However, it is exceedingly important to remember that R^2 is nothing more than PRE for this special case of testing an overall model against the simple mean model. Also, for many other questions we will want to ask about our data, the computer programs will not provide the relevant PRE for the model comparison of interest; hence, it is important to know how to calculate PRE directly from $SSE(A)$ and $SSE(C)$.

The calculation of F for this case is straightforward once $PRE = R^2$ is known; that is:

$$F = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} = \frac{R^2/(p - 1)}{(1 - R^2)/(n - p)}$$

Alternatively, in terms of the sums of squares:

$$F = \frac{SSR/(PA - PC)}{SSE(A)/(n - PA)}$$

If we make the normal assumptions that the errors ε_i have mean zero, are from one normal distribution, and are independent, then we can compare the F we calculated to the sampling distribution for the F statistic with $p - 1$ and $n - p$ degrees of freedom. Most computer programs that report R^2 will also report F . Many will also report the “significance of F ,” which is simply the probability that if the null hypothesis (Model C) were true a value of F this large or larger would be found. If this probability is lower than our desired level for α (the Type I error probability) then we can reject the null hypothesis that none of the predictors are useful.

The probability tables for PRE and F that we developed in Chapter 4 are for the special case when the number of parameters in Models A and C differ by exactly one (i.e., $PA - PC = 1$). However, Model A might have many more parameters than the simple mean model, which serves as Model C. We therefore need new tables. The logic of developing these tables is exactly the same as for the tables developed in Chapter 4. In the Appendix, Figures A.2 and A.4 give the critical values for PRE and F , respectively, for $\alpha = .05$ for selected values of $PA - PC$.

An example of an overall model test

An example will help to make the statistical inference process for the overall model less abstract. Returning to our earlier question, college admissions offices want to know whether student information such as high school percentile rank and Scholastic Aptitude Tests (SAT), both verbal and quantitative, predict performance in college. In this example, we have the first-year cumulative grade point average (GPA), high school rank (HSRank), SAT verbal (SATV), and SAT math (SATM) scores for 414 engineering students at a state university. We want to know whether we can use those variables to make useful conditional predictions of each student’s GPA or whether we would do just as well if we simply used the group mean as an unconditional prediction of everyone’s GPA. In other words, does Model A, which uses all the predictor variables, do better than Model C, which uses none of the predictors? In terms of equations we are comparing:

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \varepsilon_i$$

For this comparison, $PA = 4$ and $PC = 1$. Model A and Model C would be the same if β_1 , β_2 , and β_3 all equal zero. Thus, comparing Models C and A is equivalent to testing the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Computer programs provide these estimated models and associated SSEs:

$$\begin{aligned} \text{MODEL A: } \hat{GPA}_i &= -1.74 + 0.027HSR\text{rank}_i + 0.011SATV_i + 0.022SATM_i \\ \text{SSE(A)} &= 172.99 \end{aligned}$$

$$\begin{aligned} \text{MODEL C: } \hat{GPA}_i &= 2.70 \\ \text{SSE(C)} &= 221.78 \end{aligned}$$

Although the program producing these estimates also provided $PRE = R^2 = .22$ and $F = 38.54$ with $p < .0001$, we will calculate them to reinforce the basic process that applies to all model comparisons involving multiple predictors:

$$PRE = \frac{SSE(C) - SSE(A)}{SSE(C)} = \frac{221.78 - 172.99}{221.78} = \frac{48.79}{221.78} = .22$$

Of course, 48.79, the difference between the two sums of squared errors, is equivalent to SSR. We can now compute:

$$\begin{aligned} F_{PA - PC, n - PA} &= F_{3, 410} = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} = \frac{.22/3}{.78/410} = \frac{.073}{.0019} = 38.54 \\ &= \frac{SSR/(PA - PC)}{SSE(A)/(n - PA)} = \frac{48.79/3}{172.99/410} = \frac{16.26}{0.422} = 38.54 \end{aligned}$$

In other words, the proportional reduction in error per additional parameter in the overall model of .073 exceeds by a factor of 38.5 the proportional reduction in error per parameter of .0019 that we would expect for a useless predictor. Consulting the tables in the Appendix for $\alpha = .05$, $PA - PC = 3$, and $n - PA = 200$ (the largest value for $n - PA$ less than 410 available in the table), we find that the PRE and F for the data easily exceed the critical values of .038 and 2.65, respectively. Hence, we reject Model C in favor of Model A and conclude that the overall model is indeed significantly better than the simple mean model. Equivalently, we reject the null hypothesis that all the predictor coefficients equal zero.

Problems with overall model tests

On the whole, overall model tests such as this one are not particularly useful for two reasons. First, it is easy when testing a large overall model to be in the position of losing a needle in a haystack. For example, suppose we were testing a large model in which only one predictor was actually useful; the other predictors have all been added inadvertently and they are of no value in reducing error. We would get approximately the same value of PRE for this example whether we were testing the overall model or the simple regression model in which we used only the one good predictor. However, PRE per parameter (the numerator of F) would be much reduced for the overall model

because we would be dividing approximately the same PRE by a larger number of extra, unnecessary parameters. Simultaneously, the remaining error ($1 - \text{PRE}$) per unused potential parameter (the denominator of F) would be larger for the overall model with the extra useless predictors because we would be dividing approximately the same residual error by a smaller number of potential unused parameters ($n - p$ instead of $n - 2$). Both of these effects will cause F to be considerably smaller than it otherwise would have been without the extra useless predictors. From the tables in the Appendix we see that the magnitude of the critical value of F decreases as the numerator degrees of freedom increase, so F may sometimes still be significant. However, in general, we risk losing the needle—the one good predictor—in a haystack of useless predictors. The essence of the problem is that by throwing away parameters for useless predictors, we lose statistical power.

A second reason for avoiding overall tests of models is that the results are often ambiguous. If the null hypothesis is rejected, then we only know that at least one of the partial regression coefficients is not equal to zero, but we do not know which predictor or predictors are useful. For example, for the GPA data we concluded that β_1 , β_2 , and β_3 are not all equal to zero, but we do not know precisely which ones are not equal to zero. There are seven different ways in which the overall null hypothesis could be false:

$$\beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$$

$$\beta_1 \neq 0, \beta_2 \neq 0$$

$$\beta_1 \neq 0, \beta_3 \neq 0$$

$$\beta_2 \neq 0, \beta_3 \neq 0$$

$$\beta_1 \neq 0$$

$$\beta_2 \neq 0$$

$$\beta_3 \neq 0$$

That is, HSRank, SATV, and SATM as a group are useful predictors of GPA, but we do not know whether all are needed or what combination is best. For instance, maybe SATV and SATM are highly redundant, so only one SAT score needs to be included in the regression equation.

The question we ask with the test of the overall model is so diffuse that we are generally unsure what the answer means. For the reasons of power and removing ambiguity discussed above, it is almost always better to ask more specific or, in the words of Rosenthal and Rosnow (1985), more *focused* questions of our data. Focused questions will generally ask about a single predictor. That does not mean that we will ignore all the other predictors, because, as we shall soon see, we can ask about the usefulness of a particular predictor in the context of all the other predictors.

So for reasons of statistical power and specificity, we prefer null hypotheses that pertain to a single parameter, i.e., where PA and PC differ by 1. The F for testing such null hypotheses will always have one degree of freedom in the numerator. If the compact model representing the null hypothesis is rejected, then the answer is unambiguous—the extra parameter in Model A is required. Next we consider an important one-degree-of-freedom test.

One Predictor Over and Above the Other Predictors

We know that redundancy complicates questions about individual predictors. Regardless of whether a predictor is useful by itself in a simple regression, we often want to know whether it is useful over and above other predictors in the model. In other words, if we have already made conditional predictions based on a set of predictors, is there further improvement by adding one more predictor variable? In other words, we are comparing:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \beta_p X_{ip} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

That is, Model C and Model A are the same except that Model A is augmented by exactly one extra parameter (β_p) and predictor variable (X_p). Thus Model C has p parameters, and Model A has $p + 1$ parameters. Most importantly, then, $PA - PC = 1$.

By comparing Model A with Model C, we are asking whether we need the predictor X_p given that we are already using predictors X_1, X_2, \dots, X_{p-1} . The β values in this model comparison are partial regression coefficients in that their values depend on what other β values are included in the Model. That is, the estimated value of, say, β_2 is likely to depend on whether or not β_1 is also included in the model. Thus, we should use the more complete notation that reminds us of this dependence. In particular, the extra coefficient added to Model A is represented by $\beta_{p.12\dots p-1}$. The subscript before the dot indicates the particular coefficient, and the subscripts after the dot indicate the other coefficients also included in the Model. Clearly, we do not need the extra predictor if its partial regression coefficient equals zero, so our null hypothesis is:

$$H_0: \beta_{p.12\dots p-1} = 0$$

Note that this is merely a generalization of the test that we did for the simple regression model with one predictor. There we asked whether adding a predictor was useful over and above the simple mean model. Here we ask whether adding a predictor is useful over and above the other predictors already in the model.

One strategy for model building is to start simple and add complexity only when necessary. Thus, a frequent question is whether the addition of another predictor would be worthwhile. As always, we answer the question about whether the extra parameter is worthwhile by estimating the model parameters so as to minimize the sum of squared errors and then using PRE and F to compare the two models. The PRE for the addition of exactly one parameter has the special name “coefficient of partial determination” and is represented as:

$$r_{p.123\dots p-1}^2$$

The square root of this PRE with the sign from the partial regression coefficient is often called the “partial correlation coefficient” because it is the simple correlation between Y and X_p when controlling or holding constant the other $p - 1$ predictors.

Given F , we can compute a confidence interval for $\beta_{p.12\dots p-1}$ using the same logic we used to construct a confidence interval for β_1 in the case of simple regression in Chapter 5. That is, for a specific set of data we can ask how extreme a null hypothesis could be so that F would still be less than the critical value of F for a given α under the usual assumptions. The resulting equation is:

$$b_{p.123\dots p-1} \pm \sqrt{\frac{F_{\text{crit}; 1, n-p-1; \alpha} \text{MSE}}{(n-1)s_{X_p}^2(1-R_p^2)}}$$

As before, MSE is the mean squared error for Model A, $F_{\text{crit}; 1, n-p-1; \alpha}$ is the appropriate critical value of F from the tables in the Appendix, and $s_{X_p}^2$ is the variance of the predictor variable. For whatever level of α is chosen, the above provides the $(1 - \alpha)\%$ confidence interval for the partial regression slope. The formula for the confidence interval is the same as the one for a simple regression slope from Chapter 5 except for the addition of the term $1 - R_p^2$ in the denominator. R_p^2 is simply the R^2 or PRE when all the other $p - 1$ predictors are used to predict X_p . Thus, R_p^2 is a measure of the redundancy between X_p and the other predictors already included in the model.

Conversely, the term $1 - R_p^2$, which has the special name of *tolerance* in some computer outputs, is a measure of X_p 's uniqueness from the other predictors. Only the unique part of X_p might be useful for reducing the error further; if the tolerance $1 - R_p^2$ is low, then it will be difficult for X_p to be useful. This is reflected in the formula for the confidence interval by including it in the denominator. A low tolerance in the denominator will cause the confidence interval to be large, and a large confidence interval will be more likely to include zero. Because a low tolerance makes the confidence interval wider, some programs report the variance inflation factor (VIF), which equals $1/(1 - R_p^2)$, the inverse of tolerance.

There is an exact equivalence between using F , compared with a critical value of F for a certain α , to test the addition of a parameter and whether or not the confidence interval constructed using that same critical value of F includes zero. The confidence interval excludes zero if and only if F exceeds the critical F .

It is seldom necessary to calculate by hand the F for the addition of a parameter because most regression programs will routinely produce F (or sometimes the equivalent $t_{n-p-1} = \sqrt{F_{1, n-p-1}}$). Many programs will also give the 95% confidence interval (corresponding to $\alpha = .05$) for $\beta_{p.12\dots p-1}$. Few programs will directly give the corresponding PRE, which in this case is the proportional reduction in error due to adding the p th predictor. However, almost all programs will give the partial correlation coefficient, which for this model can be squared to yield PRE. The important things to remember therefore are not the computational formulas but that the coefficient of partial determination or, equivalently, the squared partial correlation, is simply PRE for the special case of comparing a compact model with an augmented model that has one extra predictor. Similarly, the associated F or t tests whether the partial regression coefficient for X_p is zero. Thus, the following are equivalent null hypotheses:

$$H_0 : \beta_{p.12\dots p-1} = 0 \quad \text{and} \quad H_0 : r_{p.12\dots p-1}^2 = 0$$

If one is false, then the other must be false also.

An example of an additional predictor test

We will again use the GPA data to illustrate the test for adding an additional predictor to the model. Suppose that we want to ask whether it is useful to add SATM to the model when HSRank and SATV are already in the model. This question not only has practical importance—it would be easier for the clerks in the admissions office if they only had to include HSRank and SATV in their calculations of predicted GPA—but

it also is substantively interesting to know whether quantitative skills as measured by SATM make a contribution to the predicted GPA that is independent of HSRank and SATV. These particular data are for engineering students, so one would suppose that SATM as a measure of quantitative skills that engineers would need in their courses would be required even when HSRank and SATV as measures of overall ability were already included in a model. In other words, we want to compare these two models:

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \varepsilon_i$$

Note that PA = 4 and PC = 3, so PA – PC = 1; we therefore are asking a focused, one-degree-of-freedom question. Model A and Model C would be the same if β_3 equaled zero. (Note, we no longer use more complete $\beta_{3,12}$, with the implicit understanding that any parameter depends on the model in which it appears.) Thus, the Model A/Model C comparison is equivalent to testing the null hypothesis:

$$H_0: \beta_3 = 0$$

Model A here is the same as the Model A we used when testing the overall model. Thus, we already have the parameter estimates and SSE for Model A. A multiple regression program provides the parameter estimates and SSE for Model C. So:

$$\begin{aligned} \text{MODEL A: } \hat{GPA}_i &= -1.74 + 0.027 HSRank_i + 0.011 SATV_i + 0.022 SATM_i \\ \text{SSE(A)} &= 172.99 \end{aligned}$$

$$\begin{aligned} \text{MODEL C: } \hat{GPA}_i &= -0.71 + 0.028 HSRank_i + 0.018 SATV_i \\ \text{SSE(C)} &= 181.48 \end{aligned}$$

To answer the question of whether SATM is useful over and above HSRank and SATV, we compute PRE and F in the usual way. That is:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{181.48 - 172.99}{181.48} = \frac{8.49}{181.48} = .047$$

and

$$\begin{aligned} F_{PA - PC, n - PA} = F_{1, 410} &= \frac{\text{PRE}/(PA - PC)}{(1 - \text{PRE})/(n - PA)} = \frac{.047/1}{.953/410} = \frac{.047}{.0023} = 20.4 \\ &= \frac{\text{SSR}/(PA - PC)}{\text{SSE(A)}/(n - PA)} = \frac{8.49/1}{172.99/410} = \frac{8.49}{0.422} = 20.4 \end{aligned}$$

PRE = 0.047 and $F_{1,410} = 20.4$ easily exceed their respective critical values (for $n - PA = 200$, the largest value less than 410 in the appendix tables) of 0.019 and 3.89. We therefore can reject the null hypothesis that the partial regression coefficient for SATM is zero, and we can conclude that it is worthwhile to add SATM to the model that already includes HSRank and SATV. In this case, the remaining error—the error left after using HSRank and SATV in the model—is reduced by 4.7%. The square root of PRE equals 0.22 and represents the partial correlation between GPA and SATM holding constant or “partialing out” HSRank and SATV. Knowing that the partial regression coefficient for SATM is not zero, we should therefore interpret it: When comparing students of

equivalent HSRank and SATV scores, for each point higher on SATM (measured on a 20–80 scale as SAT scores are reported to universities), we predict that a student's GPA will be 0.022 points higher. In other words, all else being equal, we would expect a student who scored 80 (800 as reported to the student) on SATM to have a GPA about two-thirds of a letter grade higher than a student who scored 50 (i.e., $(80 - 50) \cdot 0.022 = 0.66$).

The ingredients for computing the confidence interval are available from various sources: $MSE = SSE(A)/(n - PA) = 172.99/410 = 0.42$, a descriptive statistics program gives the variance of SATM as 46.21, and a multiple regression program gives the PRE or $R_{3,12}^2$ for predicting SATM with HSRank and SATV as 0.11. Hence, the 95% confidence interval for β_3 is computed as:

$$b_{SATM} \pm \sqrt{\frac{F_{crit; 1, n-PA, \alpha} MSE}{(n-1)s_{SATM}^2 (1-R_{3,12}^2)}}$$

$$0.022 \pm \sqrt{\frac{3.87(0.42)}{413(46.21)(1-0.11)}}$$

$$0.022 \pm 0.0098$$

$$[0.012, 0.032]$$

Hence, although our best estimate of the coefficient is 0.022, it might reasonably be as low as 0.012 or as high as 0.032.

The Overall Model Revisited

We noted earlier that a single test of the overall model was usually unsatisfactory because rejection of the null hypothesis does not indicate which of the many β_j significantly depart from zero. The test for adding another variable, which we have just considered, suggests a solution. The numbering of the predictor variables is arbitrary, so we can treat each predictor in turn as if it were the last variable to be added. In this way we will test separately for each predictor whether its partial regression coefficient (i.e., its coefficient with all the other variables in the model) is different from zero. Conceptually, this implies that we need to do separate regression analyses in which we first enter the other predictor variables and then test whether it is worthwhile to add the last. In practice, it is not necessary to do the separate regression analyses for each variable as the last because the required information is available when the statistical program does the computations for the overall model. The results of these computations are often summarized in an analysis of variance table like Figure 6.12.

We will examine how this table summarizes the statistical tests we have done of (a) the overall model and (b) adding SATM over and above the other variables in the model. Then we will consider how it provides information for the tests of adding either SATV or HSRank last. In our test of the overall model we found:

$$\text{MODEL A: } \hat{GPA}_i = -1.74 + 0.027 \text{ HSRank}_i + 0.011 \text{ SATV}_i + 0.022 \text{ SATM}_i$$

$$\text{SSE(A)} = 172.99$$

$$\text{MODEL C: } \hat{GPA}_i = 2.70$$

$$\text{SSE(C)} = 221.78$$

The two sums of squared errors, $SSE(A) = 172.99$ and $SSE(C) = 221.78$, correspond, respectively, to the values in the *SS* column for Error and Total. Their difference is $SSR = 221.78 - 172.99 = 48.79$, which we used in the computation of PRE:

$$PRE = \frac{SSE(C) - SSE(A)}{SSE(C)} = \frac{221.78 - 172.99}{221.78} = \frac{48.79}{221.78} = .22$$

and which is listed in the *SS* column for Regression in Figure 6.12. The values in the *df* column for Regression, Error, and Total are, respectively, $PA - PC$, $n - PA$, and $n - PC$. The other values in the Regression row summarize the test of the overall model against the simple model. Note that the rows for Regression, Error, and Total in Figure 6.12 have the same role as they did in Figure 5.11 for simple regression. The only difference is that the Regression row now describes the reduction in error due to allowing all of the partial regression slopes to differ from zero and the Error row is the remaining error in that case. The Total row describes the error when all the partial regression slopes are set to zero.

Unlike the ANOVA table for simple regression (Figure 5.11), the one for multiple regression (Figure 6.12) has a row for each variable in the model. To understand this row, we return to our earlier test of adding SATM to a model already making conditional predictions of GPA based on HSRank and SATV. Previously, we found:

$$\begin{aligned} \text{MODEL A: } \hat{GPA}_i &= -1.74 + 0.027 \text{ HSRank}_i + 0.011 \text{ SATV}_i + 0.022 \text{ SATM}_i \\ \text{SSE(A)} &= 172.99 \end{aligned}$$

$$\begin{aligned} \text{MODEL C: } \hat{GPA}_i &= -0.71 + 0.028 \text{ HSRank}_i + 0.018 \text{ SATV}_i \\ \text{SSE(C)} &= 181.48 \end{aligned}$$

The difference between these two sums of squared errors, $181.48 - 172.99 = 8.48$, appears in the *SS* column for SATM. The other columns in the row for SATM provide the statistics for comparing the above Models A and C. We computed PRE directly from these numbers before, but it is convenient when working with summary tables like Figure 6.12 to use this equivalent alternate formula:

$$PRE = \frac{SSR}{SSR + SSE(A)} = \frac{8.483}{8.483 + 172.994} = \frac{8.483}{181.477} = .047$$

One way to think about this computation is that if SATM were not in the model, then the error it reduces would need to be put back into the model error—the computation in the denominator. In other words, the denominator corresponds to the total error for a model that included all the other variables except SATM.

Once we have this understanding of the analysis of variance summary table, it is easy to do other model comparisons. For example, we might want to know, for these engineering students, whether knowing their SATV scores improves predictions of GPA over and above a model based only on HSRank and SATM. In other words, our model comparison is

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 \text{ HSRank}_i + \beta_2 \text{ SATV}_i + \beta_3 \text{ SATM}_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \beta_1 \text{ HSRank}_i + \beta_3 \text{ SATM}_i + \varepsilon_i$$

FIGURE 6.12 Analysis of variance summary table for the GPA data

Source	SS	df	MS	F	PRE	p
Regression	48.789	3	16.263	38.5	.220	.0001
SATV	2.779	1	2.779	6.6	.016	.0106
SATM	8.483	1	8.483	20.1	.047	.0001
HSRank	29.209	1	29.209	69.2	.144	.0001
Error	172.994	410	0.422			
Total	221.783	413				

Instead of fitting these two models and computing their sum of squared errors we can obtain these values from Figure 6.12. Model A is as before, so we know $SSE(A) = 172.994$ from the Error row in the table. The value of $SS = 2.779$ for SATV corresponds to SSR for the above model comparison. Hence, $SSE(C) = 2.779 + 172.994 = 175.773$. We then obtain the PRE for this model comparison as:

$$PRE = \frac{SSR}{SSR + SSE(A)} = \frac{2.779}{2.779 + 172.994} = \frac{2.779}{175.773} = .016$$

And we can compute F for this model comparison either using the formula based on PRE or the sums of square formula (from the source table):

$$F_{PA-PC, n-PA} = F_{1,410} = \frac{SSR/(PA-PC)}{SSE(A)/(n-PA)} = \frac{MSR}{MSE} = \frac{2.779}{0.422} = 6.6$$

The key values for MSR and MSE are available in the MS column of Figure 6.12. Computer programs for regression analysis often provide the probability values listed in the last column. In this case, the probability of obtaining a PRE or F this large or larger if Model C were correct is only 0.016. Using the usual cutoff of 0.05, we would reject Model C in favor of Model A; in other words, we would conclude that SATV is a useful predictor over and above HSRank and SATM.

Now that we understand the information provided in the summary table, we can readily make the model comparison asking whether information about HSRank is useful over and above the predictive information provided by SATV and SATM. When comparing:

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

we see that $SSR = 22.209$, $PRE = .144$, and $F_{1,410} = 69.2$ with $p < .0001$. Clearly, HSRank significantly reduces error in predicting GPA over and above the error reduction provided by using SATV and SATM. We might say that performance information (HSRank) provided useful information beyond that provided by the aptitude information (SAT scores).

It is important to note in the ANOVA table of Figure 6.12 that when we add the SSRs for the individual variables, the sum ($2.779 + 8.483 + 29.209 = 40.471$) is less than the total SSR we obtain when using all the predictor variables together (48.789). The reason for this is the redundancy among the predictors. The SSRs when adding each

predictor last captures only the *unique* contribution to the error reduction made by that predictor *over and above* the contribution from the other predictors. Hence, a contribution that is shared with other predictors appears in the total SSR but not in the unique SSR added by that variable. Had there been no redundancy among the predictors, then the sum of their SSRs would have equaled the SSR for the overall model. With certain patterns of positive and negative correlations among predictors, it is even possible for the sum of the individual SSRs to exceed the total SSR, a condition sometimes confusingly labeled as “suppression.” Whether the sum of the SSRs is smaller or larger than the overall SSR is not particularly important. Simply interpret the individual SSRs as the unique contribution to the error reduction of each variable and the overall SSR as the total reduction in error when using all the variables together in a model.

Testing Other Model Questions

So far we have asked questions about the overall model and about each variable if it were added last to the model. And we saw how these tests are summarized in the ANOVA table of Figure 6.12. However, this does not begin to exhaust the interesting questions we might ask. Another common question is to ask whether it is useful to add a set of variables to the model. For example, we observed above that performance information (HSRank) added useful prediction over and above aptitude information (test scores). If we want to ask the corresponding question of whether, as a set, aptitude information provides useful prediction over and above performance information, then we would compare these two models:

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \beta_1 HSRank_i + \varepsilon_i$$

In this case, the null hypothesis is that the parameters for *both* SATV and SATM are zero. That is:

$$H_0: \beta_2 = \beta_3 = 0$$

To answer this question we need only find the appropriate SSEs. Using the most basic multiple regression program, we can simply estimate both models, obtain their SSEs, and do the computations of PRE and F manually. More advanced multiple regression programs usually have options for testing sets of parameters directly. We will use the basic approach:

$$\begin{aligned} \text{MODEL A: } \hat{GPA}_i &= -1.74 + 0.027 HSRank_i + 0.011 SATV_i + 0.022 SATM_i \\ \text{SSE(A)} &= 172.99 \end{aligned}$$

$$\begin{aligned} \text{MODEL C: } \hat{GPA}_i &= 0.18 + 0.028 HSRank_i \\ \text{SSE(C)} &= 189.11 \end{aligned}$$

$$\text{PRE} = \frac{189.11 - 172.99}{189.11} = \frac{16.121}{189.11} = .085 \quad F_{2,410} = \frac{.085/2}{(1 - .085)/410} = 19.04$$

The values for PRE and F easily exceed their critical values so we conclude that, as a set, the test scores SATV and SATM provide significant error reduction over and above that for HSRank alone.

As another example of the kind of questions we might ask beyond the obvious ones provided by most multiple regression programs, we might ask whether SATV and SATM need to be weighted differently. Or, in other words, we are asking whether the observed weights of 0.011 for SATV and 0.022 for SATM in the overall model are significantly different from each other. For predicting first-year GPA for engineering students, we might suspect that more weight should be given to SATM. Answering this question implies this model comparison:

$$\text{MODEL A: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_3 SATM_i + \varepsilon_i$$

$$\text{MODEL C: } GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 SATV_i + \beta_2 SATM_i + \varepsilon_i$$

Look closely to see that the only difference in the two models is that the same coefficient (β_2) is used for *both* SATV and SATM in Model C but separate coefficients (β_2 and β_3) are used in Model A. Note that this meets our rule for Model A/Model C comparisons because the parameters used in Model C are a subset of those used in Model A, even though both models have the same variables. In this case, the null hypothesis is:

$$H_0 : \beta_2 = \beta_3$$

Unlike the prior test where we asked whether both coefficients were zero, here we ask only if they are equal, but not necessarily zero. Another way to express this null hypothesis is that their difference is zero; that is:

$$H_0 : \beta_2 - \beta_3 = 0$$

Some sophisticated multiple regression programs allow specification of complex null hypotheses like this. If not, then some ingenuity is required to obtain the necessary SSEs in order to compare the models. In this case we observe with a simple algebraic rearrangement that Model C is equivalent to:

$$GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 (SATV_i + SATM_i) + \varepsilon_i$$

$$GPA_i = \beta_0 + \beta_1 HSRank_i + \beta_2 (SATTOT_i) + \varepsilon_i$$

where SATTOT = SATV + SATM. Hence, we can get the SSE(C) by regressing GPA on HSRank and SATTOT:

$$\begin{aligned} \text{MODEL A: } \hat{GPA}_i &= -1.74 + 0.027HSRank_i + 0.011SATV_i + 0.022SATM_i \\ \text{SSE(A)} &= 172.99 \end{aligned}$$

$$\begin{aligned} \text{MODEL C: } \hat{GPA}_i &= -1.63 + 0.027HSRank_i + 0.016SATV_i + 0.016SATM_i \\ \text{SSE(C)} &= 173.86 \end{aligned}$$

$$\text{PRE} = \frac{173.86 - 172.99}{173.86} = \frac{0.87}{173.86} = .005 \quad F_{1,410} = \frac{.005}{(1 - .005)/410} = 2.06$$

Even for $n - PA = 500$, these values do not exceed the respective critical values of .008 and 3.86 for $\alpha = .05$. Hence, there is not sufficient evidence in these data to conclude that SATV and SATM ought to be weighted differently when predicting first-year GPA for engineering students.

These are but two of the many specialized, focused questions we might want to ask about our data that are beyond the obvious questions answered by the usual output from

multiple regression programs. Other questions that might be of interest would be asking whether the new model is significantly better than the old model that might have been used in the decision to admit these students. Or we might want to ask questions about the mean GPA of these students, in which case we would convert our variables to mean deviation form and do model comparisons parallel to those in Chapter 5 for the intercept. The important point is that you can answer any question about your data by framing it in terms of a Model A/Model C comparison, obtaining the SSEs for those models, and then computing PRE and F .

CAUTIONS ABOUT USING MULTIPLE REGRESSION

Multiple regression is likely the most widely used statistical procedure in the social sciences. As a consequence, it is also the most frequently abused procedure. In this section we comment on some important cautions about using and especially interpreting multiple regression.

Causal Conclusions from Nonexperiments

We interpret the partial regression coefficient b_j as the amount the predicted value \hat{Y}_i changes for each unit increase in the predictor variable, controlling for all other variables in the model. It is important to remember that this is only a prediction based on the existing data and may not represent a causal relationship between X_j and Y . Users of multiple regression sometimes forget this fact and incorrectly presume that manipulating X_j will *cause* a change in Y .

The following example highlights the causal ambiguity inherent in observational data. In almost any city with seasonal temperature variation it is easy to demonstrate that the number of ice cream cones sold in a month is a good predictor of the number of suicides in that month: the more cones sold the more suicides there will be. However, it would be ludicrous to suggest that by banning ice cream sales suicides would be reduced. The problem with such a conclusion is that ice cream is not causing suicides; rather, it is probably just a proxy variable for temperature, which might be the cause of both increased suicides and increased ice cream sales. And even temperature is probably just a proxy variable for discomfort or unpleasantness that might send a depressed person over the brink.

The causal error is obvious in the ice cream example. But many researchers do not realize that the dangers of presuming that manipulation of the predictor variable will cause a change in Y are no less serious whenever the regression analysis is based on observational data.

On the other hand, it is also incorrect to assume that causal conclusions can never be based on regression analyses. If the researcher determines the values of the predictor variables and randomly assigns those values to the cases, then it is quite proper to presume that changing the predictor variables will cause a change in Y . Such a design, involving random assignment to levels of a predictor, is commonly known as an experimental design. For example, to test the sensitivity of credit card use to the interest rate charged, a bank might randomly assign, say, one of five different interest rates to different customers. If a regression analysis revealed that the interest rate was a significant

predictor of the amount that customers charged to their credit cards, then one would reasonably presume that the manipulation of the interest rate by the bank causally affected the amount that people charged to their cards.

There are some other research designs, originally called *quasi-experiments* by Campbell and Stanley (1963), in which limited causal conclusions from regression analyses of observational data are warranted. For further treatment of these issues, see Shadish, Cook, and Campbell (2002) and Judd and Kenny (1981a). In general, however, one should be extremely cautious about making causal conclusions from regression analyses of nonexperimental data.

Relative Importance of Predictor Variables

Many researchers are tempted to make inferences about the relative importance of predictor variables in multiple regression models. It should be obvious that it is not reasonable to compare directly the relative magnitudes of b_j because they clearly depend on the unit of measurement. By making the unit of measurement for a given predictor variable smaller (e.g., using centimeters instead of inches to measure height), we can make b_j arbitrarily large without changing the fundamental impact of that variable in the model.

Sometimes predictor variables will be measured on the same scale, enabling the same units of measurement for the variables. For example, consider trying to model attitudes toward a particular political party as a function of economic status. It might be reasonable to use both income and savings as predictor variables. We can easily measure both variables in dollars, so that we could be assured of using the same units. However, even here comparing b_j values as a measure of importance would probably not be appropriate, because the ranges of the variables would be so different. One variable might have a smaller b_j —that is, a smaller predicted change in Y for a unit increase in X_j —but still have a larger overall impact because it had a larger range over which it could change.

The dependence of the overall effect of a predictor variable on its range suggests that we might normalize the partial regression coefficients by multiplying them by their ranges or, more typically, their respective standard deviations. This effectively makes the unit of measurement either the range or the standard deviation, respectively. The *standardized regression coefficients* are also traditionally normalized to include the standard deviation of Y according to the formula:

$$b_j^* = b_j \left[\frac{S_{X_j}}{S_Y} \right]$$

Some books and computer programs unfortunately use β_j or “beta weights” to refer to standardized regression coefficients. We prefer b_j^* to avoid confusion with the unknown values of the model coefficients in our models and to emphasize that standardized regression coefficients are simple transformations of the partial regression coefficients b_j .

Unfortunately, the standardized regression coefficients do not solve the problem of making inferences about relative importance. First, they still depend on the particular range and distribution of the predictor in the particular dataset. If this distribution were to change for any reason—for example, if the researcher were to include a wider range of cases—then the standardized regression coefficients could change dramatically even

if the underlying relationship between the predictor and the data remained the same (i.e., the partial regression coefficient is unchanged). Such dependence on the vagaries of a particular set of observations hardly provides a good basis for inferring relative importance. Second, relative importance is also clouded by redundancy. A regression coefficient (standardized or not) might be low not because its associated predictor variable has no relationship with Y , but because that predictor variable is at least partially redundant with the other predictors. Thus, standardized regression coefficients do not really allow us, as they were intended, to make relative comparisons of predictors. The fundamental problem is that we cannot frame a question about relative importance in terms of a Model A versus Model C comparison because both models would necessarily have the same predictors. Hence, our machinery for statistical inference cannot address the question of relative importance.

In sum, relative importance of predictor variables is a slippery concept. Any statement about relative importance must be accompanied by many caveats that it applies to this particular range of variables in this particular situation. As a result they are, in our opinion, seldom useful. Thus, although it is tempting to compare the importance of predictors in multiple regression, it is almost always best not to do so.

Automatic Model Building

Many regression programs have procedures for automatically building models. The most commonly used such procedure is *stepwise regression*, which, as its name implies, builds a model step by step. On each step a predictor variable may be added or deleted from the model. Usually the criterion for adding or deleting a variable is defined in terms of F or PRE when considering the variable as the last one in the model. On the first step the variable with the largest F is added (so long as it exceeds a preset minimum threshold). On subsequent steps, the variable not in the model that has the largest F is added to the model *or* a variable in the model whose F has fallen below the threshold is removed. The latter can happen because of redundancy. A predictor variable that had a high F when it was with few other variables in the model can have a much lower F on later steps after a number of variables, with which it is partially redundant, have been added. The stepwise procedure terminates when all the variables have F values above the threshold and all the variables not in the model have F values below the threshold. Many programs for stepwise regression allow for fine-tuning of the threshold parameters. Common variations are *forward selection*—start with the simple model and add variables in the order of their usefulness until none exceeds the criterion—or *backward selection*—start with all the variables in the model and remove variables in decreasing order of usefulness until all variables remaining in the model exceed the criterion.

Stepwise regression and its variants present the alluring promise of finding the “best” model with virtually no effort by the researcher or data analyst. However, the allure is deceptive because it can be shown that for redundant predictors the stepwise procedure will not always find the best model. For example, a stepwise procedure may be fooled by the redundancy and stop, say, with four predictor variables in the model. However, when all possible four-variable regressions are computed, another model may be found with a higher overall PRE. Stepwise regression developed at a time when computer computations for a large-scale regression analysis were expensive. With contemporary computer speeds and better algorithms it is now feasible to estimate regression models

with all possible subsets of variables. Hence, if one is going to build models automatically, it is better to use all-subset regressions instead of the stepwise procedure to avoid the latter's tendency to be misled by redundancy. However, one then has the problem of deciding between models of different size that do not have a Model C/Model A relationship because one model will not necessarily be a subset of the other. In such cases, the model with the highest adjusted R^2 (the unbiased estimated of η^2) is selected or the related Mallows's C_p criterion is used.

We do not recommend the use of any of the automatic model-building procedures for three reasons. First, an unfocused search through many possible models (sometimes pejoratively referred to as a "fishing expedition") increases the likelihood of capitalizing on chance (i.e., making Type I errors) and thereby finding a model that represents only a spurious relationship. In an exercise that we urge readers to try themselves, we generated 51 random variables with 100 observations for each. We randomly designated one of those variables as Y and then used stepwise regression to build a model with the remaining 50 variables as potential predictor variables. The estimated final model was:

$$\hat{Y}_i = 0.03 + 0.25X_{i,11} - 0.29X_{i,21} - 0.35X_{i,23} + 0.26X_{i,31} - 0.21X_{i,44}$$

with the test of the overall model against the simple model, which was the true model, having $PRE = .31$, $F_{5,94} = 8.49$, $p < .0001$. All five partial regression coefficients are significant, with the minimum being $PRE = .04$, $F_{1,94} = 5.78$, $p = 0.018$. Hence, the stepwise regression procedure produced six Type I errors. Given that there are about 50 statistical tests at each step and there is ample opportunity to make Type I errors, it is not surprising that six actual Type I errors were made.

Second, as we learned in this chapter, the interpretation of the coefficients and the meaning of the questions being asked depend on what other variables are included in the model. For example, testing whether there is a relationship between age and weight is different from testing whether there is a relationship between age and weight after controlling for or adjusting for any height differences. It seems unwise to let an automatic algorithm determine the questions we do and do not ask of our data. Some day there may indeed be artificial intelligence procedures in statistical packages, but the automatic model-building procedures are not them.

Third, it is our experience and strong belief that better models and a better understanding of one's data result from focused data analysis, guided by substantive theory. For an interesting illustration of the superiority of substantively guided data analysis over automatic model building, see Henderson and Velleman (1981). They state a fundamental principle of their philosophy of data analysis, a principle with which we agree: "The data analyst knows more than the computer." Failure to use that knowledge produces inadequate and often flawed data analyses.

STATISTICAL POWER IN MULTIPLE REGRESSION

The discussion of statistical power in the context of simple regression, using a single predictor variable, generalizes quite readily to the multiple regression case, with multiple predictors. If we have an estimate of the true proportional reduction in error, η^2 , for whatever model comparison is of interest, then we can use a software program, such as R or SAS, to estimate the statistical power for a given number of observations.

The usual sources of estimates of the true proportional reduction in error are appropriate here: past studies or experience in a research area; Cohen's small, medium, and large effects; or estimates computed from expected parameter values.

The interesting complexities in discussing power in the case of multiple predictors arise from the presence of the other predictors in the model. As we have already discussed, the denominator of the confidence interval for the coefficient for a particular predictor includes the tolerance of that predictor. As that tolerance decreases, the confidence interval for its regression coefficient increases in width. Accordingly, as the tolerance of a predictor decreases (i.e., as it becomes more redundant with the other predictors in the model), tests of its regression coefficient become less powerful.

Other predictors may also enhance the power of tests of a regression coefficient in multiple regression. Recall that in the numerator of the confidence interval is the mean square error (MSE) for the augmented model. As that mean square error decreases in size (i.e., as the predictors in the augmented model do a better job of predicting Y_i), the width of the confidence interval decreases. Factors that reduce the mean square error thus increase the power of tests of the regression coefficients included in the model. The power of any model comparison question is enhanced whenever nonredundant and error-reducing predictor variables are added to both the augmented and compact models. When the added variables are redundant with the key variables of the question there is a tradeoff: the gain in power from reducing MSE may be offset by a reduction in power due to redundancy. It is difficult to estimate this tradeoff in power. However, the implication is that to improve power one should try to add significant, nonredundant predictors to the model, assuming that their presence makes substantive sense. Of course, the presence of predictors in the compact and augmented models that do not significantly relate to Y will result in less power, because the degrees of freedom for error for augmented and compact models will be reduced, thus increasing MSE.

SUMMARY

In this chapter we have considered models with multiple continuous predictors, traditionally known as multiple regression models. Such models permit complex questions about the simultaneous roles of multiple, partially redundant predictors in explaining variation in the dependent variable. The major issue that complicates interpretation of parameter estimates in such models is that of partial redundancy among predictors. Accordingly, we spent a considerable amount of time developing notions of what it means to interpret a partial regression coefficient and, more particularly, what it means to examine the effect of one predictor controlling for others that are included in a model.

Our model comparison approach to statistical inference in such models is identical to the approach of earlier chapters, where we compare an augmented model to a compact one. In the context of multiple-predictor models, the number of possible model comparisons that might be of interest is considerably larger than in the case of simpler models considered in earlier chapters. We illustrated a number of possible model comparisons that might be of interest, many of which are typically conducted by standard multiple regression computer programs. The most frequently reported of these are tests of the overall model and tests of individual partial regression coefficients. However, there are many other possibilities that are not routinely output in standard regression

programs. The advantage of the model comparison approach is that whenever one is dealing with augmented and compact models, these can be separately estimated and then their error sums of squares compared. The range of questions we can thus ask of our data extends considerably beyond those that the computer programmers thought we would be interested in when they wrote the standard multiple regression programs.

Moderated and Nonlinear Regression Models

7

In Chapter 6 we considered in some detail models with multiple predictor variables. In this chapter we extend these models by discussing the addition of predictors that are themselves products of other predictors. Such models, as we will show, permit us to extend dramatically the range of questions that we can ask of our data. In the models of Chapter 6, our questions about individual predictor variables had the following generic form: “Is one predictor useful over and above or controlling for the set of other predictors in the model?” In this chapter we will ask a set of rather different questions about predictors. The questions considered in this chapter take the following generic form: “Does the utility of one variable as a predictor depend on the value of another variable or even of that particular variable itself?”

At this point, it may not be clear why this second sort of generic question is likely to be of interest. It probably is not even clear exactly what this second question means. To clarify things, we start this chapter with an extended example, using new data. We first recapitulate some of the major points of Chapter 6, in a somewhat different language, once again clarifying the sorts of questions about models that we asked there. We then show how the questions we will ask of our data with the models developed in this chapter differ from and extend the questions we asked earlier. Our hope is that this initial section will give an understanding of the second generic question mentioned above and will further the understanding of the sorts of questions addressed earlier in Chapter 6.

The models we consider in this chapter go by a variety of names. Occasionally they are known as *moderator regression* models. They are also known as *polynomial*, *interactive*, *nonadditive*, or *nonlinear* models. Regardless of the name we give them, the goal in this chapter is to discuss the estimation, interpretation, and tests of models that include, as predictors, products of other predictors.

Let us turn to the data that we will use to illustrate things. These are hypothetical data, supposedly gathered from 80 individuals who are all runners. Three variables have been measured. The first variable (Time) is each runner’s most recent 5-kilometer race time, measured in minutes. Next we have each runner’s age (Age). The final variable is the average number of miles per week that each runner runs when in training for a race (Miles). The data are presented in Figure 7.1.

SIMPLE AND ADDITIVE MODELS PREDICTING RACE TIMES

Our interest in these data will focus on how the age of the runner and the amount of training relate to race times in the 5-kilometer race. Let us start by examining how these

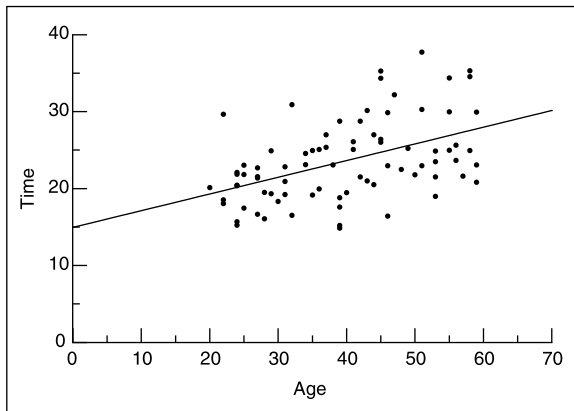
FIGURE 7.1 Runners' data

<i>Time</i>	<i>Age</i>	<i>Miles</i>	<i>Time</i>	<i>Age</i>	<i>Miles</i>
24.91	29	10	29.68	22	10
21.82	25	20	20.36	24	30
21.54	27	40	18.07	22	40
23.03	25	50	28.78	39	10
25.35	37	20	18.80	39	30
22.84	31	40	14.87	39	50
30.16	43	10	32.19	47	20
27.00	44	20	19.46	40	30
16.42	46	40	35.32	58	10
24.87	53	20	20.83	59	30
24.95	58	40	22.69	27	20
18.32	30	30	15.69	24	40
16.66	27	50	30.89	32	10
25.08	36	20	23.11	34	30
16.53	32	40	19.17	35	50
35.29	45	10	26.09	41	20
26.41	45	30	21.81	50	40
22.48	48	50	37.75	51	10
34.38	55	10	30.30	51	20
19.00	53	30	22.95	51	40
21.86	24	10	20.14	20	20
22.11	24	30	15.23	24	40
17.45	25	40	18.53	22	50
24.58	34	20	24.96	35	30
19.97	36	40	19.21	31	50
34.34	45	10	25.07	41	20
25.22	49	30	20.99	43	40
20.52	44	50	29.94	59	10
29.98	55	20	23.08	59	30
21.62	57	40	21.52	53	50
19.35	29	20	21.38	27	30
20.45	24	40	19.52	28	50
16.07	28	50	27.00	37	10
23.06	38	20	17.61	39	30
20.94	31	40	15.19	39	50
28.77	42	10	26.03	45	20
29.86	46	30	21.51	42	40
22.96	46	50	34.53	58	10
25.62	56	20	23.68	56	30
24.98	55	40	23.50	53	50

two variables relate to race time individually, using the single-predictor models developed in Chapter 5. If we regress Time on Age, the following simple regression equation results:

$$\widehat{\text{Time}} = 14.955 + .217\text{Age}$$

The sum of squared errors for this model equals 1714.720. The sum of squared errors for a single-parameter compact model in which we predict the mean time for each case equals 2211.091. Testing the null hypothesis to determine whether Age is a significant

FIGURE 7.2 Time–Age simple relationship

predictor of race time yields a value of PRE of .224, which converts to an F of 22.58 with 1 and 78 degrees of freedom. These values indicate that indeed Age is a significant predictor of Time.

This simple regression result is graphed in Figure 7.2. As the positive slope makes clear in the graph, our best prediction based on these data is that race times increase with age. Since this is a simple regression model and no other predictor variables have been included, we are unable to answer questions about what other variables may be related to Time or what would happen to the Time–Age relationship if we controlled for some other variable, such as Miles. It may be that Age and Miles are redundant or confounded in these data, such that older runners train with fewer miles. If this were the case, then if we included Miles in the model, we might reach rather different conclusions about how Age was related to Time once Miles was controlled.

The simple relationship between Time and Miles is captured by the least-squares simple regression equation:

$$\widehat{\text{Time}} = 31.911 - .280\text{Miles}$$

The sum of squared errors from this model equals 1029.013. To determine whether Miles and Time are reliably related, we want to compare the sum of squared errors from this model with the sum of squared errors from the compact single-parameter model. This compact model is the same one we used to test whether Age is reliably related to Time. Its sum of squared errors equals 2211.091. The value of PRE that we get when comparing these two models equals .535, which converts to an F statistic of 89.60 with 1 and 78 degrees of freedom. Again, this value of F is significant, and we conclude that Miles and Time are related. The simple relationship is graphed in Figure 7.3.

When we regress Time on both Miles and Age, the following multiple regression equation results:

$$\widehat{\text{Time}} = 24.605 + .167\text{Age} - .257\text{Miles}$$

The sum of squared errors for this three-parameter model equals 741.195, which gives a PRE of .665 when compared to the single-parameter compact model that includes

neither predictor. The overall F statistic to compare this three-parameter model with the single-parameter model equals 76.35 with 2 and 77 degrees of freedom.

We can also compare this three-parameter model with each of the simple regression models to determine whether each predictor variable significantly increases the accuracy of our predictions when controlling for the other predictor variable. The value of PRE when comparing compact and augmented models without and with Age equals .280, which converts to an F statistic of 29.92 having 1 and 77 degrees of freedom. The value of PRE when comparing compact and augmented models without and with Miles equals .568, which converts to an F statistic of 101.20 with 1 and 77 degrees of freedom. Hence we conclude that each of these predictor variables significantly increases the accuracy of our predictions of Time even when the other predictor is controlled. The source table in Figure 7.4 summarizes the results of the inference tests we have just conducted using this multiple regression model.

It is instructive to graph this multiple regression equation much like we did in the case of the simple equations. In Figure 7.5, we have graphed the prediction plane that this model generates in a three-dimensional space, where the vertical axis is Time, and Miles and Age are the two horizontal axes. As this figure makes clear, a specific predicted value of Time is made by the model at each pair of values for Miles and Age. The individual data points could also be included in this graph, and errors of prediction would be simply the vertical distances between each data point and the plane of prediction.

FIGURE 7.3 Time–Miles simple relationship

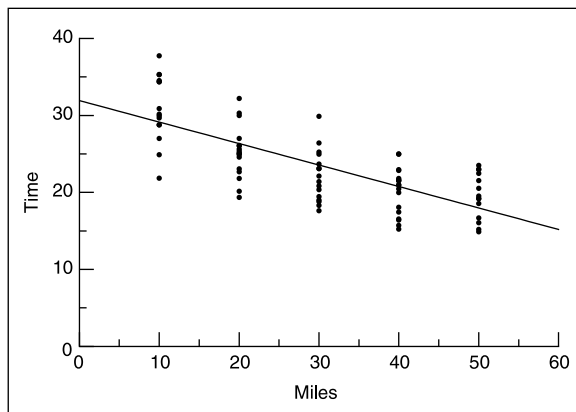
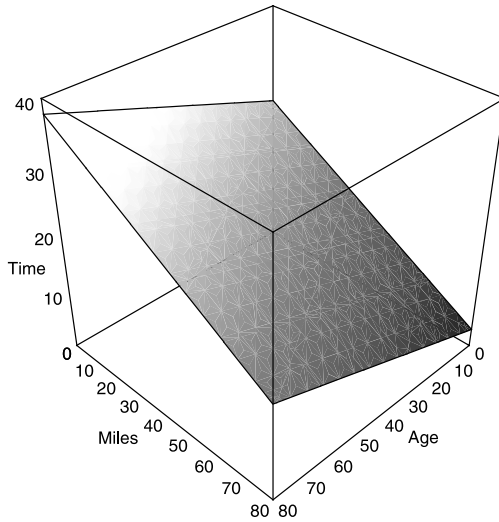


FIGURE 7.4 Multiple regression source table

Source	SS	df	MS	F	PRE	p
Model	1469.90	2	734.95	76.35	.665	<.001
Age	288.13	1	288.13	29.90	.280	<.001
Miles	974.56	1	974.56	101.20	.568	<.001
Error	741.19	77	9.63			

FIGURE 7.5 Prediction plane plot from multiple regression predicting Time from Age and Miles



“Simple” Re-expressions of the Additive Model

To make the presentation of this model a little simpler, we could also take slices of this prediction plane, treating Age, for instance, as the predictor on the horizontal axis and graphing Time–Age prediction lines at different levels of Miles. If we group terms in the multiple regression equation a little differently, we get the re-expression:

$$\widehat{\text{Time}} = (24.605 - .257\text{Miles}) + .165\text{Age}$$

We can think of this re-expression as telling us about the “simple” relationship between Time and Age, allowing the intercept in the simple regression to change as Miles changes. In other words, suppose we wanted to predict race time conditional on age for someone who trained 20 miles per week. We would substitute 20 for Miles in the above re-expression and get:

$$\begin{aligned}\widehat{\text{Time}} &= (24.605 - .257(20)) + .165\text{Age} \\ &= 19.465 + .165\text{Age}\end{aligned}$$

For a runner who trained 30 miles per week, our “simple” predicted relationship between Time and Age would be:

$$\begin{aligned}\widehat{\text{Time}} &= (24.605 - .257(30)) + .165\text{Age} \\ &= 16.895 + .165\text{Age}\end{aligned}$$

We call these re-expressions “simple” relationships, meaning that we are acting as if we are examining a “simple” regression model between Age and Time at particular values of the other variable, Miles.

If we graphed the Time–Age relationship from this multiple regression equation where we also use Miles as a predictor, we would get many different “simple” regression

prediction functions between Time and Age, one for each of the possible levels of Miles. A few such lines are graphed in Figure 7.6. Importantly, these “simple” relationships all have the same slopes: they are parallel lines that differ only in their intercepts. Accordingly, this multiple regression equation suggests that the slope or “simple effect” of Age on Time is the same regardless of the level of Miles. As Age increases, we predict the same change in Time, regardless of the level of Miles.

Since Miles is in theory a continuous variable, our plot of the Time–Age relationship within levels of Miles contains an infinite number of lines, one for each of the possible levels of Miles. All of these “simple” prediction lines combine to form the prediction plane of Figure 7.5.

Now let us turn the equation around and graph the Time–Miles “simple” relationship at different Age levels. Re-expressing our multiple regression equation, we get:

$$\widehat{\text{Time}} = (24.605 + .165\text{Age}) - .257\text{Miles}$$

For representative values of Age, we have graphed this function in Figure 7.7, putting Miles on the horizontal axis this time. Again, we see that as Age changes, the intercept for the Time–Miles “simple” relationship changes but the slope does not. In other words,

FIGURE 7.6 Time–Age “simple” relationships

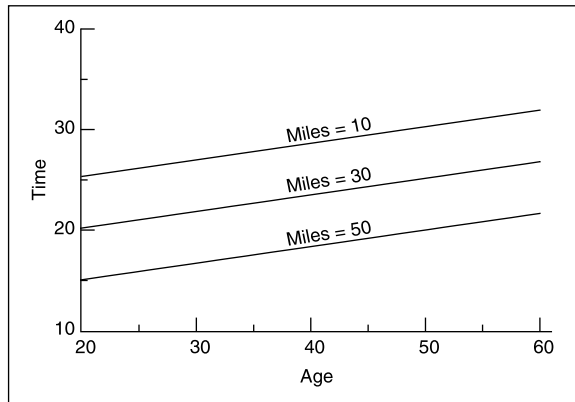
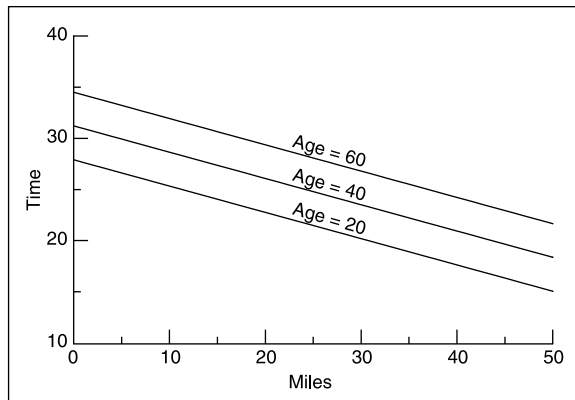


FIGURE 7.7 Time–Miles “simple” relationships



the “simple effect” of Miles on Time (i.e., its slope) is assumed to remain the same regardless of Age.

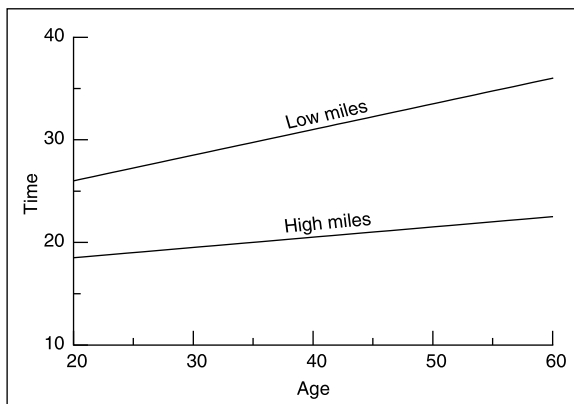
INTERACTIONS BETWEEN PREDICTOR VARIABLES

The assumption that we are making in this multiple regression equation, and that these graphs make clear, is known as the *additive* assumption. In slightly more abstract language, if we regress some variable Y_i on predictors X_{1i} and X_{2i} , we are assuming that the contribution of each predictor to the best predicted values, \hat{Y}_i , does not depend on the value of the other predictor. The two contributions are thus additive: regardless of the level of X_{1i} , the relationship (i.e., slope) between Y_i and X_{2i} does not change. Likewise, regardless of the level of X_{2i} , the relationship between Y_i and X_{1i} is presumed not to change. We can simply “add” up the “effects” of two predictors in deriving \hat{Y}_i : In figuring out what the contribution of each predictor variable is to these \hat{Y}_i values, we do not need to worry about the level or value of the other predictor.

How reasonable is this additive assumption? The answer is that its reasonableness depends on the substantive domain under examination. For the variables at hand, it may not be a very good assumption. For instance, it seems reasonable to suggest that the age difference in race times might be greater when dealing with people who do not run very much than when dealing with people who run a lot. To put this hypothesis another way, training a lot may reduce the relationship between Age and Time, such that differences in race times associated with age may be smaller with more training. This hypothesis is graphed in Figure 7.8. Notice that the slope of the prediction function linking Age and Time is hypothesized to be steeper when Miles is relatively low than when it is relatively high. In other words, the relationship between Age and Time is hypothesized to be stronger when Miles is low than when it is high.

We are suggesting that rather than Age and Miles combining *additively* in giving us our best predictions of Time, they may *interact* in affecting time. An *interaction* between two variables, say X_{1i} and X_{2i} , is said to occur whenever the slope or the “simple effect” of one of them, say X_{1i} , on the dependent variable Y_i varies with or depends on the value of the other variable, X_{2i} . In the case at hand, we are hypothesizing that Age

FIGURE 7.8 Hypothesized “simple” Time–Age relationships



and Miles interact in affecting Time. This means that we suspect that the relationship between Age and Time depends on or varies with the number of miles run.

Whether or not there is an interaction between two variables in predicting a third is an issue that is totally independent of whether or not the two predictor variables are correlated or redundant with each other. To suggest that Age and Miles may interact in predicting Time is to suggest that the “simple” relationship between Age and Time may depend on or vary with Miles. This may be true regardless of whether Age and Miles are themselves correlated or redundant predictor variables.

To test this hypothesis about the interaction between Age and Miles, we want to compare our three-parameter additive model with one in which the “simple” slope of one predictor variable is allowed to change as the other predictor variable changes. We saw earlier that the simple additive model allows the “simple” intercept of one predictor variable to change as the other predictor variable changes. However, the “simple” slope does not. The “simple” slope of Age, for instance, was invariant across changes in Miles. To allow for changing slopes, we need to build a model that allows both the intercept and the slope in the “simple” re-expressions to vary. In the following, we show how this can be done.

First, consider the additive model:

$$\widehat{\text{Time}} = b_0 + b_1\text{Age} + b_2\text{Miles}$$

and its two “simple” re-expressions, the first focusing on the Time–Age “simple” relationship at different levels of Miles and the second focusing on the Time–Miles “simple” relationship at different levels of Age:

$$\begin{aligned}\widehat{\text{Time}} &= (b_0 + b_2\text{Miles}) + b_1\text{Age} \\ \widehat{\text{Time}} &= (b_0 + b_1\text{Age}) + b_2\text{Miles}\end{aligned}$$

Now we would like to modify these “simple” expressions so that not only the “simple” intercepts vary as a function of the other variable, but so do the slopes. In other words, we want the “simple” re-expressions to have the following forms:

$$\begin{aligned}\widehat{\text{Time}} &= (b_0 + b_2\text{Miles}) + (b_1 + b_3\text{Miles})\text{Age} \\ \widehat{\text{Time}} &= (b_0 + b_1\text{Age}) + (b_2 + b_3\text{Age})\text{Miles}\end{aligned}$$

In the first of these two, examining the “simple” relationship between Time and Age at different levels of Miles, the “simple” intercept $(b_0 + b_2\text{Miles})$, varies as a function of Miles, and now so does the “simple” slope $(b_1 + b_3\text{Miles})$. That is, as Miles changes value, the “simple” slope for Age varies, to the extent that b_3 differs from zero. Likewise, in the second of the above “simple” expressions, focusing on the Time–Miles “simple” relationship, both the “simple” intercept $(b_0 + b_1\text{Age})$ and the “simple” slope $(b_2 + b_3\text{Age})$, are allowed to vary as a function of Age.

These two “simple” expressions are actually symmetrical and can be equivalently written as:

$$\widehat{\text{Time}} = b_0 + b_1\text{Age} + b_2\text{Miles} + b_3(\text{Age} \times \text{Miles})$$

This then indicates how we go about estimating models in which we allow for interactions, that is, models in which the “simple” slope for each variable is allowed to vary as a function of the value of the other variable. We simply form a new variable that is the product of the two predictors and estimate a model that includes this product variable as an additional predictor, in addition to the two component variables as predictors themselves. And the symmetry of these two “simple” effect expressions and their equivalence to the model that includes the product predictor means that interactions can be interpreted in either of two ways. If the product is a useful predictor, over and above its component variables, then that can be interpreted either as “the simple slope for Age depends on the value of Miles” or as “the simple slope for Miles depends on the value of Age.”

Accordingly, in the following model:

$$\widehat{\text{Time}} = b_0 + b_1\text{Age} + b_2\text{Miles} + b_3(\text{Age} \times \text{Miles})$$

two different, but equally correct, interpretations of the various regression coefficients are possible, depending on which “simple” relationship we focus on. If we think about this model in the context of the “simple” Time–Age relationship:

$$\widehat{\text{Time}} = (b_0 + b_2\text{Miles}) + (b_1 + b_3\text{Miles}) \text{Age}$$

then the following interpretations follow:

- b_0 : the intercept for the “simple” Time–Age relationship when Miles equals zero.
- b_1 : the slope for the “simple” Time–Age relationship when Miles equals zero.
- b_2 : the change in the intercept for the “simple” Time–Age relationship as Miles increases by one unit.
- b_3 : the change in the slope for the “simple” Time–Age relationship as Miles increases by one unit.

On the other hand, if we think about this model in terms of the “simple” Time–Miles relationship:

$$\widehat{\text{Time}} = (b_0 + b_1\text{Age}) + (b_2 + b_3\text{Age}) \text{Miles}$$

then the following equivalent interpretations follow:

- b_0 : the intercept for the “simple” Time–Miles relationship when Age equals zero.
- b_1 : the change in the intercept for the “simple” Time–Miles relationship as Age increases by one unit.
- b_2 : the slope for the “simple” Time–Miles relationship when Age equals zero.
- b_3 : the change in the slope for the “simple” Time–Miles relationship as Age increases by one unit.

Although these two sets of interpretations sound rather different, in fact they are fundamentally equivalent. In any specific context, it is likely that one set of interpretations, and accordingly one “simple” relationship, will be more compelling than the alternative. But, in a formal sense, both sets remain equally viable regardless of such preferences.

Estimation and Interpretation of the Interactive Model

Let us now illustrate this interactive model, and the resulting possible interpretations, with the data at hand. We have already estimated the three parameters of the linear additive model, including both Age and Miles as predictors:

$$\widehat{\text{Time}} = 24.605 + .167 \text{ Age} - .257 \text{ Miles}$$

Now let us estimate the four parameters of the interactive model. To do this we need to compute a new predictor variable that is the product of Age and Miles. We call this variable AgeMiles. Its value, for each observation, is the product of the runner's value on Age multiplied by his or her value on Miles. We then regress Time on Age, Miles, and this new predictor variable, AgeMiles:

$$\widehat{\text{Time}} = 18.899 + .308 \text{ Age} - .069 \text{ Miles} - .005 \text{ AgeMiles}$$

The sum of squared errors for this model equals 697.625.

Let us compare this estimated model to the results that we have already given for the additive model. We can do this in two steps: compare the models to ask whether the interactive model does significantly better than the simpler additive one; and compare the estimated coefficients of the two models, and their interpretations.

In order to compare the quality of predictions of the two models, we treat the additive one as Model C and the interactive one as Model A:

$$\text{MODEL A: Time} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Miles} + \beta_3 \text{AgeMiles} + \varepsilon_i$$

$$\text{MODEL C: Time} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Miles} + \varepsilon_i$$

permitting us a test of the null hypothesis that the parameter associated with the additional predictor (the AgeMiles product) equals zero:

$$H_0: \beta_3 = 0$$

The additive model has a sum of squared errors of 741.195 with 77 degrees of freedom for error, having estimated three parameters. We have just seen that the interactive model has a sum of squared errors of 697.625, with 76 degrees of freedom. Thus the comparison yields an SSR of 43.570 (i.e., $741.195 - 697.625$) with $PA - PC = 1$. The values of PRE and F that result from this comparison are thus:

$$\text{PRE} = \frac{43.570}{741.195} = .059$$

$$F_{1,76} = \frac{.059/1}{(1 - .059)/76} = \frac{43.570/1}{697.625/76} = 4.75$$

Both of these values exceed their critical values. Accordingly, the interactive model, including the product predictor, does a significantly better job of predicting the data than the simpler additive model. In other words, the product predictor is a useful predictor over and above the two variables that are its components, that is, over and above Age and Miles.

To help us understand what this conclusion means, and to understand exactly how this model differs from the additive one, it is important to compare the models in terms of the interpretations we can give the parameter estimates. We will do this in the context of the models' "simple" re-expressions, as defined earlier. In general, we strongly encourage interpretations of parameter estimates in the context of "simple" re-expressions, particularly in the case of interactive models.

We have already interpreted the various parameter estimates in the additive model twice: once as we examined the "simple" Time–Age relationship within levels of Miles and once as we examined the "simple" Time–Miles relationship within levels of Age. As we saw there, when we focused on the "simple" Time–Age relationship, the estimated coefficient for Miles, $-.257$, was the change in the "simple" intercept as Miles increases, while the estimated coefficient for Age, $.167$, was the "simple" slope for Age, invariant across levels of Miles. When we focused on the "simple" Time–Miles relationship, the estimated coefficient for Age ($.167$) was the change in the "simple" intercept as Age increases, while the estimated coefficient for Miles ($-.257$) was the "simple" slope for Miles, invariant across levels of Age.

Parallel interpretations can be given for the interactive model. First, looking at the "simple" Time–Age relationship within levels of Miles, we re-express the estimated interactive model as follows:

$$\widehat{\text{Time}} = (18.899 - .069\text{Miles}) + (.308 - .005\text{Miles})\text{Age}$$

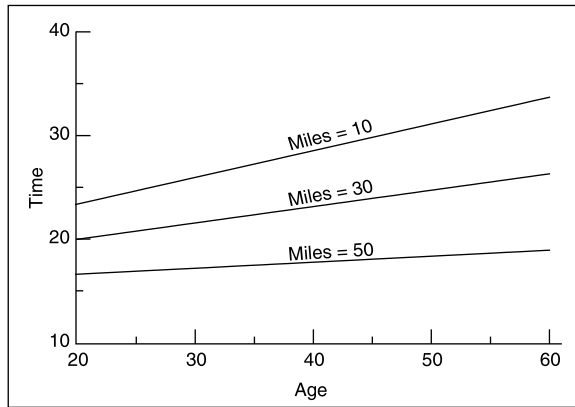
Now let us evaluate this "simple" re-expression at various levels of Miles. We start with the value of zero. Obviously, we must be cautious here in talking about the "simple" Time–Age relationship at this value, since no observation in our dataset had a Miles value of zero. Nevertheless, based on the model, we can substitute the value of zero for Miles in the above re-expression. Our predicted Time–Age "simple" relationship for this value of Miles is:

$$\widehat{\text{Time}} = 18.899 + .308\text{Age}$$

Accordingly, when interpreted from the point of view of the "simple" Time–Age relationship, the intercept in the interactive model, 18.899 , gives us our best estimate of the intercept for the "simple" Time–Age relationship when Miles equals zero. The coefficient for Age in the interactive model, $.308$, tells us the estimated "simple" slope for the Time–Age relationship when Miles equals zero. Note that this interpretation is quite different from the interpretation given for the regression coefficient for Age in the additive model. There the regression coefficient for Age was interpreted as the slope for the Time–Age relationship within levels of Miles, regardless of the actual value of Miles. In other words, there the "simple" slope did not change as Miles changed.

Now, however, in the interactive model, the "simple" slope for Age changes as the value of Miles varies. Suppose we were interested in the "simple" relationship between Time and Age for runners who average 30 miles per week. The prediction function for such runners is:

$$\begin{aligned}\widehat{\text{Time}} &= (18.899 - .069(30)) + (.308 - .005(30))\text{Age} \\ &= 16.829 + .158\text{Age}\end{aligned}$$

FIGURE 7.9 “Simple” Time–Age relationships from interactive model

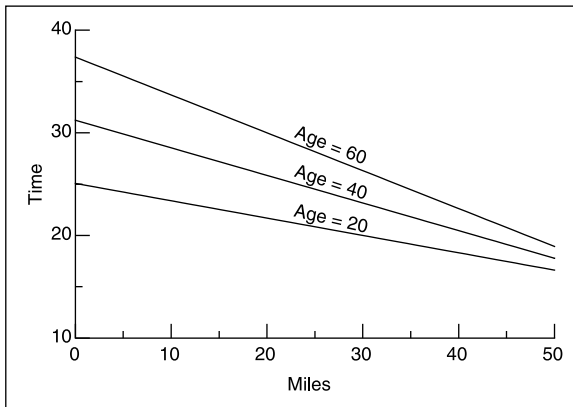
Clearly this “simple” relationship between Time and Age is not what it was for the prediction function when Miles was set equal to zero. As Miles has increased, both the “simple” intercept and the “simple” slope for Age have decreased. For each unit change in Miles, the “simple” intercept has decreased by .069 units. And for each unit change in Miles, the “simple” slope has decreased by .005.

From this perspective, focusing on the “simple” Time–Age relationship in this interactive model, we can now interpret all four parameter estimates. First, the intercept in the model, 18.899, is the “simple” intercept when Miles equals zero. The estimated coefficient for Miles, $-.069$, is the change in the “simple” intercept as Miles increases by one unit. The estimated coefficient for Age, $.308$, is the “simple” slope for Age when Miles equals zero. And the coefficient for the product predictor, $-.005$, is the change in the “simple” slope as Miles increases by one unit.

In Figure 7.9 we have graphed a few of the “simple” Time–Age prediction functions for representative values of Miles. Clearly, both the intercepts and the slopes of these “simple” prediction functions vary as a function of Miles—the intercept by $-.069$ per unit increase in Miles and the slope by $-.005$ per unit increase in Miles. In comparison to the graph of this same simple relationship from the additive model (contained in Figure 7.6), we now are allowing nonparallel “simple” prediction functions. Our earlier statistical conclusion that the product predictor variable significantly improved the quality of the predictions made by the model is equivalent to the conclusion that the graphed “simple” prediction functions in Figure 7.9 are significantly nonparallel. Substantively, the interaction represented by these plotted lines amounts to the conclusion that while 5-kilometer race times may be slower for older than for younger runners, that age difference is less pronounced with more training.

We now turn to the alternative “simple” re-expression of this interactive model, focusing on the “simple” Time–Miles relationship at different levels of Age and interpreting the estimated regression coefficients from this perspective. The re-expressed interactive model is:

$$\widehat{\text{Time}} = (18.899 + .308\text{Age}) + (-.069 - .005\text{Age})\text{Miles}$$

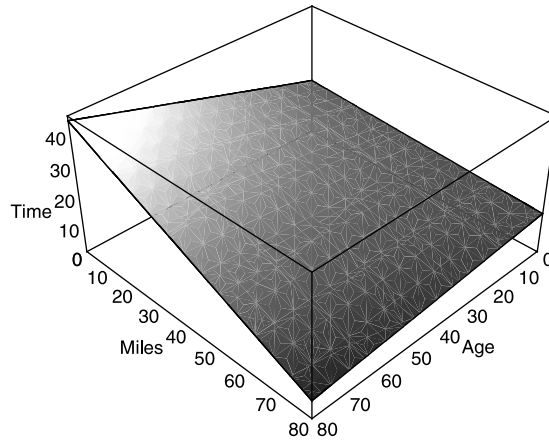
FIGURE 7.10 “Simple” Time–Miles relationships from interactive model

The value of the estimated intercept, 18.999, is the “simple” intercept when Age equals zero. The estimated coefficient for the Age predictor is the change in the “simple” intercept as Age increases by one unit. The estimated coefficient for the Miles predictor, $-.069$, is the “simple” slope for Miles when Age equals zero. And finally, the estimated coefficient for the product predictor, $-.005$, is the change in the “simple” slope for Miles as Age increases by one unit.

Again, in Figure 7.10 we have graphed a few representative prediction functions from this “simple” re-expression. Once again, our earlier statistical test, comparing the additive model as the compact one with the interactive one as the augmented one, is in essence asking whether these “simple” prediction functions are significantly nonparallel. Substantively, our conclusion from this graph of the “simple” prediction functions focuses on the effect of training and how that varies as a function of age. While people who train more run faster 5-kilometer races, that training benefit is greater among those who are older than among those who are younger.

As the above has made clear, each estimated regression coefficient in an interactive model has multiple interpretations (which in a deep sense are all equivalent). So, the coefficient associated with the Age predictor can be interpreted either as the “simple” slope for Age when Miles equals zero or as the change in the intercept for the “simple” Time–Miles relationship as Age increases. The estimated coefficient for the Miles predictor in a parallel manner can be interpreted either as the “simple” slope for Miles when Age equals zero or as the change in the intercept for the “simple” Time–Age relationship as Miles increases. And finally, the estimated coefficient associated with the product predictor can be interpreted either as the change in the “simple” slope for Age as Miles increases or as the change in the “simple” slope for Miles as Age increases.

While the choice of interpretation for the coefficient associated with the product predictor is largely arbitrary, we do have a preference about one of the two interpretations for the coefficients associated with the variables that are components of the product. In general, we find it easier to think in terms of “simple” slopes rather than “simple” intercepts. Accordingly, in the case at hand, we have a preference for interpreting the estimated coefficient for Age as its “simple” slope when Miles equals zero, and the estimated coefficient for Miles as its “simple” slope when Age equals zero. Both

FIGURE 7.11 Interactive model predictions

interpretations for the estimated coefficient for the product predictor refer to “simple” slopes, either how the Age “simple” slope varies as a function of Miles or how the Miles “simple” slope varies as a function of Age. Here, the preference of one over the other of these depends entirely on theoretical assumptions about which of the two component variables is in some sense primary and which one modifies or moderates its “effect.” In the case at hand, we have a slight preference for saying that the age difference in running times decreases as one trains more. But someone else might prefer to say that the training difference in running times is greater among older runners. Both are equally viable interpretations of the significant interaction that is present in these data.

Figure 7.11 presents a three-dimensional graph of the prediction plane generated by this interactive model. As the graphs of the “simple” relations in Figures 7.9 and 7.10 make clear, if we take slices of this prediction plane parallel to one of the horizontal axes, the slope of the resulting prediction lines change. Thus, as we move up the Age horizontal axis, the slope of the plane becomes more and more negative. Compared to the prediction plane generated by the additive model (graphed in Figure 7.5), this prediction plane seems to twist as we move across the values of one of the horizontal axes.

Tests of Other Parameters in the Interactive Model

We have already tested the null hypothesis that the value of β_3 equals zero when we compared the augmented interactive model with the compact additive one. This test of whether the parameter associated with the product predictor departs from zero is formally a test of the interaction: Does the “effect” of Age depend on Miles or, equivalently, does the “effect” of Miles depend on Age? Importantly, this model comparison must include the two component predictors in both Models A and C. A test of a product predictor is a test of the interaction only when both component predictor variables are also included in the models.

With multiple predictors in the interactive model, there are of course other Model Cs with which the interactive model might be compared, other than the additive one that

leaves out the product predictor. For instance, suppose we constructed a compact model in which the parameter associated with the Age predictor was set equal to zero and compared such a model with the full interactive one. This comparison would be testing the null hypothesis that the component variable Age is not a useful predictor over and above Miles and the product predictor. Models A and C for this comparison are estimated as follows:

$$\text{MODEL A: } \widehat{\text{Time}} = 18.889 + .308\text{Age} - .069\text{Miles} - .005\text{AgeMiles}$$

$$\text{MODEL C: } \widehat{\text{Time}} = 31.542 - .416\text{Miles} - .004\text{AgeMiles}$$

The augmented interactive model has a sum of squared errors of 697.625, as we saw previously, while this compact model has a sum of squared errors of 869.153. Their comparison yields a PRE value of .197 and an F statistic of 18.688 with 1 and 76 degrees of freedom. Clearly, we must reject the null hypothesis that the parameter associated with Age in the interactive model has a value of zero.

We can also test the null hypothesis that the parameter associated with Miles in the interactive model equals zero. Estimating a compact model with only Age and AgeMiles as predictors, we get:

$$\text{MODEL C: } \widehat{\text{Time}} = 16.784 - .357\text{Age} - .006\text{AgeMiles}$$

with a sum of squared errors of 702.953. Again, comparing this model with the augmented interactive model, we get $\text{PRE} = .008$ and $F = 0.581$ with 1 and 76 degrees of freedom. We thus cannot reject the null hypothesis that the parameter associated with Miles in the interactive model equals zero.

At the very start of this chapter we tested null hypotheses about the parameters associated with both Age and Miles in the additive model that did not include the product term. Testing whether the regression coefficient for Age in the additive model differed from zero, we obtained a PRE value of .280 and an F of 29.899 with 1 and 77 degrees of freedom. Testing whether the regression coefficient for Miles in the additive model differed from zero, we obtained a PRE value of .568 and an F of 101.143 with 1 and 77 degrees of freedom. Clearly these tests of the Age and Miles parameters in the additive model do not give the same results as the tests of the parallel parameters associated with these variables in the interactive model. For the test of the Age parameter, the resulting PRE value in the additive model was .280, while in the interactive model it equalled .197. For the tests of the Miles parameter, the resulting PRE value in the additive model was .568, while in the interactive model it equalled only .008. Why this difference?

In a general sense, we have already discussed the reason for this difference. As shown in Chapter 6, the interpretation of a partial regression coefficient depends on what other predictors are included in a given model. Since the interactive model includes a different set of predictors than the additive one, we would naturally expect the partial regression coefficient for a given predictor variable to differ between the two models.

In this case, a more specific and substantive answer to this question is extremely important, although in fact this more specific answer is just a restatement of the general one given in Chapter 6 and in the above paragraph. The partial regression coefficients for Age and Miles have quite different interpretations in the interactive model than

they do in the additive one, as we have repeatedly stressed in interpreting the interactive model. Hence, testing whether they significantly differ from zero in the two models amounts to testing entirely different substantive hypotheses. In the additive model, the Age parameter is interpreted as the slope for Age within levels of Miles or holding Miles constant. But that slope is presumed not to change in the additive model as Miles changes. Hence, we can interpret the slope as something like the average Time–Age slope within levels of Miles. On the other hand, the regression coefficient for Age in the interactive model is the slope for Age *when and only when* Miles equals zero.

Therefore in the additive model, when we test whether the Age parameter equals zero, we are testing whether Time and Age are significantly related on average within levels of Miles or holding Miles constant. On the other hand, in the interactive model, we are testing whether Time and Age are significantly related *when Miles equals zero*. Similarly, when we test whether the regression coefficient for Miles equals zero in the additive and interactive models, we are testing entirely different substantive null hypotheses. In the additive case, we are testing whether the slope for Miles on average, within levels of Age or holding Age constant, differs significantly from zero. On the other hand, in the interactive model, the test of whether the Miles parameter differs from zero amounts to testing whether Time and Miles are significantly related *when Age equals zero*.

In the present case, of course, asking questions about “simple” slopes of one predictor variable when the other one equals zero is totally uninteresting and uninformative, since our data do not include any observations with zero values on these variables. Given that we have already found that the interactive model does a better job than the additive one, and therefore that the “simple” slope of one variable changes as a function of the other variable, it seems unlikely that we would be interested in testing a “simple” slope at values outside the range of those included in the data.

In our experience of giving advice on issues of data analysis and in reviewing articles that report results from multiple regression models, misinterpreting regression coefficients associated with predictor variables that are components of interactions is the most frequent error of interpretation that we have encountered. In general, when a product predictor is included in a model, along with its components, the coefficients associated with each component estimate the “simple” effect of that component when the other component with which it is multiplied to make the product equals zero. Since it is frequently the case that values of zero are not of theoretical interest and may not even be possible values of the data, it makes no sense to interpret such regression coefficients. Certainly, they do not represent “overall” or “net effects” of predictors. Indeed, the search for some estimate of the “overall effect” of a component predictor variable is generally misguided. The very nature of a significant interaction between one variable and another means that the “effect” of either one of them “depends”; there is no constant or overall “effect.” Rather, the magnitude (and/or direction) of the effect depends on the level of the other component variable.

Deviating or Centering Component Predictors

It is possible to obtain and test more meaningful “simple” slopes of component predictor variables by transformations that change the zero values of such predictors. Recall that

in Chapter 5 we showed that by deviating a predictor from its mean, or centering it, the intercept in simple regression models came to equal the predicted value at the mean level of the predictor, which equivalently is the mean of Y_i . It then became possible to compare augmented and compact models to ask potentially interesting questions about the mean of Y_i . In a similar vein, it often makes sense in interactive models to center the two component variables to make their regression coefficients of greater interest. Importantly, one must center the predictor variables before computing the product. In other words, the product variable that is included as a predictor in the interactive model should be the product of the two centered predictors. If such an interaction model is estimated, then the regression coefficients for each component predictor variable will equal the “simple” slope for that predictor when the other centered predictor equals zero, that is, at the mean level of the other predictor. Admittedly, no overall or “net” effect of each component predictor is readily available in the context of a significant interaction. Yet one may well want to ask questions about the average “simple” effect of a component predictor, that is, its “simple” effect at the average level of the other component predictor variable.

Let us illustrate with the data at hand. In the dataset, the mean Age is 39.662 and the mean Miles trained is 29.875. We compute new centered or mean-deviated versions of both variables, Age_0 and $Miles_0$, subtracting the means of the variables from each observation’s raw values. The zero subscript is used to indicate that each variable, in its centered form, now has a mean of zero. We now compute a new product predictor variable, taking the product of the two centered components, Age_0Miles_0 .¹ We then estimate the interactive model again, this time using the centered predictor and their product:

$$\widehat{\text{Time}} = 23.431 + .166Age_0 - .258Miles_0 - .005Age_0Miles_0$$

In a deep sense, this is exactly the same interactive model that we estimated previously, without the centered variables. We have simply relocated the zero values on the two horizontal axes of Figure 7.11, putting zero at the mean values of both Age and Miles. The data themselves have not changed nor have the model’s predicted values, that is, the prediction plane in Figure 7.11. Thus, the sum of squared errors for this model is identical to what it was for the estimated interactive model with all variables in their uncentered form, that is, 697.625.

But now all of the estimated coefficients, except for that associated with the product predictor, have changed. And this is because, within the context of this interactive model, they are now estimating different things. The intercept, 23.431, is the predicted value when all three predictors equal zero, which are the mean values of Age_0 and $Miles_0$. The regression coefficient associated with Age_0 , .166, is the “simple” slope for Age when $Miles_0$ equals zero, that is, for runners who run the average number of miles in the dataset. Similarly, the regression coefficient associated with $Miles_0$, $-.258$, is the “simple” slope for Miles when Age_0 equals zero, that is, for runners who are of the average age in the dataset. Since this interactive model is one in which each of these “simple” slopes changes as a function of the other variable, these are different values from the simple slopes estimated in the model with the uncentered predictors. And the degree to which they depart from those uncentered “simple” slopes is a function of the regression coefficient

associated with the product predictor. This is therefore the one estimated coefficient that remains unchanged in the model: it continues to indicate the extent to which the “simple” slope of Age changes as Miles increases or, equivalently, the extent to which the “simple” slope of Miles changes as Age increases. Graphically, the regression coefficients associated with the predictors that are components of the product in the interactive model (Age and Miles in the original model *or* Age_0 and $Miles_0$ in the centered model) are slopes of the prediction plane at particular values along the two horizontal axes (i.e., where the values equal zero), while the coefficient associated with the product predictor is the change in these simple slopes, that is, the degree to which the predicted plane twists as values along a horizontal axis change.

Tests of the regression coefficients associated with the component predictor variables in the centered interactive model are considerably more informative than those same tests in the uncentered version of the model. Comparing the centered interactive model with a compact one that fixes the coefficient for Age_0 at zero yields an SSR of 283.351, a PRE of .289, and an F of 30.91 with 1 and 76 degrees of freedom. Comparing this same interactive model with a compact one that fixes the $Miles_0$ coefficient at zero yields an SSR of 976.934, a PRE of .583, and an F of 106.50 with 1 and 76 degrees of freedom. Conclusions from these tests suggest that within the confines of this interactive model (i.e., allowing “simple” slopes to change) the effect of each component predictor variable is significant at the mean level of the other predictor.

Note that the statistical results of these “simple” tests are similar, although not identical, to the tests of each of the predictors in the three-parameter additive model reported at the start of this chapter. There the statistical test for the additive effect of Age yielded a PRE of .280 and an F of 29.92 with 1 and 77 degrees of freedom, and the test of the additive effect of Miles yielded a PRE of .568 and $F = 101.20$ with 1 and 77 degrees of freedom. Although these results are very similar, they are fundamentally asking different questions. In the context of the additive model, we are simply asking about the “effect” of each variable controlling for the other. In the context of the centered interactive model, we are asking about the “simple effect” of each variable at the average value of the other, allowing those “simple effects” to vary as the other predictor changes values.

Since the regression coefficient associated with the product predictor has not changed as a function of centering its components, and since in a deep sense this centered version of the interactive model is identical to the uncentered version, it has to be the case that the Model A/Model C comparison to test whether the parameter associated with the product equals zero yields the same result in the centered version as in the uncentered case already examined, that is, PRE = .059 and $F = 4.75$ with 1 and 76 degrees of freedom. In other words, regardless of whether we center Age and Miles, the test of the regression coefficient for their product is always a test of whether the “simple” slope of one variable changes as the other variable changes in value, so *long as the two variables involved in the product term, Age and Miles or Age_0 and $Miles_0$, are included as separate predictors in the augmented model.* The italics are to emphasize that the regression coefficient for a product of two predictor variables estimates the effect of the interaction between those variables only when the regression coefficient is a partial one, partialling out or controlling for the component variables that were multiplied

together to yield that product. As long as the components are included as separate predictors in the model, a test of whether the regression coefficient for the product departs from zero will be invariant across all linear transformations of the component variables (Cohen, 1978).

The utility of a centering transformation for both Age and Miles in order to increase the interpretability of their regression coefficients can be generalized to other sorts of deviation transformations in order to test hypotheses about particular “simple” relationships in which one might be interested. Suppose, for instance, in the interactive model, allowing for different “simple” Time–Age relationships at various levels of Miles, we were interested in determining whether Time and Age were significantly related for runners who averaged 50 miles a week. Since the coefficient for Age in the interactive model estimates the “simple” Time–Age slope when the other predictor with which Age interacts equals zero, we could transform Miles by subtracting 50 from the value of Miles for each observation, resulting in a deviation transformation defined as:

$$\text{Miles}_{50} = \text{Miles} - 50$$

Now we regress Time on Age₀, Miles₅₀, and their product Age₀Miles₅₀:

$$\widehat{\text{Time}} = 18.244 + .070\text{Age}_0 - .258\text{Miles}_{50} - .005\text{Age}_0\text{Miles}_{50}$$

Notice that while the coefficient associated with Miles₅₀ has not changed as a result of its transformation (it still equals the same value as it did when both predictors were centered), the coefficient for Age₀ has changed considerably and now is no longer significant: PRE = .022 and $F = 1.69$ with 1 and 76 degrees of freedom. It, of course, estimates the “simple effect” of Age for runners who train an average of 50 miles per week (i.e., when Miles₅₀ equals zero). This estimated “simple effect” is not significant in these data: We cannot conclude that there is an age difference in 5-kilometer race times among runners who train 50 miles per week.

Note the curious but correct implication of this transformation exercise: A linear transformation of one component predictor variable in an interactive model (and a recomputation of the product predictor) does not affect the regression coefficient associated with that component predictor; rather, it affects the regression coefficient associated with the other component predictor. This curious fact results from the definition of these component coefficients as “simple” slopes for each component when the other equals zero. If the zero value of one is altered, the “simple” slope for the other changes, assuming that an interaction between the two exists.

One set of commonly used and recommended transformations, to present the results of an interactive model, is to estimate and test “simple” slopes of one component variable at three values of the other: its mean, one standard deviation above its mean, and one standard deviation below its mean. Although there is nothing sacred about such values, and others might be more informative in a particular context, getting into the habit of estimating and testing different “simple” slopes of variables that are components of interactions is good practice.

Power Considerations in Interactive Models

In Chapter 6, we presented the general formula for the confidence interval of a regression coefficient in multiple regression. That formula, reproduced from Chapter 6, is:

$$b_{p.12\dots p-1} \pm \sqrt{\frac{F_{crit}(\text{MSE})}{s_p^2(n-1)(1-R_p^2)}}$$

where $b_{p.12\dots p-1}$ is the regression coefficient for the p th predictor variable, F_{crit} is the critical value of the F statistic with 1 and $n - PA$ degrees of freedom, MSE is the mean squared error of the model, s_p^2 is the variance of the p th predictor, and $1 - R_p^2$ is its tolerance. This formula is the appropriate one regardless of the nature of the predictor variable. Thus, it provides the confidence intervals for the regression coefficients of products and their components in interactive models, as well as for the coefficients of additive models.

The same considerations affect the power of tests in interactive models as in additive models. Namely, factors that tend to increase the size of the confidence interval tend to decrease the power of the test of the null hypothesis that the estimated parameter equals some a priori value. Factors that decrease the size of the confidence interval increase power. Accordingly, as the mean square error of the model decreases, as the variance for the p th predictor increases, and as the tolerance of the p th predictor increases, the inferential test of its associated parameter is more powerful.

Within the context of interactive models, however, there is an interesting twist to the power issue, since the product predictor variable is computed from the other predictor variables in the model and thus is in some sense dependent upon them. Consider the following two variables and their product:

X_1	X_2	X_1X_2
3	2	6
1	3	3
4	5	20
2	1	2
5	4	20

If we were to regress some data variable Y_i on these three predictors, the confidence interval for the regression coefficient for the product variable would be a function of the critical value of F , the mean square error from the model, the variance of the product predictor, its tolerance, and the sample size. Let us think about these last three terms, all in the denominator of the formula for the confidence interval. Multiplying the variance of the product by $n - 1$ gives the sum of squares of the product around its mean. The tolerance of the product is obtained by regressing it on its two components and subtracting the resulting R^2 from 1. In this case, the total sum of squares of the product equals 328.80 and its tolerance equals .0097 (notice that the product and its components are extremely redundant here). If we multiply these two together we get 3.175,² which is the sum of squared errors resulting from the model where the product is regressed on its two components. This is the value of the denominator in the formula for the confidence interval.

Now consider what happens if we center both X_1 and X_2 , subtracting their mean values (3 in both cases), and then compute the product of the transformed variables:

X_{10}	X_{20}	$X_{10}X_{20}$
0	-1	0
-2	0	0
1	2	2
-1	-2	2
2	1	2

Now the variance of the product variable equals 1.2, $n - 1$ still equals 4, and the tolerance of the product equals .6615. As a function of the transformations of X_{10} and X_{20} , their product is considerably less variable and relatively less redundant (linearly) with its two components. Even so, by multiplying these three terms together, the product in the denominator of the confidence interval still equals 3.175 (and this number is again the error sum of squares from a model in which the product of the two centered variables is regressed on its two centered components). This is exactly as it should be, since inferential tests of a product variable's regression coefficient are not affected by linear deviation transformations of its components, as we have seen. Thus, the power of the test of the interaction or product variable is unaffected by deviation transformations of its components, *as long as those components are included as predictors in the model*.

What is interesting about this result, however, is that deviation transformations do affect the tolerance of the product variable or its linear redundancy with its components. Yet the width of the confidence interval of its regression coefficient is unaffected since the change in tolerance is exactly compensated for by the change in the variance of the product variable. This is an interesting exception to the general rule given in the last chapter that redundancy among predictors tends to reduce power. When a predictor is a product of other predictors, then various transformations can radically affect its redundancy with its components. Yet such transformations have no effect on the power of the test of the product's regression coefficient, as long as the component variables are included as predictors in the model.

For the running data, we compute the confidence interval for the regression coefficient of the product predictor for the interactive model, first with the component predictors in their original metrics, and then with them centered. In both models, the critical value of F , with 1 and 76 degrees of freedom, is 3.96 and the MSE from both models equals 9.18.

In the original uncentered model, the variance of the product, AgeMiles, equals 380,387.09 and its tolerance ($1 - R^2$ when regressed on Age and Miles) equals .064. Accordingly the confidence interval for the coefficient of the product predictor is:

$$.005 \pm \sqrt{\frac{3.96 \times 9.18}{380,387.09 \times 79 \times .064}}$$

$$.005 \pm \sqrt{\frac{36.35}{1,917,017.84}}$$

$$.005 \pm .004$$

And when the two component predictors are centered and these and their product are the predictors in the interactive model, the confidence interval for the coefficient for the product equals:

$$\begin{aligned} &.005 \pm \sqrt{\frac{3.96 \times 9.18}{24,276.01 \times 79 \times .999}} \\ &.005 \pm \sqrt{\frac{36.35}{1,917,017.84}} \\ &.005 \pm .004 \end{aligned}$$

The lesson from all of this is that while a product variable may be quite redundant, in a linear manner, with its components, and while transformations of its components may reduce that redundancy, such transformations will have no effect on the statistical inference of the product variable's regression coefficient as long as the component variables are included in the model.

A GENERAL PROCEDURE FOR THE DERIVATION OF "SIMPLE" SLOPES

As we have emphasized, we find it easiest to interpret parameter estimates in models with product predictors by thinking in terms of "simple" slopes. Up to this point we have relied on a simple heuristic to derive these "simple" slopes: we have simply rearranged terms in the model. With more complex models, this approach does not work. In this section we present a more general procedure for deriving "simple" slopes, relying on a notion in calculus: the derivative or partial derivative of a function with respect to some variable.

A derivative *is* a "simple" slope, in the way we have been using it, in so many words. We can express any estimated multiple regression equation in the following generic terms:

$$\hat{Y}_i = F(X_1, X_2, X_3, \dots, X_p)$$

meaning simply that our predicted values are some function of a set of predictor variables. The "simple" slope associated with one of these variables, say X_j , is the partial derivative of the function with respect to that variable:

$$\frac{\Delta \hat{Y}}{\Delta X_j}$$

This standard notation for a derivative means that it is a slope, in the sense that the "Δ" indicates a change or difference: We are asking about the "simple" difference in the predicted values as a function of a one-unit difference in a particular predictor, X_j .

Even if one is unfamiliar with calculus, it is easy to calculate partial derivatives of the sorts of functions that are of interest to us in this book. Three simple rules suffice:

1. The partial derivative of a sum, with respect to X_j , equals the sum of the partial derivatives of the components of that sum.

2. The partial derivative of aX_j^m with respect to X is amX_j^{m-1} , where a can be either a constant or another variable (or some product of both).
3. The partial derivative of a component of a sum, with respect to X_j , where that component does not contain X_j , is zero.

Let us take the interactive model that we have been dealing with so far in this chapter and apply these rules to derive the “simple” slopes of both Age and Miles at particular levels of the other variable.

The function is:

$$\widehat{\text{Time}} = b_0 + b_1\text{Age} + b_2\text{Miles} + b_3\text{AgeMiles}$$

To evaluate the “simple” slope for Age, we take the partial derivative of all four components of this expression with respect to Age. According to the third rule above, the partial derivative of the first component, b_0 , is zero, as is that of the third. And the partial derivative of $b_1\text{Age}$ is b_1 according to the second rule above, and that of $b_3\text{AgeMiles}$ is $b_3\text{Miles}$, also according to the second rule. This yields the following partial derivative for Age:

$$\frac{\Delta \widehat{\text{Time}}}{\Delta \text{Age}} = b_1 + b_3\text{Miles}$$

This derivation of the Age “simple” slope is, of course, identical to that discussed earlier: b_1 is the “simple” slope when Miles equals zero and b_3 is the increase in that “simple” slope as Miles increases. Unsurprisingly, a parallel process yields the “simple” slope for Miles identified earlier, that is, $b_2 + b_3\text{Age}$.

Consider now a model, which we have yet to explicate, involving a product of a predictor with itself:

$$\hat{Y}_i = b_0 + b_1X_i + b_2X_i^2$$

Let us take the derivative of this with respect to X_i , thus obtaining its “simple” slope:

$$\frac{\Delta \hat{Y}_i}{\Delta X_i} = b_1 + 2b_2X_i$$

What does this mean? This is the “simple” slope at a particular value of X_i . It says that the “simple” slope depends on the very value of X_i itself. At higher values of X_i the “simple” slope is something different than at lower values. This defines a curvilinear relationship between X_i and Y_i , such that the “simple” slope varies as a function of X_i itself. The “simple” slope is thus the slope of the line (actually curve) at a particular value of X_i , with that slope changing continuously as X_i increases. If it helps, think about this simple slope (i.e., the derivative) as the slope of a line tangent to the curve that results from the plot of the predicted values from the estimated model:

$$\hat{Y}_i = b_0 + b_1X_i + b_2X_i^2$$

We will return with a data example to illustrate this shortly.

Finally, consider a considerably more complicated model:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}^2 + b_3X_{1i}^2 + b_4X_{1i}X_{2i}$$

Again, we will worry about the specifics of such a model later. For now, we simply derive the “simple” slopes of both X_{1i} and X_{2i} :

$$\frac{\Delta \hat{Y}_i}{\Delta X_{1i}} = b_1 + 2b_3X_{1i} + b_4X_{2i}$$

$$\frac{\Delta \hat{Y}_i}{\Delta X_{2i}} = b_2 + b_4X_{1i}$$

Thus, according to this model the “simple” slope of X_{1i} depends both on its value and on the value of X_{2i} , while the “simple” slope of X_{2i} depends only on X_{1i} .

POWERS OF PREDICTOR VARIABLES

A generalization of what has been learned about interactions between two predictor variables can be used to understand models that include powers of a predictor variable. Suppose, for instance, with the data at hand, we suspect that the relationship between Time and Miles depends on the level of Miles. We suspect that when one does not train very much, the relationship between training and race times might be a lot stronger than it is when one trains a great deal. In other words, adding 5 miles more per week to one’s training regimen is likely to make much more of a difference to one’s race time if one’s base were 10 miles per week than if one’s base were 50 miles per week. Such a prediction amounts to arguments about changes in the “simple” Time–Miles relationship. This time, however, we are not saying that the magnitude of that “simple” relationship changes as Age changes. Rather, we are hypothesizing that the “simple” Time–Miles relationship changes as Miles changes. In essence, we are hypothesizing an interaction of Miles with itself in predicting Time.

We use the same functional form to incorporate such an interaction into our model; namely, we include the product of the two variables that are hypothesized to interact as a separate predictor in the regression model. But this time, since it is Miles that is thought to interact with itself, we include a product predictor that is Miles \times Miles (or Miles²):

$$\widehat{\text{Time}} = b_0 + b_1\text{Miles} + b_2\text{Miles}^2$$

Given the present data, this model is estimated as:

$$\widehat{\text{Time}} = 37.479 - .754\text{Miles} + .008\text{Miles}^2$$

with a sum of squared errors of 894.190.

To interpret this model and what its parameters mean, we do two things: we derive and plot predicted values of Time as a function of Miles; and we interpret the parameters by focusing on the “simple” slope for Miles. Since the maximum value of Miles in our

dataset is 50, we derive predicted values of Time (as a function of Miles) varying between 0 and 50. For instance, when Miles equals zero, the predicted value of Time equals:

$$37.479 - .754(0) + .008(0^2) = 37.479$$

And when Miles equals 10, the predicted value of Time equals:

$$37.479 - .754(10) + .008(10^2) = 30.739$$

Once we have a full set of such predicted values we plot them, as in Figure 7.12. As is apparent, these predicted values, as a function of Miles, no longer lie along a straight line. Although in a formal sense our model is still a “linear” one, the prediction function is now a curve. And, as we expected, the shape of the curve is such that the slope is decidedly steeper at lower levels of Miles than it is at higher levels: Training one additional mile per week has more of a benefit in reducing race times if one is at a low training base than if one is at a high training base.

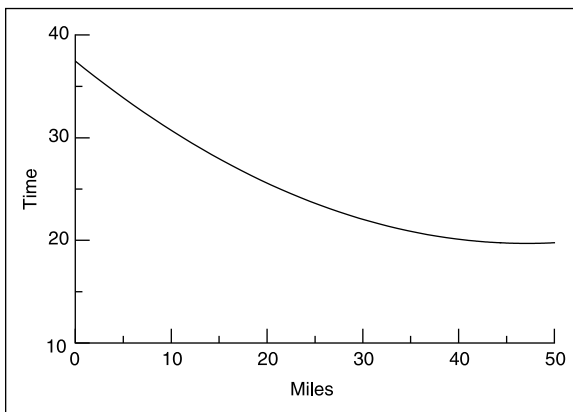
To interpret the coefficients of the model, we can start with the intercept, 37.479. This number represents what it always has, namely the predicted value when the predictors both equal zero. In this case, it is the predicted race time for someone who trains zero miles per week.

To interpret the two slope estimates in this model, we derive the “simple” slope for Miles, using the procedure explicated in the last section to get the derivative of this function with respect to Miles:

$$\frac{\Delta \widehat{\text{Time}}}{\Delta \text{Miles}} = -.754 + (2).008\text{Miles}$$

This Miles “simple” slope obviously varies as a function of Miles. It amounts to the slope of lines that are at a tangent to the prediction curve in Figure 7.12 at various points along the Miles horizontal axis. So, for instance, when Miles equals zero, the “simple” slope of the curve equals $-.754$. When talking about someone who runs 10 miles per week, the “simple” slope of the curve is $-.754 + (2).008(10) = -.594$. At the point of the curve where Miles equals 20, the “simple” slope is $-.754 + (2).008(20) = -.434$. And by the time we get up to someone who runs 40 miles per week, the “simple” slope is $-.754 + (2).008(40) = -.114$.

FIGURE 7.12 Time–Miles nonlinear model



Let us be clear about what these “simple” slopes are and are not. They are not predicted values of Time. Predicted values are what we plotted when we derived the graph in Figure 7.12. These “simple” slopes are rather the slopes of the curve at each different point along the Miles axis. They tell us the expected benefit of running one more mile per week on race times, given that one already is training by running a given number of miles each week. It may be difficult to think about the slope of this curve at particular points, since the curve is obviously changing its slope continuously. Accordingly, think about the slope of lines that are at a tangent to the curve at different values of Miles. The slopes of such lines, at particular values of Miles, are the “simple” slopes.

From the point of view of the “simple” slope, the coefficient associated with Miles, $-.754$, is the “simple” slope of the curve at only one point, namely when Miles equals zero. In the plot of Figure 7.12, it is the slope of the curve at the intercept, when the curve is steepest. And as Miles increases, the “simple” slope becomes less and less negative. For each one-unit increase in Miles, the slope becomes less negative by $2 \times .008$. This provides the interpretation for the regression coefficient associated with Miles²: $.008$ is half the change in the “simple” Miles slope as Miles increases by one unit. If you wish, it indexes the degree of nonlinearity in the curve plotted in Figure 7.12.

We can ask whether this model, with its changing “simple” slope and its quadratic prediction function, does a better job than the simpler linear model with Miles as the only predictor. Our model comparison is thus:

$$\text{MODEL A: Time} = \beta_0 + \beta_1 \text{Miles} + \beta_2 \text{Miles}^2 + \varepsilon$$

$$\text{MODEL C: Time} = \beta_0 + \beta_1 \text{Miles} + \varepsilon$$

and the null hypothesis is $H_0: \beta_2 = 0$. In words, the null hypothesis is that the Time–Miles relationship is simply a linear one; there is no diminishing marginal return for additional miles of training on 5-kilometer race times.

At the very start of this chapter, we provided the estimates for Model C:

$$\widehat{\text{Time}} = 31.911 - .280 \text{Miles}$$

with $\text{SSE}(C) = 1029.013$. Recall that $\text{SSE}(A) = 894.190$. Thus the value of SSR for this comparison is 134.823 , which converts to a PRE of $.131$ and an F of 11.61 with 1 and 77 degrees of freedom. Thus, Model A is to be preferred to Model C and we can conclude that a quadratic prediction function is a better fit to the data than a linear one. There is thus evidence that the benefit of training one more mile per week, on 5-kilometer race times, is less the higher the training base is.

We could also, of course, compare this augmented nonlinear model to a compact one in which β_1 is fixed at zero, regressing Time only on Miles². The estimates for such a model are:

$$\widehat{\text{Time}} = 28.101 - .004 \text{Miles}^2$$

with a sum of squared errors equal to 1222.243 ; the comparison generates a PRE of $.268$ and an F of 28.25 with 1 and 77 degrees of freedom. We thus conclude that the parameter associated with Miles in this augmented model is not equal to zero. But what exactly

have we tested here? Have we tested the overall linear effect of Miles on Time? The answer is clearly “no,” since in this augmented model the effect of Miles on Time varies as a function of Miles, as indicated by the Miles “simple” slope, $-.754 + (2).008\text{Miles}$. What we have just tested is not an “overall” linear effect of Miles, but rather the “simple” effect of Miles when and only when Miles equals zero. Based on these data and allowing Miles to be nonlinearly related to Time, we can conclude that running one more mile per week lowers race times when one is currently running zero miles per week. But since the effect of running more miles depends on the number of miles one already runs per week, this test does not tell us about the benefits of running one more mile given any other base per week. Additionally, since the value of zero on Miles lies outside the range of values included in the dataset, we probably should not have confidence in this prediction even if it is statistically significant.

As with the interactive model discussed previously, the interpretation of slopes of component predictor variables is a bit easier if those components are mean-deviated or centered and then the product predictor is recomputed using this centered variable. When we center Miles and regress Time on Miles_0 and Miles_0^2 (i.e., the centered variable and its square), we get the following parameter estimates:

$$\widehat{\text{Time}} = 22.054 - .279\text{Miles}_0 + .008\text{Miles}_0^2$$

In a deep sense, of course, this model is the same as that estimated before, prior to centering the predictor. All we have done is move the zero point on the horizontal axis of Figure 7.12, so that it is now at the mean value of Miles, 29.875. And at that point, the “simple” slope of Miles in this model is $-.279$. This is the interpretation of the coefficient for Miles_0 in this centered model: It is the simple slope when Miles_0 equals zero, which of course is at the mean of Miles.³

A comparison of this model with a Model C that fixes β_1 at zero tests whether the “simple” effect of Miles is different from zero when dealing with runners who train the average number of miles per week (average in the dataset). The estimated Model C is:

$$\widehat{\text{Time}} = 22.021 + .008\text{Miles}_0^2$$

with a sum of squared errors of 2070.265. This comparison gives a PRE of .568 and an F of 101.264, with 1 and 77 degrees of freedom. At the average level of Miles in the dataset, there is a highly significant Time–Miles relationship: Running one more mile per week leads to a lower 5-kilometer race time.

Note that this analysis is subtly different in both its details and its conclusion from asking the simple regression question of whether there is a significant overall linear relationship between Time and Miles. Here we are asking about a significant “simple” linear relationship at the mean level of Miles, in the context of an overall quadratic model. When we ask whether Miles is related to Time in a simple regression model, we are assuming a linear and thus constant Time–Miles relationship.

We have just provided tests of “simple” slopes both when Miles equals zero and when Miles equals its mean. We could re-estimate this nonlinear model, of course, with Miles deviated from any other particular value where the Miles “simple” slope was of interest. Suppose, for instance, that someone ran 40 miles a week and wanted to know whether the present data suggest a significant effect of running one more mile, given

this base. To test the “simple” slope of Miles when Miles equals 40, within the confines of this nonlinear model, we would simply re-estimate it with a new deviated Miles predictor, deviating it from 40, and its square:

$$\widehat{\text{Time}} = 20.042 - .118\text{Miles}_{40} + .008\text{Miles}_{40}^2$$

Again, in a deep sense, this is the same nonlinear prediction function, making the same predictions. But now the coefficient associated with Miles_{40} estimates the “simple” slope for Miles when it equals 40. And a comparison of this augmented model with a Model C that omits the Miles_{40} predictor provides a test of that “simple” slope. The resulting PRE is .057, which converts to an F of 4.62 with 1 and 77 degrees of freedom. We can conclude that there is a significant “simple” effect of Miles when Miles equals 40, allowing a nonlinear relationship between Time and Miles.

Just as we noticed when discussing interactive models, the coefficient for Miles^2 in all three versions of the quadratic model we have estimated (with Miles in its raw metric, with Miles centered, and with Miles deviated from 40) remained the same. And Model A/Model C tests of its significance yield the same result in every case. Again, assuming that the component predictor is included in the model, the test of the coefficient of the product predictor is invariant across deviation transformations of its component. It continues to examine whether the “simple” slope changes as Miles changes value.

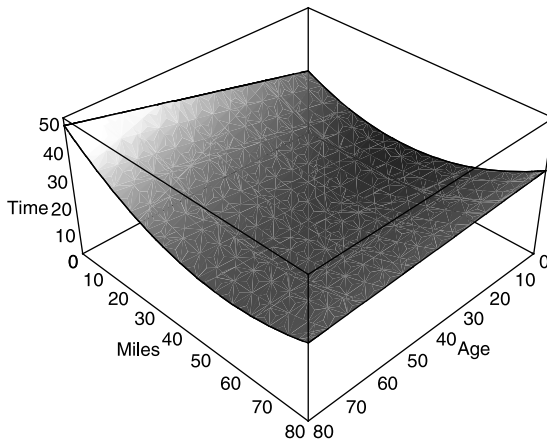
MORE COMPLEX PRODUCT MODELS

We have found evidence in our data for both an interaction between Age and Miles and a quadratic effect of Miles in predicting race times. We have examined these two nonlinear effects separately, rather than putting them both in a single model, for didactic purposes. Let us now look at a more complete model that incorporates both the Age by Miles interaction and the quadratic Miles term. The estimated model is:

$$\widehat{\text{Time}} = 25.578 + .273\text{Age} - .565\text{Miles} + .008\text{Miles}^2 - .004\text{AgeMiles}$$

The sum of squared errors for this model equals 579.724. If we compare this model with a model that omits the Miles^2 predictor, the resulting PRE equals .169 and the F , with 1 and 75 degrees of freedom, equals 15.30. Hence, the nonlinearity in the Time–Miles relationship is still found even when we control for Age and the interaction of Age with Miles. However, when we compare this model with one that deletes only the AgeMiles interaction term, the resulting PRE (.039) and F (3.06) are not significant. Hence, although we found that Age and Miles reliably interact in predicting Time when Miles^2 is not controlled, that effect ceases to be significant in this more complete model.

Nevertheless, let us continue to interpret the various coefficients of this model in order to clarify the interpretation of more complex nonlinear functional forms. As before, we want to do two things to interpret this model. First, we derive predicted values of Time for all combinations of Age and Miles and we plot these, this time in the three-dimensional graph of Figure 7.13. And second, to interpret the regression coefficients of the model, in the context of this graph of the predicted values, we derive the “simple” slopes of both Age and Miles according to this model:

FIGURE 7.13 Predictions of model with Miles squared and Miles by Age interaction

$$\frac{\widehat{\Delta \text{Time}}}{\Delta \text{Age}} = .273 - .004 \text{Miles}$$

$$\frac{\widehat{\Delta \text{Time}}}{\Delta \text{Miles}} = -.565 + (2) \cdot .008 \text{Miles} - .004 \text{Age}$$

First, let us interpret the coefficient associated with Age in the model, .273. According to the “simple” slope derived for Age, above, it is that “simple” slope of Age when and only when Miles equals zero. In the three-dimensional plot of Figure 7.13, we can think about taking slices of the prediction plane parallel to the Age axis. Such slices are prediction lines as a function of Age, at particular values of Miles. One of these slices is found when Miles equals zero. And the slope of that slice (or line) is .273. Substantively, we might say that with no training, a 1-year increase in Age is predicted to result in slower 5-kilometer race times by .273 minutes.

The coefficient associated with the Miles predictor, $-.565$, can be interpreted in the context of the derived “simple” slope for Miles, above. There we can see that it is the “simple” slope of Miles when and only when *both* Age and Miles equal zero. Graphically, it is the slope of the plane, with respect to Miles, at the point in the plot where both Age and Miles equal zero. If we met a runner who was zero years old and who ran zero miles per week, we would predict that running one more mile per week would result in a faster 5-kilometer race time by .565 minutes! Obviously, neither this coefficient nor the one associated with the Age predictor focuses on a “simple” slope of much interest, given the values in our dataset. We will revisit them, with centered predictor variables later, once we have interpreted the other two slope coefficients in this model.

The coefficient for Miles² in the model, .008, is half the change in the “simple” slope for Miles as Miles increases. Again, looking at the prediction plane in Figure 7.13, taking slices parallel to the Miles axis we get a series of curves. This coefficient tells us about the degree to which these curves depart from linearity. The Miles “simple” slope is not constant; it changes as Miles increases, becoming less negative. This interpretation is similar to the interpretation given earlier for the coefficient for Miles²

in the model that included neither Age nor AgeMiles. The difference is that here we are estimating the degree of nonlinearity in the Time–Miles “simple” relationship, controlling for Age and its interaction with Miles.

Finally, the coefficient associated with the AgeMiles product, $-.004$, can be interpreted in the context of either of the derived “simple” slopes. From the point of view of Age, it is the change in the Age “simple” slope as Miles increases. In Figure 7.13, taking slices of the prediction plane parallel to the Age axis we can see that the “simple” slope for Age is not invariant. It seems steeper (i.e., more positive: older runners have longer race times) when Miles is relatively low than when the value of Miles is higher. From the point of view of the Miles “simple” slope, it is the change in that “simple” slope as Age increases. Taking slices of the prediction plane parallel to the Miles axis, in general the predictions are going down as Miles increases (i.e., faster race times with more miles training). But the extent to which this is true seems greater among older runners than among younger ones. The linear effect of Miles is greater among older runners than among younger ones.

If we were to center Age and Miles and regress Time on the centered predictors and their product terms, it is an interesting exercise to figure out which of the coefficients in the model would change and which would not. Obviously, all we would be doing is changing where the value of zero is found on both the Age and Miles axes. Any coefficient that makes reference to the value of zero of either variable would thus change. Accordingly, it should come as no surprise that when we reestimate this model, having centered the two variables and recomputed the product terms, we get the following:

$$\widehat{\text{Time}} = 22.044 + .167\text{Age}_0 - .257\text{Miles}_0 + .008\text{Miles}_0^2 - .004\text{Age}_0\text{Miles}_0$$

The coefficients associated with Age_0 and with Miles_0 now have different values. The former is the “simple” slope for Age when Miles equals its mean value; the latter is the “simple” slope for Miles when both Miles and Age equal their mean values. The coefficients associated with Miles_0^2 and $\text{Age}_0\text{Miles}_0$ have not changed, for they make no reference to zero values on the two axes. Rather, these coefficients indicate *changes* in the “simple” slopes; they are not “simple” slopes when something else equals zero. It should by now come as no surprise that this “centered” model has the same sum of squared errors as the “uncentered” one estimated previously.

Although the data that we have been using throughout this chapter are not consistent with any more complex nonlinear functional forms than those already explored, we could extend this model to a more complex one, should our theories suggest that more complex forms are plausible. There is nothing inherently more difficult about interpreting more complex nonlinear models, so long as the basic principles that we have explained are kept in mind. In general, we want to generate predicted values as a function of the possible values of the predictors, and then graph these. To interpret the resulting coefficients, it is easiest to do so in the context of derived “simple” slopes. Doing this makes clear the general principle that the coefficient associated with any predictor variable that is a component of one or more product predictors included in the model estimates the “simple” slope of that variable when and only when the other component(s) of those products equal zero. So, in the example just discussed, the estimated Age coefficient is the “simple” slope for Age when Miles equals zero, because Age is a component of the

AgeMiles product that is included in the model. And the estimated coefficient for Miles is the “simple” slope for Miles when both Age and Miles equal zero, because Miles is a component of both the Miles² and AgeMiles products that are included in the model.

Additionally, if the interpretation of a coefficient is in terms of a “simple” effect at a zero level of some other component variable, then transformations that affect that zero level will alter the obtained coefficient. Only the coefficients of predictors in models that are not themselves components of product predictors will be unaffected by such deviation transformations.

We close the chapter with one last model that includes the product of three different predictor variables:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{1i}X_{2i} + b_5X_{1i}X_{3i} + b_6X_{2i}X_{3i} + b_7X_{1i}X_{2i}X_{3i}$$

As before, whenever we have products in a model, we want to make sure that we include all the components of those products as predictors. And in this case, all components of the triple product $X_{1i}X_{2i}X_{3i}$ include not only the single predictor variables, e.g., X_{1i} , but also the products of two of the three predictors, e.g., $X_{1i}X_{2i}$. To aid interpretation of the resulting coefficients, we will assume that all variables have been centered and the included product predictors have been computed from their centered components.

With a model such as this, a single plot is not possible because only three dimensions are visually available. Instead, we could generate separate three-dimensional plots at various levels of the fourth variable. For instance, we might put X_{1i} and X_{2i} on the horizontal axis and plot predicted values of Y_i in one graph at one value of X_{3i} , creating additional graphs for different values of X_{3i} .

To interpret the various parameter estimates, we generate the “simple” effect expressions:

$$\frac{\Delta \hat{Y}_i}{\Delta X_{1i}} = b_1 + b_4X_{2i} + b_5X_{3i} + b_7X_{2i}X_{3i}$$

$$\frac{\Delta \hat{Y}_i}{\Delta X_{2i}} = b_2 + b_4X_{1i} + b_6X_{3i} + b_7X_{1i}X_{3i}$$

$$\frac{\Delta \hat{Y}_i}{\Delta X_{3i}} = b_3 + b_5X_{1i} + b_6X_{2i} + b_7X_{1i}X_{2i}$$

These “simple” slopes inform us first of all about the estimated coefficients attached to the three predictor variables by themselves, b_1 , b_2 , and b_3 . These obviously are the “simple” effects of each variable at the mean (zero value given centering) of both of the other predictors. Each of the estimated coefficients associated with the products involving two variables, b_4 , b_5 , and b_6 , can be interpreted in the context of two different “simple” slopes. These estimate “simple” two-way interactions between the variables involved in the product at the mean level (zero level) of the third variable. For example, b_4 can be interpreted either as the change in the “simple” slope of X_{1i} as X_{2i} increases when X_{3i} equals the mean, or as the change in the “simple” slope of X_{2i} as X_{1i} increases when X_{3i} equals the mean. And finally, there are three possible interpretations for the coefficient associated with the triple product, b_7 . These three possible interpretations are perhaps best shown if we derive “simple” slopes for each of the three two-way products:

$$\frac{\Delta \hat{Y}_i}{\Delta X_{1i}X_{2i}} = b_4 + b_7X_{3i}$$

$$\frac{\Delta \hat{Y}_i}{\Delta X_{1i}X_{3i}} = b_5 + b_7X_{2i}$$

$$\frac{\Delta \hat{Y}_i}{\Delta X_{2i}X_{3i}} = b_6 + b_7X_{1i}$$

These are derived in exactly the same way as we explicated for the “simple” slopes of single predictor variables, but now we apply those rules to a product predictor. In this context, b_7 can be interpreted in the following three equivalent ways:

1. The change in the effect of the $X_{1i}X_{2i}$ interaction as X_{3i} increases.
2. The change in the effect of the $X_{1i}X_{3i}$ interaction as X_{2i} increases.
3. The change in the effect of the $X_{2i}X_{3i}$ interaction as X_{1i} increases.

These interpretations obviously imply that one has already interpreted the “simple” two-way interactions at the mean level of the third variable, as discussed above. Thus, the process of understanding a model like this, with a triple interaction, happens in steps. First, one attempts to make sense of the “simple” effects of single variables at the mean level of the others, then the two-way interactions at the mean of the third (i.e., how the “simple” effect of one variable depends on the level of a second, at the mean of the third), and finally one interprets the three-way interaction by talking about how a “simple” two-way interaction is affected by changes in the third variable.

As we have discussed, with a two-way interaction there are two equally viable interpretations (“the simple effect of X_{1i} depends on the value of X_{2i} ” or “the simple effect of X_{2i} depends on the value of X_{1i} ”), but one of these is likely to be more intuitively compelling than the other. Similarly, with a three-way interaction and its three possible interpretations, one will generally emerge as more intuitively compelling than the other two.

One final word of caution is in order about the sorts of models we have been exploring in this chapter. With three variables as predictors in a model, there are three two-way interactions and one three-way interaction that might be estimated. With four predictor variables, there are potentially six two-way interactions, four three-way interactions, and one four-way interaction. And with increasing numbers of predictor variables, the possible interactions go up geometrically. And this count does not include the possibilities that arise from including quadratic terms and higher order powers of predictors (and the interactions of these with other predictors).

Although we have explicated the general approach for testing and interpreting models involving product predictors, in general the inclusion of such predictors should be strongly dependent on compelling theoretical expectations. We do not encourage the inclusion of product predictors in models unless there are such theoretical expectations. Building more and more complex models is not our goal, even if we do have the tools to interpret them.

SUMMARY

The models that have been considered in this chapter considerably extend the range of questions that can be asked of data. They do so by allowing the slopes of predictor variables to vary as a function of the values of those variables and of other variables. In the most frequently encountered case, these models thus allow statistical interactions between predictor variables—examining whether the “effect” of one variable depends on the value of another. These models also allow nonlinear effects, examining for instance whether the “effect” of one variable changes as its own values change.

In general, these interactive and nonlinear effects are estimated by including products of predictors as additional predictor variables, making sure that all components of those products are included as predictors as well. Interpreting such models can be difficult and we have recommended a two-stage procedure, first graphing predicted values and then deriving “simple” slopes and interpreting parameter estimates in the context of such simple slopes. By focusing on these “simple” slopes, it becomes clear that the parameter estimate associated with a predictor that is a component of some product predictor estimates the “simple” effect of that predictor when and only when the other components of the product equal zero. To make these estimates generally more interpretable, we have recommended centering or mean-deviating variables prior to computations of the product predictors.

While models such as those that have been considered in this chapter (allowing interactions and nonlinear effects) dramatically increase the range of questions that one can ask of data, their complexity can become overwhelming. We therefore encourage analysts to examine such effects only when there are good theoretical reasons for them. Adding products willy-nilly to models will only result in Type I errors and overly complex models.

Notes

- 1 While this product variable is the product of two centered variables, it will not necessarily be centered itself. What is important is that the product be computed from the two centered variables, not that the product itself be centered.
- 2 This is the product of the two if they are not rounded (as they are in the text).
- 3 This value can be computed directly from the expression for the “simple” slope of Miles given earlier from the estimated nonlinear model prior to centering. There the “simple” slope was calculated as $-.754 + (2).008\text{Miles}$. If we substitute the mean level of Miles, 29.875, in this expression we get $-.754 + (2).008(29.875) = -.276$, which is the estimated coefficient for Miles₀ in the centered model (within rounding error).

8

One-Way ANOVA

Models with a Single Categorical Predictor

To this point we have relied on regression models where the predictor variables have been treated as continuous variables. Our purpose in this chapter and the following two is to examine our basic approach to data analysis when predictors are categorical variables. In the language of traditional statistics books, earlier chapters concerned multiple regression. The present chapter concerns one-way analysis of variance (ANOVA) models or, equivalently, models with a single categorical predictor. In the next chapter we consider models having multiple categorical predictors, or higher order ANOVA models. Chapter 10 is devoted to models in which some predictors are categorical and some are continuous. Such models have been traditionally labeled analysis of covariance models. By integrating these into a common approach, we will not only explore these traditional topics but also consider others that considerably extend the sorts of questions that we are able to ask of our data, in the context of categorical predictor variables. Throughout we will continue to use our basic approach to statistical inference, testing null hypotheses by the comparison of augmented and compact models.

THE CASE OF A CATEGORICAL PREDICTOR WITH TWO LEVELS

Figure 8.1 contains hypothetical data from a study in which the impact of a SAT training course was evaluated. Twenty high-school seniors were randomly assigned to either take the 2-week training course, designed to improve SAT performance, or to a control no-course condition. As we can see, 10 students wound up in each of the two groups. At the end of the 2-week period, all 20 students took the SAT test and their performance was recorded. What we would like to do is examine whether the course made a difference in subsequent performance. Our question thus is whether we can reliably predict subsequent SAT performance as a function of whether a student was in the Course group or the No Course control group.

If we are to tell the computer to specify a model in which SAT performance is predicted by which of the two groups a student was in, we need some way of coding or numerically representing the group variable. This variable is categorical rather than numerical or continuous, meaning that students in the two groups differ on whether or not they have taken the course, but no automatic numerical representation of that difference is implied. Such variables require some coding scheme to represent them numerically, so that they can be used as predictor variables in models. It turns out that any numerical representation of this group variable would do, as long as we used that

numerical representation consistently. By consistent use, we mean that if a given value on the variable that represents group (Course versus No Course) numerically is assigned to one group, then every student in that group has that same value on the variable, and no student in the other group has that value.

To illustrate, suppose we created a variable X_i to represent group numerically, arbitrarily assigning the value of -1 to students in the No Course group and $+1$ to students in the Course group. Since every student is in one group or the other, all students have values of either -1 or $+1$ on variable X_i . Notice that our purpose in creating this variable is simply to differentiate numerically between the two groups. Since group is a categorical variable, no rank order or interval information need be preserved in our coding scheme. We could just as easily have given the value of -1 to the Course group and $+1$ to the No Course group. Similarly, we could have given the value of 203 to students in the Course group and the value of -20.5 to students in the No Course group. The point is that the values that represent the categorical variable are arbitrarily defined, but they must be consistently used.

Throughout all of the rest of the book we will use a convention for coding nominal predictors known as *contrast codes*. Contrast codes are simply one of the possible arbitrary coding schemes for numerically representing categorical predictors. Two conditions define contrast codes and differentiate them from other coding schemes. For now, we will only define one of these two conditions. The other is only relevant when the categorical variable has more than two categories and will be given later in this chapter. Let us define a value on a contrast-coded categorical variable X_i as λ_k (“lambda_{*k*}”), where the subscript k refers to the level of the categorical variable being coded. In this case, k refers to the two levels of the group variable: Course versus No Course. Across levels of k , or across all categories of the variable, a contrast code is one where:

$$\sum_k \lambda_k = 0$$

Notice that we are summing here across levels or categories rather than across individual observations. In other words, the condition is that the values of the contrast variable sum to zero across the two categories, not across the individual observations in those two categories.

In our example, the values of $+1$ for students in the Course group and -1 for students in the No Course group constitute values of a contrast-coded variable, since the sum of these two values across the two categories equals zero. Another valid contrast-coded variable would use values of $+0.5$ for the Course group and -0.5 for the No Course group. Notice, however, that the values of 203 for the Course group and -20.5 for the No Course group do not meet the condition for a contrast-coded variable. Note also that in this example the values of a contrast-coded variable, say $+1$ and -1 , sum to zero not only

FIGURE 8.1 Two-group SAT data

<i>Student</i>	<i>Group</i>	<i>SAT</i>
1	Course	580
2	Course	560
3	Course	660
4	Course	620
5	Course	600
6	Course	580
7	Course	590
8	Course	640
9	Course	620
10	Course	600
11	No Course	580
12	No Course	530
13	No Course	590
14	No Course	550
15	No Course	610
16	No Course	590
17	No Course	600
18	No Course	530
19	No Course	590
20	No Course	600

across the two categories but also across the 20 students in those two categories. This will be the case when a contrast-coded variable is used and when there are an equal number of observations in the two groups or categories. Had we had more students in one of the two groups than in the other, then the sum of the values across the two groups would be zero, but the sum of the values across all the students would not be.

In the following sections, we will use two different contrast-coded predictors to predict SAT performance with the data presented in Figure 8.1. We will first use the values of +1 and -1 and then we will use the values of +.5 and -.5. These possible values of a contrast-coded predictor are simply two from an infinite number of such values that might be used. At a later point, we will also briefly discuss the estimation of models using coding conventions other than contrast codes.

Model Estimation and Inference with a Contrast-Coded Predictor

We start by estimating a model in which we predict SAT performance with a contrast-coded predictor having values of +1 for students in the Course group and -1 for students in the No Course group. SAT is thus our Y_i variable and our predictor, X_i , is the contrast-coded predictor. Our model is the simple regression model with a single predictor variable of Chapter 5. We will want to compare this model, making predictions of SAT conditional on X_i , with a compact one in which we predict the same value for every student regardless of whether they were in the Course group or the No Course group:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

Assuming that $\beta_1 \neq 0$, the augmented model makes conditional predictions of SAT performance, conditional on whether students were in the Course group or the No Course group. On the other hand, the compact model makes the same prediction for all students, regardless of their group.

The least-squares estimates for these models are:

$$\text{MODEL A: } \hat{Y}_i = 591 + 14X_i$$

$$\text{MODEL C: } \hat{Y}_i = 591$$

Both of these models, as well as the data on which they are based, are graphed in Figure 8.2. The horizontal prediction function is Model C, while the prediction line that makes differential predictions is Model A. The sums of squared errors of these two models are 16,060 for Model A and 19,980 for Model C.

The calculations of these two sums of squares are shown in Figure 8.3 where we derive for each observation the predicted value from each model, the residual, and the squared residual. The sums of these squared residuals, given in the last row of Figure 8.3, are the sums of squared errors for the two models.

The comparison of these two models, asking whether the predictions conditional on the contrast-coded predictor do a better job than the unconditional predictions, yields:

$$\text{PRE} = \frac{19,980 - 16,060}{19,980} = \frac{3920}{19,980} = .196$$

FIGURE 8.2 Models A and C for the SAT data

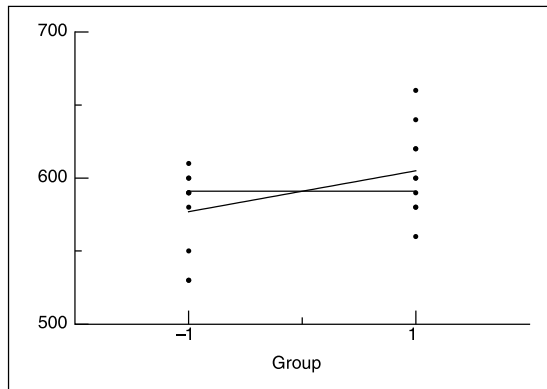


FIGURE 8.3 Calculation of SSE for Models C and A

Student	Group	X_i	SAT (Y_i)	Compact			Augmented		
				\hat{Y}_i	e_i	e_i^2	\hat{Y}_i	e_i	e_i^2
1	Course	1	580	591	-11	121	605	-25	625
2	Course	1	560	591	-31	961	605	-45	2025
3	Course	1	660	591	69	4761	605	55	3025
4	Course	1	620	591	29	841	605	15	225
5	Course	1	600	591	9	81	605	-5	25
6	Course	1	580	591	-11	121	605	-25	625
7	Course	1	590	591	-1	1	605	-15	225
8	Course	1	640	591	49	2401	605	35	1225
9	Course	1	620	591	29	841	605	15	225
10	Course	1	600	591	9	81	605	-5	25
11	No Course	-1	580	591	-11	121	577	3	9
12	No Course	-1	530	591	-61	3721	577	-47	2209
13	No Course	-1	590	591	-1	1	577	13	169
14	No Course	-1	550	591	-41	1681	577	-27	729
15	No Course	-1	610	591	19	361	577	33	1089
16	No Course	-1	590	591	-1	1	577	13	169
17	No Course	-1	600	591	9	81	577	23	529
18	No Course	-1	530	591	-61	3721	577	-47	2209
19	No Course	-1	590	591	-1	1	577	13	169
20	No Course	-1	600	591	9	81	577	23	529
SSE(C) = 19,980							SSE(A) = 16,060		

Model A contains two parameters and Model C one, hence $PA - PC$ equals 1 and $n - PA$ equals 18. Accordingly, we can compute the F statistic for this comparison either from the computed value of PRE or from the values of the sums of squares:

$$F_{1,18} = \frac{\text{PRE}/(PA - PC)}{(1 - \text{PRE})/(n - PA)} = \frac{.196/1}{(1 - .196)/18} = 4.39$$

$$F_{1,18} = \frac{\text{SSR}/(PA - PC)}{\text{SSE(A)}/(n - PA)} = \frac{3920/1}{16,060/18} = 4.39$$

These statistics fall just short of their critical values, with 1 and 18 degrees of freedom, with α set at .05. Hence, we are unable to conclude that the conditional predictions of SAT are significantly better than the unconditional one made by Model C.

So far, once we have coded our categorical predictor, there is nothing different about this simple regression model from those simple models using continuous predictors that were discussed in Chapter 5. The estimation of the model parameters and the calculations of PRE and F proceed just as before. Since all of the assumptions underlying the use of the critical values of PRE and F involve assumptions about the distribution of Y_i , or really of ε_i , and since the categorical nature of X_i has no effect on the distribution of ε_i , none of the assumptions underlying this analysis have been impacted by the use of the categorical predictor.

What has changed somewhat, however, is the interpretation of the estimated parameters of the model and, accordingly, the interpretation of the statistical inference results. It is not the case that the old interpretations (those developed in Chapter 5) are incorrect, for the value of the intercept in the augmented model, 591, is still the predicted value of Y_i when X_i equals zero. The coefficient for X_i , 14, is still a slope—the amount by which the predicted value of Y_i changes for each unit increase in X_i . And PRE and F still tell us about the reduction in errors of prediction. Rather, when we have categorical predictors, new interpretations of these statistics become possible.

To understand these new interpretations, it is helpful to consider the predicted values that the augmented model makes. These are contained in Figure 8.3. If we are dealing with a student in the Course group, the predicted value from the augmented model is:

$$\hat{Y}_{+1} = 591 + 14(+1) = 605$$

And if we are dealing with a student in the No Course group, the predicted value from the augmented model is:

$$\hat{Y}_{-1} = 591 + 14(-1) = 577$$

These predicted values turn out to be the mean SAT scores of the 10 students in each of the two groups. That is, 605 is the mean SAT score of those students who received the course, and 577 is the mean SAT score of those students who did not. And given our use of +1 and -1 as the values of the contrast-coded predictor, the slope associated with that predictor equals half the difference between the means of the two groups:

$$\begin{aligned} \bar{Y}_C - \bar{Y}_{NC} &= \hat{Y}_{+1} - \hat{Y}_{-1} = 605 - 577 \\ &= (591 + 14(+1)) - (591 + 14(-1)) \\ &= 14(+1) - 14(-1) \\ &= 2(14) \end{aligned}$$

In general, the least-squares parameter estimate or slope associated with a contrast-coded predictor is given by:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

which, in the example at hand, is evaluated as:

$$\frac{(-1)577 + (+1)605}{(-1)^2 + (+1)^2} = \frac{28}{2} = 14$$

This is a very general and useful formula for interpreting the slope associated with any contrast-coded predictor. The numerator represents a comparison between group means, in this case the difference between the mean for the Course group and that for the No Course group, and the denominator is a scaling factor dependent on the specific values used for the contrast-coded predictor. The important point is that the regression coefficient associated with any contrast-coded predictor tells us about the difference between group means, the direction of that difference being determined by which group is coded with a positive value and which group is coded with a negative value.

The estimated intercept of 591 equals, as always, the predicted value of Y_i when X_i equals zero. Since X_i equals zero halfway between the two values of +1 and -1 that code the two groups, the estimated value of the intercept is necessarily equal to the average of the two group means. This result is made clear by the graph of the model in Figure 8.2. It can also be shown algebraically as follows:

$$\hat{Y}_{+1} = \bar{Y}_C = 591 + 14(+1)$$

$$\hat{Y}_{-1} = \bar{Y}_{NC} = 591 + 14(-1)$$

Adding these two equalities gives:

$$\bar{Y}_C + \bar{Y}_{NC} = (591 + 14(+1)) + (591 + 14(-1))$$

$$\bar{Y}_C + \bar{Y}_{NC} = (2)591$$

$$\frac{\bar{Y}_C + \bar{Y}_{NC}}{2} = 591$$

Notice that this interpretation of the intercept in the augmented model, including the contrast-coded predictor, is not the same as the interpretation of the intercept in the compact model, the one making unconditional predictions. The intercept in the compact model is estimated as the mean of all the observations, what we might call the grand mean, \bar{Y} , defined as:

$$b_{0C} = \bar{Y} = \frac{\sum_i Y_i}{n}$$

On the other hand, the intercept in the augmented model is the mean of the group means, defined as:

$$b_{0A} = \frac{\sum_k \bar{Y}_k}{m}$$

where m is the total number of groups, in this case 2.

In the dataset that we have been using, the values of these two intercepts, one from the compact model and one from the augmented, are identical, i.e., 591, because there

are an equal number of students in the two groups. In general, however, they estimate different things. The intercept in the compact model, the one making unconditional predictions, is the mean of all the observations. The intercept in the augmented model, the one making predictions conditional on group, is the mean of the group means.

One last result when using contrast-coded predictors is important to know. Just as the regression coefficient for any contrast-coded predictor can be represented as a comparison among group means:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

so too can the SSR associated with any such predictor be similarly expressed. As always, the SSR associated with a predictor is the difference between the SSE(A) and SSE(C) for compact and augmented models with and without that predictor. And, as always, that SSR equals:

$$\text{SSR} = \sum (\hat{Y}_{iA} - \hat{Y}_{iC})^2$$

In the case of a categorical predictor, as we have seen, the predicted values for the augmented model in this expression are the group or category means, \bar{Y}_k , and the predicted value from the compact model is the grand mean of all the observations, \bar{Y} . As a result, it is possible to show that the SSR associated with any contrast-coded predictor can be expressed as a function of the category means (and the number of observations in each group, n_k) as follows:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k \right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

In the case at hand, we have seen that SSE(C) equals 19,980 and SSE(A) equals 16,060, resulting in an SSR associated with the contrast-coded predictor of 3920. That is equivalently computed as:

$$\frac{((-1)577 + (+1)605)^2}{((-1)^2/10) + ((+1)^2/10)} = 3920$$

Estimation with Alternative Values for the Contrast-Coded Predictor

If we had defined the values of the contrast-coded predictor as +.5 for the Course group and -.5 for the No Course group, rather than +1 and -1, the estimated model would be:

$$\text{MODEL A: } \hat{Y}_i = 591 + 28X_i'$$

where X_i' is the new contrast-coded predictor. Importantly, this model makes exactly the same predictions for students in the two groups, i.e., their respective group means:

$$\hat{Y}_{+.5} = 591 + 28(+.5) = 605 = \bar{Y}_C$$

$$\hat{Y}_{-.5} = 591 + 28(-.5) = 577 = \bar{Y}_{NC}$$

Accordingly, in a deep sense, it is the same augmented model with the same sum of squared errors. The slope for the contrast-coded predictor now equals 28 instead of 14, since now there is a one-unit difference on X_i —that separates the two groups (between $-.5$ and $+.5$) rather than the two-unit difference that separated them on X_i (between $+1$ and -1). And that slope now equals the difference between the two group means:

$$b_1 = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{(-.5)577 + (+.5)605}{(-.5)^2 + (+.5)^2} = \frac{.5(605 - 577)}{.5} = 605 - 577 = 28$$

Of course, since this Model A is in a deep sense the same as the one where the contrast-coded predictor had values of $+1$ and -1 , the model comparison of it with the compact model, making predictions that were not conditional on group, yields the exact same SSR, PRE and F statistics. Recomputing the SSR for this comparison, using these new codes and the formula given for the SSR in terms of the group means, gives us:

$$\frac{((-5)577 + (+5)605)^2}{((-5)^2/10) + ((+5)^2/10)} = 3920$$

Equivalence with ANOVA and Two-Group t -Test

The slope in the augmented model, making conditional predictions, informs us about the difference between the two group means (regardless of the values of λ_k used to construct the contrast-coded predictor). Therefore, the comparison of this augmented model with the compact one making unconditional predictions asks both whether the parameter associated with the contrast-coded predictor departs from zero and whether the two group means differ from each other. In other words, the model comparison we have examined addresses the following equivalent null hypotheses:

$$H_0: \beta_1 = 0$$

$$H_0: \mu_C = \mu_{NC}$$

where μ_C and μ_{NC} are the true but unknown means of the two groups.

In more traditional statistical textbooks, the test of a null hypothesis about the difference between two group means is usually conducted by computing a two-group ANOVA or a two-group independent samples t -test. It is therefore important to show that our model comparison and its associated PRE and F statistics are identical to those generated by these more traditional approaches.

In Figure 8.4 we give the ANOVA source table for the model comparison we have just conducted, using the formulas developed in Chapter 5 and earlier. We also provide the source table using the numeric values generated by our data example.

The formula for the SSR for the model comparison that we have given before is:

$$SSR = \sum_i (\hat{Y}_A - \hat{Y}_C)^2$$

FIGURE 8.4 ANOVA source tables

Source	SS	df	MS	F	p	PRE
Reduction	SSR	PA – PC	MSR	$\frac{MSR}{MSE(A)}$		$\frac{SSR}{SSE(C)}$
Error	SSE(A)	n – PA	MSE(A)			
Total	SSE(C)	n – 1	MSE(C)			

Source	SS	df	MS	F	p	PRE
Reduction	3920	1	3920.00	4.39	< .10	.196
Error	16060	18	892.22			
Total	19980	19	1051.58			

summing across all individual observations. In the present case, \hat{Y}_C is the mean of all of the observations, \bar{Y} , and the predicted values from the augmented model, \hat{Y}_A , are the two group means \bar{Y}_C and \bar{Y}_{NC} . Let us generically represent these group means as \bar{Y}_k , thereby indicating the group mean for the k th group. Then we can write the above expression for the sum of squares reduced as:

$$SSR = \sum_i (\hat{Y}_A - \hat{Y}_C)^2 = \sum_i (\bar{Y}_k - \bar{Y})^2 = \sum_k n_k (\bar{Y}_k - \bar{Y})^2$$

where n_k is the number of observations in the k th group.

We can also re-express the formula for SSE(A) and SSE(C) in terms of means, since those are the predicted values from each model:

$$SSE(A) = \sum_i (Y_i - \hat{Y}_{iA})^2 = \sum_i (Y_i - \bar{Y}_k)^2$$

$$SSE(C) = \sum_i (Y_i - \hat{Y}_{iC})^2 = \sum_i (Y_i - \bar{Y})^2$$

Accordingly, the formulas in the source table that we have used all along for summarizing our computations of PRE and F (given in the top half of Figure 8.4) can, in this case, be equivalently written with the formulas used to compute an analysis of variance to compare group means in traditional statistics textbooks. This revised version of the source table is given in Figure 8.5.

The names given to the rows in this version of the source table have been changed to reflect those traditionally used in analysis of variance. So, the sum of squares reduced and its mean square are traditionally called the sum of squares and mean square *between groups*, and the sum of squares and mean square from Model A are traditionally called the sum of squares and mean square *within groups*. But fundamentally and algebraically this is the same source table, yielding the exact same F and PRE, as the one we are more used to, given in Figure 8.4.

The square root of F is the t statistic (since $n - PA = 1$), with $n - 2$ degrees of freedom, that is traditionally called the two-group independent samples t -test.

All of this is simply to demonstrate that our integrated approach to statistical inference, resting on model comparisons estimated with any least-squares multiple

FIGURE 8.5 ANOVA source table

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>PRE</i>
Between	$\sum_k n_k (\bar{Y}_k - \bar{Y})^2$	1	MSB	$\frac{MSB}{MSW}$		$\frac{SSB}{SST}$
Within	$\sum_i (Y_i - \bar{Y}_k)^2$	$n - 2$	MSW			
Total	$\sum_i (Y_i - \bar{Y})^2$	$n - 1$	MST			

regression program, yields the same results as the cookbook recipes given in more traditional statistical textbooks. Our model comparison in this case, testing whether the regression coefficient for a single contrast-coded predictor departs from zero, is exactly equivalent to a two-sample *t*-test for examining whether the means of two groups differ from each other.

Confidence Interval for the Slope of a Contrast-Coded Predictor

The formula that we gave in Chapter 6 for the confidence interval for a regression coefficient continues to be applicable in the situation where predictor variables are contrast-coded. The confidence interval for the slope associated with any predictor variable was given there as:

$$b \pm \sqrt{\frac{F_{\text{crit}} \text{MSE}}{(n-1)s_x^2(\text{tol})}}$$

Since the slope of a contrast-coded predictor informs us about the magnitude of the difference between group means, so its confidence interval also informs us about the confidence interval associated with that mean difference. To see this, let us take the case where the contrast-coded predictor used values of -0.5 for the No Course group and $+0.5$ for the Course group. The resulting slope in this case equaled 28 and the confidence interval for that slope is calculated as:

$$28 \pm \sqrt{\frac{4.41(892.22)}{19(0.263)1}}$$

$$28 \pm 28.06$$

where 4.41 is the critical *F* value with 1 and 18 degrees of freedom, 892.22 is the mean square error from our Model A, $n - 1$ equals 19, the variance of the 20 individuals on the contrast-coded predictor is 0.263, and its tolerance is of course 1 since it is the only predictor in the model. This confidence interval can also be written as:

$$-0.06 \leq \beta_1 \leq 56.06$$

Since the parameter that is being estimated here, with this contrast-coded predictor, is also an estimate of the true mean difference between the two groups, this confidence interval can be equivalently expressed as

$$-0.06 \leq \mu_C - \mu_{NC} \leq 56.06$$

When the contrast-coded predictor used the values of -1 and $+1$ rather than $-.5$ and $+.5$, the estimated slope was half the difference in the group means (i.e., 14) and its confidence interval is computed as:

$$14 \pm \sqrt{\frac{4.41(892.22)}{19(1.052)1}}$$

$$14 \pm 14.03$$

The term that is different in this interval (other than the value of the estimated slope itself) is the variance of the predictor (coded $+1$ and -1), which is 1.052 rather than 0.263.

In this case, the confidence interval is given equivalently as:

$$-0.03 \leq \beta_1 \leq 28.03$$

$$-0.03 \leq \frac{\mu_C - \mu_{NC}}{2} \leq 28.03$$

Thus, it continues to tell us about the confidence interval for the mean difference, except with these codes of course it is the confidence interval for half the mean difference. If we multiply this expression by 2, we get the confidence interval for the mean difference.

CATEGORICAL PREDICTORS WITH MORE THAN TWO LEVELS

Suppose a developmental psychologist is interested in the effects of feedback about performance on subsequent motivation to do a task. She hypothesizes that subsequent motivation will decline if children are told that they earlier failed at the task. To test this hypothesis, she randomly assigns children to three conditions; in one condition they are told that they failed on the task; in a second condition they are given no feedback; and in a third condition they are told they succeeded. The experimenter then monitors the number of tasks they subsequently complete, after the differential feedback has been given. Twenty-four children are run in total, eight in each of the three conditions. The hypothetical raw data are given in Figure 8.6.

FIGURE 8.6 Hypothetical experimental data for three conditions (values represent number of tasks each subject completes)

	<i>Failure</i>	<i>No Feedback</i>	<i>Success</i>
	2	4	4
	2	3	6
	2	4	5
	3	5	4
	4	5	6
	4	2	4
	3	4	3
	4	3	3
\bar{Y}_k	3.000	3.750	4.375

Contrast Codes for Multilevel Categorical Predictors

In order to examine the effects of feedback on the number of tasks subsequently completed, we need to derive a coding scheme to represent the three levels of the categorical feedback variable. We might think that a single variable that codes all three conditions would be appropriate, giving observations from the Success condition a higher value on the variable than observations for the No Feedback condition who in turn receive a higher value than observations from the Failure condition. We could then see if such a coded variable would be predictive of Y_i . The problem with using a single variable to code the three levels of this categorical variable is that with such a coding scheme we are assuming that the categories can be ordered in an a priori manner and that the relationship between the values of the single predictor variable and Y_i is a linear one. While we may have a reason for expecting that Y_i should be lower in the Failure condition than in the other two, we do not have any reason for assigning particular values to the groups, expecting linear predictions as a function of those particular values. In other words, a single-predictor variable that codes the three conditions with particular values does not make much sense, given that we are dealing with a categorical variable whose levels do not differ in a neat linear way.

To ask whether we can predict Y_i as a function of some categorical variable having in general m levels or groups is equivalent to asking whether there are differences among the m group means (\bar{Y}_k) across those levels (with k varying from 1 to m). To answer this question, we need to employ $m - 1$ contrast-coded predictor variables in our model. We will then be able to ask about mean differences among the groups, allowing for all possible orderings of those means. To define these $m - 1$ contrast-coded predictors, it is now time to introduce the second defining condition for contrast codes. The first condition, you will recall, was that for a contrast-coded variable the sum of the λ values across the groups or levels of the categorical variable must equal zero: $\sum_k \lambda_k = 0$. When we use more than a single contrast-coded predictor to code a categorical variable having more than two levels, the second condition that must be met is that across levels of the categorical variable all pairs of contrast codes must be orthogonal to each other. Given that the first condition is met, this second condition of orthogonality will be met whenever the sum (across k or the levels of the categorical variable) of the products of the λ values from pairs of contrast codes equals zero.

In our example, we have three levels of the categorical variable. We will therefore use two contrast codes to code it. Each value of λ now has two subscripts, the first one designating which contrast code we are talking about and the second one designating the level of the categorical variable (k). The condition of orthogonality is met when:

$$\sum_k \lambda_{1k} \lambda_{2k} = 0$$

To make this second defining condition of contrast codes more understandable, let us illustrate codes that do and do not meet it for the example at hand. In Figure 8.7, two sets of codes, with two codes in each set, are given for coding the three levels of the categorical predictor variable: Failure, No Feedback, and Success.

Each of the four codes meets the first defining condition for a contrast code, in that the sum of the λ values for any given code, computed across the three levels of the categorical variable, equals zero. The second defining condition, however, is only met

FIGURE 8.7 Sets of codes for a three-level categorical predictor

	<i>Failure</i>	<i>No Feedback</i>	<i>Success</i>
Set A			
λ_{1k}	-2	1	1
λ_{2k}	0	-1	1
Set B			
λ_{1k}	-1	0	1
λ_{2k}	0	-1	1

by the codes in Set A. If we multiply the value of λ_{1k} by the value of λ_{2k} at each of the levels of the categorical predictor variable and then we add up the resulting three products, we get a sum of 0 from Set A (i.e., $0 + (-1) + 1 = 0$) and a sum of 1 from Set B (i.e., $0 + 0 + 1 = 1$). Accordingly, only the codes in Set A can legitimately be called contrast codes.

This second defining condition means that a given code cannot be defined as a contrast code in isolation. We could not, for instance, look at the code for λ_{1k} in Set A and identify it as a contrast code, unless we looked at the other code or codes with which it is used in combination to code the categorical predictor variable. For instance, if we changed the values of λ_{2k} in Set A to be -1, -1, and 2 for Failure, No Feedback, and Success, respectively, then the codes in Set A would no longer be contrast codes, even though we had not changed the values of λ for the first code. This set of codes would no longer be contrast codes since the sum of the products of the λ values across the category levels would no longer equal zero.

If our categorical predictor variable had four levels, we would need three contrast codes to code it completely. The second defining condition for contrast codes would be met in such a case if the sums of the products of the λ values for all possible pairs of codes equaled zero. Suppose, for instance, that we had a categorical variable with four levels, as in Figure 8.8. There we define three contrast codes with values of λ_{1k} , λ_{2k} , and λ_{3k} . We then have three pairs of codes, and for each of these pairs the sum of the products of the λ values must equal zero. For codes 1 and 2, the sum of the products of the λ_{jk} values equals $(-3)0 + 1(-2) + 1(1) + 1(1) = 0$. For codes 1 and 3, the sum of the products of the λ_{jk} values equals $(-3)0 + 1(0) + 1(-1) + 1(1) = 0$. And for codes 2 and 3, the sum of the products of the λ_{jk} values equals $0(0) + 0(-2) + 1(-1) + 1(1) = 0$.

With a categorical predictor having three levels, then, we need two contrast codes and a single sum of products of λ values must equal zero. With a categorical predictor having four levels, we need three contrast codes. Those three codes result in three possible pairs of codes, and hence three sums of products of λ values must equal zero. In general, with a categorical variable having m levels, we need $m - 1$ contrast codes to code it. From these $m - 1$ contrast codes, there are $(m - 1)(m - 2)/2$ pairs of codes. This many sums of products of λ values must equal zero to meet the second defining condition of contrast codes.

For any given categorical predictor, there are an infinite number of sets of contrast codes that could be used. The choice of codes to be used should be guided by some theoretical or substantive notions about how the groups defined by the categorical predictor variable are expected to differ on the dependent variable. For instance, in the

FIGURE 8.8 Codes for a four-level categorical predictor

	Level 1	Level 2	Level 3	Level 4
λ_{1k}	-3	1	1	1
λ_{2k}	0	-2	1	1
λ_{3k}	0	0	-1	1

illustration at hand, we expected subjects in the Failure condition to have lower scores than subjects in the other two conditions. Since, as we saw in the case of a categorical predictor with only two levels, the regression coefficient for a contrast code tells us about the relative mean difference between observations having different values on the contrast code, it makes sense to derive a code that will allow us to examine this prediction about mean differences on the dependent variable. In other words, given that we want to see whether the observations in the Failure condition have lower scores than observations in the other two conditions, the first contrast code we gave in Set A of Figure 8.7 is one that we may well choose to examine.

As was the case with a single contrast-coded predictor that codes a categorical variable with two levels, the regression coefficient associated with a contrast-coded predictor in the case of a categorical variable with more than two levels tells us about mean differences among the various groups or levels of the categorical variable, according to the following formula:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

But this will be the case *only* if a complete set of $m - 1$ coded predictors are included in the model and *only* if the contrast codes used meet the orthogonality condition that we have just defined. At a later point in this chapter we will discuss estimation in the presence of nonorthogonally coded predictors. For now, the important point is that slopes tell us about coded mean differences among the categories only with a complete set of codes and only with orthogonality.

Based on theoretical considerations, then, we are interested in the comparison that is made by the first contrast code of Set A in Figure 8.7. With three levels of our categorical variable and one code chosen, the second code is constrained to be one that compares the means in the No Feedback and Success conditions, like the second code in Set A. In general, with m levels of a categorical variable and $m - 1$ contrast codes, the final code is constrained once the first $m - 2$ codes have been defined, in order to meet the orthogonality condition.

We can use these two codes to define two predictor variables, X_{1i} based on the codes -2, 1, 1 (for Failure, No Feedback, and Success respectively) and X_{2i} based on the codes 0, -1, 1, and then estimate a multiple regression model in which these are used as simultaneous predictors of Y_i . If we did this, we can exactly specify the mean differences estimated by the two resulting slopes using the formula for the slope of a contrast-coded predictor in the context of a model with a full set of orthogonal contrast-coded predictors:

$$b_{X_1} = \frac{\sum_k \lambda_{1k} \bar{Y}_k}{\sum_k \lambda_{1k}^2} = \frac{(-2)\bar{Y}_F + (1)\bar{Y}_{NF} + (1)\bar{Y}_S}{(-2)^2 + (1)^2 + (1)^2} = \frac{(\bar{Y}_{NF} + \bar{Y}_S) - 2\bar{Y}_F}{6} = \frac{\left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F}{3}$$

$$b_{X_2} = \frac{\sum_k \lambda_{2k} \bar{Y}_k}{\sum_k \lambda_{2k}^2} = \frac{(0)\bar{Y}_F + (-1)\bar{Y}_{NF} + (1)\bar{Y}_S}{(0)^2 + (-1)^2 + (1)^2} = \frac{\bar{Y}_S - \bar{Y}_{NF}}{2}$$

In general, such slopes will inform us about differences among category means following the codes used, with the numerator of the above expressions representing the mean difference, and the denominator representing a scaling factor. Notice that group means for levels of the categorical variable that are coded with a zero value of λ on a particular contrast-coded predictor drop out of the numerator of the slope and thus do not figure in the comparison that is made (i.e., the group mean for the Failure condition does not play a role in the slope of the second contrast-coded variable).

To show the impact of the scaling factor in the denominator of the slope expression, had we used fractional values for λ ($-\frac{2}{3}$, $\frac{1}{3}$, $\frac{1}{3}$ for X'_{1i} , and 0 , $-\frac{1}{2}$, $\frac{1}{2}$ for X'_{2i}) rather than those defined above, then the following would be the values of the slopes:

$$b_{X'_1} = \frac{\sum_k \lambda_{1'k} \bar{Y}_k}{\sum_k \lambda_{1'k}^2} = \frac{(-\frac{2}{3})\bar{Y}_F + (\frac{1}{3})\bar{Y}_{NF} + (\frac{1}{3})\bar{Y}_S}{(-\frac{2}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2} = \left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F$$

$$b_{X'_2} = \frac{\sum_k \lambda_{2'k} \bar{Y}_k}{\sum_k \lambda_{2'k}^2} = \frac{(0)\bar{Y}_F + (-\frac{1}{2})\bar{Y}_{NF} + (\frac{1}{2})\bar{Y}_S}{(0)^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2} = \bar{Y}_S - \bar{Y}_{NF}$$

The advantage of such fractional codes is that their slopes will equal the mean differences rather than fractions of mean differences.

Without practice, it may seem difficult to come up with a set of orthogonal contrast codes, particularly when dealing with a categorical variable having more than three or so levels. Our advice is that one should initially create codes that represent mean comparisons one would like to make theoretically, and then derive the remainder of the codes so that orthogonality is preserved. One way to do this, once one or more initial codes have been defined, is to construct further contrast codes that compare category means that were tied (or received the same value of λ) on already used code(s). With some practice, deriving orthogonal codes becomes a relatively easy task.

In the absence of any motivated comparisons, one can always use a convention called Helmert codes, regardless of the number of levels. A simple algorithm generates such codes. If there are m levels of the categorical variable, one defines the first of $m - 1$ contrast codes by assigning the value of $m - 1$ to the first level and the value of -1 to each of the remaining $m - 1$ levels. For the second contrast code, the first level is given

FIGURE 8.9 Helmert contrast codes

Code	Category level						
	1	2	3	...	$m-2$	$m-1$	m
λ_{1k}	$m-1$	-1	-1	...	-1	-1	-1
λ_{2k}	0	$m-2$	-1	...	-1	-1	-1
λ_{3k}	0	0	$m-3$...	-1	-1	-1
\vdots	\vdots	\vdots	\vdots	...	\vdots	\vdots	\vdots
λ_{m-2k}	0	0	0	...	2	-1	-1
λ_{m-1k}	0	0	0	...	0	1	-1

the value of 0, the second level is given the value of $m - 2$, and all remaining levels are given the value of -1 . For the third contrast code, the first two levels of the categorical predictor are assigned values of 0, the third level is given the value of $m - 3$, and the remaining levels are given the value of -1 . One proceeds in this manner to define all $m - 1$ contrast codes, with the last one having values of 0 for all levels of the predictor variable except for the last two. These last two levels have values of 1 and -1 . The resulting code values are presented in Figure 8.9.

Estimation and Inference with Multilevel Categorical Predictors

Using the data in Figure 8.6, we estimated the parameters of the following multiple regression model, with X_{1i} and X_{2i} as contrast-coded predictors, given the values of λ defined by Set A in Figure 8.7:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

The parameter estimates are:

$$\text{MODEL A: } \hat{Y}_i = 3.7083 + .3542X_{1i} + .3125X_{2i}$$

and the sum of squared errors is 23.375.

Unsurprisingly, the predicted values from this model are the means of the three categories (given in Figure 8.6) of the categorical independent variable:

$$\hat{Y}_F = 3.7083 + .3542(-2) + .3125(0) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .3542(1) + .3125(-1) = 3.750$$

$$\hat{Y}_S = 3.7083 + .3542(1) + .3125(1) = 4.375$$

As we have said before, a model with a categorical independent variable will make predictions of the group or category level means whenever a complete set of $m - 1$ codes is used as predictors.

We have already discussed the interpretations of the two parameter estimates associated with the contrast-coded predictors in terms of the category means. Let us revisit these interpretations now that we have the numerical estimates:

$$b_{X_1} = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{\left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F}{3} = \frac{\left(\frac{3.750 + 4.375}{2}\right) - 3.000}{3} = .3542$$

$$b_{X_2} = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{\bar{Y}_S - \bar{Y}_{NF}}{2} = \frac{4.375 - 3.750}{2} = .3125$$

And just as we found with a categorical predictor with two levels, the estimated intercept in this model equals the mean of the three category means:

$$b_0 = \frac{\sum_k \bar{Y}_k}{m} = \frac{3.000 + 3.750 + 4.375}{3} = 3.7083$$

Although these interpretations for the regression coefficients in models with contrast-coded predictors are typically the most useful, interpretations we gave earlier for parameter estimates in multiple regression models continue to be entirely appropriate. Thus slopes of a predictor can be interpreted as differences in \hat{Y}_i values as the predictor increases by one unit, holding constant other predictors. In the case of the slope for X_{1i} , as we move from a score of -2 (for the Failure condition) to a score of $+1$ (for the No Feedback and Success conditions) the predicted values go from the mean of the Failure condition (3.000) to the means in the No Feedback and Success conditions (3.750 and 4.375). Thus, for a three-unit increase in X_{1i} , we go from a predicted value of 3.000 to one of 4.0625, meaning that the increase in predicted values for a one-unit increase in X_{1i} is .3542. And for X_{2i} , as we go from a score of -1 to 1 , the predicted value goes from 3.75 to 4.375. Accordingly, per unit increase in X_{2i} , we predict a .3125 increase in \hat{Y}_i . And finally, the intercept equals the predicted value when both contrast-coded predictors equal zero. When do these predictors equal zero? From the first condition used to define contrast codes, the mean of each contrast code, across categories, equals zero. Accordingly, the intercept is the predicted value for the average of the categories.

There are, of course, many Model Cs with which we can compare this model to test various null hypotheses. One obvious comparison is with the single-parameter simplest model, estimating just the intercept:

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

and predicting the grand mean, \bar{Y} , for all the observations. Since, in this example, each of the levels of the categorical variable has the same number of observations, the overall grand mean of the 24 observations is the same as the mean of the category means. Hence, it is the case that the estimated parameter in this Model C is identical to the intercept in the three-parameter Model A with which we are comparing it:

$$\text{MODEL C: } \hat{Y}_i = 3.7083$$

This estimated Model C has a sum of squared errors of 30.9583.

What exactly is the null hypothesis that is tested by this model comparison? Obviously it is that the two predictors have slopes of zero, that is, that using them as predictors does nothing to improve the quality of our predictions:

$$H_0: \beta_1 = \beta_2 = 0$$

But this null hypothesis can also be expressed in terms of the equality of the category means, since Model C predicts the grand mean, \bar{Y} , for every observation and Model A makes predictions that are conditional on category membership, predicting the category mean, \bar{Y}_k , for each observation. Accordingly, the null hypothesis can equivalently be expressed as:

$$H_0: \mu_F = \mu_{NF} = \mu_S$$

where these are the true but unknown means of the three levels of the categorical independent variable.

The comparison of these two models yields the following values of PRE and F :

$$\text{PRE} = \frac{30.9583 - 23.3750}{30.9583} = .245$$

$$F_{2,21} = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} = \frac{.245/2}{(1 - .245)/21} = 3.406$$

$$F_{2,21} = \frac{\text{SSR}/(\text{PA} - \text{PC})}{\text{SSE}(\text{A})/(n - \text{PA})} = \frac{7.583/2}{23.375/21} = 3.406$$

And these come just short of beating the critical values for 2 and 21 degrees of freedom. Hence, we cannot reject the null hypothesis that there are no mean differences among these three categories or conditions.

This conclusion does not mean, of course, that we should accept the null hypothesis of no mean differences. And in this case, since we clearly had an expectation that the mean in the Failure condition would be less than the mean in the other two conditions, we should certainly proceed to directly test that hypothesis, which is the comparison made by the first contrast-coded predictor. Within the analysis of variance tradition, it is sometimes maintained that one should not test specific focused comparisons among category means unless the overall multiple-degree-of-freedom test that we have just conducted—that there are no mean difference among the categories—is rejected. We strongly disagree with this point of view. For reasons we have explained earlier, we are generally not enamored of model comparisons where $\text{PA} - \text{PC}$ is > 1 . One of the distinct advantages of a regression-based approach to traditional analysis of variance procedures is that one is forced to construct individual one-degree-of-freedom comparisons or contrasts among group means. Many traditional ANOVA programs automatically provide only the omnibus, multiple-degree-of-freedom test, and this, we think, is a distinct disservice.

As we have seen, the regression coefficients for X_{1i} and X_{2i} estimate particular differences among category means, the first comparing the Failure mean with the average of the No Feedback and Success means, and the second comparing the No Feedback and Success means. Hence, model comparisons that test whether these two parameters depart from zero are equivalently tests of mean differences among the three categories. Specifically, one model comparison is whether the parameter associated with X_{1i} equals zero:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$\text{MODEL C1: } Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$$

with the following equivalent null hypotheses:

$$H_0: \beta_1 = 0$$

$$H_0: \mu_F = (\mu_{NF} + \mu_S)/2$$

And the other model comparison tests whether the parameter associated with X_{2i} equals zero:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$\text{MODEL C2: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

with the following equivalent null hypotheses:

$$H_0: \beta_2 = 0$$

$$H_0: \mu_{NF} = \mu_S$$

Model A for both of these comparisons is the same three-parameter augmented model that we estimated earlier, with a sum of squared errors of 23.375. Model C1 for the first comparison is estimated as follows:

$$\text{MODEL C1: } \hat{Y}_i = 3.7083 + .3125X_{2i}$$

with a sum of squared errors of 29.396. And Model C2 for the second comparison is estimated as follows:

$$\text{MODEL C2: } \hat{Y}_i = 3.7083 + .3542X_{1i}$$

with a sum of squared errors of 24.937. Note that the estimated intercept and slope in these models are unchanged from what they were in the Model A with both predictors. This results from the conjunction of two conditions: First, we have employed contrast-coded predictors, which by definition are orthogonal at the level of the three categories. Second, we have an equal number of observations in each of the three conditions. As a result of these two conditions, the contrast-coded predictors are uncorrelated with each other across the 24 individual observations. Their tolerance in Model A is 1.00.

The first model comparison, asking whether β_1 differs from zero, yields the following PRE and F statistics:

$$\text{PRE} = \frac{29.396 - 23.375}{29.396} = .205$$

$$F_{1,21} = 5.41$$

This F statistic exceeds the critical value of F with α at .05. Hence, we conclude that β_1 differs significantly from zero. Equivalently, we conclude that the mean value of Y_i in the Failure condition is significantly different from the average of the mean values in the Success and No Feedback conditions. Since the sample mean in the Failure condition is less than the average of the other two sample means, we conclude that Failure feedback in this study decreases subsequent performance relative to Success and No Feedback.

The test of the second null hypothesis, that β_2 equals zero, yields the following PRE and F statistics:

$$\text{PRE} = \frac{24.937 - 23.375}{24.937} = .063$$

$$F_{1,21} = 1.40$$

Since this F does not exceed its critical value, we conclude that β_2 does not differ significantly from zero. Equivalently, we cannot conclude that the mean performance under the Success condition is different from that under No Feedback.

Earlier in this chapter we gave a general formula for the SSR due to a contrast-coded predictor expressed in terms of the category means:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k \right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

This expression for the SSR of a contrast-coded predictor continues to apply in the case of categorical variables with more than two levels, as long as a full set of $m - 1$ contrast-coded predictors is included in Model A. Thus, in the present case, we have seen that the SSR for the Model A/Model C1 comparison that tested whether β_1 equaled zero was equal to:

$$\text{SSR}_{X_1} = \text{SSE}(C1) - \text{SSE}(A) = 29.396 - 23.375 = 6.021$$

This can be obtained equivalently in terms of the category means as:

$$\frac{((-2)3.00 + (+1)3.750 + (+1)4.375)^2}{(-2)^2/8 + (+1)^2/8 + (+1)^2/8} = 6.021$$

Likewise, we saw that the SSR for the Model A/Model C2 comparison that tested whether β_2 equaled zero was equal to:

$$\text{SSR}_{X_2} = \text{SSE}(C2) - \text{SSE}(A) = 24.937 - 23.375 = 1.562$$

This can be obtained equivalently in terms of the category means as:

$$\frac{((0)3.00 + (-1)3.750 + (+1)4.375)^2}{(0)^2/8 + (-1)^2/8 + (+1)^2/8} = 1.562$$

We have now done three different tests comparing the augmented model, which includes both contrast-coded predictors, with three different compact ones. The results of these three tests are presented in Figure 8.10. Notice that we have given labels, in parentheses, for each of these tests to indicate the questions they are examining in terms of the group means. The two-degree-of-freedom test, done first, comparing Model A to a Model C that predicted the grand mean for all observations, was an omnibus test of any group mean differences. The second, comparing models with and without X_{1i} as a predictor, examined whether the mean in the Failure condition differed from the average of the two in the other conditions. And the third, comparing models with and without X_{2i} as a predictor, examined whether the means in the No Feedback and Success conditions differed. We want to emphasize again that even though the two-degree-of-

FIGURE 8.10 Summary source table

<i>Source</i>	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model (between conditions)		7.583	2	3.792	3.406	.245
X_1 (Failure vs. No Feedback, Success)	.3542	6.021	1	6.021	5.409	.205
X_2 (No Feedback vs. Success)	.3125	1.562	1	1.562	1.403	.063
Error		23.375	21	1.113		
Total		30.958	23			

freedom test did not prove to be significant, we did find a significant mean difference when we tested the more focused contrast question represented by X_{1i} . As always, we strongly encourage focused PA – PC = 1 model comparisons.

As the sums of squares in this source table show, the SSRs for the individual predictor variables sum to the SSR for the first model comparison, where the overall augmented model was compared to a compact single-parameter model, predicting the grand mean for all observations. As we saw in Chapter 6, this will be the case whenever predictors are completely nonredundant, with tolerances of 1.0. In the present case, this results from the conjunction of two conditions: the use of contrast-coded predictors, which are by definition orthogonal at the level of the groups or categories; and the fact that each category contains the same number of observations.

A Quick Look at Alternative Contrast Codes

Earlier we stated that there were many, many possible sets of contrast codes that could be used to code a categorical predictor. Let us examine the same data that we have been focusing on using a different set of codes. Suppose we now define our contrast codes as follows:

	<i>Failure</i>	<i>No Feedback</i>	<i>Success</i>
λ_{1k}	–1	0	1
λ_{2k}	–1	2	–1

To differentiate these codes from the earlier set, we define Z_{1i} and Z_{2i} as contrast-coded predictors, assigning individuals the indicated values to represent category membership. We then regress Y_i on Z_{1i} and Z_{2i} with the estimated parameters:

$$\hat{Y}_i = 3.7083 + .6875Z_{1i} + .0208Z_{2i}$$

While the parameter estimates for the two contrast-coded predictors in this model are quite different from those that we estimated using the earlier set, in a deeper sense this model is equivalent to the model we developed under the old set of codes. Substituting for the values of Z_{1i} and Z_{2i} , we see that the group or condition means continue to be the predictions made by the model for all observations:

$$\hat{Y}_F = 3.7083 + .6875(-1) + .0208(-1) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .6875(0) + .0208(+2) = 3.750$$

$$\hat{Y}_S = 3.7083 + .6875(+1) + .0208(-1) = 4.375$$

Since the model makes the same predictions for all observations as the model with the previous set of contrast-coded predictors, the sum of squared errors is identical to what it was before, that is, 23.375.

The regression coefficients for the contrast-coded predictors have changed since the new contrast codes are making different comparisons among condition means from the comparisons made by the old set of codes. The contrast-coded predictor Z_{1i} is now comparing the means in the Success and Failure conditions. The value of its regression coefficient equals half the difference between these two group means. The second contrast-coded predictor, Z_{2i} , compares the mean in the No Feedback condition with the average of the means of the other two conditions. Its regression coefficient equals one-third of the difference between the mean in the No Feedback condition and the average of the other two means. These values for the regression coefficients are easily derived using the formula we gave earlier for the regression coefficient for a contrast-coded predictor. They also follow immediately once we realize the comparisons made by the contrasts and the number of units that separate observations in the various conditions on Z_{1i} and Z_{2i} . The intercept has not changed in value as a result of the new set of codes. It still equals the mean of the three condition means, as it will whenever a full set of contrast-coded predictors is used.

Since the change in codes has not changed the predicted values or the sum of squared errors for this model, a test of the null hypothesis that both β_1 and β_2 equal zero produces the same values of PRE and F as it did under the old set of codes. The compact single-parameter model is:

$$\text{MODEL C: } \hat{Y}_i = 3.7083$$

with a sum of squared errors of 30.958. PRE continues to equal .245, which converts to an F of 3.406 with 2 and 21 degrees of freedom. Thus, a test of the omnibus null hypothesis—that all of the condition means equal one another—reaches the same conclusion regardless of our choice of contrast codes.

Single-degree-of-freedom tests of whether β_1 or β_2 equals zero, however, reach rather different conclusions than they did before. These regression coefficients now estimate different comparisons between the condition means from those estimated with the earlier set of contrast codes. A test of whether β_1 equals zero is now equivalent to a test of whether the means in the Failure and Success conditions are equal to each other. The compact model for this test is:

$$\text{MODEL C1: } \hat{Y}_i = 3.7083 + .0208Z_{2i}$$

with a sum of squared errors of 30.938. The resulting PRE equals:

$$\frac{30.938 - 23.375}{30.938} = .244$$

which converts to an F statistic of 6.793 with 1 and 21 degrees of freedom. Since this exceeds the critical value, our test reveals both that β_1 is significantly greater than zero and that the mean in the Success condition is significantly greater than the mean in the Failure condition. As before, the sum of squares reduced associated with this contrast-coded predictor can be computed as a function of the relevant condition means that are being compared:

$$SSR_{Z_1} = \frac{[(-1)3.00 + (0) 3.750 + (+1)4.375]^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8} = 7.562$$

The test of whether β_2 differs from zero is equivalently a test of whether the No Feedback mean is significantly different from the average of the Failure and Success means. The compact model for this test is:

$$\text{MODEL C2: } \hat{Y}_i = 3.7083 + .6875Z_{1i}$$

with a sum of squared errors of 23.396. The resulting PRE equals .001, which converts to an F of 0.019 with 1 and 21 degrees of freedom. Clearly, the difference between the No Feedback mean and the average of the means in the other two conditions is not significant. As before, the SSR for this comparison can be directly calculated from the condition means:

$$SSR_{Z_2} = \frac{[(-1)3.00 + (+2)3.750 + (-1)4.375]^2}{(-1)^2/8 + (+2)^2/8 + (-1)^2/8} = 0.021$$

Since these two new contrast-coded predictors are nonredundant, just as were the earlier two, the two SSRs explained by each predictor over and above the other can be added to equal the SSR explained by them both as a set. The three tests we have just conducted can be summarized in Figure 8.11. Note that the only changes in this source table compared to the one using the earlier set of contrast codes (Figure 8.10) occur in the two rows of the table testing the specific comparisons made by the coefficients associated with Z_{1i} and Z_{2i} . All we have done is divide up the sum of squares between conditions (7.583) differently here, focusing on a different set of contrasts. As we have said before, with m category levels, there are only $m - 1$ orthogonal contrasts or comparisons that can be used. And whether we use these codes or those we discussed earlier is only a matter of theoretical preference.

FIGURE 8.11 Summary source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model (between conditions)		7.583	2	3.792	3.406	.245
Z_1 (Failure vs. Success)	.6875	7.562	1	7.562	6.793	.244
Z_2 (No Feedback vs. Failure, Success)	.0208	0.021	1	0.021	0.019	.001
Error		23.375	21	1.113		
Total		30.958	23			

Problems and Pitfalls in Using Nonorthogonal Codes

Suppose we were interested in asking the following two questions of these data, coded by the following set of codes:

	<i>Failure</i>	<i>No Feedback</i>	<i>Success</i>
λ_{1k}	-2	+1	+1
λ_{2k}	-1	0	+1

The first question is whether the Failure mean differs from the average of the means in the other two conditions. This was the first of the codes that we used in our first set, used to create the predictor X_{1i} . The second is whether the means in the Failure and Success conditions differ. This was the first of the codes that we used in our second set, used to create the predictor Z_{1i} . As we have seen, these are perfectly legitimate questions that we might want to ask of these data, but they are not orthogonal questions about the mean differences, as revealed by the fact that the second condition for contrast codes is not met by these two codes if we use them simultaneously:

$$\sum_k \lambda_{1k} \lambda_{2k} = (-2)(-1) + (+1)(0) + (+1)(+1) = +3$$

It is for this reason that we have not called them “contrast codes” when we think about them as a set. But, of course they are perfectly valid questions to ask of the data, albeit nonorthogonal.

If we were interested in these two questions, we might be tempted to use the two resulting predictors, X_{1i} and Z_{1i} , created with these codes to simultaneously predict Y_i , even though the codes themselves are not orthogonal. The resulting model would be:

$$\hat{Y}_i = 3.7083 + .0417 X_{1i} + .6250 Z_{1i}$$

This model, even with these nonorthogonal codes, is the same in a deep sense as the earlier model, in that it makes the same predictions of the condition means for every observation:

$$\hat{Y}_F = 3.7083 + .0417(-2) + .6250(-1) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .0417(+1) + .6250(0) = 3.750$$

$$\hat{Y}_S = 3.7083 + .0417(+1) + .6250(+1) = 4.375$$

And, as a result, it has the same sum of squared errors, 23.375. A model comparison between it, as Model A, and the single-parameter Model C, making a constant prediction for all observations, thus still provides the omnibus two-degree-of-freedom test about whether there are any differences among the category means. The resulting test of the overall model is summarized in the first row of the source table given in Figure 8.12 ($PRE = .245$, $F(2,21) = 3.406$). Both this row and the final two rows of the table are identical to what they were in the earlier source tables that we presented from these data. However, when we estimate the model in a regression program and examine the regression coefficients for the individual predictor variables and their respective SSRs, as given in the source table of Figure 8.12, they are not the values that we might expect.

FIGURE 8.12 Summary source table using nonorthogonal codes

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model (between conditions)		7.583	2	3.792	3.406	.245
X_1	.0417	0.021	1	0.021	0.019	.001
Z_1	.6250	1.562	1	1.562	1.403	.063
Error		23.375	21	1.113		
Total		30.958	23			

The coefficient for X_{1i} , when it was embedded in an orthogonal set of contrast-coded predictors (i.e., with X_{2i}), equaled .3542, which was shown to be:

$$\frac{\frac{\bar{Y}_{NF} + \bar{Y}_S}{2} - \bar{Y}_F}{3}$$

and its associated SSR was 6.021, which was shown to equal:

$$\frac{((-2)\bar{Y}_F + (+1)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-2)^2/8 + (+1)^2/8 + (+1)^2/8}$$

Now, however, when it is used to predict Y_i along with the nonorthogonal predictor Z_{1i} , its estimated coefficient equals .0417 and its SSR equals 0.021.

Similarly, the coefficient for Z_{1i} in this model equals .625, whereas its coefficient when used in the orthogonal set with Z_{2i} equaled .687, which was half the difference between \bar{Y}_S and \bar{Y}_{NF} . And now its SSR equals 1.562, whereas earlier, when used in the orthogonal set with Z_{2i} , its SSR was 7.562, which was shown to equal:

$$\frac{((-1)\bar{Y}_F + (0)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8}$$

The important point is that predictors that code the levels of a categorical independent variable will not yield coefficients that equal the expected mean differences and their associated SSRs unless a full set of orthogonal contrast-coded predictors is used. In other words, the following important equalities will *not* hold, unless a full set of contrast-coded predictors is included in the augmented model:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

$$SSR = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)}$$

Serious interpretative errors can ensue if one thinks that a given categorical predictor codes a particular mean difference when it is embedded in a nonorthogonal set. Orthogonality here depends solely on meeting the second defining condition of contrast codes—that the sum of the products of their coded values across category levels equals zero. As we will show, unequal numbers of observations can result in redundant contrast-coded predictors, redundant across observations, but this creates no interpretative problems as long as the codes are themselves orthogonal contrasts.

Given this, how might one proceed if one really were interested in testing the nonorthogonal questions of whether the Failure mean differs from the average of the

No Feedback and Success means (i.e., the question implicit in the X_{1i} codes) and of whether the Failure and Success means differ (i.e., the question implicit in the Z_{1i} codes)? Obviously, one could test these sequentially by specifying two models, one using both X_{1i} and X_{2i} as predictors and the other using Z_{1i} and Z_{2i} as predictors, just as we did in the earlier sections. Alternatively, one could simply rely on two bits of knowledge to test the mean differences implied by these contrasts without doing both estimations. First, as we have shown, the SSE(A) for a Model A that incorporates any complete set of contrast-coded predictors will be the same regardless of the specific set of such predictors used. Second, if a given contrast-coded predictor were included in a complete set of contrast-coded predictors, its SSR would be given by the following formula:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)}$$

Accordingly, in the case at hand, had one simply estimated the model that included X_{1i} and X_{2i} as predictors, one would have known that the SSE(A) for a model that used any full set of contrast-coded predictors would equal 23.375 with 21 degrees of freedom. One then could calculate the SSR associated with Z_{1i} if it were embedded in a full set of contrast-coded predictors, that is:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)} = \frac{((-1)\bar{Y}_F + (0)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8} = 7.562$$

Then one could calculate the values of PRE and F that would result if one tested Z_{1i} in the context of a complete set of contrast-coded predictors:

$$\text{PRE} = \frac{7.562}{23.375 + 7.562} = .244$$

$$F_{1,21} = \frac{7.562/1}{23.375/21} = 6.793$$

When doing this, one should recognize that the questions represented by these two contrast codes are not independent. The question of whether the Failure mean differs from the No Feedback and Success means is not entirely independent of the question of whether the Failure and Success means differ from each other. The answer to one is partially informative about the answer to the other.

Dummy Codes

There is another coding convention, known as dummy coding, that is widely used in some of the literature. Under this convention, one of the groups defined by the categorical variable is given values of zero on all $m - 1$ codes, and the other groups are given values

of zero on all but one of the codes. So, for instance, in the case at hand the following codes are consistent with this convention:

	<i>Failure</i>	<i>No Feedback</i>	<i>Success</i>
λ_{1k}	0	+1	0
λ_{2k}	0	0	+1

Obviously dummy codes do not meet either of the conditions that define contrast codes. As a result, it takes care to interpret exactly what is examined when one uses such codes to form predictor variables. One might think that a predictor that uses the first codes above would be comparing the No Feedback group to the other two groups and that a predictor that uses the second code above would be comparing the Success group with the other two groups, but in fact the regression coefficients for these two coded predictors, if included simultaneously, would each be asking whether the group coded zero on both codes differs from the group coded with a 1 for the predictor that is being examined. So, the predictor with the first code would examine the Failure–No Feedback difference and the predictor with the second code would examine the Failure–Success difference.

Because of the fact that interpretive mistakes can follow from the use of dummy codes, unless one is thoroughly familiar with them, we strongly recommend that researchers adopt the contrast-coding convention that we have explicated and that we will use in the remainder of this book.

CONTRAST CODES WITH UNEQUAL CELL SIZES

Historically, the procedure of ANOVA to detect mean differences was developed for data from experimental designs in which there were equal numbers of observations in every cell or condition of the design. In this sense, it was developed as an arithmetic shortcut, based on the assumption that predictors would be nonredundant. Of course, with the wide availability of computer programs that permit the estimation of linear regression models with partially redundant predictors, the assumption of nonredundant predictors is no longer necessary and, as we have just shown, ANOVA is easily implemented within general purpose multiple regression procedures, even with unequal numbers of observations in the various conditions.

In this section we present a new example with four levels of a categorical variable having unequal numbers of observations in each level or category. The bottom line is that, as long as estimation is done with a full set of contrast-coded predictors, all interpretations that we have previously given continue to be applicable, even though with unequal n values those predictors will be partially redundant across observations.

Let us assume that you are in a Psychology Department of a major university in which there are four PhD programs to which students are admitted every year: Clinical, Developmental, Experimental, and Social. The number of students who are admitted varies across the programs. Your question is whether there are mean differences in the verbal Graduate Record Examinations (GREs) of admitted students across the four programs. The data for a given year are presented in Figure 8.13, along with the

FIGURE 8.13 Hypothetical GRE scores, group means, and contrast codes

	<i>Program</i>			
	<i>Clinical</i>	<i>Developmental</i>	<i>Experimental</i>	<i>Social</i>
	750	700	640	690
	730	630	660	720
	710	620	710	750
	690		620	670
	670			650
	770			
\bar{Y}_k	720	650	657.5	696
n_k	6	3	4	5
λ_{1k}	3	-1	-1	-1
λ_{2k}	0	2	-1	-1
λ_{3k}	0	0	1	-1

four group means and the contrast-coded predictors that we will use. We simply use the Helmert coding convention here to derive these codes, as we have no strong expectations about where mean differences might be found.

We create three contrast-coded predictors, X_{1i} , X_{2i} , and X_{3i} , using these codes. We then proceed to estimate a Model A in which these three are used to predict verbal GRE scores:

$$\hat{Y}_i = 680.875 + 13.042X_{1i} - 8.917X_{2i} - 19.250X_{3i}$$

This model has a sum of squared errors of 21,595 and makes predictions of the category means for all observations.

The model's parameter estimates can be interpreted as we have done previously, based on the fact that we used a full set of contrast-coded predictors, even though those predictors are now partially redundant because of the unequal numbers of observations (i.e., the tolerances of all three predictors are less than 1.0).

The intercept, 680.875, is the mean of the four category means. It is not the mean of all 19 observations, which equals 687.778. The values of the three slopes equal differences among the category means, according to the following formulas:

$$b_1 = \frac{\sum_k \lambda_{1k} \bar{Y}_k}{\sum_k \lambda_{1k}^2} = \frac{(3)\bar{Y}_C + (-1)\bar{Y}_D + (-1)\bar{Y}_E + (-1)\bar{Y}_S}{12} = \frac{\bar{Y}_C - \frac{\bar{Y}_D + \bar{Y}_E + \bar{Y}_S}{3}}{4} = 13.042$$

$$b_2 = \frac{\sum_k \lambda_{2k} \bar{Y}_k}{\sum_k \lambda_{2k}^2} = \frac{(0)\bar{Y}_C + (2)\bar{Y}_D + (-1)\bar{Y}_E + (-1)\bar{Y}_S}{6} = \frac{\bar{Y}_D - \frac{\bar{Y}_E + \bar{Y}_S}{2}}{3} = -8.917$$

$$b_3 = \frac{\sum_k \lambda_{3k} \bar{Y}_k}{\sum_k \lambda_{3k}^2} = \frac{(0)\bar{Y}_C + (0)\bar{Y}_D + (1)\bar{Y}_E + (-1)\bar{Y}_S}{2} = \frac{\bar{Y}_E - \bar{Y}_S}{2} = -19.250$$

FIGURE 8.14 Summary source table for analysis of GRE scores

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model (between groups)		14,516.00	3	4838.70	3.14	.40
X_1 (C vs. D, E, S)	13.04	10,727.00	1	10,727.00	6.95	.33
X_2 (D vs. E, S)	-8.92	1605.00	1	1605.00	1.04	.07
X_3 (E vs. S)	-19.25	3293.89	1	3293.89	2.14	.13
Error		21,595.00	14	1542.5		
Total		36,111.00	17			

Figure 8.14 presents the source table that results from a comparison of this Model A with four different Model Cs. The first row of the table is a comparison between this Model A and the single-parameter Model C that predicts the grand mean for all observations. The null hypothesis for this comparison is that all the group means are equal to each other. The second row of the table gives the model comparison between this Model A and a Model C that omits the X_{1i} predictor. The null hypothesis here is that β_1 equals zero or, equivalently, that the mean of the Clinical group (C) is equal to the average of the means of the other three groups. The third row of the table gives the model comparison between this Model A and a Model C that omits the X_{2i} predictor. The null hypothesis here is that β_2 equals zero or, equivalently, that the mean of the Developmental group (D) is equal to the average of the means of the Experimental (E) and Social (S) groups. And the fourth row of the table gives the model comparison between this Model A and a Model C that omits the X_{3i} predictor. The null hypothesis here is that β_3 equals zero or, equivalently, that the mean of the Experimental group is equal to the mean of the Social group. All the resulting SSRs for these single-degree-of-freedom comparisons can be expressed in terms of the means, according to the formula we have frequently used before:

$$SSR = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

In short, all interpretations and computations explicated in this chapter for the analysis of a categorical independent variable apply regardless of whether there are equal numbers of observations across the levels of the categorical variable, as long as (*once again*) a full set of contrast-coded predictors is employed. The only thing that differs in this example from those presented earlier, as a function of the unequal values of n_k , is that the sums of squares for the individual single-degree-of-freedom tests in the above source table cannot be added up to equal the overall SSR for the model as a whole. In this case, the sum of the SSRs for the single-degree-of-freedom comparisons equals 15,625.89, while the overall test of the model yields an SSR of 14,516. This difference is due to the fact that across observations the predictors are now somewhat redundant. But all model comparisons and interpretations remain as they have been all along throughout the chapter.

ORTHOGONAL POLYNOMIAL CONTRAST CODES

Sometimes the observations fall into discrete categories on an independent variable of interest even though the underlying variable itself can be thought of as a continuum. For instance, suppose we were interested in age differences among elementary school children in their performance on a standardized arithmetic test. We take children from three different elementary school classes, those in the fourth, fifth, and sixth grades, and give them the standardized test. We then want to know if there are class mean differences. Our conceptual independent variable of interest is the children's age, but what we measure is their year in school and observations are clearly in three distinct categories on this measured variable.

FIGURE 8.15 Hypothetical scores on standardized arithmetic test

	Grade		
	Fourth	Fifth	Sixth
	68	68	80
	72	75	75
	76	68	78
	65	72	
	70	65	
		80	
		69	
\bar{Y}_k	70.20	71.00	77.67
n_k	5	7	3

Imagine that we had data from the 14 children given in Figure 8.15. In cases like this one, there is a special set of contrast codes that are sometimes useful for assessing trends in category means. These special codes are really just regular contrast codes, but they have the special name of “orthogonal polynomials” when values on the categorical independent variable can be ordered on some underlying continuum, as they clearly can in this case. In Figure 8.16 we present orthogonal polynomial contrast codes for categorical predictors having up to five levels.

As we indicate there, these codes have names that refer to the trend in the category means, across levels of the categorical independent variable, that they examine. So, with a two-level categorical variable, we can only examine whether the means go up or down, in essence fitting a linear function to the category means. With three levels, we can fit both a linear trend to the three category means and also ask whether the mean of the middle level is higher or lower than it ought to be given a simple linear ordering

FIGURE 8.16 Orthogonal polynomial contrast codes

Trend	Category				
	1	2	3	4	5
Linear	-1	1			
Linear	-1	2	3		
Quadratic	-1	0	1		
	-1	2	-1		
Linear	1	2	3	4	
Quadratic	-3	-1	1	3	
Cubic	1	-1	-1	1	
	-1	3	-3	1	
Linear	1	2	3	4	5
Quadratic	-2	-1	0	1	2
Cubic	2	-1	-2	-1	2
Quartic	-1	2	0	-2	1
	1	-4	6	-4	1

(the quadratic trend). With four levels, we can fit not only linear and quadratic trends, but also a cubic one, having two bends rather than one. And so forth.

With the three-level categorical variable in our data, let us use the two codes from the orthogonal polynomials to fit the linear and quadratic trends to these data. We create two contrast-coded predictors, X_{1i} and X_{2i} , using the codes specified in Figure 8.16 for a three-level categorical variable (4th grade is category level 1, etc.). The estimated model, using these to predict the standardized test scores, is:

$$\hat{Y}_i = 72.96 + 3.73X_{1i} - .98X_{2i}$$

with a sum of squared errors of 237.47. Of course this model, with a full set of contrast codes, exactly predicts the group means:

$$\bar{Y}_{4th} = 72.96 + 3.73(-1) - 0.98(-1) = 70.20$$

$$\bar{Y}_{5th} = 72.96 + 3.73(0) - 0.98(+2) = 71.00$$

$$\bar{Y}_{6th} = 72.96 + 3.73(+1) - 0.98(-1) = 77.67$$

And as always the parameter estimates can be interpreted in terms of the group means. The intercept, 72.96, is the mean of the three means, 3.73 is half the difference between \bar{Y}_{6th} and \bar{Y}_{4th} and -0.98 is one-third of the difference between \bar{Y}_{5th} and the average of the other two group means. The source table that summarizes the analysis of these data is given in Figure 8.17.

So far, there is nothing new about this model or its interpretations. So why do we refer to the codes we have used as orthogonal polynomials? The reason is that the slope of the first one is the slope that results if we were to fit a straight line to the three group means, going from the fourth grade up to the sixth grade. And the slope of the second contrast-coded predictor estimates the degree to which the group mean for the fifth grade does not lie on that prediction line, that is, the degree to which that prediction function deviates from a straight line if it is to predict all three group means. Given the significance of the coefficient associated with X_{1i} , we can conclude that there is a linear increase in performance on the standardized test as we go up from children in the fourth grade to children in the sixth grade.

An obvious question is how these results would differ from what would be obtained if we simply regressed Y_i on Grade itself, treated as a continuous variable, numerically coded as 4, 5, and 6. Such a model would be a simple regression model, asking if there is a linear relationship between Grade and test performance. It is estimated as:

$$\hat{Y}_i = 55.62 + 3.38Grade_i$$

with a sum of squared errors of 268.62. A test of whether there is a linear relationship between grade and performance yields an SSR of 88.31, $PRE = .25$, and $F = 4.27$ with 1 and 13 degrees of freedom. In this model, the slope associated with the Grade predictor variable informs us about the degree to which predicted performance on the standardized math test increases as grade goes up by one unit, that is, from fourth to fifth and from fifth to sixth.

While these are obviously different models in a variety of ways, the important conceptual difference is that in the one using the orthogonal polynomial contrast codes we are predicting the group means and asking about differences among those group means. And in the simple linear regression model, using Grade as our predictor, we are

FIGURE 8.17 Summary source table for analysis of arithmetic scores

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model (between groups)		119.47	2	59.73	3.019	.34
X_1	3.73	104.506	1	104.506	5.28	.31
X_2	-0.98	31.17	1	31.17	1.57	.12
Error		237.47	12	19.79		
Total		356.93	14			

simply fitting a linear function to all the individual observations, rather than to the group means. Given that the group sizes are very unequal, modeling the group means and modeling the individual observations yield different results.

TYPE I ERROR RATES IN TESTING MEAN DIFFERENCES

Earlier we discussed the general strategy for testing any mean difference that was of interest in designs with multiple levels (m) of a categorical independent variable. One first derives a full set of contrast codes ($m - 1$ of them) and uses them to estimate an augmented model, which predicts the category means. This model provides tests of the specific mean differences that were used as codes for the predictors. Importantly, it also provides the $SSE(A)$ and the mean square error for any model that used a full set of codes, regardless of which set was used. Then, for any additional mean comparison of interest, one calculates the SSR associated with that contrast as if it were used as a predictor in a complete set of orthogonal contrast-coded predictors:

$$SSR = \frac{\left(\sum_k \lambda_k \bar{Y}_k \right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

Dividing this SSR by the MSE from the estimated model yields the F statistic associated with the mean comparison of interest, and the PRE for that comparison can be calculated as:

$$\frac{SSR}{SSE(A) + SSR}$$

When there are more than two or three levels of the categorical variable, the number of potential mean differences that might be tested can become very large. For instance, with only four levels, one could in theory test the means of individual groups against each other, the means of all pairs of groups against each other, the mean of each triad of groups against the mean of the remaining group, and so forth. A problem that arises in this case, where many mean differences might be tested, is that the probability of a Type I statistical error may become unacceptably large. While α may be set at .05 for any one mean comparison, across many such comparisons the probability that somewhere a Type I error has been committed can quickly become quite large. For instance, if we

had four groups and we asked whether each mean differed from each other mean, that would be six contrasts tested (in addition to others included in the orthogonal set of codes used initially to generate Model A). If on each test α was set at .05, across the six tests, the probability that we would make at least one Type I error is equal to:

$$1 - (1 - .05)^6 = 1 - .95^6 = .265$$

In other words, even if the null hypothesis were true and all the true group means were equal to each other, at least one of the six mean comparisons would be significant more than a quarter of the time. In the ANOVA literature many different procedures have been developed for dealing with this issue. We focus on only two of them, and the crucial difference between these two is whether the mean comparison that is tested is a *planned comparison* or a *post hoc comparison*.

Planned comparisons

Planned comparisons are those that the researcher had theoretical or substantive reasons for examining *before* conducting the experiment. In other words, the researcher specified all the planned comparisons of interest before collecting or examining the data. Ideally, as many of these planned contrasts as possible would be included in the set of orthogonal contrast-coded predictor used to generate the initial Model A. Regardless, one adds up the number of comparisons that one intends to examine, both in the initial model and in the other models of theoretical interest. Let us say that number is c . To keep α at .05 across all c tests, one wants to compare each obtained F to a critical value of F , using α/c to determine the critical F . So, for instance, if there are six tests to be conducted, one would use a critical value of F at $.05/6 = .0083$ rather than at .05.¹

Post hoc comparisons

Post hoc comparisons are those that do not occur to us until *after* we have examined the data. Often when looking at the data certain comparisons that we did not anticipate appear to be interesting. It is natural to want to test those interesting, unanticipated contrasts. However, it is impractical to use the above procedure for planned comparisons because when looking at the data we are implicitly doing many, many comparisons—all those that do not strike us as interesting—that ought to be included in c , the total number of comparisons made. Instead of trying to count all those implicit comparisons, standard practice is to compare F to the following critical value developed by Scheffé (1959):

$$(m - 1)F_{crit; m-1, n-PA; \alpha}$$

There are two important features of using the Scheffé adjusted critical value. First, the overall probability of making at least one Type I error will remain at α no matter how many contrasts are evaluated using the adjusted critical value. Thus, the researcher can do as much snooping and exploring with contrasts as desired without undue risk of making Type I errors. Second, there will be at least one contrast whose F exceeds the Scheffé adjusted critical value if and only if the omnibus test, comparing Model A to a Model C that predicts the grand mean for all observations, is statistically significant. Thus, if the omnibus test is not significant, then there is no point in evaluating any post hoc contrasts using the Scheffé criterion.

POWER ANALYSIS FOR ONE-WAY ANOVA

Estimating Statistical Power

The advantage of consistently adhering to the model comparison approach is that all that we have learned before still applies as we consider new types of models. Thus, the methods for estimating statistical power presented in Chapter 6 for multiple regression apply unaltered to one-way ANOVA. In particular, you can use prior research to estimate the value of η^2 (the expected proportional reduction in error) for either the omnibus test of any mean differences or specific one-degree-of-freedom contrasts. As before, you should consider adjusting empirical estimates of η^2 based on the number of observations and parameters. This is especially important for experiments in which the number of observations is often small relative to the number of model parameters. Also, as before, you may use either Cohen's values for small, medium, and large effects for your power analyses, or values from the literature of your substantive research topic. For example, if we wanted to re-do the SAT coaching study (with which we began this chapter) with a larger number of observations, we would start our power analysis by adjusting the PRE of .196 reported in that study:

$$\hat{\eta}^2 = 1 - (1 - PRE) \left[\frac{n - PC}{n - PA} \right] = 1 - (1 - .196) \left[\frac{20 - 1}{20 - 2} \right] = .15$$

A quick check of the power associated with this value suggests that about 50 participants would be required to have an 80% chance of detecting an effect of this magnitude.

Earlier we also saw how estimates of the parameter values—the regression slopes and the variance of the error—can provide estimates of the effect size for power analysis. However, in one-way ANOVA, rather than having prior ideas about the values of regression parameters, a researcher more commonly has notions about the values of the group means that determine the regression slopes. Hence, it is worth examining how we can estimate the expected effect size by beginning with expectations of the cell means. We begin with our usual definition of PRE:

$$PRE = \frac{SSE(C) - SSE(A)}{SSE(C)} = \frac{SSR}{SSE(C)}$$

We have noted before that $SSE(C) = SSE(A) + SSR$ (i.e., the error for the compact model includes all the error of the augmented model plus the error that was reduced by the addition of the extra parameters in the augmented model). Hence:

$$PRE = \frac{SSR}{SSE(A) + SSR} = \frac{1}{(SSE(A)/SSR) + 1}$$

To obtain a definition of η^2 , we simply calculate $SSE(A)$ and SSR using the true parameter values (β_0, β_1 , etc., ε^2) instead of the estimated parameters (b_0, b_1 , etc., s^2). For one-way ANOVA, we can start with our expectations about what the true group means, μ_k , might be, and use those in the formula for the SSR for a contrast-coded predictor:

$$\text{SSR for a contrast} = \frac{[\sum \lambda_k \mu_k]^2}{\sum \lambda_k^2 / n_k}$$

For a complete model, SSE(A) depends only on σ^2 , the within-cell variance. Specifically:

$$\text{SSE(A)} = \left(\sum_k n_k \right) \sigma^2 = n \sigma^2$$

where n is the grand total of observations. We multiply by n instead of by $(n - PA)$ because we have not estimated any parameters from data in the calculations of SSE(A). Substituting these values for SSR and SSE(A) calculated from the presumed true parameters into the formula for PRE yields the following formula for η^2 :

$$\hat{\eta}^2 = \left(\frac{n\sigma^2(\sum \lambda_k^2/n_k)}{[\sum \lambda_k \mu_k]^2} + 1 \right)^{-1}$$

If n_k , the number of observations in each group, is equal for all groups and if m is the number of groups so that $n = mn_k$, then the above formula reduces to:

$$\hat{\eta}^2 = \left(\frac{m\sigma^2(\sum \lambda_k^2)}{[\sum \lambda_k \mu_k]^2} + 1 \right)^{-1}$$

which does not depend on either the total number of observations or the number of observations in each group. Therefore, to find a value for η^2 to use in our power calculations, we need only specify the values for σ^2 and for μ_k that we expect to obtain in our study. And, of course, we need to specify the λ_k values for the contrast code for which we want to estimate the statistical power.

As an example of this direct approach for estimating η^2 , let us again consider the feedback study described earlier in this chapter. Suppose the researcher on the basis of prior research had expected values of 3, 4, and 4 for the means of the Failure, No Feedback, and Success groups, respectively, and a within-cell variance of about 1.5. Then for the $\{-2, 1, 1\}$ contrast the expected effect size is:

$$\hat{\eta}^2 = \left(\frac{3(1.5)6}{[(-2)3 + 4 + 4]^2} + 1 \right)^{-1} = .13$$

From this we estimate that about 60 observations, 20 per group, would be necessary to provide a statistical power of .8.

Power and Research Design

When planning experiments, researchers must choose how many groups to use and how many observations should be in each group. It is important to consider the consequences that such choices have for statistical power. We first consider the design implications for the omnibus test of whether there are any differences among the means and then for the power of specific contrasts.

The statistical power of the omnibus test of whether there are any mean differences among the groups is maximized when there are an equal number of observations in each group. In that case, the omnibus F statistic equals the average of all the F statistics for

a set of orthogonal contrasts. This highlights a common mistake in research design: If too many groups are used, then there are many contrasts where no differences are expected, which in turn lowers the omnibus F . Sometimes using too many groups cannot be avoided. For example, biopsychological researchers must sometimes use multiple control groups (e.g., handling the animal, injecting with a drug vehicle, injecting with a placebo) in comparison to a single treatment group. If there are no differences expected among the different control groups, then the possible magnitude of the omnibus F statistic is reduced. In this case, the omnibus F is often misleading and researchers should simply focus on the treatment versus controls contrast.

Researchers using ordered category levels over which they expect polynomial trends often use too many groups. For example, researchers expecting linear and perhaps quadratic trends sometimes err in using as many as five levels. With equal numbers of observations at each level, the omnibus F is reduced because it is the average of not only the expected linear and quadratic contrasts but also the not-expected cubic and quartic contrasts. Even the power of the separate tests of the linear and quadratic contrasts is reduced because some of the study's valuable resources—the observations—have been allocated to test for the cubic and quartic effects. At the same time, testing more polynomial trends than expected increases the chances of making Type I errors.

For specific contrasts, allocating an equal number of observations to all groups effectively gives equal importance to all contrasts. If one or two contrasts are more important to the research purpose than the other contrasts, one may want to allocate the number of observations unequally across the groups so as to maximize the statistical power of the contrasts of greater importance. Power for a contrast is maximized when the allocation of observations to groups is proportional to the absolute values of the contrast weights. For example, for the t -test where the weights are $\{-1,+1\}$, statistical power is maximized with an equal number of observations in each group. However, for the quadratic contrast for three groups with weights $\{-1,+2,-1\}$, power is maximized by allocating a quarter of the observations to each of the extreme categories and the remaining half of the observations to the middle category. The biopsychology researcher comparing the treatment group to four control groups could maximize the power of that comparison by allocating half the total observations to the treatment group and one-eighth of the observations to each of the four control groups.

SUMMARY

In this chapter we have considered models with a single categorical predictor having two or more levels or categories. Such predictors need to be numerically coded and we have given our strong preference that the codes to be used are orthogonal contrast codes. A full set of such codes means that there are $m - 1$ contrast-coded predictors with m category levels. Additionally, given that the sum of the codes for any predictor equals one, orthogonality is assured by codes where the sum of all pairs of crossproducts equals zero.

Assuming that a full set of codes is used to predict the data variable, then the predicted values from such a model will equal the category means of that data variable. Accordingly, inferences about slopes are then equivalently inferences about difference among category means. And these models are then equivalent to traditional independent sample t -tests (given $m = 2$) and one-way ANOVA models (given $m > 2$). Although

one-way ANOVA has traditionally emphasized omnibus tests to examine whether there are any mean differences among the groups or categories, our strong preference is for single-degree-of-freedom model comparisons, testing specific focused contrasts that are of theoretical interest. We encourage analysts to ask about mean differences that are of interest to them, even when those differences are not themselves orthogonal. We discuss appropriate procedures for asking about such nonorthogonal differences, given the constraint that a full set of codes be orthogonal. Finally, we discuss procedures for avoiding inflated α levels when conducting many mean comparisons, whether a priori or post hoc.

Note

- 1 The use of α/c instead of α is based on the *Bonferroni inequality* and so some manuals for statistical programs refer to this as the Bonferroni method of multiple comparisons. Sometimes this is also referred to as the Dunn method.

Factorial ANOVA

Models with Multiple Categorical Predictors and Product Terms

In the previous chapter we developed models with one categorical predictor. In this chapter we expand our consideration to models with two or more categorical predictor variables. Our reasons for wanting to include more than one categorical variable as a predictor in our model are the same as those that motivated us to expand from simple regression models with one predictor to multiple regression models with two or more predictors in Chapter 6. Models in which predictions are conditional on two or more categorical variables may be required by our data and, more importantly, by the underlying process that generates the data. Just as in multiple regression, controlling for one categorical variable by including it in the model often allows us to have a better look at the effects of other categorical variables. Also, as in multiple regression, we are often interested in modeling the joint effect of two or more categorical variables. With categorical variables we will be especially interested in whether the effect of a given categorical variable depends on the levels of the other categorical variables; that is, we are interested in whether or not there is an interaction between the categorical variables analogous to the interactions of continuous variables in multiple regression considered in Chapter 7.

The generalization of models with one categorical predictor (one-way ANOVA of the previous chapter) to models with two categorical predictors (two-way ANOVA) and to models with more than two categorical predictors (q -way factorial ANOVA, where q is the number of categorical predictors) is straightforward. As we shall demonstrate, classical analysis of variance with two or more categorical predictors is nothing more than a simple one-way ANOVA with a specific, clever set of contrast codes. In other words, the only new thing to learn is how to generate the appropriate set of contrast codes; fitting the model and testing hypotheses are *exactly* the same as for one-way ANOVA in Chapter 8.

FACTORIAL ANOVA AS ONE-WAY ANOVA

We begin by considering the hypothetical dataset in Figure 9.1. In this hypothetical experiment clinically depressed patients either receive psychotherapy (treatment) or not (control) and receive one of three drugs (A, B, or placebo). After six months each patient completes a mood questionnaire on which higher scores mean improved mood or decreased depression.

FIGURE 9.1 Hypothetical data (mood scores) in a Drug (3) by Psychotherapy (2) experimental design

Drug	Psychotherapy					
	Treatment			Control		
A	31	31	34	17	15	19
B	25	25	28	23	18	16
Placebo	17	18	16	9	10	8

FIGURE 9.2 Hypothetical data of Figure 9.1 arrayed as a one-way design

	Group 1 (A, T)	Group 2 (B, T)	Group 3 (P, T)	Group 4 (A, C)	Group 5 (B, C)	Group 6 (P, C)
	31	25	17	17	23	9
	31	25	18	15	18	10
	34	28	16	19	16	8
Mean	32	26	17	17	19	9

Contrast codes	Group						
	Predictor 1	2	3	4	5	6	
λ_1 1 vs. 2, 3, 4, 5, 6	Z_1	5	-1	-1	-1	-1	-1
λ_2 2 vs. 3, 4, 5, 6	Z_2	0	4	-1	-1	-1	-1
λ_3 3 vs. 4, 5, 6	Z_3	0	0	3	-1	-1	-1
λ_4 4 vs. 5, 6	Z_4	0	0	0	2	-1	-1
λ_5 5 vs. 6	Z_5	0	0	0	0	1	-1

The experimental design depicted in Figure 9.1 is known as a *factorial* design because every level of one categorical variable or factor is combined with every level of the other categorical variable or factor. There are three levels of the Drug variable and two levels of the Psychotherapy variable, so there are a total of $3 \times 2 = 6$ different combinations, each defining a group or cell in the design. We often refer to this as a 3×2 *design*. In this hypothetical study, from a total of 18 patients, three patients are randomly assigned to each of the six groups.

Although it is natural to display these data in a table with three rows and two columns as in Figure 9.1, we can also display the data as a one-way layout in terms of the six groups or cells as in Figure 9.2. Seeing the data in the one-way layout makes it clear that we can use any of the sets of contrast codes developed in the previous chapter to analyze these data. To illustrate this, we will first do an analysis with one-way contrast codes that are statistically correct but that do not ask questions that are usually interesting. In this first set of contrast codes we form the first contrast λ_1 by comparing Group 1 (Drug A combined with psychotherapy Treatment: A,T) with all the other groups; the corresponding contrast-coded predictor is Z_1 .¹ The contrast λ_2 compares Group 2 against all the remaining groups except Group 1, and so on. These contrast codes are also displayed in Figure 9.2. We can verify that these codes are orthogonal by checking that the sum of each set of crossproducts is zero. For example:

$$\begin{aligned}\sum_{k=1}^6 \lambda_{1k} \lambda_{2k} &= 5(0) - 1(4) - 1(-1) - 1(-1) - 1(-1) - 1(-1) \\ &= 0 - 4 + 1 + 1 + 1 + 1 \\ &= 0\end{aligned}$$

We can regress Y_i , the mood scores, on the predictor variables Z_1, Z_2, \dots, Z_5 using a standard multiple regression program or, equivalently, we could use the one-way ANOVA formulas from Chapter 8. Figure 9.3 shows the actual data matrix we would use to regress Y_i on Z_1, Z_2, \dots, Z_5 . Note that the mean for each contrast-coded predictor is zero because there are equal numbers of observations in each group.

We can ask whether all six group means are equal by comparing these models:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

Regressing the mood scores on the five contrast-coded predictors yields the analysis in Figure 9.4. The two estimated models are:

$$\text{MODEL A: } \hat{Y}_i = 20 + 2.4Z_{1i} + 2.1Z_{2i} + 0.5Z_{3i} + 1Z_{4i} + 5Z_{5i}$$

$$\text{MODEL C: } Y_i = 20$$

FIGURE 9.3 Data matrix for analyzing mood scores using multiple regression and one-way coded predictors

Group		Mood scores	Predictors				
Drug	Psychotherapy	Y	Z_1	Z_2	Z_3	Z_4	Z_5
A	T	31	5	0	0	0	0
A	T	31	5	0	0	0	0
A	T	34	5	0	0	0	0
B	T	25	-1	4	0	0	0
B	T	25	-1	4	0	0	0
B	T	28	-1	4	0	0	0
P	T	17	-1	-1	3	0	0
P	T	18	-1	-1	3	0	0
P	T	16	-1	-1	3	0	0
A	C	17	-1	-1	-1	2	0
A	C	15	-1	-1	-1	2	0
A	C	19	-1	-1	-1	2	0
B	C	23	-1	-1	-1	-1	1
B	C	18	-1	-1	-1	-1	1
B	C	16	-1	-1	-1	-1	1
P	C	9	-1	-1	-1	-1	-1
P	C	10	-1	-1	-1	-1	-1
P	C	8	-1	-1	-1	-1	-1
Mean		20	0	0	0	0	0

FIGURE 9.4 ANOVA table of the hypothetical mood scores using the model:

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \varepsilon_i$$

Source	b_j	SS	df	MS	F	p	PRE
Between groups		960.0	5	192.0	46.1	.0001	.95
Z_1	2.4	518.4	1	518.4	124.4	.0001	.91
Z_2	2.1	264.6	1	264.6	63.5	.0001	.84
Z_3	0.5	9.0	1	9.0	2.2	.17	.15
Z_4	1.0	18.0	1	18.0	4.3	.06	.26
Z_5	5.0	150.0	1	150.0	36.0	.0001	.75
Within groups (MSE)		50.0	12	4.2			
Total		1010.0	17				

The augmented model using all five predictors reduces the error of the compact model using the mean by $960/1010 = .9505$. We calculate F in the usual way:

$$F_{5,12} = \frac{.9505/5}{(1 - .9505)/12} = 46.1$$

This omnibus tests rejects Model C in favor of Model A because the large value of PRE (.95) is statistically surprising ($F_{5,12} = 46.1, p < .0001$), which implies, as it did in one-way ANOVA, that the means for the six groups are not all equal to one another. That is, the six groups cannot be adequately represented by the overall mean.

But which groups are different from each other? The test of the Z_1 contrast indicates that the estimate $b_1 = 2.4$ is significantly different from zero and it reduces 91% of the error remaining after all the other codes are in the equation.² Remember that this test is obtained by comparing an augmented model including all the Z values with a compact model including all the Z values except Z_1 . In particular:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \beta_5 Z_{5i} + \varepsilon_i$$

$$H_0: \beta_1 = 0$$

In this case, we can reject Model C in favor of Model A and therefore reject the null hypothesis that $\beta_1 = 0$. The corresponding contrast λ_1 codes the comparison between Group 1 (Drug A with psychotherapy treatment) and the average of all the other groups; thus, we can conclude that Group 1 is significantly different from the average of all other cells in the study.

We interpret the value of the coefficient the same as we always have. Specifically, the coefficient $b_1 = 2.4$ means that for a one-unit increase in Z_1 the predicted mood score increases, on average, by 2.4 points. To verify this, reexamine Figure 9.2 to see that the value of Z_1 for all the other groups is -1 but for Group 1 the value of Z_1 is 5. Thus, there is a change of six units on Z_1 between the two comparison groups. Model A therefore predicts the mean for Group 1 to be $6 \times 2.4 = 14.4$ higher than the average in all the other groups. This is indeed the case: the mean for Group 1 is 32 and the mean for all the other groups combined is $(26 + 17 + 17 + 19 + 9)/5 = 17.6$, thus $32 - 17.6 = 14.4$. But why is Group 1 better, on average, than all the other groups? Is it because of Drug A? Or is it because patients receiving psychotherapy score higher regardless of which

drug they receive? Or is it because Drug A is especially effective for patients who also receive psychotherapy treatment? The contrast λ_1 cannot tell us. It simply indicates that there is a difference.

We can similarly interpret the results for the contrast-coded predictor Z_2 . That the estimate $b_2 = 2.1$ is significantly different from zero indicates that Group 2 (Drug B administered to those receiving psychotherapy treatment) is different from the average of all the subsequent groups (i.e., excluding Group 1). The difference in means equals $5 \times 2.1 = 10.5$. But again we do not know whether the higher average mood scores of Group 2 are due to Drug B or psychotherapy treatment or their combination.

The other contrast-coded predictor whose coefficient is significantly different from zero (using $\alpha = .05$) is Z_5 ; the corresponding contrast λ_5 compares Groups 5 and 6 or the difference between taking Drug B versus placebo for those in the Control condition. We know in this case that the higher average score for Group 5 ($2 \times 5 = 10$) is due to Drug B relative to the placebo because all patients in this comparison are in the control group not receiving psychotherapy. But is Drug B also better than the placebo for those patients who do receive psychotherapy treatment? None of the Z contrasts help to answer that question.

A BETTER SET OF CONTRAST CODES

Even though the one-way analysis of variance using the Z contrast-coded predictors is statistically correct, it has failed to answer important questions we want to ask of the data. One solution would be to use the multiple comparison procedures of Chapter 8 to address some of the unanswered questions. However, a much more efficient strategy is to begin with a set of contrast codes that do ask many of the questions we would naturally want to consider. Given the two-way layout of the data (as in Figure 9.1), there are several sets of codes that are more natural than the Z codes. “Two-way” ANOVA is nothing more than one-way ANOVA using one of these natural sets of codes.

To develop a more natural set of codes, we begin by considering each of the two categorical variables separately. The strategy is to develop contrast codes identical to the ones that we would use if each categorical variable were considered alone in its own one-way ANOVA. That is, our first step is to code each categorical variable as if the other one did not exist. For the Drug variable an interesting question is whether the drugs (either A or B), on average, do better than the placebo. The contrast code λ_1 (and its corresponding contrast-coded predictor X_1) in Figure 9.5 makes precisely that comparison with the pattern (1, 1, -2). This pattern is repeated for each level of the other categorical variable (in this case, Psychotherapy). The contrast code λ_2 (and its corresponding contrast-coded predictor X_2) then asks whether there is any difference between Drugs A and B. We can verify that λ_1 and λ_2 are orthogonal by checking whether the sum of their crossproducts equals zero. That is:

$$\begin{aligned} \sum_{k=1}^6 \lambda_{1k} \lambda_{2k} &= 1(1) + 1(-1) - 2(0) + 1(1) + 1(-1) - 2(0) \\ &= 1 - 1 + 0 + 1 - 1 + 0 \\ &= 0 \end{aligned}$$

FIGURE 9.5 Hypothetical data of Figure 9.1 coded for two-way ANOVA

	Group 1 (A, T)	Group 2 (B, T)	Group 3 (P, T)	Group 4 (A, C)	Group 5 (B, C)	Group 6 (P, C)
	31	25	17	17	23	9
	31	25	18	15	18	10
	34	28	16	19	16	8
Mean	32	26	17	17	19	9

Contrast codes		Group						
		Predictor 1	2	3	4	5	6	
λ_1	Drugs vs. Placebo	X_1	1	1	-2	1	1	-2
λ_2	Drug A vs. Drug B	X_2	1	-1	0	1	-1	0
λ_3	Treatment vs. Control	X_3	1	1	1	-1	-1	-1
λ_4	Interaction: $\lambda_1 \times \lambda_3$	X_4	1	1	-2	-1	-1	2
λ_5	Interaction: $\lambda_2 \times \lambda_3$	X_5	1	-1	0	-1	1	0

The orthogonality combined with the equal numbers of observations in each cell ensures that the coded predictors X_1 and X_2 are uncorrelated or not redundant. With three levels of the Drug categorical variable, we can only have two orthogonal codes so λ_1 and λ_2 are sufficient for the one-way analysis of that variable.

There are only two levels of the Psychotherapy categorical variable, so we need only one code. Hence, λ_3 codes the contrast between receiving psychotherapy treatment versus being in the control condition and provides the complete one-way ANOVA for that variable.

So far we have only three codes, λ_1 , λ_2 , and λ_3 , for the two separate one-way analyses of the categorical variables, but five orthogonal contrast codes are required for the complete analysis of six groups. To generate the other two necessary codes, we simply multiply the contrast codes between the two categorical variables; that is, $\lambda_4 = \lambda_1 \times \lambda_3$ and $\lambda_5 = \lambda_2 \times \lambda_3$. These are the same kind of product terms we considered in Chapter 7 when we introduced the concept of interactions between variables. In the context of two-way ANOVA these product terms ask especially interesting questions. The code λ_4 , the product of λ_1 and λ_3 , asks whether the difference coded by λ_1 (Drugs vs. Placebo) depends on the level of psychotherapy (Treatment vs. Control). In other words, might there be one drug effect for those receiving psychotherapy treatment and a different drug effect for those in the control group. Similarly, λ_5 , the product of λ_2 and λ_3 , asks whether the comparison coded by λ_2 (Drug A vs. Drug B) depends on the level of Psychotherapy (Treatment vs. Control). The order of multiplication is arbitrary, so either statement could be reversed. That is, λ_4 asks whether the effectiveness of psychotherapy (the Treatment vs. Control difference) depends on whether the patient also received a drug or the placebo. We consider the interpretation of interactions like these in greater detail later.

Note that we do not form a code by multiplying $\lambda_1 \times \lambda_2$ because that would yield another code for just the Drug categorical variable that could not be orthogonal to the other two codes for that variable. However, the two codes formed from products, λ_4 and λ_5 , are orthogonal to each other and to the other codes. You may want to compute some of the crossproducts to verify this claim. This gives a total of five orthogonal contrast codes, precisely the number we need for the analysis of six groups.

FIGURE 9.6 Data matrix for analyzing mood scores using multiple regression and two-way coded predictors

Group		Mood scores Y	Predictors				
Drug	Psychotherapy		X_1	X_2	X_3	X_4	X_5
A	T	31	1	1	1	1	1
A	T	31	1	1	1	1	1
A	T	34	1	1	1	1	1
B	T	25	1	-1	1	1	-1
B	T	25	1	-1	1	1	-1
B	T	28	1	-1	1	1	-1
P	T	17	-2	0	1	-2	0
P	T	18	-2	0	1	-2	0
P	T	16	-2	0	1	-2	0
A	C	17	1	1	-1	-1	-1
A	C	15	1	1	-1	-1	-1
A	C	19	1	1	-1	-1	-1
B	C	23	1	-1	-1	-1	1
B	C	18	1	-1	-1	-1	1
B	C	16	1	-1	-1	-1	1
P	C	9	-2	0	-1	2	0
P	C	10	-2	0	-1	2	0
P	C	8	-2	0	-1	2	0
Mean		20	0	0	0	0	0

Figure 9.6 shows the data matrix that we can analyze with a multiple regression program to model mood scores in terms of the X variables. That is, we can regress Y_i on X_1 to X_5 . Note again that the mean for each predictor is zero because we are using contrast codes and there are an equal number of observations for each group.

Figure 9.7 presents the results of the regression analysis using the X contrast-coded predictors. Note that the sum of squares for the augmented model including all the predictors, and the F and PRE for the omnibus test of the complete model are *exactly* the same as when we used the Z predictors. This must be the case because in each analysis we used a complete set of codes so that, as in one-way ANOVA, the predicted value \hat{Y} for each group is the mean \bar{Y} for that group. If the predictions of the two models are

FIGURE 9.7 ANOVA table of the hypothetical mood scores using the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

Source	b_j	SS	df	MS	F	p	PRE
Between groups		960	5	192.0	46.1	.0001	.95
Drug		453	2	226.5	53.9	.0001	.90
X_1	3.5	441	1	441	105.8	.0001	.90
X_2	1.0	12	1	12	2.9	.11	.19
Psychotherapy		450	1	450	108.0	.0001	.90
X_3	5.0	450	1	450	108.0	.0001	.90
Drug x Psychotherapy		57	2	28.5	6.8	.01	.53
X_4	0.5	9	1	9	2.2	.16	.15
X_5	2.0	48	1	48	11.5	.005	.49
Within groups (MSE)		50.0	12	4.2			
Total		1010.0	17				

the same, then the total error and the total error reduced must be the same. The only difference in the two analyses is how that total error reduction is divided into separate components. The X predictors ask different questions than the Z predictors, so the individual error reduction associated with each of the X predictors differs from that of the Z predictors, but the total error reduction must be the same.

There are nine PRE and F values in Figure 9.7. Each one corresponds to a comparison between a particular Model C and a Model A. To understand the meaning of each test, it is important to be precise about those models for each test. The omnibus test is reported in the row labeled “Between groups” and represents the comparison between the following two models:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

This corresponding null hypothesis, equivalent to assuming that all the group means are equal, is rejected by the large values of PRE and F .

The rows for each of the X_j compare a Model A that uses all the contrast-coded predictors to a Model C that includes all the predictors except X_j . For example, the PRE and F in the row labeled X_1 compare:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$H_0: \beta_1 = 0$$

We generated the predictors X_1 and X_2 as one-way codes for the Drug categorical variable for the data matrix in Figure 9.1. The row labeled “Drug” in Figure 9.7 reports the results from testing a model including all the contrast codes to one that omits *both* the drug codes X_1 and X_2 . That is, we are comparing:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$H_0: \beta_1 = \beta_2 = 0$$

In other words, we are asking whether the predictions would suffer if we were to ignore the Drug categorical variable. The large PRE and F indicate that we cannot ignore which drug a patient received. In the traditional language of ANOVA, this is known as the test of the *main effect* of the Drug categorical variable. In this case, we would conclude that overall there is a statistically significant main effect for the Drug categorical variable. Note that because X_1 and X_2 are uncorrelated in this instance, their sums of squares add to produce the sum of squares attributable to both of them.

We prefer the single-degree-of-freedom tests of the focused comparisons X_1 and X_2 to this global test of the drug effect because rejection of the hypothesis $\beta_1 = \beta_2 = 0$ is ambiguous. We do not know which part of this multiple-degree-of-freedom hypothesis is at fault. Maybe the global hypothesis is rejected because $\beta_1 \neq 0$ or because $\beta_2 \neq 0$ or because both β_1 and β_2 are not equal to zero. We present the omnibus test for the Drug variable in Figure 9.7 not as a recommended practice, but only to show what hypothesis is being tested by the “main effect” in the traditional approach to analysis of variance.

Unfortunately, in the traditional approach usually *only* the omnibus tests in Figure 9.7 are presented in a source table. Doing so ignores readily available detailed information about the data revealed by the single-degree-of-freedom contrast codes.

We can similarly assess the main effect for the Psychotherapy variable. In this case there is only one code so that test is the same as the test of the null hypothesis that $\beta_3 = 0$. Finally, we can group together the two codes X_4 and X_5 , which we constructed from products of other codes. Each of these two products involved one code for the Drug categorical variable and one code for the Psychotherapy categorical variables. So, in the traditional language of ANOVA, they are known together as the “Drug \times Psychotherapy Interaction.” Testing the Drug \times Psychotherapy interaction is equivalent to comparing:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$H_0: \beta_4 = \beta_5 = 0$$

A traditional ANOVA source table similar to the one in Figure 9.7 would contain rows only for Drug, Psychotherapy, Drug \times Psychotherapy, Within groups (or Error), and Total. Such a source table, because it aggregates individual contrasts into more global effects, omits information that is readily available in the individual contrasts. In other words, the traditional ANOVA table fails to analyze the variance as much as is possible using the focused comparisons of contrast codes.

Any standard multiple regression program can produce the analysis in Figure 9.7. Equivalently, one can use the equations from Chapter 8 expressing the parameter estimates and their corresponding sums of squares as a function of the contrast codes and the cell means. As an illustration, we use the formulas to calculate the estimate b_1 and its associated SSR. The coefficient that estimates β_1 is:

$$\begin{aligned} b_1 &= \frac{\sum_k \lambda_{1k} \bar{Y}_k}{\sum_k \lambda_k^2} \\ &= \frac{1(32) + 1(26) - 2(17) + 1(17) + 1(19) - 2(9)}{1^2 + 1^2 + (-2)^2 + 1^2 + 1^2 + (-2)^2} \\ &= \frac{42}{12} = 3.5 \end{aligned}$$

Then the sum of squared error reduced by including X_1 in the model along with all the other predictors is:

$$\begin{aligned} \text{SSR}_{b_1} &= \frac{\left(\sum_k \lambda_{1k} \bar{Y}_k\right)^2}{\sum_k \lambda_{1k}^2/n_k} \\ &= \frac{[1(32) + 1(26) - 2(17) + 1(17) + 1(19) - 2(9)]^2}{(1^2 + 1^2 + (-2)^2 + 1^2 + 1^2 + (-2)^2)/3} \\ &= \frac{42^2}{12/3} = 441 \end{aligned}$$

Note that because two-way ANOVA is exactly the same as a one-way ANOVA with a clever set of codes, the formulas and all other details from Chapter 8 apply unaltered to two-way ANOVA. Thus, in terms of parameter estimates and statistical tests there is nothing new that needs to be learned to do two-way ANOVA—it is *exactly* the same as one-way ANOVA.

INTERPRETATION OF COEFFICIENTS

The one new difficulty introduced by the generalization of one-way ANOVA to two-way and higher ANOVA is the interpretation of the coefficients. But it is simply an interpretation problem and not a difficulty involving statistical procedures. And the strategy for interpreting coefficients in two-way and higher ANOVA is the same as the strategies we have developed in Chapters 7 and 8. The new difficulty is interpretation of the coefficients for the predictor variables that code the interactions—the products between one-way codes. Newcomers to ANOVA, whether using our approach or more traditional approaches, often have initial difficulty understanding the concept of an interaction. We therefore devote considerable attention in this section to the interpretation of interactions. We do so by considering two different approaches to interpreting the coefficients in a model with two-way codes such as those in Figure 9.5. Those two approaches are (a) re-expression of the codes in terms of the cell means, as was done in Chapter 8, and (b) interpretation of product predictors as changes in simple slopes, as was done in Chapter 7. In each approach we also consider the interpretation of the coefficients for predictors not involving products because interactions are best understood in comparison.

Interpretation in Terms of Cell Means

The interpretation of the parameter estimates associated with the X predictors proceeds in the same manner as for any set of predictors in the one-way analyses of Chapter 8. The surprisingly large values, relative to Model C being correct, of $F_{1,12} = 105.8$ and $PRE = .95$ for X_1 reject the null hypothesis that $\beta_1 = 0$. To see exactly what the rejection of this null hypothesis means, we express β_1 in terms of the contrast code using the equation from Chapter 8 that expresses the parameter estimate as a function of the codes and the cell means. That is, an equivalent statement of the null hypothesis $\beta_1 = 0$ is:

$$\beta_1 = \frac{\sum_k \lambda_{1k} \mu_k}{\sum_k \lambda_{1k}^2} = \frac{\mu_{AT} + \mu_{BT} - 2\mu_{PT} + \mu_{AC} + \mu_{BC} - 2\mu_{PC}}{12} = 0$$

where μ_{AT} represents the true but unknown mean for the group that received Drug A and participated in Psychotherapy treatment, etc. Doing a little algebra (multiplying by 12 and moving the terms for the placebo conditions to the right side of the equation) yields the following equivalent statement of the hypothesis:

$$\mu_{AT} + \mu_{AC} + \mu_{BT} + \mu_{BC} = 2(\mu_{PT} + \mu_{PC})$$

Dividing each side of this equation by 2 gives:

$$\frac{\mu_{AT} + \mu_{AC}}{2} + \frac{\mu_{BT} + \mu_{BC}}{2} = 2\left(\frac{\mu_{PT} + \mu_{PC}}{2}\right)$$

If we let $\mu_A = (\mu_{AT} + \mu_{AC})/2$, which is the mean of the Drug A conditions averaged (indicated by the dot) across both the Treatment and Control levels of the Psychotherapy variable, $\mu_B = (\mu_{BT} + \mu_{BC})/2$ and $\mu_P = (\mu_{PT} + \mu_{PC})/2$, then the null hypothesis reduces to:

$$H_0: \mu_A + \mu_B = 2\mu_P.$$

Dividing each side of this equation by 2 gives:

$$\frac{\mu_A + \mu_B}{2} = \mu_P.$$

That is, concluding that $\beta_1 \neq 0$ is equivalent to concluding that the average of the drug conditions combined is not equal to the average for the placebo groups, ignoring whether the patients received psychotherapy treatment or not. Thus, unlike the ambiguous conclusions using the Z contrast codes, rejection of the null hypothesis for β_1 clearly implies that the mood scores for the drug groups, on average, differed from the mood scores of the placebo groups. With experience it is usually easy to go directly from the contrast codes to a statement of the null hypothesis without doing the formal derivation as above. For example, the $(1, -1, 0)$ code for X_2 compares the Drug A groups with the Drug B groups so that the null hypothesis $\beta_2 = 0$ is equivalent to the following null hypothesis:

$$H_0: \mu_A - \mu_B = 0 \text{ or } H_0: \mu_A = \mu_B.$$

The direction and magnitude of a difference revealed by rejecting the null hypothesis is given by b_j , the estimate of β_j . For X_1 , this estimate is $b_1 = 3.5$. As with all regression coefficients, this means that our model prediction \hat{Y}_i increases, on average, by 3.5 units for each unit increase in X_1 . The change on X_1 from the placebo groups ($X_1 = -2$) to the drug groups ($X_1 = 1$) is a change of 3 units, so the predicted difference between the placebo and drug groups is $3 \times 3.5 = 10.5$. Indeed, the average of the four drug groups, $(32 + 26 + 17 + 19)/4 = 23.5$, exceeds by 10.5 the average of the two placebo groups, $(17 + 9)/2 = 13$. Figure 9.8(a) depicts graphically the means compared by the first contrast.

We cannot conclude in this analysis that the estimate $b_2 = 1.0$ for β_2 is reliably different from zero. Nevertheless, it is useful to interpret it to see which means are not significantly different from one another. On X_2 there is a two-unit difference between those taking Drug B and those taking Drug A, regardless of whether they were also receiving psychotherapy treatment or not. Thus, the mood scores of those taking Drug A are $2 \times 1 = 2$ points higher than the mood scores of those taking Drug B, but this difference is not statistically significant. Figure 9.8(b) depicts this comparison graphically.

The coefficient $b_3 = 5.0$ is significantly different from zero. The predictor X_3 codes the difference between the two psychotherapy groups, so we can conclude that, on average, there is a significant difference between the mood scores of those receiving treatment compared to those in the control group. That is, we can reject the null hypothesis:

$$H_0: \mu_{.T} = \mu_{.C}$$

where the dots in the subscript indicate that we averaged across levels of drug. To be specific about the amount of difference in mood scores between receiving psychotherapy versus not, we need to interpret the meaning of the estimate $b_3 = 5.0$ as before. There is a two-unit difference on X_3 between receiving Treatment ($X_3 = 1$) and receiving the Control ($X_3 = -1$). Thus, the predicted average difference between the two groups is $2 \times 5 = 10$. Indeed, the average of the three groups receiving Treatment $(32 + 26 + 17)/3 = 25$, exceeds by 10 the average of the three Control groups $(17 + 19 + 9)/3 = 15$. Figure 9.8(c) depicts these means and their comparison.

Even though the coefficient for $b_4 = 0.5$ is not significantly different from zero, its interpretation provides a good introduction to understanding interactions: products of contrast codes. We begin by considering the full expression for β_4 in terms of the unknown means:

$$\beta_4 = \frac{\sum_k \lambda_{4k} \mu_k}{\sum_k \lambda_{4k}^2} = \frac{\mu_{AT} + \mu_{BT} - 2\mu_{PT} - \mu_{AC} - \mu_{BC} + 2\mu_{PC}}{12} = 0$$

Except for the signs on the last three terms, this is very similar to the expression for β_1 above. Putting the T and C terms on opposite sides of the equation gives:

$$\frac{\mu_{AT} + \mu_{BT} - 2\mu_{PT}}{12} = \frac{\mu_{AC} + \mu_{BC} - 2\mu_{PC}}{12}$$

Dividing the numerators and denominators on both sides by 2 yields:

$$\frac{1}{6} \left[\frac{\mu_{AT} + \mu_{BT}}{2} - \mu_{PT} \right] = \frac{1}{6} \left[\frac{\mu_{AC} + \mu_{BC}}{2} - \mu_{PC} \right]$$

Note that the terms in each of the two brackets are identical to the question asked by the first code: Is there a difference in mood scores between those taking either drug compared to those taking the placebo? The first coded predictor X_1 asked whether there was a difference regardless of whether the patient was also receiving psychotherapy treatment or not. The interaction code $X_4 = X_1 \times X_3$ asks whether that difference, whatever it might be, is the same for those receiving psychotherapy treatment:

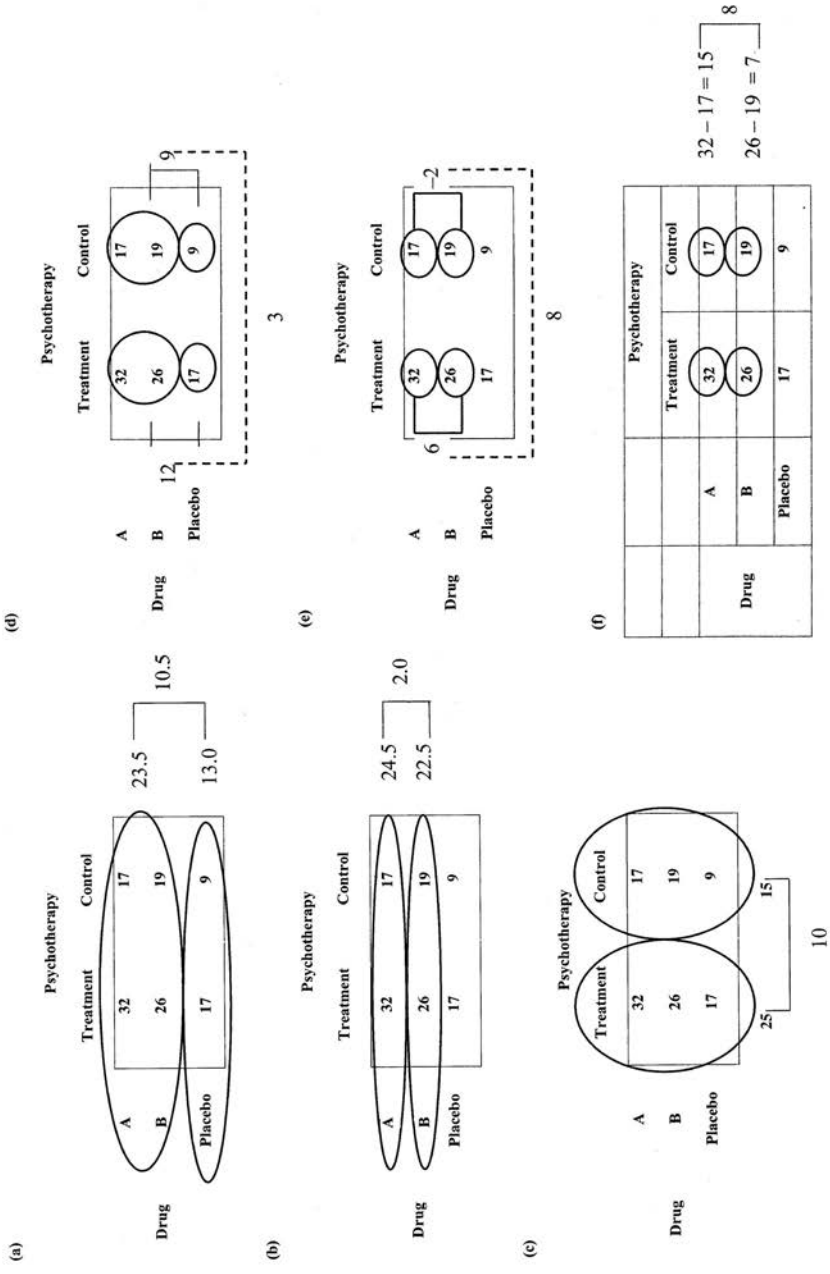
$$\frac{\mu_{AT} + \mu_{BT}}{2} - \mu_{PT}$$

as it is for those in the control condition:

$$\frac{\mu_{AC} + \mu_{BC}}{2} - \mu_{PC}$$

In other words, the null hypothesis for this and all interactions is whether two differences are equal. Different differences imply an interaction. Figure 9.8(d) depicts the means and differences involved in this comparison. For the Treatment groups, the means for those receiving a drug (either A or B) are 32 and 26 (the top-left oval) and their mean

FIGURE 9.8 Means compared by contrast codes



of 29 differs from the mean of the Treatment group taking the placebo (17) by 12 points. For the Control group, the mean of those taking either Drug A or Drug B (the top-right oval) is $(17 + 19)/2 = 18$, which exceeds the mean of those taking the placebo (9) by 9 points. The statistical test for β_4 is therefore equivalent to asking whether these two differences—12 for the Treatment group and 9 for the Control group—are significantly different from each other. Or, equivalently, whether the difference of the differences ($12 - 9 = 3$) is significantly different from zero. In this case, because we could not reject the null hypothesis that $\beta_4 = 0$, we also cannot reject the null hypothesis that the Drug versus Placebo difference is the same for the Treatment group as it is for the Control group. Note that the algebra above shows that the estimated coefficient $b_4 = 0.5$ equals one-sixth of the difference of the differences. Indeed, $0.5 \times 6 = 3$, which is the difference of the differences depicted in Figure 9.8(d).

The algebra to determine which differences are being compared in an interaction may seem complicated at first, but again with experience it is usually easy to jump from the code to a specification of the differences being compared. For example, $X_2 \times X_3$ so the null hypothesis:

$$H_0: \beta_5 = 0$$

is equivalent to testing whether the difference coded by X_2 —the difference between those taking Drug A versus Drug B—depends on the level coded by X_3 , whether the patient is in the Treatment or Control groups. Thus, the equivalent null hypothesis is:

$$H_0: \mu_{AT} - \mu_{BT} = \mu_{AC} - \mu_{BC}$$

In words, the null hypothesis is that the difference between Drugs A and B is the same for those receiving Psychotherapy treatment as it is for those in the Control group. Knowing that $b_5 = 2$ is reliably different from zero allows us to conclude that the drug effect is different for those receiving treatment than for those in the control group. Figure 9.8(e) depicts this difference and shows its magnitude. For those receiving Treatment, the difference in mood scores for those taking Drug A compared to those taking Drug B equals $32 - 26 = 6$, whereas for those in the Control group the difference is in the other direction: $17 - 19 = -2$. The difference of the differences is 8 and is significantly different from zero. Note that although it appears that Drug A is better for those receiving Treatment but Drug B is better for those in the Control group, we have no statistical test of these two individual differences. In other words, we do not know whether the drug difference of 6 for the Treatment group or the difference of -2 for the Control group differs significantly from zero. We simply know that 6 is significantly different from -2 . We return later to address the specific question of whether there are significant differences between the drugs within each treatment condition.

In Chapter 7 we noted that we can always interpret an interaction with either variable in the product as the focal variable. The same is true for products of categorical predictors. Thus, we can focus on whether the Treatment versus Control difference is the same for those taking Drug A as it is for those taking Drug B. That is, the null hypothesis is:

$$H_0: \mu_{AT} - \mu_{AC} = \mu_{BT} - \mu_{BC}$$

Figure 9.8(f) depicts this way of viewing the same interaction depicted in Figure 9.8(e) and displays the differences being compared. For those receiving Drug A, the

Treatment—Control difference equals $32 - 17 = 15$, but for those receiving Drug B the difference is only $26 - 19 = 7$. The difference of the differences is $15 - 7 = 8$, the same as before, as it must be. In other words, Treatment is more effective if one takes Drug A instead of Drug B. We therefore have two equivalent interpretations for the interaction implied by rejecting the null hypothesis that $\beta_5 = 0$. Indeed, for any interaction there will be multiple ways to express it. It will often help to understand the interaction if all the possible interpretations are considered, but it is important to realize that they are all equivalent statements of the same effect.

Interpretation in Terms of Slopes

It is also useful to examine the coefficients for two-way contrast-coded predictors using the same procedures we developed in Chapter 7 for multiple regression models involving product terms. We begin with the estimated equation for Model A:

$$\hat{Y} = 20 + 3.5X_1 + X_2 + 5X_3 + 0.5X_4 + 2X_5$$

Then we substitute for the product definitions ($X_4 = X_1X_3$ and $X_5 = X_2X_3$) to get:

$$\hat{Y} = 20 + 3.5X_1 + X_2 + 5X_3 + 0.5X_1X_3 + 2X_2X_3$$

Next we regroup the terms to express the “simple” linear relationship between \hat{Y} and one of the three predictors X_1 , X_2 , or X_3 . We will begin with X_3 ; regrouping terms yields:

$$\hat{Y} = (20 + 3.5X_1 + X_2) + (5 + 0.5X_1 + 2X_2)X_3$$

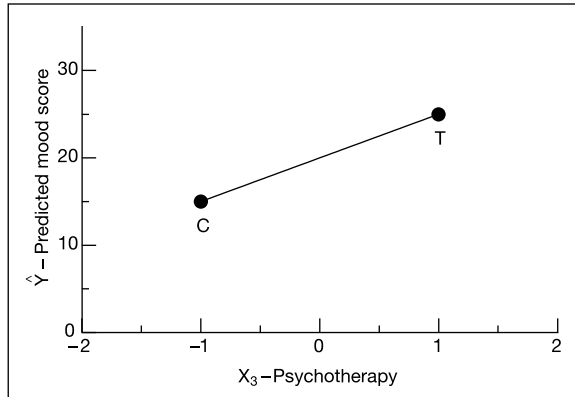
which expresses the “simple” linear relationship between \hat{Y} and X_3 . The term in the first set of parentheses represents the intercept (given specific values of X_1 and X_2), and the term in the second set of parentheses represents the slope (i.e., as always, the change in \hat{Y} given a unit change in X_3 , given specific values of X_1 and X_2). Remember that X_3 codes whether or not the patient received psychotherapy treatment, so this linear relationship reflects the effectiveness of treatment versus control—the steeper the slope, the more effective the treatment relative to the control.

To understand the equation expressing \hat{Y} as a linear function of X_3 , we begin by examining Figure 9.9, which displays the relationship between \hat{Y} and X_3 when $X_1 = 0$ and $X_2 = 0$.

Both X_1 and X_2 are contrast-coded predictors with an equal number of observations in each cell, so average values of X_1 and X_2 are zero; hence, the line in Figure 9.9 depicts the average relationship between \hat{Y} and X_3 . For $X_1 = 0$ and $X_2 = 0$, the relationship between \hat{Y} and X_3 reduces to:

$$\hat{Y} = 20 + 5 \times 3$$

The intercept of 20 is, as always, the predicted value when $X_3 = 0$. Although X_3 is never zero (it is either +1 or -1), its average value is zero. Hence, we can interpret the intercept as the average value across levels of X_3 . Indeed, 20 is the grand mean for these data. The slope for X_3 is 5, which implies that for every unit change in X_3 the predicted value \hat{Y} changes by 5. The two-unit change from $X_3 = -1$ (Control) to $X_3 = +1$ (Treatment) predicts a change in \hat{Y} (mood scores) of $2(5) = 10$ points. This, as it must be, is the same interpretation we obtained for $b_3 = 5$ in terms of cell means. On average, ignoring the Drug condition, those receiving Treatment had mood scores 10 points higher than those in the Control condition.

FIGURE 9.9 Simple relationship between \hat{Y} and X_3 when $X_1 = 0$ and $X_2 = 0$ 

What happens to the simple relationship between \hat{Y} and X_3 if we do not focus on the average relationship, but instead allow X_1 and X_2 to be different from zero? The intercept term in the “simple” relationship is:

$$(20 + 3.5X_1 + X_2)$$

so if $X_1 = 1$ (indicating Drug A or B) then the intercept will increase by 3.5 points and if $X_1 = -2$ (indicating Placebo) then the intercept will decrease by 7 points. These represent the changes in \hat{Y} due to X_1 when X_3 equals zero (which it does equal on average). On average, then, \hat{Y} for the four drug groups ought to be $3.5 + 7 = 10.5$ points higher than \hat{Y} for the two placebo groups. That is, on average, mood scores for those receiving Drug A or B are 10.5 points higher than for those in the Control group. This is necessarily the same interpretation we obtained in terms of cell means. Similarly, for $X_2 = +1$ (Drug A), $X_2 = 0$ (Placebo), and $X_2 = -1$ (Drug B), the intercept increases, stays the same, or decreases. These changes in \hat{Y} (mood scores) are similarly interpreted.

Now we turn to the slope in the “simple” relationship between \hat{Y} and X_3 and observe what happens when X_1 and X_2 are not equal to zero. The slope term is:

$$(5 + 0.5X_1 + 2X_2)$$

so the slope relating \hat{Y} and X_3 *changes* as a function of both X_1 and X_2 . If $X_2 = +1$ (Drug A), then the slope for the prediction increases by $2 \times 1 = 2$ units. If $X_2 = -1$ (Drug B) then the slope decreases by $2 \times -1 = -2$ units. The steepness of the slope for X_3 represents the relative effectiveness of Treatment versus Control. Hence, the steeper slope for Drug A indicates that the relative effectiveness of Treatment versus Control is greater for Drug A than for Drug B. This is the essence of any ANOVA interaction. If there is an interaction, *the effect of one categorical variable depends on the level of another variable*. In this case, the relative effectiveness of Treatment depends on whether the patient is taking Drug A or Drug B.

Similarly, different values of X_1 also yield different slopes. If $X_1 = +1$ (either Drug A or B), then the slope for X_3 increases slightly by $b_4 = 0.5$ units, and if $X_1 = -2$ (Placebo) then the slope decreases by $0.5(-2) = -1$ units. However, $b_4 = 0.5$ is not statistically different from zero so there is no evidence that the Treatment versus Control difference depends on whether any Drug was administered ($X_1 = -1$) as compared to the Placebo.

Figure 9.10 depicts the changes in the intercept and the slope for X_3 as X_1 and X_2 change. Clearly, the different values of X_1 (Drug versus Placebo) yield noticeably different intercepts, but the changes in the slope for X_3 (Treatment versus Control) due to changes in X_1 are minimal. Although it may not appear to be the case, the slope for the bottom line ($X_1 = -2$, or Placebo) is slightly flatter than the average of the slopes of the top two lines ($X_1 = +1$, Drugs). In contrast, the intercept changes due to X_2 (Drug A versus Drug B) are noticeable but much smaller than those due to X_1 (any Drug versus Placebo). However, the slope changes due to X_2 are obvious. Indeed, the lines cross within the range of values used in this study. That is, mood scores are higher for Drug A ($X_2 = +1$) than for Drug B ($X_2 = -1$) for those receiving Treatment ($X_3 = +1$). For those in the Control condition ($X_3 = -1$) the reverse is true—mood scores for Drug B are higher than those for Drug A. These differences in slopes are the essence of an interaction.

Similar to the above interpretation of the augmented model in terms of the simple linear relationship between \hat{Y} and X_3 , we could also do the interpretation in terms of \hat{Y} and X_2 . Although the interpretation of the interaction between X_2 and X_3 must necessarily be equivalent, the alternative interpretation in terms of the relationship between \hat{Y} and X_2 may produce different insights. We do not provide that interpretation here, but it is recommended as an exercise for the reader to construct and interpret the relevant graph similar to Figure 9.10. Finally, we could do the interpretation in terms of the simple linear relationship between \hat{Y} and X_1 . However, that is unlikely to be useful for these data because the only interaction involving X_1 is $X_4 = X_1 \times X_3$ and the coefficient b_4 for that interaction was not significantly different from zero.

Summary of Interpretation

Although both interpretation strategies yield the same interpretation, each is useful for providing a slightly different perspective. For interpreting a given ANOVA, you should use as many of the strategies as necessary until a clear interpretation of the model emerges. Below we suggest a summary interpretation that one might include in a research report:

On average, both drug treatments produced higher mood scores ($M = 23.5$) than did the placebo ($M = 13$), $PRE = .90$, $F(1, 12) = 105.8$, $p < .0001$. There was not a statistically significant difference in the average mood scores produced by Drug A ($M = 24.5$) versus Drug B ($M = 22.5$), $PRE = .19$, $F(1, 12) = 2.9$, $p = .11$. Averaged across all three drug conditions, patients receiving psychotherapy treatment had higher mood scores ($M = 25$) than did those in the control group ($M = 15$), $PRE = .90$, $F(1, 12) = 108$, $p < .0001$. However there was a significant interaction between type of drug administered (A versus B) and whether or not the patient received psychotherapy treatment, $PRE = .49$, $F(1, 12) = 11.5$, $p < .005$, such that Drug A, relative to Drug B, produced an increase in mood scores of 6 points for those receiving treatment, but a slight decrease of 2 points for those in the control group.

Useful adjuncts to the above journal summary are graphs of the cell means (which of course are also the predicted values for the full model) as a function of the drug and psychotherapy variables. Figure 9.10 (which we developed in the section on interpreting slopes) and Figure 9.11 present two views of these data. In Figure 9.10, the differences between the lines for the drug treatments represent the “Drug” differences. In particular,

FIGURE 9.10 Simple relationship between \hat{Y} and X_3 for different values of X_1 and X_2 (filled circles represent cell means)

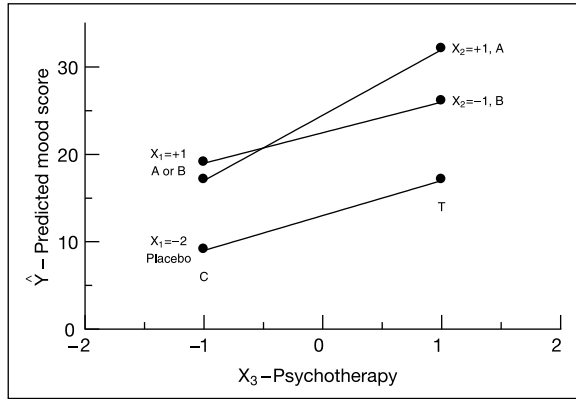
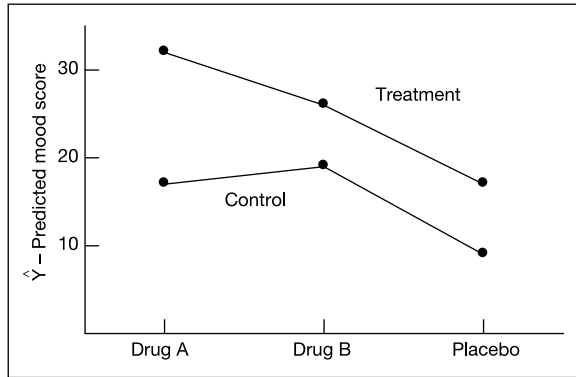


FIGURE 9.11 Cell means for mood scores by drug and psychotherapy treatment



the relatively large difference between the two lines for Drug A and Drug B and the line for Placebo corresponds to the large value for b_1 , and the small difference between the Drug A and Drug B lines corresponds to the small value for b_2 . That the three lines for the drug treatment groups are not parallel—the differences between lines are not constant—indicates the interaction between the drug and psychotherapy variables. In particular, the crossing of the Drug A and Drug B lines corresponds to the statistically significant value of b_5 ($X_5 = X_2X_3$), which asks whether the difference between the Drug A and Drug B lines is constant or dependent on whether the patient was in the Treatment or Control groups. The relatively large difference between the two treatment lines in Figure 9.11 similarly corresponds to the large value of b_3 and the non-parallelism corresponds to the statistically significant value of b_5 . Each of these graphs depicts the interaction, but from different perspectives, just as we were able to examine the interaction from different perspectives in each of the interpretation strategies presented above.

HIGHER ORDER ANOVA

Our strategy for two-way ANOVA was to generate an appropriate set of contrast codes reflecting that two-way structure and then do the statistical analysis exactly as we did for one-way ANOVA. We can use that same strategy when we have three or more categorical predictor variables. Once we have an appropriate set of contrast codes for an experimental design with more than two categorical variables, the statistical analysis proceeds just as before. We therefore only need to learn a procedure for generating the appropriate codes. That procedure is a generalization of the same procedure we used for two-way ANOVA. For a factorial design of q categorical variables, the appropriate contrast codes are constructed according to the following rules:

1. For each of the q categorical variables, develop a set of orthogonal one-way contrast codes.
2. For each *pair* of categorical variables, construct additional contrast codes by multiplying all possible pairs of contrast codes, one from each categorical variable.
3. For each *triple* of categorical variables, construct additional contrast codes by multiplying all possible triads of contrast codes, one from each categorical variable.
4. Continue this procedure for each *quadruple*, *quintuple*, etc., until products are formed using codes from all q categorical variables simultaneously.

If k_1, k_2, \dots, k_q are the respective number of categories for each of the q categorical variables, then the above algorithm for generating contrast codes will produce the complete set of $k_1 \times k_2 \times \dots \times k_q - 1$ orthogonal contrast codes.

As an example, suppose we also wanted to use the patient's gender as a predictor variable in the study of mood scores. Figure 9.12 shows the design of such a study with 12 groups, for which we will need 11 orthogonal contrast codes. We can use the five contrast codes we have already specified for the Drug by Psychotherapy design. Let λ_6 be the code for gender, with -1 indicating males and $+1$ indicating females. Then we need to form the products between that code and each code from the other variables. For the Drug by Gender interactions:

$$\lambda_7 = \lambda_1 \times \lambda_6$$

$$\lambda_8 = \lambda_2 \times \lambda_6$$

and for the Psychotherapy by Gender interaction:

$$\lambda_9 = \lambda_3 \times \lambda_6$$

Finally, for the Drug by Psychotherapy by Gender three-way interactions, we form the products of each triple of categorical codes:

$$\lambda_{10} = \lambda_1 \times \lambda_3 \times \lambda_6$$

$$\lambda_{11} = \lambda_2 \times \lambda_3 \times \lambda_6$$

This provides the 11 orthogonal contrast codes required by the 12 groups of the three-way design. Figure 9.13 shows the codes for each group. Select some codes and verify that they sum to zero as required for contrasts. Select several pairs of codes and verify that their crossproducts sum to zero as required by orthogonality.

FIGURE 9.12 A three-way factorial design: Drug (3) by Psychotherapy (2) by Gender (2)

Drug	Gender			
	Male		Female	
	Psychotherapy		Psychotherapy	
	Control	Treatment	Control	Treatment
A				
B				
Placebo				

FIGURE 9.13 Orthogonal contrast codes for the three-way design of Figure 9.12

Group	A,C,M	B,C,M	P,C,M	A,T,M	B,T,M	P,T,M	A,C,F	B,C,F	P,C,F	A,T,F	B,T,F	P,T,F
λ_1	1	1	-2	1	1	-2	1	1	-2	1	1	-2
λ_2	1	-1	0	1	-1	0	1	-1	0	1	-1	0
λ_3	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1
$\lambda_4 = \lambda_1\lambda_3$	-1	-1	2	1	1	-2	-1	-1	2	1	1	-2
$\lambda_5 = \lambda_2\lambda_3$	-1	1	0	1	-1	0	-1	1	0	1	-1	0
λ_6	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1
$\lambda_7 = \lambda_1\lambda_6$	-1	-1	2	-1	-1	2	1	1	-2	1	1	-2
$\lambda_8 = \lambda_2\lambda_6$	-1	1	0	-1	1	0	1	-1	0	1	-1	0
$\lambda_9 = \lambda_3\lambda_6$	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1
$\lambda_{10} = \lambda_1\lambda_3\lambda_6$	1	1	-2	-1	-1	2	-1	-1	2	1	1	-2
$\lambda_{11} = \lambda_2\lambda_3\lambda_6$	1	-1	0	-1	1	0	-1	1	0	1	-1	0

Interpretation of the one-way codes $\lambda_1, \lambda_2, \lambda_3$, and λ_6 and the two-way codes $\lambda_4, \lambda_5, \lambda_7, \lambda_8$, and λ_9 use the same strategy illustrated for interpreting codes in two-way ANOVA. The three-way codes λ_{10} and λ_{11} ask whether a given two-way interaction depends on the level of the third variable. For example, we can represent $\lambda_{10} = \lambda_1\lambda_3\lambda_6$ as $\lambda_1\lambda_3 \times \lambda_6$, asking whether the two-way interaction of λ_1 (Drug versus Placebo) by λ_3 (Treatment versus Control) is the same for each level of λ_6 (Female versus Male). Equivalently, the representation $\lambda_{10} = \lambda_1\lambda_6 \times \lambda_3$ shows that the three-way interaction equally asks whether the two-way interaction of λ_1 (Drug versus Placebo) by λ_6 (Female versus Male) depends on the level of λ_3 (Treatment versus Control). Finally, the representation $\lambda_{10} = \lambda_3\lambda_6 \times \lambda_1$ asks whether the two-way interaction of λ_3 (Treatment versus Control) by λ_6 (Female versus Male) depends on the level of λ_1 (Drug versus Placebo). At a fundamental level, these three questions are the same. However, depending on the focus of the study (is it about drug effects, psychotherapy effects, or gender differences?), one representation of each three-way interaction will be most appropriate to report.

The global tests, as for one-way ANOVA, are not generally as useful as the focused questions represented by the one-degree-of-freedom contrasts. However, the testing of interactions is one instance for which global tests are sometimes useful. The proliferation of interaction contrasts in factorial designs makes testing and interpreting individual contrasts unwieldy. Interpretations of three-way and higher order interactions are often so theoretically ad hoc that some data analysts recommend examining only those interactions having a priori theoretical predictions. If the higher order interactions are

not expected and not to be examined, then it might be better to eliminate them from the analysis, using only those contrasts that will be considered. This effectively includes those eliminated interaction contrast codes in the “Error” term of the source table. Rather than eliminating the higher order interactions outright, a more conservative strategy that is frequently recommended is to do a global test of the higher order interactions and then pool those contrasts in the Error term only if the null hypothesis is not rejected for the global test. This strategy protects against missing something very unusual in the data while greatly simplifying the data analysis and its presentation. An extra benefit is that the statistical power of the other contrasts will be increased if the F for the higher order interactions is < 1 , because then more will be gained by adding degrees of freedom to the within-group (Error) term than will be lost by adding extra SS in that term.

When eliminating interaction terms one must remember the rule from Chapter 7 that products represent interactions only if all the components of the product are also included in the model. This implies, for example, that if the three-way interactions are retained, then none of the one-way contrasts can be eliminated. Less obviously, it also implies that none of the two-way contrasts can be eliminated because, as illustrated above, any three-way interaction can be represented as the product of any of the two-way interactions and one of the one-way contrasts.

OTHER DETAILS IN FACTORIAL ANOVA

In all other details factorial ANOVA with two or more categorical variables is exactly the same as one-way ANOVA with one categorical variable. Specifically, confidence intervals, problems with nonorthogonal codes, handling unequal numbers of observations in each group, source tables, computational formulas for b_j and SSR_j for contrast-coded predictors, planned and post hoc comparisons, and statistical power are exactly the same. However, we provide additional details on two issues—asking other questions and assessing statistical power—in order to highlight some issues that frequently arise in factorial ANOVA.

Asking Other Questions

It is sometimes not possible to generate a set of orthogonal contrast codes addressing all the theoretical questions that one might want to ask of a given set of data. This is often true for questions that span two or more categorical variables. The strategy above for generating codes for a standard factorial design allows choice of the one-way codes for each categorical variable but then determines the remaining interaction codes as products of those one-way codes. Those products may not ask the most relevant theoretical questions. For example, in a 2×2 design there is only one choice (+1, -1) for each one-way code, so there is no choice for the interaction code, which must be the product of the two one-way codes. Thus, the necessary codes for two variables A and B are:

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
λ_1	1	1	-1	-1
λ_2	1	-1	1	-1
λ_3	1	-1	-1	1

The interaction code λ_3 asks whether the average of the A_1B_1 and A_2B_2 groups equals the average of the other two groups. However, the theoretical question of interest might be whether the mean of one group, say A_2B_2 , differs from the average of the other three groups. For example, the research hypothesis might be that both A_2 and B_2 must be present for an effect to exist. If so, the code of interest is $[-1, -1, -1, 3]$. This is not an interaction because it cannot be represented as the product of two contrast codes, one for each categorical variable. Nevertheless, if it is the question of interest, then it should be asked. We can compute its SSR with the usual formula and then compare it as a planned comparison (as in Chapter 8) to the mean within-group error (MSE) resulting from the analysis of variance using λ_1 through λ_3 . Or, if that question is the focus of the study, one may simply ignore the factorial structure of the design and analyze the data using a set of one-way codes for four groups.

Other questions may also present themselves after the initial analysis. We can use the same post hoc comparison procedures that we used in Chapter 8. For example, in the mood score data, whether Drug A was better than Drug B depended on whether the patient was also receiving psychotherapy. Examining the means, it appears that Drug A is more effective for those in Treatment but Drug B is more effective for those in the Control group. This suggests the recommendation that Drug A, even though it was superior on average, only be administered to those concurrently receiving psychotherapy. But before making such a recommendation we ought to test with post hoc comparisons whether the difference between the means within levels of Psychotherapy are statistically significant. Comparisons such as this between categories (Drug A versus Drug B) of one variable within a single level (Treatment) are known as *simple effects*. We do the calculation for this comparison as an example. The codes for the two relevant simple effects are:

	A,T	B,T	P,T	A,C	B,C	P,C
λ_6	1	-1	0	0	0	0
λ_7	0	0	0	1	-1	0

The code λ_6 asks whether there is a difference between Drug A and Drug B for those receiving Treatment and the code λ_7 asks the same drug difference question for those in the Control condition. Then:

$$SSR_6 = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k \lambda_k^2/n_k} = \frac{(32 - 26)^2}{1^2/3 + (-1)^2/3} = 54$$

The MSE (from Figure 9.4 or 9.7) is 4.2, so:

$$F = \frac{SSR}{MSE} = \frac{54}{4.2} = 12.86$$

If this were not a theoretically motivated, planned comparison, but only became a relevant question after examining the data, it would be advisable to compare F with the Scheffé adjusted critical value of:

$$(m-1)F_{m-1;n-PA;\alpha} = 5F_{5,12;05} = 5(3.11) = 15.55$$

$F = 12.86$ is below the adjusted critical value, so we would not be able to conclude for a post hoc comparison that Drug A was significantly better than Drug B for those patients receiving Treatment. The difference for λ_7 is even smaller, so we know without calculation that it too would not be statistically significant as a post hoc comparison.

Note that we have the odd situation in which we know that the effects of the drugs were different in the two conditions (the interaction coded by λ_5 was significant, so the Drug A vs. Drug B difference of 6 for Treatment is significantly different from the difference of -2 for Control), but neither simple effect is significant (for Treatment, 6 is not statistically different from 0; and for Control, -2 is not statistically different from 0). They are different questions, so in any analysis they certainly can have different answers. The remedy would be to conduct a replication study with more statistical power or with the simple effects as now legitimate planned comparisons.

Power in Factorial ANOVA

The methods we have presented before for estimating power work equally well for estimating statistical power in factorial ANOVA. It is interesting to examine how the power is affected for questions we ask about a given variable when additional categorical variables are added to the analysis. As an example, again consider the mood score data of Figure 9.1. If we were mainly interested in the questions about the effects of the drugs, we could have performed a one-way analysis of variance on these data, completely ignoring the psychotherapy variable. If we ignored whether patients were in the Treatment or Control groups, we would have three groups—Drug A, Drug B, and Placebo—each with six observations. The respective means would be $(32 + 17)/2 = 24.5$, $(26 + 19)/2 = 22.5$, and $(17 + 9)/2 = 13$. The two contrast codes would be $(1, 1, -2)$ and $(1, -1, 0)$. Applying the usual formulas from Chapter 8, we obtain $b_1 = 3.5$ with $SSR = 441$ and $b_2 = 1$ with $SSR = 12$, exactly the same values as before. However, PRE and F would be different because we are now testing the null hypothesis $\beta_1 = 0$ by comparing:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_2 X_2 + \varepsilon_i$$

This is the same basic question asked when testing $\beta_1 = 0$ in the complete two-way analysis. However, the statistical power is not the same because the question is being tested in the context of a different Model A/C comparison.

To see the likely changes in power due to ignoring the other categorical variable, let us consider Figure 9.14, which presents the complete source table for the one-way ANOVA ignoring the psychotherapy variable. The parameter estimates, the sums of squares, the degrees of freedom, and the mean square errors reduced are exactly as they were in Figure 9.7. However, the error sum of squares and the error degrees of freedom have increased substantially. The reason is that the sums of squares that were reduced by X_3 , X_4 , and X_5 in the two-way analysis are now included in the error because those predictors are not used in this analysis. Given the same SSRs for X_1 and X_2 and increased SSE, then the F and PRE values must be considerably less than they were before. For example, in the complete two-way analysis of Figure 9.7, for X_1 $F_{1,12} = 105.8$ and PRE = .90, but in the one-way analysis only considering the Drug treatment variable, $F_{1,15} = 11.9$ and PRE = .44. The differences between the means for Drug A, Drug B, and Placebo

have not changed, so clearly the two-way analysis provided considerably more statistical power for testing $\beta_1 = 0$ and $\beta_2 = 0$. This increase in power is generally the case in multiway factorial ANOVA. Using one or more other categorical variables to reduce SSE allows a more powerful test of hypothesis concerning another categorical variable. The only situation for which there would be a decrease in statistical power would be if the F values for the codes of the additional variables and the codes for the interactions were < 1.0 . If $F < 1.0$ for a contrast, then the proportional reduction in error associated with that contrast is less than we would expect for a randomly chosen parameter. Hence, including that contrast-coded predictor would not be worth the loss of a degree of freedom due to the extra parameter, and so power goes down. Therefore, one should add other categorical variables for the purpose of increasing statistical power only if one expects those variables and their interactions to themselves reduce error in the criterion variable. Interactions proliferate rapidly as other categorical variables with many levels are added. One must be cautious that those many interactions do not thereby reduce statistical power.

FIGURE 9.14 One-way ANOVA of mood scores ignoring the psychotherapy variable

Source	b_j	SS	df	MS	F	p	PRE
Between groups (Drug)		453	2	226.5	6.1	.012	.45
X_1	3.5	441	1	441	11.9	.004	.44
X_2	1.0	12	1	12	0.3	.58	.02
Within groups (MSE)		557	15	37.1			
Total		1010.0	17				

SUMMARY

In a factorial design, there are two or more categorical predictor variables, and all levels of each categorical variable are combined with all levels of other categorical variables. We analyze data from a two-way or higher order factorial design by applying the one-way ANOVA techniques from Chapter 8 to an appropriate set of contrast codes. We generate this set of codes by developing separate sets of one-way contrast codes for each categorical variable and then forming appropriate products among these codes to represent the interaction between categorical variables. Interpreting the interaction contrasts can be difficult, but the same interpretative techniques from Chapters 7 and 8 apply.

Notes

- 1 We use Z here instead of the customary X because subsequently we will do an analysis of these same data using a different set of contrast codes. The Z values are the contrast-coded predictor variables for the first analysis and the X values will be the contrast-coded predictor variables for the second analysis.
- 2 Note that the Z values are independent of each other, because they are based on orthogonal contrast codes and there are equal numbers of observations in each group. Hence, the sums of squares in Figure 9.4 for the individual contrast-coded predictors sum to the overall sum of squares for the complete model.

ANCOVA

10

Models with Continuous and Categorical Predictors

The material covered in this chapter represents an integration of the material covered in Chapters 6 and 9. In Chapter 6 we considered models with continuous predictors; Chapter 9 was concerned with models with categorical predictors. We will now consider models that include both sorts of predictors. Such models are more traditionally covered under the heading of analysis of covariance (ANCOVA). We feel that this is a confusing label for these models and prefer to think of them simply as models involving both categorical and continuous predictor variables.

In Chapter 7, we saw that an extension of models with continuous predictors included products of those predictors, or interactions among them, as predictors as well. Similarly, in the last chapter we considered as predictors interactions among categorical predictor variables. By extension, we will also consider in this chapter models that include not only categorical and continuous predictors but also interactions between the two kinds of predictor variables.

Interest in these sorts of models originally developed within the tradition of methods for the analysis of experimental and quasi-experimental research designs. In this tradition, researchers were primarily interested in the effects of experimental factors, or the categorical predictor variables, and wished to estimate those effects while controlling for some continuously measured concomitant variable, called a *covariate*. As is detailed below, this control might be desired because of the increase in power that such control might produce or because the concomitant variable was redundant with the categorical ones. Models involving both categorical and continuous predictor variables have, however, a wide range of application outside of experimental research designs. For instance, in sociology or political science, one might be interested in the effects of both a categorical variable (e.g., sex) and a continuous one (e.g., personal income) on some dependent variable, and might wish to estimate each of these effects when controlling for the other. In other words, unlike in the experimental design domain, the focus in these models need not be on the effects of the categorical variables and what happens to those effects when a continuous predictor variable is controlled. We might be just as interested in the effects of the continuous predictor variable and how those effects change when we control for a categorical one.

Given the historical tradition of these models within the context of experimental design, we start the chapter by illustrating the use of models with both categorical and continuous predictors within the context of an experimental design, where the primary interest is in the effects of the categorical variables. We use this context in order to

illustrate the reasons we might want to control for a continuously measured concomitant variable. In later sections of the chapter, we illustrate the use of these models in a context where the researcher's interest is primarily in the effects of the continuous predictor when a categorical one is controlled.

CONTROLLING FOR AN ORTHOGONAL CONTINUOUS VARIABLE IN A FACTORIAL DESIGN

Suppose we were evaluating a curriculum innovation in a secondary school. Students were randomly assigned to either the new or the old curriculum. In addition, each curriculum was taught by two different teachers, and students were assigned to one of the two teachers on a random basis. Thus, the experimental design is a two-factor crossed design, with two levels of both the curriculum and teacher factors. Ten students have been randomly assigned to each of the resulting four conditions. The dependent variable is the student's score on a standardized achievement test given at the end of the curriculum. In addition, we have a pretest measure from each student, indicating his or her achievement in the domain in question prior to exposure to either the new or old curriculum. The hypothetical raw data are given in Figure 10.1. The variable Z_i represents each student's pretest achievement score. The X_{1i} variable is a contrast-coded variable that codes curriculum (-1 if Old Curriculum; 1 if New Curriculum); X_{2i} is a contrast-coded variable that codes teacher (-1 if Teacher A; 1 if Teacher B); and Y_i represents each student's post-test achievement score.

The means for both Z_i and Y_i for each of the four cells of the design are given in Figure 10.2. Notice that the data have been constructed so that all four pretest means equal 50. Since students have been randomly assigned to the four treatment conditions

FIGURE 10.1 Hypothetical experimental data

Y_i	X_{1i}	X_{2i}	Z_i	Y_i	X_{1i}	X_{2i}	Z_i
58	1	-1	50	57	1	-1	49
63	1	-1	49	61	1	-1	52
65	1	-1	53	57	1	-1	50
56	1	-1	47	67	1	-1	51
60	1	-1	53	56	1	-1	46
50	-1	-1	49	62	-1	-1	54
58	-1	-1	51	55	-1	-1	48
52	-1	-1	50	63	-1	-1	52
55	-1	-1	47	50	-1	-1	46
57	-1	-1	52	58	-1	-1	51
61	1	1	47	59	1	1	49
71	1	1	53	65	1	1	54
68	1	1	52	60	1	1	46
58	1	1	48	65	1	1	51
68	1	1	51	65	1	1	49
47	-1	1	46	62	-1	1	51
56	-1	1	51	51	-1	1	49
63	-1	1	53	54	-1	1	51
53	-1	1	48	58	-1	1	50
52	-1	1	47	54	-1	1	54

after they took the pretest, we would expect their mean pretest scores to be very similar, although the chances are that they would not all be identical. (They have been constructed to be identical here for didactic purposes.)

The reason for constructing these data with equal pretest means is that as a result the pretest is not correlated with condition. To see this, suppose we regressed the pretest Z_i on X_{1i} , X_{2i} , and their interaction. With all three contrast codes as predictors, the predicted values of Z_i would equal the four cell means. Each of these cell means also equals the grand mean. Hence, a model with all three contrast-coded predictors generates the same predicted values for every case as a simple model in which we predict the grand mean for every case. As a result, condition is entirely unrelated to pretest.

FIGURE 10.2 Pretest and post-test means by condition

Condition	\bar{Z}_k	\bar{Y}_k
Old Curriculum, Teacher A	50	56
Old Curriculum, Teacher B	50	55
New Curriculum, Teacher A	50	60
New Curriculum, Teacher B	50	64

ANOVA of the Data

Let us conduct a straightforward two-way ANOVA on the post-test data, ignoring the pretest variable for the moment. We regress Y_i on X_{1i} , X_{2i} , and X_{3i} , where X_{3i} is the interaction contrast code, formed by multiplying X_{1i} and X_{2i} together. The resulting model is:

$$\hat{Y}_i = 58.75 + 3.25X_{1i} + 0.75X_{2i} + 1.25X_{3i}$$

with a sum of squared errors of 710. The value of the regression coefficient for X_{1i} , 3.25, equals half the difference between the average of the two mean values of Y_i under the new curriculum and the average of the two mean values of Y_i under the old curriculum. The regression coefficient for X_{2i} , 0.75, equals half the difference between the average under Teacher B and the average under Teacher A. The regression coefficient for the interaction contrast code, 1.25, equals half the difference between the average of the Old Curriculum/Teacher A and New Curriculum/Teacher B conditions and the average of the Old Curriculum/Teacher B and New Curriculum/Teacher A conditions. Conceptually it tells us that the curriculum difference is larger under Teacher B than under Teacher A.

We can calculate the sum of squares due to each contrast-coded predictor by computing a series of compact models that omit each contrast code in turn. Alternatively, we can use the formula for the SSR associated with a contrast-coded predictor that we have used before:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k \right)^2}{\sum_k \frac{\lambda_k^2}{n_k}}$$

Accordingly, the sum of squares explained by X_{1i} equals 422.5, that explained by X_{2i} equals 22.5, and that explained by X_{3i} equals 62.5. Since we have 10 observations in each of the four conditions, the three predictors are nonredundant and hence these sums of squares add up to the between condition sum of squares. The ANOVA source table for this analysis is presented in Figure 10.3.

FIGURE 10.3 Two-way ANOVA source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Between or model		507.5	3	169.17	8.58	.42
Curriculum	3.25	422.50	1	422.50	21.42	.37
Teacher	0.75	22.50	1	22.50	1.14	.03
Curriculum x Teacher	1.25	62.50	1	62.50	3.17	.08
Error		710.00	36	19.72		
Total		1217.50	39			

Including an Orthogonal Covariate

In this analysis, we have made no use of the pretest Z_i . We might decide to add it as a predictor to the model for a number of reasons. For instance, we might be interested in asking how curriculum and teacher affect post-test performance when we control for the pretest or when we look within levels of the pretest. In other words, we might be interested in these effects over and above differences in performance that existed at the time of the pretest. A seemingly different reason for including it as a predictor in the model is that we might expect it to be highly correlated with the post-test, since presumably it is only an earlier version of the post-test, measuring the same domain of achievement. If this is so, then, as we shall see, it might make our tests of curriculum and teacher effects, and their interaction, considerably more powerful.

In these data, the pretest scores are highly related to the post-test scores. The sum of squared errors from a simple regression model in which Y_i is regressed on Z_i equals 855.57. From the last or “Total” row of the source table of Figure 10.3 we see that the sum of squares of Y_i for the simplest single-parameter model equals 1217.5. Accordingly, a comparison between the two-parameter simple regression model using Z_i to predict Y_i and a one-parameter model yields a PRE of .297 and $F_{1,38} = 16.05$. Thus, the relationship between the pretest and the post-test is substantial and highly reliable.

As we have seen, the pretest Z_i is uncorrelated with condition, since all of the pretest conditions means are identical. Accordingly, the sum of squares of Y_i that could be explained by Z_i will not overlap with the sums of squares of Y_i explained by the three condition contrast codes. Because the sum of squares explained by these three contrast-coded predictors equals 507.5, the sum of squares explained by Z_i equals 361.93, and Z_i is nonredundant with the three contrast-coded predictors, the sum of squares explained by the pretest, Z_i , plus the three contrast-coded predictors ought to equal $507.5 + 361.93$ or 869.43. Accordingly, the sum of squared errors for a model in which Y_i is regressed on Z_i , X_{1i} , X_{2i} , and X_{3i} should equal $1217.5 - 869.43$ or 348.07.

In the ANOVA source table of Figure 10.3, the denominator for each of the F statistics testing the significance of the condition differences equals the mean square error within or the mean square error from the final augmented model that includes all predictor variables. If we include the pretest Z_i as an additional predictor, this mean square error ought to be reduced substantially due to the fact that the sum of squares potentially explainable by Z_i has been controlled for or removed from the sum of squared errors. Thus, by including Z_i as an additional predictor in the model, we would expect our tests of condition differences in Y_i to be more powerful, yielding larger F values.

The resulting augmented model, with both the pretest and the three contrast-coded predictors, is estimated as:

$$\hat{Y}_i = -4.52 + 1.27Z_i + 3.25X_{1i} + 0.75X_{2i} + 1.25X_{3i}$$

with a sum of squared errors of 348.07.

Notice that the coefficients in this model for the three contrast-coded predictors have not changed from the model where Z_i was not included as a predictor. Since Z_i is nonredundant with all of the contrast-coded predictors, its inclusion has no effect on the value of their coefficients or on their interpretation. The coefficient for X_{1i} continues to tell us about the magnitude of the difference in Y_i due to curriculum. The coefficient for X_{2i} continues to tell us about the magnitude of the mean teacher difference. And the coefficient for X_{3i} continues to inform us about the degree to which the curriculum difference is larger under Teacher B than under Teacher A.

We could test whether the coefficient for Z_i significantly differs from zero by comparing this augmented model with the model presented earlier that included only the contrast-coded predictors. The value of PRE that results from this comparison is:

$$\text{PRE} = \frac{710.00 - 348.07}{710.00} = .510$$

with an associated F statistic of

$$F_{1,35} = \frac{.510/1}{(1 - .510)/35} = 36.39$$

Accordingly, we can conclude that, independent of condition, or on average within condition, pretest scores significantly relate to post-test scores.

We can test whether the condition means differ by testing the regression coefficients in this model for each of the three contrast-coded predictors. Let us start with the omnibus test of whether there are any differences in the condition means of Y_i . To do this test, we want to compare the augmented model that includes Z_i and the three contrast-coded predictors with a compact one that includes only Z_i . We have already said that the sum of squared errors from this compact simple-regression model equals 855.57. Accordingly, the value of PRE for the omnibus test is:

$$\text{PRE} = \frac{855.57 - 348.07}{855.57} = .593$$

And the omnibus F statistic, having 3 and 35 degrees of freedom, is:

$$F_{3,35} = \frac{.593/3}{(1 - .593)/35} = 17.01$$

An equivalent expression for the F statistic is given in terms of the sums of squares:

$$F_{3,35} = \frac{507.50/3}{348.07/35} = 17.01$$

where 507.50 is the reduction in the sum of squared errors as we move from the compact to the augmented model, and 348.07 is the sum of squared errors of the augmented model.

Notice that this F statistic is nearly twice as large as the F statistic for the omnibus test based on the augmented model that did not include the pretest as a predictor (given in Figure 10.3). The reason for this difference is that the sum of squared errors from the augmented model is now considerably less than it was without the pretest included. It is also true that the degrees of freedom for error have been reduced by 1 as a result of the additional parameter for the pretest estimated in the augmented model. In combination, however, the substantially smaller sum of squared errors and the slightly smaller degrees of freedom for error result in a considerably smaller mean square error.

In other words, the denominator of the F ratio for the omnibus test of condition differences when the pretest is included equals $348.07/35$ or 9.94 . In the model that did not include the pretest, the denominator of the F ratio for the omnibus test of condition differences equaled $710.0/36$ or 19.72 . The net result of including the pretest as a predictor, then, has been to reduce the mean square error and increase the power of tests of the condition effects. Such an increase in power will happen whenever there is a significant relationship between the continuously measured predictor variable, in this case the pretest, and the dependent variable in the augmented model that includes the categorical predictor variables. The full source table for the model that includes the pretest and all three contrast-coded predictors is given in Figure 10.4. Let us compare this ANCOVA source table with the ANOVA source table of Figure 10.3.

First, notice that the test of the overall model—comparing this model that includes four predictor variables with the single parameter model that includes no predictors—yields substantially higher values of both PRE and F than the test of the overall model that did not include the pretest. These larger values are entirely attributable to the fact that the pretest is highly related to the dependent variable. Accordingly, the sum of squares attributable to the model, 869.43 , has dramatically increased from the ANOVA table, while the error sum of squares has dramatically decreased. The decrease in the sum of squared errors for the model, and equivalently the increase in the sum of squares explained by the model, is exactly equal to the sum of squares associated with the pretest, i.e., 361.93 .

Second, notice that the rows for the three contrast-coded variables, representing curriculum, teacher, and their interaction, have the same sums of squares, degrees of freedom, and mean squares as they did in the ANOVA source table. Because the pretest is unrelated to curriculum, teacher, and their interaction, neither the regression coefficients for these contrast-coded predictor variables nor their sums of squares are affected by the inclusion of the pretest in the model. The inclusion of the pretest, however, does have a major effect on the row in the source table referring to error in the model, as just described. As a result, all of the F statistics used to test the teacher effect, curriculum effect, and the interaction effect on the post-test dependent variable are substantially larger than they were in the ANOVA source table of Figure 10.3. Whereas the curriculum \times teacher interaction was not significant in the ANOVA source table, it now is. The positive coefficient associated with this interaction contrast code tells us that the new–old curriculum difference, while significant on average across Teachers, is significantly larger for Teacher B than it is for Teacher A.

This analysis has illustrated one of the major reasons for including a continuously measured predictor variable (equivalently called a covariate) in randomized experimental research designs. If that predictor variable is measured prior to randomization of participants to conditions, on average it will be unrelated to the contrast-coded variables

FIGURE 10.4 ANCOVA source table

<i>Source</i>	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model		869.43	4	217.36	21.87	.71
Pretest	1.27	361.93	1	361.93	36.39	.51
Between conditions		507.50	3	169.17	17.01	.59
Curriculum	3.25	422.50	1	422.50	42.48	.55
Teacher	0.75	22.50	1	22.50	2.26	.06
Curriculum x Teacher	1.25	62.50	1	62.50	6.28	.15
Error		348.07	35	9.94		
Total		1217.50	39			

that code experimental condition. The purpose of including such a variable is to increase the power of the analysis that tests for condition differences in the dependent variable. If the covariate is in fact unrelated to condition, then neither the regression coefficients for the various contrast-coded predictors that code condition differences nor their sums of squares will be affected by its inclusion in the model. The null hypothesis associated with the test of a given contrast-coded predictor will also not change. Regardless of the inclusion of the covariate, we will still be testing for differences among the condition means on the dependent variable, as coded by the λ values. Tests for condition differences will be more powerful as a result of including a covariate whenever the test of whether the covariate's regression coefficient in the full augmented model differs from zero yields a significant F statistic. When the covariate is unrelated to the dependent variable, then the decrease in the sum of squared errors resulting from the inclusion of the covariate will not offset the decrease of the degrees of freedom for error in the model. The ideal covariate, therefore, in this situation, is one that is as highly associated as possible with the dependent variable controlling for the categorical variables or within levels of the categorical variables.

Even with random assignment of participants to condition after measuring the covariate, it will almost never be the case that the covariate will be entirely nonredundant with the condition contrast codes. In other words, it will be a very rare event for all of the pretest or covariate means in the various experimental conditions to be exactly equal. Our example, then, is obviously a constructed one, designed simply to illustrate what happens in the pure case, when the covariate is completely independent of condition. In any given study, there will in all probability be some nonsignificant relationships between the covariate and the contrast codes that represent condition. Nevertheless, the inclusion of a covariate will increase the statistical power of tests of condition differences, given a covariate that is significantly related to the dependent variable within levels of the categorical variables.

As we said in the introduction to this chapter, within the context of experimental designs the usual interest in including a continuously measured predictor variable with a set of categorical ones is to examine what happens to tests of condition differences when we control for the continuously measured covariate. As we have seen, with a covariate measured prior to random assignment of participants to condition, the result will generally be an increase in statistical power for tests of condition differences. There is no necessary reason, however, for confining our interpretations of the model that includes both kinds of predictors to this typical interest. In other words, there is nothing

to prevent us from turning the interpretation of this model around—concentrating not on the tests of mean differences while controlling for the covariate, but on a test of the pretest–post-test relationship while controlling for condition differences on the post-test. If we simply regress the post-test on the pretest, the pretest’s regression coefficient equals 1.27. The sum of squared errors for this simple regression model equals 855.57. A test of the simple pretest–post-test relationship yields a PRE of .297 and an F of 16.05 with 1 and 38 degrees of freedom. When we examine the pretest–post-test relationship controlling for the categorical variables, as given in the ANCOVA source table of Figure 10.4, the pretest’s coefficient is still 1.27, but a test of whether it is reliably related to the post-test within condition levels yields a PRE of .51 and an F of 36.39 with 1 and 35 degrees of freedom. Thus, we might equivalently look at this analysis as a way of increasing the power of tests of the pretest–post-test relationship by controlling for experimental conditions.

ANALYSIS OF POST-TEST–PRETEST DIFFERENCE SCORES

In giving a rationale for the analysis that includes pretest as a predictor, we suggested that it might make sense to examine the effects of curriculum and teacher on the post-test when controlling for pretest differences or holding constant pretest performance. It might seem that an equivalent way of doing this analysis would be to examine condition differences in improvement from the pretest to the post-test. To do such an analysis, we might logically compute a new dependent variable equal to $Y_i - Z_i$, assuming they were both measures of the same thing—the pretest taken before the experiment and the post-test at its conclusion. This *difference score* tells us about each individual’s improvement in achievement during the course of the study. We would then be interested in condition differences in the mean $Y_i - Z_i$ difference scores. Since all Z_i condition means are identical, the mean differences among conditions on the $Y_i - Z_i$ difference scores will be equivalent to the mean differences among conditions on Y_i .

To examine condition effects on this improvement difference score, let us regress it on the three contrast-coded predictors that define condition. The following estimated parameters result:

$$\widehat{Y_i - Z_i} = 8.75 + 3.25X_{1i} + 0.75X_{2i} + 1.25X_{3i}$$

with a sum of squared errors of 364.00. The intercept in this model equals the average of the condition means on the difference score or, equivalently, the difference between the means of the condition means of Y_i and of the condition means of Z_i . Somewhat surprisingly, perhaps, the regression coefficients for the three contrast-coded predictors have not changed as a result of changing the dependent variable to the $Y_i - Z_i$ difference score. Both in this difference score analysis and in the analysis where the post-test was the dependent variable and the pretest was included as a predictor, the regression coefficients continue to equal what they did in the simple analysis of variance with Y_i as the dependent variable and no pretest. This invariance is once again due to the fact that the pretest is uncorrelated with condition.

To illustrate algebraically why these regression coefficients have not changed, let us examine the algebraic expression for the regression coefficients associated with

contrast-coded predictors in this difference score analysis. Using our standard formula, but substituting the difference score means for the usual \bar{Y} values, we get:

$$\frac{\sum_k \lambda_k (\overline{Y_k - Z_k})}{\sum_k \lambda_k^2} = \frac{\sum_k \lambda_k (\bar{Y}_k - \bar{Z}_k)}{\sum_k \lambda_k^2} = \frac{\sum_k \lambda_k \bar{Y}_k - \sum_k \lambda_k \bar{Z}_k}{\sum_k \lambda_k^2}$$

Since all \bar{Z}_k are identical, the expression $\sum \lambda_k \bar{Z}_k$ equals zero and this expression for the regression coefficient for each of the predictors reduces to what it is when simply Y_i is the dependent variable in the model. In sum, these regression coefficients equal the coded condition differences in the mean difference scores ($\bar{Y}_k - \bar{Z}_k$), which, given equal \bar{Z}_k , are equivalent to the coded differences in \bar{Y}_k .

This difference score analysis gives us the difference score ANOVA source table of Figure 10.5. Notice that just as the regression coefficients for the contrast-coded predictors in this difference score analysis equal what they were in the analysis that included pretest as a predictor variable, so too are their sums of squares equal to what they have been all along. Once again, this equivalence is due to the fact that Z_i is uncorrelated with condition.

This analysis, however, is different from both the ANOVA and the ANCOVA in terms of the sum of squares for error and the sum of squares total. The total sum of squares is now equal to the sum of squared variation in the $Y_i - Z_i$ difference scores, which in this case is less than the total sum of squares in Y_i . Since this total sum of squares has been reduced and since the sums of squares explained by the three contrast-coded predictors are unchanged, the error sum of squares in this analysis must be less than the sum of squared errors in the earlier analysis of variance with Y_i as the dependent variable. As a result, the F and PRE statistics for the omnibus test of any condition differences and for the three single-degree-of-freedom tests are all larger than they were in the original ANOVA source table of Figure 10.3.

Notice, however, that the sum of squared errors in this difference score source table is larger than the sum of squared errors was in the analysis that included the pretest as a predictor variable (ANCOVA source table of Figure 10.4). As a result, the F and PRE statistics testing condition differences in this difference score analysis are all slightly smaller than they were in the source table of Figure 10.4. In sum, while this difference score analysis is more powerful in this case than the simple ANOVA of Y_i , it is not as powerful as the ANCOVA in which the dependent variable was Y_i and Z_i was included as a predictor variable.

FIGURE 10.5 Difference score ANOVA source table

Source	<i>b</i>	SS	df	MS	<i>F</i>	PRE
Between or model		507.50	3	169.17	16.73	.58
Curriculum	3.25	422.50	1	422.50	41.79	.54
Teacher	0.75	22.50	1	22.50	2.22	.06
Curriculum x Teacher	1.25	62.50	1	62.50	6.18	.15
Error		364.00	36	10.11		
Total		871.50	39			

To understand why this is so, let us examine the difference score model:

$$Y_i - Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

We can re-express this model by adding Z_i to both sides of the equation:

$$Y_i = \beta_0 + Z_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

This difference score model now looks very similar to the ANCOVA model in which the pretest was used as a predictor variable. There is, however, one major difference. Instead of estimating a parameter for the pretest variable as we did in the ANCOVA model, we have set the parameter value equal to 1.0. By doing the difference score analysis, we have in effect assumed that the parameter value for the pretest equals 1.0, rather than letting it be a free parameter and deriving its least-squares estimate.

By definition, the least-squares estimates are those that minimize the sum of squared errors. Accordingly, the sum of squared errors in a model where the coefficient for the covariate is fixed at 1.0 cannot be less than the sum of squared errors in the ANCOVA model, since in the latter model the coefficient for the covariate is the least-squares estimate. Hence, the difference score analysis will generally be less powerful than the ANCOVA. Frequently, it will be substantially less powerful and may, in fact, be even less powerful than the simple ANOVA model.

Once we realize that this difference score model is identical to the ANCOVA model, except that we have fixed the coefficient for the covariate at 1.0 instead of estimating it, we can rewrite the difference score analysis source table as we have in Figure 10.6. In this revised table, the sum of squares total refers to the total sum of squares of Y_i rather than the total sum of squares in the $Y_i - Z_i$ difference score.

FIGURE 10.6 Revised difference score ANOVA source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model		853.50	3	284.5	28.14	.70
Pretest	1.00	346.0	0			
Between conditions		507.50	3	169.17	16.73	.58
Curriculum	3.25	422.50	1	422.50	41.79	.54
Teacher	0.75	22.50	1	22.50	2.22	.06
Curriculum x Teacher	1.25	62.50	1	62.50	6.18	.15
Error		364.00	36	10.11		
Total		1217.50	39			

Notice in this source table that the degrees of freedom for the pretest, Z_i , equal zero since in this model its coefficient has been set at 1.0 rather than estimated from the data. The sum of squares associated with Z_i equals the difference between the sum of squares total for Y_i and the sum of squares total for the $Y_i - Z_i$ difference score.

In these data, the difference score analysis is only slightly less powerful than the ANCOVA. This near equivalence results from the fact that the estimated parameter for the covariate in the ANCOVA model, 1.27, is rather close to the value of 1.0 at which it is fixed in the difference score analysis. We could test whether the ANCOVA model results in significantly smaller errors of prediction than the difference score model. This is equivalent to testing whether the parameter associated with the covariate in the

augmented ANCOVA model significantly differs from 1.0. For this test, the ANCOVA model, in which the covariate's parameter is estimated, is the augmented one:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 Z + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} + \varepsilon_i$$

The compact model is the difference score model in which the pretest's parameter is fixed at 1.0:

$$\text{MODEL C: } Y_i = \beta_0 + Z + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Comparing the sums of squared errors for these two models, we get:

$$\text{PRE} = \frac{364.00 - 348.07}{364.00} = .044$$

which converts to an F of 1.61 with 1 and 35 degrees of freedom. Hence, in these data, we cannot conclude that the parameter for the covariate is significantly different from 1.00. This conclusion means that the ANCOVA model for these data is not significantly more powerful than the difference score analysis.

Having said that the ANCOVA model will generally be more powerful than the difference score analysis, we should note that the ANCOVA uses up a degree of freedom in estimating the covariate's parameter while the difference score analysis does not. It is thus possible, if the estimated slope for the covariate is close to 1.00, for the mean square error from the difference score model to be smaller than that from the ANCOVA model. The difference score analysis may also be easier than the ANCOVA to describe to others who are untrained statistically.

THE CASE OF A PARTIALLY REDUNDANT CONTINUOUS PREDICTOR

So far we have been illustrating one major reason for including a continuously measured concomitant variable in an analysis of condition differences. By including a covariate that is measured prior to randomization of participants to conditions, the covariate will generally be unrelated to condition and its inclusion in the model will increase the power of the tests of mean condition differences, so long as the covariate is highly related to the dependent variable of interest.

But there are other important reasons for including a continuously measured covariate in an analysis of mean condition differences. These reasons arise precisely *because* the covariate is correlated with condition and we wish to examine the condition differences over and above, or controlling for, covariate differences. There are two primary occasions when this is of interest. Analytically, these occasions are identical—we simply will be estimating an ANCOVA model with a covariate that is partially redundant with the categorical predictor(s)—but they differ in important theoretical ways that give rise to rather different interpretations.

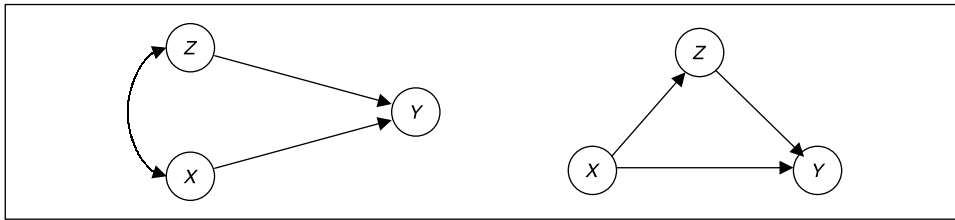
The first occasion on which controlling for a partially redundant covariate is of interest occurs when, for whatever reason, there are pre-existing differences among participants in the various conditions of the design, and the researcher wishes to attempt to examine the effects of the categorical variable(s) free from those differences. Suppose,

for instance, in the example we have been using, that students had not been randomly assigned to the four cells defined by the teacher and curriculum factors. If the decision about which students got which teacher–curriculum condition was not based on a random decision rule, but was instead based on some unknown assignment rule (e.g., keeping last year’s classes intact), then we might expect differences among the various conditions on the pretest long before the students had been exposed to the teacher or to the curriculum. We would then quite reasonably want to control statistically for these pretest differences in looking at condition differences on the post-test. In other words, we might then like to control for the pretest in our analysis so that we could look at differences among conditions on the post-test free from the pre-existing differences on the pretest.

This use of ANCOVA to control for or adjust for known pre-existing differences among participants is particularly common in what have come to be known as quasi-experimental research designs where assignment to condition has not been on a purely random basis (Judd & Kenny, 1981a; Shadish, Cook, & Campbell, 2002). Although this adjustment function of ANCOVA (attempting to equate participants on pre-existing covariate differences) is widely used in such quasi-experimental designs, it is important to recognize that this analytic approach is not a general solution to the problem of causal inference in nonexperimental research. The ability to reach causal conclusions about the effects of various independent variables depends not on the statistical analysis that one undertakes but on the research design.

The second occasion on which it becomes important to conduct an ANCOVA with a covariate that is partially redundant with condition is when one is attempting to test a mediational model (Baron & Kenny, 1986; Judd & Kenny, 1981b). Suppose one has conducted an experiment in which the effect of some treatment difference on a dependent variable is to be estimated. If there is an effect of the experimental conditions (as estimated through an analysis of variance), then there certainly must be an underlying mediating mechanism or process that is responsible for that effect. Often one has an idea about the nature of that mechanism and one may attempt to measure one or more variables that intervene in a causal chain between the experimental manipulation and the ultimate dependent variable that it affects. These intervening variables are known as mediating variables. If, in fact, the hypothesized mediating mechanism is partially responsible for the effect of the experimental treatment on the ultimate dependent variable, then that effect should no longer be as apparent once one conducts an ANCOVA, controlling for the mediating variable.

Although the analytic models in these two cases are essentially identical—to determine if there are condition differences once one controls for another variable that is known to be correlated with condition—they represent different theoretical models about the underlying causal process linking the categorical independent variable(s) with the dependent variable. This difference is perhaps most clearly shown by representing the two cases in terms of hypothesized causal models, which are shown in Figure 10.7. In these causal models, single-headed arrows are used to represent hypothesized causal effects of one variable upon another, while curved double-headed arrows are simply meant to represent a correlation between two variables with no causal claims about the process responsible for that correlation. As previously, we use X to represent a categorical independent variable (here with only two levels, thus needing only one contrast-coded predictor), Z represents the continuously measured covariate, and Y represents the dependent variable.

FIGURE 10.7 Theoretical causal models for two sorts of partially redundant covariates

The first causal model is meant to represent the quasi-experimental case, in which participants have not been randomly assigned to the two levels of the categorical independent variable and therefore there are differences between the two groups on some continuously measured variable Z (X and Z are thus partially redundant with each other). Here one wishes to estimate the effect of X on Y free from the potential confounding influence of Z on Y . The goal then is to somehow come up with a more accurate estimate of the X effect, controlling for or eliminating any confounding due to the partially redundant covariate.

The second case is one in which random assignment to levels of the categorical independent variable has been accomplished, and one measures both the ultimate dependent variable, Y , as well as another variable, Z , that is thought to be affected by X . As the model indicates, here there are two ways in which X may exert an effect on Y : the first is through the mediating variable, Z , and the second is over and above that mediating variable. In this case, one controls for the mediating variable, Z , in an ANCOVA because one is interested in examining whether there is any residual direct effect of X on Y over and above the hypothesized mediating process through Z .

As we have said, the analytic approach in these two cases is identical—estimating effects while controlling for a continuously measured covariate that is partially redundant with the categorical independent variable. Nevertheless, to illustrate both and the differences in interpretations that result from the different causal models underlying them, we provide an example of each one.

Examining Partial Effects in a Quasi-Experimental Situation

In Figure 10.8, we present data that have been somewhat modified from those used in the example earlier in this chapter. The values of Y_i are identical to what we had earlier in Figure 10.1, as are the definitions of the contrast-coded predictor variables, with X_{1i} coding curriculum and X_{2i} coding teacher. What we have modified slightly are the values of the pretest variable, Z_i , so that now there are differences among the four cells of the design (resulting from the crossing of the two factors) on the pretest measure. The four cell means for both the pretest, Z_i , and the post-test, Y_i , are given in Figure 10.9.

There is now a definite relationship in these data between condition and Z_i , the pretest, since its condition means are no longer equal to each other. If we regressed Z_i on the three contrast-coded predictors (X_{3i} being defined as earlier to capture the curriculum – teacher interaction), we get the following parameter estimates:

FIGURE 10.8 Modified pretest–post-test data

Y_i	X_{1i}	X_{2i}	Z_i	Y_i	X_{1i}	X_{2i}	Z_i
58	1	-1	51	57	1	-1	50
63	1	-1	50	61	1	-1	53
65	1	-1	54	57	1	-1	51
56	1	-1	48	67	1	-1	52
60	1	-1	54	56	1	-1	47
50	-1	-1	48	62	-1	-1	53
58	-1	-1	50	55	-1	-1	47
52	-1	-1	49	63	-1	-1	51
55	-1	-1	46	50	-1	-1	45
57	-1	-1	51	58	-1	-1	50
61	1	1	50	59	1	1	52
71	1	1	56	65	1	1	57
68	1	1	55	60	1	1	49
58	1	1	51	65	1	1	54
68	1	1	54	65	1	1	52
47	-1	1	44	62	-1	1	49
56	-1	1	49	51	-1	1	47
63	-1	1	51	54	-1	1	49
53	-1	1	46	58	-1	1	48
52	-1	1	45	54	-1	1	52

FIGURE 10.9 Modified pretest and post-test means by condition

Condition	\bar{Z}_k	\bar{Y}_k
Old Curriculum, Teacher A	49	56
Old Curriculum, Teacher B	48	55
New Curriculum, Teacher A	51	60
New Curriculum, Teacher B	53	64

$$\hat{Z}_i = 50.25 + 1.75X_{1i} + 0.25X_{2i} + .75X_{3i}$$

with a sum of squared errors of 225.94. The estimated slopes in this model tell us about the differences in the mean pretest values, according to the following frequently used expression for the slopes of contrast-coded predictor variables:

$$\frac{\sum_k \lambda_k \bar{Z}_k}{\sum_k \lambda_k^2}$$

To test whether these pretest means are significantly different from each other we could compare this as Model A with a Model C that simply predicted the grand mean of Z_i for every observation. This Model C has a sum of squared errors of 373.5, resulting in a PRE of .40 and an F of 7.84 with 3 and 36 degrees of freedom. Thus, there is significant redundancy between the pretest and the categorical variables that represent the conditions in which an observation is observed.

Turning to the analysis of post-test scores, if we conducted that analysis whilst ignoring the pretest, the exact same model would result as before:

$$\hat{Y}_i = 58.75 + 3.25X_{1i} + 0.75X_{2i} + 1.25X_{3i}$$

and the ANOVA source table that was given earlier (Figure 10.3) would continue to be found, since we have done nothing to alter the Y_i variable in this modified dataset.

However, the ANCOVA model, controlling for the partially redundant pretest variable, looks very different from what it was previously:

$$\hat{Y}_i = -4.84 + 1.26Z_i + 1.04X_{1i} + 0.43X_{2i} + 0.30X_{3i}$$

$$\text{SSE} = 348.07$$

Note that the regression coefficients for the contrast-coded predictors have been dramatically affected by the inclusion of the pretest variable Z_i . No longer does the coefficient for X_{1i} equal half the difference between the average Y_i under the new and old curricula. Similarly, the coefficients for X_{2i} and X_{3i} can no longer be interpreted as they were previously in terms of differences among various \bar{Y}_k . In other words, with the inclusion of a covariate that is partially redundant with the contrast-coded predictors, the regression coefficients for the contrast-coded predictors are no longer equal to:

$$\frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

and therefore a test of whether the parameters for these contrast-coded predictors equal zero is no longer a simple test of a comparison among the Y_i means in the various conditions.

With the inclusion of a covariate, the regression coefficient for a contrast-coded predictor variable is equal to:

$$\frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} - b_z \frac{\sum_k \lambda_k \bar{Z}_k}{\sum_k \lambda_k^2}$$

where b_z is the regression coefficient associated with the covariate in the full model that includes both the covariate and the set of contrast-coded predictors.

To illustrate this expression, let us calculate the value of the regression coefficient for X_{1i} in the model that includes Z_i . Notice that the first half of this expression equals the regression coefficient for X_{1i} in the model that did not include the covariate, i.e., 3.25. The second half of the expression equals the parallel difference coded by the contrast weights among the *covariate or pretest* condition means, weighted by the regression coefficient for the pretest. Numerically, for the coefficient for X_{1i} the second half of this expression equals:

$$b_z \frac{\sum_k \lambda_k \bar{Z}_k}{\sum_k \lambda_k^2} = 1.26 \left(\frac{53 + 51 - 48 - 49}{4} \right) = 1.26(1.75) = 2.21$$

In sum, then, according to this expression the regression coefficient for the X_{1i} contrast-coded predictor equals:

$$3.25 - 1.26(1.75) = 1.04$$

This new expression for the regression coefficient for a contrast-coded predictor in the presence of a covariate is readily interpreted. It is equal to the magnitude of the difference among the \bar{Y}_k coded by the contrast weights, adjusting for or subtracting off the magnitude of the same difference among the covariate condition means \bar{Z}_k . The degree to which the coded comparison among the \bar{Y}_k is adjusted by the same comparison among the \bar{Z}_k depends on the magnitude of the covariate's regression coefficient, b_z . In sum, the regression coefficients for contrast-coded predictors in the presence of a covariate tell us about the differences among the \bar{Y}_k coded by the contrast weights, adjusting that difference for the parallel difference that exists in the covariate condition means \bar{Z}_k . The degree to which this adjustment is performed depends on the magnitude of the within-condition relationship between the covariate and the dependent variable, i.e., the partial regression coefficient for the covariate.

In the case of the regression coefficient for X_{1i} , we know that half the difference between the mean Y_i score under the new and old curricula equals 3.25, i.e., the regression coefficient for X_{1i} not controlling for the pretest. The regression coefficient for X_{1i} controlling for the pretest equals 1.04. This equals half the difference in Y_i associated with the difference in curriculum over and above any pretest differences associated with curriculum.

An equivalent but slightly different way to think about the coefficient for a contrast-coded predictor in the presence of a covariate is that it tells us about the magnitude of differences among adjusted values of \bar{Y}_k , adjusting those condition means to get rid of differences in the covariate condition means. More precisely, we can compute for each condition the adjusted mean:

$$\bar{Y}'_k = \bar{Y}_k - b_z(\bar{Z}_k - \bar{Z})$$

where \bar{Z} is the mean of the condition means for the covariate. We can then use these adjusted means to derive the regression coefficient for a contrast-coded predictor in the model that includes the covariate, using the old formula for the regression coefficient for a contrast-coded predictor. In other words, once we have the adjusted means, the regression coefficient for a contrast-coded predictor in the model that includes the partially redundant covariate equals:

$$\frac{\sum_k \lambda_k \bar{Y}'_k}{\sum_k \lambda_k^2}$$

where \bar{Y}'_k are the adjusted cell means as just defined.

To illustrate, in Figure 10.10 the values of these adjusted cell means, \bar{Y}'_k , are given for the four teacher \times curriculum conditions of our design. These were derived using the formula for the adjusted cell means above. For example, the value of the adjusted mean for the Old Curriculum \times Teacher A condition is given by:

$$\bar{Y}'_k = 56 - 1.26(49 - 50.25) = 57.58$$

FIGURE 10.10 Post-test condition means adjusted for the pretest (\bar{Y}'_k)

Condition	\bar{Y}'_k
Old Curriculum, Teacher A	57.58
Old Curriculum, Teacher B	57.85
New Curriculum, Teacher A	59.05
New Curriculum, Teacher B	60.52

where 56 is the value of \bar{Y}_k for this condition, 1.26 is the regression coefficient for the pretest in the model that includes both the pretest and the condition contrast-coded predictors, 49 equals the pretest mean for this condition, \bar{Z}_k , and 50.25 equals the mean of the four pretest condition means.

We can now use these adjusted \bar{Y}'_k to compute the regression coefficients for the contrast-coded predictors. For instance, the coefficient for X_{1i} in the model that includes the covariate equals:

$$\frac{\sum_k \lambda_k \bar{Y}'_k}{\sum_k \lambda_k^2} = \frac{(-1)57.58 + (-1)57.85 + (+1)59.05 + (+1)60.52}{4} = 1.04$$

Conceptually, then, the coefficients for contrast-coded predictors in the presence of a covariate tell us about the magnitude of the coded differences among adjusted condition means, adjusting those dependent variable means by the extent to which the covariate means depart from each other. Obviously, if all \bar{Z}_k are equal to each other there will be no adjustment, as we saw in the case of an orthogonal covariate. But with a partially redundant covariate, the differences among the \bar{Z}_k will result in some adjustment among the dependent variable means that are compared when examining the regression coefficient for a contrast-coded predictor.

The source table for the ANCOVA model with these modified data is given in Figure 10.11. Notice that the sums of squares and, as a result, the F statistics for the omnibus test of condition differences and the individual contrast tests are all considerably smaller than they were in the ANOVA source table. This is so because they are testing different null hypotheses than they were in the ANOVA model. In the ANOVA model the omnibus test was testing whether there were any differences among the condition means \bar{Y}_k . The tests of the contrasts were testing specific comparisons among these condition means. With the inclusion of the pretest in the model, the omnibus test is now testing for the presence of differences among the adjusted condition means. Similarly, the contrast tests are now testing specific comparisons among these adjusted condition means. In other words, the tests are now examining condition differences in \bar{Y}'_k having adjusted for condition differences that existed on the pretest.

Since the pretest is now partially redundant with the various contrast codes that code condition, the sums of squares in this source table are not additive as they were in the earlier ANOVA table. Therefore, some of them must be derived through the estimation of various compact models. To get the sum of squares for the omnibus condition test, we must estimate a model with only the pretest used as a predictor, since this omnibus

FIGURE 10.11 ANCOVA source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model		869.43	4	217.36	21.87	.71
Pretest	1.26	361.93	1	361.93	36.39	.51
Between conditions		34.30	3	11.43	1.15	.09
Curriculum	1.04	27.81	1	27.81	2.80	.07
Teacher	0.43	7.44	1	7.44	0.75	.02
Curriculum x Teacher	0.30	3.29	1	3.29	0.33	.01
Error		348.07	35	9.94		
Total		1217.50	39			

sum of squares is no longer equal to the sum of the three sums of squares explained by the three contrast-coded predictors. Even though the contrast codes are still orthogonal and even though there are still an equal number of observations in each condition, the three contrast-coded predictors differ in how redundant or correlated they are with the covariate. Thus, the sum of their three individual sums of squares is not equal to the difference in the sum of squares explained if they are all omitted from the model.

Models with two or more partially redundant covariates are simple extensions to the single covariate case that we have just examined. The formula for the adjusted means, for instance, simply adjusts for or subtracts off mean differences on each covariate, each weighted by its regression coefficient. For instance, if Z_{1i} and Z_{2i} are two covariates, then the regression coefficient for contrast-coded predictors in the model that included these two covariates would be examining comparisons among the following (doubly) adjusted cell means:

$$\bar{Y}'_k = \bar{Y}_k - b_{z_1}(\bar{Z}_{1k} - \bar{Z}_1) - b_{z_2}(\bar{Z}_{2k} - \bar{Z}_2)$$

As a final comment, reiterating what we said before, while the ANCOVA permits comparisons among condition means adjusting for one or more partially redundant covariates, this is not a general solution to problems of internal validity in research designs where random assignment of observations to conditions has not been used. Adjustment for redundant covariates is a powerful procedure, but it does not solve the problems that limit causal inferences in nonexperimental research designs.

Examining Partial Effects: The Case of Mediation

All of what we have just said about the analysis of categorical predictor variables when we control for a partially redundant covariate applies as well to the second occasion when such models are of interest, namely, when we wish to examine the process that mediates the effects of some experimental categorical variable on a dependent variable. Let us consider a new example to illustrate the interpretations that ensue from such an analysis in this case. Suppose that participants were randomly assigned to a family counseling intervention designed to improve outcomes for adolescents who suffer from bipolar disorder. Either they receive the family counseling intervention along with the usual pharmacological care or they receive only the usual care. And the question is whether this intervention affects the manifestation of bipolar symptoms eight weeks later. The dependent variable is assessed by a clinical psychologist who is blind to experimental treatment, using a 10-point rating scale (higher scores equal more symptoms).

FIGURE 10.12 Hypothetical data for 20 families

	<i>Treatment</i>		<i>Control</i>	
	<i>Criticism (C_i)</i>	<i>Symptoms (S_i)</i>	<i>Criticism (C_i)</i>	<i>Symptoms (S_i)</i>
	3	4	4	7
	2	3	4	8
	4	4	3	5
	2	3	2	5
	2	4	4	7
	3	5	4	6
	1	3	5	6
	2	5	3	4
	3	7	4	5
	3	5	4	6
Mean	2.5	4.3	3.7	5.9

If the family counseling intervention is effective in reducing symptoms, it is thought that it must operate by affecting the amount of criticism the parents direct at the bipolar adolescent during the interim period. Hence, at week 7 of the study, another clinical psychologist, again blind to experimental treatment, interviews the parents and rates the degree to which the parents spontaneously criticize the adolescent, this time on a five-point scale (higher scores equal more criticism). Hence, for each of 20 families there are three variables: whether they received the experimental treatment or not, assessed parental criticism at 7 weeks, and adolescent symptoms at 8 weeks. Hypothetical scores on these variables for the families are given in Figure 10.12.

Obviously one first wishes to assess whether the treatment had an effect on the ultimate outcome variable: symptoms at 8 weeks. If it did, then the theoretical expectation is that those effects are mediated through the parental criticism variable, measured at 7 weeks. In other words, the thinking is that the treatment effect ought to work, if it does, by affecting the manifestation of parental criticism of the adolescent and this in turn is responsible for lower symptom levels.

These expectations imply the following:

1. There will be an overall treatment effect on symptoms at 8 weeks; the treatment will reduce the number of symptoms.
2. The treatment will lead to lower parental criticism.
3. Parental criticism will predict fewer symptoms at 8 weeks holding constant treatment condition.
4. The effect of the treatment on symptoms will be reduced when parental criticism is controlled.

The first two of these would be assessed by a separate two-group ANOVA model for each of the dependent variables, first symptoms at 8 weeks and then parental criticism at 7 weeks. The third and fourth conditions would be assessed by a single ANCOVA model, using both treatment and parental criticism as predictors of symptoms. Here, we expect that criticism will predict fewer symptoms and that the effect of treatment on symptoms will be reduced compared to its effect in the simple ANOVA model. These are the classic conditions for establishing mediation, as identified by Judd and Kenny (1981b) and Baron and Kenny (1986).

We first contrast-code the treatment variable ($X_i = +.5$ if treatment; $-.5$ if control) and then examine its effects on both symptoms and criticism with two simple regression models, examining the first two of the above four expectations. The overall treatment effect on symptoms is estimated as:

$$\hat{S}_i = 5.10 - 1.60X_i$$

This model, predicting the two group means, has a sum of squared errors of 27.00. The estimated slope for X_i equals the difference between the two treatment means, given the codes we have used. And a test of whether it departs from zero, and thus whether the two group means on symptoms differ, yields a PRE of .32 and an F of 8.53 with 1 and 18 degrees of freedom. Hence, there is an overall treatment effect; the treatment resulted in fewer symptoms measured 8 weeks later.

The effect of treatment on criticism at 7 weeks is examined in the following simple regression model:

$$\hat{C}_i = 3.10 - 1.20X_i$$

with a sum of squared errors of 12.60. Again, the slope for the treatment variable equals the mean difference between the treatment and control groups in criticism and this difference is significant: PRE = .36, $F_{1,18} = 10.29$.

To assess the third and fourth mediation expectations, the ANCOVA model is estimated as:

$$\hat{S}_i = 2.84 - 0.72X_i + 0.73C_i$$

with a sum of squared errors of 20.28. In the previous paragraph we showed that the means for criticism differ significantly between the two conditions, demonstrating that the two predictors in this ANCOVA model are redundant with each other. Accordingly, the coefficient for treatment no longer estimates the difference between the two condition means; rather, it estimates the difference between the two conditions when adjusting or controlling for differences between them on the criticism variable:

$$b_{SXC} = \frac{\sum_k \lambda_k \bar{S}_k}{\sum_k \lambda_k^2} - b_{SCX} \frac{\sum_k \lambda_k \bar{C}_k}{\sum_k \lambda_k^2} = -1.60 - (0.73)(-1.20) = -0.72$$

Notice in this expression that we are indicating each parameter estimate in this model with multiple subscripts, the first letter indicating the dependent variable, the second the predictor involved, and then, following the dot, any other variables controlled for in the model. The reason for this complete notation will become apparent shortly.

In this model, the test of the slope associated with the covariate, criticism, is significant (PRE = .25, $F_{1,17} = 5.63$), while that for the treatment variable no longer is (PRE = .08, $F_{1,17} = 1.40$). Accordingly, it would appear that the four conditions for establishing mediation (outlined above) have been met. Namely, the treatment has an overall effect on the ultimate outcome variable as well as on the mediator. The mediator significantly affects the outcome when controlling for treatment, whereas when controlling for the mediator the effect of treatment on symptoms is no longer significant.

However, such a demonstration does not necessarily mean that there has been a significant reduction in the treatment effect once criticism is controlled. In other words, although the treatment effect no longer differs from zero, it might not significantly differ from its effect in the initial ANOVA model. For this and other reasons, tests of whether the partial regression coefficient for the treatment in the ANCOVA model is significantly smaller than its overall or total effect have been developed. Rather than develop those tests here, we refer the interested reader to MacKinnon, Lockwood, Hoffman, West, and Sheets (2002), MacKinnon (2008), and Judd, Yzerbyt, and Muller (2015). Importantly, however, they rest on the following equality:

$$b_{SX} - b_{SX.C} = b_{SC.X} b_{CX}$$

which means that the reduction of the overall effect of the treatment when the mediator is controlled equals the product of the treatment's effect on the mediator and the mediator's partial effect on the outcome controlling for the treatment.

This equality is none other than our formula for the partial slope associated with a contrast-coded predictor when a redundant covariate is controlled, as just given:

$$b_{SX.C} = \frac{\sum_k \lambda_k \bar{S}_k}{\sum_k \lambda_k^2} - b_{SC.X} \frac{\sum_k \lambda_k \bar{C}_k}{\sum_k \lambda_k^2} = b_{SX} - b_{SC.X} b_{CX}$$

A small amount of algebraic manipulation gives the desired equality: $b_{SX} - b_{SX.C} = b_{SC.X} b_{CX}$. In other words, the tests that have been developed to determine whether the treatment effect is significantly reduced when the mediator is controlled are tests of the partial regression slope associated with the treatment variable. If such tests are significant (i.e., the regression slope is significantly smaller in the ANCOVA model), the treatment variable is said to have an indirect effect on the outcome variable via the mediator.

This mediation example is obviously rather simple and our treatment does not do justice to the extensive literature on the subject. Our point has been simply to illustrate that everything we have said about ANCOVA with a partially redundant covariate applies in the case of mediation assessment as well as in the case of simply a confounded treatment variable. These two differ not in the analyses conducted but in their underlying theoretical model.

The Homogeneity of Regression Assumption in ANCOVA

Most classic treatments of the ANCOVA specify that an assumption, referred to as the homogeneity of regression assumption, is crucial to the use and interpretation of ANCOVA results. To clarify this assumption, why it is potentially important, and what may be done if it is violated, let us return to the data example used earlier, with the two crossed independent variables being curriculum (new versus old) and teacher (A versus B) and a confounded covariate, the pretest. The data we will use were given in Figure 10.8 and the cell means, for both the post-test and the pretest, were given in Figure 10.9. We use the same notation we did earlier: the post-test is Y_i , the pretest is Z_i , X_{1i} is the contrast-coded predictor representing curriculum (+1 if new; -1 if old), X_{2i} is the contrast-coded predictor representing teacher (-1 if A; +1 if B), and X_{3i} is their product, representing the curriculum \times teacher interaction.

The ANCOVA model that we presented earlier for these data was estimated as:

$$\hat{Y}_i = -4.84 + 1.26Z_i + 1.04X_{1i} + 0.43X_{2i} + 0.30X_{3i}$$

$$\text{SSE} = 348.07$$

And we saw that the slopes associated with the contrast-coded predictors could be interpreted either as differences among Y_i cell means, adjusting for parallel differences among the Z_i cell means:

$$\frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} - b_Z \frac{\sum_k \lambda_k \bar{Z}_k}{\sum_k \lambda_k^2}$$

or as differences among the adjusted Y_i cell means, where those adjusted means are given by the expression:

$$\bar{Y}'_k = \bar{Y}_k - b_Z(\bar{Z}_k - \bar{Z})$$

The assumption of homogeneity of regression that underlies this analysis, and the interpretation of its parameter estimates, is that the relationship between the pretest and the post-test is invariant across the four conditions defined by the contrast-coded predictors. To understand this, note that in these expressions for the interpretation of the slopes associated with the contrast-coded predictors we use one value for the slope of the pretest, b_Z , rather than different values for the four different conditions. That is, when calculating the adjusted cell means that are compared, a single value of b_Z is used regardless of the value of k . If the relationship between the dependent variable, Y_i , and the covariate, Z_i , differs substantially in magnitude across the various conditions, then we should not be assuming a single adjustment weight b_Z , but instead should be allowing for different adjustment weights for the various conditions. Accordingly, in controlling for a covariate and interpreting the resulting parameter estimates for the contrast-coded predictors, it makes sense to examine whether the relationship between Y_i and Z_i is invariant or homogeneous across conditions.

To suggest that the relationship between the pretest and the post-test depends on condition is to suggest that condition and the pretest interact in affecting the post-test. Therefore, to test the homogeneity of the relationship between Y_i and Z_i across conditions, we need to test whether the interactions between Z_i , on the one hand, and the condition defining contrast-coded predictors, on the other, are significant. To examine these interactions, we follow the standard procedure of computing products of the variables whose interactions we wish to test and then entering those product variables as separate predictor variables into a model that includes the variables that are components of the products. We then test whether the augmented model that includes the product terms generates significantly better predictions than a compact model that omits all of the pretest by contrast-coded predictor interactions.

Let us illustrate this using the data contained in Figure 10.8. Since we have three contrast codes that define condition, there will be three interaction or product terms to examine the condition by pretest interaction: $X_{1i}Z_i$, $X_{2i}Z_i$, and $X_{3i}Z_i$. When these are included in the model, the parameter associated with the first will estimate the extent to

which the pretest–post-test relationship is homogeneous across the two curriculum levels, the second will estimate the extent to which it is homogeneous across the two teachers, and the third will examine the triple interaction, i.e., whether the pretest–post-test relationship is homogenous across levels of the curriculum \times teacher interaction.

The model that includes all three of these pretest \times condition interactions is estimated as:

$$\hat{Y}_i = -4.40 + 1.26Z_i + 6.11X_{1i} - 1.32X_{2i} - 3.93X_{3i} - 0.10X_{1i}Z_i + 0.03X_{2i}Z_i + 0.08X_{3i}Z_i$$

$$\text{SSE} = 344.23$$

To demonstrate that this model in fact does allow the slope for the covariate to vary between the four cells of our design, let us examine the expression for the “simple” effect of the covariate in the various cells. The model can be rewritten as:

$$\hat{Y}_i = (-4.40 + 6.11X_{1i} - 1.32X_{2i} - 3.93X_{3i}) + (1.26 - 0.10X_{1i} + 0.03X_{2i} + 0.08X_{3i}) Z_i$$

And from this, we can derive the “simple” $Y_i : Z_i$ regression models in each of the four cells of the design, substituting for the various values of X_{1i} , X_{2i} , and X_{3i} . For instance, for the New Curriculum, Teacher B cell, we get:

$$\hat{Y}_i = (-4.40 + 6.11(+1) - 1.32(+1) - 3.93(+1)) + (1.26 - 0.10(+1) + 0.03(+1) + 0.08(+1)) Z_i$$

which reduces to:

$$\hat{Y}_i = -3.54 + 1.27Z_i$$

Parallel expressions for the other three cells of the design give:

$$\text{New Curriculum, Teacher A: } \hat{Y}_i = 6.96 + 1.05Z_i$$

$$\text{Old Curriculum, Teacher B: } \hat{Y}_i = -7.90 + 1.31Z_i$$

$$\text{Old Curriculum, Teacher A: } \hat{Y}_i = -13.12 + 1.41Z_i$$

Clearly this model allows different “simple” slopes for Z_i in the four cells of the design.

Since we wish to examine whether there are any differences among conditions in these “simple” slopes, and since we have no specific expectations that predict such differences, we can conduct an omnibus three-degree-of-freedom test to examine the homogeneity of regression assumption, testing whether the set of three interaction terms leads to a significant improvement in the fit of the model. Thus this interactive model becomes Model A and the earlier ANCOVA model is the Model C with which we want to compare it. This comparison yields the following PRE and F values:

$$\text{PRE} = \frac{348.07 - 344.23}{348.07} = .011$$

$$F_{3,32} = \frac{.011/3}{(1 - .011)/32} = \frac{3.845/3}{344.226/32} = 0.119$$

There is no reason to prefer Model A over Model C. Thus there is no evidence in these data that the homogeneity of regression assumption of ANCOVA has been violated. Equivalently, we have found no evidence to suggest that the relationship between the

pretest and the post-test differs depending on which of the four cells of the research design we are looking at. Using a single slope, and thus a single adjustment weight, to compute the four adjusted means suffices.

As a check on the homogeneity of regression assumption, this omnibus model comparison seems sufficient. However, there may be times when one has prior expectations that one or more of the interactions between the pretest and the contrast-coded predictors would be found. For instance, it is possible that one might have expected the pretest–post-test relationship to be stronger under the old curriculum than under the new one. In such a case, it seems appropriate to examine the slope associated with that particular interaction term, $X_{1i}Z_i$, comparing the augmented model that includes all three interaction terms to a compact one that only leaves out this one hypothesized interaction.

Should it be found that the homogeneity of regression assumption has been violated, either through the omnibus test of the set of interaction terms, or the more focused examination of single interaction terms in the case of prior expectations about their possible existence, then one is left with the model that includes one or more covariate \times condition interactions and one is forced to interpret them. Importantly, then, the homogeneity of regression assumption is an assumption whose violation implies simply that a model more complicated than the ANCOVA model is required and one must interpret the obtained significant interaction term(s). Thus, this assumption is not like the assumptions about the distribution of residuals that we made very early on in this book (i.e., that they are normally distributed, have a single variance, and are independent). Violations of those assumptions mean that our inferential tests are biased. In this case, violations of the homogeneity of regression assumption simply mean that life is a bit more complicated and one must interpret the data in light of the resulting covariate \times condition interactions.

Suppose, for instance, in the present case, that the interaction between Z_i and X_{1i} had in fact proven to be significant. One would then be compelled to provide interpretations in the context of the more complicated interactive model that included at least this significant interaction. And in this case, the resulting interaction could be interpreted by focusing either on the extent to which the pretest–post-test relationship depends on the new versus old curriculum or on the extent to which the magnitude of the new – old curriculum difference depends on the value of the pretest. Although these two interpretations are fundamentally equivalent, given our focus on the condition differences the second is likely to be the preferred interpretation of the interaction coefficient. We will proceed to illustrate the interpretation with the present data, even though our omnibus test clearly failed to show any evidence of the interactions, and the more focused single-degree-of-freedom test of just $X_{1i}Z_i$ is also not significant in these data.

As we saw in Chapter 7, the interpretation of parameter estimates in the presence of significant interactions involving continuous predictors is complicated by the fact that slopes of variables that are components of included product predictors estimate “simple” effects at the value of zero of the other component variables included in the product predictor. Thus, in the case of the current Model A, the estimated slope for X_{1i} (i.e., 6.11) informs us about half the expected curriculum difference for someone scoring zero on the pretest variable, Z_i . Since there are no pretest values in our dataset that are close to zero, interpretations of the parameter estimates in the Model A are rendered

considerably more meaningful if one centers or mean-deviates the covariate in the model that includes the products of the contrast-coded predictors with that centered covariate. The resulting Model A in this centered case is:

$$\hat{Y}_i = 58.86 + 1.26Z_{Ci} + 1.00X_{1i} + 0.36X_{2i} + 0.27X_{3i} - 0.10X_{1i}Z_{Ci} + 0.03X_{2i}Z_{Ci} + 0.08X_{3i}Z_{Ci}$$

where Z_{Ci} is the centered version of the pretest (i.e., $Z_{Ci} = Z_i - \bar{Z}$). In a deep sense, of course, this centered model is identical to the model prior to centering, in that it makes the same predictions and has the same sum of squared errors (i.e., 344.23). But now the slope estimates associated with the three contrast-coded predictors are considerably more interpretable: they estimate the “simple” condition differences at the mean value of the pretest. For instance, the slope associated with X_{1i} in this model (i.e., 1.00) estimates half the predicted difference in post-test scores between the old and new curriculum conditions at the mean value of the pretest and allowing the magnitude of the old – new curriculum difference to depend on the value of the pretest. And the degree to which that difference does depend on the value of the pretest is indicated by the parameter estimate associated with the $X_{1i}Z_{Ci}$ product predictor. As the value of the pretest increases by one unit, the estimated “simple” difference attributable to the old versus new curriculum decreases by 0.10 units. Had this value been significant, we would have concluded that curriculum effects are larger the less well one initially performed.

An equivalent interpretation for this parameter estimate can be given by focusing on the “simple” slopes of the pretest in the two different curriculum conditions derived earlier. That “simple” slope, on average, across the two conditions is 1.26. The parameter estimate associated with the $X_{1i}Z_{Ci}$ product predictor tells us that the “simple” slope is 0.10 smaller in the new curriculum condition and 0.10 units larger in the old curriculum condition. In other words, the pretest–post-test relationship is stronger under the old curriculum than under the new one.

Our interpretation of covariate \times condition interactions in this example has been a bit forced because in fact the data suggest that the homogeneity of regression assumption is not violated. Our purpose has been simply to illustrate how one might proceed if it happened that the assumption *was* violated. One would simply live with the more complex model and interpret the resulting significant covariate \times condition interactions. But there are other conditions where such interactions are the primary focus of research, where one is precisely interested in whether the magnitude of the relationship between some continuous predictor variable (heretofore called a covariate) and the dependent variable depends on the levels of one or more categorical variables. And it is to such an example that we now turn.

MODELS WITH CONTINUOUS AND CATEGORICAL PREDICTORS OUTSIDE OF EXPERIMENTAL CONTEXTS

Our data come from a large western public university and consist of the academic records of 2740 members of the freshman class, either enrolled in the College of Arts and Sciences or in the College of Engineering. We have four variables available on these students: their combined SAT score, taken during their senior year in high school, on a scale of 20–80 (average of verbal and math SAT scores dropping the final digit); their cumulative

FIGURE 10.13 SAT scores and freshman GPAs by college and gender

	SAT		GPA	
	Male	Female	Male	Female
Engineering				
Mean	63.50	61.30	2.88	3.03
SD	(6.02)	(5.60)	(0.70)	(0.63)
<i>n</i>	329	91	329	91
Arts and Sciences				
Mean	58.39	56.62	2.68	2.93
SD	(6.64)	(6.54)	(0.76)	(0.67)
<i>n</i>	968	1352	968	1352

grade point average (GPA) at the end of the freshman year; the college in which they were enrolled; and their gender. We are interested in knowing whether applicants' SAT scores predict their freshman GPA at university and whether that relationship depends both on the college in which they enrolled and gender.

The mean combined SAT scores and freshman GPAs for each of the four groups, defined by college and gender, are given in Figure 10.13

We start by examining mean differences as a function of gender and college in each of these variables, defining X_{1i} as +1 if Female, -1 if Male, X_{2i} as +1 if Arts and Sciences, -1 if Engineering, and X_{3i} as their product. The two ANOVA models, one for each variable, are given below:

$$\begin{aligned}\widehat{SAT}_i &= 59.95 - 0.99X_{1i} - 2.44X_{2i} + 0.11X_{3i} & \text{SSE} &= 115,139.16 \\ \widehat{GPA}_i &= 2.880 + 0.100X_{1i} - 0.077X_{2i} + 0.025X_{3i} & \text{SSE} &= 1355.40\end{aligned}$$

For both variables, the gender difference is significant, although in opposite directions. Females have lower combined SAT scores on average than males ($F_{1,2736} = 124.71$, PRE = .04) but they end up with higher freshman year GPAs ($F_{1,2736} = 58.56$, PRE = .02). There is also a significant college difference for both variables, with students in Engineering having both higher SAT scores ($F_{1,2736} = 196.53$, PRE = .07) and higher freshman GPAs ($F_{1,2736} = 21.39$, PRE = .01) than Arts and Sciences students.

A simple regression model in which SAT scores are used to predict freshman GPAs yields the following parameter estimates:

$$\widehat{GPA}_i = 1.018 + 0.031SAT_i \quad \text{SSE} = 1269.97$$

Unsurprisingly, students with higher combined SAT scores when they enter university end up with better freshman year GPAs ($F_{1,2738} = 270.97$, PRE = .09).

We have already seen that SAT scores are significantly related to the two categorical variables of college and gender. Therefore, when we estimate what we have to this point called the ANCOVA model, with SAT and $X_{1i}-X_{3i}$ as predictors of GPA, these predictors will be partially redundant. The resulting model is:

$$\widehat{GPA}_i = 1.018 + 0.031SAT_i \quad \text{SSE} = 1269.97$$

FIGURE 10.14 ANCOVA source table

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Model	187.16	4	46.79	105.89	.13
Gender (X_1)	18.57	1	18.57	41.99	.02
College (X_2)	0.10	1	0.10	0.22	.00
Gender x College (X_3)	0.46	1	0.46	1.04	.00
SAT	146.91	1	146.91	332.33	.11
Error	1208.49	2735	0.44		
Total	1395.65	2739			

And the source table in Figure 10.14 summarizes the statistical results.

Clearly, SAT remains a significant predictor of freshman GPA when controlling for gender, college, and their interaction. Additionally, of the categorical predictors, only gender remains significant once SAT is controlled. The effect of college, which was significant in the ANOVA model, no longer is. It seems that once students are equated in terms of SAT performance, there is no longer a college difference in GPAs. As with any ANCOVA model, the slopes for the categorical predictors are now informing us about the magnitude of differences among the adjusted GPA cell means. These are computed as:

$$\overline{GPA}'_k = \overline{GPA}_k - b_{SAT}(\overline{SAT}_k - \overline{SAT})$$

and their values are:

For male Engineering students:	2.69
For female Engineering students:	2.92
For male Arts and Sciences students:	2.68
For female Arts and Sciences students:	2.99

As we have already discussed, this model assumes that the “simple” slope of SAT does not vary across the four groups. That is, this model makes the homogeneity of regression assumption that we discussed earlier. It is of central interest in these data to ask whether the relationship between SAT and freshman GPA is the same across all four student groups. Perhaps SAT performance is more predictive of GPA in one college or the other. If so, then the greater diagnosticity of the test among the students in one college than the other could, perhaps, be taken into account at the time of admissions, weighting more heavily the SAT scores in making admissions decisions for some students than for others. Similarly, a gender difference in the diagnosticity of the test would indicate predictive bias, that is, that the test makes better predictions for one gender than the other, which would likely need to be addressed in other ways, for example by revising the test itself. In any case, if SAT scores are more diagnostic among some student groups than others, then it implies that the homogeneity of regression assumption would be violated in these data: we would need different SAT slopes for students who differ in their gender, their course of study, or the interaction of these two factors.

To test whether SAT is differentially related to GPA among the four different groups, we estimate a model that includes as predictors all products of SAT with the contrast-coded predictors that code the four categories of students: SAT_iX_{1i} , SAT_iX_{2i} , and SAT_iX_{3i} .

But this time, since we suspect that SAT performance will vary in its diagnosticity across the four groups, we will examine the individual contributions of each of these product predictors rather than simply conducting the omnibus test of whether they, as a set, increase the explanatory power of the model. The estimated Model A is:

$$\widehat{GPA}_i = 0.200 - 0.260X_{1i} + 0.676X_{2i} + 0.289X_{3i} + 0.044SAT_i + 0.007SAT_iX_{1i} \\ - 0.011SAT_iX_{2i} - 0.004SAT_iX_{3i}$$

$$SSE = 1204.18$$

Tests of the individual product predictors are summarized in the partial source table of Figure 10.15. Although the effects sizes are not large here, there is clearly evidence to suggest that the relationship between SAT and freshman GPA depends on whether one is an Engineering student or an Arts and Sciences student.

FIGURE 10.15 Portion of source table testing interactions between continuous and categorical predictors

Source	SS	df	MS	F	PRE
SAT x Gender ($SATX_1$)	1.45	1	1.45	3.28	.001
SAT x College ($SATX_2$)	3.92	1	3.92	8.88	.003
SAT x Gender x College ($SATX_3$)	0.64	1	0.64	1.46	.001
Error	1204.18	2732	0.44		

To interpret these differences, let us re-express the model in terms of the “simple” effect of SAT for each of the four groups:

$$\widehat{GPA}_i = (0.200 - 0.260X_{1i} + 0.676X_{2i} + 0.289X_{3i}) + (0.044 + 0.007X_{1i} - 0.011X_{2i} \\ - 0.004X_{3i})SAT_i$$

This re-expression can be used to generate the “simple” GPA : SAT prediction functions for each of the four cells by substituting the appropriate values for the contrast-coded predictors.

For male Engineering students:

$$\widehat{GPA}_i = (0.200 - 0.260(-1) + 0.676(-1) + 0.289(+1)) + (0.044 + 0.007(-1) \\ - 0.011(-1) - 0.004(+1))SAT_i \\ = 0.073 + 0.044SAT_i$$

For female Engineering students:

$$\widehat{GPA}_i = (0.200 - 0.260(+1) + 0.676(-1) + 0.289(-1)) + (0.044 + 0.007(+1) \\ - 0.011(-1) - 0.004(-1))SAT_i \\ = -1.025 + 0.066SAT_i$$

For male Arts and Sciences students:

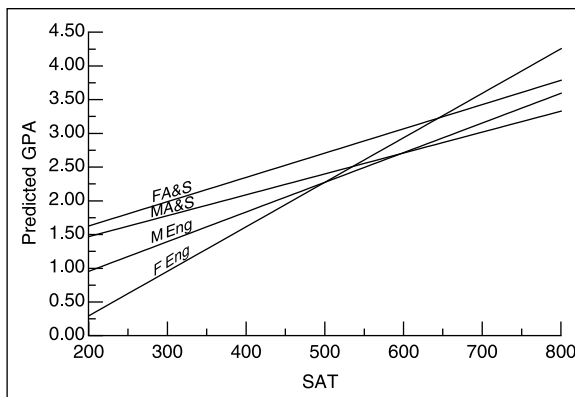
$$\begin{aligned}\widehat{GPA}_i &= (0.200 - 0.260(-1) + 0.676(+1) + 0.289(-1)) + (0.044 + 0.007(-1) \\ &\quad - 0.011(+1) - 0.004(-1))SAT_i \\ &= 0.847 + 0.031SAT_i\end{aligned}$$

For female Arts and Sciences students:

$$\begin{aligned}\widehat{GPA}_i &= (0.200 - 0.260(+1) + 0.676(+1) + 0.289(+1)) + (0.044 + 0.007(+1) \\ &\quad - 0.011(+1) - 0.004(+1))SAT_i \\ &= 0.906 + 0.036SAT_i\end{aligned}$$

In Figure 10.16 we have graphed these four “simple” relationships. As the differences in these graphed slopes make clear, the two Engineering groups have steeper slopes than do the two Arts and Sciences groups. This is the implication of the significant $SAT \times$ college interaction, with its coefficient, -0.011 , equaling half the difference between the average of the two “simple” slopes for the Arts and Sciences student groups and the average of the two “simple” slopes for the Engineering student groups. In other words, when averaging across the two gender groups, SAT performance is more diagnostic of freshman year GPA among Engineering students than it is among Arts and Sciences students.

FIGURE 10.16 Four-group simple relationships between SAT and GPA



A result that is perhaps somewhat surprising is that if we were to conduct four simple regressions, one for each of the four student groups and regressing freshman GPA on SAT scores for just the students in those four groups, we would get exactly the above parameter estimates that we have just calculated as the “simple” coefficients from this model that includes the $SAT \times$ interaction terms. That is, we split our sample into the four groups, and in each group separately we regress GPA on SAT, with the following results:

For male Engineering students:

$$\widehat{GPA}_i = 0.073 + 0.044SAT_i \quad SSE = 135.75$$

For female Engineering students:

$$\widehat{GPA}_i = -1.025 + 0.066SAT_i \quad SSE = 23.77$$

For male Arts and Sciences students:

$$\widehat{GPA}_i = 0.847 + 0.031SAT_i \quad SSE = 520.25$$

For female Arts and Sciences students:

$$\widehat{GPA}_i = 0.906 + 0.036SAT_i \quad SSE = 524.41$$

In a deep sense, then, the full model that includes all three contrast-coded predictors and their interactions with SAT is equivalent to four simple regression models, one from each of the four student groups, regressing GPA on SAT. Accordingly, if we add up the four sums of squared errors from these four within-group simple regression models, we get the sum of squared errors from the overall interactive model (i.e., $135.75 + 23.77 + 520.25 + 524.41 = 1204.18$).

While each of the within-group simple regressions provides us with a test of whether there is a significant association between SAT and GPA in that group (and all four are significant in this case), it is only by testing the various interaction terms in the full model, using all the data, that we are able to examine statistically whether the various simple slopes differ from each other across the four groups. And in this case, while all four simple slopes are significant, the two for the Engineering students are significantly larger than those from the Arts and Sciences groups.

In general, then, to examine whether two variables are related more strongly to each other in some groups than they are in others, we would encourage a test of whether the group \times continuous predictor interactions are significant in a model using the data from all groups together. This might be accompanied by separate model estimates in each of the groups, to test whether the group-specific simple slopes differ from zero. But to determine whether these simple slopes are different from each other, the test of the interactions in the full model, with all the data, is necessary.

Before leaving this final example, we re-estimate the full interactive model, this time centering or mean-deviating SAT (represented as SAT_{Ci}), both as a predictor and in computing the product predictors:

$$\begin{aligned} \widehat{GPA}_i = & 2.784 + 0.124X_{1i} + 0.045X_{2i} + 0.033X_{3i} + 0.044SAT_{Ci} + 0.007SAT_{Ci}X_{1i} \\ & - 0.011SAT_{Ci}X_{2i} - 0.004SAT_{Ci}X_{3i} \\ SSE = & 1204.18 \end{aligned}$$

We do this simply to make the parameter estimates associated with the contrast-coded predictors (X_{1i} – X_{3i}) more interpretable. Recall that the slopes associated with components of product predictors represent “simple” slopes of that component variable when and

only when the other component equals zero. In the model with SAT uncentered, the parameter estimates associated with the contrast-coded predictors thus inform us about “simple” differences among the students in the four cells of the design (defined by gender \times college) when and only when $SAT = 0$, which is an impossible value. To render these coefficients more interpretable, we therefore center the SAT variable. Thus, in this re-estimated model, the slope associated with X_{1i} estimates the gender difference in freshman GPA for students whose SAT scores are at the average for the sample. In a deep sense, of course, this model is identical to the one estimated prior to centering. Note, additionally, that we do not need to center the contrast-coded predictors. Although their means are not equal to zero (given the very unequal n values of the cells), the zero values do represent the mean of the gender and college categories.

SUMMARY

In many ways, this chapter marks the final point in the process of developing more complex models for our data in order to ask more complex, and perhaps more interesting, questions of those data. Throughout the development of these more complex models we have kept the same basic machinery to determine whether the increase in complexity is worthwhile as more parameters are added to the model. This machinery depends on comparisons of augmented and compact models in a manner that, by now, ought to be totally routine. While this machinery has remained constant across the chapters, our models have developed from the simplest one involving a single parameter (Chapter 4) to ones making predictions conditional on a single continuous variable (Chapters 5) to ones involving multiple continuous predictors, including product terms (Chapters 6 and 7), to models with categorical variables including products of those categorical variables (Chapters 8 and 9), and finally to models involving continuous and categorical predictor variables and their products (this chapter). This is as far as we wish to extend the complexity of the models we have considered. We believe that nearly every interesting substantive question that social science researchers might like to ask of their data can be answered by using the range of models that have been explored. This is not to say that the limits of model complexity have been reached. Rather, this is to suggest that the models we have considered are those that are most likely to be of use to the data analyst. Further, when other more complicated models seem appropriate, we hope that the reader by now is equipped to adapt the model comparison approach that we have used to these other more complicated situations.

So what remains to be done? The remaining chapters of the book are devoted to problems that are frequently encountered in data that violate various assumptions underlying the model comparison approach we have developed. In the next two chapters, we consider how our models and data analyses need to be modified when dealing with data for which we cannot assume independence of errors. In the final chapters, we focus on violations of the assumptions that residuals are normally distributed and come from a single population, having a single variance.

11

Repeated-Measures ANOVA

Models with Nonindependent Errors

The regression model that has been used in all of the preceding chapters is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{p-1} X_{1p-1} + \varepsilon_i$$

In this model, we have made no assumptions about the distributions of the predictor variables, the X_i . They can be continuously measured, they can be contrast-coded predictors used to represent categorical variables, or they can be products of variables. We have, however, made assumptions all along about the residuals, ε_i , in this model. Specifically, we have assumed three things about ε_i throughout the preceding chapters:

1. They are normally distributed.
2. They have constant variance.
3. They are independent of each other.

In this and subsequent chapters we consider necessary modifications to the regression model when these assumptions are violated. Specifically, in this chapter and the next we consider violations of the third assumption, that concerning the independence of ε_i , or, more colloquially, the independence of observations. Subsequent chapters consider violations of the normality and constant variance assumptions.

DEFINING NONINDEPENDENCE

Independence of observations is formally defined using conditional probabilities (Kenny & Judd, 1986). Two observations are nonindependent if the conditional probability of one of them having a particular value, given knowledge of the other one, is different from the unconditional probability. All this means is that if two observations are nonindependent, then knowing the value of one of them gives us information about the value of the other, allowing us to make a reasonable guess about the other's value. To illustrate, suppose we collected data from married couples, asking both spouses to rate, on a scale of 1–7, how satisfied they are with their marriage. Presumably, there would be some agreement between spouses on the quality of their marriage. The resulting data might look as follows:

<i>Couple</i>	<i>Male spouse</i>	<i>Female spouse</i>
1	1	1
2	4	3
3	6	7
4	5	6

The relative agreement between spouses in these data results in nonindependent observations. Observations within couples on average are more similar than observations between couples. Or, in terms of conditional probabilities, if we know that the score of one spouse is below the mean, then that gives us information about whether the score of the other spouse is above or below the mean. Since couples partially agree about the quality of their marriage in these data, if one spouse's score is below the mean, the other's is likely to be as well. Similarly, if one spouse's score is above the mean, the other's is likely to be as well. Observations that are linked, in the sense that they come from the same couple, are more similar on average than observations that are not linked, i.e., that come from different couples. This illustrates what we call *positive nonindependence*: greater similarity between linked observations than between nonlinked observations.

Negative nonindependence occurs when linked observations are more dissimilar on average than nonlinked observations. Consider hypothetical ratings from these same four couples about the percentage of housework each does:

<i>Couple</i>	<i>Male spouse</i>	<i>Female spouse</i>
1	30%	80%
2	40%	75%
3	70%	40%
4	20%	100%

If there were perfect agreement between spouses in these data, the sum of each couple's percentage estimates should be 100%. That is, as one spouse estimates that he or she does a higher percentage of housework, the other estimates a lower percentage. While there is not perfect agreement in these data, there is considerable nonindependence of observations within couples. If one spouse's estimated percentage is below the mean, the other's is quite likely to be above it. Observations that come from the same couple are on average more dissimilar than are observations from different couples. They are negatively nonindependent.

Nonindependence in data is likely to be found in two classic situations: grouped data and sequential data. The couples data that we have just used to define nonindependence illustrate nonindependence due to grouped data. Observations are grouped together in the sense that pairs of them come from the same couples. Another common grouping of observations occurs when each individual provides multiple observations: each person is measured more than once or provides more than a single response. It is then likely that observations taken from the same individual will be more or less similar on average than observations taken from different individuals. Thus, the observations are nonindependent because they are grouped within individuals. A further example of grouped data is typically encountered in social psychological research on small group

interaction. If different individuals within interacting groups are measured, it is exceedingly likely that nonindependence will be a problem.

Like grouped data, sequential data are also commonly encountered. Whenever lots of data are collected over time from one or more individuals, nonindependence due to sequence is likely to be encountered: Observations that are close together in the sequence or in time are quite likely to be more or less similar on average than observations that are further away in time. As a simple illustration, suppose we measured the number of pages read by a college student every day for 50 days. Because of exams, varying academic pressure, and other factors, we would expect cycles in the data: groups of days right before exams would be heavy reading periods; groups of days right after exams might show fewer pages read. In other words, there is likely to be more similarity in the number of pages read on adjacent days, on average, than on days that are nonadjacent. Thus, the observations would be nonindependent due to time sequence.

Since nonindependence is likely to be present in both grouped and sequential data, the usual regression model that we have employed needs to be modified in some way to deal with the fact that one of its underlying assumptions has been violated. For both kinds of data, the corrective strategy consists of removing the nonindependence from the data through some sort of transformation of those data. How this is done, however, typically depends on the origin of the nonindependence. Problems of nonindependence in sequential data are typically handled with somewhat different transformations from those employed in grouped data. We focus in this chapter on nonindependence problems in grouped data. Included here are data in which a few observations are gathered from each of a set of subjects over time. When many observations are gathered from each of the same set of subjects over time, then procedures, known as multilevel analyses for longitudinal data, are used (e.g., Hoffman, 2015). (An introduction to multilevel models, which have a number of applications, is provided in Chapter 12.) When many observations are gathered from a single subject over time, then procedures, known as *time-series analysis*, are recommended for dealing with nonindependence due to sequence.

Let us return to our married couple example to make one further distinction relevant to our discussion of nonindependence due to grouped data. Suppose in our example we are interested in knowing whether male spouses give reliably different ratings of marital satisfaction than do female spouses. If we worried not at all about the violation of the independence assumption in these data, we would construct a contrast code to represent gender of spouse, and regress our dependent variable (Y_j , each individual's rating) on this contrast-coded predictor. In this case, the groups that give rise to the nonindependence problem are said to be *crossed* with the independent variable of interest (i.e., gender). That is, each couple or group provides one and only one rating under each level of that independent variable: From within each couple there is one observation coded -1 on the contrast-code predictor and one observation coded $+1$. Another way of saying this is that the independent variable of interest varies *within* each couple, that is, gender varies *within* the groups that give rise to the nonindependence (i.e., couples).

Now suppose that the independent variable we are interested in is not gender of spouse but, rather, some characteristic of couples. Say, for instance, that we have two sets of couples, some newly married and others who have been married more than 10 years, and we wanted to know whether the two sets of couples give reliably different ratings of marital satisfaction. Again, if we did not worry about the nonindependence

in the data, we might code duration of marriage as short or long using a contrast code, and then regress ratings, one for each individual, on the resulting contrast-coded predictor. This time, however, the groups or couples that give rise to nonindependence are *nested* under the levels of the independent variable of interest. That is, each couple or group provides two ratings of marital satisfaction, both of which receive the same code on the predictor variable. If one member of the couple is coded as -1 , meaning that he or she comes from a relatively recent marriage, the other necessarily has the same code on the predictor variable. The independent variable of interest in this case is said to vary *between* couples.

In general, if groups of nonindependent observations are nested under some independent variable of interest, then observations from within a group all have the same value on that variable and the variable varies *between* groups. If groups of observations are crossed with the independent variable, then the observations within a group are under different levels of the variable and the variable varies *within* groups.

To clarify the *between–within* distinction further, consider a second example. Suppose that a researcher was interested in exploring differences in short-term memory as a function of stress levels. He intends to manipulate stress levels by testing subjects' memories in the presence of loud white noise, and in a control condition without any noise. He further decides that he will measure each subject's memory twice. These two observations from any given subject can either be crossed with the experimental treatment variable of interest (i.e., the presence or absence of white noise) or nested under it. In the case where the treatment is crossed with subjects, each subject's memory would be measured once with the white noise present and once with it not present. The treatment variable in this case varies *within subjects*. In the case where the two observations from each subject are nested under the treatment, each subject's memory would be measured twice, but now it would be measured both times either with the noise present or with it not present. In this case the treatment variable varies only *between subjects*.

For didactic purposes we will start by treating these two different cases separately. That is, we will first consider how nonindependence problems are overcome when groups of nonindependent observations are nested under the independent variable or factor whose effect is to be estimated. We will then consider solutions to nonindependence when groups are crossed with a factor of interest. In both cases, we will assume in this chapter that these independent variables, whether they vary between or within groups, are categorical and that each group has the same number of observations. With these restrictions we will thus be considering what have come to be known in the experimental design literature as repeated-measures ANOVA models. Our consideration will also include multiple independent variables that vary within or between groups.

In the next chapter, we will expand our consideration of nonindependence to include predictor variables that vary continuously rather than categorically and that may vary both within and between groups of nonindependent observations simultaneously. To illustrate such variables, imagine in our couples example that we were interested in the effects of age, rather than years of marriage, on marital satisfaction. The mean ages of couples are likely to vary between couples. But there is also likely to be some variation in age within couples—spouses are likely to differ from each other in age. This then would be an independent variable that varies both between and within couples. We will leave the consideration of such variables to the next chapter.

We remind the reader again that nonindependence can arise whenever observations are grouped, regardless of whether they come from the same person, the same small group, the same family, the same classroom, or whatever. While the ANOVA literature on repeated measures often assumes that observations are nonindependent because they come from the same person or research participant (thus, independent variables are known as between-*subject* or within-*subject* variables), nonindependence may also arise from other sorts of groupings.

NONINDEPENDENCE IN NESTED DESIGNS

Consider the following study. A researcher is interested in whether competition in small task groups affects satisfaction with the group product. Groups of three participants come into the laboratory. They are given some task to complete as a group. Half the groups are told that their performance will be compared with the performance of other groups and that they will be rewarded accordingly. The other half of the groups are not told this. Thus, half of the groups think they are competing with other groups, while the other half do not. After completing the task, all participants are asked to rate their satisfaction with the group product on a 1–9 scale, with higher numbers indicating greater satisfaction. The resulting data, with three groups (of three subjects each) in each of the two experimental conditions, are given in Figure 11.1. Unfortunately, our notation must become a bit more complicated, with each observation designated as Y_{hi} . The first subscript, h , refers to the individual observation (in this case person) within each group; it varies between 1 and s . The second subscript, i , refers to the group from which the observation comes, with a total of n groups (in this case six). Additionally, as before, we will use k to refer to condition or the level of the categorical independent variable under which groups are nested. There is a total of m such levels, which in this case equals 2.

Two issues are worth discussing about this notation. First, as will become clear shortly, it is no accident that we are now using i to refer to group rather than to the individual observations within each group. Second, the value of h that refers to the individual observation within each group is completely arbitrary, because the participants within each group are interchangeable. That is, there are no systematic factors that distinguish one person in a group from another.

We want to know whether the mean satisfaction level under the competition instructions is significantly different from the mean level under the no-competition instructions

FIGURE 11.1 Data for nested design

	<i>Instructions</i>					
	<i>Competition</i>			<i>No competition</i>		
	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>	<i>Group 6</i>
Y_{1i}	8	6	7	6	7	5
Y_{2i}	9	7	8	6	6	5
Y_{3i}	7	6	8	5	7	4
\hat{Y}_k	7.33			5.67		

(i.e., whether the two mean values of 7.33 and 5.67 differ). We will first proceed to analyze these data using the methods of Chapter 8, treating all 18 observations as independent. In fact, however, it is highly unlikely that the observations within the groups are independent of each other; there are probably some groups in which interactions proceed more smoothly and thus the mean satisfaction levels are higher than other groups. Once we have analyzed these data ignoring this possible nonindependence, we will conduct the appropriate analysis that corrects the possible assumption violation.

To examine the effect of the categorical independent variable (i.e., instructions), we code it using a contrast-coded predictor, X_{hi} , having values of +1 for competition and -1 for no competition. The inappropriate analysis, ignoring the potential nonindependence due to groups in these data, regresses Y_{hi} on X_{hi} across all 18 observations in the dataset. The resulting model is:

$$\hat{Y}_{hi} = 6.50 + .83X_{hi} \quad \text{SSE} = 16.00$$

As explained in earlier chapters, the intercept in this model equals the mean of the two condition means and the coefficient for the contrast-coded predictor equals half the difference between the mean of the competition condition and that of the no-competition condition. To examine whether these two means differ from each other, we want to compare this augmented model with a compact one in which the mean of all 18 observations is the predicted model, setting the parameter associated with the contrast-coded predictor equal to zero:

$$\hat{Y}_{hi} = 6.50 \quad \text{SSE} = 28.50$$

This comparison generates the source table of Figure 11.2. According to this analysis that inappropriately treats all 18 observations as independent, there is a significant difference between the two condition means, with higher satisfaction in the competition condition.

In order to demonstrate the nonindependence that the analysis above ignores, let us examine the residuals from Model A for these data. Recall that the independence assumption focuses on these residuals. In Figure 11.3 we give the residuals from this

FIGURE 11.2 Source table for analysis that ignores nonindependence

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Condition	.83	12.50	1	12.50	12.50	.44
Error		16.00	16	1.00		
Total		28.50	17			

FIGURE 11.3 Residuals from analysis of nested data

	<i>Instructions</i>					
	<i>Competition</i>			<i>No competition</i>		
	<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>	<i>Group 5</i>	<i>Group 6</i>
e_{1i}	0.67	-1.33	-0.33	0.33	1.33	-0.67
e_{2i}	1.67	-0.33	0.67	0.33	0.33	-0.67
e_{3i}	-0.33	-1.33	0.67	-0.67	1.33	-1.67

Model A for all 18 observations. These, of course, equal each observation minus the appropriate condition mean.

Notice that the residuals within each group tend to have the same sign. In Groups 2 and 6 all the residuals are negative, while in Group 5 they are all positive. In other words, if one observation within a group is lower than the respective condition mean, then the others in that same group tend to be lower as well. And the same tends to be true for observations that are above their respective condition mean. This is characteristic of positive dependence due to groups.

How do we handle the assumption violation? The general strategy is to do something to the data so that we are no longer dealing with nonindependent observations. In other words, we need to transform the data in some way so that the nonindependence is removed. Suppose, for instance, that instead of including three scores from each group in the analysis we only analyzed one randomly selected observation from each group. Clearly, in such an analysis nonindependence would cease to be a problem, since rather than having three linked observations from each group there would only be a single one, unlinked to every other observation in the analysis.

But clearly we would not want to throw out a randomly selected two-thirds of the data. Nevertheless, the general goal of having a single score, rather than three linked scores, to represent each group seems a viable way to handle the nonindependence problem. One obvious solution, to maintain all the information in the data and yet include only one score from each group in the analysis, is to compute the mean value for each group and to analyze the six resulting group means. In essence this is what we will do, except that rather than computing the mean of the three scores from each group we will calculate a composite score that we call W_{0i} , defined as:

$$W_{0i} = \frac{\sum_h \delta_h Y_{hi}}{\sqrt{\sum_h \delta_h^2}}$$

where all values of δ_h equal 1. We are then adding up the three δ scores from any given group and dividing this sum by the square root of three, rather than dividing by three (which would give the simple average of the three scores):

$$W_{0i} = \frac{Y_{1i} + Y_{2i} + Y_{3i}}{\sqrt{1^2 + 1^2 + 1^2}}$$

We compute (and analyze) this W_{0i} composite score rather than the simple average so that we can compare the sums of squares from the appropriate analysis that corrects the nonindependence problem to the inappropriate analysis that we have just conducted. If we analyzed the simple average we would get the appropriate test statistics (F and PRE), but the sums of squares would differ. We square the values of δ_h because at a later point we will compute and analyze other W_i scores where the values of δ_h will be contrast codes, with some of them negative and some positive. The simple sum of these δ_h values would thus always equal zero. And, again, we want to keep the sums of squares in the appropriate metric to help us understand the different sources of variation in our data. Thus, we will compute W_{0i} using the above formula with all values of δ_h set to 1.

For the first group, then, we compute W_{0i} as:

$$W_{0i} = \frac{8 + 9 + 7}{\sqrt{1^2 + 1^2 + 1^2}} = \frac{24}{\sqrt{3}} = 13.856$$

Values on this composite variable for all groups are given in Figure 11.4. With these W_{0i} scores, we have now combined all three scores from each group into a single score, and we can now ask whether the two experimental conditions, competition versus no competition, have different mean values on this composite W_{0i} score. Those mean values are given in the last row of Figure 11.4.

We can now use our regular procedures to answer this question, since we have eliminated the nonindependence problem in these data by combining the three non-independent observations from each group into a single composite one. Regressing W_{0i} on X_i (with values of +1 for competition and of -1 for no competition), we get the following model:

$$W_{0i} = 11.258 + 1.443X_i \quad \text{SSE} = 10.667$$

Since each group only contributes a single value of W_{0i} to this analysis, we have omitted the subscript h from this model. In other words, the group has become our unit of analysis. There are now six observations contributing to this analysis rather than the 18 in the earlier analysis that ignored the potential nonindependence problem.

The coefficients in this model can be interpreted just as all coefficients for models with contrast-coded predictors. The only complication is that they tell us about values of W_{0i} rather than about values of Y_{hi} . In other words, the value of b_0 in this model, 11.258, equals the mean of the two condition means of W_{0i} , and the value of b_1 , 1.443, equals half the difference between those two mean values. Accordingly, the same expression for the slope of a contrast-coded predictor (in terms of the condition means) that we gave in Chapter 8 continues to apply in the present case, except of course that it makes reference to the condition means of W_{0i} rather than the condition means of Y_{hi} :

$$b_1 = \frac{\sum_k \lambda_k \bar{W}_{0k}}{\sum_k \lambda_k^2}$$

FIGURE 11.4 Data for nested example with W_{0i} computed

	Instructions					
	Competition			No competition		
	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Y_{1i}	8	6	7	6	7	5
Y_{2i}	9	7	8	6	6	5
Y_{3i}	7	6	8	5	7	4
W_{0i}	13.856	10.970	13.279	9.815	11.547	8.083
W_{0k}		12.702			9.815	

Since the metric of W_{0i} is not the same as the metric of Y_{hi} , it makes sense to convert these parameter estimates back into the Y_{hi} metric. If we were to divide each W_{0i} by its denominator, $\sqrt{\sum_h \delta_{hs}^2}$, which in this case is $\sqrt{3}$, we get back to the Y_{hi} metric. We can do the same to the parameter estimates. Thus, if we take the value of b_0 from this W_{0i} model, i.e., 11.258, and divide it by $\sqrt{3}$ we get 6.50, which equals the mean of the two condition means in the original Y_{hi} metric. And if we take the value of b_1 , 1.443, and divide it by $\sqrt{3}$ the resulting value (.833) equals half the difference between the two \bar{Y}_k values.

Comparing this augmented model with a compact single-parameter model provides a test of whether X_i is a significant predictor of W_{0i} . This is equivalent to a test of whether the two values of \bar{W}_{0k} are significantly different from each other. And since these mean values are perfect functions of the mean values of \bar{Y}_k (dividing each by $\sqrt{3}$), this test is equivalent to asking whether the two values of \bar{Y}_k differ significantly. The resulting source table is presented in Figure 11.5. Notice that the value of the regression coefficient for X_i that we have included in this table is given in the metric of Y_{hi} rather than in the less meaningful metric of W_{0i} . We will do this throughout this chapter when we report regression coefficients in source tables. Notice also that, compared with the incorrect analysis that failed to remove the nonindependence in the data (the source table of Figure 11.2), the error sum of squares and degrees of freedom have been substantially affected by correcting the problem, while the sum of squares and degrees of freedom due to the independent variable have not been affected.¹

In essence, what we have done here is remove from the analysis, and from the sum of squares and degrees of freedom for error, the within-group variability. We are now only examining variability between groups rather than variability that is partly between and partly within groups. In other words, the original Y_{hi} scores varied within groups (across the individual observations) as well as between them. By collapsing observations within groups into a single W_{0i} score, we have eliminated any variability within groups from this analysis. All that is left is variability between groups.

Recognizing that we have divided up the total variation in the Y_{hi} scores into within- and between-group variation and then analyzed the effect of the independent variable, which varies between groups, we can now write out the full source table, integrating both between and within sources of variation in Y_{hi} . This source table is presented in Figure 11.6. Notice that in these data there is relatively little variability within groups. The sum of squares within groups equals 5.33, which is the difference between the total sum of squares in the original Y_{hi} scores (i.e., 28.50) and the total sum of squares in the W_{0i} scores that ignored within-group variability (i.e., 23.17).

We can derive this value for the within-group sum of squares by going back to the original Y_{hi} values and actually calculating the sum of the squared deviations of each individual Y_{hi} score from the mean of the group from which it was taken, \bar{Y}_i . Thus the three observations for the first group are 8, 9, and 7, with a mean of 8. The sum of the squared deviations of these three scores around their mean is thus 2 [i.e., $(7 - 8)^2 + (9 - 8)^2 + (8 - 8)^2$]. If we do this calculation for each of the six groups and then add up the resulting values, we get the total sum of squares within groups of 5.33. We do not have to actually calculate the sum of squares within groups in this manner, since we can get it by subtracting the total sum of squares in W_{0i} from the total sum of squares in Y_{hi} .

FIGURE 11.5 Analysis of nested data solving the nonindependence problem

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Condition	.83	12.50	1	12.50	4.68	.54
Error		10.67	4	2.67		
Total		23.17	5			

FIGURE 11.6 Full source table for nested example

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Condition	.83	12.50	1	12.50	4.68	.54
Error between groups		10.67	4	2.67		
Total between groups		23.17	5			
Total within groups		5.33	12	0.44		
Total		28.50	17			

Nevertheless, it is useful to know that the within-group sum of squares calculated in this way is equal to the sum of the squared deviations of all individual observations from their group means.

Comparing this analysis to the earlier incorrect analysis that failed to remove the within-group variability, the sum of squares for error has been reduced slightly, by the magnitude of the within-group variability (i.e., 5.33). The degrees of freedom for error, however, have been reduced substantially, from 16 to 4, since 12 of the original degrees of freedom in Y_{hi} were within groups rather than between. Since the error sums of squares has been reduced proportionally less than the error degrees of freedom, our correct analysis that removes the nonindependence from the data results in a substantially smaller value of F for testing the treatment effect. With positive nonindependence (i.e., linked observations more similar on average than nonlinked observations) and an independent variable that varies between groups (i.e., groups of linked observations are nested under its levels), it will always be the case that if one conducts an analysis that inappropriately ignores the nonindependence in the data the value of F will be too large, resulting in too many Type I errors. Should the nonindependence be negative, with relatively less similarity of observations within groups than between, then the F that results from the analysis that inappropriately ignores the nonindependence will always be too small, resulting in too many Type II errors.

If one examines the mean square error between groups and the mean square within groups, it is clear why nonindependence of observations within groups is a problem in these data. If the observations were independent of each other, then variation between observations within the six groups would be approximately the same as variation of the observations between groups. In such a case, these two sources of variation might be combined or pooled in a common error term, as is done in the analysis that ignores the nonindependence issue. And, in fact, the difference between the mean square error between groups and the mean square within groups provides a measure of the degree to which there is nonindependence due to groups. Since the mean square error within groups is substantially less than the mean square between groups, substantial positive

dependence due to groups is indicated: Observations on average within groups are not as variable as observations between groups on average.

A statistic known as the *intraclass correlation* provides an index of the degree of nonindependence due to grouped observations, based on the rationale just explained. That statistic is computed as follows:

$$\frac{MS_b - MS_w}{MS_b + MS_w (s - 1)}$$

where MS_b and MS_w are the mean squares between and within groups and s is the total number of observations within a group. This intraclass correlation varies between +1 and $-1/(s - 1)$, with positive values indicating that observations within groups are on average more similar than between them and negative values indicating the reverse. For the present data, the intraclass correlation equals:²

$$\frac{2.67 - .44}{2.67 + .44(2)} = .63$$

Thus these data show evidence of considerable positive nonindependence due to groups: Observations are more similar within groups than between, and the test of the between-group condition effect is substantially biased if this nonindependence is ignored (i.e., the incorrect F with 1 and 16 degrees of freedom equals 12.50, while the correct F with 1 and 4 degrees of freedom equals only 4.68). As this example illustrates, violations of the independence assumption can lead to substantial biases in the computed test statistics.³

We have illustrated this analysis with a very simple example. The same general strategy holds, however, whenever the groups that give rise to nonindependence are nested under the independent variable(s) of interest, regardless of the number of such variables and regardless of the numbers of their levels. Suppose, for instance, a study were conducted to examine impression formation between dyads under various conditions. Eighteen pairs of previously unacquainted participants are brought into the laboratory. They each fill out a questionnaire about themselves and then engage in a structured interaction. Each dyad is randomly assigned to one of six conditions, defined by two independent variables, one with two levels and one with three, that are crossed with each other. The two-level variable is whether they see each other's completed questionnaire prior to the interaction. The three-level variable is whether the interaction is competitive, cooperative, or no instructions. At the end of the interaction, they rate their impression of the other person in the dyad, thus yielding 36 observations. The design is diagrammed in Figure 11.7, with three dyads randomly assigned to each of the six

FIGURE 11.7 Design with two independent variables under which dyads are nested

Dyad	Other's questionnaire seen?																	
	Yes									No								
	Competitive			Cooperative			None			Competitive			Cooperative			None		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

cells, and two observations from each dyad.

In this case, the dyad is the grouping that likely induces nonindependence in the observations, and both independent variables vary between dyads. From the dependent variable, measured at the level of the individual observations, we would compute a W_{0i} score for each dyad, collapsing across the two scores from each. These W_{0i} scores would then be regressed on five

contrast-coded predictors (one coding the questionnaire factor, two coding the instructions factor, and two coding the interaction between questionnaire and instructions). Given that three dyads were randomly assigned to each of the six cells of this design, the rows of the resulting source table (including the appropriate degrees of freedom) would be as indicated in Figure 11.8 (with only the omnibus rows for the instructions factor and the interaction).

FIGURE 11.8 Rows of the source table for the analysis of Figure 11.7 design

Source	<i>df</i>
Between dyads	
Questionnaire seen	1
Instructions	2
Questionnaire × Instructions	2
Error between	12
Total between	17
Total within	18
Total	35

NONINDEPENDENCE IN CROSSED DESIGNS

In the introduction to this chapter, we said that nonindependence of observations can occur whenever observations are grouped in some way, be they from the same individual, the same family, the same small group, or the same classrooms. In the two illustrations that we used in explicating the analysis for nested designs, the grouping factor was first a small group of three participants and then a dyad. In many designs, individual subjects provide multiple observations in different experimental conditions over time. In this case, the individual subjects give rise to nonindependent observations because subjects are repeatedly measured in different conditions. And it is from this sort of design that “repeated-measures” ANOVA takes its name. Although we will use data examples that illustrate crossed designs with repeated measures on the same subjects, we again remind the reader that the procedures that we explicate would apply to any sort of grouping factor that induces nonindependence, when that grouping factor is crossed with one or more independent variables of interest.

Suppose we were interested in response facilitation effects, showing that the presence of another enhances performance. Participants are brought into the laboratory and asked to complete two sets of math puzzles. For one set, the experimenter is present during the task; for the other set, the experimenter is absent. Performance is measured on both puzzle sets, with higher numbers indicating better performance.⁴ It is hypothesized that performance should be better when the experimenter is present.

The hypothetical data are given in Figure 11.9. Here we have 16 individual observations, Y_{hi} , taken from eight different subjects in the two experimental conditions. The subscript i refers to the subject from whom the observation is taken; it varies from 1 to 8 (n). The subscript h continues to refer to which observation within a set of grouped observations (i.e., within a subject) we are looking at. Now, however, it also refers to

FIGURE 11.9 Data for crossed design

<i>Subject</i>	<i>Experimenter absent</i>	<i>Experimenter present</i>
1	7	8
2	5	5
3	6	6
4	7	9
5	8	8
6	7	7
7	5	6
8	6	8
\bar{Y}_h	6.375	7.125

the levels of the independent variable of interest, since that variable now varies within subjects. Thus, in these data h either equals 1 or 2, referring to the two levels of the independent variable.

As we did before, we will first test the effect of the experimental treatment variable while ignoring the likely nonindependence in these data that derives from repeated observations from the same subjects. We will then compare these results with those of the appropriate analyses.

Treating all 16 observations as if they were independent, we create a contrast-coded predictor, X_{hi} , with values of -1 if the experimenter is absent and $+1$ if present. The dependent variable, Y_{hi} , is then regressed on this contrast-coded predictor across all 16 observations:

$$\hat{Y}_{hi} = 6.75 + .375X_{hi} \quad \text{SSE} = 20.75$$

As always with a simple ANOVA using a contrast-coded predictor, the intercept equals the mean of the two condition means, and the regression coefficient for the contrast-coded predictor equals half their mean difference.

A comparison of this model with a compact single-parameter model gives the source table of Figure 11.10. We conclude from this erroneous analysis that the treatment effect is not significant since the obtained F does not exceed the critical value of F at $\alpha = .05$.

The residuals from this analysis for all of the observations are presented in Figure 11.11. Notice the pattern in these residuals. With two exceptions, if subjects have a negative residual in one treatment, they do in the other as well. Similarly, positive residuals are found in the same subjects across treatments. Since the mean of the residuals is zero, if one residual from a subject is below the mean, the other is exceedingly likely to be as well. Positive nonindependence due to the grouping of observations within subjects is thus illustrated by these residuals.

To eliminate this nonindependence problem, we need to do an analysis on a single summary score from each group of observations (i.e., each subject), just as we did in the case where groups of nonindependent observations were nested under treatment. Here, subjects are crossed with treatment, so rather than using a single summary score that is something like the average or sum of the within-subject observations, as we did in the nested case (i.e., W_{0i}), we want to compute something like a within-subject difference score, using what amount to within-subject contrast weights. In other words, if we want

FIGURE 11.10 Source table for analysis that ignores nonindependence

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
X_1	.375	2.25	1	2.25	1.52	.10
Error		20.75	14	1.48		
Total		23.00	15			

FIGURE 11.11 Residuals from analysis of crossed data

Subject	<i>Experimenter absent</i>	<i>Experimenter present</i>
1	0.625	0.875
2	-1.375	-2.125
3	-0.375	-1.125
4	0.625	1.875
5	1.625	0.875
6	0.625	-0.125
7	-1.375	-1.125
8	-0.375	0.875

to ask whether the difference between the two treatment means is significant, we might do this by computing for each subject a composite score that is essentially a difference score between the subject's performances in the two treatments, and then ask whether this within-subject difference score on average differs from zero.

Because we want to compare the results we obtain in this way with those that we obtained in the inappropriate analysis that ignored the nonindependence problem, rather than analyzing a simple difference score for each subject, we will compute a W score for each subject, using the same formula that we used earlier:

$$W_i = \frac{\sum_h \delta_h Y_{hi}}{\sqrt{\sum_h \delta_h^2}}$$

But this time we will set the values of δ_h equal to +1 if the observation is taken with the experimenter present and to -1 if the observation is taken with the experimenter absent. These are the same contrast values used to construct the contrast-coded predictor in the analysis just described that ignored the nonindependence in these data. But now, the comparison is within each subject and our composite score reflects the difference within each subject. To indicate that this within-subject composite score is a contrast or difference composite, we will refer to it as W_{1i} rather than the W_{0i} that was used in the nested example where the composite score was like an average of all of the within-subject (or within-group) observations, with all values of δ_h set to 1.

Setting the values of δ_h to -1 if the experimenter is absent and to +1 if the experimenter is present, the computation of W_{1i} for each subject is accomplished as follows:

$$W_{1i} = \frac{(-1)Y_{1i} + (+1)Y_{2i}}{\sqrt{(-1)^2 + (+1)^2}}$$

FIGURE 11.12 Values of W_{1i} : crossed design

Subject	Y_{1i}	Y_{2i}	$\sum_h \delta_h Y_{hi}$	W_{1i}
1	7	8	1	0.707
2	5	5	0	0.0
3	6	6	0	0.0
4	7	9	2	1.414
5	8	8	0	0.0
6	7	7	0	0.0
7	5	6	1	0.707
8	6	8	2	1.414
Mean	6.375	7.125	0.75	0.530

where Y_{1i} is the i th subject’s score when the experimenter is absent and Y_{2i} is the observation when the experimenter is present.

In Figure 11.12 we have calculated W_{1i} for each subject. In the column labeled $\sum_h \delta_h Y_{hi}$, we have calculated the numerator of W_{1i} . This numerator amounts to the simple difference between Y_{2i} and Y_{1i} for each subject. Since the denominator of W_{1i} is a constant for all subjects, W_{1i} will be perfectly correlated with the simple within-subject difference scores. It can also equal zero only when the simple difference score equals zero.

If the treatment difference is significant, then the average of these W_{1i} scores across subjects should be significantly different from zero, because the average of subjects’ difference scores equals the difference between the treatment averages. In other words, the mean value of W_{1i} equals $(\bar{Y}_2 - \bar{Y}_1)/\sqrt{2}$, and hence $(\bar{Y}_2 - \bar{Y}_1)$ will equal zero only when the mean value of W_{1i} equals zero.

We will use the simplest single-parameter model of Chapter 4 to test whether the mean of the W_{1i} scores differs from zero. The augmented and compact models are:

MODEL A: $W_{1i} = \beta_0 + \varepsilon_i$

MODEL C: $W_{1i} = 0 + \varepsilon_i$

A comparison of these two models is a test of whether the mean W_{1i} score differs from zero. Equivalently, this provides a test of whether \bar{Y}_h with the experimenter present differs from \bar{Y}_h with the experimenter absent. Note that because each subject appears only once in this analysis, the dependence problem (due to repeated observations taken from the same subject) is no longer problematic.

Compact and augmented models are estimated as:

MODEL A: $\hat{W}_{1i} = 0.53 \quad \text{SSE} = 2.75$

MODEL C: $\hat{W}_{1i} = 0 \quad \text{SSE} = 5.00$

Since we have a total of eight W_{1i} scores, one from each subject, the degrees of freedom for error equal 7 for the augmented model and 8 for the compact one. Comparing these two models gives PRE and F values of:

$$\text{PRE} = \frac{5.00 - 2.75}{5.00} = \frac{2.25}{5.00} = .45$$

$$F_{1,7} = \frac{.45}{(1 - .45)/7} = \frac{2.25}{2.75/7} = 5.73$$

The sum of squares reduced, 2.25, can of course be derived from the expression:

$$\sum_i (\hat{W}_{1iA} - \hat{W}_{1iC})^2$$

where \hat{W}_{1iA} and \hat{W}_{1iC} are the predicted values from the augmented and compact models. Since each of these models makes the same prediction for all subjects, the sum of squares reduced can be equivalently expressed as:

$$n(\hat{W}_{1iA} - \hat{W}_{1iC})^2$$

And since \hat{W}_{1iA} equals \bar{W}_1 and \hat{W}_{1iC} equals zero, this reduces to:

$$n(\hat{W}_{1iA})^2 = n(\bar{W}_1)^2 = 8(0.53)^2 = 2.25$$

Before presenting all of this in a source table, let us once again interpret the value of the estimated parameter in the augmented model. As we saw in the nested design, if we take the estimated parameters and divide them by the denominator of W_i , $\sqrt{\sum_h \delta_{hi}^2}$, we get a parameter estimate in the same metric as Y_{hi} rather than in the W_i metric. In this case, the denominator of W_{1i} equals $\sqrt{2}$. If we divide the single-parameter estimate in the augmented model (0.530, which is also \bar{W}_1) by $\sqrt{2}$, we get 0.375. This value is the same as the value of b_1 in the model where the nonindependence in the data was inappropriately ignored. Again, it equals half the difference between the mean Y_{hi} when the experimenter is present and the mean Y_{hi} when the experimenter is absent.

The source table that summarizes our test is presented in Figure 11.13. Notice that the sum of squares for the treatment difference in this source table is identical to what it was before in the source table of Figure 11.10, which resulted when we ignored the nonindependence in the data.⁵

As in the nested design, eliminating the nonindependence in the data has dramatically affected the sum of squared errors and the degrees of freedom for error. Instead of having a sum of squared errors of 20.75, as we had in the inappropriate analysis of Figure 11.10, we now have a sum of squared errors of 2.75. The degrees of freedom for error have been reduced from 14 to 7. Rather than counting each value of Y_{hi} as an observation, we now have only eight observations, one from each subject, hence the degrees of freedom have been cut in half.

Conceptually, the sum of squared errors in this source table can be thought of as the interaction of treatment with subjects. If all of the between-treatment differences were the same for all subjects, then there would be no variation in the values of W_{1i} around their mean, and hence the augmented model that predicts the mean value of W_{1i} for each case would fit the data without error. This amounts to saying that the treatment effect would not vary from subject to subject. The sum of squared errors from the augmented model derives from the fact that the magnitude of the difference between the treatment conditions varies from subject to subject. Hence, the error term in this source table can be thought of as the sum of squares associated with the treatment \times subject interaction.

FIGURE 11.13 Preliminary crossed source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Treatment	.375	2.25	1	2.25	5.73	.45
Error		2.75	7	0.39		
Total		5.00	8			

There is still considerable variation in the Y_{hi} scores that is not represented in this source table. In the source table of Figure 11.10 where nonindependence was ignored, we saw that the total sum of squares of Y_{hi} equals 23.00. The difference between this value and the total sum of squares for W_{1i} in the source table that we have just presented, 5.00, equals the total sum of squares between subjects. In other words, we have once again divided up the sum of squares of Y_{hi} into two portions: one between subjects and one within subjects. Since the treatment variable in this crossed example varies within subjects rather than between them, we have looked only at the within-subject variation in the data to estimate and test the treatment effect. We have entirely omitted the between-subject variation from this analysis. In the earlier section of this chapter, we considered the case where groups of nonindependent observations were nested under the treatment variable or, equivalently, the treatment variable varied between them. There, we conducted our analysis by only examining between-group variation in the data. Here, we have conducted our analysis only by examining the within-subject variation (i.e., within linked pairs of observations) in the data, since the treatment variable varies within subjects. What we now need to do, as we did in the earlier section, is to report both the between- and within-subject variation in the values of Y_{hi} in a single source table.

To calculate the between-subject sum of squares in these data, let us use the same procedure we did earlier when considering a treatment variable that varied between groups of linked observations. There, to capture only the between-group variation, we computed values of W_{0i} for each group, assigning α_n weights of 1 to all observations within groups. Let us do the same thing here, where our groups of observations come from the same subject. The resulting W_{0i} values are given in Figure 11.14. The total sum of squares of these W_{0i} values around their mean equals 18, the difference between the total sum of squares of Y_{hi} and the total within-subject sum of squares. This sum of squares total for W_{0i} equals the total between-subject sum of squares, just as it did in the earlier design where groups of linked observations were nested under treatment conditions.

FIGURE 11.14 Values of W_{0i}

Subject	Y_{1i}	Y_{2i}	$\sum_n \delta_n Y_{hi}$	W_{0i}
1	7	8	15	10.61
2	5	5	10	7.07
3	6	6	12	8.49
4	7	9	16	11.31
5	8	8	16	11.31
6	7	7	14	9.90
7	5	6	11	7.78
8	6	8	14	9.90
Mean	6.375	7.125	13.50	9.55

FIGURE 11.15 Full crossed source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Total between subjects		18.00	7	2.57		
Treatment	.375	2.25	1	2.25	5.73	.45
Error within		2.75	7	0.39		
Total within subjects		5.00	8	0.63		
Total		23.00	15			

We can now write out the full source table for these data, incorporating variation both within and between subjects, and including the sum of squares due to treatment as a component of the within-subject variation. This source table is presented in Figure 11.15.

Observations in this example are linked because they come from the same subject. This induces positive dependencies in the data: observations within subjects are more similar on average than are observations between them. This is revealed by a comparison of the mean squares within and between subjects from this full source table, using the formula given earlier for the intraclass correlation:

$$\frac{MS_b - MS_w}{MS_b + MS_w (s - 1)} = \frac{2.57 - 0.39}{2.57 + 0.39(1)} = .74$$

The positive nonindependence in these data, due to observations being linked within subjects, means that the error term was too large when the nonindependence was inappropriately ignored. That is, the *F* was considerably smaller under the incorrect analysis of Figure 11.10 than in the analysis of Figure 11.15, which more correctly uses only the within-subject variation as the error term. If there were negative dependencies among linked observations, the within-group error term would be larger than the between-group error term and the situation would be reversed.

SUMMARY OF ANALYSES

Before dealing with more complicated designs, with multiple independent variables, let us summarize what we have discussed so far in this chapter. First, we have defined nonindependence as links between observations that are grouped in some manner, which makes them either more similar (positive dependence) or less similar (negative dependence) than nonlinked observations. The groupings that give rise to nonindependence can derive from many sources: observations may come from the same subject, the same family, the same small group, the same class, and so forth. What matters is simply that they are grouped and that, as a result of the groupings, the variation between observations is not homogeneous. Observations are either less variable within the groups than between them (positive dependence) or more variable within the groups than between them (negative dependence).

Second, the independent variables that we have considered to this point vary either within sets of grouped or linked observations or between them. In the former case, the groups that give rise to the nonindependence are said to be crossed with the independent

FIGURE 11.16 Direction of bias in F if nonindependence is ignored

	<i>Nonindependence</i>	
	<i>Positive</i>	<i>Negative</i>
Groups nested under independent variable	F too large	F too small
Groups crossed with independent variable	F too small	F too large

variable of interest. In the latter case, the groups are said to be nested under the independent variable of interest.

When the independent variable of interest varies between groups of linked observations (i.e., observations are nested under the levels of the independent variable), one should examine its effects by examining only the between-group variation in the data, regressing W_{0i} (where i indicates group) on a contrast-coded predictor(s) that codes the between-group independent variable. When the independent variable of interest varies within groups of linked observations (i.e., observations are crossed with levels of the independent variable), one should examine its effects by examining only the within-group variation in the data. Here one computes a within-group contrast or difference score (i.e., W_{1i}) and asks whether its mean differs from zero across groups.

If dependencies in grouped data are ignored in the analysis, the bias in the resulting F statistics is a function both of whether the nonindependence is positive or negative and whether groups are crossed with the independent variable or nested under it. The table in Figure 11.16 summarizes the resulting bias. As we have seen, the potential biases can be far from trivial. One should attend conscientiously to violations of the independence assumption.

MULTIPLE CROSSED INDEPENDENT VARIABLES

In the example discussed above (nonindependence in crossed designs) there was only a single independent variable that varied within the groups of nonindependent observations (i.e., within subjects). The procedure that we discussed there for handling nonindependence (testing whether a within-subject difference contrast on average differs from zero) generalizes readily to multiple independent variables that are crossed with groups of observations.

Consider a design from cognitive psychology in which the effects of study time and word abstraction on memory are examined. There are six lists of 20 words to which each subject is exposed, with three of the lists consisting of abstract nouns (e.g., hope) and three consisting of concrete nouns (e.g., chair). This independent variable is crossed with the amount of time that subjects are given to study each list: 1 minute, 2 minutes, and 3 minutes. Subjects are shown each list for the appropriate amount of time and, after a short filler task, asked to recall as many words on the list as they can. They then repeat this procedure for all six lists, counterbalancing order. The dependent variable is the number of words they recall from each list. The resulting hypothetical data for five subjects are given in Figure 11.17, and the last row also gives the means of the observations in each of the six conditions.

FIGURE 11.17 Data for a 3 x 2 within-subject design

Subject	Study time					
	1 minute		2 minutes		3 minutes	
	Abstract	Concrete	Abstract	Concrete	Abstract	Concrete
1	10	13	12	14	16	17
2	8	12	9	12	11	13
3	12	13	14	14	16	16
4	15	17	16	17	19	20
5	12	13	15	16	16	17
Mean	11.4	13.6	13.2	14.6	15.6	16.6

If the 30 observations (five subjects \times six memory scores from each) were independent of each other, then the dependent variable would be regressed on five contrast code predictors, two to code study time (1 minute vs. 2 vs. 3), one to code list type (abstract versus concrete) and two to code their interactions. But since there are certainly stable individual differences in the ability to memorize these lists of words (notice, for instance, that Subject 4 consistently does better than the other subjects), there are undoubtedly positive dependencies in these memory scores. Accordingly, we must estimate and test these various contrasts within groups of linked observations (i.e., within subjects) rather than between them. And to do this, we proceed exactly as we did in the last example, asking whether within-subject contrasts or differences differ from zero. The only difference is that in this design, with six conditions, there are five within-subject contrasts that will need to be estimated and tested.

In Figure 11.18, we give the contrast codes (δ_h) that will be used to test the within-subject effects. Since study time clearly can be ordered, the first and second of these codes (W_{1i} and W_{2i}) code the linear and quadratic effects of study time; W_{3i} codes list type; W_{4i} codes the linear study time \times list type interaction; and W_{5i} codes the quadratic study time \times list type interaction. Using these contrast codes, we compute for each subject the five within-subject contrast scores, using the previously given formula:

$$W_i = \frac{\sum_h \delta_h Y_{hi}}{\sqrt{\sum_h \delta_h^2}}$$

FIGURE 11.18 Within-subject contrast codes (δ_h)

Subject	Study time					
	1 minute		2 minutes		3 minutes	
	Abstract	Concrete	Abstract	Concrete	Abstract	Concrete
W_{1i}	-1	-1	0	0	+1	+1
W_{2i}	-1	-1	+2	+2	-1	-1
W_{3i}	-1	+1	-1	+1	-1	+1
W_{4i}	+1	-1	0	0	-1	+1
W_{5i}	+1	-1	-2	+2	+1	-1

FIGURE 11.19 Subject scores on within-subject W_i

Subject	W_{1i}	W_{2i}	W_{3i}	W_{4i}	W_{5i}	W_{0i}
1	5.00	-1.15	2.45	-1.00	0.00	33.48
2	2.00	-0.58	3.67	-1.00	0.00	26.54
3	3.50	-0.29	0.41	-0.50	-0.29	34.70
4	3.50	-1.44	1.63	-0.50	-0.29	42.46
5	4.00	1.15	1.22	0.00	0.00	36.33
Mean	3.60	-0.46	1.88	-0.60	-0.12	34.70

The resulting values of $W_{1i}-W_{5i}$ for each of the five subjects, and their means, are given in Figure 11.19. We have also included in this table each subject's score on W_{0i} . As always, in computing W_{0i} we set all values of δ_{hi} equal to +1; accordingly, it captures the between-subject variation in the data.

Each of the within-subject contrast scores, $W_{1i}-W_{5i}$, codes a within-subject difference in memory scores as a function of the experimental conditions. And for each, we want to ask a Model A/Model C question to determine whether the mean within-subject difference coded by that contrast differs significantly from zero. Our estimated models and their associated sums of squares are given below:

$$\text{MODEL A: } \hat{W}_{1i} = 3.60 \quad \text{SSE} = 4.70$$

$$\hat{W}_{2i} = -0.46 \quad \text{SSE} = 4.10$$

$$\hat{W}_{3i} = 1.88 \quad \text{SSE} = 6.20$$

$$\hat{W}_{4i} = -0.60 \quad \text{SSE} = 0.70$$

$$\hat{W}_{5i} = -0.12 \quad \text{SSE} = 0.10$$

$$\text{MODEL C: } \hat{W}_{1i} = 0 \quad \text{SSE} = 69.50$$

$$\hat{W}_{2i} = 0 \quad \text{SSE} = 5.17$$

$$\hat{W}_{3i} = 0 \quad \text{SSE} = 23.83$$

$$\hat{W}_{4i} = 0 \quad \text{SSE} = 2.50$$

$$\hat{W}_{5i} = 0 \quad \text{SSE} = 0.17$$

In each case, Model A estimates the mean within-subject contrast and Model C fixes that mean within-subject contrast at zero. The sums of squared errors for Model A equal the sum of the squared W_i scores around their mean, the sums of squared errors for Model C equal the sum of the squared W_i scores around zero, and the SSRs in each case equal $n(\bar{W}_i)^2$. To illustrate these SSRs, the mean of W_{1i} is 3.60, which, when squared and multiplied by 5 (n) equals 64.80. This is the SSR for the W_{1i} model comparison, taking the difference between SSE(C) (69.50) and SSE(A) (4.70).

Notice that we are doing a series of Model A/Model C comparisons here, one for each within-subject contrast that we are interested in. And in any one of these model comparisons the number of observations is the number of subjects (5) and hence the observations in each of these model comparisons are independent of each other.

In addition to these within-subject contrasts (and associated sums of squares), there is variation between subjects in these scores and this variation is captured by W_{0i} , which has a sum of squares around its mean of 131.00.

From the above, we summarize the variation in the 30 Y_{hi} scores (and the five tests of the within-subject mean differences) with the source table in Figure 11.20. The values of the regression coefficients (b) that we have included in this table come from the means for each of the W_i , putting those estimates back into the metric of Y_{hi} by dividing each W_i mean by $\sqrt{\sum_h \delta_h^2}$. Thus, for instance, the mean value of W_{1i} , 3.60, is the parameter estimate from its Model A. If we divide that by $\sqrt{\sum_h \delta_h^2}$, which for W_{1i} equals $\sqrt{4}$, the resulting value is 1.80. In terms of \bar{Y}_h , this value is half the difference between the mean number of words recalled in the two 3-minute conditions and the mean number recalled in the two 1-minute conditions, i.e., $1.80 = ((15.6 + 16.6)/2 - (11.4 + 13.6)/2)/2$. Thus, once these parameter estimates are returned to the Y_{hi} metric, they can be interpreted in the same way that we interpreted the parameter estimates associated with contrast-coded predictors in Chapters 8 and 9.

The results of this analysis suggest that memory for the word lists improves with longer study time (the significant linear effect of time), that memory for words from the abstract lists is on average lower than memory for the words from the concrete lists, and that the concrete versus abstract difference is smaller at longer study times. These derive from the signs of the respective parameter estimates and knowledge of the contrast weights that were used to derive the various W_i . What is perhaps surprising here is that the effects are quite large in spite of the relatively few numbers of subjects from whom data were collected. In part this is due to the fact that we made up these data so that effects would be significant. But in part this is also due to the fact that we have employed a within-subject design with considerable positive dependencies in the data due to subjects. Had we employed a design in which each observation came from a different subject (thus 30 subjects in total), permitting us to treat them as independent, the substantial between-subject variation in these data would have been included in the error term for the test of the between-subject effects of interest. With the present design, however, that substantial between-subject variation in the data is controlled or removed from the error term(s) used to test each effect. It will frequently be the case that within-subject designs enjoy power benefits compared to between-subject designs for exactly this reason (i.e., substantial positive nonindependence due to subjects).

Up to this point, every source table that we have presented has relied on a single Model A, which then is compared with numerous Model Cs to test the various effects of interest. As a result, in every source table up to this point there has been only a single error term, involving the sum of squared errors and mean squared error of Model A. Now, however, when testing contrasts that vary within groups of linked observations, as in this within-subject design, each individual contrast, each W_i , is tested with its own unique Models A and C, resulting in unique error terms for each single-degree-of-freedom contrast. Thus, in this situation, there is an additional reason to refrain from omnibus multiple-degree-of-freedom tests. Up to this chapter, we expressed our dislike of such tests because if Model C is rejected in favor of Model A we are unable to specify what it is about Model A that makes it preferable to Model C. Now, if each single-degree-of-freedom contrast is tested with its own unique Models A and C, it makes little sense to conduct pooled tests, since doing so might force us to pool or combine potentially rather different error terms.

FIGURE 11.20 Source table for data in Figure 11.17

<i>Source</i>	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Between subjects		131.00	4	32.75		
Within subjects						
Linear time	1.80	64.80	1	64.80	55.15	.93
Linear time error		4.70	4	1.17		
Quadratic time	-.13	1.07	1	1.07	1.04	.21
Quadratic time error		4.10	4	1.02		
List type	.77	17.63	1	17.63	11.38	.74
List type error		6.20	4	1.55		
Linear time x List type	-.30	1.80	1	1.80	10.29	.72
Linear time x List type error		0.70	4	0.17		
Quadratic time x List type	-.03	0.07	1	0.07	2.67	.41
Quadratic time x List type error		0.10	4	0.03		
Total within		101.17	25			
Total		232.17	29			

To illustrate the problem, in the present design one might be tempted to pool together the two-degree-of-freedom effects for time and the two interaction degrees of freedom to test omnibus null hypotheses about time differences and time – list type interactions. In Figure 11.21 we have done just this, reporting only the pooled or combined two-degree-of-freedom tests for the time effect and the time \times list type interaction. These rows were computed by adding up the sums of squares and degrees of freedom from the respective rows of the source table in Figure 11.20.

The difficulty with these pooled multiple-degree-of-freedom tests is that they cannot be expressed as model comparisons. That is, there is no Model A/Model C comparison that we can write out that represents these tests; rather, they emerge from collapsing or averaging across two different Model A/Model C comparisons. And those two may have very different MSEs that are inappropriately averaged. Consider, for instance, the two error terms for the two single-degree-of-freedom interaction comparisons that are reported in the full source table of Figure 11.20. The MSE for the Linear time – List type interaction was 0.17. The MSE for the Quadratic time – List type interaction was 0.03. Clearly the magnitudes of these errors are very different, one being more than five times the size of the other, emerging from two very different Model As. Yet, in the pooled source table of Figure 11.21 we have ignored this distinction and simply averaged them together.

In ANOVA textbooks that deal with repeated-measures designs (and experimental factors that vary within subjects), an assumption often called *sphericity* or *compound symmetry* is deemed to be very important. This assumption is equivalent to the assumption that when pooling error terms to conduct an omnibus test the various error terms (from different Model As) are homogeneous. By relying on single-degree-of-freedom within-subject contrasts, each tested with its own unique Model A and Model C and hence its own unique error term, this assumption is not necessary. Our regression-based approach forces the researcher to examine single-degree-of-freedom within-subject contrasts. This is not the case if repeated-measures designs are analyzed using widely available programs for repeated-measures ANOVA, which routinely provide omnibus tests that pool potentially rather different error terms.

FIGURE 11.21 Pooled source table for data in Figure 11.17

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Between subjects	131.00	4	32.75		
Within subjects					
Time	65.87	2	32.93	29.94	.88
Pooled time error	8.80	8	1.10		
List type	17.63	1	17.63	11.38	.74
List type error	6.20	4	1.55		
Time x List type	1.87	2	0.93	9.35	.70
Pooled time x List type error	0.80	8	0.10		
Total within	101.17	25			
Total	232.17				

NONINDEPENDENCE IN MIXED DESIGNS

We have now dealt with multiple independent variables that vary either between or within groups of nonindependent observations. But so far we have not discussed cases in which the groups of linked observations are crossed with one or more independent variables while simultaneously being nested under levels of others. Such designs are widely used in some behavioral disciplines and are commonly known as mixed designs, with some independent variables varying within and some between groups of linked observations.

To return to an example in which observations are linked because they come from the same groups, rather than the same subject, consider a research design in which heterosexual married couples are asked to rate their marital satisfaction on a 1–9 point scale, with higher numbers indicating greater satisfaction. Some of these couples have been married 15 years and some of them 30 years. And among each group some had two or more children and others had none. Among the couples with children, those married for 15 years have young children at home and those married for 30 years have grown children who live elsewhere. There are thus three independent variables of potential interest in this design: the gender of the spouse, the length of the marriage, and whether there are children. The first of these varies within couple. The other two vary between them. Observations within couples are exceedingly likely to show positive dependencies: If one member of the couple reports high satisfaction, for instance, it is quite likely that the other member of the couple will as well. The hypothetical data from 16 couples are given in Figure 11.22. As this figure makes clear, there are four couples in each of the four cells defined by length of marriage and the presence or absence of children.

If we were to incorrectly analyze these data as independent observations, we would regress the 32 Y observations on seven contrast-coded predictors, to code the eight cells of the design. One of these would capture gender, one would represent length of marriage, one would represent children or no children, three product predictors would capture the three two-way interactions, and one triple-product predictor would capture the three-way interaction.

Because of the dependencies induced by two observations collected from each couple, one of the independent variables of interest varies within couples while the other two factors vary between couples. In the analyses that examine mean differences as a function of these variables and their interactions, each couple should appear in each Model A/Model C comparison only once, to assure independence of observations.

As in the case when independent variables varied between groups of linked observations, we will analyze the effects of length of marriage, the presence or absence of children, and their interaction using only the between-couple variation in the data. Thus, for these three effects we will construct between-couple contrast-coded predictors: X_{1i} codes length (+1 if 30 years; -1 if 15 years); X_{2i} codes whether or not the couple has children (+1 if yes; -1 if no); and X_{3i} is their product, which will capture the interaction. And just as we did in the first section of the chapter, when we examined the effects of independent variables that varied between groups of linked observations, we will examine the effects of these between-couple contrast-coded predictors on the between-couple variation in the data, creating a W_{0i} with both values of δ_h equal to +1, thus in essence averaging across the satisfaction scores of the male and female spouses in each couple.

Since gender of spouse varies within couples rather than between, we must follow the strategy developed in the earlier sections of the chapter devoted to the analysis of independent variables that are crossed with groups of linked observations. Namely, we will calculate a W_{1i} score for each couple that captures the gender of spouse difference, using a δ_h value of +1 for the female satisfaction rating and -1 for the male satisfaction rating. To determine whether the male and female mean satisfaction ratings are significantly different from zero, we will compare a Model A that estimates the mean of W_{1i} to a Model C that fixes it at zero, just as we did in the case of crossed independent variables earlier.

The approach just outlined permits us to examine the within-couple effects (captured by W_{1i} , the within-couple contrast) and the between-couple effects (both the length of marriage and children main effects and their interaction, estimated by examining whether the three between-couple contrast-coded predictors, X_{1i} - X_{3i} , are useful predictors of W_{0i}).

Missing from this outlined analysis, however, are tests of the interactions between independent variables that vary within and those that vary between couples. For instance, it may be the case that there is a bigger difference in satisfaction ratings as a function of length of marriage for male than for female spouses. This is an entirely appropriate question that we might ask of our data, since gender of spouse and length of marriage

FIGURE 11.22 Data for hypothetical mixed example

No children						Children					
15 Years			30 Years			15 Years			30 Years		
Couple	M	F	Couple	M	F	Couple	M	F	Couple	M	F
1	5	6	5	6	5	9	5	5	13	7	6
2	7	6	6	6	5	10	4	5	14	8	7
3	7	6	7	6	4	11	5	4	15	7	7
4	4	5	8	8	6	12	6	5	16	8	7
\bar{Y}	5.75	5.75		6.50	5.00		5.00	4.75		7.50	6.75

are crossed with each other (permitting a test of their interaction) even though one factor (gender) varies within couples and one factor (marriage length) varies between them.

Since every interaction can be equivalently expressed in more than one way, the above question about whether the length of marriage difference in satisfaction is different for male and female spouses is equivalent to asking whether the gender of spouse difference is larger or smaller for couples married for 30 years than for those married 15 years. Rephrasing this interaction question in this way gives a hint about how we might examine it, while overcoming the nonindependence problem inherent in these data. That is, the within-couple contrast code, W_{1i} , captures the gender difference in satisfaction. To ask whether this within-couple gender difference depends on the length of the marriage, we regress it on the between-couple contrast-coded predictor that codes marriage length, X_{1i} , and test whether the slope of this predictor differs from zero. If it does, then that would mean that the magnitude of the gender difference depends on the length of marriage, meaning that the gender difference is different for couples who have been married longer than for couples who have been married less long. This is the classic definition of an interaction.

In general, then, to test the interactions between the within-couple gender difference and the between-couple independent variables, we will regress W_{1i} on the three between-couple contrast-coded predictors, X_{1i} , X_{2i} , and X_{3i} . In such a model, the slope associated with each of these contrast-coded predictors estimates the degree to which the within-couple gender difference, coded by W_{1i} , differs as a function of each of the between-couple independent variables and their interaction. Additionally, because, X_{1i} , X_{2i} , and X_{3i} are all contrast-coded predictors with means of zero across the four between-couple conditions, to test whether the average within-couple gender difference across the four between-couple conditions equals zero, one can simply test whether the intercept in this W_{1i} regression model differs from zero.

Let us then specify two between-couple regression models, separately regressing W_{0i} and W_{1i} on the three contrast-coded predictors that code the two between-couple independent variables and their interaction:

$$W_{0i} = \beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + \beta_{03}X_{3i} + \varepsilon_{0i}$$

$$W_{1i} = \beta_{10} + \beta_{11}X_{1i} + \beta_{12}X_{2i} + \beta_{13}X_{3i} + \varepsilon_{1i}$$

In the W_{0i} model, we are examining only variation in the data between couples, analyzing a composite score that averages the scores from the two spouses in a couple. In this model, the slopes associated with each of the contrast-coded predictors estimate the degree to which the between-couple independent variables (and their interaction) affect the average responses of each couple. In the W_{1i} model, we are examining gender differences within couples. The intercept in this model, β_{10} , estimates the magnitude of the average gender difference, within couples, on average across the four cells of the between-couple design. And each of the slopes for the between-couple contrast-coded predictors in this model tests whether the gender difference (within couples) in satisfaction depends on the between-couple independent variables (and their interaction).

From these data these two models are estimated as:

$$\hat{W}_{0i} = 8.31 + 0.80X_{1i} + 0.18X_{2i} + 0.80X_{3i}$$

$$\hat{W}_{1i} = -0.44 - 0.35X_{1i} + 0.09X_{2i} + 0.18X_{3i}$$

FIGURE 11.23 Mixed model ANOVA source table

Source	<i>b</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>PRE</i>
Between couples (W_{0i})						
Length (X_{1i})	.56	10.12	1	10.12	9.54	.44
Children (X_{2i})	.12	0.50	1	0.50	0.48	.04
Length x Children (X_{3i})	.56	10.12	1	10.12	9.54	.44
Error between		12.75	12	1.06		
Total between		33.50	15			
Within couples (W_{1i})						
Gender (intercept)	-.31	3.12	1	3.12	8.82	.42
Gender x Length (X_{1i})	-.25	2.00	1	2.00	5.66	.32
Gender x Children (X_{2i})	.06	0.12	1	0.12	0.35	.03
Gender x Length x Children (X_{3i})	.12	0.50	1	0.50	1.42	.11
Error within		4.25	12	0.35		
Total within		10.00	16			
Total		43.50	31			

Each of the coefficients in these models can be put back into the Y metric by dividing by the denominator of the respective W , which in both cases is $\sqrt{2}$, yielding the parameter estimates that are presented in the second column of the source table given in Figure 11.23.

The various rows of this table provide tests of the various effects, either between couples (from the W_{0i} equation) or within couples (from the W_{1i} equation). And for each of these tests the appropriate error term (either between or within) comes from the sum of squared errors from one of the two models. Thus, each test represents a model comparison, setting a particular parameter equal to zero in the Model C, with which the above Model A is compared. The between-couple tests examine the significance of each of the slopes associated with the contrast-coded predictors. The within-couple tests examine the significance of the intercept (i.e., whether the gender difference on average differs from zero) as well as the significance of each of the slopes. Let us go through the various tests summarized in this table.

First, between couples, there is a significant length of marriage difference in satisfaction ratings. This F and PRE results from a comparison of the W_{0i} Model A with a Model C that fixes the slope of X_{1i} at zero. The estimated slope of X_{1i} in Model A equals 0.80. In the source table, we have converted this back into the original Y metric, dividing 0.80 by $\sqrt{2}$. The result, .56, equals half the difference between the mean satisfaction ratings for couples married for 30 years and those married for 15 years. There also is a significant length of marriage \times children interaction in the mean satisfaction ratings given by each couple, tested by comparing the W_{0i} Model A with a Model C that fixes the slope of X_{3i} at zero. Given our codes and the resulting parameter estimates, it appears that only among couples with children does satisfaction seem to increase with the length of the marriage.

Turning to the within-couple effects, first there is a significant gender difference, tested by comparing the W_{1i} Model A to a Model C that fixes the intercept at zero. The coefficient reported for this test in the source table, i.e., $-.31$, equals half the difference between the mean satisfaction rating of the female spouses and that of the males. The

other significant effect within couples is the gender \times length of marriage interaction, tested by comparing the W_{1i} Model A to a Model C that fixes the slope of X_{1i} at zero. It appears that the gender difference, with males reporting higher satisfaction than females, depends on the length of the marriage: It is particularly found among couples who have been married 30 years.

Once again, to overcome the problem of nonindependence due to couples, each couple occurs in each Model A, the Model A for W_{0i} and the Model A for W_{1i} , only once. Thus, the residuals in these models come from couples rather than from observations within couples. Additionally, the variances of these residuals (i.e., the mean square errors within couples and between them) are rather unequal (1.06 between versus .35 within), indicative of the expected positive dependence of observations due to couples. Between-couple effects are appropriately tested by comparing them to the magnitude of the between-couple error variance; within couple effects (and that includes the interactions of a within-couple difference with between-couple factors) are appropriately tested by comparing them with the magnitude of the within-couple error variance.

SUMMARY

The strategy that we have just employed in analyzing the data from a mixed design, with one independent variable varying within couples (gender) and two varying between them (length of marriage and the presence or absence of children), can readily be generalized to other, more complicated designs in which there are grouped observations, giving rise to potential nonindependence, with some factors varying within groups and others varying between them. Let us review the general strategy.

When there are multiple observations that are likely to be nonindependent because they are grouped in some way (e.g., they come from the same subject, the same couple, the same small group, etc.), one needs to combine those observations, analyzing composite scores in order to meet the assumption that the residuals are independent of each other. With s linked observations from each group, one forms a total of s composite W_i scores and these composite W_i scores are used as dependent variables in a series of between-group regression models. The first of these W_i scores, W_{0i} , is defined by setting all values of δ_h equal to 1, adding up the s scores from each group and then dividing by the square root of s . This W_{0i} composite score captures all of the between-group variation in the observations. One regresses this W_{0i} composite score on the between-group contrast-coded predictors (or continuous predictors as well; see Chapter 12) that code any between-group factors of interest. In this way, one tests the main effects and interactions of these between-group factors, just as they were tested in Chapters 8 and 9.

The remaining $s-1$ W_i composite scores are computed by using within-group contrast codes as values of δ_h in order to code the within-group factors of interest. Thus, each remaining W_i (from the set $W_{1i}, W_{2i}, \dots, W_{s-1i}$) codes a specific within-group comparison or difference between observations, capturing the main effects and interactions of the within-group factors of interest. Each of these W_i composite scores represents a single-degree-of-freedom within-group comparison. One then regresses these W_i scores on the contrast-coded predictors that code any between-group factors of interest. In each of these regression models, a test of whether the intercept differs from zero is a test of the within-group comparison or contrast coded by that particular W_i . Tests of the regression

coefficients for the between-group contrast-coded predictors in each of these models are tests of interactions between the within-group contrast coded by that particular W_i and the between-group contrasts.

Importantly, there is one error term for the test of all between-group independent variables, while each within-group contrast has its own unique error term (from its own Model A) that is used to test that particular within-group contrast as well as the interactions of that contrast with the between-group factors. What may seem strange initially about these tests of interactions of a within-group contrast and one or more between-group factors is that they are not examined through the use of product predictors. Rather, they are examined simply as slopes of the between-group contrast-coded predictors, asking whether a particular within-group difference (i.e., W_i) depends on those predictors.

To this point, we have insisted on computing W_i and then reporting coefficients that have been put back into the Y_{hi} metric by dividing each by $\sqrt{\sum_h \delta_{hi}^2}$. Our insistence here has been motivated by the desire to compare the source table (and sums of squares) from the correct analysis that appropriately handles the nonindependence in the observations with an incorrect one that ignores the grouping of observations. Thus, we wanted to show that the incorrect analysis simply combined the within-group and between-group errors into a single error term, ignoring the fact that observations within groups are more or less similar than observations between them. At this point, however, it is important to say that the use of W_i scores is not necessary. Equally appropriately, one can simply analyze the average of each group's s scores, instead of W_{0i} , to capture the between-group variation in the data, and simple within-group differences to capture the within-group contrasts of interest. This analysis of within-group averages and differences will yield the same F and PRE statistics as does the analysis of W_i scores, although the resulting sums of squares will not add up to give the total sums of squares in the Y observations. The additional advantage of using simple averages and differences, over and above computational ease, is that the resulting parameter estimates will be in the correct Y metric rather than having to be converted from the W_i metric.

We give one last example to illustrate all that we have just summarized, this time without data. We give simply the design and outline the resulting source table. Imagine that a study in cognitive psychology were conducted, exposing subjects to word lists of varying length for a fixed amount of time. One set of lists consists of 10 words, one of 20 words, and one of 30 words. In addition, half of these word lists consist of nouns and half consist of adjectives. These two independent variables vary within subjects, so that each subject is exposed to six different lists, created by crossing the three levels of list length with the two levels of word type (noun versus adjective). In addition, there is an independent variable that varies between subjects: study time. Some subjects are given 1 minute to study each list, some 2 minutes, and some 3 minutes. Thus, the design is a three (list length) by two (word type) by three (study time) factorial experimental design, with the first two factors varying within subjects and the last varying between them.

Each subject provides six observations, one memory score for each of the six lists of words that he or she is shown. These six observations are likely to manifest positive nonindependence due to being grouped within subjects: some subjects exhibit better memory across all lists than do other subjects. Accordingly, with six observations we

FIGURE 11.24 Values of δ_h for composite variable in word list memory example

Composite variable	10 Words		20 Words		30 Words	
	Noun	Adjective	Noun	Adjective	Noun	Adjective
	δ_h	δ_h	δ_h	δ_h	δ_h	δ_h
Y_{ave}	+1	+1	+1	+1	+1	+1
Y_{dif1}	-1	-1	0	0	+1	+1
Y_{dif2}	-1	-1	+2	+2	-1	-1
Y_{dif3}	-1	+1	-1	+1	-1	+1
Y_{dif4}	+1	-1	0	0	-1	+1
Y_{dif5}	+1	-1	-2	+2	+1	-1

FIGURE 11.25 Rows of the source table for the word list memory example

Source	df
Between subjects	
Linear time	1
Quadratic time	1
Error between	57
Total between	59
Within subjects	
Linear list length	1
Linear list length x Linear time	1
Linear list length x Quadratic time	1
Linear list length error	57
Quadratic list length	1
Quadratic list length x Linear time	1
Quadratic list length x Quadratic time	1
Quadratic list length error	57
Word type	1
Word type x Linear time	1
Word Type x Quadratic time	1
Word type error	57
Linear list length x Word type	1
Linear list length x Word type x Linear time	1
Linear list length x Word Type x Quadratic time	1
Linear list length x Word type error	57
Quadratic list length x Word type	1
Quadratic list Length x Word type x Linear time	1
Quadratic list length x Word type x Quadratic time	1
Quadratic list length x Word type error	57
Total within	300
Total	359

need six composite scores from each subject. Each of these six composite scores is defined using the values of δ_h in Figure 11.24. The first of these codes, designated Y_{ave} , is the average of each subject's six memory scores across the six lists. It captures all the between-subject variation in the data. The remaining five composite scores code within-subject contrasts or differences and are labelled $Y_{dif1}-Y_{dif5}$. The values of δ_h for these within-subject contrasts indicate that the first codes the linear effect of list length, the second codes the quadratic effect of list length, the third codes the word type difference, the fourth codes the linear list length by word type interaction, and the fifth codes the quadratic list length by word type interaction.

To code the between-subject independent variable, study time, we construct two between-subject contrast-coded predictors: X_1 with values of λ_i of -1 , 0 , and $+1$ to code the linear study time difference; and X_2 with values of λ_i of -1 , 2 , and -1 to code the quadratic study time difference. Then we estimate six regression models, regressing each Y composite score (Y_{ave} , and $Y_{dif1}-Y_{dif5}$) on the two between-subject contrast-coded predictors, X_1 and X_2 . In the model where Y_{ave} is the dependent variable, we are looking only at between-subject variation. Here we are interested in the slopes of the two between-subject contrast-coded predictors, X_1 and X_2 , which estimate the effects of study time on average across all six types of memory lists. Then in the five models with dependent variables that code within-subject differences, we are interested in both the intercepts and the slopes associated with the two between-subject contrast-coded predictors. Each intercept estimates the mean within-subject difference coded by the particular Y_{dif} dependent variable, and each of the slopes for X_1 and X_2 estimate the degree to which that within-subject difference is moderated by study time (linearly and quadratically).

Assuming that there are 20 subjects in each of the three between-subject conditions (study time), this analysis gives rise to the source table summarized in Figure 11.25, indicating only the sources of variance and corresponding degrees of freedom.

CONCLUSION

In this chapter, we have considered models for the analysis of categorical independent variables when observations are nonindependent because they are grouped in some way, and the independent variables vary either between groups of linked observations or within them. In the vocabulary of typical experimental design books in psychology, such analyses are called repeated-measures and mixed-model ANOVA. Although our approach to such ANOVA models yields results that are identical to those that would be obtained using more traditional ANOVA programs, the advantage to our approach, once again, is that we only need to estimate Models A and C, using any least-squares regression procedure, once the important data transformations to define both the dependent and predictor variables have been done.

In addition to relying on our by-now thoroughly familiar Model A and Model C comparisons, our approach to these repeated-measures situations has the additional advantage that we are readily able to extend these traditional topics by including independent variables that are measured continuously rather than categorically. It is to these extensions that we turn in the next chapter.

Notes

- 1 The consistency of this sum of squares for the condition, as we have already said, is due to having put $\sqrt{\sum_h \delta_{hi}^2}$ in the denominator of W_{0i} rather than more simply $\sum_h \delta_{hi}^2$. Had we used the simple average for each group, rather than W_{0i} , the sums of squares would not have worked out so nicely, although the resulting F and PRE would have been unaffected.
- 2 The values of MS_b and MS_w that are used in computing the intraclass correlation are the mean squares between and within groups after having removed any effects due to the independent variables of interest. In the present case, since the independent variable varies between groups, its effects are removed from the variation between groups.
- 3 A test for whether the intraclass correlations differs from zero is provided by the F ratio of the mean square between to the mean square within in the case of positive dependence. In the case of negative dependence, the F ratio is computed as the mean square within divided by the mean square between. While the degree of nonindependence can be tested in this way, we recommend being very conservative. Whenever data are linked, it is generally wise to assume nonindependence.
- 4 We presume that the two puzzle sets and the order of conditions are both counterbalanced across the two levels of the independent variable. Thus, each puzzle set is done equally often with the experimenter present and absent. In addition, half of the subjects complete the puzzles with the experimenter present first and half complete the puzzles with the experimenter absent first.
- 5 This equivalence is ensured by the fact that the denominator of W equals $\sqrt{\sum_h \delta_{hi}^2}$. Had we just computed a simple difference score for each subject (i.e., the numerator of W_{1i}), the sum of squares would not have been equivalent even though the resulting PRE and F values would have been correct.

12 Incorporating Continuous Predictors with Nonindependent Data

Towards Mixed Models

One of the distinct advantages of our model comparison approach is that we have been able to integrate the traditionally rather separate analytic frameworks underlying the analysis of designs with categorical independent variables (ANOVA) and the analysis of correlational data, with more or less continuously measured independent variables (multiple regression/correlation). In terms of model estimation and comparison, the model comparison approach is appropriate regardless of the measurement metric of predictor variables. One simply has to learn new ways in which parameter estimates can be interpreted (i.e., as mean differences or adjusted mean differences in the case of categorical variables as well as the traditional slope interpretations). This same advantage for our approach extends to the treatment of designs in which observations are nonindependent because they are grouped in some way. As we will show in this chapter, one can readily incorporate more or less continuously varying independent variables into the models underlying the analysis of nonindependent observations.

We start with the simple case where one or more continuously measured independent variables varies between the groups of linked observations; in other words, the groups of linked observations are nested under levels of the continuously varying predictors. We then turn to the case in which continuous predictors vary within the groups of linked observations. And, finally, we examine cases in which the independent variables of interest vary both within and between grouped observations. The exposition will permit us to show how our approach finds a ready extension to recent developments in models known variously as multilevel, hierarchical, and mixed effects models. We provide an introduction to such models, involving multiple random effects that can either be hierarchically nested or crossed.

CONTINUOUS PREDICTORS VARYING BETWEEN GROUPS OF LINKED OBSERVATIONS

Suppose a grocery store chain is interested in alternative marketing strategies for a particular product, which we will refer to as Product X. Each of ten stores in the grocery chain markets Product X in four different ways over the course of four weeks. Two of

the ways involve a displayed discount price and the other two involve displaying the product along with a coupon that must be redeemed. Crossed with that, on two weeks Product X is displayed at the end of an aisle in the grocery store and on the other two weeks it is displayed in the middle of an aisle. Thus there are two crossed independent variables that vary within stores across time: whether the discount price is displayed or a coupon must be redeemed; and whether Product X is displayed at the end of the aisle or in the middle. The dependent variable is the number of units of Product X sold each week by each store, measured four different times. We assume that the specific week for each type of marketing strategy is randomly determined for each store, so that time and the two independent variables are unconfounded. Hypothetical data from the ten stores are given in Figure 12.1.

With four observations from each store, we presume nonindependence in these data due to store. And that assumption seems clearly confirmed by the large store-to-store differences in the total number of Product X purchases. Compare the difference, for instance, between store 3 and store 4, with the former having many more sales of Product X than the latter.

Our analysis of these data accordingly treats each of the independent variables as within-store factors. Therefore we need to compute four composite scores for each store to enable the appropriate analyses that overcome the dependence issues in these data. These four composite scores, along with the original data, are given in Figure 12.2. The first composite score, Y_{ave} , is the simple average of each store's sales of Product X across the four conditions (and weeks). Its variation from store to store captures the between-store variation in these sales data. The second composite score, Y_{dc} (with $\delta_h = 1$ if Discount and -1 if Coupon), captures the within-store sales difference as a function of whether the discount price was shown or whether a coupon was required. The third composite score, Y_{em} (with $\delta_h = 1$ if Aisle End and -1 if Aisle Middle), captures the within-store sales difference as a function of whether the display was at the end of the aisle or in the middle. And the final composite score, Y_{int} , captures the within-store interaction difference, examining whether the discount – coupon difference is different for end of aisle versus middle of aisle displays.

To analyze the effects of the independent variables on the sales data, we examine three different Model A / Model C comparisons, one for each of the three composite

FIGURE 12.1 Hypothetical product x sales data from 10 stores

Store	Discounted Price		Coupon	
	Aisle End	Aisle Middle	Aisle End	Aisle Middle
1	45	43	42	37
2	56	55	43	43
3	71	68	65	59
4	23	24	22	21
5	49	50	50	48
6	67	64	67	60
7	32	30	29	29
8	51	50	48	43
9	45	44	45	47
10	58	56	57	53
Mean	49.70	48.40	46.80	44.00

FIGURE 12.2 Product x sales data plus composite scores

Store	<i>Discounted Price</i>		<i>Coupon</i>		Y_{ave}	Y_{dc}	Y_{em}	Y_{int}
	<i>Aisle End</i>	<i>Aisle Middle</i>	<i>Aisle End</i>	<i>Aisle Middle</i>				
1	45	43	42	37	41.75	9	7	-3
2	56	55	43	43	49.25	25	1	1
3	71	68	65	59	65.75	15	9	-3
4	23	24	22	21	22.50	4	0	-2
5	49	50	50	48	49.25	1	1	-3
6	67	64	67	60	64.50	4	10	-4
7	32	30	29	29	30.00	4	2	2
8	51	50	48	43	48.00	10	6	-4
9	45	44	45	47	45.25	-3	-1	3
10	58	56	57	53	56.00	4	6	-2
Mean	49.70	48.40	46.80	44.00	47.23	7.30	4.10	-1.50

difference scores, asking whether each within-store mean difference differs significantly from zero. The model comparisons are:

<p>Model(s) A</p> $Y_{dc} = \beta_0 + \varepsilon$ $Y_{em} = \beta_0 + \varepsilon$ $Y_{int} = \beta_0 + \varepsilon$	<p>Model(s) C</p> $Y_{dc} = 0 + \varepsilon$ $Y_{em} = 0 + \varepsilon$ $Y_{int} = 0 + \varepsilon$
---	---

The model estimates and SSE(A)s and SSE(C)s are:

Model(s) A	SSE(A)	Model(s) C	SSE(C)
$\hat{Y}_{dc} = 7.30$	572.1	$\hat{Y}_{dc} = 0$	1105
$\hat{Y}_{em} = 4.10$	140.9	$\hat{Y}_{em} = 0$	309
$\hat{Y}_{int} = -1.50$	58.5	$\hat{Y}_{int} = 0$	81

The first of these model comparisons suggests that the mean difference due to the discount versus coupon is significant ($F_{1,9} = 8.38$, PRE = .48), such that more sales ensue from advertising the discounted price rather than using coupons. The second also suggests that the mean difference due to end versus middle of aisle product placement is significant ($F_{1,9} = 10.74$, PRE = .54), with greater sales if Product X is placed at the end of the aisle than in the middle. The model comparison for the test of the interaction is not significant, offering no evidence that the discount versus coupon difference in sales depends on the aisle placement of the product.

So far, we have used only the procedures of Chapter 11 to analyze the effects of our two within-store categorical independent variables. But now we introduce an additional variable, *GS*, which varies between stores: the gross sales volume (in hundreds of thousands of dollars) per week of the store in question. This variable, which is a continuous rather than categorical one, is given in the final column of Figure 12.3 where we reproduce the data and the composite scores given in Figure 12.2. What we would now like to examine is how the gross sales of these ten stores (*GS*) affects the results we have so far examined.

FIGURE 12.3 Product x sales data plus composite scores and gross sales (GS)

Store	Discounted Price		Coupon		Y_{ave}	Y_{dc}	Y_{em}	Y_{int}	GS
	Aisle End	Aisle Middle	Aisle End	Aisle Middle					
1	45	43	42	37	41.75	9	7	-3	2.3
2	56	55	43	43	49.25	25	1	1	1.7
3	71	68	65	59	65.75	15	9	-3	2.2
4	23	24	22	21	22.50	4	0	-2	1.4
5	49	50	50	48	49.25	1	1	-3	1.9
6	67	64	67	60	64.50	4	10	-4	2.4
7	32	30	29	29	30.00	4	2	2	1.6
8	51	50	48	43	48.00	10	6	-4	1.8
9	45	44	45	47	45.25	-3	-1	3	1.9
10	58	56	57	53	56.00	4	6	-2	2.0
Mean	49.70	48.40	46.80	44.00	47.23	7.30	4.10	-1.50	1.92

We can first ask about the relationship between gross sales (GS) and sales of Product X on average between stores. It seems likely that stores with larger total sales volume are likely to sell more of the product in question, Product X, than stores with smaller total sales volume. And indeed when Y_{ave} is regressed on GS , we find a significant between-stores relationship:

$$\text{Model A: } \hat{Y}_{ave} = -16.91 + 33.40GS \quad \text{SSE(A)} = 669.5$$

$$\text{Model C: } \hat{Y}_{ave} = 47.23 \quad \text{SSE(C)} = 1669.3$$

This comparison yields an F of 11.95, with 1 and 8 degrees of freedom, and a PRE of .60. Not surprisingly those stores with higher gross sales volume sell more of Product X.

We can also ask about interactions between the gross sales volume of the stores and the independent variables that vary within stores. Perhaps, for instance, the end of aisle placement, compared to the middle aisle placement, makes an even bigger difference in sales of Product X in stores with greater total sales volume. Or perhaps, the difference in the sales of Product X as a function of displaying the discounted price versus giving coupons is different in stores with greater total sales than in those whose total sales is smaller. And finally, even though overall there was no interaction between the aisle placement of Product X and how its discount was managed (display discount versus redeeming a coupon), it may be the case that this interaction of the two within-store independent variables depends on the overall gross sales of the store.

To examine these interactions of the within-store categorical independent variables and the between-store continuous gross sales (GS) independent variable, we want to regress each within-store difference score on GS and test whether the slope associated with GS differ significantly from zero. As we saw in Chapter 11, when a within-subject (in this case within-store) difference is regressed on a between-subject (between-store) predictor, the resulting slope informs us about the interaction of the “within” independent variable and the “between” one. This remains true even when the ‘between’ independent variable is continuous, as it is in this example.

The only slight complication in estimating these models is that if we regress the within differences on GS , the intercepts will no longer equal the mean of those difference

scores. Rather, they will equal the predicted difference for hypothetical stores whose gross sales volume equal zero. To make these intercepts more meaningful, we want to mean-deviate or center the predictor variable. The intercepts in the resulting models will then continue to tell us about the overall or mean effects associated with the within-store manipulations.

The resulting three Model As are (with GSc being the centered GS variable, i.e., $GSc = GS - 1.92$):

$$\begin{aligned}\hat{Y}_{dc} &= 7.30 - 0.18GSc & SSE(A) &= 572.1 \\ \hat{Y}_{em} &= 4.10 - 9.91GSc & SSE(A) &= 52.9 \\ \hat{Y}_{int} &= -1.50 - 3.79GSc & SSE(A) &= 45.6\end{aligned}$$

As must be the case when predictors are centered, the intercepts in these models equal the means of the three composite difference scores that are the dependent variables. These models can be compared with compact ones that fix the intercepts at zero to test the average effect of the within-store manipulations (and their interaction). Tested in the context of these models, these comparisons will yield somewhat different results than those already reported for the repeated-measures ANOVA in which GS was ignored. Rather than having an error term with 9 degrees of freedom, as in the earlier model comparisons, each of these Model As estimates two parameters and hence their error degrees of freedom equal 8. Additionally, to the extent that the GSc predictor variable explains variation in the difference scores, the SSEs for these Model As will be reduced compared to what they were when no predictor variable was included. (See Chapter 5 regarding tests of β_0 in simple regression.)

To test the interactions of GS with the within-store manipulations, we want to compare these Model As with Model Cs that omit the predictor variable and only estimate the intercepts. These were given earlier, although they were called Model As in the earlier context where we were asking whether the intercept was different from zero (with no predictor in the model) – i.e., testing the effects of the within-store manipulations without the between-store GSc predictor.

Only in the case of aisle placement difference score (Y_{em}) is the slope of GSc significantly different from zero ($F_{1,8} = 13.31$; $PRE = .62$). What this means is that the aisle placement effect reported earlier (more sales if Product X is displayed at the end of the aisle than in the middle) depends on the gross sales volume of the store. Since the estimated slope of GSc is positive, it appears that in stores with larger gross sales volume, the aisle placement of Product X makes a bigger difference than in stores with smaller gross sales volume. For stores having sales volume at the mean, the difference in Product X sales as a function of aisle placement is estimated as 4.10 units per week. But as total sales volume of the stores goes up by .1 unit (10 thousand dollars more sales), the aisle placement difference is predicted to increase to 5.19 (i.e., $4.10 + 9.91 \cdot .1$). Aisle placement makes a bigger difference at higher volume stores.

As we saw when interpreting interactions with continuous predictors in Chapter 7, it is typically informative to derive “simple” effects at representative values of a component predictor. In the present case this is equivalent to deriving predicted values of the within-store difference at representative values of GSc , for instance one standard deviation above and below its mean. The standard deviation of the GS variable is 0.32.

Thus the predicted values of the aisle end vs. aisle middle sales difference of Product X at one standard deviation above and below the mean of *GS* equal:

$$\hat{Y}_{em,+1sd} = 4.10 + 9.91(0.32) = 7.27$$

$$\hat{Y}_{em,-1sd} = 4.10 + 9.91(-0.32) = 0.93$$

By deviating *GS* from values other than its mean, the significance of the simple aisle end vs. aisle middle sales difference could be tested at particular values of *GS* other than its mean. For instance, rather than deviate *GS* from its mean of 1.92, we might deviate it from the value of *GS* that is one standard deviation above its mean (i.e., $1.92+0.32$). If Y_{em} were then regressed on this new deviated predictor, the intercept would equal the predicted difference for stores one standard deviation above the *GS* mean (i.e., the value of 7.27 computed above) and it could be tested for significance by comparing it, as Model A, to a Model C that fixed the intercept at zero.

As this example illustrates, examining the effects of continuous predictor variables that vary between groups of nonindependent observations (in this case between stores) is entirely straightforward. One simply includes them as predictors of the various composite dependent variables. In general, it will make sense to center the continuous predictors so that the intercepts continue to estimate the mean within-group contrast scores. For a model in which the composite average is the dependent variable, the slope of a continuous between-group predictor tells us about the “main effect” of that predictor, on average, across levels of the within-group factors. For a model in which a within-group contrast or difference is the dependent variable, the slope for the continuous predictor tells us about the interaction of the within-group difference with the between-group predictor. Multiple between-group continuous predictors (as well as categorical ones and their interactions) may be used simultaneously, both in predicting the composite average scores and in predicting within-group contrast difference scores. Of course, when there are multiple predictors, there is also likely to be redundancy among them. The slope of a predictor in such models would then tell us about the partial effect of a predictor, partialling out the others.

“CONTINUOUS” PREDICTORS THAT VARY ONLY WITHIN GROUPS

Suppose we were examining the effects of two different diets on weight gain in eight newly born laboratory animals. Half of the animals were randomly assigned to one diet and half to the other, and their weights were recorded every day for 10 days. The hypothetical data are given in Figure 12.4. The observations are linked because they come from the same animal. Day thus varies within animals, whereas Diet varies between animals.

Day really is a categorical variable, with 10 levels. One could accordingly define one composite score that is the average of all 10 observations for each animal and then 9 within-animal contrasts or differences. Given the intrinsic ordering of Day, a set of nine orthogonal polynomial codes would probably be of most interest, coding the linear Day effect, the quadratic Day effect, the cubic Day effect, and so forth. Of course, one would probably not be interested in much beyond the linear effect: asking simply

FIGURE 12.4 Day x Diet data

Animal	Diet	Day									
		1	2	3	4	5	6	7	8	9	10
1	-1	10	12	12	15	13	16	17	19	17	20
2	-1	8	7	8	9	10	10	12	14	13	15
3	-1	12	13	16	18	17	19	22	23	23	26
4	-1	13	12	14	14	15	17	17	18	20	21
5	1	12	12	13	13	14	14	15	14	15	15
6	1	8	7	7	8	9	9	10	10	12	11
7	1	10	10	11	11	12	13	13	14	14	15
8	1	13	12	12	13	15	14	15	15	16	15

whether the animals' weights increased over the course of the 10 days and whether the increase depended on the diet they were fed.¹ Values of the within-animal contrast code (δ_h) that codes the linear component of Day, having 10 levels, are $-4.5, -3.5, -2.5, -1.5, -0.5, +0.5, +1.5, +2.5, +3.5,$ and $+4.5$.

In Figure 12.5 we have reproduced the data along with two composite scores: the average of each animal's weight across the 10 days, Y_{ave} , and the linear difference using the values of δ_h given in the previous paragraph, $Y_{diffLin}$. Using the procedures specified in the last chapter, the between-animal main effect of Diet is estimated by regressing Y_{ave} on the contrast-coded Diet variable:

$$\hat{Y}_{ave} = 13.725 - 1.450\text{Diet} \quad \text{SSE} = 50.755$$

A comparison of this model with a Model C that fixes the slope of Diet at zero yields $F_{1,6} = 1.99$ and $\text{PRE} = .25$. With only eight animals in the dataset, the effect of Diet on the mean weight (i.e., across the 10 days) is not significant.

To examine the within-animal linear effect of Day and its interaction with Diet, we regress the linear difference, $Y_{diffLin}$, on Diet:

$$\hat{Y}_{diffLin} = 63.750 - 25.875\text{Diet} \quad \text{SSE} = 1719.875$$

We are interested in two questions here. First, the intercept estimates the average linear difference across all eight animals. Testing whether its value is significantly different from zero yields $F_{1,6} = 113.42$ and $\text{PRE} = .95$. Thus, unsurprisingly these animals gained weight across the 10 days. Second, the slope for the between-animal predictor variable, Diet, estimates the Linear Day \times Diet interaction. A test of whether it differs from zero yields $F_{1,6} = 18.69$ and $\text{PRE} = .76$. Thus, it appears that the first four animals, fed with Diet -1 , gained weight at a significantly faster rate than did the animals fed with Diet $+1$.

So far, we have simply applied the procedures of Chapter 11 for mixed models to these data, treating Day as a categorical within-animal predictor variable. Day has 10 levels and could thus be coded into 9 orthogonal within-animal contrasts. We have only examined one of these, the linear effect of Day. But one could similarly examine each of the other 8 within-animal contrasts and their interactions with Diet if one had reasons to do so. What, then, makes this an analysis of a "continuous" independent variable that varies within groups of linked nonindependent observations?

FIGURE 12.5 Day x Diet data with computed composite scores

Animal	Diet	Day										Y_{ave}	$Y_{diffLin}$
		1	2	3	4	5	6	7	8	9	10		
1	-1	10	12	12	15	13	16	17	19	17	20	15.1	84.5
2	-1	8	7	8	9	10	10	12	14	13	15	10.6	72.0
3	-1	12	13	16	18	17	19	22	23	23	26	18.9	122.5
4	-1	13	12	14	14	15	17	17	18	20	21	16.1	79.5
5	1	12	12	13	13	14	14	15	14	15	15	13.7	29.5
6	1	8	7	7	8	9	9	10	10	12	11	9.1	41.5
7	1	10	10	11	11	12	13	13	14	14	15	12.3	47.5
8	1	13	12	12	13	15	14	15	15	16	15	14.0	33.0

FIGURE 12.6 Transposed Diet x Day data matrix

Day_h	Y_{h1}	Y_{h2}	Y_{h3}	Y_{h4}	Y_{h5}	Y_{h6}	Y_{h7}	Y_{h8}
1	10	8	12	13	12	8	10	13
2	12	7	13	12	12	7	10	12
3	12	8	16	14	13	7	11	12
4	15	9	18	14	13	8	11	13
5	13	10	17	15	14	9	12	15
6	16	10	19	17	14	9	13	14
7	17	12	22	17	15	10	13	15
8	19	14	23	18	14	10	14	15
9	17	13	23	20	15	12	14	16
10	20	15	26	21	15	11	15	15

Let us think about these same data in a different way. In Figure 12.6 we have taken the data matrix of Figure 12.4 and transposed it, so that animals are now the columns rather than the rows. The first column designates the value of the Day variable (for each of its h levels), the second column gives the Y values for the first animal, Y_{h1} , the third column gives the Y values for the second animal, Y_{h2} , and so forth. If we wanted to treat Day_h as a continuous within-animal predictor variable, we could estimate its effects by regressing weight on the values of Day_h (i.e., 1–10) separately for each animal. In other words, starting with animal 1, we could regress Y_{h1} on Day_h across the 10 days. The resulting model for the first animal is:

$$\hat{Y}_{h1} = 9.467 + 1.024 \text{ Day}_h$$

In this model, the intercept, 9.467, is the predicted weight for this animal on Day 0 (given a linear model) and 1.024 is the expected linear increase in weight across each additional day. To make the intercept more interpretable, we can center Day_h ($DayC_h$) around its mean, 5.5, with the resulting parameter estimates:

$$\hat{Y}_{h1} = 15.100 + 1.024 \text{ Day}C_h$$

The slope is, of course, unchanged and the intercept now equals the mean weight across the 10 days for the first animal.

The interpretation of this within-animal slope (i.e., as the expected linear increase in weight for this animal for each additional day) is similar to the meaning of the Y_{difLin} within-animal difference score. The difference score captured the linear difference in weight for each animal across the 10 days. In fact, one can show that the within-animal slope, when Y_{hi} is regressed on Day_h , has a numerator that is equal to the within-animal difference score, Y_{difLin} , that we computed earlier. The formula for the least-squares slope in simple regression was given in Chapter 5 as:

$$b_{yx} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

In the present case, the values of the predictor variable are 1–10, with a mean of 5.5. Therefore, the values of $(X - \bar{X})$ are $-4.5, -3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5,$ and 4.5 , which are the same values that we used in constructing the within-animal difference score, Y_{difLin} . Hence, defining these as the values of δ , the within-animal contrast codes, the formula for the slope becomes:

$$b = \frac{\Sigma\delta(Y - \bar{Y})}{\Sigma\delta^2} = \frac{\Sigma\delta Y - \Sigma\delta\bar{Y}}{\Sigma\delta^2}$$

Since $\Sigma\delta$ equals zero, this expression for the slope reduces to:

$$b = \frac{\Sigma\delta Y}{\Sigma\delta^2}$$

The numerator is our within-animal linear difference score, Y_{difLin} , and the denominator is simply the sum of the squared values of δ , which in this case equals 82.5. Hence if we take the value of the slope for this first animal, 1.024, and multiply it by 82.5 we get 84.5, which is that animal's value on the composite linear difference score, Y_{difLin} , given in Figure 12.5. Thus, if we were to conduct eight within-animal regressions (i.e., one regression for each animal), regressing each animal's 10 Y_{hi} scores on the centered Day_h variable, the intercepts in the resulting models would equal each animal's mean weight (what we earlier called the Y_{ave} composite score) and the slopes would equal each animal's composite linear difference score (Y_{difLin}) divided by a constant: $\Sigma\delta^2$.

In Figure 12.7 we give for each animal the values of the intercepts (b_{0i}) and slopes (b_{1i}) that result from the eight within-animal regressions, taking each animal's 10 weight scores, Y_{hi} , and regressing them on the centered Day_h variable. Also included here is the Diet between-animal contrast-coded independent variable.

Now, we want to repeat our Model A/Model C comparisons to ask about the main effect of the between-animal variable Diet, the Linear effect of Day (within-animal), and the Diet \times Linear Day interaction. But this time, we are going to use each animal's intercept and each animal's slope from the within-animal regressions as our dependent variables, regressing each one on Diet.

Since the intercept (b_{0i}) exactly equals each animal's composite average score (Y_{ave}), the first of these regressions is identical to what we obtained previously using Y_{ave} as the dependent variable:

$$\hat{b}_{0i} = 13.725 - 1.450 \text{ Diet}_i \quad \text{SSE} = 50.755$$

FIGURE 12.7 Intercepts and slopes from within-animal regressions

<i>Animal</i>	<i>Diet</i>	b_{0i}	b_{1i}
		(Y_{ave})	$(Y_{difLin}/\sum\delta^2)$
1	-1	15.1	1.024
2	-1	10.6	0.873
3	-1	18.9	1.485
4	-1	16.1	0.964
5	1	13.7	0.358
6	1	9.1	0.503
7	1	12.3	0.576
8	1	14.0	0.400

A comparison of this model with a Model C that fixes the slope of Diet at zero once again yields $F_{1,6} = 1.99$ and $PRE = .25$. As before, this model comparison examines whether the animals given one diet differ in average weight (i.e., averaging across the 10 days) from the animals given the other diet.

The second Model A is one in which the within-animal slopes (b_{1i}) are regressed on Diet:

$$\hat{b}_{1i} = 0.773 - 0.314\text{Diet}_i \quad \text{SSE} = 0.253$$

In a deep sense this model is identical to the one estimated earlier where Y_{difLin} was the dependent variable. Since the value of each animal's estimated slope is simply that animal's value of Y_{difLin} divided by $\sum_h \delta_h^2$, the coefficients in this model equal the estimated coefficients from the earlier Y_{difLin} model divided by this same factor. Accordingly, the intercept in this model estimates the average linear effect of day ($F_{1,6} = 113.42$ and $PRE = .95$) and the parameter associated with the Diet predictor variable estimates the Linear Day \times Diet interaction ($F_{1,6} = 18.69$ and $PRE = .76$). In other words, these models, using the intercepts and slopes from the within-animal regressions as the dependent variables, estimate exactly the same effects as those estimated using the procedures developed in Chapter 11 for mixed-factor designs. The two procedures for dealing with designs in which some independent variables (e.g., Day) are crossed with groups of linked observations (e.g., from the same animal) and others vary between groups of linked observations (e.g., Diet) are thus equivalent. This will always be the case whenever the intervals between levels of the within-subject factor are equal in magnitude and there are no missing data.

Why have we gone to these lengths to show that the analysis of slopes and intercepts that derive from the within-animal regression models gives us the exact same information as the procedures developed for mixed designs in Chapter 11? The answer is that the analysis of such slopes and intercepts, derived from within groups of linked observations, provides us with tremendous flexibility in examining the effects of independent variables of interest that vary both within and between groups of linked observations. Let us review the procedure once again before examining more complicated scenarios.

With observations that are nonindependent because they are grouped in some way (e.g., they come from the same subject, the same family, the same group, etc.), one can estimate the effects of independent variables that vary within the grouped observations.

One can do so by estimating within-group regression models, regressing the dependent variable for each set of linked observations on the independent variables for that group. From such within-group models, assuming that the independent variables have been centered, the intercepts will tell us about the mean of the linked observations for each group and the slopes will estimate the differences in those observations associated with the within-group independent variables of interest. (These may involve partial associations if multiple within-group independent variables that are partially redundant are used as predictors). Such independent variables can vary either continuously or categorically and interactions among them may also be included in within-group models, if such interactions are theoretically expected.

The intercepts and slopes from these within group regression models become the data values for further analysis, where the groups are the units of analysis rather than the individual observations. In essence one regresses the within-group intercepts and slopes, in separate models, on any between-group independent variables that are of interest. In the model where the intercepts are the data variable, assuming centering of within-group predictors, the slopes associated with the between-group predictors estimate the effects of those predictors “on average,” averaging across the levels of the within-group predictors. In the slope model(s), where the within-group slopes constitute the data variables, the slopes associated with the between-group predictors estimate interactions between the within-group predictors and the between group predictors.

This description of the relevant models is both abstract and dense, but the important point is that this general approach to the analysis of data that exhibit dependencies because of grouping is a very general and flexible one that has found widespread usage in recent years under the name of multilevel modeling, albeit with different and simultaneous estimation procedures. In the following we provide a general introduction to these models.

MULTILEVEL MODELING

To introduce these models initially in a rather simplified context, imagine that we have data from students who are grouped in classrooms. From each student we have a standardized achievement test score and we want to analyze these as a function of different independent variables, some of which vary within the classrooms and some of which vary between them. We start by assuming that classrooms induce dependence in the test scores, caused by a variety of potential factors such as within-classroom interaction, common experiences, etc. We will use Y_{hi} to represent the test score of the h th student in the i th classroom, thus keeping the same subscripts in this example that we have used all along: the i subscript refers to the group of linked observations and the h subscript refers to the individual observations within the group.

Let’s start by assuming that we have one independent variable, student gender, that varies from student to student within classrooms, and another independent variable, teacher gender, that varies between classrooms (but not within – assuming each class is taught by one teacher). The first of these we will refer to as SG_{hi} (contrast-coded with values of +1 if the student is female and –1 if male). The teacher gender variable is called TG_i , and is contrast-coded in a parallel way. Notice that the student gender variable has both h and i subscripts because it varies from student to student within classrooms,

while the teacher gender variable only has an i subscript since it varies only between classrooms.

We start by estimating the effect of the student gender variable on achievement test scores by estimating within-classroom slopes and intercepts from the following within-classroom model:

$$Y_{hi} = \beta_{0i} + \beta_{1i}SG_{hi} + \varepsilon_{hi}$$

Notice that the parameters in this model have i subscripts, meaning that each classroom has its own intercept and slope. The intercept is the mean achievement test score in a given classroom, averaging across the male and female students in the classroom, and the slope is the student gender difference (actually half the difference given our coding) between the test scores of the female students and the male students in the classroom. In the language of multilevel models, the within-classroom model is referred to as the level-one model, meaning that it captures the mean and student gender effect at the lowest level of the nested data structure, across the students who are nested within a classroom.

Then the parameters from this level-one model are in turn modeled at level-two, between classrooms, with one level-two model for the level-one intercepts and one level-two model for the slopes:

$$\beta_{0i} = \alpha_{00} + \alpha_{01}TG_i + \mu_{0i}$$

$$\beta_{1i} = \alpha_{10} + \alpha_{11}TG_i + \mu_{1i}$$

Both of these level-one parameters are modeled as a function of the level-two independent variable, teacher gender. In other words, both the intercepts and slopes are regressed separately on teacher gender. To differentiate the parameters in these level-two models from those in the level-one model, we use α instead of β , and each parameter has two subscripts; the first subscript indicates whether it comes from the level-one intercept or slope model and the second indicates whether it is an intercept or a slope in the level-two model. So, for instance, the parameter α_{01} comes from the level-one intercept model and it is the slope of the level-two TG predictor. In other words, α_{01} refers to the slope associated with teacher gender in the regression of the level-one intercepts on teacher gender. Finally, to differentiate the errors in these level-two models from those in the level-one model we have used μ_{0i} and μ_{1i} for the residuals in the level-one intercepts and slopes, respectively.

In Figure 12.8 we provide interpretations of the parameters and error terms for these models.

The level-two models can be used to test specific null hypotheses by comparison of these models with appropriate compact models. For instance, a test of whether α_{01} differs from zero tests whether on average (or more precisely averaging across male and female students) there is a mean difference in the test scores of classrooms taught by female and male teachers. A test of whether α_{10} differs from zero tests the mean score difference (collapsing across male and female teachers) between male and female students. And a test of whether α_{11} differs from zero tests the student gender by teacher gender interaction – does the student gender difference within classrooms depend on the teacher's gender – are student scores higher when there is a gender match between the student and teacher than a gender mismatch?

FIGURE 12.8 Interpretations of parameters and error terms for models predicting achievement test scores

β_{0i}	Mean test score for the i th classroom, on average across the male and female students.
β_{1i}	Half the female–male mean difference in the i th classroom.
ε_{hi}	Individual test score residual around the predicted gender mean for the i th classroom.
α_{00}	Mean of the classroom average test scores, on average across the male and female teachers.
α_{01}	Half the difference in classroom average test scores between those classes with female teachers and those with male teachers.
μ_{0i}	Residual in classroom mean test scores over and above that predicted by the model.
α_{10}	Mean of the classroom student gender differences, on average across the male and female teachers.
α_{11}	Half the difference between the female and male teachers in the within-classroom difference in test scores due to student gender.
μ_{1i}	Residual in classroom student gender difference in test scores over and above that predicted by the model.

These models could all be estimated sequentially, first at level-one, estimating each classroom's intercept and slope, and then at level-two, estimating the level-two intercepts and slopes. But doing this is not optimal, particularly if classrooms differ in size and in variability. One would prefer a simultaneous estimation procedure that somehow produces best estimates, given the overall variability in the data attributable to the various sources. The full multilevel model that we would like to estimate is found by substituting the level-two models for the parameters into the level-one model. We reproduce these models below and then we do the substitution in the level-one model:

$$\text{Level-one: } Y_{hi} = \beta_{0i} + \beta_{1i}SG_{hi} + \varepsilon_{hi}$$

$$\text{Level-two: } \beta_{0i} = \alpha_{00} + \alpha_{01}TG_i + \mu_{0i}$$

$$\beta_{1i} = \alpha_{10} + \alpha_{11}TG_i + \mu_{1i}$$

Substituting level-two models for the parameters in the level-one model:

$$Y_{hi} = (\alpha_{00} + \alpha_{01}TG_i + \mu_{0i}) + (\alpha_{10} + \alpha_{11}TG_i + \mu_{1i})SG_{hi} + \varepsilon_{hi}$$

and then rearranging:

$$Y_{hi} = \alpha_{00} + \alpha_{01}TG_i + \alpha_{10}SG_{hi} + \alpha_{11}TG_i * SG_{hi} + \mu_{0i} + \mu_{1i}SG_{hi} + \varepsilon_{hi}$$

What we now have is a single model that integrates both the level-one and level-two models that we would like to estimate. In some ways, this integrated multilevel model should look very familiar; in other ways however it looks rather different. First, the part of the model with the α parameters seems to make sense. First we are estimating an intercept, which is like the grand mean (given contrast coding), then the effect of teacher gender, then that of student gender, and finally their interaction. But then the model gets rather strange since we have three different residual or error terms, with these defined in Figure 12.8. First there is residual error at the classroom level in the classroom means, μ_{0i} ; second, there is also residual error at the classroom level in the effect of

student gender, $\mu_{1i}SG_{hi}$; and finally there is residual error at the level of the individual student test scores over and above the predicted values from the model, ε_{hi} . These different error terms very likely have different variances, meaning that they cannot be simply lumped together into one error term.

As a result, the estimation of this model cannot be done using those least squares procedures that have been used throughout the book to this point. This is because we don't have a single error term that we are attempting to minimize. Rather we have multiple residual error terms and we need to estimate them (and the model parameters) in some optimal sense. And that optimal sense is by finding the values of the parameters and residual variance that maximize the joint likelihood of the data, assuming a normal distribution. Back in Chapter 3 we provided a brief introduction to maximum likelihood estimation and how, in the context of most of the book's models, least squares estimates and maximum likelihood estimates were one and the same. But now in the present context, with multiple error or residual variances, they diverge and no least squares estimate is obtainable. But an iterative approach that finds the parameter and residual variance values that maximize the joint likelihood of the data is possible.² Widely available software, commonly known as Mixed Models procedures, will do the estimation.

Another equivalent way to write the mixed model is informative in specifying the random components of the model:

$$Y_{hi} = (\alpha_{00} + \mu_{0i}) + (\alpha_{10} + \mu_{1i})SG_{hi} + \alpha_{01}TG_i + \alpha_{11}TG_i * SG_{hi} + \varepsilon_{hi}$$

According to this specification, the model's intercept (in the first set of parentheses) has two components – the overall mean, α_{00} , and a “random” component that allows the mean score to vary from classroom to classroom, μ_{0i} . Likewise the slope for the SG predictor (in the second set of parentheses) has two components, its average slope across classrooms, α_{10} , and another “random” component that allows the slope of SG to vary across classrooms. The slopes for TG and the SG by TG interaction do not have random components, since TG varies only between classrooms and not within them. For an effect to vary across classrooms, that effect has to be estimable at the lowest level, within classrooms.

To do the mixed model estimation, programs generally require that the data be in “long” form, meaning that each row of data is an individual observation (in this case from an individual student within a classroom). Thus the data file in this case would consist of as many rows of data as there are students, specifying for each an individual Y_{hi} score, a student identifier, h , a classroom identifier, i , the student's gender, SG_{hi} , and the teacher's gender in that classroom, TG_i . Then the mixed model specification indicates that Y_{hi} scores are to be modeled as a function of student gender, teacher gender, and their interaction, with “random” intercepts and student gender slopes for classrooms. The output generally gives the parameter estimates and the random error variances (in this case the classroom intercept variance, the classroom slope variance for student gender, and the level-one residual variance). It is also possible to estimate the covariance between the classroom random intercepts and slopes. These variances (and covariance) can frequently be of substantive interest. For instance, in the present context it might be interesting to know to what extent classrooms vary in the magnitude of the student gender difference, or whether classrooms with higher average scores show smaller student gender differences (i.e., the intercept – slope covariance).

Because there are multiple error variances in the model, calculations of degrees of freedom and standard errors for parameter estimates are complex and beyond the scope of what we can here cover. Most mixed model programs provide estimated degrees of freedom and approximate t tests for individual parameters under what is called the Satterthwaite approximation. Additionally, model comparisons can be made by comparing deviance statistics of nested models, with differences in deviance values distributed approximately as chi-square statistics. Details on these issues can be found in texts devoted exclusively to multilevel mixed models (Hox, 2010; Snijders & Bosker, 2012).

There are a number of distinct advantages in using a general multilevel modeling approach to the estimation of models for nonindependent data (compared with the least-squares approaches outlined earlier in this chapter and the preceding one). First, one need not make restrictive assumptions about the numbers of observations in the groups that give rise to the nonindependence. For instance, in the present context, the number of students in each classroom could vary considerably, whereas our previous approach assumed a constant number of observations in each group (and only one observation for each level of the within-group predictors). Second, and relatedly, estimation is possible in the presence of missing data, whereas with repeated measures ANOVA approaches, if a group of data is missing one or more observations typically the whole group has to be omitted from the analysis. Third, as we next illustrate, continuous predictors that vary both within and between groups, at both level-one and level-two, can be incorporated. And finally, a simultaneous estimation of the level-one and level-two models is both efficient and statistically optimal.

A Second Multilevel Model with Continuous Predictors

Suppose a researcher were interested in the relationship between how much one exercised and one's mood. Ideally, if she were interested in seeing whether there was a causal effect of exercise on mood, she would experimentally manipulate exercise level, randomly assigning participants to different levels and then later measuring their mood. More practically and to get an initial assessment of how exercise and mood are related, she might do a longitudinal observational study, measuring both exercise level and mood repeatedly over time in some sample of participants. Figure 12.9 presents some of the data that might have been collected.

In these hypothetical data, occasions of measurement are sampled within participants with the number of occasions varying from participant to participant. On each occasion, participants indicate their daily level of exercise (1–10 self rating scale) and the positivity of their mood after exercising (again 1–10 self rating scale). In practice, many different participants and many different occasions within participants would be sampled.

We would like to build a model in which the exercise variable is used to predict mood. Observations are grouped in the sense that numerous occasions come from the same participant and dependency due to that grouping is likely. Further, the independent variable, exercise, varies both within and between participants. To illustrate what this means, consider participants 1 and 2 in Figure 12.9. Clearly within these two participants, exercise varies from occasion to occasion. Additionally, however, participant 1 exercises on average less than participant 2 (participant 1 average: 2.67;

FIGURE 12.9 Hypothetical exercise and mood data

<i>Participant</i>	<i>Occasion</i>	<i>Exercise</i>	<i>Mood</i>
1	1	3	5
1	2	1	4
1	3	4	4
2	1	5	6
2	2	7	5
2	3	3	8
2	4	5	5
3	1	1	3
3	2	3	7

participant 2 average: 5.00); thus exercise varies between them as well. To fully understand the relationship between exercise and mood in these data, we need to examine that relationship both within and between participants.

We start the specification of the multilevel model by writing out the model for relating exercise to mood within each participant, i.e., the level-one within-participant model:

$$M_{hi} = \beta_{0i} + \beta_{1i}E_{hi} + \varepsilon_{hi}$$

With occasions grouped or nested within participants subscript h now refers to the occasion and subscript i refers to the participant. According to this model, each participant has an intercept and a slope. Because the exercise variable has not been centered, each participant's estimated intercept will be his or her predicted mood on days he or she fails to exercise. An individual participant's slope will estimate the difference in his or her mood as a function of having exercised one unit higher on a given occasion. Given that we will be predicting these level-one, within-participant parameters estimates at level two, between participants, it certainly makes sense to center the level-one predictor in this case and importantly around each participant's own mean level of exercise. That is what we have done in the following revised level-one model, where \bar{E}_i equals the mean exercise score of the i th participant across his or her observed occasions.

$$M_{hi} = \beta_{0i} + \beta_{1i}(E_{hi} - \bar{E}_i) + \varepsilon_{hi}$$

In this revised level-one within-participant model, each participant's estimated intercept will equal his or her mean mood score across occasions.³

At level-two, we model the level-one parameters as a function of the participant's mean exercise level:

$$\beta_{0i} = \alpha_{00} + \alpha_{01}\bar{E}_i + \mu_{0i}$$

$$\beta_{1i} = \alpha_{10} + \alpha_{11}\bar{E}_i + \mu_{1i}$$

In order to make the intercepts in these models more interpretable, we will also center the predictor in these models, centering each participant's mean exercise level around the grand mean:

$$\beta_{0i} = \alpha_{00} + \alpha_{01}(\bar{E}_i - \bar{E}.) + \mu_{0i}$$

$$\beta_{1i} = \alpha_{10} + \alpha_{11}(\bar{E}_i - \bar{E}.) + \mu_{1i}$$

Thus, in the level-one model, the predictor is the difference between a participant's exercise score on a given occasion and his or her mean exercise level, while in the level-two model, the predictor is the difference between each participant's mean exercise level and the grand mean of exercise.

We can now integrate the level-one and level-two models (both with their centered predictors) into a single multilevel model, substituting the level-two models for the parameters in the level-one model:

$$M_{hi} = (\alpha_{00} + \alpha_{01}(\bar{E}_i - \bar{E}.) + \mu_{0i}) + (\alpha_{10} + \alpha_{11}(\bar{E}_i - \bar{E}.) + \mu_{1i})(E_{hi} - \bar{E}_i) + \varepsilon_{hi}$$

Rearranging terms, to group together the parameters (also called the "fixed effects") and the random error terms, yields:

$$M_{hi} = \alpha_{00} + \alpha_{01}(\bar{E}_i - \bar{E}.) + \alpha_{10}(E_{hi} - \bar{E}_i) + \alpha_{11}(\bar{E}_i - \bar{E}.) (E_{hi} - \bar{E}_i) + \mu_{0i} + \mu_{1i}(E_{hi} - \bar{E}_i) + \varepsilon_{hi}$$

Thus, the model to be estimated includes fixed parameters for the grand mean, α_{00} , the between-participant effect of mean exercise level, α_{01} (do people who habitually exercise more than the average have higher moods?), the within-participant mood effect, α_{10} (on occasions when exercise levels are higher or lower than is habitual do people report higher moods?) and their interaction, α_{11} (do people who habitually exercise more show greater effects of variation in exercise across occasions on their moods?). In addition to these fixed effects, the model specifies random variation in participant means or intercepts and in their slopes for day-to-day fluctuations in exercise levels (and presumably their covariance). Output from most mixed-model procedures would generally include the estimated fixed parameters (and their associated standard errors) and the variances of the random components.

The interaction term in the model is included as a result of including the between-participant predictor in the level-two model for the slope. If one respecified the level-two slope model without the predictor:

$$\beta_{0i} = \alpha_{00} + \alpha_{01}(\bar{E}_i - \bar{E}.) + \mu_{0i}$$

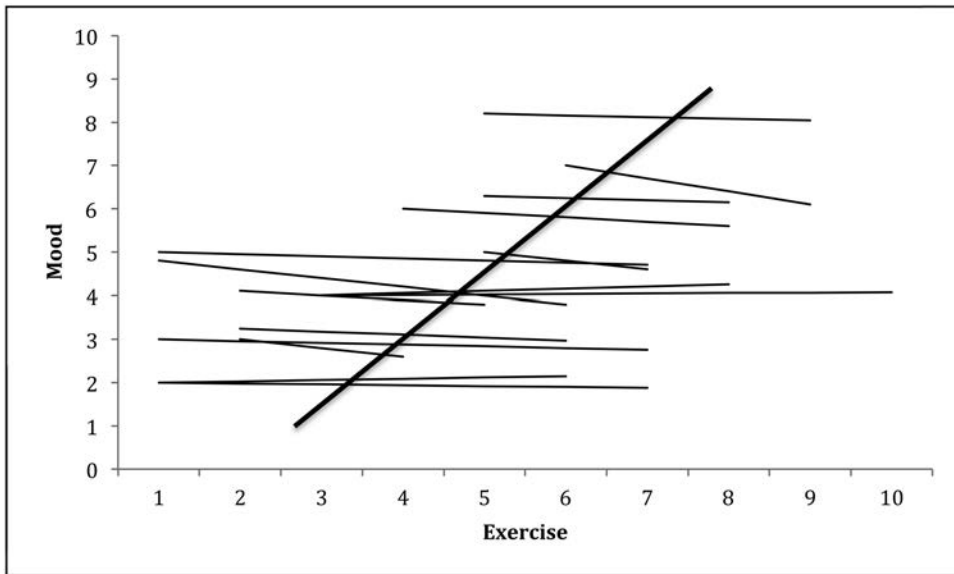
$$\beta_{1i} = \alpha_{10} + \mu_{1i}$$

one would be omitting the interaction from the full mixed model, which becomes:

$$M_{hi} = \alpha_{00} + \alpha_{01}(\bar{E}_i - \bar{E}.) + \alpha_{10}(E_{hi} - \bar{E}_i) + \mu_{0i} + \mu_{1i}(E_{hi} - \bar{E}_i) + \varepsilon_{hi}$$

It is entirely possible, given the correlational nature of the observational data that underlies this example, that the two exercise effects (one due to within-participant fluctuations and one due to between-participant variation) have very different associated slopes. For instance, day-to-day fluctuations in exercise might be negatively predictive of day-to-day fluctuations in mood while nevertheless chronic exercisers are chronically in better moods.

Figure 12.10 presents a hypothetical plot of possible results from this mixed model. Each thin line is the level-one model for a given participant, showing the within-participant relationship between exercise and mood. Many of these lines have negative slopes. At the same time, on average across the participants, it seems from these hypothetical data that those participants who on average exercise more have higher mood scores, as indicated by the thick line in the figure.

FIGURE 12.10 Exercise-mood hypothetical relationships from multilevel model

Heavy line depicts the between-participant relationship between mean exercise level and mean mood. Lighter lines depict within-participant relationships between exercise and mood on different occasions.

Crossed Random Effects Models

We have used two data examples in the previous section to illustrate multilevel models. The first involved students (level-one) nested within classrooms (level-two). The second involved occasions (level-one) nested within participants (level-two). Because these models incorporate error variances at both levels, appropriate analyses permit generalization of effects with hypothetical new samples at both levels. In other words, if we found a significant effect of exercise on mood, either within participants across occasions or between participants, and if the assumptions underlying the analysis were met, then we could reasonably generalize conclusions to other samples of participants and other samples of occasions that might have been used, assuming these were sampled in the same way as those included in the original data set. This sort of generalization has been the goal of null hypothesis testing throughout the models used in this book. We want to know whether parameter estimates might reasonably be expected to differ from a-priori values (usually zero) if other error tickets (i.e., other samples of errors) had been used. But in the present context, there are error tickets accruing from multiple “random” factors at both levels one and two.

Frequently in psychology experiments, random factors are crossed with each other rather than being nested one within the other, as in the previous examples. Suppose, for instance, that one were interested in whether stereotypic judgments (e.g., “how dangerous is this individual?”) of Black male faces differed significantly from such judgments of White male faces. One might ask a sample of participants to judge a sample of faces, half of which were Black and half White, with the goal of determining whether Black faces are judged on average differently from the White faces. Just as the participants

FIGURE 12.11 Hypothetical data from participants judging Black and White faces on a stereotypic dimension of judgment

Participant	Black Faces					White Faces				
	Face1	Face2	Face3	Face4	Face5	Face6	Face7	Face8	Face9	Face10
1	7	8	5	9	4	4	5	3	7	8
2	6	5	4	8	3	5	6	4	6	7
3	4	6	2	7	1	4	3	2	4	4
4	6	7	3	7	4	6	4	3	6	7

should properly be considered a sample of possible participants, so too the faces used should be considered just a sample of possible faces that might have been used to represent the two ethnic categories.

Consider the hypothetical data in Figure 12.11. Each row of data represents the judgments of ten faces by an individual participant, with the first five judgments of black faces and the second five of white faces. In this example, faces are crossed with participants, meaning that every face is judged by every participant. In this sense, this design, with two sampled factors, is different from those considered earlier where the two sampled factors (for instance students in classrooms) were nested, one under the other.

The standard data analytic approach to this design would be to compute two mean scores for each participant, one for the judgments of the Black faces and one for the judgments of White faces, and then to ask whether there is a significant difference between these two means on average across participants, using a within-participant repeated measures ANOVA. What this standard analysis does is treat participant as a random factor but it ignores the variation in the data that accrues from individual faces, since it averages across the faces within ethnicity. For instance, in the data of Figure 12.11, within the Black faces, it seems that ratings of Face4 are rather high on average and ratings of Face5 rather low on average. This difference, due to faces, is lost if one simply computes and analyzes average ratings for each participant.

Clark (1973) and Judd, Westfall, and Kenny (2012) have shown that this standard analytic approach, ignoring the variation due to faces, results in substantially inflated test statistics, if the goal is to reach conclusions about the ethnicity difference that generalize to studies involving different samples of both participants and faces. In other words, “significant” effects due to ethnicity may fail to replicate in studies that involve other faces sampled in the same way from the two ethnic categories. In essence both participants and faces induce random variation (dependencies) in the data and these must be taken into account in the analysis if one is to reach appropriate statistical conclusions.

To understand this bias if the variation due to faces is ignored, simply imagine that different samples of faces were used. In one study, all the Black faces, by luck of the draw, were like faces 2 and 4 with relatively high ratings. And in another study, all the Black faces, again simply by chance, were more like faces 3 and 5, with relatively low ratings. Because of these chance differences each of these studies may have reached very different conclusions if they failed to take into account the sampling variability inherent in the data due to the mean face differences.

The appropriate analysis is a mixed model one that has “error” terms deriving from both random factors, participants and faces. However, these two random factors are not

nested, as in the previous examples, so one cannot specify the appropriate mixed model by talking about random factors and models at levels one and two. Instead, the mixed model specifies the fixed parameters that one wishes to estimate along with all other components of random variation that are estimable in the data. For the current design, the fixed effects are the grand mean and the mean difference due to face ethnicity. The grand mean is the intercept in a linear model with ethnicity contrast-coded and the mean difference due to face ethnicity is the slope for the ethnicity predictor. Participants and face are both random factors, for which one can estimate a mean: in the matrix of ratings of Figure 12.11, the row means are the participant means and the column means are the face means. Additionally, for participants, but not for faces, one can calculate an ethnicity slope or difference. That is, because participants are crossed with ethnicity, a ethnicity slope can be estimated for each participant. But this is not true for the random factor of face, since each face occurs in only one ethnicity condition rather than in both.

Hence, for this design, the full mixed model that should be estimated, including all variance components attributable to both random factors (participants and faces), is:

$$Y_{ij} = \beta_0 + \beta_1 Eth_{ij} + \mu_{0i} + \mu_{0j} + \mu_{1i} Eth_{ij} + \varepsilon_{hi}$$

$$Y_{ij} = (\beta_0 + \mu_{0i} + \mu_{0j}) + (\beta_1 + \mu_{1j}) Eth_{ij} + \varepsilon_{hi}$$

In this model, the i subscript is used to refer to participant and j to face. Y_{ij} is the i th participant's judgment of the j th face and Eth_{ij} is a contrast coded predictor representing ethnicity. We have given the model in two versions, first with the fixed effects given initially followed by the four random effects, and second with the intercept components in the first parentheses and with the slope components in the second parentheses. The model specifies an overall intercept (or mean) with variations around that due to participant and face means; then an overall difference due to ethnicity, allowing that slope to vary randomly from participant to participant; and finally a random error at the level of the individual observation. One can estimate this mixed model using maximum likelihood mixed modeling procedures, again with the data in "long" form (i.e., one rating of one face by one participant per row of data) rather than in the form given in Figure 12.11. The output from the estimation procedure would include the two parameter estimates (and associated standard errors) and variance estimates attributable to all four random components of variance (along with the covariance between participant intercepts and slopes).

With nested designs, i.e., one random factor (e.g., occasions) nested under the other (e.g., participants), we saw that independent variables could vary either between levels or within levels of each one. This is also true in the case of designs with two random factors that are crossed (as in the participant by face example). Independent variables can vary either between or within the levels of each random factor. In the design of Figure 12.11 and the mixed model specification just given, the independent variable varies within participants but between faces. Many other designs are possible and the appropriate mixed model specification for each is given by Judd, Westfall, and Kenny (2012). The basic rule is that random intercepts are generally estimable for both random factors while random slopes are estimable for only those independent variables that vary within the levels of a particular random factor.

As a more complicated example involving crossed random factors, consider a well-known study on the "shooter bias" effect (Correll, Park, Judd, & Wittenbrink, 2002).

In a videogame-like format, participants were shown targets emerging from complicated background scenes. The targets were either Black or White males and they were either carrying a handgun or some other object (e.g., cell phone, coke can). Participants had to quickly decide whether to “shoot” the target (a correct decision if he was armed) or “not shoot” the target (correct if he was unarmed). Individual targets (Black or White males) appeared in both armed trials and unarmed ones. The question was whether correct responses were faster when Black rather than White targets were armed and faster when White rather than Black targets were unarmed, consistent with a stereotypic bias towards seeing Black males as dangerous.

The original reported analyses treated participants as the only random factor, computing four mean latencies for correct judgments for each participant: Black armed, Black unarmed, White armed, and White unarmed. Significant effects were a main effect of armed versus unarmed (faster correct responses to armed targets) and a ethnicity of target by armed versus unarmed interaction (in line with the prediction). A more appropriate analysis treats both participants and individual targets as crossed random factors. Targets are nested under ethnicity but crossed with armed versus unarmed so that target intercepts and slopes for the armed – unarmed factor could be estimated. Participants are crossed with both the ethnicity of the target and with whether the target was armed or unarmed. Hence, for participants, intercepts and slopes for target ethnicity, armed versus unarmed, and their interaction are all estimable. The appropriate mixed model specification and the resulting output from this model are given in Judd et al. (2012).

SUMMARY AND CONCLUSION

We began this chapter by showing how continuous predictors that vary between groups of linked observations can be incorporated into models that examine differences within groups. The strategy was a simple extension of the analyses described in Chapters 5 and 11. That is, variables that assess within group differences are regressed on the continuous predictors. Tests of the intercepts in these models provide tests of the within group differences so long as the continuous predictors are centered. Tests of the slopes tell us whether the magnitudes and directions of the within group differences depend on the level of the continuous predictor(s).

We then examined how to incorporate continuous predictors that vary only within groups of linked observations. In this chapter, for example, rats’ weights were recorded daily for 10 days and we examined whether the rats gained weight during the 10 days and whether that weight gain depended on diet—a categorical variable that varied between subjects (i.e., rats). We described two ways to analyze these data, noting that they yield the same conclusions whenever the within subject intervals (e.g., day) are equal and there are no missing data. The first method involved computing a composite variable reflecting the linear change in weight for each rat, using the methods described in Chapter 11. The second involved estimating a separate regression equation for each rat. That is, for each rat, weight was regressed on a continuous day variable having values of 1 to 10. The intercepts and slopes from these equations then became the dependent variables in further analyses. This second method, commonly known as multilevel modeling, is a simple extension of our modeling approach to repeated-measures ANOVA.

We then extended our discussion of multilevel modeling, noting that they are special cases of what are called linear mixed models. Such models allow simultaneous estimation of effects at multiple levels, with appropriate error variances accruing from random factors at the different levels. Estimation is based on maximum likelihood procedures rather than least-squares, allowing efficient and optimal estimation.

Multilevel models offer distinct advantages over traditional repeated measures analyses. First, they can incorporate continuous as well as categorical independent variables and these independent variables may be partially redundant with each other.

Second, as we have indicated, there may be missing data, unequal within subject intervals, and unequal numbers of observations in the groups of nonindependent observations. In the weight-gain example, for instance, some rats may have data for only eight or nine days rather than all 10 days.

Third, the predictor variables (whether continuous or categorical) can vary both within groups of linked observations and between them. Consider again the example in which exercise was used to predict mood. The amount of exercise varied within participants from day to day. It also varied between participants; some participants exercised more overall than did others. Multilevel modeling provides a way of simultaneously estimating both within participant and between participant effects.

Finally, we introduced crossed random effects models in which multiple random factors are crossed with each other rather than being nested, as in multilevel models. In the first crossed random effects model, for example, we described how both participants and stimuli (i.e., faces) could, and indeed, should, be treated as random factors. The goal of research is to reach conclusions that generalize to a broader group of participants and to a broader group of stimuli; crossed random effects models permit that sort of generalization.

We have sought to illustrate the flexibility available through multilevel analysis and to encourage researchers to think creatively about the sorts of questions they can ask about grouped or linked observations. Although we have highlighted the considerable advantages of multilevel programs, we believe that the intelligent data analyst can accomplish a great deal with a simple general-purpose multiple regression program. Thinking carefully about the model comparisons that make the most theoretical sense prior to embarking on the further complexities of multilevel modeling programs, will provide a more thorough understanding of the data and the models that are theoretically reasonable.

Notes

- 1 Since each within-animal contrast has its own unique error term, as discussed in the previous chapter, the test of any one of them is completely independent of the test of any and all others. Thus, we can test only the within-animal contrasts in which we are interested, ignoring the other within-animal contrasts.
- 2 In fact, for most multilevel models the estimation involves a variation of maximum likelihood that is called 'restricted' maximum likelihood. What this means is that the parameters and the residual variances are estimated sequentially rather than simultaneously.
- 3 Centering around each participant's own mean rather than the grand mean is important so that the level-one intercepts are the participant's own mean mood rather than his or her predicted mood at the mean exercise level of all participants and occasions.

13 Outliers and Ill-Mannered Error

To this point we have generally assumed that error in our basic equation:

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

is well-behaved. In the two previous chapters, we examined the consequences of nonindependent errors and developed an analytic strategy to address the problem. In this chapter, we examine other ways in which error may be ill-mannered. In particular, we consider violations of our original assumption that ε_i is normally distributed with homogeneous variance. These problems are generally not as serious as those associated with nonindependence. However, in many instances the problems caused by non-normality and heterogeneous variance are substantial and, in any case, an examination of the assumptions almost always leads to a better understanding of one's data and the construction of better models.

There are two broad classifications of violations of assumptions about error: systematic and idiosyncratic. As we shall see, the systematic violations are often best identified visually by examining a variety of specialized graphs. Violations sufficient to make us doubt either the estimated model or to question the appropriateness of the probability values for the test statistics usually stand out in these specialized graphs. Our remedy, just as it was for nonindependence, is to transform the data so that we may use our full range of techniques for building models. Idiosyncratic violations arise from wild observations, commonly referred to as *outliers*, that for one reason or another are dramatically inconsistent with the other observations. The presence of outliers frequently causes violations of normality and homogeneous variance assumptions. Hence, we begin by considering methods for detecting and resolving the problems of idiosyncratic outliers and then turn to the visual methods for detecting assumption violations. As a practical matter, it is usually best to address the idiosyncratic outlier issues first, and then consider possible systematic violations of the normality and homogeneous variance assumptions.

OUTLIERS

Outliers are extreme observations that for one reason or another do not belong with the other observations in the data. There are many ways in which outliers can be introduced into the data. The first cause we consider is scientifically uninteresting but quite troublesome and common. Data recording and data entry errors can put wild values into

the data and the predictor variables used in models. The use of computers for data analysis increases the possibility of producing such outliers in our data and at the same time reduces our chances of finding them, unless we look carefully for them. Consider, for example, the effects of entering the heights in inches and the weights in pounds for a number of individuals and then for one individual reversing the numbers for height and weight. If we were doing our calculations with the aid of a calculator, we would likely notice such a reversal. However, when using the computer we are so far removed from our data that we would be unlikely to notice such a mistake.

Throughout we have used the minimization of SSE as the method for estimating model parameters and comparing models. As we noted in Chapter 2, SSE is especially sensitive to outliers. A single outlier can sometimes dramatically alter estimates of group means or regression parameters. Hence, a large data entry error would likely bias the parameter estimates and/or inflate SSE. A cursory examination of the usual output from statistical regression programs would not suggest the presence of such an error. Allowing an outlier to “grab” the model parameters is obviously undesirable. Perhaps more insidious, even if the model is not distorted, is that inflation of SSE substantially increases the difficulty of testing other more complex models by examining changes in SSE. When a few outliers contribute a large proportion of the total error, they overshadow any reductions in error achieved by increasing model complexity. We obviously need statistical techniques for identifying erroneous observations so that they can be removed or corrected in the data.

As an example of the deleterious effects of a data recording error, let us again consider the following two sets of numbers that we examined in Chapter 2:

$$\text{Set 1: } 1359 \ 14 \quad \bar{Y} = 6.4 \quad \text{MSE} = s^2 = 26.8$$

$$\text{Set 2: } 1359 \ 140 \quad \bar{Y} = 31.6 \quad \text{MSE} = s^2 = 3681$$

Set 2 is identical to Set 1 except that a data entry error has added an extra digit to 14 to make it 140. As we noted before, this error has a dramatic impact on the mean, which is our estimate of β_0 for the simple model. In this case, the estimate changes from 6.4, which is in the middle of the data values, to 31.6, which is not very representative of any of the data values. The effects on SSE and, consequently, MSE are at least as bad. The inflated MSE makes any hypothesis testing extremely difficult. Examine the 95% confidence interval for β_0 for both sets of numbers:

$$\text{Set 1: } [0, 12.8]$$

$$\text{Set 2: } [-43.7, 106.9]$$

The confidence interval for Set 2 is much wider than that for Set 1; in fact, the confidence interval for Set 1 is entirely included within the confidence interval for Set 2. Hence, many hypotheses about β_0 that could be easily rejected for Set 1 will not be rejected for Set 2. The data entry outlier thus not only produces a misleading parameter estimate but also reduces the statistical power of any inferences. The outlier has so inflated the SSE that it is much more difficult to detect any proportional reduction in error produced by using β_0 instead of B_0 in a model for these data. The same would be true for more complex models.

Outliers are, of course, not necessarily erroneous values such as those that result from data entry errors. A second cause of outliers is that the observations are not a single

homogenous set to which a single model will apply, but rather a heterogeneous set of two or more types of observations, one of which is much more frequent. The infrequent cases of the other types will appear as outliers. Discovering that there are really two kinds of things in our data when only one was expected is almost always interesting scientifically. An example is identifying children who thrive and excel despite being raised in adverse environments. Examining other characteristics of extreme observations, especially if they are outliers, may provide clues for building more complex models. In this case, we again need statistical techniques for identifying extreme or outlier observations, not necessarily so that they can be removed or corrected but so that they can be examined with great care to extract their full informational value.

A third cause of outliers is what statisticians refer to as error distributions with “thick tails” in which extreme errors occur with greater frequency than expected for a normal distribution. We examine the normality assumption in greater detail later. For now, it suffices to note that just a few observations from the thick tails of a non-normal error distribution might appear as outliers.

Detecting Outliers

There are three separate outlier questions. Given that our analyses nearly always have predictor variable(s) X_{i1} (X_{i2}, \dots, X_{ip-1}) and always have a data or outcome variable Y_i , two outlier questions are obvious. Is the value for the predictor variable (or the set of predictor values) unusual? Is the value for the outcome variable unusual? The third question pertains to the joint effect of unusual predictor and outcome values by asking about the influence of the observation on joint inferences about all the parameters in the model. That is, does the observation distort or have undue influence on the overall regression model? Each of these three outlier issues is considered in turn after we consider an example, which we use throughout as an illustration.

An outlier example

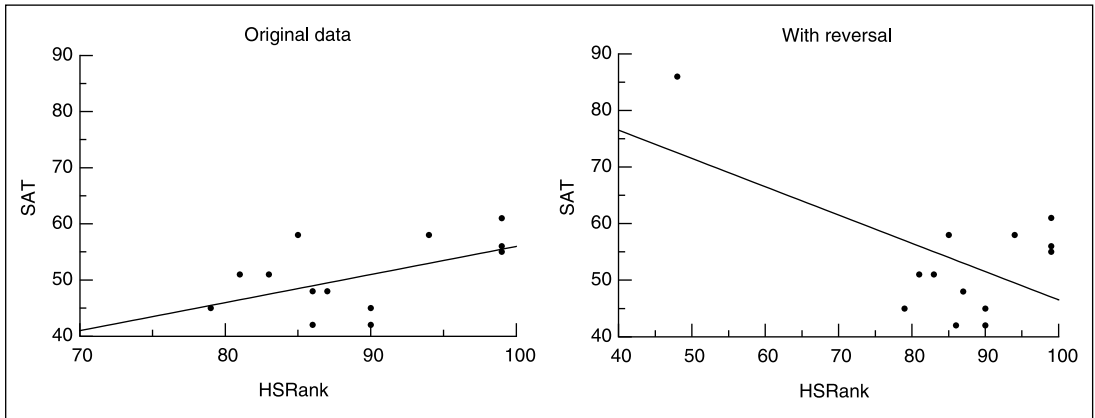
We now consider a detailed example to demonstrate the serious consequences of even a single outlier and to motivate the detection techniques to be presented later. Figure 13.1 displays the SAT verbal scores and high school ranks for 13 students. The data values on the left side of the figure are the correct values. The values on the right side are identical except that we have simulated an outlier (such as could be caused by a data entry error) by reversing the SAT and high school rank values for the sixth student. Below each dataset are the results of simple regression analyses using high school rank to predict SAT verbal scores. Also given are the values for PRE and F when asking whether high school rank is a useful predictor of SAT verbal scores. Note that PRE and, consequently, F are slightly larger for the dataset that contains the data reversal for the sixth student and the parameter estimates are dramatically different. The intercept changes from about 6 to 97, and the slope changes sign from +0.5 to -0.5. The PRE and F are statistically significant (at $\alpha = .05$) for the dataset with the data entry error, so we would conclude that as high school rank increases, SAT verbal scores decrease! This nonsensical result might cause us to check the data further, but nothing in the basic regression results or in a quick scan of the data in tabular form indicates that a serious problem exists.

FIGURE 13.1 SAT verbal and high school rank scores with and without data error (reversal of scores for Student 6)

Student	Original data		With reversal	
	SAT	HSRank	SAT	HSRank
1	42	90	42	90
2	48	87	48	87
3	58	85	58	85
4	45	79	45	79
5	45	90	45	90
6	48	86	86	48
7	51	83	51	83
8	56	99	56	99
9	51	81	51	81
10	58	94	58	94
11	42	86	42	86
12	55	99	55	99
13	61	99	61	99
	$b_0 = 5.95$		$b_0 = 96.55$	
	$b_1 = 0.50$		$b_1 = -0.50$	
	PRE = 0.29		PRE = 0.33	
	$F_{1,11} = 4.53$		$F_{1,11} = 5.35$	

How might we detect the outlier in the second dataset? It is almost always easier to see outliers in graphs than in tables. Hence, the first step is to examine the scatterplot for SAT and HSRank. The scatterplots along with the best-fitting lines for the datasets with and without the reversal of the scores for the sixth observation are displayed in Figure 13.2. The vertical axes in both scatterplots are drawn to the same scale so that the size of an error (i.e., deviation from the least-squares regression line) is comparable.

The scatterplot for the original data in the top half of Figure 13.2 has the pattern we would expect when the model is appropriate. The data values are clustered around a straight line that has a slope of 0.5, as revealed by the regression analysis in Figure 13.1. Note that none of the observations is unusually far from the line or from the other observations. Thus, no outlier is apparent in the top scatterplot. In contrast, an outlier stands out clearly in the bottom scatterplot for the dataset with the reversed observation. Again, none of the observations is particularly far from the bestfitting line (hence the significant value of PRE = .33). However, the outlier observation in the upper left-hand corner of the bottom scatterplot is unusual in three ways. First, the value of this outlying observation on the predictor variable HSRank is unusually small. If we examine the values on the predictor variable for all observations, there is only one large gap in those values—between the predictor value for this unusual observation (48) and the next lowest predictor value (79). Second, the value on the data variable (SAT) for the outlier is unusually large. Again, there is only one large gap between the data value for the unusual observation (86) and the next highest data value (61). Third, if the outlier case were omitted from the analysis, then the best-fitting line would have a positive slope (as in the first scatterplot) rather than a negative slope. This makes the outlier case unusual because it does not appear that omitting any other observation would have nearly as

FIGURE 13.2 Scatterplots and best-fitting lines for data without (left) and with (right) reversal of scores for sixth observation

large an impact on the slope for the regression line. These three ways in which the outlier is unusual suggest that to detect outliers we will want to ask the following three questions for each observation (note: we use the subscript i to represent a general observation and subscript k to represent an observation being considered as an outlier):

1. Is X_{k1} or, more generally, is the set of predictors $X_{k1}, X_{k2}, \dots, X_{kp-1}$ unusual?
2. Is Y_k unusual?
3. Would omission of the observation produce a dramatic change in the parameter estimates $b_0, b_1, b_2, \dots, b_{p-1}$, or equivalently in the predictions \hat{Y}_i ?

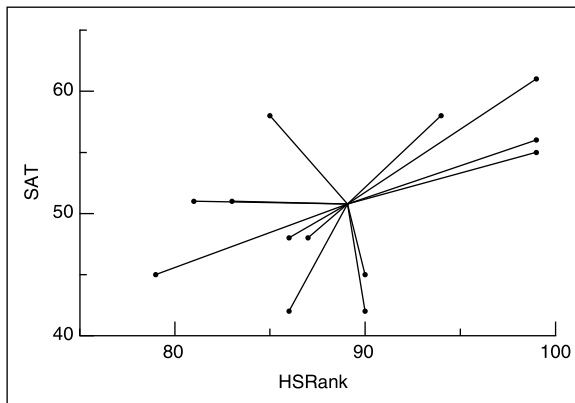
If the answer to any of these questions is yes, then we have an outlier requiring attention. For the HSRank and SAT data it appears as if the answer to all three questions is yes for the artificial outlier we produced by reversing the scores.

We now consider formal techniques for answering the three outlier questions. The statistical literature abounds with various mathematical indices and graphical procedures for each question. Except in the most unusual circumstances, most of these different procedures give essentially the same answers, so we consider only a small subset of the possible techniques. We have selected those techniques that are most consistent with the general approach to data analysis via model comparisons.

Is the predictor value X_k unusual?

All observations contribute equally to estimating the mean: that is, each observation has a weight of $1/n$. In contrast, each observation does not contribute equally to the estimate of the slope in regression. To gain insight into this issue and its importance, it is useful to reconsider an expression from Chapter 5 for estimating the slope of the best-fitting least-squares line. That is:

$$b_1 = \sum w_i \left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}} \right], \text{ where } w_i = \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

FIGURE 13.3 Individual slopes for each data point for original data in Figure 13.1

The term in brackets is the slope “suggested” by each observation or data point—the change in Y between the point and the mean for the variable Y divided by the change in X between the point and the mean for the variable X —that is, the slope of a line between the data point and the mean point. The estimated slope for the regression is simply the weighted average of all the individual slopes suggested by each point. Figure 13.3 illustrates the individual slopes for each data point in the original data of Figure 13.1.

The weight (w_i) given to each data point when calculating the overall slope estimate is based on the unusualness of the data point in terms of the predictor variable X . The further the predictor variable is from the mean, the greater its weight. This makes sense because the slopes of lines with a long horizontal component in Figure 13.3 are unlikely to be affected much by small changes, perhaps caused by error, in the outcome variable Y . Hence, our confidence in the estimated slope is greater for long lines. In contrast, small changes in Y for short lines could dramatically alter the slope. Hence, our confidence in the estimated slope is much less for short lines. One can think of each observation as having a vote on the slope that is to apply to all the observations, but with the votes of some observations counting more. If all the observations are telling more or less the same story, then all the observations ought to be voting for essentially the same slope. In the case of a perfect relationship, all the individual slopes would be identical and would equal the overall slope. It is undesirable for a single observation to have a very large percentage of the total weight, because in that case the slope votes of all the other data points are ignored in calculating the “overall” slope. In that case, the “overall” slope is really a description of only one data point. This occurs when one observation has a very unusual (relative to the other observations) predictor value.

Most modern regression programs report the lever (sometimes unhelpfully referred to as the diagonal of the hat matrix) h_i , which represents the weight or leverage that an observation has in determining the overall model. For simple regression, the lever is defined as:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

For simple regression, the model consists of two parts—the mean (or intercept) and the slope. Each observation has equal weight in the voting for the mean (i.e., $1/n$) and has a weight proportional to the squared deviation of the predictor variable from its mean in the voting for the slope (i.e., w_i). The lever, as defined above, is simply the sum of those two weights.

Unusually high values for the levers are undesirable because they imply that the estimated model in effect only applies to the one or two observations with unusually high levers. It is inappropriate and misleading to report a regression equation as if it applies to all n observations when it actually only applies to the few observations that have particularly high levers.

In judging the magnitude of levers it is useful to note that the sum of the levers necessarily equals the number of parameters, which is two for the case of simple regression. Hence, the average lever equals $2/n$ for simple regression. If a single lever is near 1.0, then it implies that one of the two parameters of the simple regression model is allocated to predict that single observation; this is clearly undesirable. In Figure 13.4 there are columns for the lever values for the original data of Figure 13.1 and for the lever values when there is the data entry error. For a simple regression based on 13 observations, the average lever equals $2/n = .15$. For the original data, which conform to our notions of a typical regression, the levers range around this average from .08 to .25, as expected. However, for the data where we introduced an outlier by reversing the predictor and outcome variables for one observation, the levers range from .08 to .15 except for one extreme value of .76. This implies that of the two parameters in the simple regression, 0.76 of a parameter is effectively allocated to predict a single observation, leaving only 1.24 effective parameters to make the predictions for the other 12 observations. Having such a high proportion of the model focused on a single observation is clearly inconsistent with the goal of having the model describe all the data.

Most researchers know that restricting the range (more properly, restricting the variance) of the predictor variable attenuates the correlation. The converse applies when there is a single outlier with an extreme predictor value: that observation artificially increases the range (variance) and thus inflates the correlation. Allowing one observation to inflate the correlation, or even to create one when otherwise there would be no relationship, obviously increases the chances of making Type I errors—rejecting the null hypothesis when the null hypothesis is in fact correct. Observations with unusual predictor values, assuming they do not also have unusual outcome values, often make Model As appear better than they actually are. In those cases, the story told by the regression model really only pertains to that one observation and it is very misleading to pretend that the regression story applies to all the data.

The generalization of levers from models with one predictor to models with multiple predictors is the same as the generalization of parameter estimates in Chapter 6. If there is no redundancy among the predictor variables, then additional terms for each predictor variable are added to the lever equation. If there is redundancy, then the computations are best left to the computer algorithms. Redundancy also introduces a new wrinkle conceptually for the lever. An observation's predictor values may not be unusual with respect to each predictor variable. However, the pattern across all the predictors might be unusual. For example, if we were using actual height and weight to predict satisfaction with body image among a sample of adolescent girls, a height of 5' 9" and a weight

FIGURE 13.4 Outlier indices for the data of Figure 13.1

Observation	Original data			Data with reversed observation		
	Lever	RStudent	Cook's D	Lever	RStudent	Cook's D
1	.08	-1.91	0.125	.08	-1.03	0.04
2	.08	-0.31	0.005	.08	-0.53	0.01
3	.11	1.96	0.182	.08	0.38	0.01
4	.26	-0.14	0.004	.10	-1.35	0.10
5	.08	-1.18	0.056	.08	-0.70	0.02
6*	.09	-0.22	0.003	.76	4.63	11.86
7	.14	0.61	0.033	.08	-0.43	0.01
8	.25	0.05	0.001	.15	0.95	0.08
9	.19	0.84	0.086	.09	-0.54	0.02
10	.12	0.89	0.055	.11	0.86	0.05
11	.09	-1.41	0.094	.08	-1.27	0.06
12	.25	-0.15	0.004	.15	0.83	0.06
13	.25	1.09	0.195	.15	1.60	0.21

of 95 pounds would not be particularly unusual, but the combination would be. The generalization of the lever detects this kind of unusualness quite well.

Is the data value Y_k unusual?

If we want to identify those Y_i that are unusual, we must ask “unusual with respect to what?” The obvious answer in the model comparison approach is that we want to find those Y_i that are unusual with respect to the model we are considering. Obviously, the errors:

$$e_i = Y_i - \hat{Y}_i$$

tell us how unusual the i th observation is with respect to the model. If the absolute value of e_i is small, then the model makes a good prediction for that Y_i , so it is not unusual. On the other hand, if e_i is large, then the model makes a bad prediction for that Y_i , so it is unusual. The individual error e_i is often referred to as the *residual* because it is the part of the original observation Y_i that is “left over” after the prediction \hat{Y}_i has been subtracted. Hence, we often refer to an examination of the individual error terms as an “analysis of the residuals.”

It is generally difficult to identify outliers by examining the absolute magnitude of e_i for two reasons. First, the importance of an error of a given magnitude is relative to the other errors and the magnitude of the prediction. We therefore need a way to transform the residuals to a common scale on which we can judge small and large values. We might, for instance, standardize them by dividing each one by the root-mean-square error.

A second problem is that there is a paradox in using the standardized residual for identifying unusual values of Y_i . Extreme data values tend “to grab” the model so that the estimated parameters minimize e_i for those extreme values. We are asking whether a particular data value, say Y_k , is unusual with respect to the model, but the extreme Y_k has itself been used to determine the model. If it has seriously altered or biased the parameter estimates in the model, then Y_k might not be unusual in terms of the magnitude

of its residual. Instead, we ought to determine whether Y_k is unusual with respect to a model determined by all the other observations *except* Y_k .

Mathematical statisticians have developed a multitude of transformations of the residuals e_i to solve these two problems in interpreting the magnitude of particular residuals. Most regression programs in computer statistical packages can produce a large variety of these transformed residuals. Some of these transformed residuals only solve the scaling problem, others eliminate the paradox by removing the effect of the k th observation when considering the k th residual, and some do both. We consider only one of these many residual indices, the *studentized deleted residual*, for three reasons. First, the studentized deleted residual both solves the scaling problem and eliminates the paradox, so there are good theoretical reasons for choosing it as an index of whether Y_k is unusual. Second, the studentized deleted residual has a natural interpretation in the context of the model comparison approach used in this book. We will demonstrate that the square of the studentized deleted residual is simply the F for comparing appropriately chosen Models C and A. Third, it is very unlikely that the studentized deleted residual would fail to detect an unusual Y_k that could be detected by any of the other transformed residuals.

We develop the studentized deleted residual by considering a specific model for an outlier. If an observation is so extreme that it is unlike the other observations, then we should be able to reduce the error appreciably by adding a specific parameter to the model just for that one observation. Designating a parameter for a single observation will ensure that there will be no error in our prediction for that observation. The other parameters are thereby freed to describe the remainder of the data. Because all error due to the outlier is eliminated and the other parameters generally provide a better fit to the remaining data, our overall measure of error will inevitably be less. In effect, we will be considering whether the outlier is fundamentally different from the other data observations. We will refer to an original multiple regression model augmented by the addition of a specific parameter for the suspected outlier as an *outlier model*. To screen data for outliers, we will test the outlier model for each observation. That is, we will test one observation at a time to see whether it is so extreme that we should allocate a specific parameter to it.

As with all other models, the question inherent in the outlier model is whether the additional complexity is worth it. Adding a parameter for the suspected outlier to the model will inevitably reduce the error, but will the reduction in error justify the increased complexity of the model? We can use PRE and F to answer this question. To be more formal, the model comparison for the outlier model is:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \beta_p X_{ip} + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$H_0 : \beta_p = 0$$

$$\text{where } X_{ip} = \begin{cases} 1 & \text{for } i = k \\ 0 & \text{for } i \neq k \end{cases}$$

where Model C is the original regression model and Model A is the outlier model having an additional parameter that estimates the effect of a single observation. If the original model really describes all the observations, then it should not be useful to isolate a single observation with its own parameter.

As an example, consider the sixth observation in the data of Figure 13.1 with the data reversal error. Let $X_i = 1$ if $i = 6$ and 0 otherwise. Then the estimation of the above models reveals:

$$\text{MODEL A: } \hat{SAT}_i = 6.71 + 0.50\text{HSRank}_i + 55.49X_i$$

$$\text{MODEL C: } \hat{SAT}_i = 96.55 - 0.50\text{HSRank}_i$$

$$\text{PRE} = .68 \quad F_{1,10} = 21.4 \quad p < .01$$

Thus, devoting one parameter to the sixth observation or, equivalently, omitting the sixth observation from the data used to fit the model reduces the error by 68%. With 11 degrees of freedom for error left after estimating b_0 and b_1 , we would expect that the omission of any one observation would, on average, reduce the error by only about $1/11 = 9\%$. The very much larger proportional reduction in error obtained by omitting the sixth observation suggests that it is a very unusual observation with respect to the model for the data. This is confirmed by the very large value of $F_{1,10} = 21.4$. Note also that the sign of the coefficient for HSRank changed between Models C and A. Indeed, omitting the sixth observation caused a dramatic difference in the estimation of both the intercept and the coefficient for HSRank. We return to this fact later.

It is interesting to consider what would happen if we were to test the outlier model for an observation that is not an outlier. Suppose, for example, that we tested the outlier model for the first observation in the data with the reversal error in Figure 13.1. Estimating the coefficients in the compact and augmented models yields (where now $X_i = 1$ if $i = 1$ and 0 otherwise):

$$\text{MODEL A: } \hat{SAT}_i = 95.71 - 0.48\text{HSRank}_i - 10.67X_i$$

$$\text{MODEL C: } \hat{SAT}_i = 96.55 - 0.50\text{HSRank}_i$$

$$\text{PRE} = .096 \quad F_{1,10} = 1.06 \quad \text{n.s.}$$

Devoting a special parameter to the first observation is clearly not worth the added complexity. The proportional reduction in error of 9.6% is little more than we would expect by chance (9%), which is confirmed by the low, nonsignificant value of F . Note that in contrast to omitting the sixth observation, omitting the first observation produces minimal changes in b_0 and b_1 .

The above results suggest that we can use the outlier model as a means of screening our data for possible outliers. We simply test the outlier model n times, each time letting a different observation be the “odd one out.” Although the outlier model is conceptually clear, it would be tedious to specify and test n different models for each regression. Fortunately, this is not necessary because Belsley, Kuh, and Welsch (1980) have shown that F for testing the outlier model for the k th observation is given by:

$$F = \frac{e_k^2(n - PA - 1)}{\text{SSE}(1 - h_k) - e_k^2}$$

All of the components for computing F for the outlier model are therefore available either from the original regression or from the levers h_k . Most computer programs for multiple regression will provide $t = \sqrt{F}$ as the studentized deleted residual for each observation. The name reflects a different derivation of the same statistic. In the model

comparison approach, it makes sense to think of the studentized deleted residual as a test of the outlier model. The studentized deleted residuals for each observation in our example are listed in Figure 13.4 under its common abbreviation “RStudent.” For the original data without an outlier, none of the values would be statistically significant using $\alpha = .05$. However, in the dataset with the reversed observation, the sixth observation has a very large value of $RStudent = 4.63$ (which squared equals 21.1, identical to the F for testing the outlier model for this observation). If the model is describing all the data, then there should not be such a large value for RStudent. The sixth observation is definitely an outlier.

We have now answered the question of whether an observation’s Y value is unusual with respect to a model of the other observations. The studentized deleted residual is a t statistic testing the outlier model. If the resulting t is large, then we conclude that allocating an additional parameter for that particular observation is worthwhile. This conclusion is equivalent to saying that that particular observation’s Y value is quite unusual with respect to the other observations. In other words, one model does not apply to all the observations.

We must be careful about using the outlier model to screen the data for outliers. If we examined the studentized deleted residuals for all observations, we would be performing the equivalent of n statistical tests on the same set of observations. If we use $\alpha = .05$ to determine the critical values of PRE and F , or equivalently RStudent, for rejecting the original model in favor of the outlier model, then we have a 5% chance of making a Type I error for *each* of the n model tests we perform. The probability that we do *not* make a Type I error in n trials equals $.95^n$. For our example with 13 observations, the probability that we do not make a Type I error is $.95^{13} = .51$. So the probability of making at least one Type I error when screening the data with the outlier model equals $1 - .51 = .49$. This is unacceptably large. One solution is to use the *Bonferroni inequality*. According to the Bonferroni inequality, the probability of making at least one Type I error will be less than α if the critical value for each test is chosen so that $\alpha' = \alpha/n$. For our example with 13 observations, $\alpha' = .05/13 = .004$, which would imply a critical value for F (t) of about 13 or 14 (3.6 or 3.7). The F (or RStudent) for the sixth observation easily exceeds this more conservative standard and so would be declared a statistically significant outlier.

In general, we do not recommend that the squared studentized deleted residual or the F that results from evaluating the outlier model be used routinely as a formal statistical test unless one has external information questioning the validity or reliability of a particular observation. Instead we suggest that large values of RStudent be used to indicate observations requiring closer scrutiny. An easy rule of thumb applies to RStudent. For reasonably large n , only about 1% of the studentized deleted residuals should be less than -3 or greater than $+3$. So, if the absolute value of RStudent is > 3 , careful attention to that observation is required. Absolute values of studentized deleted residuals of > 4 ought to be extremely rare and even after the Bonferroni adjustment will be statistically significant. Hence, all alarm bells ought to sound. An RStudent that large is a clear indication of an unusual value of Y_k . Using a model for which a studentized deleted residual is > 4 could be very misleading.

Would omission of Y_k dramatically change the predictions?

While considering the lower graph in Figure 13.2 and examining the results of the outlier model applied to the sixth observation, we noted that omitting that observation causes a dramatic change in the estimate of b_1 from -0.5 when the sixth observation is included to $+0.5$ when it is omitted. It did not appear that omitting any other observation would have an appreciable effect on the estimate b_1 and certainly would not change the sign of the estimate. This suggests that we search for outliers by considering the change in the parameter estimates resulting from the deletion of each observation from the analysis. The deletion of a “typical” observation should have little or no effect on the parameter estimates, but the deletion of an unusual observation that may have “grabbed” the estimates might result in a large change in the parameter estimates. As we noted when defining SSR conceptually, differences in parameter estimates are reflected directly in different predicted values. Hence, to assess whether the set of parameters changes dramatically as a function of one observation, we can compare the predictions for the model based on the entire set of observations to the predictions for the model estimated when one observation is omitted. Let $\hat{Y}_{i,[k]}$ represent the predicted value for the i th observation from a model estimated with the k th observation omitted. Cook’s D (1977, 1979) compares the two sets of predictions as:

$$D_k = \frac{\sum_i (\hat{Y}_i - \hat{Y}_{i,[k]})^2}{\text{PA(MSE)}}$$

A large value of Cook’s D indicates that the model predictions are very different depending on whether the k th observation is included or not. For any particular observation that was of concern, it would be relatively simple to compute the two regressions with and without the k th observation and then compute the sum of squared differences for the two sets of predictions. However, this again would be tedious as a screening procedure for outliers if we had to perform n separate regressions, each time omitting a different observation. Fortunately, as was the case for the outlier model and studentized deleted residuals, this is not necessary because the following equivalent formula for Cook’s D shows that it can be calculated from basic components available from the original regression. That is:

$$D_k = \frac{e_k^2}{\text{PA(MSE)}} \left[\frac{h_k}{(1 - h_k)^2} \right]$$

For example, for the sixth observation the only quantity we have not determined is MSE, but this is easily obtained as $\text{SSE}/(n - \text{PA}) = 1085.3/11 = 98.66$, so:

$$D_6 = \frac{13.3^2}{2(98.66)} \left[\frac{.76}{(1 - .76)^2} \right] = 11.83$$

This value (within rounding error) as well as the values of Cook’s D for all the observations are listed in Figure 13.4. The only very large value of Cook’s D is the one for the sixth observation in the dataset in which its predictor and outcome values are reversed. As we noted before, that observation grabs the model and dramatically changes the predictions for all the other observations, which is clearly undesirable in a model that purports to be a description of all the data.

There are only informal guidelines for thresholds above which values of Cook's D should require attention. One suggested rule is to consider any value greater than 1 or 2 as indicating that an observation requires a careful look. Another suggestion is to look for gaps between the largest values of Cook's D . Usually, as in Figure 13.4, a value truly requiring attention is obvious.

Many regression programs will report Cook's D , so the above computational formula will seldom be needed. However, the formula is conceptually important because it demonstrates that Cook's D is a multiplicative function of the squared error for that observation and the lever for that observation. Thus, the most influential observations in terms of their effects on the parameter estimates will be those having large squared errors (an unusual Y_k value) *and* high leverage (an unusual set of predictor values). Conversely, if either Y_k is not unusual with respect to the model for the other observations (i.e., low value of $RStudent$) *or* if the set of predictors for the k th observation is not unusual with respect to the sets of predictors for the other observations (i.e., a low value of the lever h_k), then Cook's D will not be large. But also note that if Y_k is not quite unusual enough to attract our attention and if h_k is also not quite large enough to be considered a high lever, then Cook's D can and often will be large enough to require attention.

Dealing with Outliers

Once we have identified outliers in a batch of data (by examining the levers, studentized deleted residuals, and Cook's D statistics), we have the problem of what remedial action, if any, should be taken. Although the tools to detect outliers are not particularly controversial, decisions about how to proceed in the face of clearly outlying observations have been a more contentious subject, with divergent and sometimes strong opinions devoted to the subject. If an outlying observation clearly originated because of data recording or coding errors, as in the example we have been using, then everyone would agree that it makes sense to correct the observation if possible, and redo the analysis with the corrected observation included. In many cases, however, it may not be at all clear why a particular observation is an outlier and, as a result, no clear corrective strategy is available for dealing with the problematic observation or observations. In this case, controversies may arise about how one should proceed.

In the early days, before the development of sound statistical procedures for identifying outliers, it was of course somewhat natural to question analyses that left out problematic observations based on seemingly ad hoc rules. One naturally feared, in the absence of clear rules for identifying outliers, that observations might simply have been omitted because they were problematic for the researcher in arguing for a favored hypothesis. Because of such fears, a tradition emerged in psychological research in which it was generally considered unethical to report analyses based on partial datasets that omitted problematic observations.

We believe that the development of sound statistical procedures for identifying outliers, using tools such as the lever, the studentized deleted residual, and Cook's D , has rendered this point of view obsolete. In fact, we would argue that it is unethical to include clearly outlying observations that "grab" a reported analysis, so that the resulting conclusions misrepresent the majority of the observations in a dataset. The task of data analysis is to build a story of what the data have to tell. If that story really derives from

only a few overly influential observations, largely ignoring most of the other observations, then that story is a misrepresentation.

At the very least, in the presence of clear outliers, we encourage researchers to analyze their data twice, once including the outliers and once not. If the resulting story is generally the same regardless of the inclusion of the outliers, then it is obviously simpler to include the unusual cases. On the other hand, if the story is dramatically different and the outlier statistics are clearly indicative of one or more problematic cases, then we would recommend reporting the story without the outliers, while clearly indicating the steps that were undertaken for outlier detection. And an interesting question, worthy of further pursuit, is why the unusual cases occurred in the first place and what they may be able to tell us further about the theoretical issues that are being explored. While the majority of observations may have a single story to tell, those few that make for a very different story can often be intriguing cases to follow further, to identify the limits of the story that one has told and how it needs to be expanded for other populations.

SYSTEMATIC VIOLATIONS OF ERROR ASSUMPTIONS

Outliers are observations that are idiosyncratic in a dataset, arising for unknown reasons, and leading to potentially very different stories of what the data have to say. In addition to problems that thus arise idiosyncratically, there may also be systematic tendencies of all the errors in a dataset to violate the assumptions underlying statistical inference. In Chapters 11 and 12 we dealt with the most consequential of these systematic violations, data in which errors or residuals are not independent of each other. The other two assumptions that also may be seriously violated are the assumptions that errors are normally distributed and come from a single distribution, thus having a constant variance.

It is essential to realize that the assumptions of normality and constant variance pertain only to the errors in a given model and not to any of the variables themselves, neither the data variable, Y , nor the predictors, X , in the model. We clearly have extensively used predictors whose values are not normally distributed and there is nothing in all we have done that makes any distributional assumptions about the predictors in a model. As far as the data variable, Y , is concerned, here too there are no assumptions about the distribution of this variable, as long as its errors in the models predicting it are normally distributed with constant variance. To illustrate the distinction, consider the case of a categorical predictor variable, with two levels, used to predict some Y , and there exists a very substantial mean difference between the two groups. In essence, then, our Y variable would have one common distribution that came from two within-group distributions. And if the mean difference between the two groups was substantial enough, the overall distribution of the Y variable, ignoring the two groups, would be far from normal (i.e., bimodal). Yet, within each group the distribution of Y might very well be nicely behaved (i.e., normal in each group, with the same variance). These within-group distributions are the distributions of the errors in the model that predicts the two group means. Sometimes, of course, the Y variable might be distributed in such a way that there is no possible way that its errors are normally distributed. We consider such models, that is, models in which the Y variable is dichotomous, in Chapter 14. The important point is that it is the distribution of the errors that is crucial, not the distribution of Y .

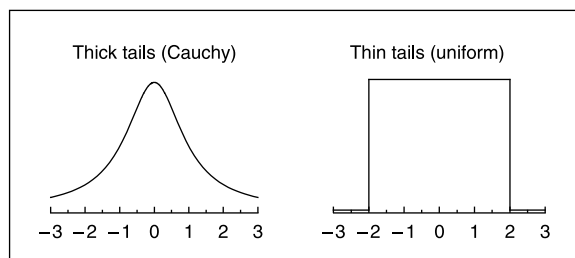
The assumptions that errors are normally distributed with a constant variance are often seen as “robust” assumptions, in the sense that the accepted advice is that violations of the assumptions do not result in major statistical errors. While these assumptions are clearly more robust than the independence assumption considered in the last two chapters, it nevertheless is true that violations of normality and constant variance in many cases can result in inferential errors (both Type I and Type II errors). Therefore it makes sense for the researcher to attend to these issues, particularly in cases where there is good reason to suspect violations—cases that we detail below.

As outlined above, the detection of outliers involves motivated statistics and model comparisons that give clear indexes of the degree to which particular cases might pose analytic problems. The detection of non-normality and heterogeneous variances of errors is a much less precise science and we will largely emphasize data plots and inspections of them to detect violation problems. Although there are statistical tests of departures from normality and of heterogeneous variance, we do not emphasize these because they often involve other assumptions that may be problematic. It is our experience that plots of data will often make clear when these assumptions are violated substantially enough that corrective action is warranted. In this sense, we regard these assumptions as robust: if their violations are likely to cause problems, they should be apparent based on relatively nonexact methods of data inspection.

Diagnosing Non-Normal Error Distributions

As already noted, the assumption that errors are normally distributed is a relatively robust assumption. But the validity of this conclusion depends on the nature of the departure of the distribution of errors from the normal one. Departures of distributions from the normal distribution occur primarily in two ways. On the one hand, the normal distribution is symmetric, with tails of equal thickness both above its central tendency and below it. Other distributions may depart from it because they are skewed in one direction or the other, either with a thicker positive than negative tail or the other way around. On the other hand, distributions may depart from a normal one because they are too peaked, with tails that are too thin, or insufficiently peaked with tails that are too thick. Although it is important to correct error distributions that are skewed compared to the normal one, the most serious problems are encountered when distributions of errors have substantially thicker tails than the normal distribution. Figure 13.5 portrays two distributions. On the left is what is called a Cauchy distribution, resulting from the ratio of two normally

FIGURE 13.5 Cauchy distribution with thick tails (left) and uniform distribution with thin tails (right)



distributed variables. On the right is a uniform distribution. Although the Cauchy distribution looks roughly normal, its tails are actually considerably thicker than those of a normal distribution. On the other hand, the uniform distribution has tails that are considerably thinner than the normal distribution. Because least-squares estimation particularly attends to extreme errors (by minimizing the sum of *squared* errors), thick-tailed distributions, like the Cauchy distribution, are potentially much more problematic than other non-normal distributions such as the uniform one.

In our judgment, the most useful tool for detecting whether the distribution of errors departs from a normal one is the normal quantile-quantile plot, provided by all major software programs in procedures that plot univariate distributions of data.¹ The normal quantile-quantile plot displays the percentile distribution of the data (i.e., the errors) against the percentile distribution that would be expected if the distribution were normally distributed. Thus, the data values are rank-ordered and the percentile of each score in the distribution, based on the number of observations, is computed. These rank-ordered percentile scores for each observation constitute the values on the vertical axis of the normal quantile-quantile plot. Along the other axis are the z scores for each percentile coming from the normal distribution. Thus, each observation is located by a point in the two-dimensional plot, with its value on the vertical axis being its percentile in the actual distribution of errors and its value on the horizontal axis being what its z score would be if the percentiles came from a normal distribution.

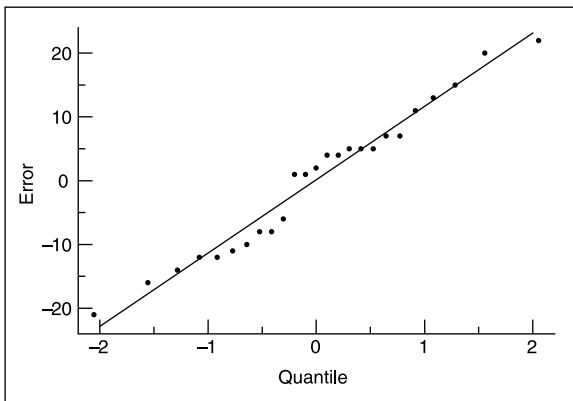
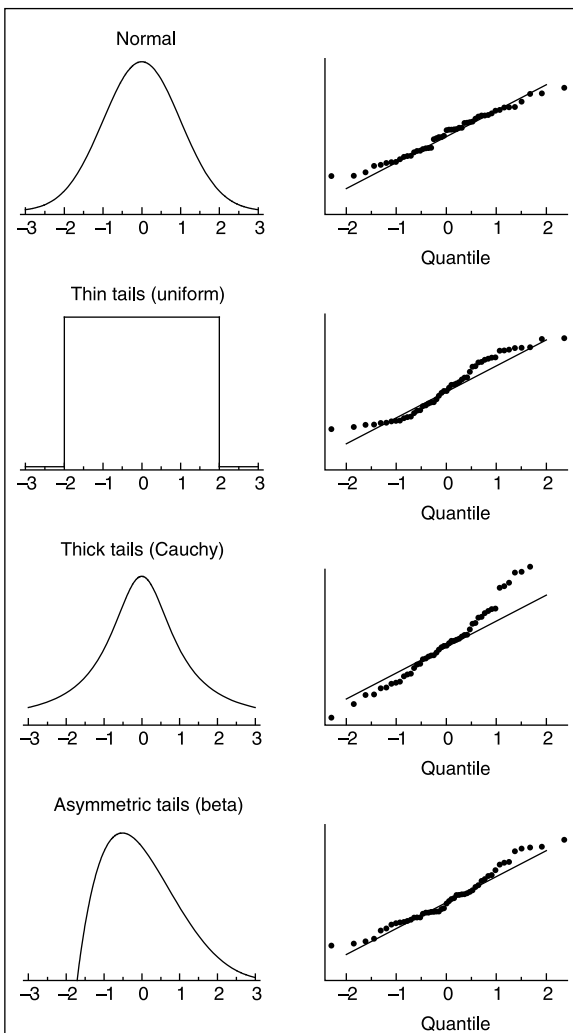
To illustrate, the values in Figure 13.6 were generated by sampling from a normal distribution with a mean of 0 and a standard deviation of 10. We may think of them as values for the error for some model. In this sample, the mean of the 25 scores is actually 0.2 and the sample standard deviation is 11.5. The rank-ordered data values or raw scores are given in the far left column of Figure 13.6. In the next column, the value i is given for each score, which is simply a variable that indicates the rank order of the score (hence it goes from 1 to 25). In the third column, the percentile in the sample for each score is given, using the formula $F = (i - .5)/n$. The .5 is included in this formula just to avoid percentiles of 0 and 1.0. These are the scores that are plotted along the vertical axis of the normal quantile-quantile plot for each observation. Then in the final column of Figure 13.6, the z score associated with the given percentile, coming from the normal distribution, is given. Thus, for instance, the third lowest error of -14 in the distribution is the 10th percentile score in the distribution of the data (i.e., $F = (3 - .5)/25 = .10$). And in a normal distribution of data, the z score corresponding to the 10th percentile score is -1.28 . The values in this final column are plotted along the horizontal axis of the normal quantile-quantile plot and thus each observation is located by a point in the plot corresponding to its percentile in the sample (vertical axis) and what its z score would be if the distribution were normal (horizontal axis). The resulting normal quantile-quantile plot for these data is given in Figure 13.7.

If the data were perfectly normally distributed they would lie along a straight line in the plot. Since these data were in fact sampled from a normal distribution, the fit of the points to a straight line is exceedingly good. In Figure 13.8 we give four normal quantile-quantile plots for errors sampled from four known distributions. The first row shows another example for the normal distribution, with the original distribution on the left and the normal quantile-quantile plot on the right. The second row depicts a sample from a uniform distribution, with each error value between -2 and $+2$ being equally likely. In its normal quantile-quantile plot, the plotted points seem to fall along a straight

line in the middle of the distribution but then the slope becomes very flat at the ends. The flat slope at the ends indicates that the most extreme scores in the sample (having the highest and lowest percentiles) are not as extreme as they would be if the data came from a normal distribution. This distribution has thinner tails than data from a normal distribution. The third row contains the Cauchy distribution. Note that in the plot on the left side its tails are thicker than those for the normal distribution in the first row. Here, once again, the slope of the points in the middle of the distribution is as it should be, but now at the ends the slope gets very steep, indicating that the data values at the high and low end of the distribution are more extreme than they would be if they came from a normal distribution of data. Normal quantile-quantile plots that look like this are a cause of serious concern. Finally, the last row in Figure 13.8 depicts errors sampled from a transformed beta distribution, which in this case has a positive skew. As the normal quantile-quantile plot makes clear, this distribution has a thicker positive tail than a normal distribution (the slope of the plot becomes very steep at the upper end) and a thinner negative tail than a normal distribution (the slope becomes very flat at the lower end). Again, it is the thick tail on the positive end that would be cause for concern.

FIGURE 13.6 Empirical errors randomly sampled from a normal distribution with their normal quantiles

<i>Empirical error values</i>	<i>i</i>	<i>f_i</i>	<i>Normal quantile z(f_i)</i>
-21	1	.02	-2.05
-16	2	.06	-1.55
-14	3	.10	-1.28
-12	4	.14	-1.08
-12	5	.18	-0.92
-11	6	.22	-0.77
-10	7	.26	-0.64
-8	8	.30	-0.52
-8	9	.34	-0.41
-6	10	.38	-0.31
1	11	.42	-0.20
1	12	.46	-0.10
2	13	.50	0
4	14	.54	0.10
4	15	.58	0.20
5	16	.62	0.31
5	17	.66	0.41
5	18	.70	0.52
7	19	.74	0.64
7	20	.78	0.77
11	21	.82	0.92
13	22	.86	1.08
15	23	.90	1.28
20	24	.94	1.55
22	25	.98	2.05

FIGURE 13.7 Normal quantile-quantile plots for errors in Figure 13.6**FIGURE 13.8** Normal quantile-quantile plots for errors sampled from normal (first row), uniform (second row), Cauchy (third row), and transformed beta with skew (fourth row) distributions

Diagnosing Errors with Heterogeneous Variances

As was the case when examining whether errors are normally distributed, we have found that graphical inspection methods are generally most useful for examining whether they manifest a constant variance. Most frequently, violations of this assumption occur when the variance of the errors depends systematically on the magnitude of the predicted values, \hat{Y}_i . For instance, if we were predicting the sales price of makes of automobiles, then being off by \$1000 may be a much more substantial error when the predicted sales price is \$20,000 than when the predicted sales price is \$100,000.

Accordingly, a useful plot for determining whether errors show heterogeneous variances is to graph the errors as a function of the predicted values, \hat{Y}_i . The most common violation of the constant variance assumption in such plots, consistent with the example given in the previous paragraph, is a funnel-shaped plot in which the errors become larger in absolute value (both positive and negative errors) as the predicted values become larger. That is, the errors are more spread out (around zero) at higher predicted values. This is perhaps even easier to see if one plots not the errors themselves as a function of the predicted values, but their absolute value (or more frequently, the square root of their absolute value)—such a plot is called a “spread-location plot”). In this plot, if errors are more variable or spread out at higher predicted values, then the absolute values of the errors should become larger at higher predicted values, and the resulting plot will have a generally positive slope.

FIGURE 13.9 Bivariate scatterplots (first row), plots of residuals \times predicted values (second row), and spread-location plots (third row) for homogeneous (left) and heterogeneous (right) errors

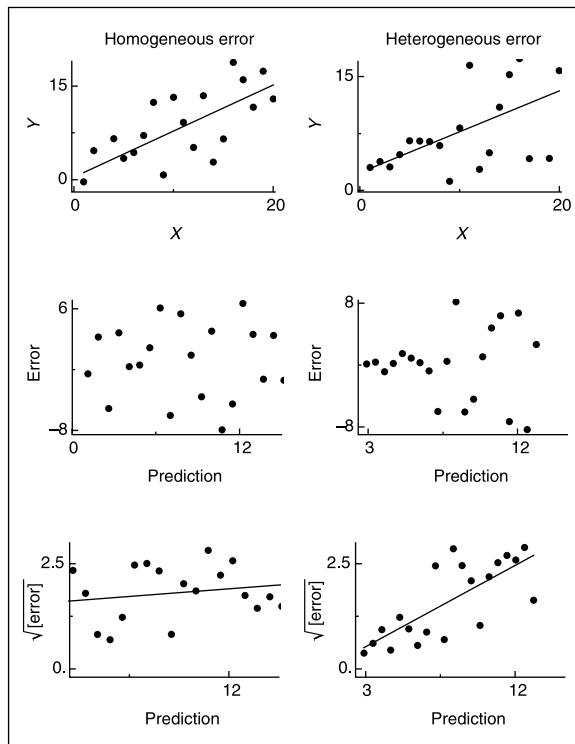


Figure 13.9 illustrates these kinds of plots with data where the errors have a constant variance (on the left) and with data where the errors were generated to be proportional to the size of the predicted values (on the right). The top row of scatterplots is simply the bivariate scatterplot, using a single predictor variable (X) to predict Y . From these plots, it is very hard indeed to see any problems. But things become clearer in the residual by predicted plots in the second row of the figure. On the left, the plotted points show a nice cloud-like structure with no noticeable funnel shape. However, such a funnel shape is much more apparent in the right-side plot where it seems that the errors are more spread out at higher values of the predicted value. The third row presents the spread-location plots. Here, it is very evident that as the predicted value increases, the average absolute error (actually its square root) also increases for the heterogeneous errors in the right column. In contrast, the absolute error for homogeneous errors shows very little increase as the predicted values increase.

While this sort of funnel-shaped pattern is typical, other patterns of Residual \times Predicted plots (or spread-location plots) may also be found and would also indicate violations of the constant variance assumption. For instance, when the Y values are proportions, there tends to be less variance in the errors at both very high predicted values (proportions near 1) and very low predicted values (proportions near 0). In such a case, the Error \times Predicted value plot may show a diamond shape, with less spread at both ends, and the spread-location plot may show a nonlinear pattern, with higher values of the absolute errors in the mid-range of the predicted values.

Resolving Violations of Normality and Constant Variance

As we suggested at the start of this chapter, it is important to deal with individual problematic observations (i.e., outliers) before addressing the more systematic violations of error assumptions (i.e., normality and constant variance). Once these more systematic problems have been diagnosed, there are various remedial actions that can be taken. In general, we will take an approach that is similar conceptually to our approach for dealing with violations of the independence assumption. There, we computed various transformations of our data variable, referring to them as W . Once we had done these transformations, combining nonindependent observations in various ways, the dependence problem was eliminated and we could proceed with our full repertoire of analytic models developed in the earlier chapters. Here, we take a similar approach. That is, we will recommend various transformations of the data variable, of the general form $W_i = f(Y_i)$, where $f(\)$ represents some function, to be defined. For different sorts of systematic violations, different functional forms will be required in order to correct the problems in the distribution of the errors. Once an appropriate transformation has been found, we can proceed with the full repertoire of analytic models developed earlier in this book, just as we did in the case of data that violate the independence assumption.

There is an alternative to the approach that we are recommending, which is widely used. It involves using what are called *nonparametric* statistical tests. We do not recommend nonparametric procedures for two primary reasons. First, it is commonly assumed that such tests make no assumptions about the error distributions. This, unfortunately, is not the case. While it is true that they effectively deal with non-normally distributed errors, they routinely make very strong assumptions about

homogeneity of variance. Thus, they do not represent the sort of generic all-purpose solution to assumption violations that they are often judged to be. Second, many nonparametric procedures are equivalent to the parametric ones we have relied upon, done not on the raw data variable but on a particular transformation of the data variable, the rank transformation. Thus, parametric procedures typically analyze a transformed score, $W_i = f(Y_i)$, where the $f(\)$ is simply the rank transform, ordering the data values and giving the lowest score a value of 1, the next lowest a value of 2, and so forth. In our judgment, other transformations are better than the rank transformation, both because they are likely to preserve more information contained in the raw data and, more importantly, because they actually can bring the data much more in line with normal distributions with constant variance. The rank transformation results in data that are uniformly, rather than normally, distributed. If one recognizes that most nonparametric procedures are simply parametric procedures applied to data that have been rank-transformed, then it becomes obvious that if there are better transformations, better in the sense that they are more likely to deal with problems of non-normality and heterogeneous variances, then these should be used instead of nonparametric statistical procedures.

Nonlinear but monotonic transformations of the data variable can often lead to more normally distributed errors with constant variances. Additionally, data that have been so transformed to alleviate assumption violations may often yield the benefit of simpler models. That is, many times higher order interactions or nonlinear effects in data are eliminated once appropriate transformations have been applied to data that otherwise violate the assumptions we have been considering. And, as always, we prefer models that are more parsimonious.

For certain kinds of data, having certain characteristic distributions, there are widely accepted transformations that generally have the desired effect of correcting assumption violations. For instance, in the case of data variables that are counts (e.g., the number of times someone does something or the number of children in a family), the appropriate transformation is generally a square root transformation, $W_i = \sqrt{Y_i}$. For data that are reaction times, the appropriate transformation is generally the log transformation, $W_i = \log(Y_i)$. For data that are proportions, the logit transformation is often the most appropriate, $W_i = \log [Y_i/(1 - Y_i)]$. And for data that are correlations, Fisher's Z transformation, $W_i = 1/2 \log [(1 + Y_i)/(1 - Y_i)]$, is generally most appropriate. Of course, while these are the routinely applied transformations in these cases, we recommend checking whether the chosen transformation has indeed resulted in errors that have corrected the original distributional problems, again by examining normal quantile-quantile plots and Residual \times Predicted plots, but this time on the residuals from the transformed data variable.

To correct problems of skew in other sorts of distributions, we recommend exploring transformations that derive from what has come to be known as the ladder of powers, $W = Y_i^\lambda$, assuming that a constant has been added to Y_i such that all values of Y_i are > 1 . If $\lambda = 1$, then of course $W_i = Y_i$ and no transformation is effected. If $\lambda > 1$, then larger values of Y_i become even larger, so that the upper end of the distribution becomes more spread out, relative to the lower end, thus reducing negative skew. And the higher the value of λ , assuming $\lambda > 1$, the more this is true. For values of $\lambda < 1$, the opposite happens: the lower end of the distribution becomes more spread out, relative to the upper end, thus reducing positive skew. Y^0 yields a value of 1 for all cases and so is not used; $\log(Y_i)$ is

the transformation that fits in the family of power transformations at that point. Thus, to reduce positive skew, one might start with the square root transformation, $W_i = \sqrt{Y_i} = Y_i^{.5}$. If positive skew remains, then one might try the log transformation. Even more extreme would be the inverse of the square root $W_i = Y_i^{-.5}$, and then the inverse $W_i = Y_i^{-1}$, and so forth. Finding the correct value of λ to reduce skew and render the distribution more like a normal distribution may be a trial and error procedure. One works either up or down the ladder of powers, depending on whether the data variable is positively or negatively skewed, examining the normal quantile-quantile plot for each transformation until the one that yields the most approximately normal distribution is identified.

Importantly, in addition to violating the assumption of normality of errors, skewed error distributions are also routinely characterized by heterogeneous variances. In the case of positive skew, for instance, when dealing with reaction times, it is generally the case that errors are more variable at longer times. Power transformations that reduce skew are thus likely to correct both the non-normality and heterogeneous variance problems.

EXAMPLE

A detailed example will help to elucidate the diagnostic and remedial tools described in this chapter. These data pertain to the relationship between the number of grants in the physical and social sciences awarded by the National Science Foundation (NSF) to various universities. These data were assembled by a colleague whose university's administration was concerned about the low number of NSF grants they received. They had decided to remedy the situation by providing extra resources to departments in the physical sciences. This colleague hoped to convince his administration that it was equally important to support social science departments because there was a strong relationship between the number of grants received in the physical and social sciences at peer institutions. The data analyzed here are the number of NSF grants of each type for the given state's two primary public universities and their self-identified peer institutions. Each university is categorized as the state's flagship or land grant university. Figure 13.10 shows the data and best-fitting regression lines for the flagship and land grant universities. Despite the apparent divergence of the slopes, the difference in slopes (i.e., a test of the interaction between university category and number of social science grants) is not significant: $F(1,15) = 1.9$, $PRE = .11$, $p = .19$. However, the plot reveals a potential outlier—a flagship university that has by far the greatest number of physical science grants—that may be distorting the regression line for the flagship universities. A visual check of the regression assumptions as well as an outlier analysis is needed.

Figure 13.11 displays the normal quantile plot (on the left) and the spread-location plot (on the right). One point, which corresponds to the apparent outlier in Figure 13.10, is far away from the normality line in the normal quantile plot; its steep slope relative to the other points identifies this as a “thick-tail” normality violation that could have substantial impact on the analysis. The spread-location plot clearly shows that the square root of the absolute value of the residuals is increasing with the size of the predicted values from the regression, thus the assumption of homogeneity of variance is violated. The studentized deleted residual for the unusual observation is 6.78, its lever is .32, and its Cook's D is 1.33, which is about twice as large as the next value of Cook's D .

FIGURE 13.10 Relationship between NSF physical and social science grants at flagship (○) and land grant (●) peer institutions

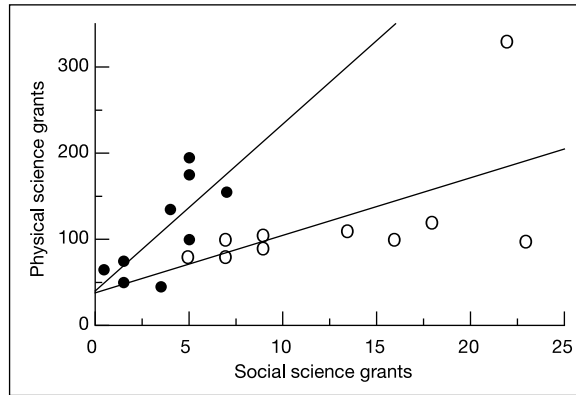
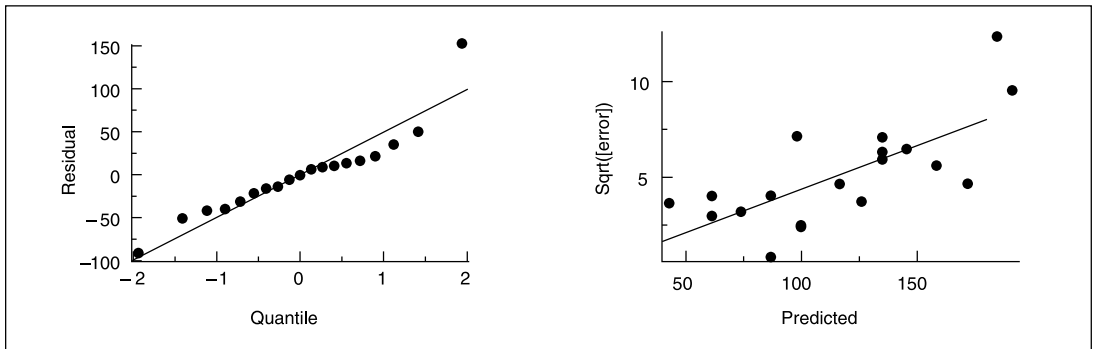


FIGURE 13.11 Normal quantile plot (left) and spread-location plot (right) for grants data of Figure 13.10



Its lever does not identify it as a particularly unusual observation in terms of its predictor values. However, the large studentized deleted residual ($p < .0001$) suggests that this university is telling a very different story about the relationship between the number of physical and social science NSF grants; the large value of Cook's D indicates that it is having a disproportionate effect on the overall model. If the outlier is omitted, the interaction is significant: $F(1,14) = 15.38$, $PRE = .53$, $p = .0015$. In other words, the story we tell about whether the relationship is different for flagship and land grant universities depends entirely on whether we include this one university. We should not allow one observation to dominate the story we tell about *all* the data! In this particular case, the unusual university has by far the highest total number of grants. And its number of physical science and social science grants fits neither the pattern of the land grant universities nor the pattern of the other flagship universities. In retrospect, this might not be a legitimate peer institution; it may have been included, wishfully, as a peer only because it was in the same intercollegiate athletic conference. For all these reasons, it is appropriate to conduct an analysis with that outlier removed. If the analysis changes appreciably without that university, that of course would not prove any of our post hoc suppositions above. Instead, those suppositions might provide hypotheses to be explored in a larger study including all the state universities.

Removing the one clear outlier does not repair the violation of homogeneity of variance apparent in the right panel of Figure 13.11 (the new graph is not presented here). Hence, a transformation may be appropriate. Also, there are a priori reasons for anticipating the need for a transformation of these data. The scale for the number of grants is not likely to be linear with an underlying scale of institution quality. For example, is the functional difference between, say, 5 and 10 grants equivalent to the difference between, say, 105 and 110 grants? We are likely to judge the second difference to be negligible while considering the first difference to be quite large. Analyzing the raw data implicitly treats these two differences as if they were equal. Also, as suggested above, counts are likely to require a square root (or similar) transformation for models of counts to have errors with homogeneous variance. Although one need not transform on both sides, it seems appropriate in this case to transform both the criterion (number of physical science grants) and the predictor (number of social science grants). Figure 13.12 shows the normal quantile plot of the residuals and the spread-location plot after the square root transformation has been applied with the outlier omitted (it remains an outlier, although not as extreme, after the transformation). Both plots suggest that the analysis of the transformed data without the outlier reasonably satisfies the normality and homogeneity of variance assumptions. Any weaker power transformation (i.e., a power between .5 and 1) leaves a positive slope in the spread-location plot, whereas any stronger transformation (such as the log or the inverse) induces a negative slope. Hence, the square root transformation (power = .5) is best for correcting the variance problems in these data.

An analysis of the transformed data reveals a second problematical observation—a land grant university that has 56 NSF grants in the physical sciences but none in the social sciences. Some land grant universities have remained closer to their roots as agricultural and mechanical universities and so do not have the full complement of social science departments; this university may be such an instance. Fortunately, the story does not change appreciably if this observation is included or omitted. So, we will stop with the analysis of the square root transformed data with the first outlier omitted; this final analysis is depicted in Figure 13.13. There is no evidence for a relationship between the number of physical and social grants at flagship universities (slope = 0.46, $F(1,14) = 0.6$, PRE = .04, $p = .45$) but there is a relationship for land grant universities (slope = 2.4, $F(1,14) = 15.65$, PRE = .53, $p = .0014$). The difference between the two slopes is statistically significant (slope difference = 1.93, $F(1,14) = 5.15$, PRE = .27, $p = .04$). (Omitting the second problematical observation would strengthen the land grant

FIGURE 13.12 Normal quantile and spread-location plots after square root transformation of grants data

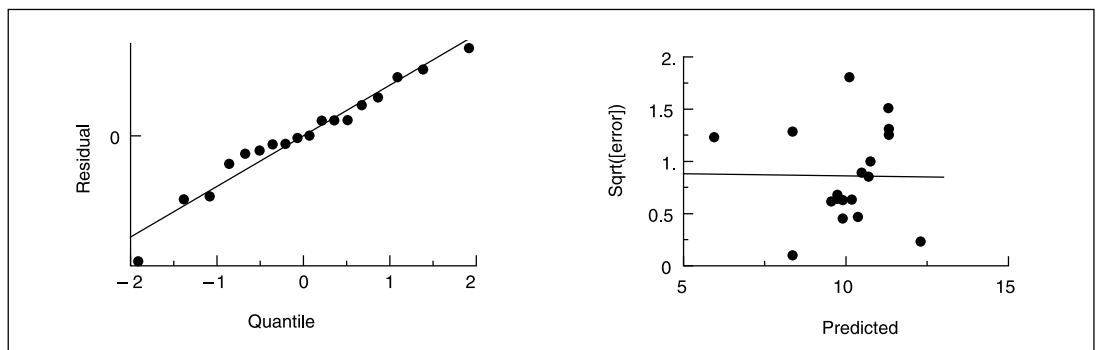
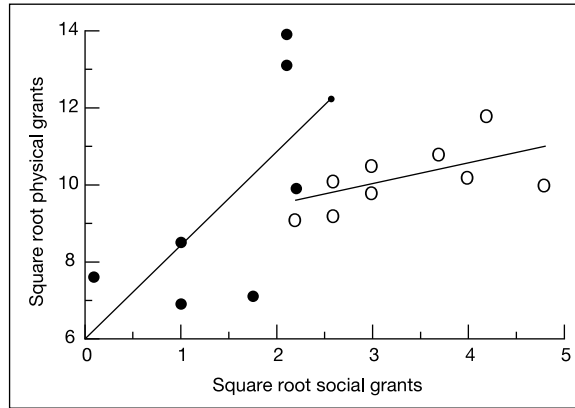


FIGURE 13.13 Final analysis of square-root-transformed data with one outlier omitted for flagship (○) and land grant (●) peer institutions



relationship and enhance the slope difference.) Transforming the original data and omitting the outlier yielded not only an analysis that satisfied the important statistical assumptions but also a clear and consistent story for these data. Including the outlier and not transforming the data produced an analysis that violated the major assumptions underlying the analysis and produced a muddled and inconsistent story for the data.

SUMMARY

In this chapter we have examined two important topics. The first concerns how one identifies and deals with observations that are outliers. We have seen that outliers can be defined in three ways: observations with unusual values on the data or outcome variable, observations with unusual values on the predictors, and observations that are unusual on both. Generally the first sort of outlier results in inflated error terms and Type II errors; the second tends to result in an overabundance of Type I errors; and the third sort is especially pernicious with outlying observations having the potential to seriously bias parameter estimates. We recommend examining the studentized deleted residual to identify outliers of the first sort, the lever to identify outliers of the second sort, and Cook's D to identify outliers of the third sort. Outliers that result from obvious errors in recording or coding data should be corrected if possible. Otherwise, analyses should be conducted without them. One should not let a few pernicious observations dictate the story and distort it from what it would be without them.

The second half of this chapter discussed graphic displays that can be used to detect violations of the assumptions that residuals are normally distributed with a constant variance. These displays include the normal quantile-quantile plot for detecting non-normality and Residual \times Predicted and spread-location plots for detecting heterogenous errors. We concluded by discussing transformations that can be used to deal with the problems of errors that display these assumption violations.

Note

- 1 But remember that it is the distribution of the errors that one wants to plot, not the distribution of the Y variable.

Logistic Regression

Dependent Categorical Variables

14

In the prior chapters we have considered models with both categorical and continuous predictor variables. However, all the dependent or outcome variables have been continuous until now. Sometimes the dependent variable is an either/or categorization. For example, the variable we want to predict might be won or lost the contest, has an ailment or not, voted or not in the last election, or graduated from college or not. There might even be more than two categories such as voted Republican, Democrat, or Independent or never smoker, former smoker, or current smoker. On the whole, continuous outcome variables are preferred to categorical dependent variables when both are available. For example, more information is available modeling exam scores on a continuous scale than modeling the pass/fail outcome of those scores. However, sometimes a continuous measure is not available so instead we must model a categorical outcome variable.

Unlike the models considered before, there are special problems when the dependent variable is dichotomous (having two outcome values). Just as we provided numerical codes for categorical predictor variables, we can assign numerical codes of, say, 0 and 1, or -1 and $+1$ to the two categories of a dependent variable, such as 0 if someone does not have a specific illness and 1 if they do. Alas, however, the procedures we have developed so far for testing model comparisons do not work for categorical dependent variables for several reasons. A simple model would be:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

If the only possible values for Y are 0 and 1, how could we interpret predicted values of, say, $\hat{Y}_i = 0.74$ or perhaps $\hat{Y}_i = 1.22$? Furthermore, with possible values being only 0 and 1 it would be difficult for the errors ($e_i = Y_i - \hat{Y}_i$) to have a normal distribution, as assumed for the statistical tests to be valid. The solution is to model the probability or odds of a particular outcome instead of trying to predict the outcomes of 0 or 1.

Categorical Predictor

As an example, consider Figure 14.1, which displays the results of a health department survey after an outbreak of an intestinal illness in a community. Of the 111 people surveyed, 55 reported being sick with the illness and 56 reported being well. So the probability of being sick (i.e., having a score of 1 on the dependent variable) is $p = 55/111 = 0.495$ and the probability of being well (having a score of 0 on the dependent variable) is $1 - p = 56/111 = 0.505$. The odds of being sick is the ratio of these two probabilities, that is,

FIGURE 14.1 Number of respondents who were sick or well and their drinking water source

		Drinking water source		
		Tap	Bottled	Total
Health status	Sick	49	6	55
	Well	32	24	56
	Total	81	30	111

$$\text{odds}_{\text{sick}} = \frac{p}{1-p} = \frac{0.495}{1-0.495} = 0.98$$

Odds near 1.0 imply the two probabilities are approximately equal, while those more extreme (very small or very large) imply one of the two outcomes is much more likely than the other.

In terms of odds, it is easy to specify the same kinds of models we have used throughout to ask questions about our data. For example, we can compare these two models to ask whether the current odds of being sick with the intestinal illness differs from the historical average odds of 0.02 using:

$$\text{Model A: } \text{odds}_{\text{sick}} = \beta_0 + \varepsilon_i$$

$$\text{Model C: } \text{odds}_{\text{sick}} = 0.02 + \varepsilon_i$$

Or we might ask whether being sick was a function of the source of one's water (bottle or tap). We might code the variable Water as +1 for bottle water and -1 for tap, then comparing these models ask whether the odds of being sick are different for the two groups:

$$\text{Model A: } \text{odds}_{\text{sick}} = \beta_0 + \beta_1 \text{Water}_i + \varepsilon_i$$

$$\text{Model C: } \text{odds}_{\text{sick}} = \beta_0 + \varepsilon_i$$

Indeed, all the modeling techniques we have used before are applicable here. For example, to ask whether the odds of being sick as a function of water source depended on the person's age, we could compare these two models:

$$\text{Model A: } \text{odds}_{\text{sick}} = \beta_0 + \beta_1 \text{Water}_i + \beta_2 \text{Age}_i + \beta_3 \text{Water}_i * \text{Age}_i + \varepsilon_i$$

$$\text{Model C: } \text{odds}_{\text{sick}} = \beta_0 + \beta_1 \text{Water}_i + \beta_2 \text{Age}_i + \varepsilon_i$$

We have two additional problems to solve when modeling odds as the dependent variable. The first is heteroscedasticity—the errors depend on the magnitude of the odds. Odds near 1 to 1 (0.50/0.50 probabilities) have much bigger errors than more extreme odds, say near 1 to 9 or 9 to 1. To understand this issue intuitively, consider what happens to, say, a 5% change in the odds ratio near 1.0 compared to more extreme odds ratios. If the odds ratio is 1.0, then the probabilities are 0.5, and 0.5. A 20% increase in the odds ratio to 1.20 would correspond to probabilities of 0.545 and 0.455. However, if the original probabilities were 0.9 and 0.1 for an odds ratio 9, then a 20% increase to 10.8 would correspond to probabilities of 0.915 and 0.085, a much smaller change in the probabilities. Hence, extreme probabilities are more stable (i.e., have less error) than middle probabilities. The solution to this heteroscedasticity problem is the log transformation because it makes the same percentage changes equivalent no matter what

the starting value of the odds or the probabilities. So in the above models instead of modeling *odds* as the dependent variable, we will use *log odds*, also known as the *logit*, defined as:

$$\text{logit}_i = \log\left(\frac{p_i}{1-p_i}\right) = \log(\text{odds}_i)$$

We will explain later how after we estimate our model for the *logits*, we can convert the results to more understandable terms of odds and probabilities.

The last problem to solve is that because we have made so many transformations to get to the *logits* that we model, the logic of minimizing the sum of squared errors no longer is appropriate. Instead, we will use *maximum likelihood* procedures, introduced in Chapter 3 and discussed in Chapter 12, to estimate the models, and we will use the χ^2 (“chi-squared”) statistic instead of *t* and *F* to evaluate the model comparisons. Fortunately, the meta-logic remains the same. We express a model comparison in terms of Models A and C for the *logits* and then use χ^2 and its probability (assuming Model C to be true) to evaluate whether the extra complexity of Model A is worth it in terms of reducing error. This technique, because it models *logits*, is often referred to as “logistic regression.”

For the simple example of predicting sickness from the source of drinking water, we can work out the details of the model comparison. This is not possible for more complicated models for which we will need to rely on the computer to do the estimation. For the simple case, the model comparison is:

$$\text{Model A: } \text{logit}_{\text{sick}} = \beta_0 + \beta_1 \text{Water}_i + \varepsilon_i$$

$$\text{Model C: } \text{logit}_{\text{sick}} = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = 0$$

To compare these two models we need to find the parameter estimates giving the maximum likelihoods for each model and then compare the maximum likelihoods of both models. Conceptually, this is essentially the same as finding the parameter estimates giving the minimum sum of squared errors and then comparing those two sums of squared errors. Only the underlying mathematics and estimation algorithms differ. Just as using a full set of codes yielded the cell means as the predicted values in ANOVA, so too using a full set of codes in logistic regression yields the cell probabilities as the predicted values. For Model C, the predicted probability of being sick is $\hat{p} = 55/111 = 0.495$, for which the predicted log odds are $b_0 = \text{logit}_{\text{sick}} = \log [0.495/(1 - 0.495)] = -0.018$. For Model A, for those drinking tap water (−1), the predicted probability of being sick is $\hat{p} = 49/81 = 0.605$, corresponding to $\text{logit}_{\text{sick}} = \log [0.605/(1 - 0.605)] = 0.426$; and for those drinking bottled water (+1), the predicted probability of being sick is $\hat{p} = 6/30 = 0.2$, corresponding to $\text{logit}_{\text{sick}} = \log [0.2/(1 - 0.2)] = -1.386$. Then it is easy to compute the parameter estimates for Model A: b_0 is the average of the two predicted cell values and b_1 is the distance of each cell value from this average, or, equivalently, half the distance between the two cell values (assuming codes of +1 and −1). That is:

$$b_0 = \frac{0.426 - 1.386}{2} = -0.48$$

$$b_1 = -0.426 - 0.48 = -1.386 - (-0.48) = \frac{-0.426 + (-1.386)}{2} = -0.906$$

To summarize,

$$\text{Model A: } \log\hat{it}_{sick} = -0.48 - 0.906\text{Water}_i$$

$$\text{Model C: } \log\hat{it}_{sick} = -0.018$$

In this simple case, we know these are the maximum likelihood estimates for the two models. Now it remains to compute those maximum likelihoods so that we can compare the two models. For binary outcomes with probability p of an event (being sick in our example) and k events in N cases, the likelihood equals:

$$L = p^k(1 - p)^{N-k}$$

This is usually a very small number so we normally use

$$-2 \log(L) = -2 \log (p^k(1 - p)^{N-k})$$

both because it will be a more reasonably sized number and will approach having a chi-squared distribution as the number of observations N increases. So the “deviance” for Model C is given by

$$\begin{aligned} -2 \log(L) &= -2 \log (p^k(1 - p)^{N-k}) \\ &= -2 \log (0.495^{55} * 0.505^{56}) \\ &= 153.87 \end{aligned}$$

For Model A we multiply the likelihoods of the two groups (tap water and bottled water) to obtain the deviance as:

$$\begin{aligned} -2 \log(L) &= -2 \log [(0.2^6 * 0.8^{24}) * (0.605^{49} * 0.398^{32})] \\ &= 138.719 \end{aligned}$$

Just as with the case for computing the sums of squared errors, we will generally depend on computer programs to compute these deviances or model fits. For more complicated models, we *must* rely on the computer algorithms. For example, Figure 14.2 shows the output for the logistic regression model for the water data. Note that the coefficients for the logistic model are exactly those we computed for Model A above. Also, the “Null Deviance” equals 153.87, corresponding to Model C so we will denote it as DEV(C), and the “Residual Deviance” equals 138.72, corresponding to Model A, which we will denote as DEV(A). The degrees of freedom are computed exactly as they were for models compared with sums of squared errors. For Model C, there is one parameter so $df(C) = N - 1 = 111 - 1 = 110$ and for Model A there is one extra parameter so

FIGURE 14.2 Basic output provided by most logistic regression programs

	<i>Estimate</i>	<i>Standard Error</i>	<i>Wald Chi-Sq</i>	<i>Pr(> z)</i>
(Intercept)	-0.4801	0.2549	3.546	0.059679
water	-0.9062	0.2549	12.634	0.000379
Null deviance:	153.87 on 110 degrees of freedom			
Residual deviance:	138.72 on 109 degrees of freedom			

$df(A) = N - 2 = 111 - 2 = 109$. Finally, we are ready to compare Model C and Model A based on the following information:

$$\text{Model A: } \widehat{\text{logit}}_{\text{sick}} = -0.48 - 0.906 \text{Water}_i \quad PA = 2 \quad DEV(A) = 138.72$$

$$\text{Model C: } \widehat{\text{logit}}_{\text{sick}} = -0.018 \quad PC = 1 \quad DEV(C) = 153.87$$

Very similar to how we compute SSE when the dependent variable was continuous, we compute the reduction in the deviance as:

$$DEV(C) - DEV(A) = 153.87 - 138.72 = 15.15$$

Mathematical statisticians have shown that this difference, as N increases, asymptotically approaches a chi-square distribution with degrees of freedom equal to $PA - PC$. Just as we used the F distribution to determine whether the extra complexity of Model A was worth it, we can use the chi-square distribution to decide whether the extra complexity of Model A is worth it in logistic regression. Figure A.5 in the Appendix provides the critical values for a variety of probabilities and values of $PA - PC$. In this case, we want to compare the obtained chi-square value of 15.15 to the critical value for $\alpha = 0.05$ and $PA - PC = 1$. That critical value is 3.84 and arises so frequently in logistic regression that it should be memorized. Looking across the row in Figure A.5 we can see that if the null hypothesis were true ($\beta_1 = 0$), obtaining a chi-square value in the data as high as 15.15 would be extremely rare. We therefore reject Model C in favor of Model A: the source of one's drinking water was related to the probability of being ill, with those drinking tap water more likely to be ill.

Just as we computed PRE by comparing the achieved improvement $SSR = SSE(C) - SSE(A)$, we can compare the deviances. In this case, the proportional reduction in deviance (PRD) equals:

$$PRD = \frac{DEV(C) - DEV(A)}{DEV(C)} = \frac{153.87 - 138.72}{153.87} = \frac{15.15}{153.87} = 0.098$$

There are other ways of generalizing the PRE concept to logistic regression and they have slight differences. Statisticians often refer to one of these measures as being "pseudo- R^2 ." However, for our purposes, PRD provides a useful measure, like PRE , of how much better Model A is than Model C.

Now that we know that Model A, predicting different sickness rates conditional on water source, is better than Model C, we need to interpret Model A. We know the predicted values for Model A and we used them above to derive the parameter estimates. Such derivations will not be possible for more complicated models so we will go through the exercise of interpreting the coefficients as an example of what to do with more complicated models. First we compute the predicted logits of being sick for both water groups:

$$\begin{aligned} \text{Bottle: } \text{logit} &= -0.48 - 0.906 \text{Water}_{\text{bottle}} \\ &= -0.48 - 0.906(+1) \\ &= -1.386 \end{aligned}$$

$$\begin{aligned}\text{Tap: } \logit &= -0.48 - 0.906\text{Water}_{\text{tap}} \\ &= -0.48 - 0.906(-1) \\ &= 0.426\end{aligned}$$

Most of us do not think very well in terms of *logits* so the next step is to convert the predicted *logits* to predicted odds using:

$$\begin{aligned}\text{Bottle: } odds_{\text{sick}} &= e^{\logit} = e^{-1.386} = 0.25 \\ \text{Tap: } odds_{\text{sick}} &= e^{\logit} = e^{0.426} = 1.53\end{aligned}$$

That is, the odds of someone being sick who had been drinking bottled water was only 1 in 4, while the odds for someone who had been drinking tap water was about 1.5 to 1 or 3 to 2. For every two people in the tap water group who were well, there were three people who were sick. It is often useful to report the odds ratio:

$$\frac{odds_{\text{tap}}}{odds_{\text{bottle}}} = \frac{1.53}{0.25} = 6.12$$

That is, someone drinking tap water was 6.12 times as likely to be sick as someone drinking bottled water. This clearly points to tap water as being the source of the illness, but like all non-experimental data we cannot know for sure. However, the survey results did motivate the city to check the water treatment plant where they found a malfunction.

In many disciplines, such as medicine and public health, results are discussed in terms of odds and odds ratios. However, we can also return to probabilities using these computations:

$$\begin{aligned}\text{Bottle: } \hat{p} &= \frac{odds_{\text{sick}}}{1 + odds_{\text{sick}}} = \frac{0.25}{1 + 0.25} = 0.20 \\ \text{Tap: } \hat{p} &= \frac{odds_{\text{sick}}}{1 + odds_{\text{sick}}} = \frac{1.53}{1 + 1.53} = 0.605\end{aligned}$$

Knowing that Model A, which allowed separate probabilities of being sick conditional on water source, is better than Model C, which had a single probability of being sick, we also know that these two probabilities, log odds, and *logits* are significantly different.

Continuous Predictor

Logistic regression uses the same kind of modeling techniques as we have seen previously for least-squares regression. Models may contain continuous variables, coded-categorical variables, and products of those variables. We will first consider a single continuous predictor and then illustrate more complex models. In all these models we will depend on the computer to perform the complex maximum likelihood estimations.

Colleges and universities are concerned about their graduation rates. A success is usually defined as graduation within six years of entry into the college or university. To make admission decisions, many schools compute a predicted grade point average (PGPA) for the student's first year, using test scores, quality of high school, high school GPA, etc. as predictors in a multiple regression model of the kind we have seen before.

Although PGPA is designed only to predict first-year success, it seems plausible that it might also predict graduation success under the theory that better students are more likely to be successful students. We can test this theory by comparing these two models:

$$\text{Model A: } \text{Grad}_i = \beta_0 + \beta_1 \text{PGPA}_i + \varepsilon_i$$

$$\text{Model C: } \text{Grad}_i = \beta_0 + \varepsilon_i$$

The question is whether knowing a student's PGPA would improve the predicted probability that the student would graduate. Figure 14.3 displays common statistics from a logistic regression of a very large dataset (15,790 observations) from a major research university, summarized in our terms as:

$$\text{Model A: } \widehat{\text{logit}}_{\text{grad}_i} = -3.17 + 1.33 \text{PGPA}_i \quad PA = 2 \quad DEV(A) = 19398$$

$$\text{Model C: } \widehat{\text{logit}}_{\text{grad}_i} = 0.695 \quad PC = 1 \quad DEV(C) = 20080$$

FIGURE 14.3 Basic output provided by most logistic regression programs using predicted grade point average (GPA) to predict graduation within six years

	<i>Estimate</i>	<i>Standard Error</i>	<i>Wald Chi-Sq</i>	<i>Pr(> z)</i>
(Intercept)	-3.16996	0.15229	433.47	<2e-16
PGPA	1.32960	0.05246	642.12	<2e-16
Null deviance:	20080 on 15789 degrees of freedom			
Residual deviance:	19398 on 15788 degrees of freedom			

The overall graduation probability in these data is 0.667. The question is whether knowing a student's PGPA improves the predicted probability that the student would graduate. The answer is yes—Model A is significantly better than Model C: $\chi^2 = 20080 - 19398 = 682$, $p < 0.0001$, but not by a lot:

$$PRD = \frac{20080 - 19398}{20080} = 0.034$$

The interpretation of the Model A in terms of *logits* is exactly the same as for simple regression. The intercept -3.17 is the predicted *logit* when $\text{PGPA} = 0$ and the slope 1.33 is the expected increase in the *logit* for each unit increase (i.e., each grade level) in PGPA. In more familiar terms, $e^{-3.17} = 0.04$ is the odds that a person with $\text{PGPA} = 0$ will graduate and the predicted probability of graduating within six years is $\hat{p} = 0.04/(1+0.04) = 0.04$. It seems unlikely that a student with $\text{PGPA} = 0$ would be admitted so, just as with least-squares models, it is useful to use mean deviation to render the intercept in the model more interpretable. The new model, using $\text{PGPA}_{dev} = \text{PGPA} - \overline{\text{PGPA}}$, is:

$$\text{Model A: } \widehat{\text{logit}}_{\text{grad}_i} = 0.732 + 1.33 \text{PGPA}_{dev}_i \quad PA = 2 \quad DEV(A) = 19398$$

$$\text{Model C: } \widehat{\text{logit}}_{\text{grad}_i} = 0.695 \quad PC = 1 \quad DEV(C) = 20080$$

Unfortunately, the estimates of the intercepts in these two models are not identical as they would be with least-squares regression models. The slight difference is because maximum likelihood algorithms do not find an exact solution but instead search until the solutions converge to a result that is “close enough” to the maximum likelihood. In this case, the differences are trivial: from Model A $e^{0.732} = 2.08$ with a predicted graduation probability of $\hat{p} = 2.08/(1+2.08) = 0.675$ and from Model C $e^{0.695} = 2.00$ with a predicted graduation probability of $\hat{p} = 0.667$. Such slight differences are common when using logistic regression and its estimation procedures. These are the predicted values for the odds and probabilities of graduating when the *PGPA* is at its mean value.

The slope of 1.33 was the same whether using the *PGPA* model or the mean-deviated model. It says that for every one-unit increase (i.e., a one-level grade increase, C to B, B to A, etc.) in *PGPA*, the predicted *logit* increases by 1.33 units. Converting to odds, this is the same as saying the odds of graduating increase by a factor of $e^{1.33} = 3.78$ for each grade-level increase in *PGPA*. When working with odds, addition is replaced by multiplication. Hence, an increase of two grade-levels in *PGPA* would imply an increase in the odds of graduation by a factor of $3.78^2 = 14.29$. We cannot make easy statements about the increase in the probability of graduation as *PGPA* increases because it is necessarily non-linear. It is easiest to see all these relationships graphically. Figure 14.4 shows graphs of Models A and C for *logits*, odds, and probabilities. Clearly, the predicted *logit*, the odds, and the probability of graduating all increase as *PGPA* increases.

For least-squares models, the statistics for the coefficients (their standard errors and associated *t* statistics) provided the same information as a model comparison in which the parameter was set to zero in Model C. For logistic regression, the two tests are not quite the same. For logistic regression, squaring the ratio of the coefficient to its standard error provides a chi-squared test for the coefficient. For *PGPA*,

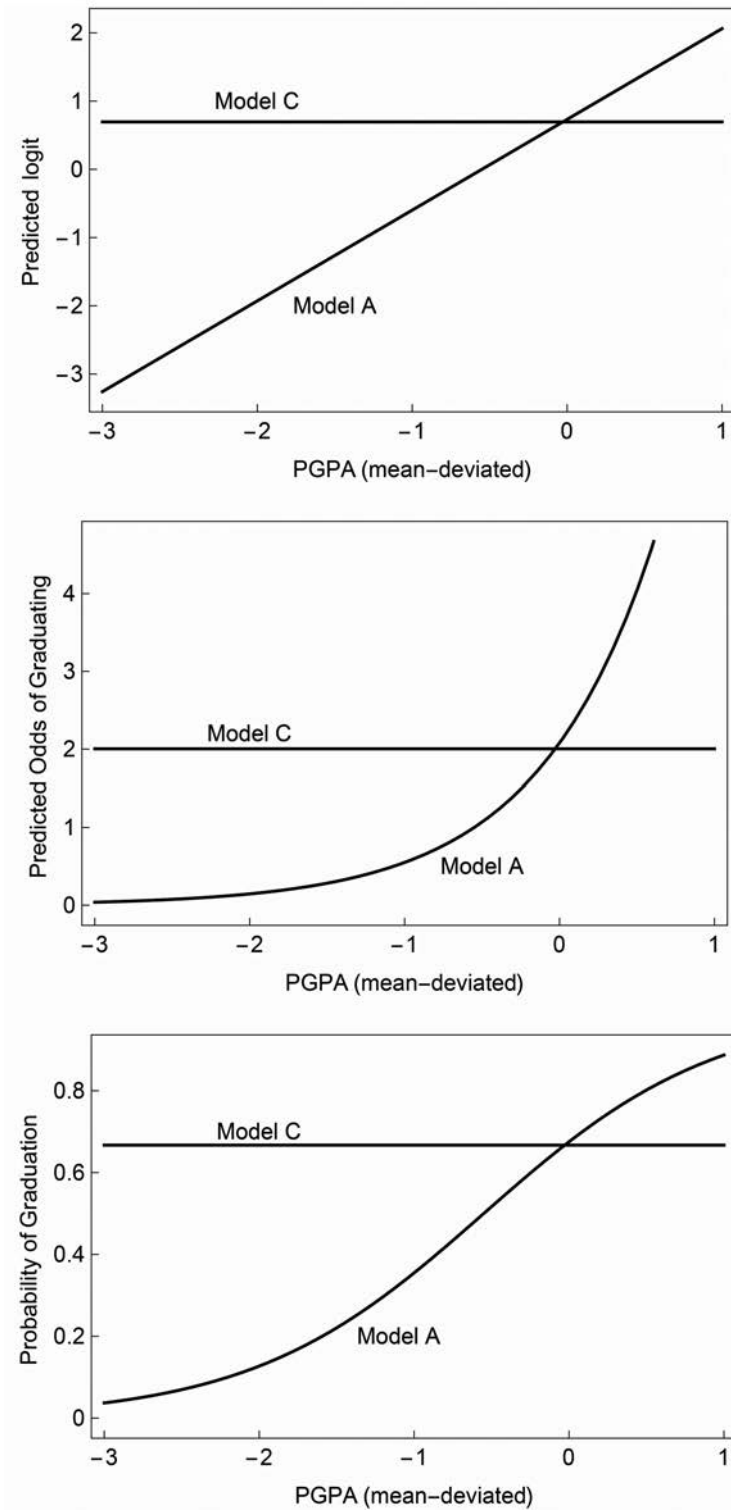
$$\chi_1^2 = \left[\frac{b_1}{s_{b_1}} \right]^2 = \left[\frac{1.3296}{0.05246} \right]^2 = 642.12$$

The difference in the deviances of the two models ($\chi^2 = 682$) and the squared ratio of the coefficient to its standard error ($\chi_1^2 = 642.12$) estimate the same chi-squared value asymptotically. However, even with almost 16,000 observations in this case, they seldom converge exactly. Almost always they yield the same conclusion in terms of deciding whether the coefficient is significantly different from zero. To distinguish the two, the squared ratio of the coefficient and its standard error are known as the “Wald chi-square.”

The standard error of the coefficient is also useful in computing confidence intervals in logistic regression. Instead of using the critical value of *t* as in the least-squares confidence intervals, we use the square root of the critical value of chi-squared, which is $\sqrt{\chi_{1,\alpha=0.05}^2} = \sqrt{3.84} = 1.96$. So the 95% confidence interval for the logit slope is

$$\begin{aligned} b_j \pm 1.96s_{b_j} &= 1.329 \pm 1.96(0.05246) \\ &= 1.329 \pm 0.103 = [1.226, 1.432] \end{aligned}$$

which can easily be converted to odds using $[e^{1.226}, e^{1.432}] = [3.41, 4.19]$. The best estimate of the change in odds of graduating for a one grade-level increase was 3.78, but those odds might reasonably be as low as 3.41 or as high as 4.19. The large number of observations provides a rather narrow confidence interval for the odds. Note that while the null hypothesis for the logit is that the slope equals zero, this corresponds to the null

FIGURE 14.4 Plots of the graduation models in terms of logits, odds, and probabilities

hypothesis for the odds being one because $e^0 = 1$. Thus, determining whether 1.0 is inside or outside the 95% confidence interval is the same as the statistical test of whether the *logit* slope equals zero.

More Complex Models

With the basic tools of logistic regression outlined above we can compare the same types of more complex models that we investigated using least-squares methods. We provide a few examples here, again using the graduation rate data. It is easy to generalize these procedures to other types of models. There are additional variables in the graduation dataset that might be useful for predicting who graduates. The data are from a state university so one potentially useful predictor is the student's residency (RES). In-state students have lower tuition and proximity to social support from family and friends whereas out-of-state students have higher tuition and are farther from social support of family and friends. Hence, in-state students might be more likely to graduate. There are beliefs that women are more diligent in their college classes than men so gender (GEN) might be a useful predictor of graduation. College administrators believe that students who arrive with a plan, including a preferred major, are more likely to graduate than students who select "Open Option" (OPEN) as their major at entry. Figure 14.5 lists a number of models using these variables, their parameter estimates, and their deviances. Comparisons between these models allow us to ask a number of questions about graduation rates.

Simple Logistic Regression

Above we asked whether predicted GPA (PGPA) was a useful predictor of graduation rates. In terms of Figure 14.5, asking this question is equivalent to comparing Models 1 and 2. Model 1 is the simple model and Model 2 adds a continuous predictor so it corresponds to simple regression. In this case, $\chi_1^2 = Dev_1 - Dev_2 = 20080 - 19398 = 682$. Although significant ($p < 0.001$), the magnitude of the relationship is not large because the proportional reduction in deviance is small:

$$PRD = \frac{Dev_1 - Dev_2}{Dev_1} = \frac{20080 - 19398}{20080} = 0.034$$

More complicated model comparisons are addressed in the same way.

Multiple Logistic Regression

Comparing the simple model (Model 1) to a model with all the predictors (PGPA, Dev, Res, Gen, and Open in Model 3) corresponds to multiple regression. In this comparison, $\chi_5^2 = Dev_1 - Dev_3 = 20080 - 19353 = 727$. Again, this improvement of Model 3 over Model 1 is definitely significant ($p < 0.001$, note that the degrees of freedom = $PA - PC = 4$). However, what would correspond to R^2 is not large:

$$PRD = \frac{20080 - 19353}{20080} = 0.036$$

FIGURE 14.5 Logistic models predicting graduation rates**Model 1:**

$$\text{logit}_i = \beta_0 + \varepsilon_i \quad PA = 1$$

$$\widehat{\text{logit}}_i = 0.698$$

$$Dev = 20080$$

Model 2:

$$\text{logit}_i = \beta_0 + \beta_1 PGPA Dev_i + \varepsilon_i \quad PA = 2$$

$$\widehat{\text{logit}}_i = 0.732 + 1.33 PGPA Dev_i$$

$$Dev = 19398$$

Model 3:

$$\text{logit}_i = \beta_0 + \beta_1 PGPA Dev + \beta_2 Res_i + \beta_3 Gen_i + \beta_4 Open_i + \varepsilon_i \quad PA = 5$$

$$\widehat{\text{logit}}_i = 0.716 + 1.24 PGPA Dev + 0.046 Res_i + 0.107 Gen_i + 0.01 Open_i$$

$$Dev = 19353$$

Model 4:

$$\text{logit}_i = \beta_0 + \beta_1 PGPA Dev + \beta_2 Res_i + \beta_3 Gen_i + \varepsilon_i \quad PA = 4$$

$$\widehat{\text{logit}}_i = 0.716 + 1.24 PGPA Dev_i + 0.045 Res_i + 0.108 Gen_i$$

$$Dev = 19353$$

Model 5:

$$\text{logit}_i = \beta_0 + \beta_1 PGPA Dev + \beta_2 Res_i + \beta_3 Gen_i + \beta_4 PGPA Dev * Res_i + \varepsilon_i \quad PA = 5$$

$$\widehat{\text{logit}}_i = 0.688 + 1.22 PGPA Dev_i + 0.067 Res_i + 0.108 Gen_i + 0.387 PGPA Dev_i * Res_i$$

$$Dev = 19302$$

Thus, we can conclude that, as a set, all the predictors provide a significant reduction in deviance over the simple model.

Evaluating One Predictor among Many

As with models of continuous outcome variables, we often want to ask whether a particular predictor variable is needed over and above or when controlling for other predictor variables. Comparing Model 4, which does not include the *Open* predictor, to Model 3, which includes *Open* and additional predictors, asks whether *Open* is a useful predictor over and above the other predictors. Since $\chi_1^2 = Dev_4 - Dev_3 = 19353 - 19353 = 0$, *Open* (assessing whether the student had committed to a major or not at entry) is not a useful predictor over and above predicted GPA, residency status, and gender.

Interactions

Just as we used model comparisons to evaluate interactions (products of two variables) and polynomials (products of a variable with itself) in least-squares models, we can use the same modeling techniques in logistic regression. Predicted GPA is expected to be a more potent predictor for state residents than for out-of-state students. Comparing Model 4, which includes the predictors without the interaction, to Model 5, which adds the product between $PGPADev$ and Res , assesses whether predicted GPA is indeed a more potent predictor for in-state students. Note that the same model comparison rule applies: all the components of the interaction product must be included in both models. In this case, $\chi^2 = Dev_4 - Dev_5 = 19353 - 19302 = 51$, $p < 0.001$. Thus, the model including the interaction is a significantly better model than the one that does not.

We interpret the interaction, in terms of logits, exactly as we did for least-squares models. Res was coded $+1/2$ for in-state students and $-1/2$ for out-of-state students. Substituting these values into the predicted equation for Model 5 yields:

In-State:

$$\widehat{\text{logit}}_i = 0.688 + 1.22PGPADev_i + 0.045(\frac{1}{2}) + 0.108Gen_i + 0.387(\frac{1}{2})PGPADev_i$$

$$\widehat{\text{logit}}_i = 0.711 + 1.41PGPADev_i + 0.108Gen_i$$

Out-of-State:

$$\widehat{\text{logit}}_i = 0.688 + 1.22PGPADev_i + 0.045(-\frac{1}{2}) + 0.108Gen_i + 0.387(-\frac{1}{2})PGPADev_i$$

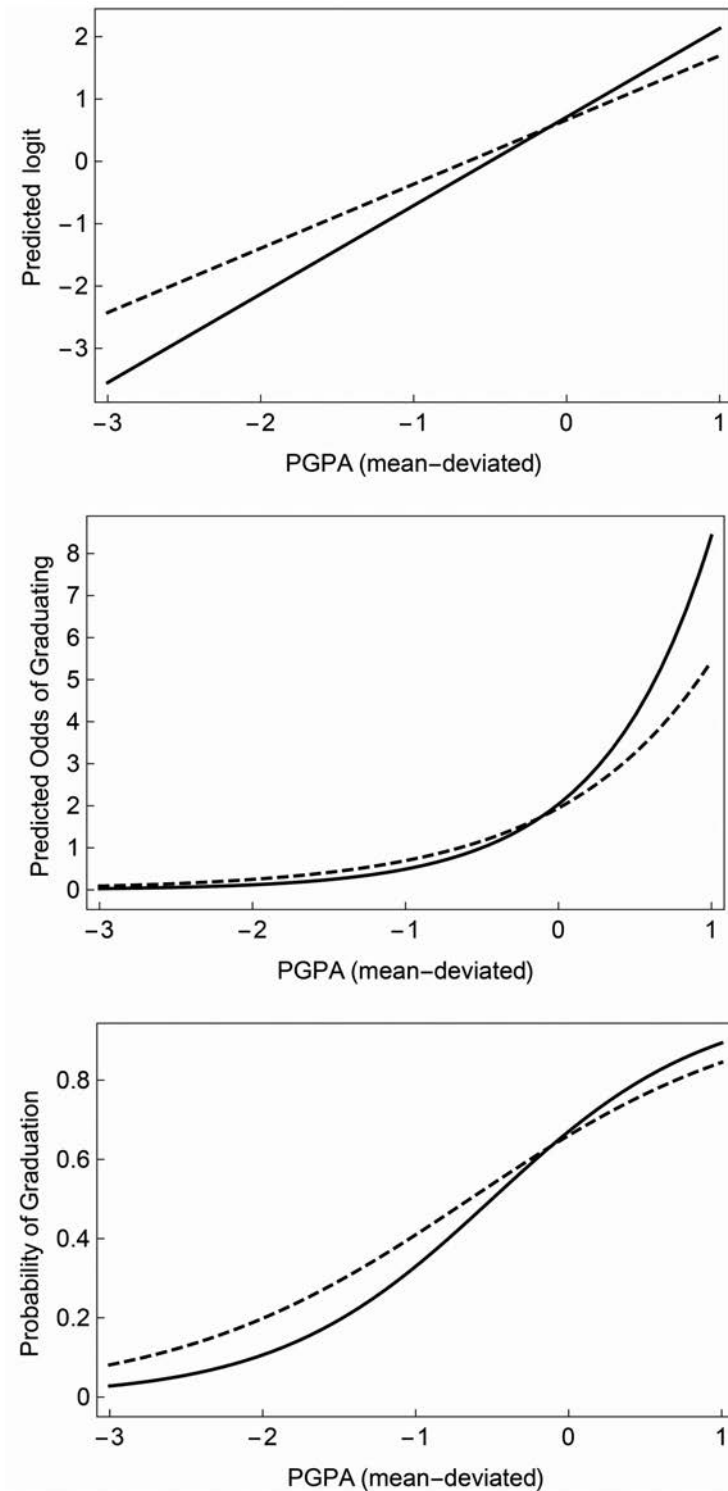
$$\widehat{\text{logit}}_i = 0.665 + 1.03PGPADev_i + 0.108Gen_i$$

The slope relating $PGPADev$ to the logit (log odds of graduating) is larger for the in-state students (1.42) than for the out-of-state students (1.03) and we know this difference is significant because of the comparison of the deviances of Models 4 and 5. The comparison of the coefficients shows how much more potent predicted grade point average is for in-state than out-of-state students. The top panel of Figure 14.6 shows the two lines (in-state solid, out-of-state dashed) for the predicted logits (substituting 0 for Gen , which was coded $+1/2$ and $-1/2$). Non-parallel lines are characteristic of an interaction. Note that the slope relating $PGPADev$ to the logit of graduation rate is steeper for in-state students. The middle panel transforms the logits to odds of graduating. Consistent with predicted GPA being more potent for in-state students, the odds curve changes more for the in-state students. Finally, the lower panel transforms to probability of graduating. In this graph the graduation rates for in-state students are lower than those for out-of-state students when predicted GPA is low but at about the mean of predicted PGA (i.e., $PGPADev = 0$), the probability of graduating for in-state students overtakes that of out-of-state students. Again, the enhanced potency for predicted GPA for in-state students is reflected in the curve that changes more rapidly than for out-of-state students.

Other Model Comparisons

The examples in Figure 14.5 are intended merely to indicate the possibilities. The data analyst using logistic regression may use the same variety of models as in least-

FIGURE 14.6 Plots of the interaction model predictions in terms of logits, odds, and probability. Solid lines are in-state students; dashed lines are out-of-state students



squares regression. Models involving combinations and products of continuous and contrast-coded categorical predictors may be used. Again, model comparisons when $PA - PC = 1$ are preferred because there is less ambiguity when interpreting the model comparison.

Outliers

Checking for unusual predictor values (i.e., observations with high leverage) is exactly the same as it was for least-squares models and described in the previous chapter. There are corresponding techniques for identifying unusual Y values and observations having unusual impact on the overall model. However, they are too complex for discussion here. Most computer programs provide tools for identifying unusual observations.

Statistical Power for Logistic Regression

Statistical power is important for the same reasons it was in least-squares regression. With low power, there will be increased chances of making Type II errors. Just as the computation of the maximum-likelihood solutions for logistic models is best left to the computer algorithms, so too the computation of statistical power for logistic regression model comparisons is best left to the computer. Most computer statistical packages provide tools for estimating power for logistic analyses. The same guidelines for enhancing statistical power apply. Collect data of better quality (less error), increase before number of observations, and increase the variability of the predictor variables. Additionally, all else being equal, the least-squares analysis of a continuous outcome variable will be more powerful than a logistic regression analysis of a dichotomous variable. So, whenever possible use a continuous outcome variable instead of a categorical outcome. For example, it would be more powerful to use actual scores from a performance test rather than a simple pass/fail measure.

SUMMARY

In this chapter, we considered models for the analysis of dichotomous categorical dependent variables, that is, dependent variables that rely on either/or categories, such as pass/fail or voted/did not vote. Such analyses present two challenges: the predicted values are difficult to interpret and the errors are unlikely to be normally distributed. Logistic regression analyses solve these problems by modeling logits rather than the individual dichotomous values of the dependent variable. What this means is that we convert the dichotomous observations to probabilities given the model and then these probabilities to odds. And finally to get to logits we take the log of those odds. It is these logits that we model. Predictor variables can be continuous, categorical, or both. Maximum likelihood estimation, as described in Chapters 3 and 12, is then used to estimate parameters and error (or deviance). The amount by which the deviance for Model A is smaller than the deviance for Model C represents the reduction in error. The proportional reduction in deviance (PRD) can also be calculated in the same way that PRE would be calculated. We also described the interpretation of parameter estimates,

which uses familiar procedures, but is more complicated because the relationship between the predictor variable and the dependent categorical variable is nonlinear. Finally, we concluded with a brief discussion of statistical power, noting that ordinary least squares analyses of a continuous dependent variable is generally more desirable than logistic regression analyses of a dichotomous dependent variable as the former is more statistically powerful.

References

- 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.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally. (Also published as *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.)
- Clark, H. H. (1973). The language-as-fixed effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.
- Cohen, J. (1978). Partialled products are interactions: Partialled powers and curve components. *Psychological Bulletin, 85*, 858–866.
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15–18.
- Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association, 74*, 169–174.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer’s dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329.
- Henderson, H., & Velleman, P. F. (1981). Building multiple regression models interactively. *Biometrics, 37*, 391–411.
- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Hoboken, NJ: John Wiley.
- Judd, C. M., & Kenny, D. A. (1981a). *Estimating the effects of social interventions*. Cambridge: Cambridge University Press.
- Judd, C. M., & Kenny, D. A. (1981b). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602–619.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54–69.
- Judd, C. M., Yzerbyt, V., & Muller, D. (2014). Mediation and moderation. In H. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 653–676). New York, NY: Cambridge University Press.
- Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99*, 422–431.

- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Taylor & Francis.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Maruyama, G., & Ryan, C. S. (2014). *Research methods in social relations* (8th ed.). Oxford, UK: Wiley-Blackwell.
- Reis, H., & Judd, C. M. (Eds.). (2014). *Handbook of research methods in social and personality psychology* (2nd ed.). New York, NY: Cambridge University Press.
- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge: Cambridge University Press.
- Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis*. Boston: McGraw Hill.
- SAS Institute (2015). *SAS/STAT software*. Cary, NC: SAS Institute, Inc. (www.sas.com/stat).
- Scheffé, H. A. (1959). *The analysis of variance*. New York: Wiley.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Appendix

FIGURE A.1 Critical values of PRE for $PA - PC = 1$

$n - PA$	α							
	.5	.2	.1	.05	.025	.01	.005	.001
1	.500	.905	.976	.994	.998	1.000	1.000	1.000
2	.250	.640	.810	.902	.951	.980	.990	.998
3	.163	.472	.649	.771	.853	.919	.949	.982
4	.121	.370	.532	.658	.753	.841	.887	.949
5	.096	.303	.448	.569	.667	.765	.820	.904
6	.079	.257	.386	.499	.595	.696	.756	.855
7	.067	.222	.339	.444	.536	.636	.699	.807
8	.059	.196	.302	.399	.486	.585	.647	.761
9	.052	.175	.272	.362	.445	.540	.602	.717
10	.047	.158	.247	.332	.410	.501	.562	.678
11	.042	.145	.227	.306	.379	.467	.526	.642
12	.039	.133	.209	.283	.353	.437	.495	.608
13	.036	.123	.194	.264	.330	.411	.467	.578
14	.033	.114	.181	.247	.310	.388	.441	.550
15	.031	.107	.170	.232	.292	.367	.419	.525
16	.029	.100	.160	.219	.277	.348	.398	.502
17	.027	.095	.151	.208	.262	.331	.379	.480
18	.026	.090	.143	.197	.249	.315	.362	.461
19	.024	.085	.136	.187	.238	.301	.346	.443
20	.023	.081	.129	.179	.227	.288	.332	.426
22	.021	.074	.118	.164	.208	.265	.307	.395
24	.019	.067	.109	.151	.192	.246	.285	.369
26	.018	.062	.101	.140	.179	.229	.266	.346
28	.016	.058	.094	.130	.167	.214	.249	.325
30	.015	.054	.088	.122	.157	.201	.234	.307
35	.013	.046	.075	.105	.135	.175	.204	.269
40	.011	.041	.066	.093	.119	.155	.181	.240
45	.010	.036	.059	.083	.107	.138	.162	.216
50	.009	.033	.053	.075	.096	.125	.147	.196
55	.008	.030	.048	.068	.088	.115	.135	.180
60	.008	.027	.044	.063	.081	.106	.124	.166
80	.006	.020	.033	.047	.061	.080	.094	.127
100	.005	.016	.027	.038	.049	.065	.076	.103
150	.003	.011	.018	.025	.033	.043	.051	.070
200	.002	.008	.013	.019	.025	.033	.039	.053
500	.001	.003	.005	.008	.010	.013	.016	.021

FIGURE A.2 Critical values of PRE for $\alpha = .05$

<i>n</i> - <i>PA</i>	<i>PA</i> - <i>PC</i>								
	1	2	3	4	5	6	7	8	9
1	.994	.997	.998	.999	.999	.999	.999	.999	1.000
2	.902	.950	.966	.975	.980	.983	.985	.987	.989
3	.771	.864	.903	.924	.938	.947	.954	.959	.964
4	.658	.776	.832	.865	.887	.902	.914	.924	.931
5	.569	.698	.764	.806	.835	.856	.872	.885	.896
6	.499	.632	.704	.751	.785	.811	.831	.847	.860
7	.444	.575	.651	.702	.739	.768	.791	.810	.825
8	.399	.527	.604	.657	.697	.729	.754	.775	.792
9	.362	.486	.563	.618	.659	.692	.719	.742	.761
10	.332	.451	.527	.582	.624	.659	.687	.711	.731
11	.306	.420	.495	.550	.593	.628	.657	.682	.703
12	.283	.393	.466	.521	.564	.600	.630	.655	.677
13	.264	.369	.440	.494	.538	.574	.604	.630	.653
14	.247	.348	.417	.471	.514	.550	.580	.607	.630
15	.232	.329	.397	.449	.492	.527	.558	.585	.608
16	.219	.312	.378	.429	.471	.507	.538	.564	.588
17	.208	.297	.361	.411	.452	.488	.518	.545	.569
18	.197	.283	.345	.394	.435	.470	.501	.527	.551
19	.187	.270	.331	.379	.419	.454	.484	.510	.534
20	.179	.259	.317	.364	.404	.438	.468	.495	.518
22	.164	.238	.294	.339	.377	.410	.439	.466	.489
24	.151	.221	.273	.316	.353	.385	.414	.440	.463
26	.140	.206	.256	.297	.332	.363	.391	.417	.440
28	.130	.193	.240	.279	.314	.344	.371	.396	.418
30	.122	.181	.226	.264	.297	.326	.353	.377	.399
35	.105	.157	.198	.232	.262	.289	.314	.336	.357
40	.093	.139	.176	.207	.234	.259	.282	.304	.323
45	.083	.125	.158	.186	.212	.235	.257	.277	.295
50	.075	.113	.143	.170	.194	.215	.235	.254	.272
55	.068	.103	.131	.156	.178	.198	.217	.235	.252
60	.063	.095	.121	.144	.165	.184	.202	.219	.234
80	.047	.072	.093	.111	.127	.142	.157	.171	.184
100	.038	.058	.075	.090	.103	.116	.128	.140	.151
150	.025	.039	.051	.061	.070	.080	.088	.096	.104
200	.019	.030	.038	.046	.053	.060	.067	.074	.080
500	.008	.012	.015	.019	.022	.025	.028	.030	.033

FIGURE A.3 Critical values of F for $PA - PC = 1$

$n - PA$	α							
	.5	.2	.1	.05	.025	.01	.005	.001
1	1.00	9.47	39.9	161.00	648.00	4052.00	16,210.00	405,284.00
2	0.67	3.56	8.53	18.51	38.51	98.50	198.00	998.00
3	0.59	2.68	5.54	10.13	17.44	34.12	55.55	167.00
4	0.55	2.35	4.54	7.71	12.22	21.20	31.33	74.14
5	0.53	2.18	4.06	6.61	10.01	16.26	22.78	47.18
6	0.51	2.07	3.78	5.99	8.81	13.75	18.63	35.51
7	0.51	2.00	3.59	5.59	8.07	12.25	16.24	29.25
8	0.50	1.95	3.46	5.32	7.57	11.26	14.69	25.41
9	0.49	1.91	3.36	5.12	7.21	10.56	13.61	22.86
10	0.49	1.88	3.29	4.96	6.94	10.04	12.83	21.04
11	0.49	1.86	3.23	4.84	6.72	9.65	12.23	19.69
12	0.48	1.84	3.18	4.75	6.55	9.33	11.75	18.64
13	0.48	1.82	3.14	4.67	6.41	9.07	11.37	17.82
14	0.48	1.81	3.10	4.60	6.30	8.86	11.06	17.14
15	0.48	1.80	3.07	4.54	6.20	8.68	10.80	16.59
16	0.48	1.79	3.05	4.49	6.12	8.53	10.58	16.12
17	0.47	1.78	3.03	4.45	6.04	8.40	10.38	15.72
18	0.47	1.77	3.01	4.41	5.98	8.29	10.22	15.38
19	0.47	1.76	2.99	4.38	5.92	8.18	10.07	15.08
20	0.47	1.76	2.97	4.35	5.87	8.10	9.94	14.82
22	0.47	1.75	2.95	4.30	5.79	7.95	9.73	14.38
24	0.47	1.74	2.93	4.26	5.72	7.82	9.55	14.03
26	0.47	1.73	2.91	4.23	5.66	7.72	9.41	13.74
28	0.47	1.72	2.89	4.20	5.61	7.64	9.28	13.50
30	0.47	1.72	2.88	4.17	5.57	7.56	9.18	13.29
35	0.46	1.71	2.85	4.12	5.48	7.42	8.98	12.90
40	0.46	1.70	2.84	4.08	5.42	7.31	8.83	12.61
45	0.46	1.69	2.82	4.06	5.38	7.23	8.71	12.39
50	0.46	1.69	2.81	4.03	5.34	7.17	8.63	12.22
55	0.46	1.68	2.80	4.02	5.31	7.12	8.55	12.09
60	0.46	1.68	2.79	4.00	5.29	7.08	8.49	11.97
80	0.46	1.67	2.77	3.96	5.22	6.96	8.33	11.67
100	0.46	1.66	2.76	3.94	5.18	6.90	8.24	11.50
150	0.46	1.66	2.74	3.90	5.13	6.81	8.12	11.27
200	0.46	1.65	2.73	3.89	5.10	6.76	8.06	11.15
500	0.46	1.65	2.72	3.86	5.05	6.69	7.95	10.96

FIGURE A.4 Critical values of F for $\alpha = .05$

$n - PA$	$PA - PC$								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90

FIGURE A.5 Critical values of χ_{df}^2

<i>df</i>	α						
	0.5	0.2	0.1	0.05	0.01	0.005	0.001
1	0.45	1.64	2.71	3.84	6.63	7.88	10.83
2	1.39	3.22	4.61	5.99	9.21	10.60	13.82
3	2.37	4.64	6.25	7.81	11.34	12.84	16.27
4	3.36	5.99	7.78	9.49	13.28	14.86	18.47
5	4.35	7.29	9.24	11.07	15.09	16.75	20.52
6	5.35	8.56	10.64	12.59	16.81	18.55	22.46
7	6.35	9.80	12.02	14.07	18.48	20.28	24.32
8	7.34	11.03	13.36	15.51	20.09	21.95	26.12
9	8.34	12.24	14.68	16.92	21.67	23.59	27.88
10	9.34	13.44	15.99	18.31	23.21	25.19	29.59
11	10.34	14.63	17.28	19.68	24.72	26.76	31.26
12	11.34	15.81	18.55	21.03	26.22	28.3	32.91
13	12.34	16.98	19.81	22.36	27.69	29.82	34.53
14	13.34	18.15	21.06	23.68	29.14	31.32	36.12
15	14.34	19.31	22.31	25.00	30.58	32.80	37.70
16	15.34	20.47	23.54	26.30	32.00	34.27	39.25
17	16.34	21.61	24.77	27.59	33.41	35.72	40.79
18	17.34	22.76	25.99	28.87	34.81	37.16	42.31
19	18.34	23.9	27.2	30.14	36.19	38.58	43.82
20	19.34	25.04	28.41	31.41	37.57	40.00	45.31

Author Index

Baron, R. M. 240, 248
Belsley, D. A. 323
Bosker, R. J. 306

Campbell, D. T. 2, 130, 240
Clark, H. H. 310
Cohen, J. 61, 62, 133, 152
Cook, R. D. 325
Cook, T. D. 2, 240
Correll, J. 312

Henderson, H. 132
Hoffman, J. M. 249
Hoffman, L. 262
Hox, J. J. 306

Judd, C. M. 2, 130, 240, 248, 249, 260, 310, 312

Kenny, D. A. 2, 130, 240, 248, 260, 310, 312
Kuh, E. 323

Lockwood, C. M. 249

MacKinnon, D. P. 249
Maruyama, G. 2
Muller, D. 249

Park, B. 312

Reis, H. T. 2
Rosenthal, R. 2, 120
Rosnow, R. L. 2, 120
Ryan, C. S. 2

SAS Institute i, 132
Scheffé, H. A. 200, 226
Shadish, W. R. 2, 130, 240
Sheets, V. 249
Snijders, T. A. B. 306
Stanley, J. C. 2, 130

Velleman, P. F. 132

Welsch, R. E. 323
West, S. G. 249
Westfall, J. 310, 312
Wittenbrink, B. 312

Yzerbyt, V. 249

Subject Index

$1-\beta$ 57; *see also* power

absolute error 12–13

additive assumption 139–141

adjusted R^2 132

α *see* significance level

alternative hypothesis 6

ANCOVA 74, 168, 229–259; adjusted means 244–246, 250; adjustment function of 239–246; to examine mediation 240–241, 246–249; homogeneity of regression assumption 249–253; multiple covariates 246; power function of 230–236; in repeated measures designs 292–297

ANOVA: factorial 195, 209–228; mixed model 283–290; one-way 168–204; repeated measures 260–283, 292–297; *see also* multilevel models; simple effects 226–227; source table 46–47, 55–56, 97–98, 124–127, 176–177, 209–214; unequal n 192, 194–196, 203, 225, 259

augmented model 5–7, 86–87, 89

B_0 10, 43

b_0 11; *see also* notation; β_0 : estimates of

β_0 3, 10–11, 15, 17, 18; *see also* notation; estimates of 10–15, 17–18, 21–23, 43–44; 75, 78, 86; *see also* centered predictors; intercept; simple regression

β_1 3–4; *see also* notation; slope

β_j 7; partial 104–105

beta weights *see* standardized regression coefficient

between-subject treatment variable 262–264

bias 38–39, 252; *see also* properties of estimators; if nonindependence ignored 269–271, 278–279, 310–311; in PRE 48–49, 61–62; resulting from violations of assumptions 314–315, 321, 338

Bonferroni inequality for multiple comparisons 200, 324

Cauchy distribution 328–331

CE *see* count of errors

centered predictors 83–85, 150–153, 296; *see also* multilevel models; in logistic regression 345–346

central limit theorem 29–30

central tendency 11, 15, 21; *see also* mean; median;

mode

chi-square 306, 341, 343, 346

coefficient of multiple determination 117

coefficient of partial determination 121

coefficient of variation (CV) 21–22

compact model 5–7, 43, 85–86, 89

compound symmetry 282

conditional predictions 2–4, 72–73, 103

confidence intervals 67–68, 70; effect of outlier on 315; for slope of a contrast-coded predictor 177–178; in interactive models 153–156; in logistic regression 346–348; in multiple regression 121–122, 133; in simple regression 92–93, 99–100

consistency *see* properties of estimators

constant variance *see* homogeneity of variance; errors: assumptions about distribution of contrast codes 169–170, 179–193; in factorial ANOVA 209–212; Helmert codes 182–183, 195; orthogonal polynomial contrasts 197–199; unequal n 194–196

Cook's D 325–327, 335, 338

correlation coefficient 91

count of errors (CE) 15, 18, 20, 21

covariance 79–80

covariate 98, 229; interactions *see* ANCOVA:

homogeneity of regression assumption; interactions; nonorthogonal to condition 239–246; orthogonal to condition 232–239; in within-subject designs 294–297

critical values 52

crossed design 262, 278, 279–283; *see also*

ANOVA: repeated measures; nonindependence: in crossed designs

CV *see* coefficient of variation

DATA 1–4, 6–8

degrees of freedom (df) 56; in logistic regression 342–343

δ_i 266

dependence *see* errors: assumptions about distribution of; nonindependence

descriptive statistics 11, 19, 24; *see also* simple model

deviance 306, 342–343, 345, 346, 348, 352; proportional reduction in 343, 345

dichotomous dependent variables 339; analysis of *see* logistic regression

difference scores: analysis of 236–239

- dummy codes 193–194
Dunn method of multiple comparisons 200
- e_i 11–12; *see also* notation; errors
efficiency *see* properties of estimators
 ε_i 10, 15, 25; *see also* notation; errors
ERROR 1–4, 75; (A) 5–7; (C) 5–7; definitions of 10–24; per remaining parameter 20
errors 1, 15, 25, 75–76, 81–82; assumptions about distribution of 28–30, 37–39, 85, 260, 314, 327; metaphorical representation of 26–27; violations of assumptions about distribution of 252, 260, 270, 314, 327–338
 η^2 46, 57–60, 93–95; unbiased estimate of 61–63
- F 52–56, 71
factorial design *see* ANOVA: factorial
Fisher's z transformation 334
- general linear model 104
grand mean 174
- Helmert contrast codes *see* contrast codes: Helmert codes
heterogeneity of variance 38, 332–338
heteroscedasticity 38, 340
hierarchical linear models *see* multilevel models
homogeneity of regression assumption *see* ANCOVA
homogeneity of variance 38, 314, 327–328; resolving violations of 333–338
homoscedasticity 38
- identically distributed *see* homogeneity of variance
independence *see* errors: assumptions about distribution of; nonindependence
independent samples t -test *see* t -test: for independent samples
indicator function 15
inferential statistics 6
interactions 135; between categorical predictors 205, 210, 214–222, 224–225; *see also* ANOVA: factorial; between continuous and categorical predictors 255–259; *see also* ANCOVA: homogeneity of regression assumption; between continuous predictor variables 141–148; in complex product models 162–166; deviating or centering continuous component predictors 150–153, 258–259; in logistic regression 349–351; in repeated measures ANOVA 279–290; tests of other parameters in models with 148–150, 258–259
intercept 76, 80–81, 86–87, 101; *see also* simple regression; in additive models 139–141; in interactive models 141–143, 219–221; interpretation when predictor is centered 83–85; interpretation with categorical predictor 172–174, 184, 186, 189; in logistic regression 345–346; in mixed designs 285–289
intraclass correlation 270, 277–278
- ladder of powers *see* power transformations
 λ 160
least-squares coefficient 77, 238; derivation of 78–79
levers 318–321, 336, 338
linear model 74–77; *see also* general linear model
location, measures of *see* central tendency
logistic regression 339–341; compared to least squared regression 339, 352; confidence interval 346–348; interactions in 349–351; with categorical predictor 339–344; with centered predictors 345–346; with continuous predictor 344–347, 348–349; with multiple predictors 348–351
logits 341–342, 344, 346–347, 350–351
log odds *see* logits
log transformation 334–335
- MAD *see* median absolute deviation
main effect 212
Mallow's C_p 132
maximum likelihood estimation 39–41, 42, 305, 311, 313, 341–342, 346
mean 14–15, 18, 19, 24, 35–36; *see also* properties of estimators; simple model; sampling distribution; proof that it minimizes SSE 17–19
mean deviated predictors *see* centered predictors
mean square error (MSE) 20, 22, 39, 42, 55–56, 83, 87, 89, 93, 133, 176–177; *see also* @IND:ANOVA: repeated measures
mean squares: between subjects (MSB) 167–168; reduced (MSR), 55–56 167–168; within subjects (MSW) 167–168
median 12, 14–15, 18, 19, 21, 24, 35–36; *see also* sampling distribution; properties of estimators
median absolute deviation (MAD) 18, 20–21, 24
mediation 240–241, 246–249
modal error 20, 24
mode 18, 19, 21, 24
Model 1, 2–4, 8–9; A *see* augmented model; C *see* compact model; simple *see* simple model
moderator regression *see* interactions
MSB *see* mean squares: between subjects
MSE *see* mean square error
MSR *see* mean squares: reduced
MSW *see* mean squares: within subjects
 μ 31, 33, 39, 175
multilevel models 272, 302–313; advantages over repeated measures models 306, 313; crossed random effects models 310–313; equivalence to mixed model ANOVA 297–302, 313; with categorical predictor variables 302–306, 313; with continuous predictor variables 306–309, 313
Multiple R^2 *see* coefficient of multiple determination
multiple regression 103–104; cautions about using 129–132; estimating parameters 105–107; interpreting parameters 107–116; problems with tests of overall models 119–120; relative importance of predictor variables 130–131; statistical inference 116–129

- nested design 262–263; *see also* ANOVA: repeated measures; nonindependence: in nested designs
 nonadditive models *see* interactions
 nonindependence 260–264, 278–279, 293; *see also* errors: assumptions about distribution of; multilevel models; in crossed designs 271–279; direction of bias if ignored 278; in mixed designs 283–288; in nested designs 264–271, 278–279
 nonlinear models 135; *see also* polynomial models
 non-normality of errors 327–331, 333–338; *see also* errors: assumptions about distribution of
 nonparametric statistical tests 333–334
 normal distribution 28–29; formal specification of 30–33
 normal quantile-quantile plot 329–331, 334–338
 notation 7–9
 null hypothesis 6, 8, 43, 48–49
- odds 339–341, 344, 345, 346–348, 350–351
 odds ratio 344
 one-sample *t*-test *see t*-test: one-sample; simple model: equivalence to one-sample *t*-test
 outliers 36, 38, 314–316; *see also* errors: violations of assumptions about distribution of; dealing with 326–327; detecting 12–15, 45, 316–326; in logistic regression 352
- PA 54; *see also* degrees of freedom
 parameter 1–5, 7–9, 10
 partial correlation coefficient 121–122
 partial regression coefficient *see* multiple regression
 partial regression plots 115–116
 PC 54; *see also* degrees of freedom
 Pearson *r* 91
 planned comparison 200, 225–227
 polynomial contrasts *see* contrast codes: orthogonal
 polynomial contrasts
 polynomial models 135; *see also* powers of predictor variables
 pooling error term within subjects 282–283
 post hoc comparison 200–201, 225–227
 power 57, 93; in ANOVA 201–203; effects of heteroscedasticity on 315, 328; effects of including a covariate on 229, 232–236; estimating 57–63, 93–95, 132–133; in factorial designs 224–226, 227–228; improving 64–67; in interactive models 153–156; in logistic regression 352; and research design 202–203; in within-subjects designs 281–282
 power transformations 334–338
 powers of predictor variables 158–162; in complex product models 162–166
 PRE 5–6, 8–9, 70; *see also* ?; per parameter 6, 18, 53–55, 119; sampling distribution of 48–52, 57–60
 probability of binary outcome 339–341, 344, 346–347, 350–351
 properties of estimators 27–28, 35–36
 proportional reduction in deviance (PRD) *see* deviance
- proportional reduction in error *see* PRE
 pseudo- R^2 343
p value 56
- R^2 *see* coefficient of multiple determination
 random effects models *see* multilevel models
 redundancy: of continuous and categorical predictors 229–230, 232, 233, 241–246, 253–255; in interactive models 153–156; of predictor variables 104–106, 121–122, 126–127, 131–133; in ANOVA 195–196, 208, 209–210
 regression *see* simple regression; multiple regression
 residual *see* errors
 residual plots 332–333, 335–338; *see also* normal quantile-quantile plot; spread-location plot
 restriction of range 320
 root-mean-squared-error (RMSE) 20, 22, 39
 RStudent *see* studentized deleted residual
- s* *see* standard deviation
 s^2 *see* variance
 SAE *see* sum of absolute errors
 sample 26
 sampling distribution 26–27, 33–35; of PRE 48–52, 57–60
 Satterthwaite approximation 306
 scatterplot 74–77, 79–81, 85, 108–109, 317–318
 Scheffé method of multiple comparisons 200–201, 227
 SE *see* sum of errors
 σ^2 31, 39
 significance level 57
 simple effects *see* ANOVA; “simple” relationships
 simple model 3, 10–11, 15, 21, 25, 26; equivalence to one-sample *t*-test 68–69; statistical inference 43–56
 simple regression 72–74, 100–103; centered predictor 83–85; defining parameters 74–77; estimating parameters 77–83; statistical inference about intercept (β_0) 95–99; statistical inference about slope (β_1) 85–92
 “simple” relationships 139; in additive models 139–141; in complex product models 162–166; derivation of 156–158; in factorial ANOVA 219–222; in models with continuous and categorical predictors 251–252, 256–259; *see also* ANCOVA: homogeneity of regression assumption; in models with continuous predictors 142–153; in models with powers of predictor variables 158–162
 skewed distribution 35–36, 328, 334–335
 slope 76–77, 101; *see also* multiple regression; simple regression; “simple” relationships; associated with contrast-coded predictor 170–175, 179–183, 214–222; associated with contrast-coded predictor in ANCOVA 241–245; in logistic regression 345–346
 sphericity assumption 282
 spread-location plot 332–333, 335–338
 spread, measures of *see* variability, measures of

- squared partial correlation 122
- SSB *see* sum of squares: between
- SSE *see* sum of squared errors
- SSE(A) *see* sum of squared errors: for Model A
- SSE(C) *see* sum of squared errors: for Model C
- SSR *see* sum of squares: reduced
- SSW *see* sum of squares: within
- standard deviation 20–22, 24, 39
- standard error of prediction 83
- standardized regression coefficient 130–131
- standard normal deviates 33
- Stem and Leaf plot 45
- stepwise regression 131–132
- studentized deleted residual 322–325, 335–338
- sum of absolute errors (SAE) 12–14, 16, 18, 20, 21, 35
- sum of errors (SE) 16
- sum of squared errors (SSE), 13–15, 16–18, 20–22, 24, 35; decomposition of 46–47; for Model A (SSE(A)) 45–46; for Model C (SSE(C)) 45–46; proof that mean minimizes 17–19
- sum of squares 13–15; accounted for by a contrast-coded predictor 174; between (SSB) 176–177; reduced (SSR) 47–48, 90, 96; in repeated measures designs 266–270, 275–277, within (SSW) 176–177
- suppression 127

- thick-tailed distribution 36, 316, 328–338
- time-series 37, 262
- tolerance 122, 133, 154; *see also* redundancy
- transformations 332, 333–335, 337
- true PRE *see* η^2

- t*-test: equivalence to *F* 68–69, 91; for correlation coefficient 91; for independent samples 175–177; in multilevel models 306; one-sample 43, 68–69, 98
- Type I error 56–57, 65–66, 71, 93, 118, 132, 320, 324, 328; rates in tests of mean differences 199–200
- Type II error 57–59, 64–66, 71, 93, 328

- unbiased *see* errors: assumptions about the distribution of; properties of estimators
- uniform distribution 30, 328–331

- variability, measures of 11, 15–17, 18, 20–21
- variance 20, 22, 39
- variance inflation factor (VIF) 122

- W* 266; *see also* within-subject contrasts; conversion to original *Y* metric 268, 275
- Wald chi-square 346
- weighted mean 18
- weighted sum of squared errors (WSSE) 16–17, 18
- within-subject contrasts 272–273; multiple 279–281
- within-subjects treatment variable 37, 262–264
- WSSE *see* weighted sum of squared errors

- X_i 4, 72–73; *see also* notation

- \hat{Y} *see* notation
- Y_i 3; *see also* notation

- z* score 31–33; and normal quantile-quantile plot 329