

**MULTIVARIABLE
MODELING AND
MULTIVARIATE
ANALYSIS
FOR THE
BEHAVIORAL
SCIENCES**

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Series Editors

A. Colin Cameron
University of California, Davis, USA

J. Scott Long
Indiana University, USA

Andrew Gelman
Columbia University, USA

Sophia Rabe-Hesketh
University of California, Berkeley, USA

Anders Skrondal
London School of Economics, UK

Aims and scope

Large and complex datasets are becoming prevalent in the social and behavioral sciences and statistical methods are crucial for the analysis and interpretation of such data. This series aims to capture new developments in statistical methodology with particular relevance to applications in the social and behavioral sciences. It seeks to promote appropriate use of statistical, econometric and psychometric methods in these applied sciences by publishing a broad range of reference works, textbooks and handbooks.

The scope of the series is wide, including applications of statistical methodology in sociology, psychology, economics, education, marketing research, political science, criminology, public policy, demography, survey methodology and official statistics. The titles included in the series are designed to appeal to applied statisticians, as well as students, researchers and practitioners from the above disciplines. The inclusion of real examples and case studies is therefore essential.

Published Titles

Analysis of Multivariate Social Science Data, Second Edition

David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane I. Galbraith

Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition

Jeff Gill

Foundations of Factor Analysis, Second Edition

Stanley A. Mulaik

Linear Causal Modeling with Structural Equations

Stanley A. Mulaik

Multiple Correspondence Analysis and Related Methods

Michael Greenacre and Jorg Blasius

Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences

Brian S. Everitt

Statistical Test Theory for the Behavioral Sciences

Dato N. M. de Gruyter and Leo J. Th. van der Kamp

**Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series**

**MULTIVARIABLE
MODELING AND
MULTIVARIATE
ANALYSIS
FOR THE
BEHAVIORAL
SCIENCES**

BRIAN S. EVERITT



CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20110725

International Standard Book Number-13: 978-1-4398-0770-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

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

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Dedication

To the memory of my parents, Emily Lucy Everitt
and Sidney William Everitt

Contents

Preface.....	xiii
Acknowledgments	xvii

1. Data, Measurement, and Models.....	1
1.1 Introduction	1
1.2 Types of Study	2
1.2.1 Surveys	3
1.2.2 Experiments	4
1.2.3 Observational Studies	5
1.2.4 Quasi-Experiments	6
1.3 Types of Measurement	7
1.3.1 Nominal or Categorical Measurements.....	7
1.3.2 Ordinal Scale Measurements	8
1.3.3 Interval Scales.....	8
1.3.4 Ratio Scales	9
1.3.5 Response and Explanatory Variables.....	9
1.4 Missing Values.....	10
1.5 The Role of Models in the Analysis of Data.....	11
1.6 Determining Sample Size	14
1.7 Significance Tests, p-Values, and Confidence Intervals.....	16
1.8 Summary	19
1.9 Exercises	19
2. Looking at Data.....	21
2.1 Introduction	21
2.2 Simple Graphics—Pie Charts, Bar Charts, Histograms, and Boxplots	22
2.2.1 Categorical Data	22
2.2.2 Interval/Quasi-Interval Data	30
2.3 The Scatterplot and Beyond.....	35
2.3.1 The Bubbleplot.....	38
2.3.2 The Bivariate Boxplot	40
2.4 Scatterplot Matrices	44
2.5 Conditioning Plots and Trellis Graphics	45
2.6 Graphical Deception	52
2.7 Summary	58
2.8 Exercises	58

3. Simple Linear and Locally Weighted Regression.....	61
3.1 Introduction	61
3.2 Simple Linear Regression	62
3.2.1 Fitting the Simple Linear Regression Model to the Pulse Rates and Heights Data	64
3.2.2 An Example from Kinesiology	65
3.3 Regression Diagnostics	68
3.4 Locally Weighted Regression	72
3.4.1 Scatterplot Smoothers	73
3.5 Summary	79
3.6 Exercises	80
4. Multiple Linear Regression	81
4.1 Introduction	81
4.2 An Example of Multiple Linear Regression	84
4.3 Choosing the Most Parsimonious Model When Applying Multiple Linear Regression	89
4.4 Regression Diagnostics	96
4.5 Summary	100
4.6 Exercises	100
5. The Equivalence of Analysis of Variance and Multiple Linear Regression, and an Introduction to the Generalized Linear Model	103
5.1 Introduction	103
5.2 The Equivalence of Multiple Regression and ANOVA.....	103
5.3 The Generalized Linear Model	110
5.4 Summary	112
5.5 Exercises	113
6. Logistic Regression	115
6.1 Introduction	115
6.2 Odds and Odds Ratios	115
6.3 Logistic Regression	117
6.4 Applying Logistic Regression to the GHQ Data	120
6.5 Selecting the Most Parsimonious Logistic Regression Model....	124
6.6 Summary	128
6.7 Exercises	128
7. Survival Analysis	131
7.1 Introduction	131
7.2 The Survival Function.....	132
7.3 The Hazard Function	136
7.4 Cox's Proportional Hazards Model.....	138

7.5	Summary	143
7.6	Exercises	144
8.	Linear Mixed Models for Longitudinal Data	145
8.1	Introduction	145
8.2	Linear Mixed Effects Models for Longitudinal Data.....	146
8.3	How Do Rats Grow?	150
8.3.1	Fitting the Independence Model to the Rat Data	151
8.3.2	Fitting Linear Mixed Models to the Rat Data	153
8.4	Computerized Delivery of Cognitive Behavioral Therapy— Beat the Blues.....	157
8.5	The Problem of Dropouts in Longitudinal Studies	162
8.6	Summary	165
8.7	Exercises	166
9.	Multivariate Data and Multivariate Analysis	169
9.1	Introduction	169
9.2	The Initial Analysis of Multivariate Data.....	170
9.2.1	Summary Statistics for Multivariate Data.....	170
9.2.2	Graphical Descriptions of the Body Measurement Data	173
9.3	The Multivariate Normal Probability Density Function.....	174
9.4	Summary	180
9.5	Exercises	181
10.	Principal Components Analysis	183
10.1	Introduction	183
10.2	Principal Components Analysis (PCA).....	183
10.3	Finding the Sample Principal Components	185
10.4	Should Principal Components Be Extracted from the Covariance or the Correlation Matrix?	188
10.5	Principal Components of Bivariate Data with Correlation Coefficient r	190
10.6	Rescaling the Principal Components	192
10.7	How the Principal Components Predict the Observed Covariance Matrix	193
10.8	Choosing the Number of Components	193
10.9	Calculating Principal Component Scores.....	195
10.10	Some Examples of the Application of PCA.....	196
10.10.1	Head Size of Brothers	196
10.10.2	Crime Rates in the United States	200
10.10.3	Drug Usage by American College Students.....	205

10.11 Using PCA to Select a Subset of the Variables	208
10.12 Summary	209
10.13 Exercises	210
11. Factor Analysis	211
11.1 Introduction	211
11.2 The Factor Analysis Model.....	212
11.3 Estimating the Parameters in the Factor Analysis Model.....	215
11.4 Estimating the Numbers of Factors.....	217
11.5 Fitting the Factor Analysis Model: An Example	218
11.6 Rotation of Factors	220
11.6.1 A Simple Example of Graphical Rotation.....	222
11.6.2 Numerical Rotation Methods.....	223
11.6.3 Rotating the Crime Rate Factors	226
11.7 Estimating Factor Scores.....	227
11.8 Exploratory Factor Analysis and Principal Component Analysis Compared	228
11.9 Confirmatory Factor Analysis.....	229
11.9.1 Ability and Aspiration	230
11.9.2 A Confirmatory Factor Analysis Model for Drug Usage	233
11.10 Summary	235
11.11 Exercises.....	236
12. Cluster Analysis.....	239
12.1 Introduction	239
12.2 Cluster Analysis	241
12.3 Agglomerative Hierarchical Clustering.....	241
12.3.1 Clustering Individuals Based on Body Measurements	243
12.3.2 Clustering Countries on the Basis of Life Expectancy Data	246
12.4 <i>k</i> -Means Clustering	250
12.5 Model-Based Clustering.....	253
12.6 Summary	258
12.7 Exercises	259
13. Grouped Multivariate Data	261
13.1 Introduction	261
13.2 Two-Group Multivariate Data.....	262
13.2.1 Hotelling's T^2 Test	262
13.2.2 Fisher's Linear Discriminant Function.....	265
13.3 More Than Two Groups	270
13.3.1 Multivariate Analysis of Variance (MANOVA).....	270
13.3.2 Classification Functions	273

13.4 Summary	277
13.5 Exercises	277
References	279
Appendix: Solutions to Selected Exercises.....	285
Index	299

Preface

The *Encyclopedia of Statistics in the Behavioral Sciences* (Everitt and Howell, 2005) opens with the following paragraph:

Forty years ago there was hardly a field called “behavioral science.” In fact, psychology largely was the behavioral sciences, with some help from group theory in sociology and decision making in economics. Now, of course, psychology has expanded and developed in a myriad of ways, to the point where behavioral science is often the most useful term. Physiological psychology has become neuroscience, covering areas not previously part of psychology. Decision-making has become decision science, involving people from economics, marketing, and other disciplines. Learning theory has become cognitive science, again exploring problems that were not even considered 40 years ago. And developments in computing have brought forth a host of new techniques that were not possible in the days of manual and electronic calculators. With all these changes, there have been corresponding changes in the appropriate statistical methodologies.

Despite the changes mentioned in the last sentence of this quotation, many statistical books aimed at psychologists and others working in the behavioral sciences continue to cover primarily simple hypothesis testing, using a variety of parametric and nonparametric significance tests, simple linear regression, and analysis of variance. Such statistical methodology remains important in introductory courses, but represents only the first step in equipping behavioral science students with enough statistical tools to help them on their way to success in their later careers. The aim of this book is to encourage students and others to learn a little more about statistics and, equally important, how to apply statistical methods in a sensible fashion. It is hoped that the following features of the text will help it reach its target:

- The central theme is that statistics is about solving problems; data relevant to these problems are collected and analyzed to provide useful answers. To this end, the book contains a large number of real data sets arising from real problems. Numerical examples of the type that involve the skiing activities of belly dancers and politicians are avoided as far as possible.
- Mathematical details of methods are confined to numbered and separated Technical Sections. For the mathematically challenged, the most difficult of these displays can, at least as a last resort, be ignored. But the study of the relevant mathematical material (which on occasion will include the use of vectors and matrices)

will undoubtedly help in the reader's appreciation of the corresponding technique.

- Although many statistical methods require considerable amounts of arithmetic for their application, the burden of actually performing the necessary calculations has been almost entirely removed by the development and wide availability of powerful and relatively cheap personal computers and associated statistical software packages. It is assumed, therefore, that all students will be using such tools when undertaking their own analyses. Consequently, arithmetic details are noticeable largely by their absence, although a little arithmetic is included where it is considered helpful in explaining a technique.
- There are many challenging data sets both in the text and in the exercises provided at the end of each chapter. All data sets, both in the body of the text and in the exercises, are given on the Web site associated with the book, as are the answers to all the exercises. (Because the majority of data sets used in the book are available on the book's Web site, (<http://www.crcpress.com/product/isbn/978143980769>) tables of data in the text only give a small subset of each data set.)

As mentioned in the penultimate bullet point above, the text assumes that readers will be using one or other of the many available statistical software packages for data analysis. This raises the thorny question for the author of what information should be provided in the text about software. Would, for example, screen dumps from SPSS be useful, or listings of STATA code? Perhaps, but neither are included here. Instead, all the computer code used to analyze the many examples to be found in the text is given on the book's Web site, and this code is in the R language, where R is a software system for statistical computing, data analysis, and graphics. This may appear a strange choice for a book aimed at behavioral scientists, but the rationale behind the choice is first that the author uses R in preference to other statistical software, second that R can be used to produce many interesting and informative graphics that are difficult if not impossible to produce with other software, third that R is free and can be easily downloaded by students, and fourth, R has a very active user community and recently developed statistical methods become available far more quickly than they do with other packages. The only downside with R is that it takes a little more time to learn than say using "point-and-click" SPSS. The initial extra effort, however, is rapidly rewarded. A useful book for learning more about R is Everitt and Hothorn (2008).

The material covered in the book assumes the reader is familiar with the topics covered in introductory statistics courses, for example, population, sample, variable, parameter, significance test, p-value, confidence interval, correlation, simple regression, and analysis of variance. The book is primarily about methods for analyzing data but some comments are made in

Chapter 1 about the various types of study that behavioral researchers may use and their design. And it is in Chapter 1 that the distinction between multivariable and multivariate—both of which appear in the book’s title—will be explained.

It is hoped that the text will be useful in a number of different ways, including:

- As the main part of a formal statistics course for advanced undergraduates and postgraduates in all areas of the behavioral sciences.
- As a supplement to an existing course.
- For self-study.
- For researchers in the behavioral sciences undertaking statistical analyses on their data.
- For statisticians teaching statistics to psychologists and others.
- For statisticians using R when teaching intermediate statistics courses both in the behavioral sciences and in other areas.

B. S. Everitt

Dulwich, U.K.

References

- Everitt, B. S. and Hothorn, T (2009). A Handbook of Statistical Analyses Using R, 2nd edition, Chapman and Hall/CRC, Boca Raton, Florida.
- Everitt, B. S. and Howell, D (2005). Encyclopedia of Statistics in the Behavioral Sciences, Wiley, Chichester, U.K.

Acknowledgments

Thanks are due to Dr. Deepayan Sarkar, the author of *Lattice: Multivariate Data Visualization with R*, and Springer, the publishers of the book, for permission to use Figures 2.8, 2.9, 4.5, 4.6, 4.7, and 5.16 from the book in Chapter 2 of this book. I would also like to thank an anonymous reviewer who made many helpful suggestions that have greatly improved a number of sections in the book. Finally, I would like to thank Rob Calver of Taylor & Francis for his constant support during the writing of this book, and the magnanimous manner in which he has dealt with the move of Harry Redknapp, the one-time manager of his beloved football team Portsmouth, to the more glamorous and successful Tottenham Hotspurs team, a team supported by the writer of this book for the last 40 years.

1

Data, Measurement, and Models

1.1 Introduction

Statistics is a general intellectual method that applies wherever data, variation, and chance appear. It is a fundamental method because data, variation and chance are omnipresent in modern life. It is an independent discipline with its own core ideas, rather than, for example, a branch of mathematics Statistics offers general, fundamental and independent ways of thinking.

Journal of the American Statistical Association

Quintessentially, statistics is about solving problems; data (measurements or observations) relevant to these problems are collected, and statistical analyses are used to provide useful answers. But the path from data collection to analysis and interpretation is often not straightforward. Most real-life applications of statistical methodology have one or more nonstandard features, meaning in practice that there are few routine statistical questions, although there are questionable statistical routines. Many statistical pitfalls lie in wait for the unwary. Indeed, statistics is perhaps more open to misuse than most other subjects, particularly by the nonstatistician with access to powerful statistical software. The misleading average, the graph with “fiddled axes,” the inappropriate p-value, and the linear regression fitted to nonlinear data are just four examples of horror stories that are part of statistical folklore.

Statisticians often complain that many of those working in the behavioral sciences put undue faith in significance tests, use complex methods of analysis when the data merit only a relatively simple approach, and sometimes abuse the statistical techniques they are employing. Statisticians become upset (and perhaps feel a little insecure) when their advice to, say, “plot a few simple graphs,” is ignored in favor of a multivariate analysis of covariance or similar statistical extravagance.

However, if statisticians are at times horrified by the way in which behavioral scientists apply statistical techniques, behavioral scientists may be no less horrified by many statisticians’ apparent lack of awareness of what stresses behavioral research can place on an investigator. A statistician may, for example, demand a balanced design with 30 subjects in each cell so as to

achieve some appropriate power for the analysis. But it is not the statistician who is faced with the frustration caused by a last-minute phone call from a subject who cannot take part in an experiment that has taken several hours to arrange. Again, the statistician advising on a longitudinal study may call for more effort in carrying out follow-up interviews so that the study avoids statistical problems produced by the presence of missing data. It is, however, the behavioral researcher who must continue to persuade people to talk about potentially distressing aspects of their lives, who must confront possibly dangerous respondents, or who arrives at a given (and often remote) address to conduct an interview, only to find that the person is not at home. Many statisticians often do not appear to appreciate the complex stories behind each data point in many behavioral studies. One way of improving the possible communication problems between behavioral scientist and statistician is for each to learn more about the language of the other. There is already available a plethora of, for example, "Statistics for Psychologists" books, but sadly, (as far as I know) no "Psychology for Statisticians" equivalent. Perhaps there should be?

Having outlined briefly a few caveats about the possible misuse of statistics and the equally possible conflict between statistician and behavioral scientist, it is time to move on to consider some of the basics of behavioral science studies and their implications for statistical analysis.

1.2 Types of Study

It is said that, when Gertrude Stein lay dying, she roused briefly and asked her assembled friends, "Well, what's the answer?" They remained uncomfortably quiet, at which she sighed, "In that case, what's the question?"

Research in the behavioral science, as in science in general, is about searching for the answers to particular questions of interest. Do politicians have higher IQs than university lecturers? Do men have faster reaction times than women? Should phobic patients be treated by psychotherapy or by a behavioral treatment such as flooding? Do children who are abused have more problems later in life than children who are not abused? Do children of divorced parents suffer more marital breakdowns themselves than children from more stable family backgrounds?

In more general terms, scientific research involves a sequence of asking and answering questions about the nature of relationships among variables (e.g., How does A affect B? Do A and B vary together? Is A significantly different from B? and so on). Scientific research is carried out at many levels that differ in the types of question asked and therefore in the procedures used to answer them. Thus, the choice of which methods to use in research is largely determined by the kinds of questions that are asked.

Of the many types of investigation used in behavioral research, the most common are perhaps the following:

- Surveys
- Experiments
- Observational studies
- Quasi-experiments

Some brief comments about each of these four types are given below; a more detailed account is available in the papers by Stretch, Raulin, and Graziano, and by Dane, all of which appear in the second volume of the excellent *Companion Encyclopedia of Psychology* (see Colman, 1994).

1.2.1 Surveys

Survey methods are based on the simple discovery that “asking questions is a remarkably efficient way to obtain information from and about people” (Schuman and Kalton, 1985, p. 635). Surveys involve an exchange of information between researcher and respondent; the researcher identifies topics of interest, and the respondent provides knowledge or opinion about these topics. Depending upon the length and content of the survey as well as the facilities available, this exchange can be accomplished via written questionnaires, in-person interviews, or telephone conversations; and, in the 21st century, surveys via the Internet are increasingly common.

Surveys conducted by behavioral scientists are usually designed to elicit information about the respondents’ opinions, beliefs, attitudes, and values. Perhaps one of the most famous surveys of the 20th century was that conducted by Alfred Charles Kinsey, a student of human sexual behavior in the 1940s and 1950s. The first Kinsey report, *Sexual Behavior in the Human Male*, appeared in 1948 (see Kinsey et al., 1948), and the second, *Sexual Behavior in the Human Female*, in 1953 (see Kinsey et al., 1953). It is no exaggeration to say that both reports caused a sensation, and the first quickly became a bestseller.

Surveys are often a flexible and powerful approach to gathering information of interest, but careful consideration needs to be given to several aspects of the survey if the information is to be accurate, particularly when dealing with a sensitive topic. Having a representative sample, having a large-enough sample, minimizing nonresponse, and ensuring that the questions asked elicit accurate responses are just a few of the issues that the researcher thinking of carrying out a survey needs to consider. Readers are referred to Sudman and Bradburn (1982), and Tourangeau, Rips, and Rasinski (2000) for a detailed account of survey methodology.

Examples of data collected in surveys and their analysis are given in several later chapters.

1.2.2 Experiments

According to Sir Ronald Fisher, perhaps the greatest statistician of the 20th century, “experiments are only experience carefully planned in advance and designed to form a secure basis of new knowledge.” The essential feature of an experiment is the large degree of control in the hands of the experimenters, and in designed experiments the goal is to allow inferences to be drawn about the effects of an intervention of interest that are logically compelled by the data and hence allow assessment of a causal relationship. In many cases the “intervention” will be some form of therapy in which case the experiment is usually called a clinical trial.

In an experiment, the researcher controls the manner in which subjects are allocated to the different levels of the experimental factors. In a comparison of a new treatment with one used previously, for example, the researcher would have control over the scheme for allocating subjects to the two treatments. The manner in which this control is exercised is of vital importance if the results of the experiment are to lead to a largely unambiguous assessment of the effect of treatment. The objective in allocation is that the groups to be compared should be alike in all respects except the intervention (treatment) received. Comparable groups prior to the intervention ensure that differences in outcomes after the intervention reflect effects of the intervention in an unbiased fashion. Let us begin by considering two flawed allocation procedures that are unlikely to achieve the desired degree of similarity of the two groups.

- Perhaps the first subjects to volunteer to take part in the experiment should all be given the new treatment, for example, and the later ones the old treatment? The two groups formed in this way may differ in level of motivation and so subsequently in performance. Observed treatment differences would be confounded with differences produced by the allocation procedure. Alternatively, early volunteers might be more seriously ill, those desperate to find a new remedy that works, and again, this might lead to a bias in the measured difference between the two treatments.
- So what about putting alternate subjects into each group? The objection to this is that the experimenter will know who is receiving what treatment and may be tempted to “tinker” with the scheme to ensure that his patients who are most ill receive the new treatment.

So, how should we form the groups that will be used to assess an experimental intervention? The answer is deceptively simple—use randomization. The group to which a participant in the experiment is allocated is decided by chance. It could be arranged by flipping a coin each time a new eligible patient arrives, and allocating the patient to the new treatment if the result is a head, or to the old treatment if a tail appears. In practice, of course, a more sophisticated randomization procedure will be used. The essential feature, however, is

randomization, rather than the mechanism used to achieve it. Randomization was introduced into scientific experiments far more recently, when in 1926 Fisher randomly assigned individual blocks or plots of land in agricultural experiments to receive particular types of “treatment”—different amounts of fertilizer. The primary benefit that randomization has is the chance (and therefore impartial) assignment of extraneous influences among the groups to be compared, and it offers this control over such influences whether or not they are known by the experimenter to exist. Note that randomization does not claim to render the two samples equal with regard to these influences; if, however, the same procedure was applied to repeated samples from the population, equality would be achieved in the long run. Thus, randomization ensures a lack of bias, whereas other methods of assignment may not. In a properly conducted, randomized, experiment the interpretation of an observed group difference is largely unambiguous; its cause is very likely to be the different treatments or conditions received by the groups.

Several of the data sets introduced and analyzed in later chapters arise from experimental studies, often clinical trials.

1.2.3 Observational Studies

Suppose a researcher is interested in investigating how smoking cigarettes affects a person’s systolic blood pressure. Using the experimental approach described earlier, people would have to be allocated at random to two groups, the members of one group being asked to smoke some quantity of cigarettes per day, and the members of the other group required not to smoke at all. Clearly, no ethical committee would approve of such a study. So, what can be done? An approach that would get ethical approval is to measure the systolic blood pressure of naturally occurring groups of individuals who smoke, and those who do not, and then compare the results. This would then be what is known as an observational study, defined by Cochran (1965) as follows:

An empiric comparison of “treated” and “control” groups in which the objective is to elucidate cause-and-effect relationships but where it is not possible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign patients at random to different procedures.

Many observational studies involve recording data on the members of naturally occurring groups, generally over a period of time, and comparing the rate at which a particular event of interest occurs in the different groups (such studies are often referred to as prospective). If, for example, an investigator was interested in the health effects of a natural disaster such as an earthquake, those who experienced the earthquake could be compared, on some outcome variable of interest, with a group of people who did not.

Another commonly used type of observational study is the case-control investigation. Here, a group of people (the cases) all having a particular characteristic

(a certain disease perhaps) are compared with a group of people who do not have the characteristic (the controls), in terms of their past exposure to some event or risk factor. The cases and controls are usually matched one-to-one for possible confounding variables. An example of such a study is reported in Lehman, Wortman, and Williams (1987). Here the researchers collected data following the sudden death of a spouse or a child in a car crash. They matched 80 bereaved spouses and parents to 80 controls drawn from 7582 individuals who came to renew their driver's license. Specifically, they matched for gender, age, family income before crash, education level, and number and ages of children.

The types of analyses suitable for observational studies are often the same as those used for experimental studies. Unlike experiments, however, the lack of control over the groups to be compared in an observational study makes the interpretation of any difference between the groups detected in the study open to a variety of interpretations. In the smoking and systolic blood pressure study, for example, any difference found between the blood pressures of the two groups would be open to three possible interpretations:

- Smoking causes a change in systolic blood pressure.
- Level of blood pressure has a tendency to encourage or discourage smoking.
- Some unidentified factors play a part in determining both the level of blood pressure and whether or not a person smokes.

In the design of an observational study, an attempt is made to reconstruct some of the structure and strengths of an experiment. But the possible ambiguity in interpretation of the results from an observational study, however well designed, means that the observational approach is not as powerful as a designed experiment. A detailed account of observational studies is given in Rosenbaum (2002).

1.2.4 Quasi-Experiments

Quasi-experimental designs resemble experiments proper but are weak on some of the characteristics. In particular (and as in the observational study), the ability to manipulate the groups to be compared is not under the investigator's control. But, unlike the observational study, the quasi-experiment involves the intervention of the investigator in the sense that he or she applies a variety of different "treatments" to naturally occurring groups. In investigating the effectiveness of three different methods of teaching mathematics to 15 year olds, for example, a method might be given to all the members of a particular class in a school. The three classes that receive the different teaching methods would be selected to be similar to each other on most relevant variables, and the methods would be assigned to classes on a chance basis.

For more details of quasi-experiments see Cook and Campbell (1979).

1.3 Types of Measurement

The measurements and observations made on a set of subjects comprise the basic material that is the foundation of all behavioral science investigations. These measurements provide the data for statistical analysis from which the researcher will draw his or her conclusions. Clearly, not all measurements are the same. Measuring an individual's weight is qualitatively different from measuring that person's response to some treatment on a two-category scale: "improved" and "not improved," for example. Whatever measurements are made, they need to be objective, precise, and reproducible for reasons nicely summarized in the following quotation from Fleiss (1986):

The most elegant design of a study will not overcome the damage caused by unreliable or imprecise measurement. The requirement that one's data be of high quality is at least as important a component of a proper study design as the requirement for randomization, double blinding, controlling where necessary for prognostic factors, and so on. Larger sample sizes than otherwise necessary, biased estimates, and even biased samples are some of the untoward consequences of unreliable measurements that can be demonstrated.

Measurement scales are often differentiated according to the degree of precision involved. If it is said that an individual has a high IQ, it is not as precise as the statement that the individual has an IQ of 151. The comment that a woman is tall is not as accurate as specifying that her height is 1.88 m. Certain characteristics of interest are more amenable to precise measurement than others. Given an accurate thermometer, a subject's temperature can be measured very precisely. Quantifying the level of anxiety or depression of a psychiatric patient or assessing the degree of pain of a migraine sufferer are, however, more difficult measurement tasks.

Four levels of measurement scales are generally distinguished.

1.3.1 Nominal or Categorical Measurements

Nominal measurements allow patients to be classified with respect to some characteristic. Examples of such measurements are marital status, sex, and blood group. The properties of a nominal scale are

- The categories are mutually exclusive (an individual can belong to only one category).
- The categories have no logical order—numbers may be assigned to categories but merely as convenient labels.

1.3.2 Ordinal Scale Measurements

The next level of measurement is the ordinal scale. This scale has one additional property over those of a nominal scale—a logical ordering of the categories. With such measurements, the numbers assigned to the categories indicate the amount of a characteristic possessed. A psychiatrist may, for example, grade patients on an anxiety scale as “not anxious,” “mildly anxious,” “moderately anxious,” or “severely anxious,” and use the numbers 0, 1, 2, and 3 to label the categories, with lower numbers indicating less anxiety. The psychiatrist cannot infer, however, that the difference in anxiety between patients with scores of, say, 0 and 1 is the same as the difference between patients assigned scores 2 and 3. The scores on an ordinal scale do, however, allow patients to be ranked with respect to the characteristic being assessed.

The following are the properties of an ordinal scale:

- The categories are mutually exclusive.
- The categories have some logical order.
- The categories are scaled according to the amount of a particular characteristic they indicate.

1.3.3 Interval Scales

The third level of measurement is the interval scale. Such scales possess all the properties of an ordinal scale plus the additional property that equal differences between category levels, on any part of the scale, reflect equal differences in the characteristic being measured. An example of such a scale is temperature on the Celsius (C) or Fahrenheit (F) scale; the difference between temperatures of 80°F and 90°F represents the same difference in heat as that between temperatures of 30° and 40° on the Fahrenheit scale. An important point to make about interval scales is that the zero point is simply another point on the scale; it does not represent the starting point of the scale or the total absence of the characteristic being measured. The properties of an interval scale are as follows:

- The categories are mutually exclusive.
- The categories have a logical order.
- The categories are scaled according to the amount of the characteristic they indicate.
- Equal differences in the characteristic are represented by equal differences in the numbers assigned to the categories.
- The zero point is completely arbitrary.

1.3.4 Ratio Scales

The final level of measurement is the ratio scale. This type of scale has one further property in addition to those listed for interval scales, namely, the possession of a true zero point that represents the absence of the characteristic being measured. Consequently, statements can be made about both the differences on the scale and the ratio of points on the scale. An example is weight, where not only is the difference between 100 and 50 kg the same as between 75 and 25 kg, but an object weighing 100 kg can be said to be twice as heavy as one weighing 50 kg. This is not true of, say, temperature on the Celsius or Fahrenheit scales, where a reading of 100° on either scale does not represent twice the warmth of a temperature of 50° . If, however, two temperatures are measured on the Kelvin scale, which does have a true zero point (absolute zero or -273°C), then statements about the ratio of the two temperatures can be made.

The properties of a ratio scale are

- The categories are mutually exclusive.
- The data categories have a logical order.
- The categories are scaled according to the amount of the characteristic they possess.
- Equal differences in the characteristic being measured are represented by equal differences in the numbers assigned to the categories.
- The zero point represents an absence of the characteristic being measured.

In many statistical textbooks, discussion of different types of measurements is often followed by recommendations as to which statistical techniques are suitable for each type. For example, analyses on nominal data should be limited to summary statistics such as the number of cases, the mode, etc., and for ordinal data, means and standard deviations are said to be not suitable. But Velleman and Wilkinson (1993) make the important point that restricting the choice of statistical methods in this way may be a dangerous practice for data analysis. In essence, the measurement taxonomy described is often too strict to apply to real-world data. This is not the place for a detailed discussion of measurement, but we take a fairly pragmatic approach to such problems. For example, we would not agonize too long over treating variables such as measures of depression, anxiety, or intelligence as if they were interval scaled, although strictly, they fit into the ordinal level described earlier.

1.3.5 Response and Explanatory Variables

This is a convenient point to mention a further classification of measurements that is used in many studies, and that is the division of measured

variables into response or dependent variables (often also referred to as outcome variables), and independent (a misnomer; the variables are not independent of one another, and therefore a term to be avoided) or explanatory variables (also occasionally called predictor variables); in this book we shall stick to explanatory. Essentially, response variables are those that appear on the left-hand side of the equation defining the proposed model for the data, with the explanatory variables thought to possibly affect the response variable appearing on the right-hand side of the model equation. Chapters 3 to 8 of this book will be concerned with data sets in which there is a response variable, and possibly several explanatory variables, and the aim of the analysis is to assess which explanatory variables are related to the response; only the response is considered a random variable. Such data sets are best termed multivariable to contrast them from data sets in which all variables are on the same footing, that is, are not divided into response and explanatory variables and all are considered to be random variables; such data are labeled multivariate. Techniques for analyzing multivariate data will be the subject of Chapters 9 to 13.

1.4 Missing Values

A problem that frequently arises when collecting data is that missing values occur, that is, observations and measurements that should have been recorded, but for one reason or another, were not; for example, a subject may simply not turn up for a planned measurement session. When faced with missing values, many researchers simply resort to analyzing only complete cases since this is what most statistical software packages do automatically. If data are being collected on several variables, for example, the researcher might omit any case with a missing value on any of the variables. When the incomplete cases comprise only a small fraction of all cases (say, 5% or less), then case deletion may be a perfectly reasonable solution to the missing data problem. But when there are many cases with missing values, omitting them may cause large amounts of information, that is, the variable values on which a case has been measured, to be discarded, which would clearly be very inefficient. However, the main problem with complete-case analysis is that it can lead to serious biases in both estimation and inference unless the missing data are missing completely at random in the sense that the probabilities of response do not depend on any data values observed or missing (see Chapter 8 and Little and Rubin, 1987, for more details). In other words, complete-case analysis implicitly assumes that the discarded cases are like a random subsample. So, at the very least, complete-case analysis leads to a loss, and perhaps a substantial loss, in power (see [Section 1.5](#)), but worse, analyses based just on complete cases might in some cases be misleading.

So, what can be done? One answer is to consider some form of imputation, which is the practice of “filling in” missing data with plausible values. At one level this will solve the missing-data problem and enable the investigator to progress normally. But from a statistical viewpoint careful consideration needs to be given to the method used for imputation; otherwise, it may cause more problems than it solves. For example, imputing an observed variable mean for a variable’s missing values preserves the observed sample means but distorts the variance of a variable and the correlation of this variable with others, biasing both toward zero. On the other hand, imputing predicted values from regression models tends to inflate observed correlations, biasing them away from zero. Further, treating imputed data as if they were “real” in estimation and inference can lead to misleading standard errors and p-values since they fail to reflect the uncertainty due to the missing data.

Perhaps the most appropriate way to deal with missing values is by a procedure suggested by Rubin (1987), known as multiple imputation. This is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (say 3–10). Each of the simulated complete data sets is then analyzed by the method appropriate for the investigation at hand, and the results are later combined to produce, say, parameter estimates and confidence intervals that incorporate missing-data uncertainty. Details are given in Rubin (1987) and, more concisely, in Schafer (1999). The great virtues of multiple imputation are its simplicity and generality; the user may analyze the data by virtually any technique that would be appropriate if the data were complete. However, one should always bear in mind that the imputed values are not real measurements. We do not get something for nothing, and if there is a substantial proportion of the individuals with large amounts of missing data, one should clearly question whether any form of statistical analysis is worth the bother.

1.5 The Role of Models in the Analysis of Data

Models attempt to imitate the properties of “real” objects or situations in a simpler or more convenient form. A road map, for example, models part of the earth’s surface, attempting to reproduce the relative positions of towns, roads, and other features. Chemists use models of molecules to mimic their theoretical properties, which, in turn, can be used to predict the behavior of real compounds. A good model follows as accurately as possible the relevant properties of the real object while being convenient to use.

Statistical models allow inferences to be made about an object, or activity, or a process by representing some associated observable data. Suppose, for example, a child has scored 20 points on a test of verbal ability, and after studying a dictionary for some time, scores 24 points on a similar test. If it

is believed that studying the dictionary has caused an improvement, then a possible model of what is happening is

$$20 = \{\text{person's initial score}\}$$

$$24 = \{\text{person's initial score}\} + \{\text{improvement}\}$$

The improvement can now be found by simply subtracting the first score from the second. Such a model is, of course, very naive since it assumes that verbal ability can be measured exactly. A more realistic representation of the two scores, which allows for possible measurement error, is

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + \delta + \varepsilon_2$$

where x_1 and x_2 represent the two verbal ability measurements, γ represents the “true” initial measure of verbal ability, and δ is the value of the improvement made in verbal ability. The terms ε_1 and ε_2 represent the measurement error for verbal ability made on the two occasions of testing. Here the improvement score can only be estimated as $\hat{\delta} = x_2 - x_1$. (The “hat” over a parameter indicates an estimate of that parameter.)

A model gives a precise description of what the investigator assumes is occurring in a particular situation; in the foregoing case it says that the improvement, δ , is considered to be independent of γ and is simply added to it. (An important point that needs to be noted here is that if you do not believe in a model, you should not perform operations and analyses on the data that assume the model to be true.)

Suppose now that it is believed that studying the dictionary does more good if a child already has a fair degree of verbal ability, so that the initial ability effect is multiplied by the dictionary effect and that the various random influences that affect the test scores are also dependent on the true scores, so also enter the model multiplicatively. Then an appropriate model would be

$$x_1 = \gamma \varepsilon_1$$

$$x_2 = \gamma \delta \varepsilon_2$$

Now the parameters are multiplied rather than added to give the observed scores x_1 and x_2 . Here, δ might be estimated by dividing x_1 by x_2 .

A further possibility is that there is a limit, λ , to improvement, and studying the dictionary improves performance on the verbal ability test by some proportion of the child’s possible improvement, $\lambda - \gamma$. Here, a suitable model would be

$$x_1 = \gamma + \varepsilon_1$$

$$x_2 = \gamma + (\lambda - \gamma) \delta + \varepsilon_2$$

With this model there is no way to estimate δ from the data unless a value of λ is given or assumed. One of the principal uses of statistical models is to attempt to explain variation in measurements. This variation may be due to a variety of factors, including variation from the measurement system, variation due to environmental conditions that change over the course of a study, variation from individual to individual (or experimental unit to experimental unit), etc.

The decision about an appropriate model should be largely based on the investigator's prior knowledge of an area. In many situations, however, additive linear models are invoked by default since such models allow many powerful and informative statistical techniques to be applied to the data. Such models appear in several later chapters.

Formulating an appropriate model can be a difficult problem. The general principles of model formulation are covered in detail in books on scientific method but include the need to collaborate with appropriate experts, to incorporate as much background theory as possible, etc. Apart from those models formulated entirely on a priori theoretical grounds, most models are, to some extent at least, based on an initial examination of the data, although completely empirical models are rare. The more usual intermediate case arises when a class of models is entertained a priori, but the initial data analysis is crucial in selecting a subset of models from the class. In regression analysis, for example, the general approach is determined a priori, but a scatter diagram (see Chapter 2) will be of crucial importance in indicating the "shape" of the relationship, and residual plots (see Chapter 3) will be essential for checking assumptions such as normality, etc.

The formulation of a preliminary model from an initial examination of the data is the first step in the iterative, formulation/criticism cycle of model building. This can produce some problems since formulating a model and testing it on the same data is not generally considered good science. It is always preferable to confirm whether a derived model is sensible by testing on new data. But when data are difficult or expensive to obtain, some model modification and assessment of fit on the original data are almost inevitable. Investigators need to be aware of the possible dangers of such a process.

Perhaps the most important principle to have in mind when testing models on data is that of parsimony, that is, the "best" model is one that provides an adequate fit to data with the fewest number of parameters. This principle is often known as Occam's razor, which in its original form in Latin is *entia non sunt multiplicanda praeter necessitatem*, which translates roughly as "a plurality of reasons should not be posited without necessity."

One last caveat about statistical models: according to George Box, "all models are wrong, but some are useful." Statistical models are always simplifications, but some models are useful in providing insights into what is happening in complex, real-world situations.

1.6 Determining Sample Size

One of the most frequent questions faced by a statistician dealing with investigators planning a study is, “How many participants do I need to recruit?” Answering the question requires consideration of a number of factors, for example, the amount of time available for the study, the likely ease or difficulty in recruiting the type of subject required, and the possible financial constraints that may be involved. But the statistician may, initially at least, largely ignore these important aspects of the problem and apply a statistical procedure for calculating sample size. To make things simpler, we will assume that the investigation the researcher is planning is an experimental intervention with two treatment groups. To calculate the sample size, the statistician and the researcher will first need to identify the response variable of most interest and the appropriate statistical test to be used in the analysis of the chosen response; then they will need to decide on values for the following quantities:

- The size of the type I error, that is, the significance level.
- The likely variance of the response variable.
- The power they would like to achieve. (For those readers who have forgotten, or perhaps never knew, the power of a statistical test is the probability of the test rejecting the null hypothesis when the null hypothesis is false.)
- A size of treatment effect that the researcher feels is important, that is, a treatment difference that the investigators would not like to miss being able to declare to be statistically significant.

Given such information, the calculation of the corresponding sample size is often relatively straightforward, although the details will depend on the type of response variable and the type of test involved (see the following text for an example). In general terms, the sample size will increase as the variability of the response variable increases, and decrease as the chosen clinically relevant treatment effect increases. In addition, the sample size will need to be larger to achieve a greater power and a more stringent significance level.

As an example of the calculations involved in sample-size determination, consider a trial involving the comparison of two treatments for anorexia nervosa. Anorexic women are to be randomly assigned to each treatment, and the gain in weight in kilograms after three months is to be used as the outcome measure. From previous experience gained in similar trials, it is known that the standard deviation (σ) of weight gain is likely to be about 4 kg. The investigator feels that a difference in weight gain of 1 kg (Δ) would be of clinical importance and wishes to have a power of 90% when the appropriate two-sided test is used with significance level of 0.05 (α). The

formula for calculating the number of women required in each treatment group (n) is

$$n = \frac{2(Z_{\alpha/2} + Z_{\beta})^2 \sigma^2}{\Delta^2}$$

where β is 1-Power, and

- $Z_{\alpha/2}$ is the value of the normal distribution that cuts off an upper tail probability of $\alpha/2$. So, for $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$.
- Z_{β} is the value of the normal distribution that cuts off an upper tail probability of β . So, for a power of 0.90, $\beta = 0.10$ and $Z_{\beta} = 1.28$.

Therefore, for the anorexia trial,

$$n = \frac{2 \times (1.96 + 1.28)^2 \times 4^2}{1} = 336 \text{ women per treatment group}$$

The foregoing example is clearly simplistic in the context of most psychiatric clinical trials in which measurements of the response variable are likely to be made at several different time points, during which time some patients may drop out of the trial (see Chapter 8 for a discussion of such longitudinal data and the drop out problem). Fortunately, the last decade or so has produced a large volume of methodology useful in planning the size of randomized clinical trials with a variety of different types of outcome measures and with the complications outlined; some examples are to be found in Lee (1983), McHugh and Lee (1984), Schoenfield (1983), Sieh (1987), and Wittes and Wallenstein (1987). In many cases, tables are available that enable the required sample size for chosen power, significance level, effect size, etc., to be simply read off. Increasingly, these are being replaced by computer software for determining sample size for many standard and nonstandard designs and outcome measures.

An obvious danger with the sample size determination procedure just mapped out is that investigators (and, in some cases, even their statisticians) may occasionally be led to specify an effect size that is unrealistically extreme (what Senn, 1997, has described with his usual candor as “a cynically relevant difference”) so that the calculated sample size looks feasible in terms of possible pressing temporal and financial constraints. Such a possibility may be what led Senn (1997) to describe power calculations as “a guess masquerading as mathematics,” and Pocock (1996) to comment that they are “a game that can produce any number you wish with manipulative juggling of the parameter values.” Statisticians advising on behavioral investigations need to be active in estimating the degree of difference that can be realistically expected for a study based

on previous studies of a similar type or, when such information is lacking, perhaps based on subjective opinions of investigators not involved in the putative study.

Getting the sample size right in a scientific study is generally believed to be critical; indeed, according to Simon (1991), discussing in particular clinical trials,

An effective clinical trial must ask an important question and provide a reliable answer. A major determinant of the reliability of the answer is the sample size of the trial. Trials of inadequate size may cause contradictory and erroneous results and thereby lead to an inappropriate treatment of patients. They also divert limited resources from useful applications and cheat the patients who participated in what they thought was important clinical research. Sample size planning is, therefore, a key component of clinical trial methodology.

Studies with inadequate sample sizes risk missing important intervention differences, a risk nicely summarized in the phrase “absence of evidence is not evidence of absence.” The case against studies with inadequate numbers of subjects appears strong, but as Senn (1997) points out, sometimes only a small study is possible. Also, misinterpreting a nonsignificant effect as an indication that a treatment effect is not effective, rather than as a failure to prove that it is effective, suggests trying to improve statistical education rather than totally abandoning small studies. In addition, with the growing use of systematic reviews and meta-analysis (see, for example, Everitt and Wessely, 2008), the results from small studies may prove valuable in contributing to an overview of the evidence of intervention effectiveness, a view neatly summarized by Senn in the phrase “some evidence is better than none.” Perhaps size really is not always everything.

1.7 Significance Tests, p-Values, and Confidence Intervals

Although we are assuming that readers have had an introductory course in statistics that covered simple significance tests, p-values, and confidence intervals, a few more words about these topics here will hopefully not go amiss.

For many behavioral science students and researchers the still-ubiquitous p-value continues to be the Holy Grail of their research efforts, and many see it as the *raison d'être* of statistics and statisticians. Despite the numerous caveats about p-values in the literature (e.g., Gardner and Altman, 1986), many behavioral scientists still seem determined to experience a “eureka moment” on finding a p-value of 0.049, and despair on finding one of 0.051. The p-value retains a powerful hold over the average behavioral researcher and student; there are a number of reasons why it should not.

The first is that the p-value is poorly understood. Although p-values appear in almost every account of behavioral science research findings, there is evidence that the general degree of understanding of the true meaning of the term is very low. Oakes (1986), for example, put the following test to 70 academic psychologists:

Suppose you have a treatment which you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further suppose you use a simple independent means t-test and your result is $t = 2.7$, $df = 18$, $P = 0.01$. Please mark each of the following statements as true or false.

- You have absolutely disproved the null hypothesis that there is no difference between the population means.
- You have found the probability of the null hypothesis being true.
- You have absolutely proved your experimental hypothesis.
- You can deduce the probability of the experimental hypothesis being true.
- You know, if you decided to reject the null hypothesis, the probability that you are making the wrong decision.
- You have a reliable experiment in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

The subjects were all university lecturers, research fellows, or postgraduate students. The results presented in Table 1.1 are illuminating.

Under a relative frequency view of probability, all six statements are in fact false. Only 3 out of the 70 subjects came to this conclusion. The correct interpretation of the probability associated with the observed t-value is

TABLE 1.1

Frequencies and Percentages of “True” Responses in Test of Knowledge about p-Values

Statement	Frequency	Percentage
1. The null hypothesis is absolutely disproved.	1	1.4
2. The probability of the null hypothesis has been found.	25	35.7
3. The experimental hypothesis is absolutely proved.	4	5.7
4. The probability of the experimental hypothesis can be deduced.	46	65.7
5. The probability that the decision taken is wrong is known.	60	85.7
6. A replication has a 0.99 probability of being significant.	42	60.0

The probability of obtaining the observed data (or data that represent a more extreme departure from the null hypothesis) if the null hypothesis is true.

Clearly, the number of false statements described as true in this experiment would have been reduced if the true interpretation of a p-value had been included with the six others. Nevertheless, the exercise is extremely interesting in highlighting the misguided appreciation of p-values held by a group of behavioral researchers.

The second reason for researchers and students to be careful using p-values is that a p-value represents only limited information about the results from a study. Gardner and Altman (1986) make the point that the excessive use of p-values in hypothesis testing, simply as a means of rejecting or accepting a particular hypothesis at the expense of other ways of assessing results, has reached such a degree that levels of significance are often quoted alone in the main text and abstracts of papers with no mention of other more relevant and important quantities. The implications of hypothesis testing—that there can always be a simple “yes” or “no” answer as the fundamental result from a research study—is clearly false and, used in this way, hypothesis testing is of limited value.

The most common alternative to presenting results in terms of p-values in relation to a statistical null hypothesis is to estimate the magnitude of some parameter of interest along with some interval that includes the population value of the parameter with a specified probability. Such confidence intervals can be found relatively simply for many quantities of interest (see Gardner and Altman, 1986, for details), and although the underlying logic of interval estimation is essentially similar to that of significance tests, they do not carry with them the pseudoscientific hypothesis testing language of such tests. Instead, they give a plausible range of values for the unknown parameter. As Oakes (1986) rightly comments: “The significance test relates to what the population parameter is not: the confidence interval gives a plausible range for what the parameter is.”

So, should the p-value be abandoned completely? Many statisticians would, grumpily, answer yes, but I think a more sensible response, at least for behavioral scientists, would be a resounding “maybe.” The p-value should rarely be used in a purely confirmatory way, but in an exploratory fashion, p-values can be useful in giving some informal guidance on the possible existence of an interesting effect even when the required assumptions of whatever test is being used are known to be invalid. It is often possible to assess whether a p-value is likely to be an under- or overestimate, and whether the result is clear one way or the other. In this text, both p-values and confidence intervals will be used; purely from a pragmatic point-of-view, the former are needed by behavioral students since they remain of central importance in the bulk of the behavioral science literature.

1.8 Summary

- Statistical principles are central to most aspects of a psychological investigation.
 - Data and their associated statistical analyses form the evidential parts of behavioral science arguments.
 - Significance testing is far from the be-all and end-all of statistical analyses, but it does still matter because evidence that can be discounted as an artifact of sampling will not be particularly persuasive. But p-values should not be taken too seriously; confidence intervals are often more informative.
 - Good statistical analysis should highlight those aspects of the data that are relevant to the substantive arguments; do so clearly and fairly, and be resistant to criticisms.
 - Experiments lead to the clearest conclusions about causal relationships.
 - Variable type often determines the most appropriate method of analysis, although some degree of flexibility should be allowed.
 - Sample size determination to achieve some particular power is an important exercise when designing a study, but the result of the statistical calculation involved needs to be considered in terms of what is feasible from a practical viewpoint.
-

1.9 Exercises

- 1.1 The Pepsi-Cola Company carried out research to determine whether people tended to prefer Pepsi Cola to Coca Cola. Participants were asked to taste two glasses of cola and then state which they preferred. The two glasses were not labeled Pepsi or Coke for obvious reasons. Instead, the Coke glass was labeled Q, and the Pepsi glass was labeled M. The results showed that “more than half chose Pepsi over Coke” (Huck and Sandler, 1979, p. 11). Are there any alternative explanations for the observed difference, other than the taste of the two drinks? Explain how you would carry out a study to assess any alternative explanation you think possible.
- 1.2 You develop a headache while working for hours at your computer. You stop, go into another room, and take two aspirins. After about 15 min, your headache has gone and you return to work. Can you infer a definite causal relationship between taking the aspirin and curing the headache? If not, why not?

- 1.3 You are interested in assessing whether or not laws that ban purchases of handguns by convicted felons reduce criminal violence. What type of study would you carry out, and how would you go about the study?
- 1.4 In reading about the results of an intervention study, you find that alternate subjects have been allocated to the two treatment groups. How would you feel about the study and why?
- 1.5 Attribute the following quotations about statistics and statisticians:
- a. To understand God's thoughts we must study statistics, for these are a measure of his purpose.
 - b. You cannot feed the hungry on statistics.
 - c. A single death is a tragedy, a million deaths is a statistic.
 - d. Thou shall not sit with statisticians nor commit a Social Science.
 - e. Facts speak louder than statistics.
 - f. I am one of the unpraised, unrewarded millions without whom statistics would be a bankrupt science. It is we who are born, marry, and who die in constant ratios.
- 1.6 What is the ratio of two measurements of warmth, one of which is 25°C and the other of which is 110°F ?

2

Looking at Data

2.1 Introduction

According to Chambers et al. (1983), “there is no statistical tool that is as powerful as a well-chosen graph.” Certainly, graphical display has a number of advantages over tabular displays of numerical results, not least in creating interest and attracting the attention of the viewer.

But just what is a graphical display? A concise description is given by Tufte (1983):

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

Graphical displays are very popular; it has been estimated that between 900 billion (9×10^{11}) and 2 trillion (2×10^{12}) images of statistical graphics are printed each year. Perhaps one of the main reasons for such popularity is that graphical presentation of data often provides the vehicle for discovering the unexpected; the human visual system is very powerful in detecting patterns, although the following caveat from the late Carl Sagan should be kept in mind:

Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.

Some of the advantages of graphical methods have been listed by Schmid (1954):

- In comparison with other types of presentation, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time since the essential meaning of large measures of statistical data can be visualized at a glance.
- Charts and graphs provide a comprehensive picture of a problem that makes for a more complete and better-balanced understanding than could be derived from tabular or textual forms of presentation.

- Charts and graphs can bring out hidden facts and relationships, and can stimulate, as well as aid, analytical thinking and investigation.

Schmid's last point is reiterated by the legendary John Tukey in his observation that "the greatest value of a picture is when it forces us to notice what we never expected to see."

The prime objective of a graphical display is to communicate to ourselves and others. Graphic design must do everything it can to help people understand. In some cases a graphic is required to give an overview of the data and perhaps to tell a story about the data. In other cases a researcher may want a graphical display to suggest possible hypotheses for testing on new data and, after some model has been fitted to the data, a graphic that criticizes the model may be what is needed. In this chapter we will consider graphics primarily from the story-telling angle; graphics that help check model assumption will be discussed in later chapters.

During the last two decades, a wide variety of new methods for displaying data graphically have been developed; these will hunt for special effects in data, indicate outliers, identify patterns, diagnose models, and generally search for novel and perhaps unexpected phenomena. Large numbers of graphs may be required, and computers are generally needed to supply them for the same reasons they are used for numerical analyses, namely, they are fast and they are accurate.

So, because the machine is doing the work, the question is no longer "shall we plot?" but rather "what shall we plot?" There are many exciting possibilities, including dynamic graphics (see Cleveland and McGill, 1987), but graphical exploration of data usually begins with some simpler, well-known methods, and it is these that we deal with in Section 2.2.

2.2 Simple Graphics—Pie Charts, Bar Charts, Histograms, and Boxplots

2.2.1 Categorical Data

Newspapers, television, and the media in general are very fond of two very simple graphics for displaying categorical data, namely, the pie chart and the bar chart. Both can be illustrated using the data shown in [Table 2.1](#), which show the percentage of people convicted of five different types of crime. In the pie charts for drinkers and abstainers (see [Figure 2.1](#)), the sections of the circle have areas proportional to the observed percentages. In the corresponding bar charts (see [Figure 2.2](#)), percentages are represented by rectangles of appropriate size placed along a horizontal axis.

TABLE 2.1
Crime Rates for Drinkers and Abstainers

Crime	Drinkers	Abstainers
Arson	6.6	6.4
Rape	11.7	9.2
Violence	20.6	16.3
Stealing	50.3	44.6
Fraud	10.8	23.5

Note: Figures are percentages.

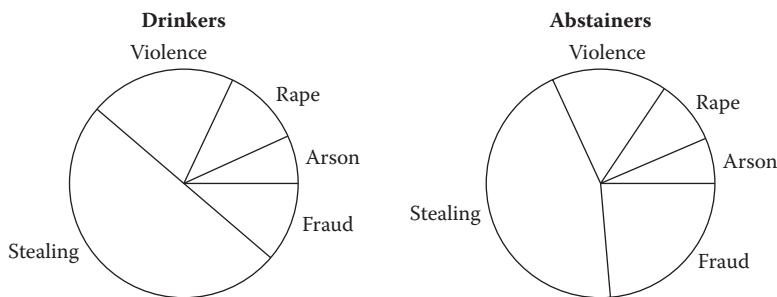


FIGURE 2.1
Pie charts for drinkers' and abstainers' crime percentages.

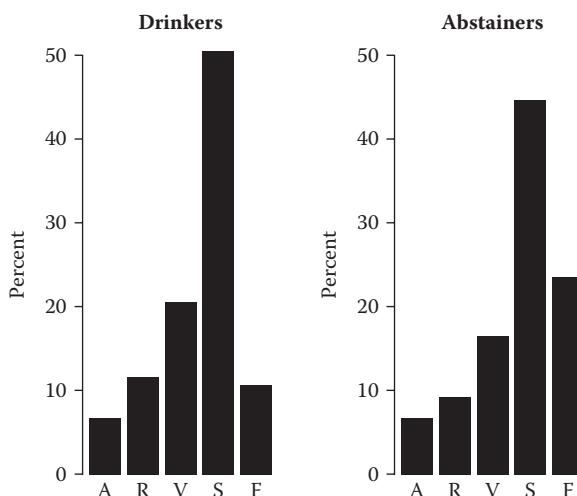


FIGURE 2.2
Bar charts for drinkers' and abstainers' crime percentages.

Despite their widespread popularity, both the general and, in particular, the scientific use of pie charts has been severely criticized. Tufte (1983), for example, comments that “tables are preferable to graphics for many small data sets. A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them ... pie charts should never be used.” A similar lack of affection is shown by Bertin (1981), who declares that “pie charts are completely useless,” and more recently by Wainer (1997), who claims that “pie charts are the least useful of all graphical forms.” Certainly in regard to the data in [Table 2.1](#), the numerical data in the table are as informative, or perhaps even more informative, than the associated pie charts in [Figure 2.1](#), and the bar chart in [Figure 2.2](#) seems no more necessary than the pie chart for these data.

Two examples that illustrate why both pie charts and bar charts are often (but not always—as will be seen later) of little more help in understanding categorical data than the numerical data themselves and how other graphics are frequently more useful are given in Cleveland (1994). The first example compares the pie chart of 10 percentages with an alternative graphic, the dot plot. The plots are shown in [Figures 2.3](#) and [2.4](#). The 10 percentages represented by the two graphics have a bimodal distribution; odd-numbered observations cluster around 8%, and even-numbered observations cluster around 12%. Furthermore, each even value is shifted with respect to the preceding odd value by about 4%. This pattern is far easier to spot in the dot plot than in the pie chart.

Dot plots for the crime data in [Table 2.1](#) are shown in [Figure 2.5](#), and these are also more informative than the corresponding pie charts in [Figure 2.1](#).

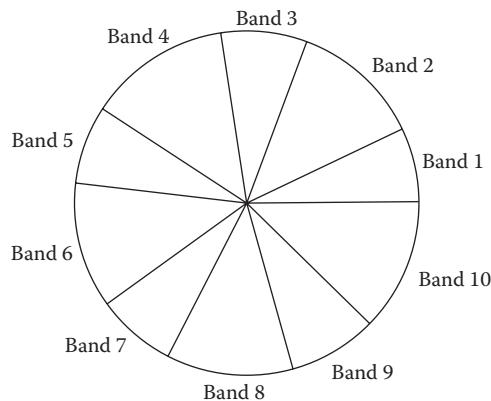


FIGURE 2.3

Pie chart for 10 percentages. (Suggested by Cleveland, 1994. Used with permission from Hobart Press.)

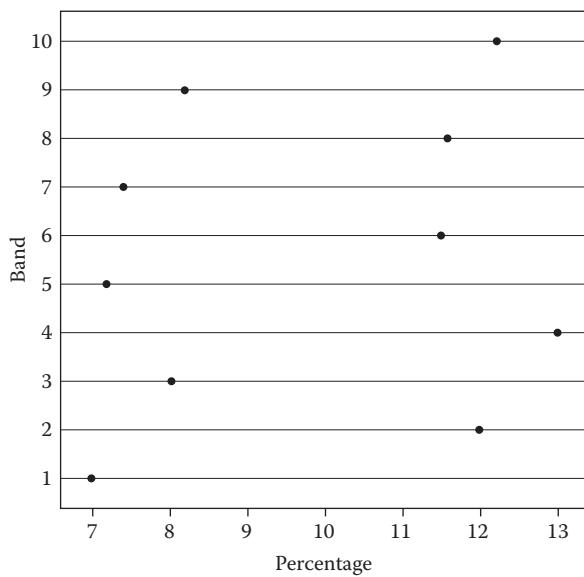


FIGURE 2.4
Dot plot for 10 percentages.

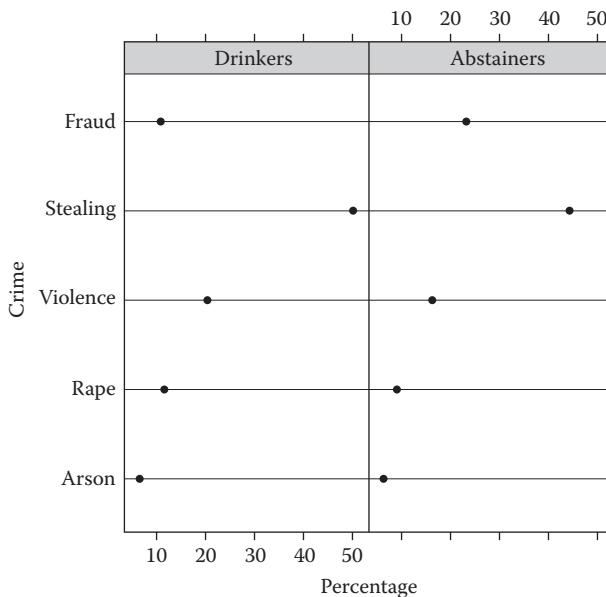


FIGURE 2.5
Dot plots for drinkers' and abstainers' crime percentages.

The second example given by Cleveland begins with the diagram shown in Figure 2.6, which originally appeared in Vetter (1980). The aim of the diagram is to display the percentages of degrees awarded to women in several disciplines of science and technology during three time periods. At first glance the labels on the diagram suggest that the graph is a standard divided bar chart with the length of the bottom division of each bar showing the percentage for doctorates, the length of the middle division showing the percentage for master's degrees, and the top division showing the percentage for bachelor's degrees. A little reflection shows that this interpretation is not correct since it would imply that, in most cases, the percentage of bachelor's degrees given to women is lower than the percentage of doctorates. Closer examination of the diagram reveals that the three values of the data for each discipline during each time period are determined by the three adjacent vertical dotted lines. The top of the left-hand line indicates the value for doctorates, the top end of the middle line indicates the value for master's degrees, and the top end of the right-hand line indicates the value for bachelor's degrees.

Cleveland (1994) discusses other problems with the diagram in Figure 2.6; in particular, he points out that the manner of the diagram's construction makes it hard to connect visually the three values of a particular type of degree for a specific discipline, thus making it difficult to see changes over time.

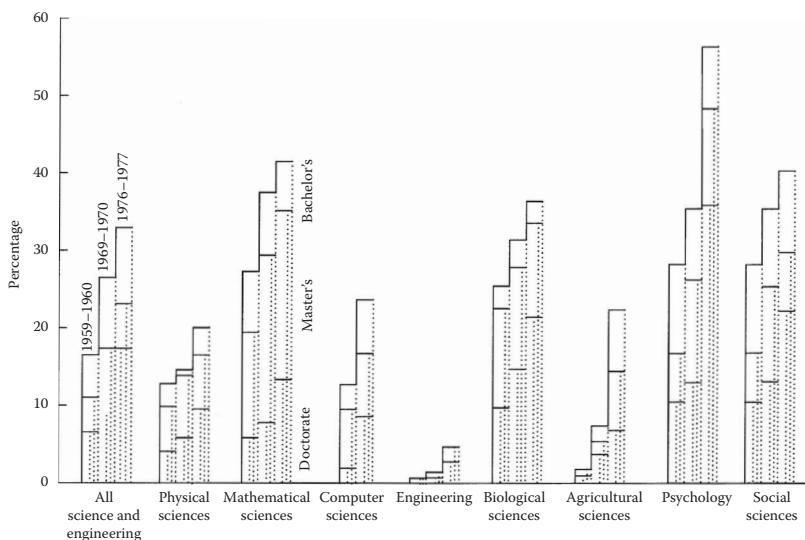
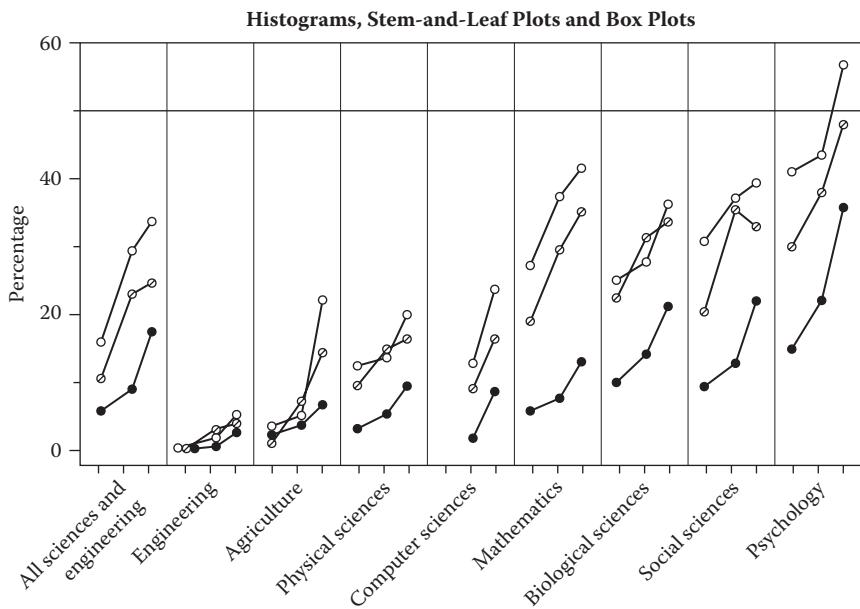


FIGURE 2.6

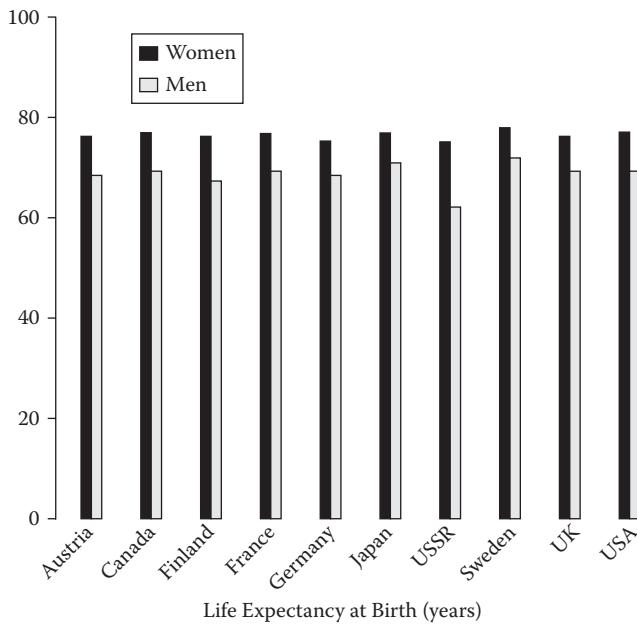
Proportion of degrees in science and technology earned by women in the periods 1959–1960, 1969–1970, and 1976–1977. (Reproduced with permission from Vetter, B. M., 1980, *Science*, 207, 28–34. © 1980 American Association for the Advancement of Society.)

**FIGURE 2.7**

Percentage of degrees earned by women for three degrees (○ bachelor's degree; □ master's degree; • doctorate), three time periods, and nine disciplines. The three points for each discipline and degree indicate the periods 1959–1960, 1969–1970, and 1976–1977.

Figure 2.7 shows the data represented by [Figure 2.6](#) replotted by Cleveland in a bid to achieve greater clarity. It is now clear how the data are represented, and this diagram allows viewers to see easily the percentages corresponding to each degree, in each discipline, over time. Finally the figure caption explains the content of the diagram in a comprehensive and clear fashion. All in all Cleveland appears to have produced a graphic that would satisfy even that doyen of graphical presentation, Edward R. Tufte, in his demand that “excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency.”

Wainer (1997) gives a further demonstration of how displaying data as a bar chart can disguise rather than reveal important points about data. [Figure 2.8](#) shows a bar chart of life expectancies in the middle 1970s, divided by sex, for ten industrialized nations. The order of presentation is alphabetical (with the U.S.S.R. positioned as Russia). The message we get from this diagram is that there is little variation and women live longer than men. But by displaying the data in the form of a simple stem-and-leaf plot (see [Figure 2.9](#)), the magnitude of the sex difference (7 years) is immediately clear as is the unusually short life expectancy for men in the U.S.S.R., whereas Russian women have life expectancy similar to women in other countries.

**FIGURE 2.8**

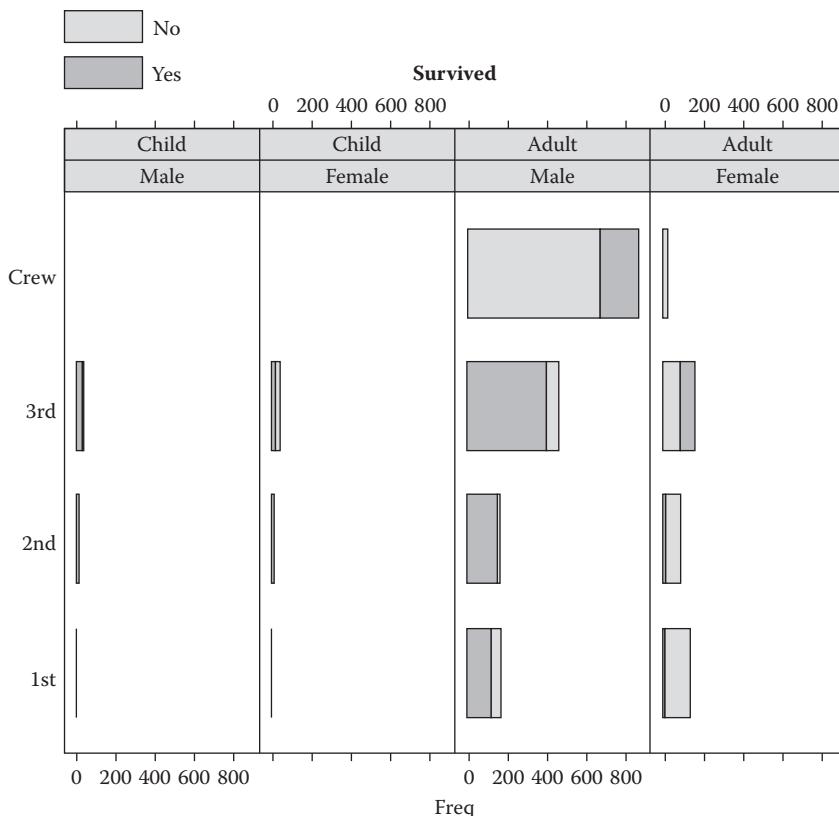
Bar chart showing life expectancies at birth by sex and by country.

Women	Age	Men
Sweden	78	
France, US, Japan, Canada	77	
Finland, Austria, UK	76	
USSR, Germany	75	
	74	
	73	
	72	Sweden
	71	Japan
	70	
	69	Canada, UK, US, France
	68	Germany, Austria
	67	Finland
	66	
	65	
	64	
	63	
	62	USSR

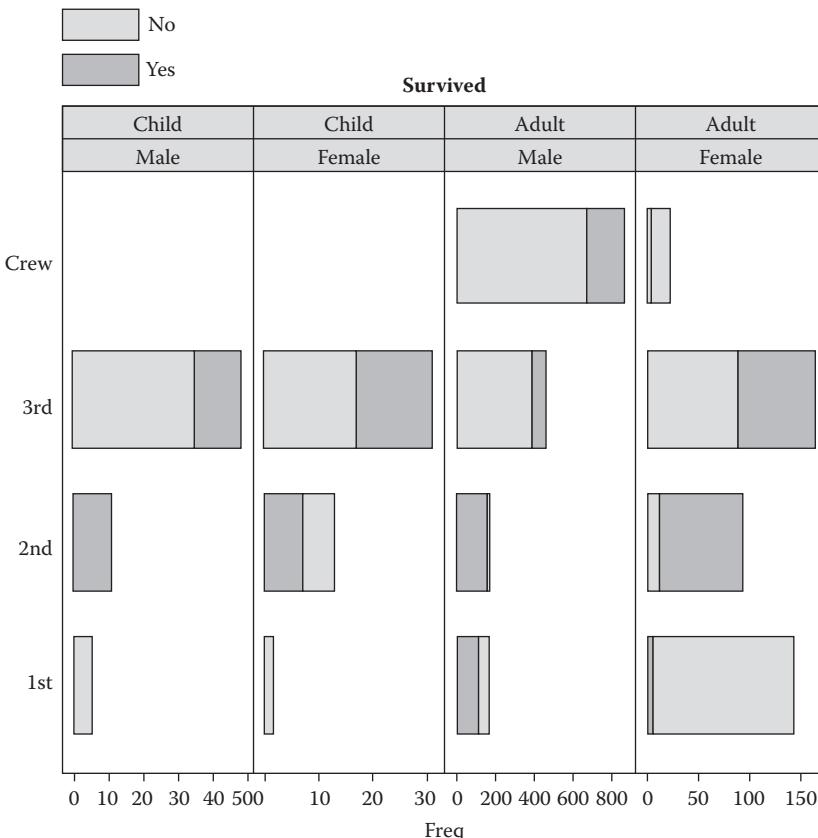
FIGURE 2.9

An alternative display of life expectancies at birth by sex and by country

To be fair to the poor old bar chart, we will end this subsection by illustrating how a sophisticated adaptation of the graphic can become an extremely effective tool for displaying a complex set of categorical data. The example is taken from Sarkar (2008) and uses data summarizing the fates of the 2201 passengers on the Titanic. The data are categorized by economic status (class of ticket, 1st, 2nd, or 3rd, or crew), sex (male or female), age (adult or child), and whether they survived or not (the data are available on Sarkar's Web site, <http://lmdv.r-forge.r-project.org/>). The first diagram produced by Sarkar (using R—again the code is on the Web site) is shown in Figure 2.10. This plot looks impressive but is dominated by the third “panel” (adult males) as heights of bars represent counts, and all panels have the same limits. So, sadly, all the plot tells us is that there were many more males than females aboard (particularly among the crew, which is the largest group), and that there were even fewer children. The plot becomes more illuminating about what really happened to the passengers if the proportion of survivors is plotted and by allowing

**FIGURE 2.10**

A bar chart summarizing the fate of passengers of the Titanic, classified by sex, age, and whether they survived or not.

**FIGURE 2.11**

Survival among different different subgroups of passengers on the Titanic, with a different horizontal scale in each panel.

independent horizontal scales for the different “panels” in the plot; this plot is shown in Figure 2.11. This plot emphasizes the proportion of survivors within each subgroup rather than the absolute numbers. The proportion of survivors is lowest among third-class passengers, and the diagram makes it very clear that the “women and children first” policy did not work very well for this class of passengers.

2.2.2 Interval/Quasi-Interval Data

The data shown in Table 2.2 come from an observational study described in Howell and Huessy (1981, 1985), in which a researcher asked 50 children to tell her about a given movie. For each child the researcher recorded the number of “and then ...” statements.

TABLE 2.2

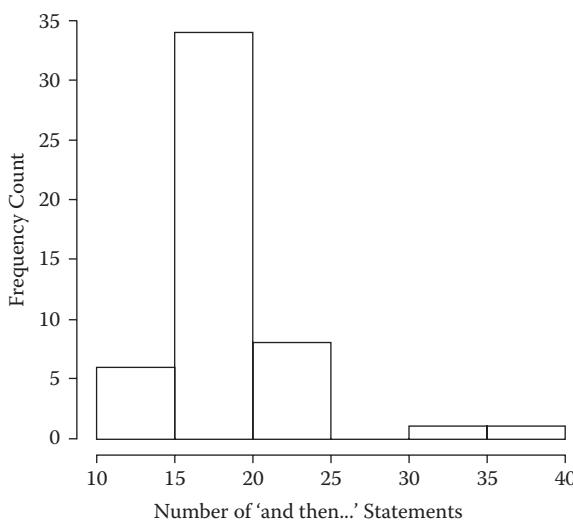
Number of “and Then ...” Statements Made by 50 Children Recalling the Story of a Movie They Had Just Seen

18	15	22	19	18	17	18	20	17	12	16	16	17	21	23	18	20	21	20	20	15	18	17	19	20
23	22	10	17	19	19	21	20	18	18	24	11	19	31	16	17	15	19	20	18	18	40	18	19	16

Let us begin by constructing that old favorite, the histogram, for these data; it is shown in Figure 2.12. Here the histogram is based on a relatively small number of observations and tells us little about the data except that there is some degree of skewness perhaps and possibly two “outliers.”

The histogram is a widely used graphical method that is at least a 100 years old. But Cleveland (1994) makes the point that “maturity and ubiquity do not guarantee the efficacy of a tool.” The histogram is generally used for two purposes: counting and displaying the distribution of a variable. However, according to Wilkinson (1992), “it is effective for neither.” Histograms can often be misleading about a variable’s distribution because of their dependence on the number of classes chosen, and simple tallies of the observations to give a numerical frequency distribution table are usually preferable for counting. Finally, the histogram is a poor method for comparing groups of univariate measurements (Cleveland, 1994).

A more useful graphical display of a variable’s distributional properties is the boxplot. This is obtained from the five-number summary of a data set, the five numbers in question being the minimum, the lower quartile (LQ), the median, the upper quartile (UQ), and the maximum. The distance between the upper

**FIGURE 2.12**

Histogram of count of “and then ...” statements by 50 children.

and lower quartiles, the interquartile range (IQR), is a measure of the spread of a variable's distribution. The relative distances from the median of the upper and lower quartiles give information about the shape of a variable's distribution; for example, if one distance is much greater than the other, the distribution is skewed. In addition, the median and the upper and lower quartiles can be used to define arbitrary but often useful limits, L and U, that maybe helpful in identifying possible outliers. The two limits are calculated as follows:

$$U = UQ + 1.5 \text{ IQR}$$

$$L = LQ - 1.5 \text{ IQR}$$

Observations outside the limits can be regarded as potential outliers (they are sometimes referred to specifically as outside observations), and such observations may merit careful attention before undertaking any analysis of a data set because there is always the possibility that they can have undue influence on the results of the analysis.

The construction of a boxplot is described in Figure 2.13.

The boxplot of the data in [Table 2.2](#) is shown in [Figure 2.14](#). The diagram indicates a number of possible outliers and also highlights the skewness in the data.

In [Table 2.3](#), there is a similar data set as in [Table 2.2](#), but here the observations were collected from adults. A question of interest is whether children and adults recall stories in the same way. At some stage this may require a formal procedure such as the construction of a confidence interval for, say, the mean difference in the number of "and then ..." statements between children and adults. But here we will see how far we can get with a graphical approach namely comparing the boxplots of each data set. The side-by-side boxplots are shown in [Figure 2.15](#). The diagram clearly demonstrates that the

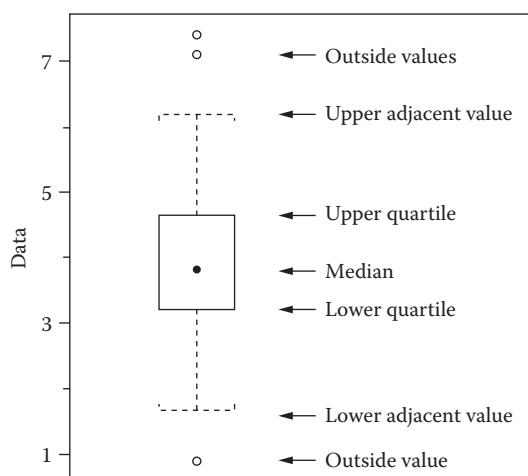


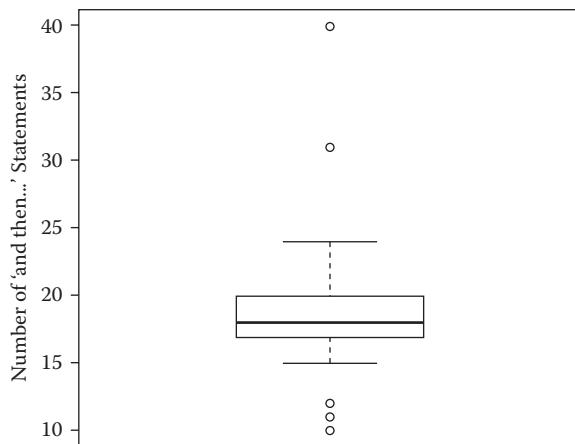
FIGURE 2.13

The construction of a boxplot.

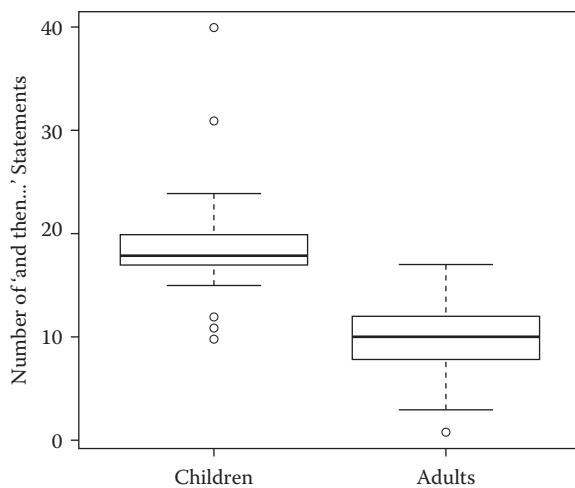
TABLE 2.3

Number of “and Then ...” Statements Made by 50 Adults Recalling the Story of a Movie They had Just Seen

10	12	5	8	13	10	12	8	7	11	11	10	9	9	11	15	12	17	14	10	9	8	15	16	10
14	7	16	9	1	4	11	12	7	9	10	3	11	14	8	12	5	10	9	7	11	14	10	15	9

**FIGURE 2.14**

Boxplot of count of “and then ...” statements by children.

**FIGURE 2.15**

Side-by-side boxplots of counts of “and then ...” statements by children and adults.

adults generally use less “and then ...” statements than children and also suggests that the distribution of the adults’ observations is closer to being symmetric than that of the children.

Although the boxplots in Figure 2.15 give some information about the distributions of the observations in each group, it may be useful to delve a little further and use probability plots to assess the normality of the observations for both children and adults. Probability plots are described in Technical Section 2.1.

Technical Section 2.1: Probability Plots

The classic example of such a plot is that for investigating the assumption that a set of data is from a normal distribution; here the ordered sample values, $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ are plotted against the quantiles of a standard normal distribution, that is, $\Phi^{-1}(p_i)$, where

$$p_i = \frac{i - \frac{1}{2}}{n}, \quad \text{and} \quad \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{u^2}{2}\right)} du$$

Departures from linearity in the plot indicate that the data do not have a normal distribution.

The normal probability plots for the children and the adults are shown in Figure 2.16. The plot for the children’s observations shows a marked

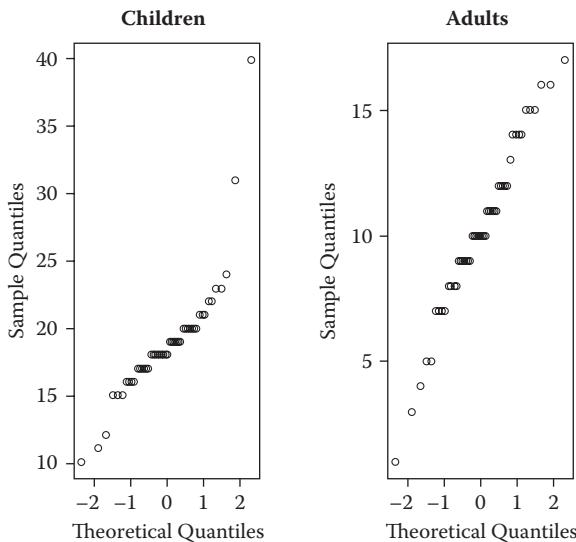


FIGURE 2.16

Probability plots of counts of “and then ...” statements by children and adults.

departure from linearity but the plot for the adults' data looks linear. These findings might need to be considered before any formal test or procedure is applied to the data set, although here constructing the usual normality based confidence interval is unlikely to be misleading.

Probability plots have been around for a long time, but they remain a useful technique for assessing distributional assumptions in some cases as here for the raw data, but also for the residuals that are used to assess the assumptions when fitting models to data, as we shall see in Chapter 3.

2.3 The Scatterplot and Beyond

The simple xy scatterplot has been in use since at least the 18th century and has many advantages for an initial exploration of data. Indeed, according to Tufte (1983):

The relational graphic—in its barest form the scatterplot and its variants—is the greatest of all graphical designs. It links at least two variables encouraging and even imploring the newer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y .

Let us begin by looking at a straightforward use of the scatterplot using some of the data in Table 2.4. These data were collected from a sample of 24 primary school children in Sydney, Australia. Part of the data is given in Table 2.4. Each child completed the Embedded Figures Test (EFT), which measures “field dependence,” that is, the extent to which a person can abstract the logical structure of a problem from its context. Then the children were allocated to one of two experimental groups, and they were timed as they constructed a 3×3 pattern from nine colored blocks, taken from the Wechsler Intelligence Scale for Children (WISC). The two groups differed in the instructions they were given for the task: the “row” group was told

TABLE 2.4

Field Dependence Measure and Time to Complete a Task from the WISC for Children in Two Experimental Groups

Row Group			Corner Group		
Child	Time	EFT	Child	Time	EFT
1	317	59	1	342	43
2	464	33	2	222	23
3	525	49	3	219	9
4	298	69	4	513	128
5	491	65	5	295	44

to start with a row of three blocks, and the “corner” group was told to start with a corner of three blocks. The experimenter was interested in whether the different instructions produced any change in the average time to complete the picture and whether this time was affected by field dependence. We shall analyze these data more formally in Chapter 4; here we will examine primarily the relationship between completion time and EFT.

So, to begin, Figure 2.17 shows the scatterplots of completion time against EFT for both the row and the corner groups. The first thing to notice about the two plots is the obvious outlier in the plot for the row experimental group, and the relationship between time to completion and EFT appears to be stronger in the row group than in the corner group although the outlier in the row group may be entirely responsible for this apparent difference. Other than this, the plots are not particularly informative, and we perhaps cannot expect too much from them given the rather small samples involved. We could, of course, calculate correlations, but this is left as an exercise for the reader (see Exercise 2.1). Here we wish to concentrate on the information we can get from graphics alone. It should be remembered that calculating a correlation coefficient between two variables without looking at the corresponding scatterplot is very poor data analysis practice because a correlation coefficient can, on occasions, be badly affected by outliers, and the scatterplot is needed to spot the offending observations. An example will be given later.

To make the scatterplots a little more informative, we can add the linear regression fit (see Chapter 3) of time to completion against EFT to each plot. The result is Figure 2.18. The plots now demonstrate more clearly that completion time appears to have a stronger linear relationship to EFT in the row group than in the corner group (but remember that outlier).

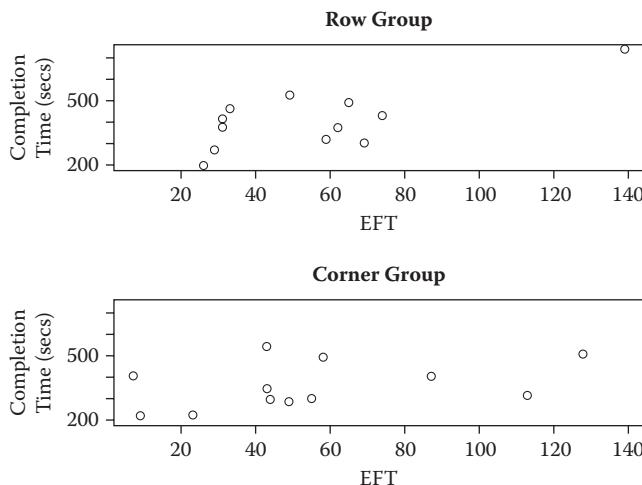
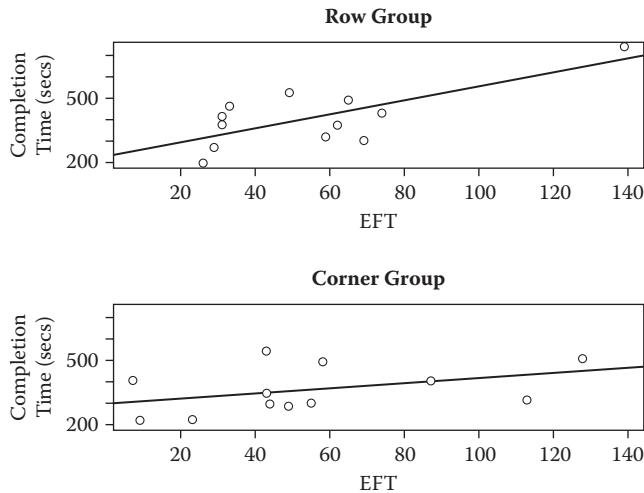


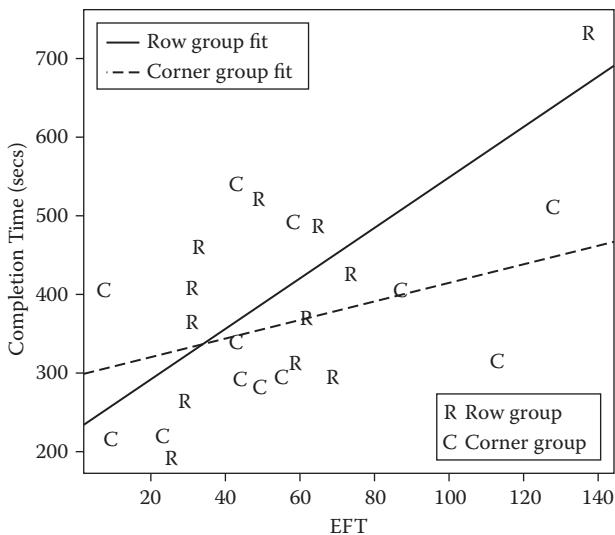
FIGURE 2.17

Scatterplots of time to completions against EFT score for row and corner groups.

**FIGURE 2.18**

Scatterplots of time to completion against EFT for row and corner groups with added linear regression fit.

Another possible way to plot the data in [Table 2.4](#) is to simply combine all the data in one scatterplot, identifying the row and corner group observations in some way. This is what we have done in Figure 2.19.

**FIGURE 2.19**

Scatterplot of completion time against EFT with observations labeled as row or corner group and linear regression fits for each group shown.

TABLE 2.5

Data Collected About Time Spent Looking After Car for First Five Subjects Out of 40

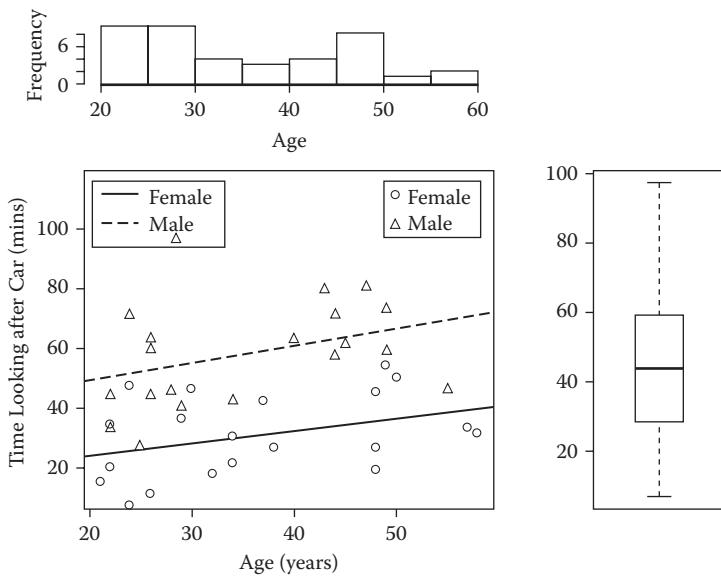
Subject	Sex	Age	Extro	Time
1	Male	55	40	46
2	Male	43	45	79
3	Female	57	52	33
4	Male	26	62	63
5	Female	22	31	20

Now let us move on to consider a larger set of data, part of which is given in Table 2.5. These data are taken from Miles and Shevlin (2001), and give the sex, age, extroversion score, and the average number of minutes per week a person spends looking after his or her car, for 40 people. People may project their self-image through themselves or through the objects they own, such as their cars. Therefore, a theory could be developed that predicts that people who score higher on a measure of extroversion are likely to spend more time looking after their cars. This possibility will be examined in Chapter 4; here we will see how much information about the data we can derive from some scatterplots. Any information about the data collected at this point may be very helpful in fitting formal models to the data.

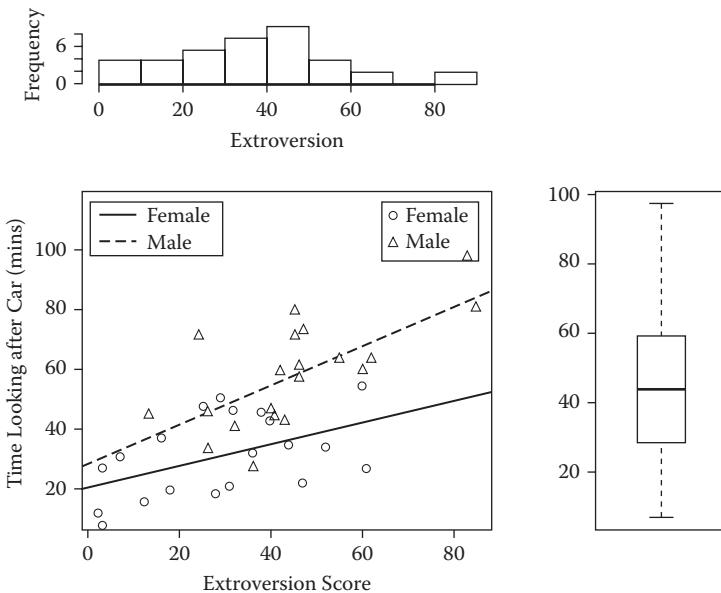
To begin, we shall construct scatterplots of time spent by people looking after their cars, against age and extroversion score. Often when using scatterplots to look at data, it is helpful to add something about the marginal distributions of the two variables, and this we will do here. Further, we will add to each plot the appropriate linear regression fits separately for men and women. The two plots are shown in [Figures 2.20](#) and [2.21](#). The plot in [Figure 2.20](#) shows that the relationship between time spent looking after car and age is approximately the same for men and women and time increases a little with age. [Figure 2.21](#) shows that time spent looking after car increases with an increase in extroversion score for both men and women, but that the increase appears greater for men. This has implications for how these data might be modeled as we shall see in Chapter 4.

2.3.1 The Bubbleplot

The scatterplot can only display two variables. However, there have been a number of suggestions as to how extra variables may be included. In this subsection we shall illustrate one of these, the bubbleplot, in which three variables are displayed; two are used to form the scatterplot itself, and then the values of the third variable are represented by circles with radii proportional to these values and centered on the appropriate point in the scatterplot.

**FIGURE 2.20**

Scatterplot of time spent looking after car, against age, showing marginal distributions of the two variables and fitted linear regressions for men and women.

**FIGURE 2.21**

Scatterplot of time spent looking after car, against extroversion, showing marginal distributions of the two variables and fitted linear regressions for men and women.

For the data in [Table 2.5](#), Figure 2.22 shows a bubbleplot with time spent looking after the car, against age as the scatterplot, and the extroversion scores represented by circles with appropriate radii. Gender is also displayed on the plot, so essentially, Figure 2.22 displays all four variables in the data set. Whether more information can be gleaned from this than from the plots given earlier is perhaps a moot point. But one observation does stand out: an approximately 30-year-old, extroverted man who spends almost 100 min per week looking after his car. Perhaps some counseling might be in order!

A plot a little like a bubbleplot is used by Bickel et al. (1975) to analyze the relationship between admission rate and the proportion of women applying to various academic departments at the University of California at Berkeley. The scatterplot of percentage of women applicants against percentage of applicants admitted is shown in [Figure 2.23](#); the plots are enhanced by “boxes,” the sizes of which indicate the relative number of applicants. The negative correlation indicated by the scatterplot is due almost exclusively to a trend for the large departments. If only a simple scatterplot had been used here, vital information about the relationship would have been lost.

2.3.2 The Bivariate Boxplot

A further helpful enhancement to the scatterplot is often provided by the two-dimensional analog of the boxplot for univariate data, known as the bivariate boxplot (Goldberg and Iglewicz, 1992). This type of boxplot may be

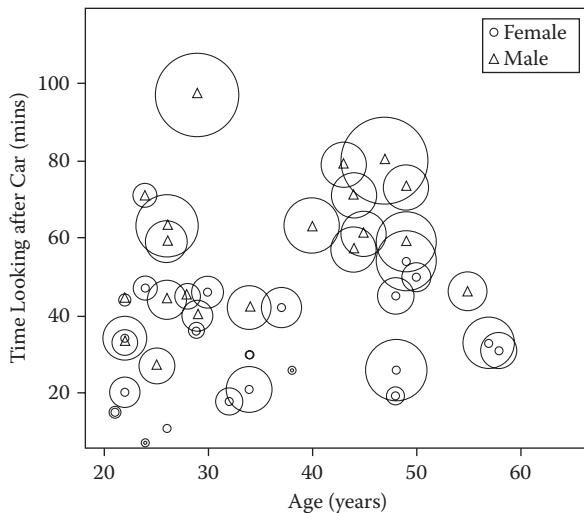
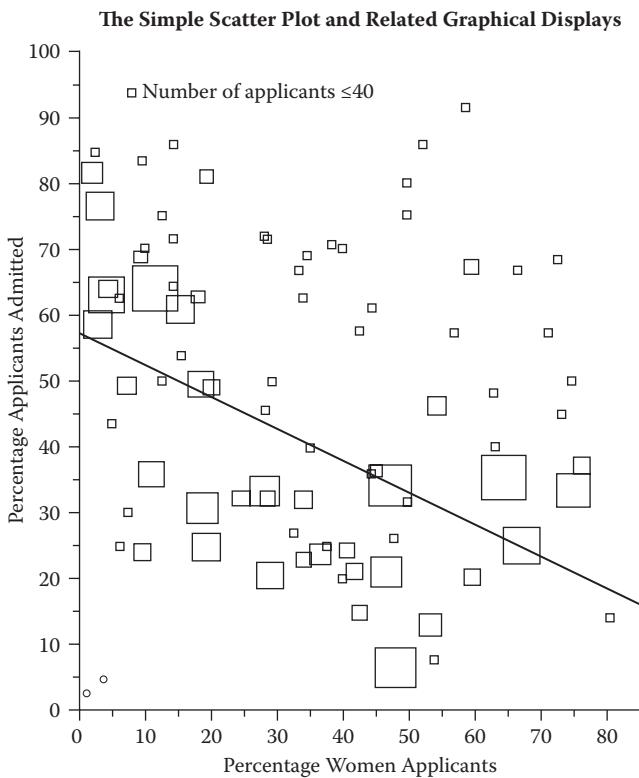


FIGURE 2.22

Bubbleplot of time spent looking after car, against age, with extroversion represented as circles.

**FIGURE 2.23**

Scatterplot of the percentage of female applicants versus percentage of applicants admitted to 85 departments at the University of California at Berkeley. (Reproduced with permission from Bickel, P. J. et al., 1975, *Science*, 187, 398–404.)

useful in indicating the distributional properties of the data and in identifying possible outliers. The bivariate boxplot is based on calculating robust measures of location, scale, and correlation; it consists essentially of a pair of concentric ellipses, one of which (the “hinge”) includes 50% of the data, and the other (called the “fence”) which delineates potential troublesome outliers. In addition, resistant regression lines of both y on x and x on y are shown, with their intersection showing the bivariate locations estimator. The acute angle between the regression lines will be small for a large absolute value of correlations and large for a small one. Details of the construction of a bivariate boxplot are given in Technical [Section 2.2](#).

Technical [Section 2.2](#): Constructing a Bivariate Boxplot

The bivariate boxplot is the two-dimensional analog of the familiar boxplot for univariate data and consists of a pair of concentric ellipses, the “hinge” and the “fence.” To draw the elliptical fence and hinge, location (T_x^*, T_y^*) , scale

(S_x^*, S_y^*) , and correlation (R^*) estimators are needed and, in addition, a constant D that regulates the distance of the fence from the hinge. In general, $D = 7$ is recommended, since this corresponds to an approximate 99% confidence bound on a single observation. In general, robust estimators of location, scale, and correlation are recommended since they are better at handling data with outliers, or with density or shape differing moderately from the elliptical bivariate normal. Goldberg and Iglewicz (1992) discuss a number of possibilities.

To draw the bivariate boxplot, first calculate the median E_m and the maximum E_{\max} of the standardized errors E_i , which are essentially the generalized distances of each point from the center (T_x^*, T_y^*) . Specifically, the E_i are defined by

$$E_i = \sqrt{\frac{X_{si}^2 + Y_{si}^2 - 2R^* X_{si} Y_{si}}{1 - R^{*2}}}$$

where $X_{si} = (X_i - T_x^*) / S_x^*$ is the standardized X_i value, and Y_{si} is similarly defined.

Then

$$E_m = \text{median } \{E_i : i = 1, 2, \dots, n\}$$

and

$$E_{\max} = \text{maximum } \{E_i : E_i^2 < DE_m^2\}$$

To draw the hinge, let $R_1 = E_m \sqrt{\frac{1+R^*}{2}}$, $R_2 = E_m \sqrt{\frac{1-R^*}{2}}$
For $\theta = 0$ to 360 in steps of 2, 3, 4, or 5 degrees, let

$$\Theta_1 = R_1 \cos \theta,$$

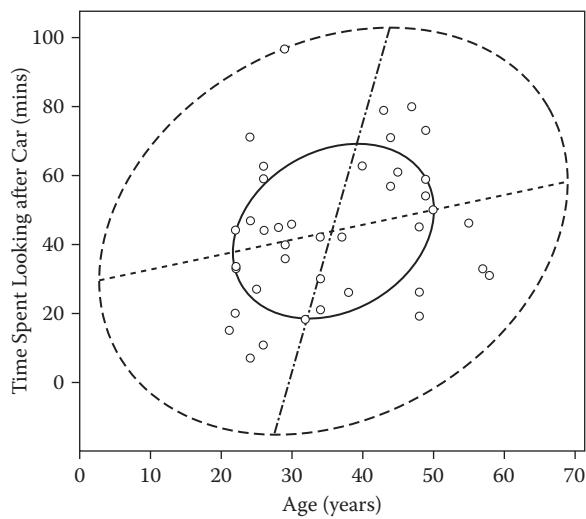
$$\Theta_2 = R_2 \sin \theta,$$

$$X = T_x^* + (\Theta_1 + \Theta_2) S_x^*,$$

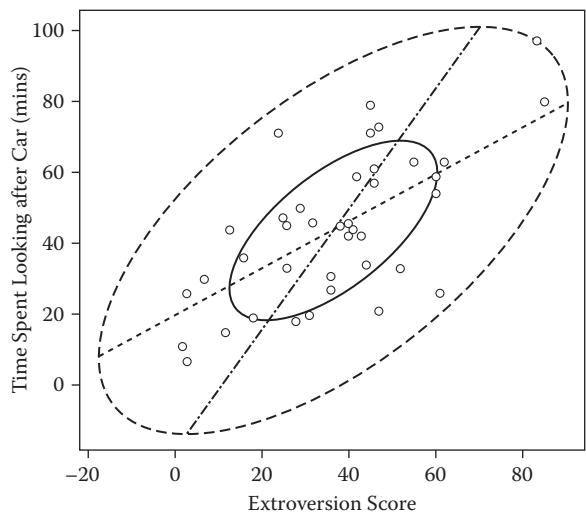
$$Y = T_y^* + (\Theta_1 - \Theta_2) S_y^*$$

Finally, plot X, Y.

[Figure 2.24](#) shows the bivariate boxplot of time spent looking after car and age, and [Figure 2.25](#) shows the corresponding diagram for time and extroversion. Neither diagram shows any clear outliers, that is, observations that fall outside the dotted ellipse. But in both diagrams there is one observation that lies on the dotted ellipse.

**FIGURE 2.24**

Bivariate boxplot of time spent looking after car and age.

**FIGURE 2.25**

Bivariate boxplot of time spent looking after car and extroversion.

2.4 Scatterplot Matrices

When there are many variables measured on all the individuals in a study, an initial examination of all the separate pairwise scatterplots becomes difficult. For example, if 10 variables are available, there are 45 possible scatterplots. But all these scatterplots can be conveniently arranged into a scatterplot matrix that then aids in the overall comprehension and understanding of the data.

A scatterplot matrix is defined as a square, symmetric grid of bivariate scatterplots. The grid has q rows and columns, each one corresponding to a different variable. Each of the grid's cells shows a scatterplot of two variables. Variable j is plotted against variable i in the ij th cell, and the same variables appear in cell ji with the x - and y -axes of the scatterplots interchanged. The reason for including both the upper and lower triangles of the grid, despite the seeming redundancy, is that it enables a row and a column to be visually scanned to see one variable against all others, with the scales for the one variable lined up along the horizontal or the vertical.

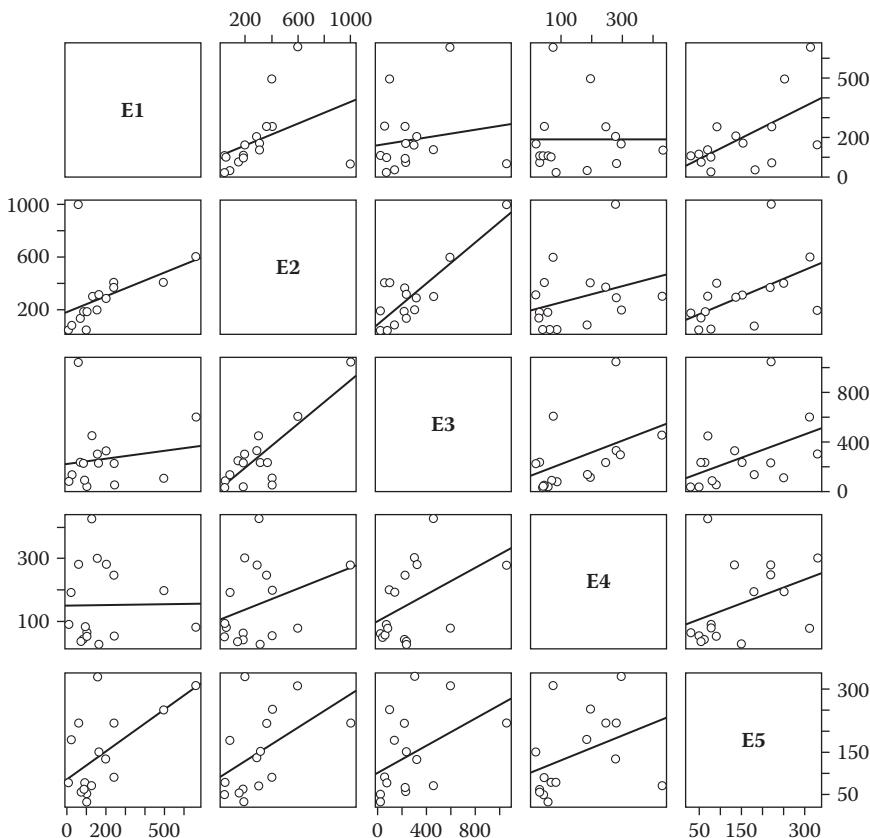
To illustrate the use of a scatterplot matrix, we shall use the data shown in Table 2.6. These data arise from an experiment in which five different types of electrode were applied to the arms of 16 subjects and the resistance measured (in kilohms). The experiment was designed to see whether all electrode types performed similarly. The scatterplot matrix for the data is shown in [Figure 2.26](#); each of the scatterplots in the diagram has been enhanced by the addition of the linear fit of the y variable on the x variable. The diagram suggests the presence of several outliers, the most extreme of which is subject 15; the reason for the two extreme readings on this subject was that he had very hairy arms. [Figure 2.26](#) also indicates that the readings on particular pairs of electrodes, for example, electrode 1 and electrode 4, are hardly related at all.

We can use the plot of results for the first and second electrodes to demonstrate how the bivariate boxplot looks when there are probable outliers in the data

TABLE 2.6

Measure of Resistance (Kilohms) Made on Five Different Types of Electrode for Five of the 16 Subjects

Subject	E1	E2	E3	E4	E5
1	500	400	98	200	250
2	660	600	600	75	310
3	250	370	220	250	220
4	72	140	240	33	54
5	135	300	450	430	70

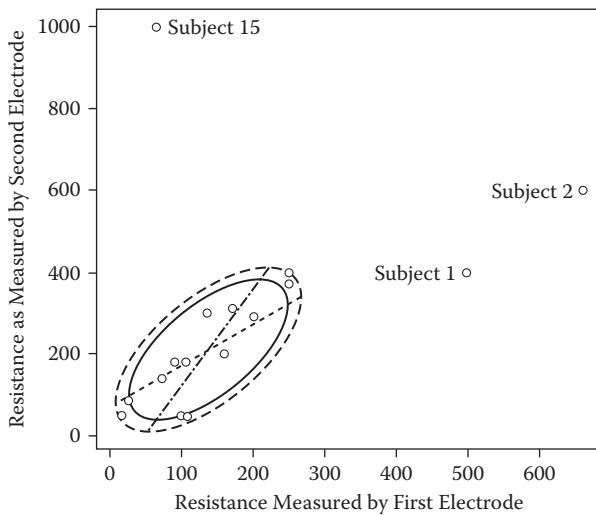
**FIGURE 2.26**

Scatterplot matrix for data on measurements of skin resistance made with five different types of electrodes.

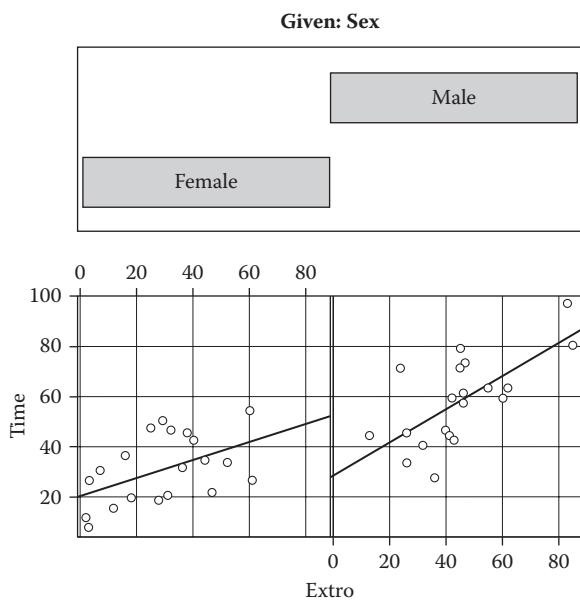
(see [Figure 2.27](#)). Three outliers are identified by the bivariate boxplot. If we calculate the correlation coefficient between the two variables using all the data, we get a value of 0.41; if we recalculate the correlation after removing subjects 1, 2, and 15, we get a value of 0.88—more than double the previous value. This example underlines how useful the bivariate boxplot can be, and also underlines the danger of simply calculating a correlation coefficient without examining the relevant scatterplot.

2.5 Conditioning Plots and Trellis Graphics

The conditioning plot or coplot is a potentially powerful visualization tool for studying how, say, a response variable depends on two or more explanatory variables. In essence, such plots display the bivariate relationship between

**FIGURE 2.27**

Bivariate boxplot for data on electrodes one and two.

**FIGURE 2.28**

Coplot of time spent looking after car against extroversion conditioned on sex.

two variables while holding constant (or “conditioning upon”) the values of one or more other variables. If the conditioning variable is categorical, then the coplot is no more than, say, a scatterplot of two variables for each level of the categorical variable. As an example of this type of simple coplot, Figure 2.28 shows plots of time spent looking after car against extroversion score conditioned on sex; each scatterplot is enhanced by a linear regression fit. The plot highlights what was found in an earlier plot (Figure 2.21)—that the relationship between time spent looking after car and extroversion is different for men and women.

As a more complicated coplot, Figure 2.29 shows time spent looking after car against extroversion conditioned on age. In this diagram, the panel at the top of the figure is known as the given panel; the panels below are dependence panels. Each rectangle in the given panel specifies a range of values of population size. On a corresponding dependence panel, time is plotted against age for those people whose ages lie in the particular interval. To match age intervals to dependence panels, the latter are examined in order

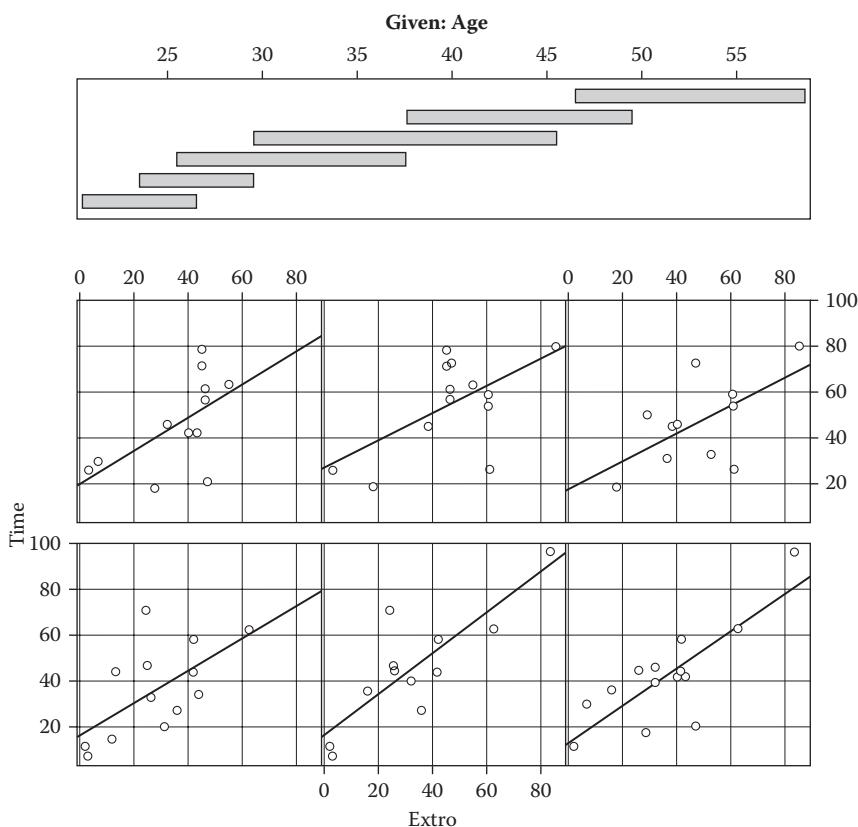


FIGURE 2.29
Coplot of time against extroversion conditioned on age.

from left to right in the bottom row and, then again, from left to right in subsequent rows. The plot suggests that the relationship between time and extroversion is much the same over the age range observed in the data set.

Conditional graphical displays are simple examples of a more general scheme known as trellis graphics (Becker and Cleveland, 1994). This is an approach to examining high-dimensional structure in data by means of one-, two-, and three-dimensional graphs. The problem addressed is how observations of one or more variables depend on the observations of the other variables. The essential feature of this approach is the multiple conditioning that allows some type of plot to be displayed for different values of a given variable (or variables). The aim is to help in understanding both the structure of the data and how well proposed models describe the structure. An excellent recent example of the application of trellis graphics is given in Verbyla et al. (1999).

As a relatively simple example of what can be done with trellis graphics, we will again use the data on time spent looking after car and produce a three-dimensional scatter plot for time, age, and extroversion conditioned on sex (see Figure 2.30). This diagram makes the generally longer times spent looking after their cars by men very apparent, although whether it adds anything to earlier plots is a question I leave for the reader.

Let us now look at two more complex examples of trellis graphics, taken from Sarkar (2008). The first involves data collected in a survey of doctorate degree recipients in the United States. The data are shown in [Table 2.7](#). Any graphic for the data has to involve the proportions of reasons across fields of study rather than the counts because the latter do not tell us very much, except, for example, that the “Biological Sciences” subject area contributes the majority of postdocs. A stacked bar chart of the data based on the proportions rather than the counts in [Table 2.7](#) is shown in [Figure 2.31](#). An alternative display for the proportions, a multipanel dot plot, is shown

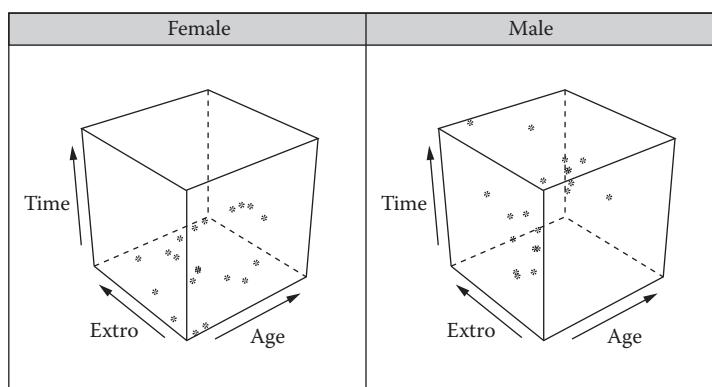


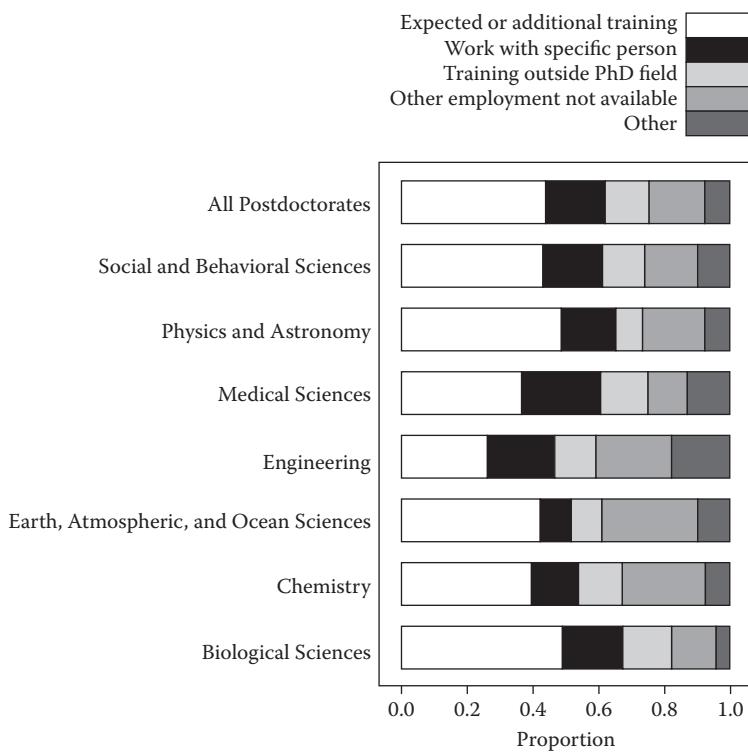
FIGURE 2.30

Three-dimensional scatterplot of time, age, and extroversion conditioned on sex.

TABLE 2.7

Reasons for Choosing a Postdoctoral Position After Graduating from U.S. Universities by Area of Study

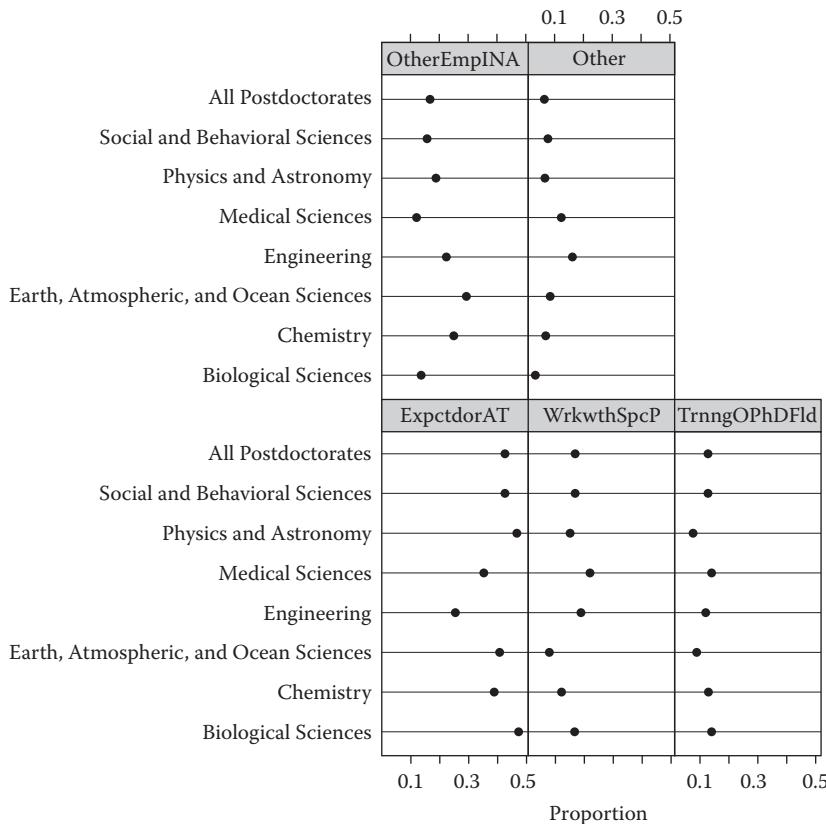
Subject	Expected or Additional Training	Work with Specific Person	Training Outside PhD Field	Other Employment not Available	Other
Biological sciences	6404	2427	1950	1779	602
Chemistry	865	308	292	551	168
Earth, Atmospheric, and Ocean Sciences	343	75	75	238	80
Engineering	586	464	288	517	401
Medical sciences	205	137	82	68	74
Physics and astronomy	1010	347	175	399	162
Social and behavioral sciences	1368	564	412	514	305
All postdoctorates	11197	4687	3403	4406	1914

**FIGURE 2.31**

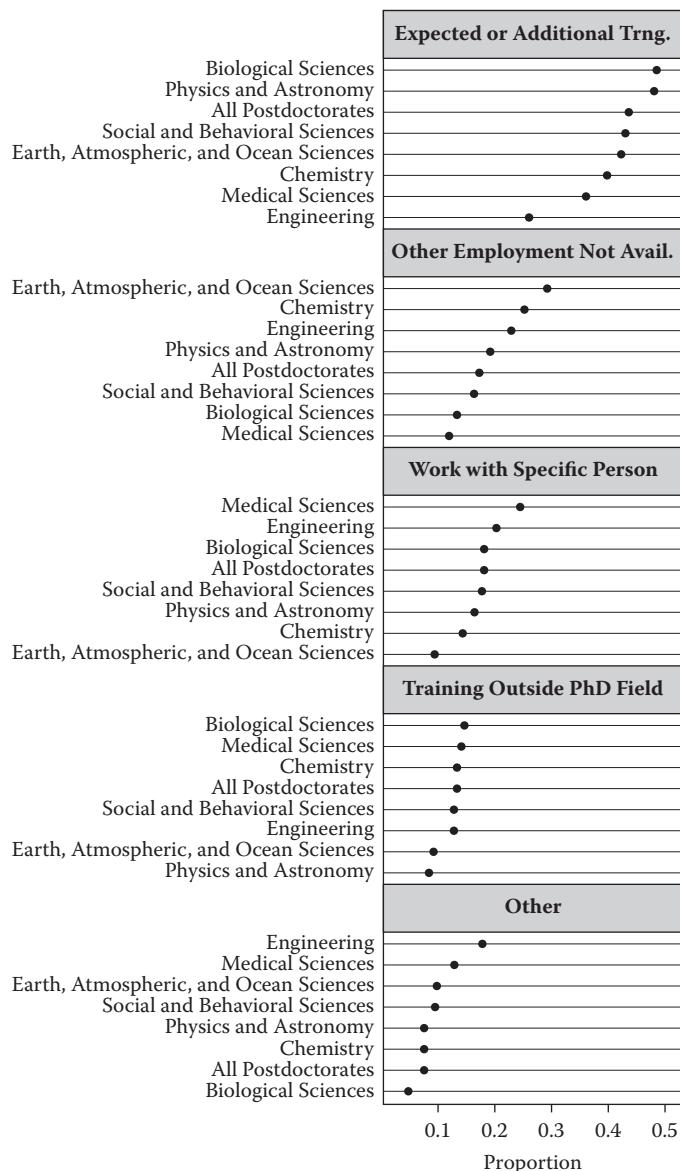
A stacked bar chart showing the proportion of reasons for choosing a postdoc by field of study.

in Figure 2.32. For comparing the proportions of reasons across areas of study, the dot plot seems preferable because it is more easily judged by eye. The multipanel dot plot becomes even more informative if the proportions are ordered from low to high within each panel, as shown in Figure 2.33. We see that the most popular reason for choosing a postdoctoral position is “Expected or Additional Training,” and that this applies to all areas of study. For “Earth, Atmospheric, and Ocean Sciences,” postdocs appear to mostly take a job because other employment is not available. Figure 2.33 provides an easy-to-use and informative display of the data.

The last example is also taken from Sarkar (2008) and is shown in Figure 2.34. The diagram gives the scatterplot matrix of violent crime rates in the 50 states of the United States in 1973, conditioned on geographical region. Each scatterplot in the diagram is enhanced by a locally weighted regression fit, an

**FIGURE 2.32**

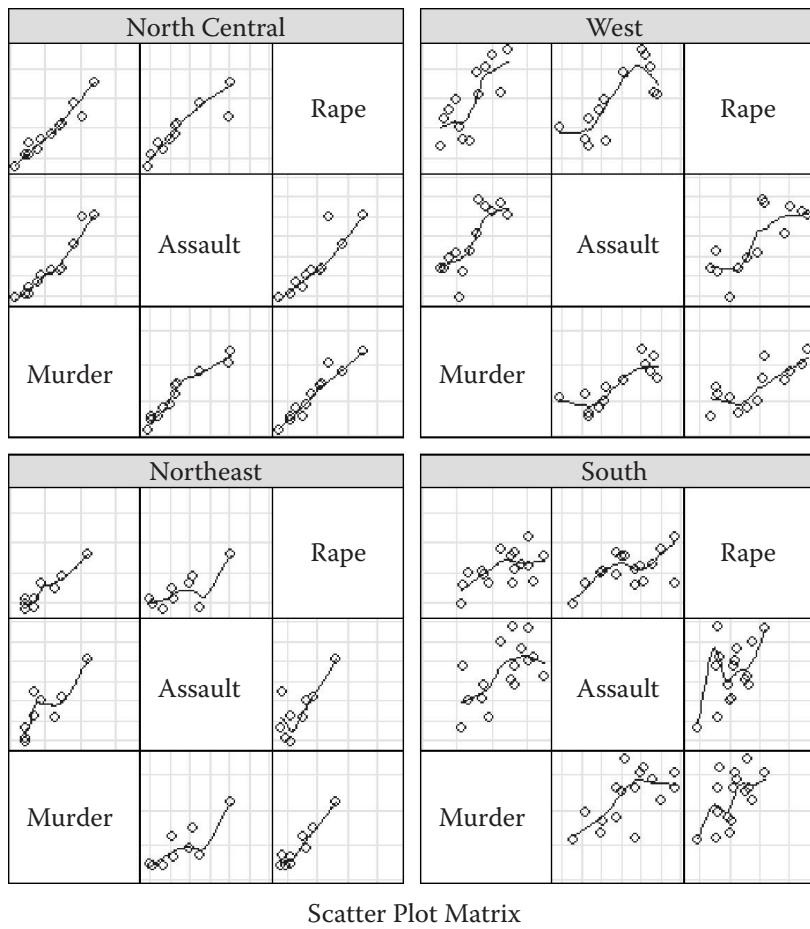
A multipanel dot plot showing the proportion of reasons for choosing a postdoc by field of study.

**FIGURE 2.33**

Reasons for choosing a postdoc position.

alternative to linear regression, to be discussed in Chapter 3. The relationship between each pair of crimes appears to be pretty similar in all four regions.

Trellis graphics is a potentially very exciting and powerful tool for the exploration of data from behavioral studies. However, a word of caution is perhaps



Scatter Plot Matrix

FIGURE 2.34

Scatterplot matrices of violent crime rates conditioned on geographical region.

in order. With small or moderately sized data sets, the number of observations in each panel may be too few to make the panel graphically acceptable. A further caveat is that trellis graphics can be seductive with the result that simpler graphics, which in many cases may be equally informative about a data set, may be ignored.

2.6 Graphical Deception

In general, graphical displays of the kind described in previous sections are extremely useful in the examination of data; indeed, they are almost

essential both in the initial phase of data exploration and in the interpretation of results from more formal statistical procedures, as will be seen in later chapters. Unfortunately, it is relatively easy to mislead the unwary with graphical material, and not all graphical displays are as honest as they should be. For example, consider the plot of the death rate per million from cancer of the breast for several periods over the last three decades, shown in Figure 2.35. The rate appears to show a rather alarming increase. However, when the data are replotted with the vertical scale beginning at zero, as shown in Figure 2.36, the increase in the breast cancer death rate is altogether less startling. This example illustrates that undue exaggeration or compression of the scales is best avoided when drawing graphs (unless, of course, you are actually in the business of deceiving your audience).

A very common distortion introduced into the graphics most popular with newspapers, television, and the media in general is when both dimensions of a two-dimensional figure or icon are varied simultaneously in response to changes in a single variable. The examples shown in Figure 2.37, both taken from Tufte (1983), illustrate this point. Tufte quantifies the distortion with what he calls the lie factor of a graphical display, which is defined as the size of the effect shown in the graph divided by the size of the effect in the data. Lie factor values close to unity show that the graphic is probably representing the underlying numbers reasonably accurately. The lie factor for the “oil barrels” is 9.4 since a 454% increase is depicted as 4280%. The lie factor for the “shrinking doctors” is 2.8.

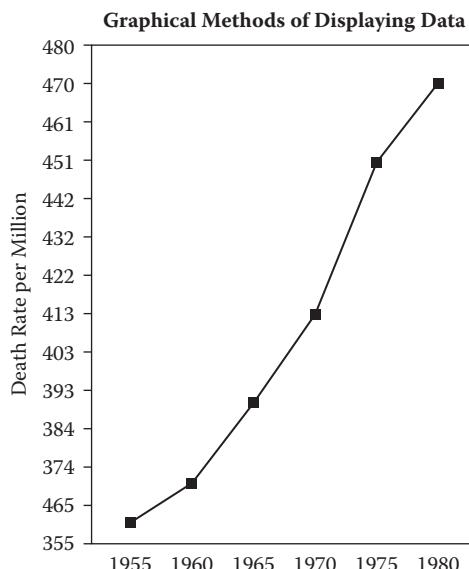
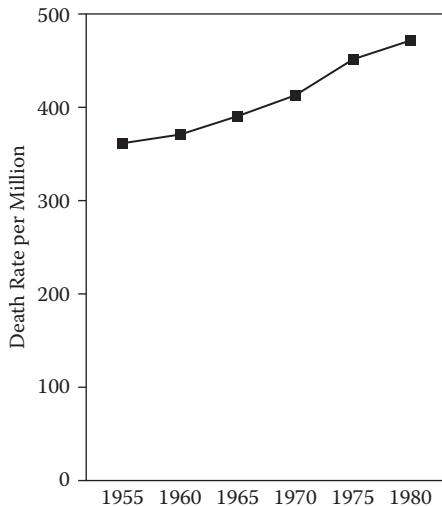
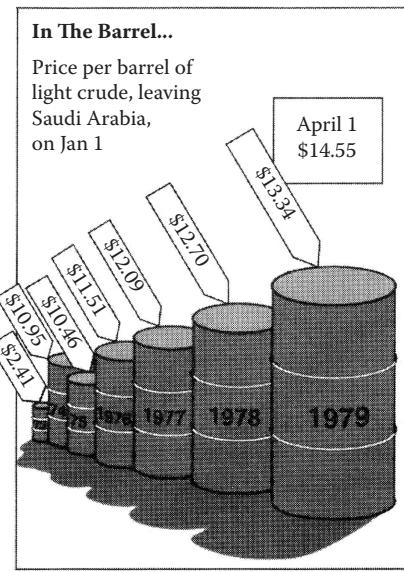


FIGURE 2.35

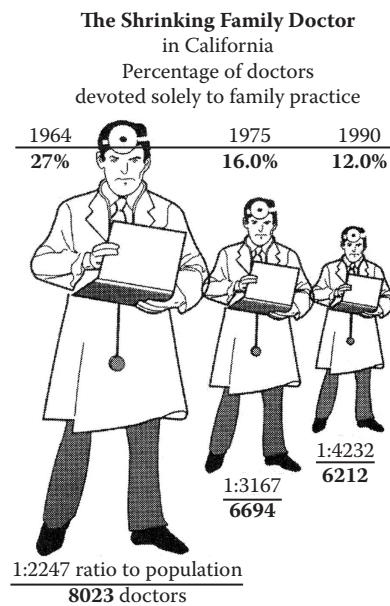
Death rates from cancer of the breast where the y-axis does not include the origin.

**FIGURE 2.36**

Death rates from cancer of the breast where the y-axis does include the origin.



(a)



(b)

FIGURE 2.37

Graphics exhibiting lie factors of (a) 9.4 and (b) 2.8.

A further example given by Cleveland (1994) and reproduced here in Figure 2.38 demonstrates that even the manner in which a simple scatterplot is drawn can lead to misperceptions about data. The example concerns the way in which judgment about the correlation of two variables made on the basis of looking at their scatterplot can be distorted by enlarging the area in which the points are plotted. The coefficient of correlation in the diagram on the right in Figure 2.38 appears greater.

Some suggestions for avoiding graphical distortion, taken from Tufte (1983), are

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
- Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
- To be truthful and revealing, data graphics must bear on the heart of quantitative thinking: “compared to what?” Graphics must not quote data out of context.
- Above all else, show the data.

Being misled by graphical displays is usually a sobering but not a life-threatening experience. However, Cleveland (1994) gives an example where using the wrong graph contributed to a major disaster in the American space program, namely, the explosion of the Challenger space shuttle

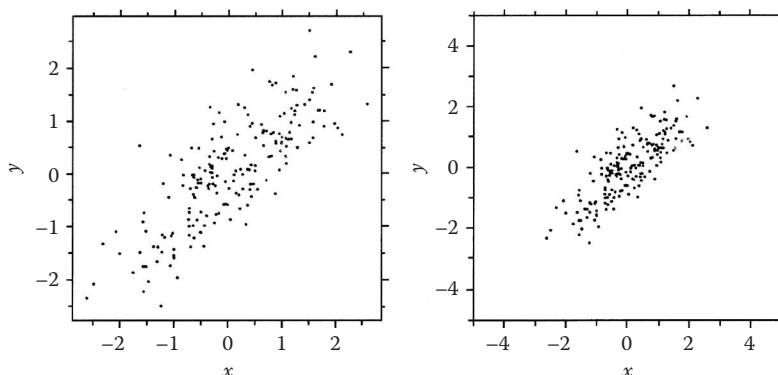


FIGURE 2.38

Misjudgment of size of correlation caused by enlarging the plot area.

and the deaths of the seven people on board. To assess the suggestion that low temperature might affect the performance of the O-rings that sealed the joints of the rocket motor, engineers studied the graph of the data shown in Figure 2.39. Each data point was from a shuttle flight in which the O-rings had experienced thermal distress. The horizontal axis shows the O-ring temperature, and the vertical scale shows the number of O-rings that had experienced thermal distress. On the basis of these data, Challenger was allowed to take off when the temperature was 31°F, with tragic consequences.

The data for “no failures” are not plotted in Figure 2.39 because the engineers involved believed that these data were irrelevant to the issue of dependence. They were mistaken, as shown by the plot in [Figure 2.40](#), which includes all the data. Here a pattern does emerge, and a dependence of failure on temperature is revealed.

To end on a less somber note and to show that misperception and miscommunication are certainly not confined to statistical graphics, see [Figure 2.41](#)!

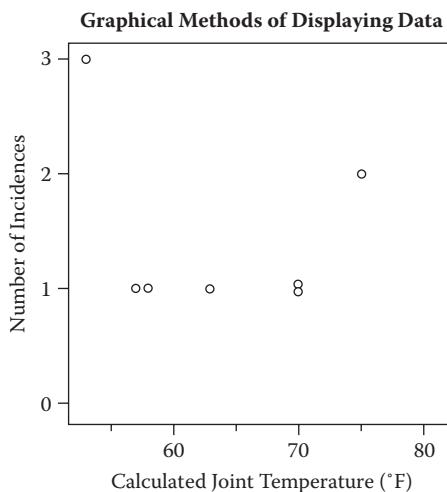
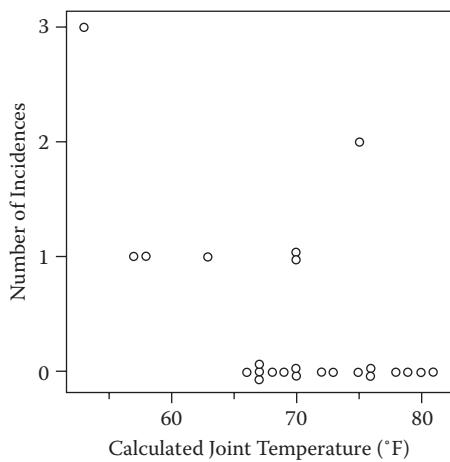


FIGURE 2.39

Data plotted by space shuttle engineers the evening before the Challenger accident to determine the dependence of O-ring failure on temperature.

**FIGURE 2.40**

A plot of the complete O-ring data.

**FIGURE 2.41**

Misperception and miscommunication are sometimes a way of life! (Drawing by Charles E. Martin, © 1961, 1969, *The New Yorker Magazine*. Used with permission.)

2.7 Summary

- Graphical displays are an essential feature in the analysis of empirical data. The prime objective is to communicate to ourselves and others.
- Graphic design must do everything it can to help people understand the subject.
- In some cases, a graphical “analysis” may be all that is required (or merited) for a data set.
- Pie charts and bar plots are rarely more informative than a numerical tabulation of the data.
- Boxplots are more useful than histograms for displaying most data sets and are very useful for comparing groups. In addition, they are useful for identifying possible outliers.
- Scatterplots are the fundamental tool for examining relationships between variables. They can be enhanced in a variety of ways to provide extra information. Scatterplots are always needed when considering numerical measures of correlation between pairs of variables.
- Scatterplot matrices are a useful first step in examining data with more than two variables.
- Trellis graphs can look very enticing and may in many, but not all, cases give greater insights into patterns in the data than simpler plots.
- Beware of graphical deception.
- Unless graphs are relatively simple, they are not likely to survive the first glance.

2.8 Exercises

2.1 According to Cleveland (1994), “The histogram is a widely used graphical method that is at least a century old. But maturity and ubiquity do not guarantee the efficiency of a tool. The histogram is a poor method.”

Do you agree with Cleveland? Give your reasons.

2.2 Shortly after metric units of length were officially introduced in Australia, each of a group of 44 students was asked to guess, to the nearest meter, the width of the lecture hall in which they were

sitting. Another group of 69 students in the same room were asked to guess the width in feet, to the nearest foot. (The true width of the hall was 13.1 m or 43.0 ft). The data are in exer_22.txt.

Construct suitable graphical displays for both sets of guesses with the aim of answering the question “which set of guesses is most accurate?”

2.3 Figure 2.42 shows the traffic deaths in a particular area before and after stricter enforcement of the speed limit by the police. Does the graph convince you that the efforts of the police have had the desired effect of reducing road traffic deaths? If not, why not?

2.4 The data set ex_24.txt contains values of seven variables for 10 states in the United States. The seven variables are

1. Population size divided by 1000
2. Average per capita income
3. Illiteracy rate (% population)
4. Life expectancy (years)
5. Homicide rate (per 1000)
6. Percentage of high school graduates
7. Average number of days per year below freezing

- Construct a scatterplot matrix of the data, labeling the points by state name.
- On each panel of the scatterplot matrix show the corresponding bivariate boxplot.

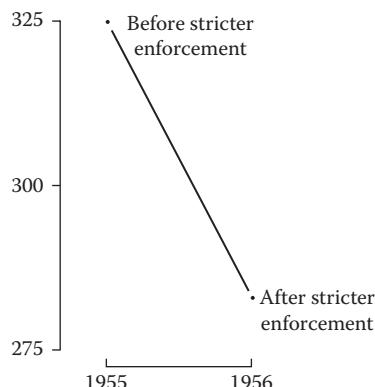


FIGURE 2.42

Traffic deaths before and after introduction of stricter enforcement of speed limit.

- Construct a coplot of life expectancy and homicide rate conditional on average per capita income.
- 2.5 Mortality rates per 100,000 from male suicides for a number of age groups and a number of countries are given in ex_25.txt. Construct side-by-side boxplots for the data from different age groups, and comment on what the graphics tell us about the data.

3

Simple Linear and Locally Weighted Regression

3.1 Introduction

Table 3.1 shows the heights (in centimeters) and the resting pulse rates (beats per minute) for 5 of a sample of 50 hospital patients (data sets with two variables are often referred to as bivariate data). Is it possible to use these data to construct a model for predicting pulse rate from height, and what type of model might be used? Such questions serve to introduce one of the most widely used statistical techniques: regression analysis. In very general terms, regression analysis involves the development and use of statistical techniques designed to reflect the way in which variation in an observed random variable changes with changing circumstances. More specifically, the aim of a regression analysis is to derive an equation relating a dependent and an explanatory variable or, more commonly, several explanatory variables. The derived equation may sometimes be used solely for prediction, but more often its primary purpose is as a way of establishing the relative importance of the explanatory variables in determining the response variable, that is, in establishing a useful model to describe the data. (Incidentally, the term regression was first introduced by Galton in the 19th century to characterize a tendency toward mediocrity, that is, more average, observed in the offspring of parents.)

In this chapter, we shall concern ourselves with regression models for a response variable that is continuous and for which there is a single explanatory variable. In Chapter 4, we will extend the model to deal with the situation in which there are several explanatory variables, and then, in Chapter 6, we shall consider suitable models for categorical response variables.

No doubt most readers will have covered simple linear regression for a response variable and a single explanatory variable in their introductory statistics course. Despite this, it may be worthwhile reading [Section 3.2](#) both as an aide-memoire and as an initial step in dealing with the more complex procedures needed when several explanatory variables are considered, a situation to be discussed in Chapter 4. It is less likely that readers will have been exposed to locally weighted regression, which will also be covered in

TABLE 3.1
Pulse Rates and Heights Data

Subject	Heights (cm)	Pulse Rates (beats/min)
1	160	68
2	167	80
3	162	84
4	175	80
5	185	80

this chapter and which can often serve as a useful antidote to the (frequently unthinking) acceptance of the simple linear model per se.

3.2 Simple Linear Regression

The technical details of the simple linear regression model are given in Technical [Section 3.1](#).

Technical [Section 3.1](#): Simple Linear Regression

Assume that y_i represents the value of what is generally known as the response variable on the i th individual and x_i represents the individual's values on what is most often called an explanatory variable; the simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_0 is the intercept, and β_1 is the slope of the linear relationship assumed between the response variable y and the explanatory variable x , and ε_i is an error term measuring the amount by which the observed value of the response differs from the value predicted by the fitted model. ("Simple" here means that the model contains only a single explanatory variable; we shall deal with the situation where there are several explanatory variables in Chapter 4.) The error terms are assumed to be statistically independent random variables having a normal distribution with mean 0 and the same variance σ^2 at every value of the explanatory variable. The parameter β_1 measures the change in the response variable produced by a change of one unit in the explanatory variable.

The regression coefficients β_0 and β_1 may be estimated as $\hat{\beta}_0$ and $\hat{\beta}_1$ using least-squares estimation in which the sum of squared differences between the observed values of the response variable y_i and the values

“predicted” by the regression equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is minimized, leading to the following estimates:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\end{aligned}$$

The predicted values of y_i from the model are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

This fitted equation could be used to predict the value of the response variable for some particular value of the explanatory variable, but it is very important to note that trying to predict values of the response variable outside the observed range of the explanatory is a potentially dangerous business.

The variability of the response variable can be partitioned into a part that is due to regression on the explanatory variable, the regression mean square (RGMS) given by RGMS = $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, and a residual mean square (RMS) given by RMS = $\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-2)$. The RMS gives an estimate of σ^2 , and the F-statistic given by $F = RGMS/RMS$ with 1 and $n-2$ degrees of freedom (DF) gives a test that the slope parameter β_1 is 0. (This is of course equivalent to testing that the correlation of the two variables is 0.)

The estimated variance of the slope parameter estimate is

$$\text{Var}(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The estimated variance of a predicted value y_{pred} at a given value of x , say x_0 , is

$$\text{Var}(y_{\text{pred}}) = s^2 \left[\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} + 1 \right]$$

where s^2 is the RMS value defined earlier. (Note that the variance of the prediction increases as x_0 gets further away from \bar{x} .)

A confidence interval for the slope parameter can be constructed in the usual way from the estimated standard error of the parameter, and the

variance of a predicted value can be used to construct confidence bounds for the fitted line.

In some applications of simple linear regression, a model without an intercept is required (when the data is such that the line must go through the origin), that is, a model of the form

$$y_i = \beta x_i + \varepsilon_i$$

In this case, application of least squares gives the following estimator for β :

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

3.2.1 Fitting the Simple Linear Regression Model to the Pulse Rates and Heights Data

Fitting the simple linear regression model to the data in Table 3.1 gives the results shown in Table 3.2. Figure 3.1 shows the fitted line on a scatterplot of the data and a 95% confidence interval for the predicted values calculated from the relevant variance term given in Technical Section 3.1 above; the diagram also contains some graphical displays giving information about the marginal distributions of each of the two variables. The results in Table 3.2 show that there is no evidence of any linear relationship between pulse rate and height. The multiple R-squared, which in this example with a single explanatory variable is simply the square of the correlation coefficient between pulse rate and height, is 0.0476, so that less than 5% of the variation in pulse rate is explained by the variation in height. Figure 3.1 shows that the fitted line is almost horizontal and that a horizontal line could easily be placed within the two dotted lines indicating the confidence interval for predicted values. Clearly, the fitted linear regression would be very poor if used to predict pulse rate from height.

TABLE 3.2
Results from Fitting a Simple Linear Regression to the Pulse and Heights Data

	Coefficients			
	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	46.9069	22.8793	2.050	0.0458
Heights	0.2098	0.1354	1.549	0.1279

Note: Residual standard error: 8.811 on 48 DF; multiple R-squared: 0.04762; F-statistic: 2.4 on 1 and 48 DF; p-value: 0.1279.

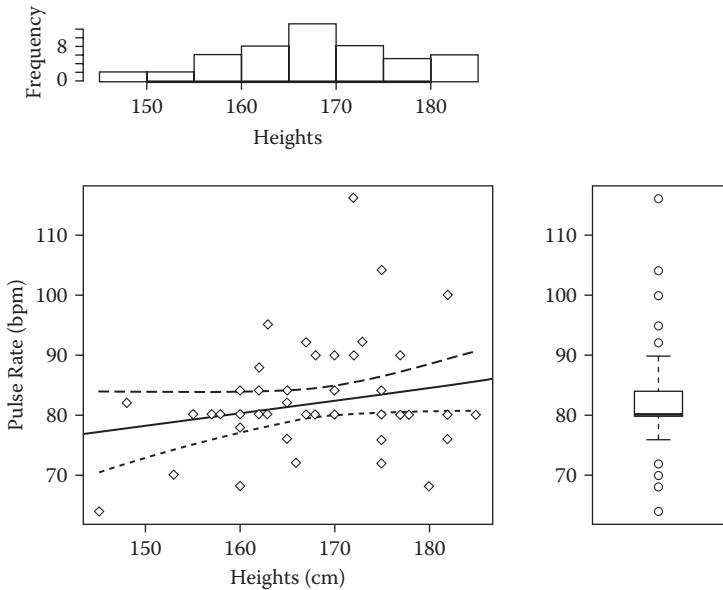


FIGURE 3.1
Scatterplot and fitted linear regression for pulse and heights data.

Figure 3.1 also shows that pulse rate has a very skewed distribution. Because of the latter, it may be of interest to repeat the plotting and fitting process after some transformation of pulse rate (see Exercise 3.1).

3.2.2 An Example from Kinesiology

For our second example of simple linear regression, we will use some data from an experiment in kinesiology, a natural care system that uses gentle muscle testing to evaluate many functions of the body in the structural, chemical, neurological, and biological realms. A subject performed a standard exercise at a gradually increasing level. Two variables were measured: (1) oxygen uptake and (2) expired ventilation, which is related to the exchange of gases in the lungs. Part of the data is shown in Table 3.3 (there are 53 subjects in the full data set), and the researcher was interested in assessing the relationship between the two variables. A scatterplot of the data along with the fitted simple linear regression is shown in Figure 3.2. The estimated regression coefficient in Table 3.4 is highly significant, but Figure 3.2 makes it very clear that the simple linear model is not appropriate for these data; we need to consider a more complicated model. The obvious choice here is to consider a model that, in addition to the linear effect of oxygen uptake, includes a quadratic term in this variable, that is, a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

TABLE 3.3

Data on Oxygen Uptake and Expired Volume
(in liters)

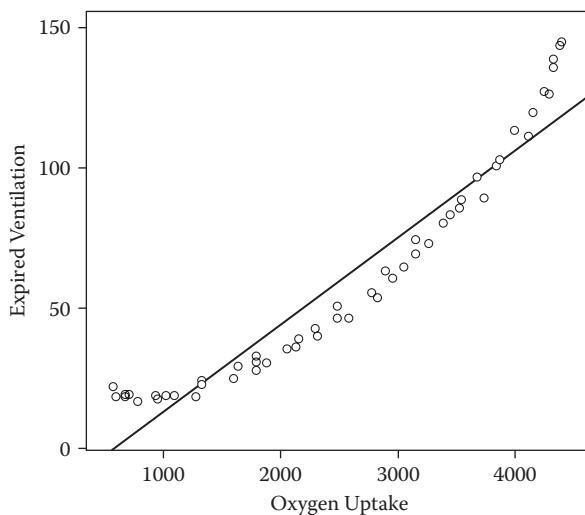
Subject	Oxygen Uptake	Expired Ventilation
1	574	21.9
2	592	18.6
3	664	18.6
4	667	19.1
5	718	19.2

TABLE 3.4

Results from Fitting a Simple Linear Regression
to the Kinesiology Data

	Coefficients			
	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	-18.448734	3.815196	-4.836	<0.001
Oxygen	0.031141	0.001355	22.987	<0.001

Note: Residual standard error: 11.96 on 51 DF; multiple R-squared: 0.912; adjusted R-squared: 0.9103; F-statistic: 528.4 on 1 and 51 DF; p-value: < 0.001.

**FIGURE 3.2**

Scatterplot of expired ventilation against oxygen uptake with fitted simple linear regression.

Such a model can easily be fitted by least squares to give estimates of its three parameters. One point to note about this model that may seem confusing is that it remains, similar to the simple model outlined in Technical Section 3.1, a linear model despite the presence of the quadratic term. The reason for this is that “linear” in linear regression models refers to the parameters rather than the explanatory variables. An example of a nonlinear model would be

$$y_i = \beta_1 x_i + \exp(\beta_2 x_i) + \varepsilon_i$$

We will not deal with such models in this book. It is worth mentioning here that including polynomial terms, for example, x and x^2 , in a linear regression model can sometimes lead to a problem known as collinearity, which will be discussed in Chapter 4. This can often be overcome by what is known as centering the explanatory variable, that is, using the original variable with its mean subtracted as the explanatory variable. Kleinbaum et al. (1988) provide an example of the effectiveness of such an approach for correcting collinearity.

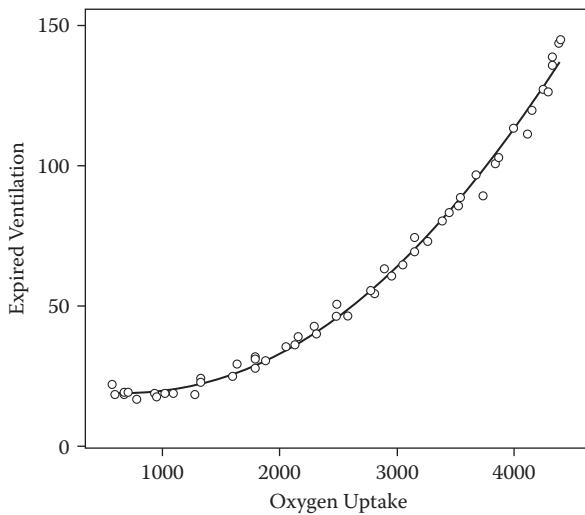
Fitting the model containing the quadratic term in oxygen uptake gives the numerical results shown in Table 3.5 and shows that the regression coefficient for the quadratic term is highly significant. The numerical results are summarized in Figure 3.3, which shows a scatterplot of the data with the addition of the fitted quadratic curve. Clearly, the new model provides an excellent fit.

Note that the numerical results in Table 3.5 are written in scientific notation where, for example, 1.5e-3 means 1.5×10^{-3} ; the reason for this is that values of oxygen squared are very large, so the corresponding estimated regression coefficient and its standard error are very small.

TABLE 3.5
Results from Fitting a Linear Regression Model to the
Kinesiology Data with Linear and Quadratic Terms
for Oxygen Uptake

	Coefficients			
	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	2.427e+01	1.940e+00	12.509	<2e-16
Oxygen	-1.344e-02	1.762e-03	-7.628	6.27e-10
Oxygen ²	8.902e-06	3.444e-07	25.850	<2e-16

Note: Residual standard error: 3.186 on 50 DF; multiple R-squared: 0.9939; F-statistic: 4055 on 2 and 50 DF; p-value: <2.2e-16.

**FIGURE 3.3**

Kinesiology data showing a fitted linear regression model that includes both oxygen uptake and oxygen uptake squared as explanatory variables.

3.3 Regression Diagnostics

Having fitted a simple regression model to our data and estimated and interpreted the regression coefficients, there still remains work to be done. We need to assess the model to see whether, for example, the assumption that the variance of the response does not change with the values of the explanatory variable, the constant variance assumption, is reasonable. Further, we need to discover if the model we have used is sensible for the data at hand. Not checking assumptions or assessing if, say, a more complex model is needed and fitting a model that is, in one way or another, unsuitable for the data, are likely to lead to incorrect inferences and conclusions. One way of investigating both the assumptions made and the possible failings of a model is an examination of residuals, that is, the difference between an observed value of the response variable y_i and the fitted value \hat{y}_i . (The residuals essentially estimate the error terms in the model.)

In regression analysis, there are various ways of plotting residual values that can be helpful in assessing particular components of the regression model. The most useful plots are as follows:

- A boxplot or probability plot of the residuals can be useful in checking for symmetry and specifically the normality of the error terms in the regression model.

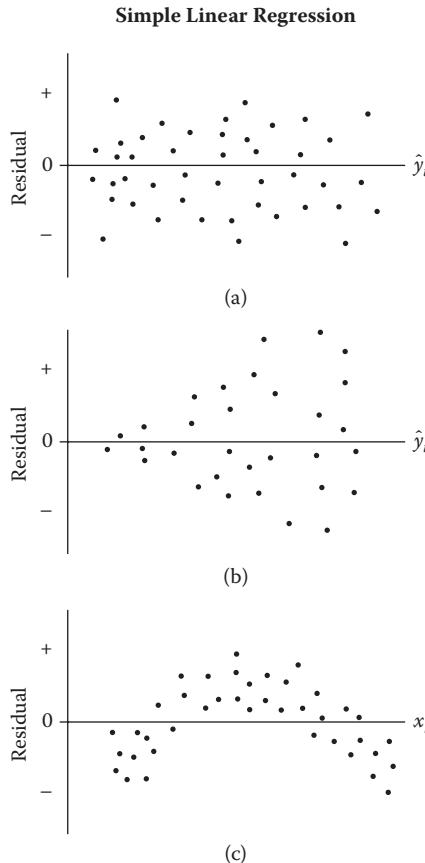
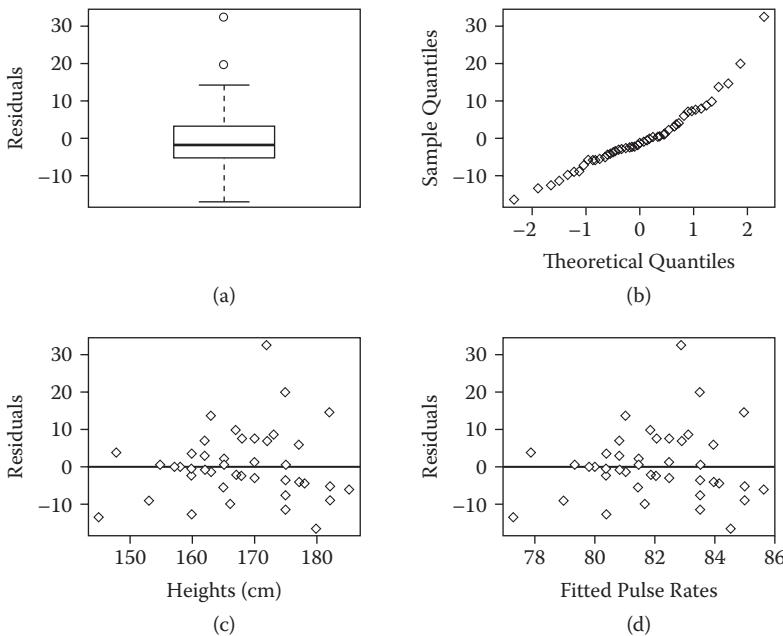


FIGURE 3.4
Idealized residual plots.

- Plotting the residuals against the corresponding values of the explanatory variable. Any sign of curvature in the plot might suggest that, say, a quadratic term in the explanatory variable should be included in the model.
- Plotting the residuals against the fitted values of the response variable (not the response values themselves for reasons spelt out in Rawlings et al., 2001). If the variability of the residuals appears to increase with the size of the fitted values, a transformation of the response variable prior to fitting is indicated.

Figure 3.4 shows some idealized residual plots that indicate particular points about models:

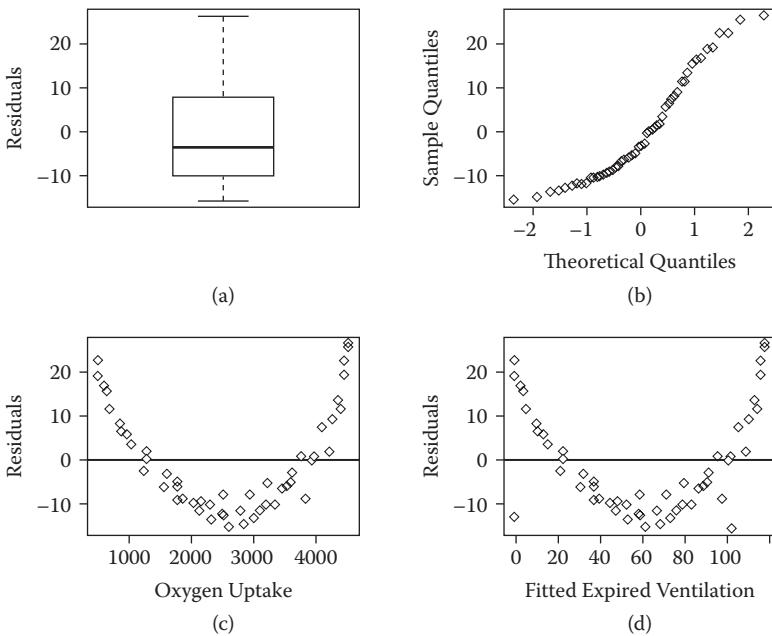
**FIGURE 3.5**

Residual plots for the pulse rates and heights data from fitting a simple linear regression model:
 (a) boxplot of residuals, (b) probability plot of residuals, (c) plot of residuals against height, and
 (d) plot of residuals against fitted pulse rates.

- **Figure 3.4a** is what is looked for to confirm that the fitted model is appropriate.
- **Figure 3.4b** suggests that the assumption of constant variance is not justified so that some transformation of the response variable before fitting might be sensible.
- **Figure 3.4c** implies that the model needs a quadratic term in the explanatory variable.

(In practice, of course, the residual plots obtained might be somewhat more difficult to interpret than these idealized plots!)

Let us now look at some residual plots for the two examples considered earlier. For the pulse rate and heights data, Figure 3.5 shows four residual plots. The boxplot indicates two very large residuals, but the probability plot shows little evidence of a departure from linearity, so there is no evidence of a departure from normality. Both Figure 3.5c and 3.5d suggest that the variance of the residuals increase both with height and the fitted values of the pulse rate; the constant variance assumption seems questionable for these data, and a transformation of the response may be helpful (again, see Exercise 3.1).

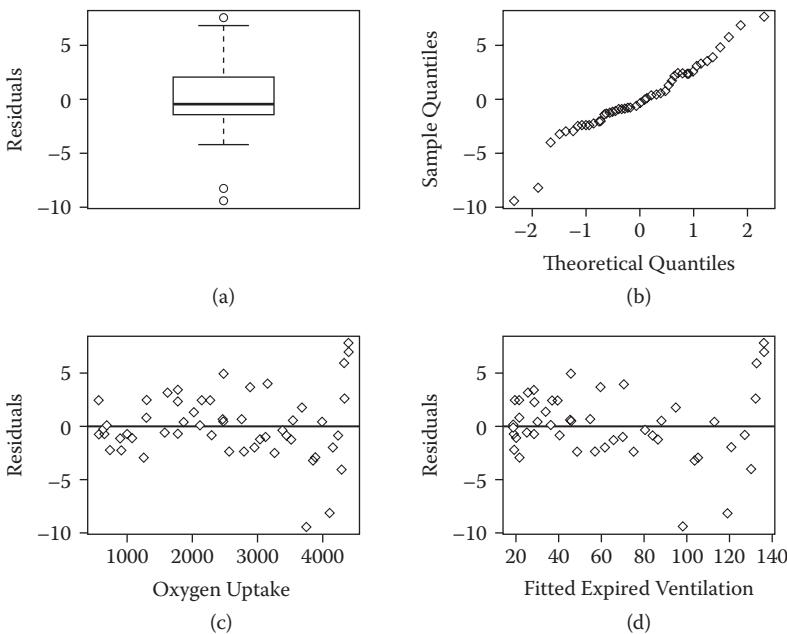
**FIGURE 3.6**

Residual plots for the oxygen uptake and expired ventilation data from fitting a simple linear regression model: (a) boxplot of residuals, (b) probability plot of residuals, (c) plot of residuals against height, and (d) plot of residuals against fitted pulse rates.

Moving on to the kinesiology data, Figure 3.6 shows the same four residual plots for a simple linear regression fitted to these data as the plots in Figure 3.7. The probability plot indicates that the residuals do not have a normal distribution, and the plots of residuals against oxygen uptake and fitted expired ventilation show very clearly that a model with a quadratic term in oxygen uptake is needed. For these data, the need for a quadratic term was clear by looking at the scatterplot of expired ventilation against oxygen uptake; but this will not always be the case, and in many cases, the residual plots will uncover problems or the need for other terms in the current model that are not apparent in the scatterplot of the data.

In Figure 3.7, the same four residual plots are given for the kinesiology data after fitting the model with both a linear and a quadratic term for oxygen uptake. We see that now the residuals are far better behaved than in Figure 3.6; clearly, this more complicated model is a far better fit than the simple linear regression model.

The “raw” residuals used here suffer from certain problems that make them less helpful in investigating fitted models than they might be with some relatively simple adjustments. These adjustments, along with a number of other diagnostic tools for regression models, will be discussed in Chapter 4.

**FIGURE 3.7**

Residual plots for the oxygen uptake and expired ventilation data from fitting a linear regression model that includes both a linear and quadratic term for oxygen uptake: (a) boxplot of residuals, (b) probability plot of residuals, (c) plot of residuals against height, and (d) plot of residuals against fitted pulse rates.

3.4 Locally Weighted Regression

When investigating the relationship between two variables, the first stop is the simple linear regression model. A scatterplot of the data that also shows the fitted line provides an excellent first graphic for studying the dependence of two variables. After looking at this graph and also viewing the residual plots, we may perhaps decide to add, say, a quadratic explanatory variable term. But, instead of assuming we know the functional form for a regression model, is there any way to allow the data themselves to suggest the appropriate functional form? The secret is to replace the global estimates from the regression models considered earlier in this chapter with local estimates in which the statistical dependency between two variables is described not with a single parameter such as a regression coefficient but with a series of local estimates. For example, a regression might be estimated between the two variables for some restricted range of values for each variable and the process repeated across the range of each variable. The series of local estimates is then aggregated by drawing a line to summarize the relationship between

the two variables. In this way, no particular functional form is imposed on the relationship. Such an approach is particularly useful when

- The relationship between the variables is expected to be of a complex form not easily fitted by standard linear or nonlinear models.
- There is no a priori reason for using a particular model.
- We would like the data themselves to suggest the appropriate functional form.

This approach is essentially an example of exploratory data analysis, in which the form of any functional relationship emerges from a set of data rather than from, say, a theoretical construct. The starting point for a local estimation approach to fitting relationships between variables is the scatterplot smoother, which is described in Section 3.4.1.

3.4.1 Scatterplot Smoothers

The local estimation procedures to be discussed here are essentially nonparametric because, unlike a parametric technique, for example, linear regression, they do not summarize the relationship between two variables with a parameter such as a regression or correlation coefficient. Instead, nonparametric smoothers summarize the relationship between two variables using a line drawing. The simplest of this collection of nonparametric smoothers is a locally weighted regression or lowess fit, first suggested by Cleveland (1979). Technical details are given below.

Technical Section 3.2: Lowess Fit

The locally weighted regression approach assumes that the explanatory variable x and the response variable y are related in the following way:

$$y_i = g(x_i) + \varepsilon_i$$

where g is a p -degree polynomial function in the predictor variable x , and ε_i are random variables with mean 0 and constant scale. Values y_i are used to estimate y_i at each x_i and are found by fitting the polynomials using weighted least squares with large weights for points near to x_i and small otherwise.

Two parameters control the shape of a lowess curve. The first is a smoothing parameter, a (often known as the span), with larger values leading to smoother curves—typical values are $1/4$ to 1. In essence, the span decides the amount of the trade-off between reduction in bias and increase in variance. If the span is too large, the nonparametric regression estimate will be biased, but if the span is too small, the estimate will be overfitted with inflated variance. Keele (2008) gives an extended discussion of the influence of the choice of span on nonparametric regression.

The second parameter, γ is the degree of the polynomials that are fitted by the method; γ can be 1 or 2. In any specific application, the change of the two

parameters must be based on a combination of judgment and trial and error. Residual plots may be helpful in judging a particular combination of values.

Our first example of using a locally weighted regression approach involves again the data on pulse rates and heights given in [Table 3.1](#). A scatterplot of the data that shows both the simple linear regression fit and the locally weighted regression fit is shown in Figure 3.8. The lowess fit shows some degree of curvature explained perhaps by the locally weighted approach being less influenced by the observations with large pulse rates. This possible curvature may be worth investigating by fitting a linear model with both a linear and a quadratic term for height (see Exercise 3.1). This demonstrates one way of using locally weighted regression fits—they may indicate a possible parametric model for the data.

An alternative smoother that can often be usefully applied to bivariate data is some form of spline function. (A spline is a term for a flexible strip of metal or rubber used by a draftsman to draw curves.) Such functions are described in the following technical section.

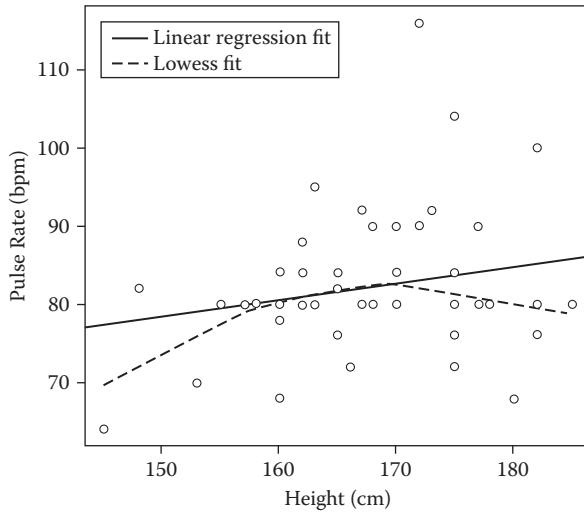


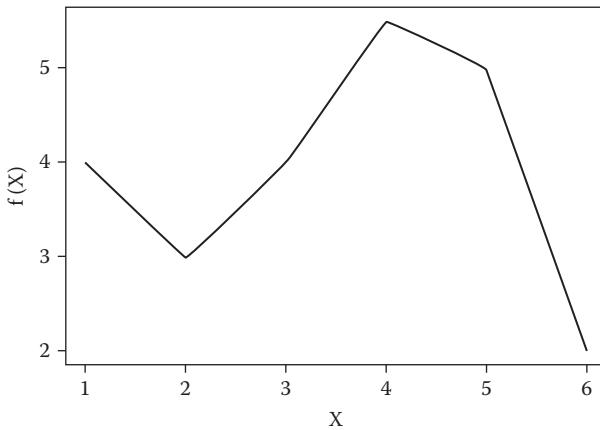
FIGURE 3.8

Scatterplot of pulse rate against height showing fitted linear and locally weighted regression fits.

Technical [Section 3.3: Spline Smoothers](#)

Spline functions are polynomials within intervals of the x variable that are connected across different values of x . [Figure 3.9](#), for example, shows a linear spline function, that is, a piecewise linear function, of the form

$$f(x) = \beta_0 + \beta_1 X + \beta_2(X - a)_+ + \beta_3(X - b)_+ + \beta_4(X - c)_+$$

**FIGURE 3.9**

A linear spline function with knots at $a = 1$, $b = 3$, and $c = 5$.

$$\begin{aligned} (u)_+ &= u & u > 0 \\ \text{where } & \\ &= 0 & u \leq 0 \end{aligned}$$

The interval endpoints a , b , and c are called knots. The number of knots can vary according to the amount of data available for fitting the function.

The linear spline is simple and can approximate some relationships, but it is not smooth and so will not fit highly curved functions well. This problem is overcome by using piecewise polynomials—in particular, cubics, which have been found to have beneficial properties with good ability to fit a variety of complex relationships. The result is a *cubic spline*.

Again, we wish to fit a smooth curve, $g(x)$, that summarizes the dependence of y on x . A natural first attempt might be to try to determine g by least squares as the curve that minimizes

$$\sum [y_i - g(x_i)]^2$$

But this would simply result in an interpolating curve and would not be smooth at all. Instead, an amended least-squares criterion can be used to determine g , namely,

$$\sum [y_i - g(x_i)]^2 + \lambda \int g''(x)^2 dx$$

where $g''(x)$ represents the second derivation of $g(x)$ with respect to x . Although when written formally this criterion looks a little formidable, it is really nothing more than an effort to govern the trade-off between

the goodness of fit of the data (as measured by $\sum[y_i - g(x_i)]^2$) and the “wiggliness” or departure of linearity of g (as measured by $\int g''(x)^2 dx$); for a linear function, this part of the fitting criterion would be zero. The parameter λ governs the smoothness of g , with larger values resulting in a smoother curve.

The function that minimizes the amended least-squares fitting criterion is known as a cubic spline and is essentially a series of cubic polynomials joined at the unique observed values of the explanatory variables x_i (for more details, see Keele, 2008).

The spline smoother does have a number of technical advantages over the lowess smoother, such as providing the best mean square error and avoiding overfitting, which can cause smoothers to display unimportant variation between x and y , which is of no real interest. But, in practice, the lowess smoother and the cubic spline smoother will give very similar results on many examples.

As an example of using a spline smoother, Figure 3.10 shows the pulse rate and height data yet again with, in this case, added linear, lowess, and spline fits. The latter two fits are very similar for these data.

As a further example of the use of scatterplot smoothers, we will use data from Jacobson and Dimmock's (1994) study of the 1992 U.S. House elections, an example also used in Keele (2008). In the 1992 House elections, many incumbents were defeated, and Jacobson and Dimmock investigated the factors that contributed to the unusually high number of incumbents who were

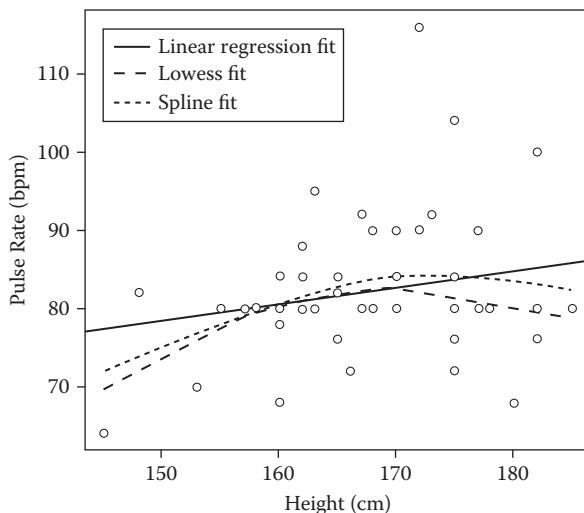


FIGURE 3.10

Scatterplot of pulse rate against height showing linear, lowess, and spline fits.

TABLE 3.6
First Five Observations of Challenger Vote
and Perot Vote

District	Challenger Vote (%)	Perot Vote (%)
1	37.92675	11.68032
2	38.24330	10.75909
3	29.76948	11.89173
4	32.75101	14.80878
5	53.76603	9.22018

beaten that year. In 1992, dissatisfaction with Congress was high because of a weak economy and a number of congressional scandals. Jacobson and Dimmock suggest that one possible indicator of such dissatisfaction was the percentage of vote for H. Ross Perot in the 1992 presidential election. The district-level vote between the president and members of the house is highly correlated, and Jacobson and Dimmock explored whether the level of support for Perot in the presidential election increased the support for the challengers in the House elections. The first five observations (out of a total of 312) of the challenger's percentage vote and the percentage vote for Perot in each congressional district in the 1992 election are shown in Table 3.6.

A scatterplot of the Jacobson and Dimmock data is shown in Figure 3.11. Jacobson and Dimmock assumed the relationship between the two variables

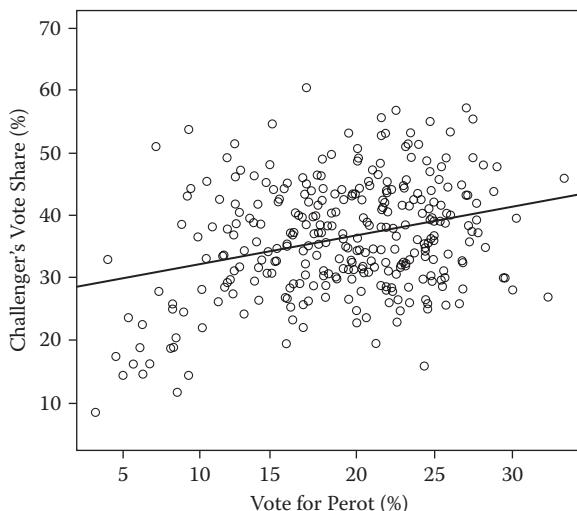
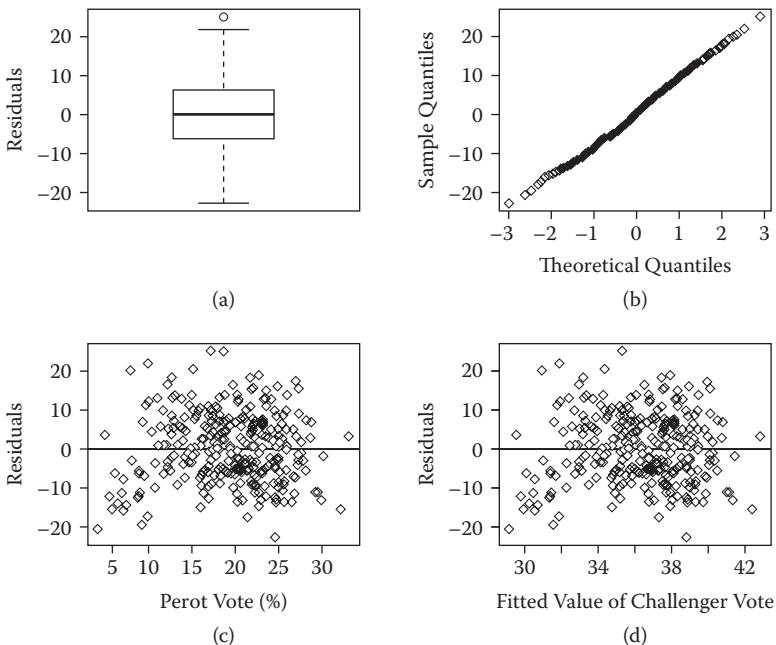
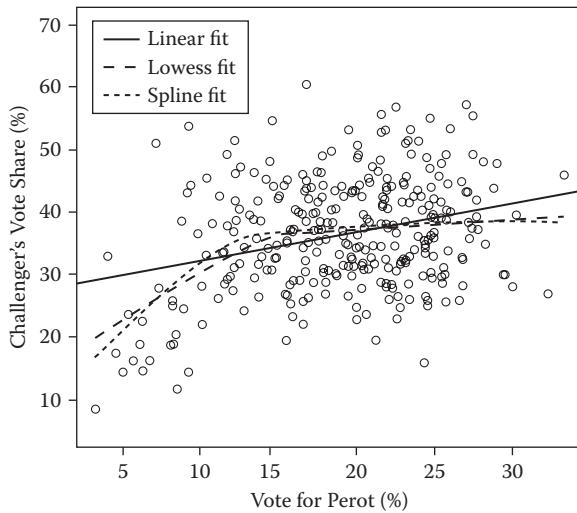


FIGURE 3.11
Scatterplot of challenger vote (%) and Perot vote (%) showing fitted simple linear regression.

**FIGURE 3.12**

Residual plots from the fitting of a simple linear regression model to the congressional vote data.

to be linear, and the fitted simple linear regression of challenger vote on Perot vote is also shown in Figure 3.11. We can investigate the fitted model by looking at the four residual plots used previously; they are shown in Figure 3.12. These plots give no cause for concern about the simple linear regression model; the probability plots give no evidence that the residuals depart from normality, and Figures 3.12c and 3.12d give no indication that higher-order terms in the explanatory variable are needed in the model. Nevertheless, we shall fit both a lowess and a spline smoother to the data, with the result shown in Figure 3.13. Both smoothers suggest some leveling off of the relationship between challenger vote and Perot vote; given the evidence from the residual plots in Figure 3.12, we might conclude that the lowess and spline fits are misleading in this case. But Keele (2008) shows this is not true by carrying out some inferential tests, which clearly demonstrate that there is significant curvature in the relationship between challenger vote and Perot vote. So, here is an example in which use of a locally weighted regression approach leads to a discovery not apparent with the use of simple linear regression and associated residual plots. It appears that increases in the challenger's vote share does not occur uniformly as support for Perot increases (which is implied by the simple linear regression model); rather, support for Perot has a diminishing effect on the challenger's vote share.

**FIGURE 3.13**

Scatterplot of challenger's vote share and Perot vote 1992, showing linear, lowess, and spline fits.

In this section we have given only a brief, relatively informal account of scatterplot smoothers that hopefully demonstrates how they might be useful in practice. We have not covered issues of inference, overfitting, etc., and for these and other details of the approach, readers are referred to the excellent book by Keele (2008).

3.5 Summary

- The scatterplot and a fit of a simple linear regression model are the first points to be considered when dealing with a set of bivariate data.
- The assumptions of the fitted model must be checked by looking at a variety of residual plots. In this way the assumptions of normality and constant variance can be assessed.
- Residual plots may also be helpful in indicating the possible need for higher-order terms in the explanatory variable to be included in the model.
- Scatterplot smoothers can be a useful additional tool in the exploration of a bivariate data set. They often provide a helpful antidote to the unthinking application of simple linear regression.

3.6 Exercises

- 3.1 Reanalyze the pulse rates and heights data after taking a log transformation of pulse rate. Contrast and compare the results with those described in the text, remembering that using a log transformation changes the scale of this variable.
- 3.2 The data in exer_32.txt gives the final examination scores (out of 75) and corresponding exam completion times (seconds) of 134 individuals. Construct a scatterplot of the data that shows the simple linear regression fit of exam score on time and also gives suitable graphics for the marginal distributions of each variable. Use residual plots to check the assumptions of the model and whether a more complex model might be needed for these data.
- 3.3 The data in exer_33.txt gives the average vocabulary size of children at various ages. Construct the scatterplot of the data and use the scatterplot and knowledge of the data to fit a suitable model.
- 3.4 The data in exer_34.txt gives marriage and divorce rates (per 1000 population per year) for 14 countries. Derive the linear regression equation of divorce rate on marriage rate and show the fitted line on a scatterplot of the data. On the basis of the regression line, predict the divorce rate for a country with a marriage rate of 8 per 1000 and also for a country with a marriage rate of 14 per 1000. How much conviction do you have in each prediction?
- 3.5 The data in exer_35.txt gives the average percentage memory retention measured against passing time (minutes). The measurements were taken five times during the first hour after subjects memorized a list of disconnected items, and then at various times up to a week later. Plot the data (after a suitable transformation if necessary) and investigate the relationship between retention and time using a suitable regression model.

4

Multiple Linear Regression

4.1 Introduction

Multiple linear regression represents a generalization, to more than a single explanatory variable, of the simple linear regression procedure described in Chapter 3. It is now that the relationship between a response variable and several explanatory variables becomes interesting. The adjective “multiple” indicates that at least two explanatory variables are involved in the modeling exercise. At the onset, it is important to note that the explanatory variables are strictly assumed to be fixed and under the control of the investigator, that is, they are not considered to be random variables; only the response variable is considered to be a random variable. In practice, of course, this assumption is unlikely to be true, in which case the results from a multiple linear regression are interpreted as being conditional on the observed values of the explanatory variables, and the inherent variation in the explanatory variables is ignored. Because there are no distributional assumptions about the explanatory variables, they may be nominal, categorical with more than two categories (such variables need to be coded in an appropriate way—see Exercise 4.2, and Chapters 5 and 6), ordered categorical, or interval. The goals of a multiple regression may be to determine whether the response variable and one or more explanatory variables are associated in some systematic way or to predict values of the response variables from values of the explanatory variables, or both.

Details of the model, including the estimation of its parameters by least squares and the calculation of standard errors, are given in Technical Section 4.1.

Technical Section 4.1: Multiple Linear Regression

The multiple linear regression model for a response y with observed values y_1, y_2, \dots, y_n and q explanatory variables x_1, x_2, \dots, x_q with observed values $x_{i1}, x_{i2}, \dots, x_{iq}$ for $i = 1, 2, \dots, n$ is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i$$

The regression coefficient β_i measures the change in the mean response associated with a unit change in the corresponding explanatory variable, provided the values of all the other explanatory variables do not change. This is often referred to as partialling out or controlling for other variables, although such terms are probably best avoided. The “linear” in multiple linear model refers to the parameters rather than to the explanatory variables, as discussed in Chapter 3.

The error terms in the model $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are assumed to have a normal distribution with zero mean and the same variance σ^2 for all values of the explanatory variables. This assumption implies that, for given values of the explanatory variables, the response variable is normally distributed with a mean that is a linear function of the explanatory variables and a variance that is not dependent on them.

The least-squares estimation process is used to estimate the parameters in the multiple linear regression model, and the resulting estimators are most conveniently described with the use of a matrix and vector notation. So we introduce a vector $\mathbf{y}' = [y_1, y_2, \dots, y_n]$ and an $n \times (q + 1)$ matrix \mathbf{X} given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

Now we can write the multiple linear regression model for all n observations as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]$ and $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \dots, \beta_q]$. The least-squares estimators of the parameters in the multiple linear regression model are given by the set of equations

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

These matrix manipulations are easily performed on a computer, but you must ensure that there are no linear relationships between the explanatory variables, such as one variable is the sum of several others; otherwise, your regression software will complain because the inverse of the matrix $\mathbf{X}'\mathbf{X}$ will be singular. (More details of the model in matrix form and the

least-squares estimation process are given in Rawlings et al., 2001.) The estimated regression coefficients have the same interpretation as given earlier for the population values of these parameters. Of course, each estimated coefficient and its interpretation are only applicable within the range of values of the corresponding explanatory variable that has been used in fitting the multiple linear regression model.

The variation in the response variable can be partitioned into a part due to regression on the explanatory variables and a residual as in the case of simple linear regression. These can be arranged in an analysis of variance table as follows:

Source	Df	Sum of Squares	Mean Square	F-statistic
Regression	q	RGSS	$\text{RGMS} = \text{RGSS}/q$	RGMS/RSMS
Residual	$n - q - 1$	RSS	$\text{RSMS} = \text{RSS}/(n - q - 1)$	

The F-statistic gives a test of the omnibus null hypothesis that all the regression coefficients are zero, that is, none of the explanatory variables are associated with the response variable; in most practical situations, this is a relatively uninteresting hypothesis. The residual mean square s^2 is an estimator of σ^2 , and the estimator of the covariance matrix (see Chapter 9) of the parameters is

$$\mathbf{S}_{\hat{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

The diagonal elements of this matrix give estimates of the variances of the estimated regression coefficients, and the off-diagonal elements give estimates of the estimated covariances. The estimated variances are used to assess the statistical significance of the regression coefficients and to construct confidence intervals for them.

A measure of the fit of the model is provided by the multiple correlation coefficient R , which is defined as the correlation between the observed values of the response variable y_i and the values predicted by the fitted model \hat{y}_i , which are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_q x_{iq}$$

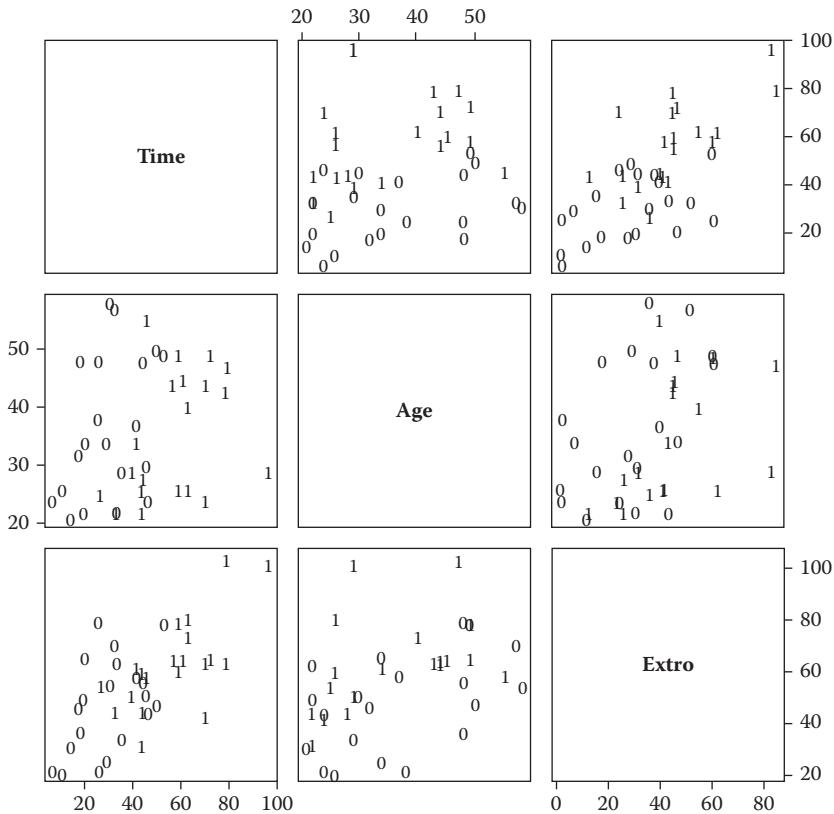
The value of R^2 gives the proportion of variability in the response variable accounted for by the explanatory variables.

4.2 An Example of Multiple Linear Regression

As our first example of fitting a multiple linear regression model, we will return to the data introduced in Chapter 2 (see Table 2.5) concerned with how long each week people spend looking after their cars. The interest lies in investigating which of the three explanatory variables—age, a measure of extroversion, and gender—are most important in determining the amount of time people spend looking after their cars. Note that, in this example, one of the explanatory variables, gender, is a binary variable, but as explained earlier, since no distributional assumptions are made about explanatory variables, such a variable causes no problems when applying multiple linear regression. (Categorical explanatory variables with more than two categories also cause no problems, but they have to be coded appropriately as we shall see in later examples.)

Let us begin with constructing a scatterplot matrix for the three variables—time, age, and extroversion—labeling on each panel of the plot the gender of a subject. The resulting diagram is shown in [Figure 4.1](#). It is clear from this diagram that men spend more time looking after their cars and that time is strongly related to extroversion; perhaps it is less strongly related with age. In addition, the diagram suggests that age and extroversion scores are related, and that there may be two potentially troublesome outliers.

The results from fitting the multiple linear regression model are shown in [Table 4.1](#). The omnibus F-test for testing the hypothesis that all three regression coefficients are zero has a very low associated p-value; there is strong evidence that not all three coefficients are zero. The square of the multiple correlation coefficient is 0.6377; the three explanatory variables together account for about 64% of the variation in time spent looking after the car. The size of the “raw” regression coefficients in [Table 4.1](#) should not be used to judge the relative importance of the explanatory variables in predicting the response variable, although what are known as standardized values of these coefficients can, partially at least, be used in this way. The standardized values might be obtained by applying the regression model to the values of the response variable and explanatory variables, standardized by (divided by) their respective standard deviations. In such an analysis, each regression coefficient represents the change in the standardized response variable associated with a change of one standard deviation unit in the explanatory variable, again conditional on the other explanatory variables remaining constant. The standardized regression coefficients can, however, be found without undertaking this further analysis, by simply multiplying the raw regression coefficient by the standard deviation of the appropriate explanatory variable and dividing by the standard deviation of the response variable. For the time spent looking after car data, the relevant standard deviations are time (20.79), age (11.39), extroversion (19.67), and gender (0.51); so, the required standardized regression coefficients are

**FIGURE 4.1**

Scatterplot matrix for the time spent looking after car data, with gender labeled 1 = male, 0 = female.

$$\text{Age: } 20.07 \times 0.51 / 20.79 = 0.49$$

$$\text{Extroversion: } 0.16 \times 11.39 / 20.79 = 0.09$$

$$\text{Gender: } 0.46 \times 19.67 / 20.79 = 0.44$$

Now, it looks like extroversion and gender are more important than age in predicting the time spent looking after the car. For binary explanatory variables such as gender in this example, the unstandardized regression coefficients are more directly interpretable than the standardized versions. This is because the unstandardized coefficients for such explanatory variables simply estimate the difference in the average value of the response between the two categories defined by the variable, holding the other explanatory variables in the model constant.

The t-values associated with each explanatory variable are obtained by simply dividing the estimated regression coefficient by the standard error of the

estimate, and it might be thought that the associated significance levels would indicate the importance of the explanatory variables. Here, the values of these associated p-values appear to imply that gender and extroversion are strongly associated with time spent looking after car, whereas age seems to not be associated with the response variable. But, this rather simplistic interpretation of the t-statistics is not always appropriate as we shall make clear later in the chapter.

The estimated regression coefficients give the changes in the value of the response variable when the corresponding explanatory variable changes by one unit; for a binary variable such as gender in this example, this statement means a change from one category to the other, so the regression coefficient gives an estimated difference between the two categories conditional on the other variables. Here, the estimated difference in time spent looking after the car between men and women, conditional on age and extroversion staying constant, is 20 min longer for men than for women, with a 95% confidence interval of $20 - 2.04 \times 4.65, 20 + 2.04 \times 4.65$, that is, [10.5,29.5] (2.04 is the value of a t-statistic with 36 degrees of freedom [DF] for a 0.05 significance level).

If here we accept, for the moment, the results from the t-statistics at face value, then we might conclude that a model that includes only the explanatory variables, extroversion score and gender, will be adequate for these data, thus providing a more parsimonious model for the data. If extroversion score and gender were both independent of age, then their regression coefficients in the new model would be the same as they are in Table 4.1. But because age is certainly related to extroversion (see [Figure 4.1](#)), the model with only extroversion and gender as explanatory variables needs to be fitted anew to get the correct regression coefficients for the gender and extroversion score explanatory variables. The results of fitting this simpler model are shown in [Table 4.2](#). The regression coefficients for gender and extroversion have changed a little from those given in Table 4.1, but the t-statistics for both variables remain highly significant. The square of the multiple correlation is now 0.63, implying that the two explanatory variables in this model account for 63% of the variation in time spent looking after the car, only a very small reduction from the model with three explanatory variables.

TABLE 4.1

Results from Fitting Multiple Linear Regression Model to Time Spent Looking After the Car Data, with Age, Extroversion Score, and Gender as Explanatory Variables

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	11.3063	7.3153	1.546	0.130956
Gender	20.0711	4.6514	4.315	0.000119
Age	0.1556	0.2062	0.754	0.455469
Extroversion	0.4643	0.1303	3.564	0.001053

Note: Residual standard error: 13.02 on 36 degrees of freedom (DF); multiple R-squared: 0.6377; F-statistic: 21.13 on 3 and 36 DF; p-value: 4.569e-08.

TABLE 4.2

Results from Fitting Multiple Linear Regression Model to Time Spent Looking After the Car Data, with Extroversion Score and Gender as Explanatory Variables

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	15.6797	4.4365	3.534	0.001118
Gender	19.1801	4.4727	4.288	0.000124
Extroversion	0.5093	0.1151	4.423	8.24e-05

Note: Residual standard error: 12.95 on 37 DF; multiple R-squared: 0.632; F-statistic: 31.77 on 2 and 37 DF; p-value: 9.284e-09.

The fitted model with gender and extroversion as explanatory variables is

$$\text{time} = 15.68 + 19.18 \times \text{gender} + 0.51 \times \text{extroversion}$$

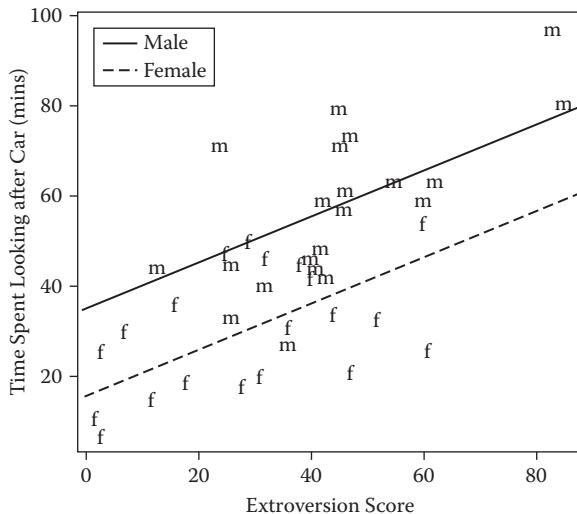
So, for men (gender = 1) this becomes

$$\text{time} = 15.68 + 19.18 + 0.51 \times \text{extroversion}$$

and for women (gender = 0) the model is

$$\text{time} = 15.68 + 0.51 \times \text{extroversion}$$

The fitted model is seen to be equivalent to two simple regression fits, each with the same slope but with a different intercept for men and women. The model is conveniently summarized in Figure 4.2.

**FIGURE 4.2**

Plot illustrating the multiple linear regression model fitted to time spent looking after the car, with extroversion and gender as the explanatory variables.

TABLE 4.3

Results from Fitting a Multiple Linear Model to the Time Spent Looking after the Car Data, with Explanatory Variables, Extroversion, Gender, and Extroversion \times Gender Interaction

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	20.0182	5.4560	3.669	0.000782
Gender	7.8178	9.5705	0.817	0.419379
Extroversion	0.3607	0.1590	2.268	0.029430
Gender: Extroversion	0.3052	0.2279	1.339	0.188970

Note: Residual standard error: 12.81 on 36 DF; multiple R-squared: 0.6495; F-statistic: 22.23 on 3 and 36 DF; p-value: 2.548e-08.

Another model we might consider is one in which an extroversion \times gender interaction is allowed. The results from fitting such a model are shown in Table 4.3. The fitted model is now

$$\text{time} = 20.02 + 7.82 \times \text{gender} + 0.36 \times \text{extroversion} + 0.31 \times (\text{gender} \times \text{extroversion})$$

So, for males (gender = 1) this becomes

$$\text{time} = 20.02 + 7.82 + 0.36 \times \text{extroversion} + 0.31 \times \text{extroversion}$$

And for females (gender = 0)

$$\text{time} = 20.02 + 0.36 \times \text{extroversion}$$

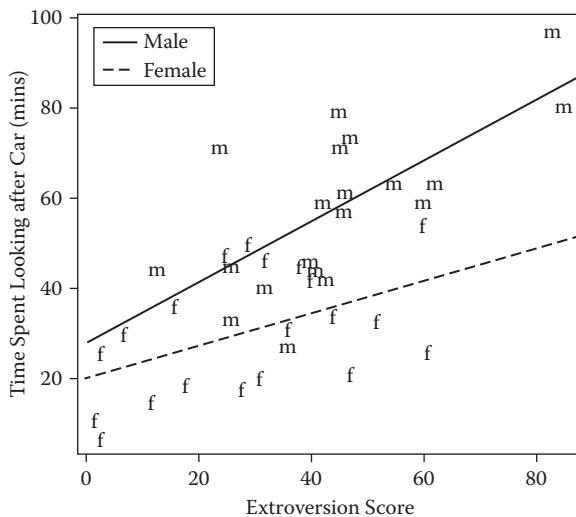
**FIGURE 4.3**

Diagram illustrating the results given in Table 4.3.

In this case, the model allows the fitted simple linear regression fits for men and women to have both different slopes and different intercepts. [Figure 4.3](#) illustrates the results of fitting this model. Of course, [Table 4.3](#) shows that the interaction term is not significant and so is not needed for these data; the simple model with parallel fits is to be preferred. However, the more complex model is illustrated simply as a useful teaching aid.

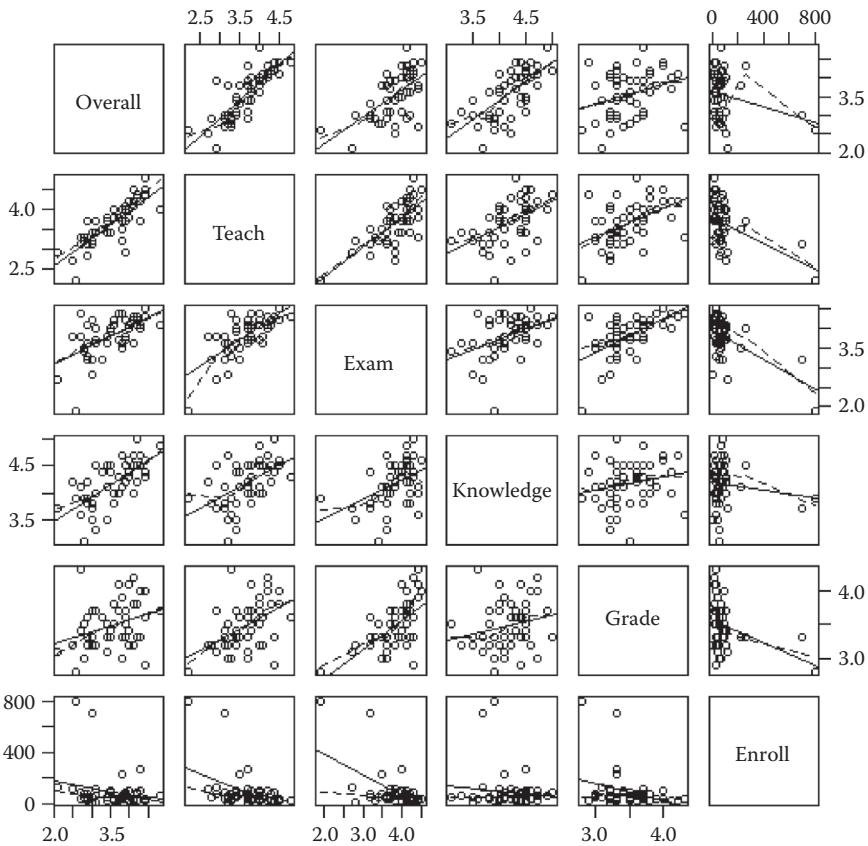
4.3 Choosing the Most Parsimonious Model When Applying Multiple Linear Regression

Now we introduce some data taken from Howell (2002), which arise from an evaluation of several hundred courses taught at a large university during the preceding semester. Students in each course had completed a questionnaire in which they rated a number of different aspects of the course on a five-point scale (1 = failure, very bad; ...; 5 = excellent, exceptional). The data we will use are the mean scores on six variables for a random sample of 50 courses; the scores for the first five chosen courses are shown in [Table 4.4](#). The six variables are (1) overall quality of lectures (overall), (2) teaching skills of the instructor (teach), (3) quality of the tests and exams (exam), (4) instructor's perceived knowledge of the subject matter (knowledge), (5) the student's expected grade in the course (grade, where higher means better), and (6) the enrollment of the course (enroll). Interest lies in how variables 2 to 5 associate with or predict the overall variable.

Before we begin the model-fitting exercise, we should examine the data graphically, and [Figure 4.4](#) shows a scatterplot matrix of the six variables, each individual scatterplot being enhanced with both a linear and a lowess fit. The plot indicates that the overall rating is related to teach, exam, knowledge, and grade, and that these explanatory variables are also related to each other. For all the scatterplots in [Figure 4.4](#), the fitted linear and lowess regressions are very similar, suggesting that for none of these explanatory variables is it necessary to consider quadratic or higher-order terms in any model. The enroll variable is problematic because of the presence of two very obvious outliers, one of which is course number 3 with an enroll value of 800, and the other is course 45 with an enroll value of 700. For the moment, we will not remove these two observations and will consider

TABLE 4.4
Course Evaluation Data

Course	Overall	Teach	Exam	Knowledge	Grade	Enroll
1	3.4	3.8	3.8	4.5	3.5	21
2	2.9	2.8	3.2	3.8	3.2	50
3	2.6	2.2	1.9	3.9	2.8	800
4	3.8	3.5	3.5	4.1	3.3	221
5	3.0	3.2	2.8	3.5	3.2	7

**FIGURE 4.4**

Scatterplot matrix of course evaluation data showing both linear regression and lowess fits for each pair of variables.

another problem that can occur when using multiple linear regression in practice, which we have not considered up to this point, namely, *multicollinearity*. The term is used to describe situations in which there are moderate to high correlations among some or all of the explanatory variables. Multicollinearity gives rise to a number of difficulties when multiple regression is applied:

- It severely limits the size of the multiple correlation coefficient because the explanatory variables largely attempt to explain much of the same variability in the response variable (see Dizney and Gromen, 1967, for an example).
- It makes determining the importance of a given explanatory variable difficult because the effects of explanatory variables are confounded due to their intercorrelations.

- It increases the variances of the regression coefficients, making the use of the fitted model for prediction less stable. The parameter estimates become unreliable.

Spotting multicollinearity among a set of explanatory variables may not be easy. The obvious course of action is to simply examine the correlations between these variables, but while this is often helpful, it is by no means foolproof; more subtle forms of multicollinearity may be missed. An alternative and generally far more useful approach is to examine what are known as the variance inflation factors of the explanatory variables. The variance inflation factor VIF_j for the j th variable is given by

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j is the multiple correlation coefficient from the regression of the j th explanatory variable on the remaining explanatory variables. The variance inflation factor of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables. A rough rule of thumb is that variance inflation factors greater than 10 give some cause for concern. For the course evaluation data, the required variance inflation factors are as follows:

Teach: 2.38

Exam: 3.12

Knowledge: 1.49

Grade: 1.61

Enroll: 1.54

It appears that multicollinearity is not a problem for the course evaluation data. In situations where multicollinearity may be a problem, what should be done? One possibility is to combine in some way explanatory variables that are highly correlated; an alternative approach is simply to select one of the set of correlated variables. Two more complex possibilities are regression on principal components and ridge regression, both of which are described in Chatterjee, Hadji, and Price (1999).

But here, we can now go ahead and fit the multiple linear regression model to give the results shown in [Table 4.5](#). Together, the five explanatory variables account for about 76% of the variation in the overall rating. The omnibus F-test has a very low associated p-value, so there is very strong evidence that not all the five regression coefficients are zero. From the t-statistics, it is probably a good bet that the two most important variables for predicting the overall rating are the teaching skills of the instructor and the instructor's

TABLE 4.5

Results of Fitting a Multiple Linear Regression Model to the Course Evaluation Data

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	-1.1951810	0.6311922	-1.894	0.064875
Teach	0.7632345	0.1329150	5.742	8.06e-07
Exam	0.1320347	0.1627995	0.811	0.421716
Knowledge	0.4889675	0.1365333	3.581	0.000849
Grade	-0.1842549	0.1654897	-1.113	0.271586
Enroll	0.0005259	0.0003901	1.348	0.184555

Note: Residual standard error: 0.3202 on 44 DF; multiple R-squared: 0.7555; F-statistic: 27.19 on 5 and 44 DF; p-value: 1.977e-12.

perceived knowledge of the subject matter. But as mentioned earlier, the use of t-statistics in this simplistic way is not really appropriate, the reason being that if say we were to drop exam from the model because its associated t-test has the highest p-value, we would need to refit the model with the remaining four explanatory variables before making any further statements about their importance because the estimated regression coefficients will now change. Of course, if the explanatory variables happened to be independent of one another, there would be no problem and the t-statistics could be used in selecting the most important explanatory variables. This is, however, of little consequence in most practical applications of multiple linear regression.

Before moving on, we should ponder the question of how the results in Table 4.5 are affected by removing the two outlier courses—course 3 and course 5—from the data and refitting the model. The answer is “not very much” as readers can verify by carrying out the task themselves.

So, if using the simple t-statistics identifying a more parsimonious model, that is, one with fewer explanatory variables but still providing an adequate fit, might be suspect in many practical applications of multiple linear regression, what are the alternatives? One approach is *all subsets regression* in which all possible models, or perhaps a subset of possible models, are compared using some suitable numerical criterion; when there are q explanatory variables, there are a total of $2^q - 1$ models (each explanatory variable can be in or out of a model, and the model in which they are all out is excluded). The course evaluation data has five explanatory variables, and so there are 31 possible models to consider. With larger numbers of explanatory variables, the number of models to consider rapidly becomes large; for example, for $q = 12$ there are 4095 models to consider. Special search algorithms are used to make this method feasible. We shall not consider this method any further.

Software packages frequently offer automatic methods of selecting variables for a final regression model from a list of candidate variables. There are three typical approaches:

- Forward selection
- Backward elimination
- Stepwise regression

The forward selection approach begins with an initial model that contains only a constant term, and successively adds explanatory variables to the model until the pool of candidate variables remaining contains no variables that, if added to the current model, would contribute information that is statistically important concerning the mean value of the response. The backward elimination method begins with an initial model that contains all the explanatory variables under investigation and successively removes variables until no variables among those remaining in the model can be eliminated without adversely affecting the predicted value of the mean response in a statistical sense. Various criteria have been suggested for assessing whether a variable should be added to an existing model in forward selection or removed in backward elimination—for example, the change in the residual sum of squares that results from the inclusion or exclusion of a variable.

The stepwise regression method of variable selection combines elements of both forward selection and backward elimination. The initial model of stepwise regression is one that contains only a constant term. Subsequent cycles of the approach involve first the possible addition of an explanatory variable to the current model, followed by the possible elimination of one of the variables included earlier if the presence of new variables has made its contribution to the model no longer important.

In the best of all possible worlds, the final model selected by applying each of the three procedures outlined here would be the same. Often this does happen, but it is in no way guaranteed. Certainly, none of the automatic procedures for selecting subsets of variables are foolproof. For example, if two explanatory variables are highly correlated with each other, it is highly unlikely that any of the usual automatic methods of model selection will produce a final model that includes both variables. In one way, this is good because it avoids the problem of collinearity discussed earlier. But the final model that automatic selection produces hides the fact that another line of modeling exists based on the second of the two highly correlated variables, and the end results of pursuing that direction might be equally satisfactory, statistically or scientifically—it may even be better (Matthews, 2005). Automatic model selection methods must be used with care, and the researcher using them should approach the final model selected with a healthy degree of skepticism. Agresti (1996) nicely summarizes the problems:

Computerized variable selection procedures should be used with caution. When one considers a large number of terms for potential inclusion in a model, one or two of them that are not really important may look impressive simply due to chance. For instance, when all the true effects are weak, the largest sample effect may substantially overestimate its

true effect. In addition, it often makes sense to include variables of special interest in a model and report their estimated effects even if they are not statistically significant at some level.

(See McKay and Campbell, 1982a, 1982b, for some more thoughts on automatic selection methods in regression.)

With all these caveats in mind, we will illustrate how the backward elimination approach works on the course evaluation data using what is known as Akaike's information criterion (AIC) to decide whether a variable can be removed from the current candidate model. The AIC index takes into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this degree of fit, by imposing a penalty for increasing the number of parameters. In a series of competing models, "lower" values of the AIC are preferable; in what follows, the judgment necessary will be made informally.

The AIC for a model is defined explicitly as minus twice the maximized log-likelihood of the model plus twice the number of parameters in the model; as the log-likelihood of the model gets larger, the AIC goes down, and as the number of parameters of the model increases, so does the value of the AIC.

The results of the backward elimination approach using the AIC are as follows:

Start: AIC = -108.28

Explanatory variables in the model are teach, exam, knowledge, grade, and enroll.

Step 1: Removing one explanatory variable at a time and leaving the other four in the model

Remove exam: Model AIC = -109.54

Remove grade: Model AIC = -108.89

Remove enroll: Model AIC = -108.26

Remove knowledge: Model AIC = -97.49

Remove teach: Model AIC = -82.32

Removing exam leads to a model containing the other four explanatory variables, and this model has a lower AIC than the original five-explanatory-variable model. Consequently, we drop the exam variable and start afresh with the model containing the variables grade, enroll, knowledge, and teach.

Current: AIC = -109.54

Explanatory variables in the model are now teach, knowledge, grade, and enroll.

Step 2: Removing one explanatory variable at a time and leaving the other three in the model

Remove grade: Model AIC = -110.74

Remove enroll: Model AIC = -110.14

Remove knowledge: Model AIC = -97.22

Remove teach: Model AIC = -76.54

Removing grade leads to a model containing the other three explanatory variables, and this model has a lower AIC than the current four-explanatory-variable model. Consequently, we drop the grade variable and start afresh with the model containing the variables enroll, knowledge, and teach.

Current: AIC = -110.74

Explanatory variables in the model are now enroll, knowledge, and teach.

Step 3: Removing one explanatory variable at a time and leaving the other two in the model

Remove enroll: Model AIC = -110.98

Remove knowledge: Model AIC = -98.58

Remove teach: Model AIC = -77.38

Removing enroll leads to a model containing the other two explanatory variables, and this model has a lower AIC than the current three-explanatory-variable model. Consequently, we drop the enroll variable and start afresh with the model containing the variables knowledge and teach.

Current: AIC = -110.98

Explanatory variables in the model are now teach and knowledge.

Step 4: Removing one explanatory variable at a time and leaving the other one in the model

Remove knowledge: Model AIC = -97.81

Remove teach: Model AIC = -77.12

Removal of either one of the two variables, teach and knowledge, results in a model with a far higher value of the AIC than the model containing both these variables. Consequently, we accept this as our final model.

We now need to fit the chosen model to the data to get the relevant estimated regression coefficients, etc. The results are shown in [Table 4.6](#). We see that an

TABLE 4.6

Results of Fitting the Multiple Linear Regression Model with the Two Explanatory Variables, Teach and Knowledge, to the Course Evaluation Data

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	-1.2984	0.4773	-2.720	0.009121
Teach	0.7097	0.1011	7.021	7.6e-09
Knowledge	0.5383	0.1319	4.082	0.000172

Note: Residual standard error: 0.3202 on 47 DF; multiple R-squared: 0.7388; F-statistic: 66.47 on 2 and 47 DF; p-value: 1.991e-14.

increase in one unit in teach leads to an estimated increase of 0.71 overall, conditional on knowledge; and an increase of one unit in knowledge leads to an increase of 0.54 overall, conditional on teach. The square of the multiple correlation coefficient for this model is 0.74, only a little less than its value of 0.76 in the five-variable model.

4.4 Regression Diagnostics

Having selected a more parsimonious model, there still remains one further important aspect of a regression analysis to consider, and that is to check the assumptions on which the model is based. We have already described in Chapter 3 the use of residuals for this purpose, but in this section we shall go into a little more detail and introduce several other useful regression diagnostics that are now available. These diagnostics provide ways for identifying and understanding the differences between a model and the data to which it is fitted. Some differences between the data and the model may be due to isolated observations; one, or a few, observations may be outliers, or may differ in some unexpected way from the rest of the data. Other differences may be systematic; for example, a term may be missing in a linear model. Technical [Section 4.2](#) describes a number of regression diagnostics.

Technical [Section 4.2](#): Regression Diagnostics

To begin, we need to introduce what is known as the hat matrix \mathbf{H} , defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where \mathbf{X} is the matrix introduced earlier in the chapter in Technical [Section 4.1](#) dealing with estimation of the multiple linear regression model.

In a multiple linear regression, the predicted values of the response variable can be written in matrix form as $\hat{\mathbf{y}} = \mathbf{Hy}$ so that \mathbf{H} “puts the hats” on \mathbf{y} . The diagonal elements of \mathbf{H} , h_{ii} , $i = 1, \dots, n$, are such that $0 \leq h_{ii} \leq 1$, and have an average value of q/n . Observations with large values of h_{ii} are said to have high leverage, and such observations have the most effect on the estimation of the model parameters. It is often helpful to produce a plot of h_{ii}

against i , an index plot, to identify any observations that may have undue influence on fitting the model.

The raw residuals introduced in Chapter 3 are not independent of one another, nor do they have the same variance because the variance of $r_i = y_i - \hat{y}_i$ is $\sigma^2 = (1 - h_{ii})$. Both properties make the raw residuals less useful than they might be when amended a little. Two alternative residuals are the standardized residual and the deletion residual; both are based on the raw residual r_i and are defined as follows:

$$r_i^{\text{std}} = \frac{r_i}{\sqrt{s^2(1 - h_{ii})}}$$

$$r_i^{\text{del}} = \frac{r_i}{\sqrt{s_{(i)}^2(1 - h_{ii})}}$$

where $s_{(i)}^2$ is the residual mean square estimate of σ^2 after the deletion of observation i .

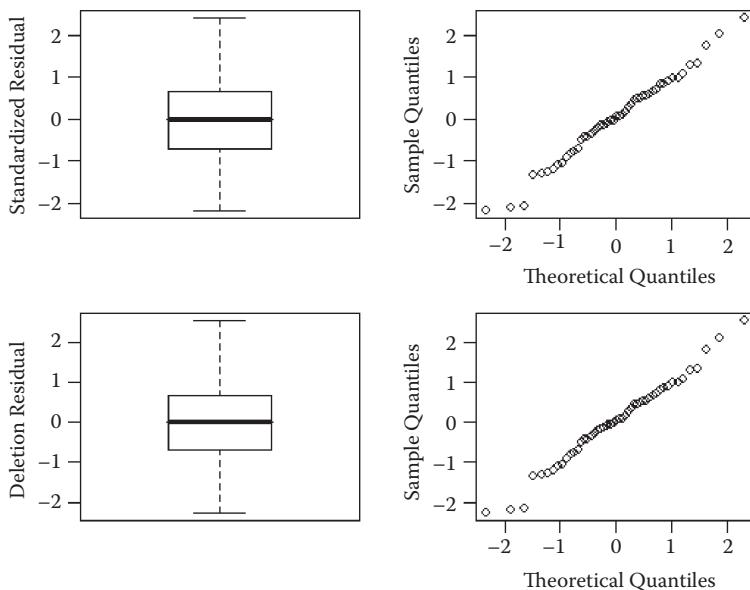
The deletion outliers are often particularly helpful for the identification of outliers. A further useful regression diagnostic is Cook's distance, D_i , defined as

$$D_i = \frac{r_i h_{ii}}{\sqrt{qs^2(1 - h_{ii})}}$$

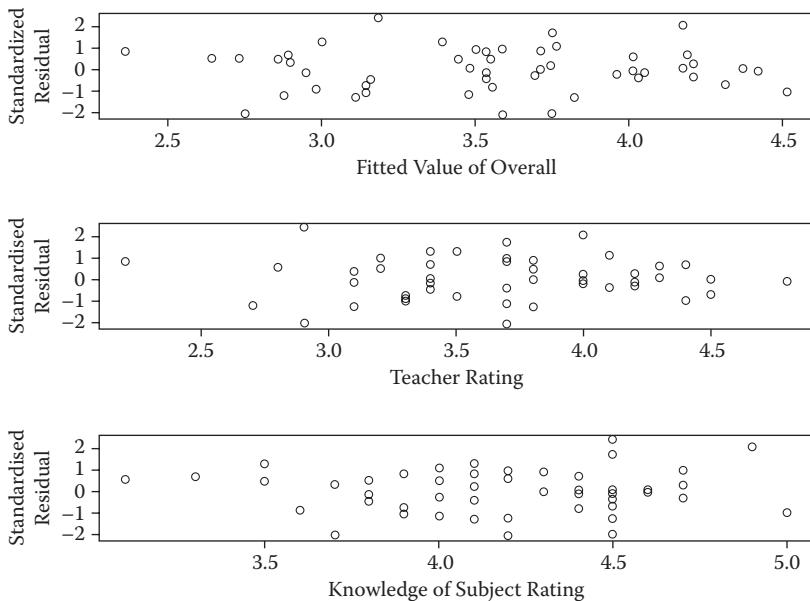
Cook's distance measures the influence of observation i on the estimation of all the parameters in the model. Values greater than 1 suggest that the corresponding observation has undue influence on the estimation process.

A full account of regression diagnostics is given in Cook and Weisberg (1982).

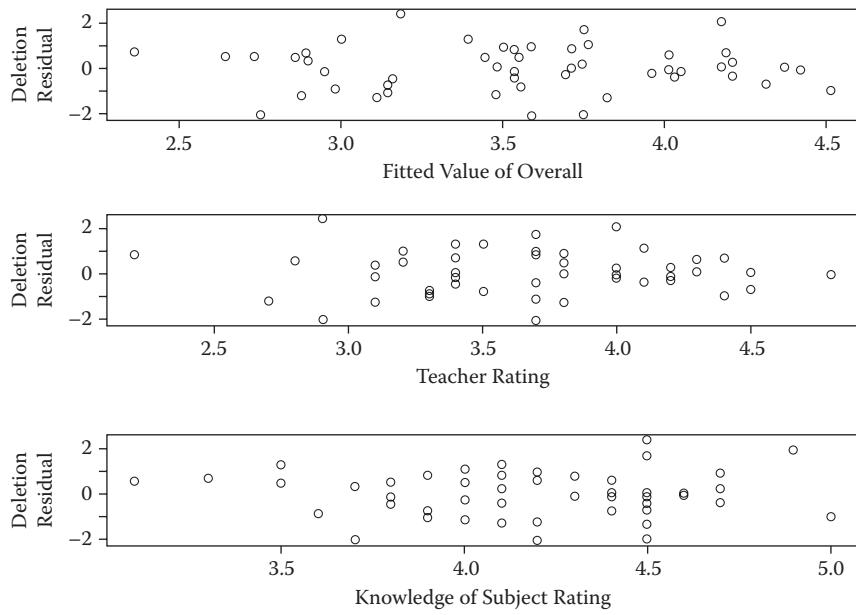
We can now take a look at these regression diagnostics using the final model chosen for the course evaluation data, namely, a model containing only the two explanatory variables: teach and knowledge. Figure 4.5 shows boxplots and normal probability plots for both the standardized and deletion residuals. The corresponding plots look very similar and give no cause for concern for the model fitted. Figure 4.6 shows plots of the standardized residuals against the fitted value of the overall rating: the rating of teaching ability and perceived knowledge. Figure 4.7 shows the same three plots using deletion residuals. Again, the two sets of plots are very similar and raise no issues about the fitted model. Figure 4.8 shows an index plot of Cook's distances; none is greater than 1, and so, once again, this diagnostic, like the others, gives us some confidence that we have not violated any assumptions when fitting the chosen model.

**FIGURE 4.5**

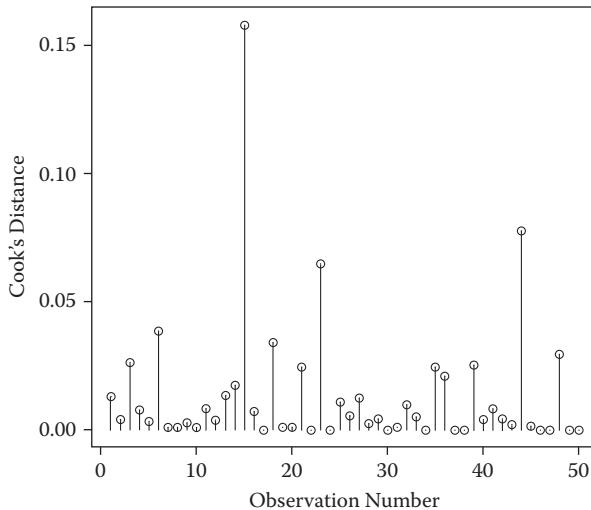
Boxplots and normal probability plots for both standardized and deletion residuals from the final model chosen for the course evaluation data.

**FIGURE 4.6**

Plots of standardized residuals against the fitted value of the overall rating—the rating of teaching ability and perceived knowledge for the course evaluation data.

**FIGURE 4.7**

Plots of deletion residuals against the fitted value of the overall rating—the rating of teaching ability and perceived knowledge for the course evaluation data.

**FIGURE 4.8**

Index plot of Cook's distances for the final model chosen for the course evaluation data.

4.5 Summary

- Multiple linear regression is used to assess the relationship between a set of explanatory variables and a continuous-response variable.
- The response variable is assumed to be normally distributed with a mean that is a linear function of the explanatory variables, and a variance that is independent of them.
- The explanatory variables are strictly assumed to be fixed. In practice, where this is almost never the case, the results of multiple regression are to be interpreted conditional on the observed values of these variables.
- It may be possible to find a more parsimonious model for the data, that is, one with fewer explanatory variables using all subsets of regression or one of the “stepping” methods. Care is required when using the latter.
- An extremely important aspect of a regression analysis is the inspection of a number of regression diagnostics in a bid to identify any departures from assumptions, outliers, etc.

4.6 Exercises

4.1 The data in `ex_41.txt` were collected to investigate the determinants of pollution. For 41 cities in the United States, seven variables were recorded:

SO ₂ :	SO ₂ content of air in micrograms per cubic meter
Temp:	Average annual temperature in degrees Fahrenheit
Manuf:	Number of manufacturing enterprises employing 20 or more workers
Pop:	Population size (according to 1970 census) in thousands
Wind:	Average annual wind speed in miles per hour
Precip:	Average annual precipitation in inches
Days:	Average number of days with precipitation per year

Construct a scatterplot matrix of the data and use it to guide the fitting of a multiple linear regression model with SO₂ as the response variable and the remaining variables as explanatory. Find the variance inflationary factor for each explanatory variable and use the factors to decide if there are any problems in using all six of the explanatory

variables. Use the procedure involving the AIC described in the text to search for a more parsimonious model for the data. For the final model chosen, use some regression diagnostics to investigate the assumptions made in fitting the model.

- 4.2 The data in ex_42.txt arise from a study of the quality of statements elicited from young children reported by Hutcheson et al. (1995). The variables are statement quality, child's gender, age and maturity, how coherently the children gave their evidence, the delay between witnessing an incident and recounting it, the location of the interview (the child's home, school, a formal interviewing room, or an interview room specially constructed for children), and whether or not the case proceeded to prosecution. Carry out a complete regression analysis on these data to see how statement quality depends on the other variables, including selection of the best subset of explanatory variables and examining residuals and other regression diagnostics. Pay careful attention to how the categorical explanatory variables with more than two categories are coded.
- 4.3 In ex_43.txt, four sets of bivariate data are given. Fit a simple linear regression to each data set. What do you find? Now construct regression graphics and describe what you conclude from these plots.
- 4.4 In ex_44.txt, the age, percentage fat, and gender of 20 normal adults are given. Investigate multiple linear regression models with the percentage of fat as the response variable, and age and gender as explanatory variables. Illustrate the models you fit with informative graphics.

5

The Equivalence of Analysis of Variance and Multiple Linear Regression, and an Introduction to the Generalized Linear Model

5.1 Introduction

The phrase “analysis of variance” (ANOVA) was coined by arguably the most famous statistician of the 20th century, Sir Ronald Aylmer Fisher, who defined the technique as “the separation of variance ascribable to one group of causes from the variance ascribable to other groups.” ANOVA is probably the piece of statistical methodology most widely used by behavioral scientists, but there is no chapter simply entitled analysis of variance in this book. Why not? The primary reason is that the multiple linear regression model described in Chapter 4 is essentially ANOVA in disguise, and so there is really no need to describe each technique separately. Further, showing the equivalence of ANOVA and multiple linear regression, as we will in Section 5.2, will also enable us to say a few words about what is known as the generalized linear model, a powerful method for the analysis of many types of data, as we shall see later in this chapter and in Chapter 6.

5.2 The Equivalence of Multiple Regression and ANOVA

In a study of fecundity of fruit flies, per-diem fecundity (average number of eggs laid per female per day for the first 14 days of life) for 25 females of each of three genetic lines of the fruit fly *Drosophila melanogaster* was recorded. The lines labeled RS and SS were selectively bred for resistance and susceptibility to the pesticide DDT, and the line NS as a nonselected control strain. The results for the first three fruit flies of each genetic line are shown in Table 5.1. Of interest here is whether the data give any evidence of a difference in fecundity of the three strains.

In this study, the effect of a single independent factor (genetic strain) on a response variable (per-diem fecundity) is of interest. The data arise from

TABLE 5.1
Fecundity of Fruit Flies

Resistant (RS)	Susceptible (SS)	Nonselected (NS)
12.8	38.4	35.4
21.6	32.9	35.4
14.8	48.5	19.3

what is generally known as a one-way design and would usually be dealt with by analysis of variance based on the following model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where y_{ij} represents the value of the j th observation in the i th genetic line, μ represents the overall mean of the response, α_i is the effect of the i th genetic line ($i = 1, 2, 3$), and ε_{ij} are random error terms assumed to have a normal distribution with mean zero and variance σ^2 . The model has four parameters to describe three group means and is said to be overparameterized, which causes problems because it is impossible to find unique estimates for each parameter. This aspect of ANOVA models is discussed in detail in Maxwell and Delaney (2003), but essentially, overparameterization is overcome by imposing constraints on the parameters. In the fruit fly example, we will assume that $\alpha_1 + \alpha_2 + \alpha_3 = 0$. The usual analysis of variance table for the fruit fly data is shown in Table 5.2. The F -test is highly significant, and there is strong evidence that the average number of eggs laid per day differs among the three lines.

How can this analysis be undertaken using multiple linear regression? First, we introduce two dummy variables x_1 and x_2 defined below, which are used to label the three genetic lines:

	Genetic Line		
	RS	SS	NS
x_1	1	0	-1
x_2	0	1	-1

TABLE 5.2
Analysis of Variance (ANOVA) Table for Fruit Fly Data

Source	Sum of Squares	Df	Mean Square	F	p-Value
Between lines	1362.21	2	681.11	8.67	< 0.001
Within lines (error)	5659.02	72	78.60	—	—

The usual one-way ANOVA model for this situation is the one described earlier:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \text{ with } \alpha_1 + \alpha_2 + \alpha_3 = 0$$

This can be rewritten in terms of the variables x_1 and x_2 as

$$y_{ij} = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon_{ij}$$

and this is exactly the same form as a multiple linear regression model with two explanatory variables. So, applying multiple regression and regressing average number of eggs laid per day on x_1 and x_2 , what do we get? The regression sum of squares is 1362.62 with 2 degrees of freedom, and the residual sum of squares is 5659.02 with 72 degrees of freedom. The results are identical to those from ANOVA, and the estimates of the regression coefficients from the regression analysis are

$$\hat{\mu} = 27.42, \hat{\alpha}_1 = -2.16, \hat{\alpha}_2 = -3.79$$

The estimates of α_1 and α_2 are simply the differences between each genetic line mean and the grand mean.

Now let us consider a 2×2 factorial design with factors A and B both at two levels: A1 and A2, and B1 and B2. The usual ANOVA model for such a design is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where y_{ijk} represents the k th observation in the ij th cell of the design; α_i represents the effect of the i th level of factor A; β_j represents the effect of the j th level of factor B; γ_{ij} represents the interaction of A and B; and, as always, ε_{ijk} represents random error terms with the usual distributional assumptions. The usual constraints on the parameters to deal with overparameterization in this case are

$$\sum_{i=1}^2 \alpha_i = 0, \sum_{j=1}^2 \beta_j = 0, \sum_{i=1}^2 \gamma_{ij} = \sum_{j=1}^2 \gamma_{ij} = 0$$

These constraints imply that the parameters in the model are such that

$$\alpha_1 = -\alpha_2, \beta_1 = -\beta_2, \gamma_{1j} = -\gamma_{2j}, \gamma_{i1} = -\gamma_{i2}$$

The last two equations imply that

$$\gamma_{12} = -\gamma_{11}, \gamma_{21} = -\gamma_{11}, \gamma_{22} = \gamma_{11}$$

showing that there is only a single interaction parameter. The model for the observations in each of the four cells of the design can now be written explicitly as follows:

	A1	A2
B1	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$	$\mu - \alpha_1 + \beta_1 - \gamma_{11}$
B2	$\mu + \alpha_1 - \beta_1 - \gamma_{11}$	$\mu - \alpha_1 - \beta_1 + \gamma_{11}$

Now we define two variables as follows:

$$\begin{aligned}x_1 &= 1 \text{ if first level of A, } x_1 = -1 \text{ if second level of A.} \\x_2 &= 1 \text{ if first level of B, } x_2 = -1 \text{ if second level of B.}\end{aligned}$$

The original ANOVA model for the design can now be written as

$$y_{ijk} = \mu + \alpha_1 x_1 + \beta_1 x_2 + \gamma_{11} x_3 + \varepsilon_{ijk}, \text{ where } x_3 = x_1 \times x_2$$

We can now recognize this as a multiple linear regression model with three explanatory variables, and we can fit it in the usual way. Here, the fitting process can be used to illustrate the difference in analyzing a balanced 2×2 design (equal number of observations per cell) and an unbalanced design (unequal number of observations per cell). To begin, we will apply the multiple regression model to the data in [Table 5.3](#).

So, for fitting the multiple regression model, all observations in cell A1,B1 have $x_1 = 1$ and $x_2 = 1$; all observations in cell A1,B2 have $x_1 = 1$, $x_2 = -1$; and so on for the remaining observations in [Table 5.4](#). To begin, we will fit the model with the single explanatory variable x_1 to give the following results:

Source	Sum of Squares	Df	Mean Square
Regression	12.25	1	12.25
Residual	580.75	14	41.48

and $\hat{\mu} = 28.75$, $\hat{\alpha}_1 = -0.875$.

The regression sum of squares 12.25 is what would be the between levels of A sum of squares in an ANOVA table.

Now fit the regression with x_1 and x_2 as explanatory variables to give the following results:

Source	Sum of Squares	Df	Mean Square
Regression	392.50	2	196.25
Residual	200.50	13	15.42

and $\hat{\mu} = 28.75$, $\hat{\alpha}_1 = -0.875$, $\hat{\beta}_1 = -4.875$.

The difference between the regression sums of squares for the two-variable and one-variable models gives the sum of squares for factor B that would be obtained in an ANOVA.

TABLE 5.3
A Balanced 2×2 Data Set

	A1	A2
B1	23	22
	25	23
	27	21
	29	21
B2	26	37
	32	38
	30	40
	31	35

TABLE 5.4
Unbalanced 2×2 Data Set

	A1	A2
B1	23	22
	25	23
	27	21
	29	21
	30	19
	27	23
	23	17
	25	—
B2	26	37
	32	38
	30	40
	31	35
	—	39
	—	35
	—	38
	—	41
	—	32
	—	36
	—	40
	—	41
	—	38

Finally, we can fit a model with three explanatory variables to give the following:

Source	Sum of Squares	Df	Mean Square
Regression	536.50	3	178.83
Residual	56.50	12	4.71

and $\hat{\mu} = 28.75$, $\hat{\alpha}_1 = -0.875$, $\hat{\beta}_1 = -4.875$, $\gamma_{11} = 3.000$.

The difference between the regression sums of squares for the three-variable and two-variable models gives the sum of squares for the $A \times B$ interaction that would be obtained in an analysis of variance. The residual sum of squares in the final step corresponds to the error sum of squares in the usual ANOVA table. (Readers might like to confirm these results by running an analysis of variance on the data.)

Note that, unlike the estimated regression coefficients in the examples considered in Chapter 4, the estimated regression coefficients for the balanced 2×2 design do not change as extra explanatory variables are introduced into the regression model. The factors in a balanced design are independent; a more technical term is that they are orthogonal. When the explanatory variables are orthogonal, adding variables to the regression model in a different order than the one used earlier will alter nothing; the corresponding sums of squares and regression coefficient estimates will be the same. Is the same true of an unbalanced example? To answer this question, we shall use the data in [Table 5.4](#).

Again, we will fit regression models first with only x_1 , then with x_1 and x_2 , and finally with x_1 , x_2 , and x_3 .

Results for x_1 model:

Source	Sum of Squares	Df	Mean Square
Regression	149.63	1	149.63
Residual	1505.87	30	50.19

and $\hat{\mu} = 29.567$, $\hat{\alpha}_1 = -2.233$.

The regression sum of squares gives the sum of squares for factor A.

Results for x_1 and x_2 model:

Source	Sum of Squares	Df	Mean Square
Regression	1180.86	2	590.42
Residual	476.55	29	16.37

and $\hat{\mu} = 29.667$, $\hat{\alpha}_1 = -0.341$, $\hat{\beta}_1 = -5.997$.

The difference in the regression sums of squares for the two-variable and one-variable models gives the sum of squares due to factor B, conditional on A already being in the model.

Results for x_1 , x_2 , and x_3 model:

Source	Sum of Squares	Df	Mean Square
Regression	1474.25	3	491.42
Residual	181.25	28	6.47

and $\hat{\mu} = 28.606$, $\hat{\alpha}_1 = -0.667$, $\hat{\beta}_1 = -5.115$, $\hat{\gamma}_{11} = 3.302$.

The difference in the regression sums of squares for the three-variable and two-variable models gives the sum of squares due to the $A \times B$ interaction, conditional on A and B being in the model.

For an unbalanced design the factors are no longer orthogonal, and so the estimated regression parameters change as further variables are added to the model, and the sums of squares for each term in the model are now conditional on what has entered the model before them. If variable x_2 was entered before x_1 , then the results would differ from those given earlier, as readers should confirm by repeating the fitting process as an exercise.

So, using the regression approach clearly demonstrates why there is a difference between analyzing a balanced design (not just a 2×2 design as in the example) and an unbalanced design. In the latter, the order of entering effects is important. From the need to consider order, a great deal of confusion has arisen. For example, some authors have suggested that, in a two-way unbalanced design with factors A and B, the main effects of A and B can be entered after the $A \times B$ interaction to give what are called type III sums of squares; indeed, this is the default in many software packages. However, this approach is heavily criticized by Nelder (1977) and Aitkin (1978). The arguments are relatively subtle, but they go something like this:

- When fitting models to data, the principle of parsimony is of critical importance. In choosing among possible models, we do not want to adopt complex models for which there is no empirical evidence.
- Thus, if there is no convincing evidence of an $A \times B$ interaction, we do not retain this term in the model. Thus, additivity of A and B is assumed unless there is convincing evidence to the contrary.
- So, the argument proceeds that type III sum of squares for, say, A, in which it is adjusted for the $A \times B$ interaction, makes no sense.
- First, if the interaction term is necessary in the model, then the experimenter will usually want to consider simple effects of A at each level of B separately. A test of the hypothesis of no A main effect would not usually be carried out if the $A \times B$ interaction is significant.

- If the $A \times B$ interaction is not significant, then adjusting for it is of no interest and causes a substantial loss of power in testing A and B main effects.

The arguments of Nelder and Aitkin against the use of type III sums of squares are persuasive and powerful. Their recommendation to use what are generally known as type I sums of squares in which interaction terms are considered after the main effects of the factors in the interaction term, perhaps considering main effects in a number of orders, as the most suitable way in which to identify a suitable model for a data set is also convincing and strongly endorsed by this author.

Now we have seen, in a series of examples, the equivalence of ANOVA and multiple linear regression models, we can move on to look at a general framework for linear models, which will allow the appropriate models to be fitted to response variables that do not satisfy the assumptions required by such models.

Note that, although ANOVA models can be expressed as multiple linear regression models, we are not suggesting that behavioral researchers should stop using the ANOVA module in whatever software they use, because that module will conveniently take care of the conversion to a multiple linear regression model and print out the usual analysis of variance table that is required by the researcher.

5.3 The Generalized Linear Model

The term generalized linear model (GLM) was first introduced in a landmark paper by Nelder and Wedderburn (1972), in which a wide range of seemingly disparate problems of statistical modeling and inference were set in an elegant unifying framework of great power and flexibility. To begin this account of such models, let us first return to the multiple linear regression model described in Chapter 4. The model has the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon$$

The error term ε is assumed to have a normal distribution with zero mean and variance σ^2 . An equivalent way of writing the multiple linear regression model is as

$$y \sim N(\mu, \sigma^2)$$

where $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$. This makes it clear that this model is only suitable for continuous-response variables conditional on the values of the explanatory variables, a normal distribution with a mean that is a linear

function of the explanatory variables, and a variance that is not dependent on the values of the explanatory variables. (ANOVA is essentially exactly the same model, with x_1, \dots, x_q , being dummy variables coding factor levels; analysis of covariance is also the same model with a mixture of continuous and categorical explanatory variables.)

The assumption of the conditional normality of a continuous-response variable is one that is probably made more often than it is warranted, and there are many situations in which such an assumption is clearly not justified. In behavioral research, response variables that are neither continuous nor normally distributed are common. The most obvious examples are binary responses such as experimental task completed/experimental task not completed, or agree/disagree with a particular statement, etc. Another example is one where the response is a count, for example, the number of correct answers in a testing situation. The question then arises as to how the multiple linear model might be modified to allow such responses to be related to the explanatory variables of interest. The generalization of the multiple linear regression model in generalized linear models consists of allowing the following two assumptions associated with the former model to be modified:

- The response variable is normally distributed with a mean that is a linear function of the explanatory variables (remember this can include nonlinear terms in the explanatory variables). The effects of the explanatory variables on the mean of the response are additive.
- The variance of the response remains the same for all values of the explanatory variables.

In a GLM, some transformation of the mean of the response is modeled by a linear function of the explanatory variables, and the distribution of the response about its mean (often referred to as the error distribution) is generalized usually in a way that fits naturally with a particular transformation. Technical [Section 5.1](#) gives a brief account of GLMs.

Technical [Section 5.1](#): GLMs

The first component of a GLM is a linear predictor, η , formed from the explanatory variables $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$. The next component is a transformation of the mean, μ , of the response variable called the link function $g(\mu)$. In a GLM, it is $g(\mu)$ that is modeled by the linear predictor to give $g(\mu) = \eta$. In multiple linear regression and ANOVA, the link function is the identity function. Other common link functions include the log function and the logit function, the latter being the basis of logistic regression for a binary response, a topic to be covered in detail in Chapter 6. The final component of a GLM is the distribution of the

response variable given its mean μ , and this is assumed to be one of a class of distribution functions called the exponential family. Some distributions that are contained in this family are normal distribution, binomial distribution, Poisson distribution, gamma distribution, and exponential distribution. Particular link functions are naturally associated with particular error distributions, for example, the identity link with the normal distribution, the log link with the Poisson, and the logit with the binomial (see Chapter 6).

The choice of probability distribution determines the relationship between the variance of the response variable conditional on the explanatory variables and its mean. The general form of the relationship is $\text{Var}(\text{response}) = \phi V(\mu)$, where $V(\mu)$ is known as the variance function that specifies how the variance depends on the mean, and ϕ is a constant. For three of the error distributions mentioned earlier, this general form becomes

- Normal: $V(\mu) = 1$, $\phi = \sigma^2$; here, the variance does not depend on the mean, and so can be freely estimated.
- Binomial: $V(\mu) = \mu(1 - \mu)$, $\phi = 1$.
- Poisson: $V(\mu) = \mu$, $\phi = 1$.

In the case of a Poisson variable, we see that the mean and variance are equal, and in the case of a binomial variable, where the mean is the probability of, say, the event of interest π , the variance is $\pi(1 - \pi)$. Both the Poisson and binomial have variance functions that are completely determined by the mean. This can be a problem when fitting GLMs using the binomial or Poisson error distributions to some data sets where there is a failure to fully account for the empirical variance in the data. Such a phenomenon is usually known as overdispersion; we will not deal with it in this book.

The parameters in a GLM are estimated by maximum likelihood, and full details of the estimation process and GLMs themselves are given in McCullagh and Nelder (1989), and a more concise description in Dobson and Barnett (2008).

Although GLMs include a wide range of statistical models, we shall deal with only one in detail in the book, and that is logistic regression, the subject of Chapter 6.

5.4 Summary

- The models used in ANOVA are equivalent to those used in multiple linear regression.

- By using dummy variables to appropriately code the levels of the factors in an ANOVA design, the model for the design can be put in the form of a multiple linear regression model.
 - ANOVA software essentially transforms the required analysis into a regression format and then gives results in the form of an ANOVA table. Consequently, such software remains useful for undertaking the required analysis.
 - The generalized linear model allows a suitable transform of the mean of the response variable to be modeled as a linear function of the explanatory variables, and to have an error distribution appropriate for the type of response involved. The model will be developed in Chapter 6 in the context of a binary response.
-

5.5 Exercises

- 5.1 The data in *ex_51.txt* arise from a survey of systolic blood pressure in individuals classified according to smoking status and family history of circulation and heart problems. Analyze the data using multiple linear regression and find the residuals from the fitted model, and use them to check the assumptions made by the model you have fitted.
- 5.2 The data in *ex_52.txt* were collected in a clinical trial of the use of estrogen patches in the treatment of postnatal depression. Using posttreatment depression score as the response, formulate a suitable model for examining the effects of the baseline measurements and treatment group on the response. Construct a suitable 95% confidence interval for the treatment effect and state your conclusions based on the analyses you have carried out.

6

Logistic Regression

6.1 Introduction

In a study of a psychiatric screening questionnaire called the GHQ (General Health Questionnaire; see Goldberg, 1972), the men and women participating were given a score on the GHQ and also categorized as being a psychiatric case or not. Here, the question of interest to the researcher is how being judged to be a “case” is related to gender and GHQ score. Part of the data is shown in [Table 6.1](#); in this table the binary responses (case/not case) of individuals with the same values of the two explanatory variables, GHQ score and gender, have been grouped together.

How we go about analyzing these data in an appropriate manner will be taken up later in the chapter, but before that we need a small digression to look at odds and odds ratios because these quantities will be important in interpreting results from fitting the models we shall apply to the data.

6.2 Odds and Odds Ratios

If we collapse the psychiatric caseness data over the GHQ score, we get the following 2×2 contingency table of caseness against gender.

	Case	Noncase
Male	25	79
Female	43	131

Such a table would usually be analyzed to assess the independence or otherwise of gender and caseness using a chi-squared test. However, here we will use the table to explain the meanings of the terms odds and odds ratios.

First odds, which is defined for an event with probability p as $p/(1 - p)$. For women in the 2×2 table in the preceding text, the probability of being judged a case is estimated to be $43/174 = 0.247$, and so, for women, the odds of being judged a case versus being judged a noncase is $0.247/(1 - 0.247) = 0.328$. The same calculations for men show that the probability of being judged a case is 0.240, and the corresponding value for the odds is 0.316. It is easy to see that

TABLE 6.1
Psychiatric Caseness Data

GHQ Score	Gender	Number of Cases	Number of Noncases
0	F	4	80
1	F	4	29
2	F	8	15
	
10	F	1	0
0	M	1	36
1	M	2	25
2	M	2	8
	
10	M	2	0

the odds for women can be calculated directly from the frequencies in the 2×2 table as $43/131$; similarly, for men, the odds are $25/79$. Further, having found the odds for caseness versus noncaseness for women and for men, the odds ratio is simply what it says—the ratio of the two odds, that is, $0.316/0.328 = 0.963$. When the two variables forming the contingency table are independent, the odds ratio in the population will be 1. So, is it possible to use the estimated odds ratio to test the hypothesis that the population value is 1 and, more important, is it possible to construct a confidence interval (CI) for the odds ratio? Technical Section 6.1 shows how to do the latter.

Technical Section 6.1: CI for the Odds Ratio

Consider the general 2×2 table given by

		Variable 1	Variable 2
		Category 1	Category 2
Category 1		a	b
Category 2		c	d

The odds ratio in the population will be denoted by ψ , and it can be estimated from the observed frequencies in the table as

$$\hat{\psi} = \frac{ad}{bc}$$

A CI for ψ can be constructed relatively simply by using the following estimator of the variance of $\log(\psi)$:

$$\hat{\text{var}}(\log \psi) = 1/a + 1/b + 1/c + 1/d$$

So, an approximate 95% confidence interval for $\log(\psi)$ is given by

$$\log(\hat{\psi}) \pm 1.96 \times \sqrt{\text{var}(\log \hat{\psi})}$$

If the limits of the CI for $\log(\psi)$ obtained in this way are ψ_L, ψ_U , then the corresponding confidence interval for ψ is $\exp(\psi_L)$ and $\exp(\psi_U)$.

We can illustrate the construction of the CI for the odds ratio using the data from the 2×2 table of gender and caseness given earlier in the chapter. First, the odds ratio is estimated to be

$$\hat{\psi} = \frac{25}{43} \times \frac{131}{79} = 0.964$$

and so $\log(\hat{\psi}) = -0.037$ and the estimated variance of $\log(\psi)$ is $1/43 + 1/131 + 1/25 + 1/79 = 0.084$, leading to a 95% CI for $\log(\psi)$ of $[-0.036 - 1.96 \times 0.290, -0.036 + 1.96 \times 0.290]$, that is, $[-0.604, 0.531]$.

Finally, the CI for ψ itself is found as $[\exp(-0.604), \exp(0.531)]$, that is, $[0.546, 1.701]$.

As this interval contains the value 1, we can conclude that there is no evidence of an association between caseness and gender.

6.3 Logistic Regression

In any regression problem, the key quantity is the population mean (or expected value) of the response variable given the values of the explanatory variables. As we have learned in Chapter 4, in multiple linear regression, the mean of the response variable is modeled directly as a linear function of the explanatory variables, that is, using the E operator to denote expected value

$$E(y | x_1, x_2, \dots, x_q) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

where $E(y | x_1, x_2, \dots, x_q)$ represents the expected value (population mean) of the response variable given the values of the explanatory variables. However, we now want to consider how to model appropriately a binary response variable with its two categories labeled 0 and 1. The first question we need to ask is: what is the expected value of such a variable? It is easy to show that the mean (expected value) in this case is simply the probability that the response variable takes the value 1. (We shall denote this probability by π .) So, to investigate the effects of a set of explanatory variables on a binary response, why not simply continue to use the multiple linear regression approach and consider a model of the form

$$\pi = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

There are two problems with this model:

- The predicted value of the probability π must satisfy $0 \leq \pi < 1$, whereas a linear predictor can yield values from minus infinity to plus infinity.
- The observed values of the response variable y , conditional on the values of the explanatory variables, will not now follow a normal distribution with mean π but rather what is known as a Bernoulli distribution.

Details of how these problems are overcome by what is known as logistic regression are described below.

Technical Section 6.2: Logistic Regression

We have a binary response variable y , coded 0 or 1, and a set of explanatory variables x_1, x_2, \dots, x_q . The mean of the response variable given the values of the explanatory variables is the probability that it takes the value 1; we represent this probability as π . Because of the problems identified earlier, π cannot be modeled directly as a linear function of the explanatory variables; instead, some suitable transformation of π must be modeled. The transformation most often used is the logit function of the probability, which is simply the log(odds), that is, $\log(\frac{\pi}{1-\pi})$, and this leads to the logistic regression model having the form

$$\text{logit}(\pi) = \log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

The logit transformation is chosen because, from a mathematical point of view, it is extremely flexible and, from a practical point of view, it leads to meaningful and convenient interpretation, as we will discuss later when looking at examples of the application of logistic regression. The logit transformation of π can take values ranging from $-\infty$ to $+\infty$ and thus overcome the first problem associated with modeling π directly. In a logistic regression model, the parameter β_i associated with the explanatory variable x_i represents the expected change in the log odds when x_i is increased by one unit, conditional on the other explanatory variables remaining the same. Interpretation is simpler using $\exp(\beta_i)$, which corresponds to an odds ratio, as we will see later when discussing some examples.

The preceding logistic regression model can be rearranged to give the following model for π :

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q)}$$

This equation can be used to predict the probability that the response variable takes the value 1 for any values of the explanatory variables, but it would, of course, only make sense for values in the observed range of the explanatory variables in the data set being modeled.

In linear regression, the observed value of the outcome variable is expressed as its expected value, given the explanatory variables plus an error term. The error terms are assumed to have a normal distribution with mean 0 and a variance that is constant across levels of the explanatory variables. With a binary response, we can express an observed value y in the same way as

$$y = \pi(x_1, x_2, \dots, x_q) + \varepsilon$$

but here ε can only assume one of two possible values (note that here we have introduced a slight change in the nomenclature to remind us that the expected value of the response is dependent on the explanatory variables). If $y = 1$ then $\varepsilon = 1 - \pi(x_1, x_2, \dots, x_q)$ with probability $\pi(x_1, x_2, \dots, x_q)$, and if $y = 0$, then $\varepsilon = \pi(x_1, x_2, \dots, x_q)$ with probability $1 - \pi(x_1, x_2, \dots, x_q)$. Consequently, ε has a distribution with mean 0 and variance equal to $\pi(x_1, x_2, \dots, x_q)(1 - \pi(x_1, x_2, \dots, x_q))$. So, the distribution of the response variable conditional on the explanatory variables follows what is known as a Bernoulli distribution (which is simply a binomial distribution for a single trial) with probability $\pi(x_1, x_2, \dots, x_q)$.

Maximum likelihood is used to estimate the parameters in the logistic regression model (for details, see Collett, 2003a). Intuitively, what the estimation procedure tries to do is to find estimates of the regression parameters that make the predicted probabilities as close as possible, in some sense, to the observed probabilities.

The lack of fit of a logistic regression model can be measured by a term known as the *deviance*, which is essentially the ratio of the likelihoods of the model of interest to the saturated model that fits the data perfectly (see Collett, 2003a, for a full explanation). Differences in deviance can be used to compare alternative nested logistic regression models. For example,

$$\text{Model 1 (Deviance } D_1\text{): } \text{logit}(\pi) = \beta_0 + \beta_1 x_1$$

$$\text{Model 2 (Deviance } D_2\text{): } \text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The difference in the two deviances, $D_1 - D_2$, reflects the combined effect of explanatory variables x_2 and x_3 , and under the hypothesis that these variables have no effect (i.e., β_2 and β_3 are both 0), the difference has an approximate chi-squared distribution with degrees of freedom (DF) equal to the difference in the number of parameters in the two models, which in this example is 2.

6.4 Applying Logistic Regression to the GHQ Data

To begin, we will fit both a logistic regression and a linear regression to the data using the GHQ score as the single explanatory variable. So, for the linear model, the probability of being a case is modeled as a linear function of the GHQ score, whereas in the logistic model, the logit transformation of the probability of being a case is modeled as a linear function of the GHQ score. The results from fitting both models are shown in Table 6.2. The results from both show that the GHQ score is a highly significant predictor of the probability of being judged a case. In the linear model, the estimated regression coefficient is 0.10; the estimated increase in the probability of being a case is 0.10 for each increase of 1 in the GHQ score. For an individual with a GHQ score of, say, 10, the linear model would predict that the probability of the individual being judged a case is $0.114 + 0.100 \times 10 = 1.114$; with this model, fitted values of the probabilities are not constrained to lie in the interval (0,1). Now consider the fitted logistic regression model, that is,

$$\log[\text{pr(case)}/\text{pr(not case)}] = -2.71 + 0.74 \times \text{GHQ score}$$

This equation can be rearranged to give the predicted probabilities for the fitted logistic regression model as

$$\text{pr(case)} = \exp(-2.71 + 0.84 \times \text{GHQ score})/[1 + \exp(-2.71 + 0.84 \times \text{GHQ score})]$$

For the individual with a GHQ score of 10, this model predicts the probability of the individual being judged a case as 0.99.

TABLE 6.2

Results of Fitting a Linear and a Logistic Regression Model to the GHQ Data with Only a Single Explanatory Variable, the GHQ Score

Logistic Model ^a				
	Estimate	Standard Error	z-Value	Pr(> z)
Intercept	-2.71073	0.27245	-9.950	<2e-16
GHQ	0.73604	0.09457	7.783	7.1e-15
Linear Model ^b				
	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	0.11434	0.05923	1.931	0.0678
GHQ	0.10024	0.01001	10.012	3.1e-09

^a Null deviance: 130.306 on 21 DF; residual deviance: 16.237 on 20 DF; AIC: 56.211; number of Fisher scoring iterations: 5.

^b Residual standard error: 0.1485 on 20 DF; multiple R-squared: 0.8337; F-statistic: 100.2 on 1 and 20 DF; p-value: 3.099e-09

Now let us consider the fitted logistic regression model for individuals with GHQ scores of, say, S and S + 1; the corresponding models for the logits can be written as

$$\log[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S] = -2.71 + 0.74 \times S$$

$$\log[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S + 1] = -2.71 + 0.74 \times (S + 1)$$

So, subtracting the first equation from the second, we get

$$\begin{aligned}\log[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S + 1] - \log[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S] \\ = 0.74\end{aligned}$$

This can be rewritten as

$$\log[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S + 1]/\text{pr}[(\text{case})/\text{pr}(\text{not case})|\text{GHQ} = S] = 0.74$$

And thus, by exponentiating, we get

$$[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S + 1]/[\text{pr(case)}/\text{pr(not case)}|\text{GHQ} = S] = \exp(0.74)$$

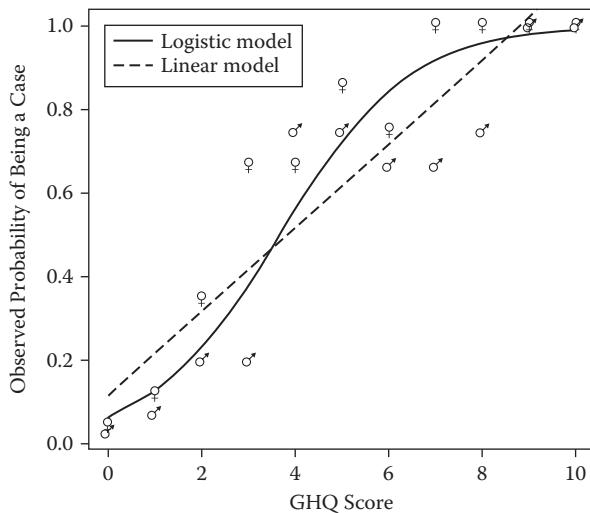
The left-hand side is simply the odds ratio for being rated a case for a subject with a GHQ score 1 higher than another subject, and this is estimated to be $\exp(0.74) = 2.10$. The corresponding approximate 95% CI is calculated as $[\exp(0.74 - 1.96 \times 0.09), \exp(0.74 + 1.96 \times 0.09)]$ to give [1.76, 2.50]. The increase in the odds of being judged a case associated with a 1 unit increase in the GHQ is estimated to be between 76% and 150%.

The null deviance given in [Table 6.2](#) is for a model with no explanatory variables, and the residual deviance is that for a model with GHQ as the single explanatory variable; the reduction in deviance is considerable, and could be tested as a chi-squared with 1 DF. The value of Akaike's fit criterion (see Chapter 4) is also given, and this might be useful when comparing competing models as we shall see later. Finally, the number of iterations used in getting the maximum likelihood estimates is shown to be 5.

In [Figure 6.1](#), we plot the predicted probabilities of being a case from each model against GHQ, along with the observed probabilities labeled by gender. (The observed probabilities are found from the data by dividing number of cases by number of cases plus number of noncases.) The plot again demonstrates that the linear regression model leads to predicted probabilities greater than 1 for some values of GHQ.

Next, we will fit a logistic regression model that includes only gender as an explanatory variable. The results are shown in [Table 6.3](#). The estimated regression coefficient of -0.037 is the odds ratio calculated earlier in the chapter for the gender \times case/not case cross-classification, and the estimated standard error of the estimated regression coefficient is also the same as we calculated earlier.

The next logistic model we will consider for the GHQ data is one where the gender and GHQ scores are both included as explanatory variables.

**FIGURE 6.1**

Plot of predicted probabilities of caseness from both linear and logistic regression models with GHQ score as the single explanatory variable and observed probabilities labeled by gender.

The results are shown in [Table 6.4](#). In this model, both GHQ and gender are significant. The regression coefficient for gender shows that, conditional on GHQ score, the log odds of caseness for men is -0.94 lower than for women, which gives an estimated odds ratio of $\exp(-0.94) = 0.39$ with 95% CI of [0.167, 0.918]. For a given GHQ score, the odds of a man being diagnosed as a case is between about 0.17 and 0.92 of the corresponding odds for a woman, although we know from previous analyses that the overall odds ratio, ignoring the GHQ score, does not differ significantly from 1. We might ask: why the difference? The reason is that the overall odds ratio is dominated by the large number of cases for the lower GHQ scores.

TABLE 6.3

Results from Fitting a Logistic Regression Model to the GHQ Data with Only a Single Explanatory Variable, Gender

	Estimate	Standard Error	z-Value	$\text{Pr}(> z)$
Intercept	-1.11400	0.17575	-6.338	2.32e-10
sexM	-0.03657	0.28905	-0.127	0.9

Note: Null deviance: 130.31 on 21 DF; residual deviance: 130.29 on 20 DF;
AIC: 170.26.

TABLE 6.4

Results from Fitting a Logistic Regression Model to the GHQ Data with Gender and GHQ Scores as Explanatory Variables

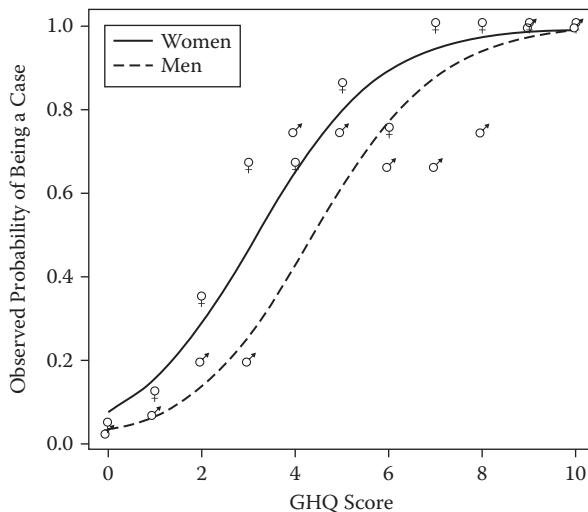
	Estimate	Standard Error	z-Value	Pr(> z)
Intercept	-2.49351	0.28164	-8.854	<2e-16
sexM	-0.93609	0.43435	-2.155	0.0311
GHQ	0.77910	0.09903	7.867	3.63e-15

Note: Null deviance: 130.306 on 21 DF; residual deviance: 11.113 on 19 DF; AIC: 53.087.

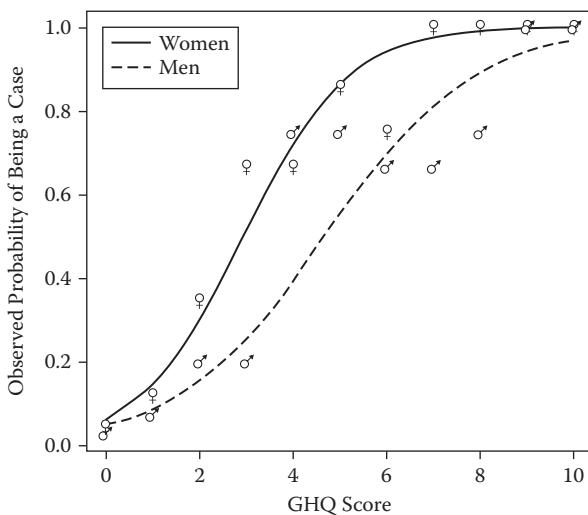
The estimated regression coefficient for GHQ conditional on gender is 0.779. This is very similar to the value for the model having only the GHQ score, and so the interpretation of the conditional coefficient is very similar to that given previously.

The AIC value of 53.1 is lower than that for the models with only the GHQ score or gender, so the model with both is considered a better model. The fitted model is displayed graphically in Figure 6.2.

Finally, for the GHQ data we will fit a model with gender, GHQ score, and gender \times GHQ score interaction. The result is shown in Table 6.5. The model is illustrated graphically in Figure 6.3. Although the AIC value is a little lower than for the previously fitted model with no interaction term, the

**FIGURE 6.2**

Plot of predicted probabilities of caseness from logistic regression model with gender and GHQ score as explanatory variables and observed probabilities labeled by gender.

**FIGURE 6.3**

Plot of predicted probabilities of caseness from logistic regression model with gender, GHQ score, gender \times GHQ as explanatory variables, and observed probabilities labeled by gender.

TABLE 6.5

Results from Fitting a Logistic Regression Model to the GHQ Data with Gender, GHQ Score, and the Interaction of Gender and GHQ Score as Explanatory Variables

	Estimate	Standard Error	z-Value	Pr(> z)
Intercept	-2.7732	0.3586	-7.732	1.06e-14
sexM	-0.2253	0.6093	-0.370	0.712
GHQ	0.9412	0.1569	6.000	1.97e-09
sexM:GHQ	-0.3020	0.1990	-1.517	0.129

Note: Null deviance: 130.306 on 21 DF; residual deviance: 8.767 on 18 DF; AIC: 52.741.

regression coefficient for the interaction term in Table 6.5 is not significant, and a reasonable conclusion is that the model with only gender and GHQ score provides the best description of these data.

6.5 Selecting the Most Parsimonious Logistic Regression Model

As with the fitting of multiple linear regression models, the aim when fitting logistic regression models is to select the most parsimonious model that gives an adequate description of the data. We can use exactly the

TABLE 6.6
Do-It-Yourself Data

Work	Tenure	Response	Accommodation					
			Apartment Age			House Age		
			<30	31–45	46+	<30	31–45	46+
Skilled	Rent	Yes	18	15	6	34	10	2
		No	15	13	9	28	4	6
	Own	Yes	5	3	1	56	56	35
		No	1	1	1	12	21	8

same approach as given in Chapter 4, as we shall illustrate on data collected from a survey of employed men aged between 18 and 67 years who were asked whether in the preceding year they had carried out work on their home for which they would have employed a craftsman previously. The response variable is their answer, yes/no, to this question. In addition to age, the respondents' accommodation type, apartment or house; whether this accommodation was rented or owned; and their type of work, skilled, unskilled, or office were recorded. Part of the data is shown in Table 6.6.

The first point to consider about these data is how to deal with the categorical variables, work and age, that have more than two categories. We could simply label the categories for work one, two, three, and likewise the age categories, and use these numerical values in the model-fitting process. But this would be a mistake, particularly for the work variable; such coding would imply that changing, say, from work category one to work category two, and from work category two to work category three, has an equal effect on the probability of responding yes in the survey, which is not necessarily the case. The appropriate method is to use dummy variable coding for both work and age. So, for the work variables, we define two dummy variables D1 and D2 to label the three categories such that

Work	D1	D2
Skilled	0	0
Unskilled	1	0
Office	0	1

So, in the fitted logistic regression model, the estimated regression coefficient for D1 will quantify the difference between unskilled and skilled workers, and D2 will quantify the difference between office and skilled workers. A similar coding will be used for the age variable. We will now fit a logistic model for the probability of responding yes in the survey with work, tenure,

TABLE 6.7

Results from Fitting a Logistic Regression Model to the Do-It-Yourself Data with Explanatory Variables Work, Tenure, Type, and Age

	Estimate	Standard Error	z-Value	Pr(> z)
Intercept	0.30606	0.15428	1.984	0.0473
Work (D1)	-0.76267	0.15197	-5.018	5.21e-07
Work (D2)	-0.30535	0.14088	-2.167	0.0302
Type	-0.00249	0.14717	-0.017	0.9865
Tenure	1.01570	0.13787	7.367	1.74e-13
Age (D1)	-0.11304	0.13697	-0.825	0.4092
Age (D2)	-0.43661	0.14059	-3.106	0.0019

Note: Null deviance: 158.884 on 35 DF; residual deviance: 29.671 on 29 DF;
AIC: 167.87.

accommodation type, and age as explanatory variables. The results are shown in Table 6.7. We will wait to interpret the estimated regression coefficients until we have explored whether a simpler model might be adequate for these data. This we will do by using a backward search procedure using the AIC criterion to guide the search, that is, the same method was used in Chapter 4 for multiple linear regression models.

The results of the backward search are as follows:

Start: AIC = 167.87

Explanatory variables in the model: work, tenure, type, age.

Step 1: Remove one explanatory variable at a time and leave the other three in the model.

Remove type: AIC = 165.87.

Remove age: AIC = 174.76.

Remove work: AIC = 191.72.

Remove tenure: AIC = 221.80.

The variable type can be removed because the model without type but with the other three explanatory variables has a lower AIC value than the four variable model.

Current AIC = 165.87.

Explanatory variables currently in the model: work, tenure, age.

Step 2: Remove one explanatory variable at a time and leave the other two in the model.

Remove age: AIC = 172.81.

Remove work: AIC = 189.19.

Remove tenure: AIC = 244.98.

Removing any of the three explanatory variables currently in the model leads to an AIC value greater than that for the current model and so we accept the current model that has work, age, and tenure as explanatory variables. Fitting this model gives the results shown in Table 6.8. In Table 6.9 we list the estimated odds ratios and 95% CIs for each variable. The results show the following:

- The odds of unskilled workers responding yes to the questions asked are between about 35% and 63% of the odds of skilled workers.
- The odds of office workers responding yes to the question asked are between about 56% and 97% of the odds of skilled workers.
- The odds of home owners responding yes to the question asked are between about twice to three-and-a-half times the odds of non-home owners.

TABLE 6.8

Results from Fitting the Model with Work, Age, and Tenure as Explanatory Variables

	Estimate	Standard Error	z-Value	Pr(> z)
Intercept	0.3048	0.1347	2.262	0.02370
Work (D1)	-0.7627	0.1520	-5.019	5.21e-07
Work (D2)	-0.3053	0.1408	-2.168	0.03012
Tenure	1.0144	0.1144	8.866	<2e-16
Age (D1)	-0.1129	0.1367	-0.826	0.40877
Age (D2)	-0.4364	0.1401	-3.116	0.00183

Note: Null deviance: 158.884 on 35 DF; residual deviance: 29.671 on 30 DF; AIC: 165.87.

TABLE 6.9

Estimated Odds Ratios and CIs of Each Explanatory Variable in the Final Model Chosen for the Do-It-Yourself Data

Variable	Estimated Odds Ratio	95% CI
Work (D1)	0.466	[0.346,0.628]
Work (D2)	0.737	[0.559,0.972]
Tenure	2.757	[2.205,3.447]
Age (D1)	0.893	[0.683,1.168]
Age (D2)	0.647	[0.491,0.851]

- The odds of respondents in the age range 31–45 responding yes to the question asked does not differ from the odds of those respondents less than 30.
 - The odds of respondents aged over 46 responding yes to the question asked are between about 50% and 85% of the odds of respondents less than 30.
-

6.6 Summary

- The logistic regression model can be used to assess the effects of a set of explanatory variables on a binary response variable.
 - The estimated parameters in the model can be interpreted in terms of odds and odds ratios.
 - Parsimonious models can be selected by the same approach as for the multiple linear regression model.
 - There are a number of diagnostics available for logistic regression that can be used to assess various aspects of the model. Details are available in Collett (2003a), however, the binary nature of the response variable often makes the use of these diagnostics somewhat difficult to interpret.
 - Logistic regression is a member of the generalized linear models family as are analysis of variance (ANOVA), analysis of covariance, and multiple linear regression. Several other types of regression that may be useful in behavioral research are subsumed in the general model, and interested readers are referred to McCullagh and Nelder (1989) for full details.
-

6.7 Exercises

- 6.1 The data in `exer_61.txt` arise from a survey carried out in 1974/1975 in which each respondent was asked if he or she agreed or disagreed with the statement “Women should take care of running their homes and leave running the country up to men.” The years of education of each respondent was also recorded. Use logistic regression to assess the effects of gender and years of education on the probability of agreeing with the statement, and construct suitable graphics for illustrating the models you fit. What are your conclusions?

- 6.2 Return to the data on do-it-yourself used in the text and use a backward search approach to assess models that allow interactions between each pair of explanatory variables. What conclusion do you reach, and how do these differ from those given in the text?
- 6.3 The data in exer_63.txt were obtained from a study of the relationship between car size and car accident injuries. Accidents were classified according to their type, severity, and whether or not the driver was ejected. Using severity as the response variable, derive and interpret a suitable logistic model for these accounts.
- 6.4 The data in exer_64.txt (taken from Johnson and Albert, 1999) are for 30 students in a statistics class. The response variable y indicates whether or not the student passed ($y = 1$) or failed ($y = 0$) the statistics examination at the end of the course. Also given are the student's scores on a previous mathematics test and their grade for a prerequisite probability course. Fit a logistic model to the data with mathematics test and probability course grade as explanatory variables. Cross-tabulate the predicted passes and failures with the observed passes and failures.
- 6.5 The data in exer_65.txt relate to a sample of girls in Warsaw, the response variable indicating whether or not the girl has begun menstruation and the exploratory variable age in years (measured to the month). Plot the estimated probability of menstruation as a function of age, and show the linear and logistic regression fits to the data on the plot.

7

Survival Analysis

7.1 Introduction

In many studies, particularly in medicine, the main outcome variable is the time from a well-defined time origin to the occurrence of a particular event or end point. In medical studies, the end point is frequently the death of a patient, and the resulting data are, quite literally, survival times. Behavioral research studies are, fortunately, rarely life threatening, but end points other than death may be of importance in such studies—for example, the time to relief of symptoms, to the recurrence of a specific condition, or simply, to the completion of an experimental task. Such observations should perhaps properly be referred to as time-to-event data, although the generic term survival data is commonly used even when literal survival is not the issue. Where time-to-event data are collected, the interest generally lies in assessing the effects of some explanatory variables on survival times, and it might be thought that multiple linear regression would fit the bill; but it would not because survival data require special techniques for their analysis for two main reasons:

- The distribution of survival times is very often positively skewed, and so, assuming normality for an analysis (as is done in multiple linear regression, for example) is almost always not reasonable.
- More critical than the probable normality problem, however, is the likely presence of censored observations, which occur when, at the completion of the study or experiment, the end point of interest has not been reached. When censoring occurs, all that is known about the individual's survival time is that it is larger than the time elapsed up to censoring.

An example of survival data from a behavioral study is provided by an investigation of the age at which women experience their first sexual intercourse (Schmidt et al., 1995). Data were collected from 84 women in two diagnostic groups: restricted anorexia nervosa (RAN) and normal controls (NC). Part of the data is shown in [Table 7.1](#). Some women at the time the study took place had not had intercourse; consequently, these observations are censored (to be precise, the observations are right censored).

TABLE 7.1

Age at First Sexual Intercourse for Women
in Two Diagnostic Groups

	Diagnosis	Age at First Sex	Age	Status
1	RAN	—*	30	0
2	RAN	—	24	0
3	RAN	12	18	1
4	RAN	21	42	1
5	RAN	—	19	0
70	NC	16	19	1
71	NC	19	22	1
72	NC	19	22	1
73	NC	18	21	1
74	NC	17	19	1
75	NC	19	21	1

* status 0 = censored; first sex has not yet taken place at time of interview; 1 = age at first sex, occurred at age given.

When the response variable of interest is a survival time, the censoring problem requires special regression techniques for modeling the relationship of the survival time to the explanatory variables of interest. A number of procedures are available, but the most widely used by some margin is the one known as Cox's proportional hazards model, or Cox regression for short. Introduced by Sir David Cox in 1972 (Cox, 1972), the method has become one of the most commonly used in medical statistics, and the original paper one of the most heavily cited. But, before discussing Cox regression, we must give an account of two approaches most often used to characterize and describe survival times, namely, the survival function and the hazard function.

7.2 The Survival Function

As with all data, an initial step in the analysis of a set of survival data is to construct an informative graphic for the data. However, the presence of censored observations makes the use of conventional descriptive graphics such as boxplots rather problematic, and survival data containing censored observations are best displayed graphically using an estimate of what is known as the data's survival function. This function and how it is estimated from sample data is described in Technical Section 7.1.

Technical Section 7.1: The Survival Function

The survivor function $S(t)$ is defined as the probability that the survival time T is greater than or equal to t , that is,

$$S(t) = \Pr(T > t)$$

A plot of an estimate of $S(t)$ against t is often a useful way of describing the survival experience of a group of individuals. When there are no censored observations in the sample of survival times, a nonparametric survivor function can be estimated simply as

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times } \geq t}{\text{Number of individuals in the data set}}$$

As every subject is “alive” at the beginning of the study and no one is observed to “survive” longer than the largest of the observed survival times t_{\max} , then

$$\hat{S}(0) = 1 \text{ and } \hat{S}(t) = 0 \text{ for } t > t_{\max}$$

The estimated survivor function is assumed to be constant between two adjacent times of death so that a plot of $\hat{S}(t)$ against t is a step function that decreases immediately after each “death.” Because the estimator above is simply a proportion, confidence intervals (CIs) can be obtained for each time t by using the usual estimate of the variance of a proportion to give

$$\text{var}[\hat{S}(t)] = \frac{\hat{S}(t)[1 - \hat{S}(t)]}{n}$$

The simple proportion estimator cannot be used to estimate the survivor function when the data contains censored observations. In the presence of censoring, the survivor function is generally estimated using the Kaplan–Meier estimator which is based on the calculation and use of conditional probabilities. Assume again we have a sample on n observations from the population for which we wish to estimate the survivor function. Some of the survival times in the sample are right censored—for those individuals we know only that their true survival times exceed the censoring time. We denote by $t_1 < t_2 \dots$ the times when deaths are observed and let d_j be the number of individuals who “die” at t_j . The Kaplan–Meier estimator for the survival function then takes the form

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j} \right)$$

where r_j is the number of individuals at risk, that is, alive and not censored, just prior to time t_j . If there are no censored observations, the estimator reduces to the simple proportion given earlier. The essence of the

Kaplan–Meier estimator is the use of the continued product of a series of conditional probabilities. For example, to find the probability of surviving, say, 2 years, we use $\text{Pr}(\text{surviving 2 years}) = \text{Pr}(\text{surviving 1 year}) \times \text{Pr}(\text{surviving 2 years} | \text{having survived 1 year})$, and the probability of surviving 3 years is found from $\text{Pr}(3) = \text{Pr}(3|2) \times \text{Pr}(2) = \text{Pr}(3|2) \times \text{Pr}(2|1) \times \text{Pr}(1)$. In this way, censored observations can be accommodated correctly.

The variance of the Kaplan–Meier estimator is given by

$$V[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}$$

When there is no censoring, this reduces to the standard variance of a proportion used earlier for the simple proportion estimator of the survival function.

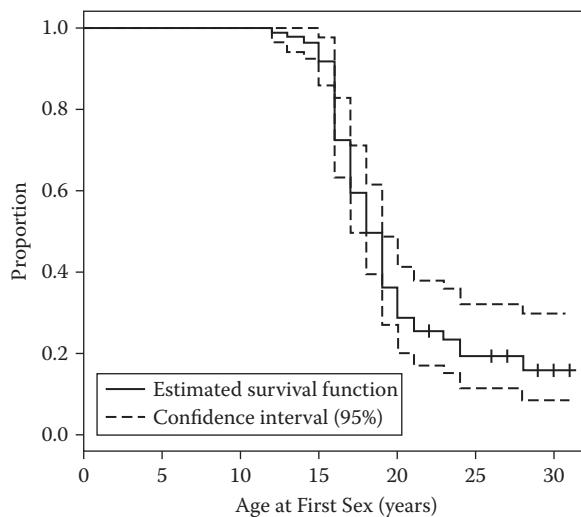
Let us begin by constructing and plotting the Kaplan–Meier estimated survival function of all the ages at first sexual intercourse, ignoring for the moment the two diagnostic groups. Table 7.2 shows the numerical results, and [Figure 7.1](#) is a plot of the estimated survival function against age at first sex. Also shown in [Figure 7.1](#) are the upper and lower 95% CI bounds for the survival function. Table 7.2 shows the number at risk at the time of each “event” (age at first sex in this example), the estimated survival function at this time, and the 95% CI. The plot in [Figure 7.1](#) summarizes the numerical information in Table 7.2. From this plot, we can read off the median event time (the median is the preferred measure of location for survival time data because of their likely skewness) as 18 with 95% CI of about [17,19].

TABLE 7.2

Kaplan–Meier Estimate of the Survival Function of all the Ages at First Sex Data

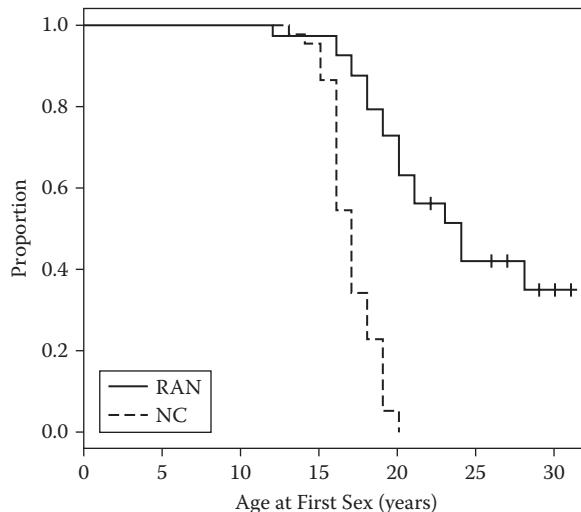
Time	N.risk	N.event	Survival	Standard Error	Lower 95% CI	Upper 95% CI
12	84	1	0.988	0.0118	0.9652	1.000
13	83	1	0.976	0.0166	0.9441	1.000
14	82	1	0.964	0.0202	0.9254	1.000
15	81	4	0.917	0.0302	0.8594	0.978
16	77	16	0.726	0.0487	0.6368	0.828
17	61	11	0.595	0.0536	0.4990	0.710
18	47	8	0.494	0.0551	0.3969	0.615
19	34	9	0.363	0.0551	0.2697	0.489
20	24	5	0.288	0.0530	0.2003	0.413
21	18	2	0.256	0.0517	0.1719	0.380
23	12	1	0.234	0.0516	0.1521	0.361
24	11	2	0.192	0.0503	0.1147	0.320
28	6	1	0.160	0.0510	0.0854	0.299

Note: N. risk = Number of subjects at risk; N. event = Number of events; Survival = Proportion “survived”

**FIGURE 7.1**

Plot of estimated survival function for age at first sex data, along with the 95% confidence bands for the survival function.

We can now estimate and plot the survival curves for age at first sexual intercourse in each diagnostic group for the data, again using the Kaplan-Meier estimator. The resulting plot is shown in Figure 7.2. This plot suggests very strongly that the time to first sexual intercourse in women

**FIGURE 7.2**

Estimated survival functions for age at first sex for two diagnostic groups.

diagnosed as RAN is later than in NC. For those not convinced by the plot, a formal test is available to test the hypothesis that the time to first intercourse is the same in each diagnostic group; the required test is known as a log-rank test, which assesses the null hypothesis that the survival functions of the two groups are the same. Essentially, this test compares the observed number of “deaths” occurring at each time point with the number to be expected if the survival experience of the groups being compared is the same. Details of the test are given in Collett (2003b); here, we shall simply give the value of the chi-squared test statistic that arises from the log-rank test, namely, 46.4 with one degree of freedom and associated p-value <0.00001. There is very convincing evidence that the age at first sex differs in the two diagnostic groups.

7.3 The Hazard Function

In the analysis of survival data, it is often of interest to assess which periods have high or low chances of death (or whatever the event of interest maybe) among those still active at the time. A suitable approach to characterizing such risks is the hazard function $h(t)$, which is described in Technical Section 7.2. (The hazard function is also known as the intensity function, instantaneous failure rate, and age-specific failure rate.)

Technical Section 7.2: Hazard Function

The hazard function $h(t)$ is defined as the probability that an individual experiences the event of interest in a small time interval s , given that the individual has survived up to the beginning of the interval, when the size of the time interval approaches zero. Mathematically, this is written as

$$h(t) = \lim_{s \rightarrow 0} \Pr(t \leq T \leq t+s | T \geq t)$$

where T is the individual’s survival time. The conditioning feature of this definition is very important. For example, the probability of dying at age 100 is very small because most people die before that age; in contrast, the probability of a person dying at age 100 for one who has reached that age is much greater. In practice, the hazard function may increase, decrease, remain constant, or have a more complex shape. The hazard function for death in human beings, for example, has the “bathtub” shape shown in Figure 7.3. It is relatively high immediately after birth, declines rapidly in the early years, and then remains approximately constant before beginning to rise again during late middle age.

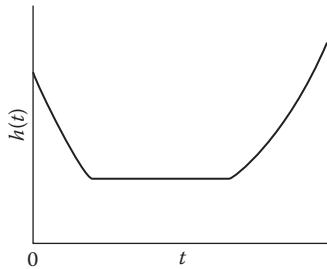


FIGURE 7.3
Bathtub hazard function.

The hazard function can also be defined in terms of the cumulative distribution $F(t)$ and the probability density function $f(t)$ of survival times as

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

It then follows that

$$h(t) = -\frac{d}{dt}[\ln S(t)]$$

and so

$$S(t) = \exp[-H(t)]$$

where $H(t)$ is the integrated or cumulative hazard function and is given by

$$H(t) = \int_0^t h(u) du$$

The cumulative hazard function gives the accumulated risk until a specific time.

The hazard function can be estimated as the proportion of individuals experiencing the event of interest in an interval per unit time, given that they have survived to the beginning of the interval. That is,

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}$$

where d_j is the number of individuals experiencing an event in the interval beginning at time t , and n_j is the number of patients surviving at time t . The sampling variation in the estimate of the hazard function

within each interval is usually considerable, and so, it is rarely plotted directly. Instead, plots of the cumulative hazard function are used because they are typically smoother and easier to interpret. Everitt and Rabe-Hesketh (2001) show that the cumulative hazard function can be estimated as $\hat{H}(t) = \sum_j \frac{d_j}{n_j}$.

The estimated survival function is more helpful than the estimated hazard function for an initial assessment of survival data, but the hazard function assumes great importance when we consider how to investigate the effect of explanatory variables on survival times, as we shall see in Section 7.4.

7.4 Cox's Proportional Hazards Model

One of the main aims of many behavioral studies is the investigation of how a number of explanatory variables of interest relate to the chosen response variable. The same is true for studies involving survival time data, but because of the special features of such data, the regression methods considered in earlier chapters are no longer appropriate. A number of models that can be applied to survival data have been developed, of which perhaps the most successful is that described by Cox (1972). In these models, it is the hazard function that serves as the response variable because this is a simpler vehicle for modeling the joint effects of explanatory variables as it does not involve the cumulative history of events. Here, we shall concentrate on the model first suggested by Cox.

To introduce the basic concepts of the Cox model, suppose first that there are two explanatory variables of interest x_1 and x_2 . We might begin by considering a model in which the hazard function is a linear function of the two explanatory variables. But, it is easy to see that such a model is unsuitable because the hazard function is restricted to being positive, whereas the postulated linear function need not necessarily be likewise constrained (compare Chapter 6 and the discussion of logistic regression). A more sensible model is one that models the log of the hazard function as a linear function of the explanatory variables, and it is a description of this model that begins Technical Section 7.3.

Technical Section 7.3: Cox Regression

A possible model for the log of the hazard function when there are q explanatory variables is

$$\log[h(t)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

In this model the hazard function does not depend on time, which can be shown to imply that the survival times have an *exponential distribution*.

Such a model is very restrictive because hazard functions that increase or decrease with time are much more likely; a suitable model is

$$\log[h(t)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \alpha t$$

Depending on the sign of this model, it can represent both a hazard function that increases with time and one that decreases with time. But, what if the hazard function has some more complicated form, for example, the bathtub shape illustrated in [Figure 7.3](#)? In most practical situations, it will be very difficult to determine the correct function of time to include in the model, but this problem is overcome in Cox regression where the dependence of the hazard function on time does not have to be specified explicitly. In Cox regression, the model is

$$\log h(t) = \log h_0(t) + \beta_1 x_1 + \cdots + \beta_q x_q$$

where $h_0(t)$ is known as the baseline hazard function, being the hazard function for individuals with all explanatory variables equal to zero (this makes more sense if you suppose that each explanatory variable is centered at its sample mean). The model can be rewritten as

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)$$

Written in this way, it is possible to show that the Cox model forces the hazard ratio between two individuals with vectors of covariate values \mathbf{x}_1 and \mathbf{x}_2 to be constant over time:

$$\frac{h(t | \mathbf{x}_1)}{h(t | \mathbf{x}_2)} = \frac{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_1)}{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_2)} = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_1)}{\exp(\boldsymbol{\beta}' \mathbf{x}_2)}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. So, we see that the Cox model implies that if, at some early time point, say, an individual has a risk of “death” that is twice as high as another individual, then at all later times, the risk of death remains twice as high. Hence, Cox regression is often labeled *Cox's proportional hazards model*. While the proportionality assumption may sound complicated, it is actually very simple and implies that the single regression coefficient associated with each explanatory variable represents the effect of this variable throughout the time period. If the risk of outcome associated with a particular variable is higher at one point in time and lower at another, a single coefficient cannot represent that relationship.

In the Cox model, the baseline hazard describes the common shape of the survival time distribution for all individuals, while the relative risk function $\exp(\boldsymbol{\beta}' \mathbf{x})$ gives the level of each individual's hazard. The interpretation of a particular element of the vector $\boldsymbol{\beta}$, say the parameter β_j , is that $\exp(\beta_j)$ gives the relative risk change associated with an increase

of one unit in x_j , all other explanatory variables remaining constant. An alternative aid to interpretation is to calculate $100(\exp(\beta_j) - 1)$ to get the percentage change in the hazard function with each unit change in the appropriate explanatory variable, conditional on the other explanatory variables remaining constant.

The parameters in a Cox model can be estimated by maximizing what is known as a *partial likelihood*. Details are given in Kalbfleisch and Prentice (1980). The partial likelihood is derived by assuming continuous survival times. In reality, however, survival times are measured in discrete units, and there are often ties. There are three common methods for dealing with ties, which are described briefly in Everitt and Rabe-Hesketh (2001).

The Cox model can be extended to allow the baseline hazard function to vary with the levels of a stratification variable. Such a *stratified proportional hazards model* is useful in situations where the stratifier is thought to affect the hazard function but is not of primary interest. A common example of a stratifier variable is gender.

As a first example of applying Cox regression, we will apply it to the data on age at first sex. The estimated regression coefficient for the single explanatory variables diagnosis coded 0 for RAN and 1 for NC is 2.14, with an estimated standard error of 0.346; dividing 2.14 by 0.346 gives a z-statistic with a value of 6.18 and an associated very, very small p-value. The log hazard for first sex is estimated to be 2.14 greater in the NC than in the RAN patients. The value 2.14 can be exponentiated to give a value of 8.47 for the relative risk; a 95% CI for the relative risk is found as $[\exp(2.14 - 1.96 \times 0.346), \exp(2.14 + 1.96 \times 0.346)]$, giving [4.3, 16.70]. The hazard function for first sex for NC is about 4 to 17 times that for the women diagnosed as RAN.

Our second example of the use of Cox regression will involve data collected on the retention of heroin addicts in maintenance treatment using methadone. The end point in this study is methadone cessation (either by the choice of the patient or the treating doctor) and departure from the treating clinic. The time in days until methadone cessation was recorded for 238 heroin addicts. Censored observations occurred when patients departed for reasons other than methadone cessation (status = 1 if methadone cessation, and 0 otherwise). So, in this example, methadone cessation is equivalent to “death” in a study in which the end point is actual death. In addition, the maximum methadone dose prescribed (milligrams per day) and whether or not the addict had a prison record were recorded (0 = no prison record, 1 = prison record). The patients were treated in one of two clinics (coded as clinic 1 = 0, clinic 2 = 1). The data for five patients from each of the two clinics are shown in [Table 7.3](#).

To begin, let us take a look at the estimated survival functions for the two clinics (see [Figure 7.4](#).) The plot demonstrates that the second clinic appears to be more successful in keeping heroin addicts on their methadone treatment. From the graph, the median methadone treatment in clinic 1 can

TABLE 7.3
Data for Heroin Addicts Being Treated with Methadone

	Clinic	Status	Time (Days)	Prison	Dose (mg/Day)
1	Clinic 1	1	428	No prison record	50
2	Clinic 1	1	275	Prison record	55
3	Clinic 1	1	262	No prison record	55
4	Clinic 1	1	183	No prison record	30
5	Clinic 1	1	259	Prison record	65
98	Clinic 2	1	708	Prison record	60
99	Clinic 2	0	713	No prison record	50
100	Clinic 2	0	146	No prison record	50
101	Clinic 2	1	450	No prison record	55
102	Clinic 2	0	555	No prison record	80

Note: Status = 1 implies methadone cessation; otherwise, status = 0.

be estimated as 428 days; the estimate is not available for clinic 2 because more than 50% of the patients continued treatment throughout the study period.

But our real interest for these data lies in assessing the effects of all the explanatory variables by using Cox regression. The results of fitting the

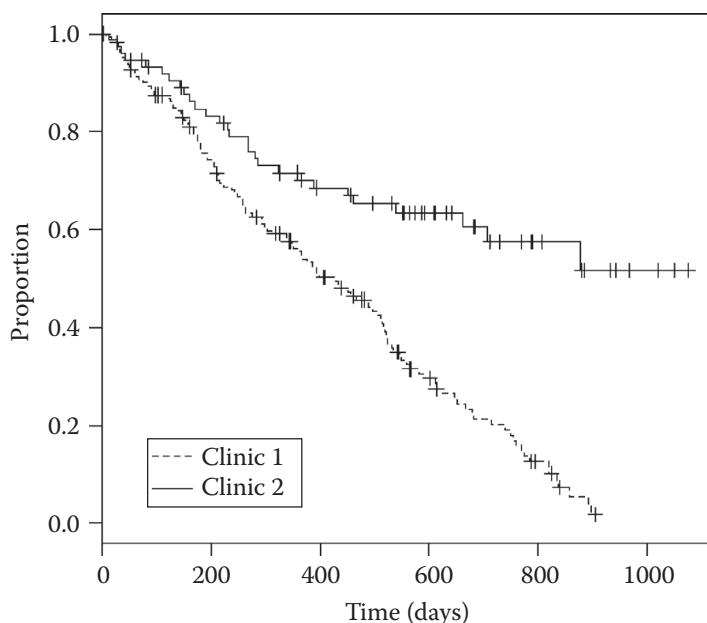


FIGURE 7.4
Estimated survival functions for the two clinics in the heroin addicts data.

TABLE 7.4

Results from Fitting the Cox Regression Model to the Data on Heroin Addicts

	Coef	Exp(coef)	Se(coef)	z	p
Prison record	0.3266	1.386	0.16722	1.95	5.1e-02
Dose	-0.0354	0.965	0.00638	-5.54	2.9e-08
Clinic	-1.0099	0.364	0.21489	-4.70	2.6e-06
	Exp(coef)	Exp(–coef)	Lower 95% CI	Upper 95% CI	
Prison record	1.386	0.721	0.999	1.924	
Dose	0.965	1.036	0.953	0.977	
Clinic	0.364	2.745	0.239	0.555	

Note: CI = Confidence interval.

model for the heroin data are shown in Table 7.4. Looking at the exponentiated results given in the lower part of Table 7.4, we see that

- Heroin addicts with a criminal record are estimated to have a hazard of immediate methadone cessation of between about 0.999 and 1.924 times the corresponding hazard for addicts without a criminal record, conditional on clinic and dose. As the CI contains the value 1, there is no strong evidence that a prison record has any real effect on the hazard function, although there may be a tendency for addicts with a criminal record to have an increased risk of immediate methadone cessation.
- Each milligram per day increase in maximum methadone dose prescribed leads to a decrease in the hazard function of immediate methadone cessation of between about 2% and 5%, conditional on prison record and clinic. Higher doses of methadone tend to keep addicts on treatment.
- In clinic 2, the hazard function of immediate methadone cessation is about 0.3 to 0.6 times the corresponding hazard function in clinic 1, conditional on dose and prison record. Clinic 2 is more successful in keeping addicts on treatment; the risk of immediate methadone cessation is higher in clinic 1.

As in multiple linear regression and logistic regression, we might often wish to find a more parsimonious Cox regression model than the one containing all the explanatory variables under investigation. We can use the same backward elimination approach as we used for both multiple linear and logistic regressions, and if we do, we get the following results:

Start: AIC = 1352.52

Step 1: Model with all three explanatory variables and one variable at a time considered for removal.

Drop prison: AIC = 1354.3

Drop clinic: AIC = 1376.9

Drop dose: AIC = 1 1381.3

We see in this example that dropping any of the three explanatory variables results in a model with larger AIC value than the model that includes all three variables. Consequently, it is this model that is accepted despite the regression coefficient for the prison variable not being significant at the 5% level.

The Cox model is based on the assumption of proportional hazards, but this assumption will not hold for all sets of survival data. For example, if two individuals with a heart condition receive different treatments, one medical and one surgical, then the individual treated surgically may be at higher risk initially because of the possibility of operative mortality. At a later stage, however, the risk may become the same or even less than that for a medically treated individual. In this case, if one of the explanatory variables is a dummy variable indicating which treatment a person receives, then the proportional hazards assumption will not hold, and the regression coefficient for treatment effect will change over time. Examining whether or not any of the regression estimates in a Cox regression change over time is one way of checking the proportional hazards assumption. Therneau and Grambsch (2000) describe a test of the hypothesis of constant regression coefficients over time that can be applied individually to each coefficient and globally to the set of coefficients. The test statistic is a chi-squared; however, we will not give details of the test here but simply look at the results, which are as follows:

	chisq	p
Prison	0.22	0.64
Dose	0.70	0.40
Clinic	11.19	<0.01
Global	12.62	<0.01

It appears that the proportional hazards function does not hold for the clinics in this example. Consequently, a better model for the data may be one in which the clinic is used as a stratifier. Consideration of such a model is left as an exercise for the reader (see Exercise 7.1).

7.5 Summary

- Survival data need specialized techniques for their analysis because of the presence of censored observations. Such data arise frequently in medical studies but less often in behavioral research. However,

behavioral researchers need to recognize such data and be aware of the correct methods of analysis.

- The first step in the analysis of survival data is, generally, to plot the estimated survival functions for particular groups of observations.
 - The hypothesis that two survival functions are the same can be tested formally using the log-rank test.
 - Cox regression is the model most often used to assess the effects of explanatory variables on survival times. The assumption of proportional hazards on which the model is based needs to be checked.
 - Diagnostics for survival analysis, in particular for checking the assumptions made by Cox regression, are described in Collett (2003).
-

7.6 Exercises

- 7.1 Fit a Cox regression model to the data on heroin addicts using the clinic as a stratifier. How do the results compare with those derived in the text?
- 7.2 The data in exer_71.txt are the survival times (in months) after mastectomy of women with breast cancer. The cancers are classified as having metastasized or not based on a histochemical marker. Censoring is indicated by the event variable that takes the value TRUE in the case of death and FALSE otherwise. Plot the Kaplan-Meier estimated survival functions for each type of cancer on the same graph and comment on the differences. Use a log-rank test to formally compare the survival experience of each class.
- 7.3 Grana et al. (2002) report the results from a nonrandomized clinical trial investigating a novel radioimmunotherapy in malignant glioma patients. The overall survival, that is, the time from the beginning of therapy to the disease-caused death of the patient, is compared for two groups of patients—one, the standard therapy plus the novel therapy; and the other, the standard therapy alone. In addition to the treatment groups a number of other explanatory variables are recorded, namely age, gender, and histology. The data are given in exer_73.txt. Fit a Cox model to the data and find a CI for the treatment effect conditional on age, gender, and histology.

8

Linear Mixed Models for Longitudinal Data

8.1 Introduction

Chapters 3 to 7 have looked at ways to model and analyze different types of *multivariable data* in which there is a single response variable and a number of explanatory variables, and only the response is considered a random variable. Chapters 9 to 11 look at ways to explore and uncover possible structure in *multivariate data* in which there may be many variables but no division into response and explanatory variables, and *all* the variables are random variables. In this chapter we look at ways of modeling a type of data that comes somewhere between the multivariable and the multivariate. As with the former, there *is* a division into explanatory and response variables, but the response variable (and possibly some of the exploratory variables) is observed more than once on each individual in the study; consequently, the response is multivariate.

In behavioral research, data with a multivariate response are common; for example, a response variable may be measured under a number of different experimental conditions, leading to what are generally called repeated measures data, or a response variable may be recorded on several different occasions over some period of time, in which case we have longitudinal data. (Although I think distinguishing two different types of data here is useful, it has to be said that the repeated measures label is often used for both types of data.) The variation among the repeated measures of the response is *within-subject variation*. But often one of the covariates will be a factor such as gender or treatment that will give rise to *between-subject variation*.

As several observations of the response variable are made on the same individual, it is likely that the repeated measurements of the response will be correlated rather than independent even after conditioning on the explanatory variables. Consequently, for the analysis of repeated measures data and for longitudinal data, models are needed that can both assess the effects of explanatory variables on the multiple measures of the response variable and account for the likely correlations between these multiple measures. Linear mixed-effects models seek to deal with both requirements as we shall see in [Section 8.2](#) where such models are described in the context of analyzing longitudinal data.

8.2 Linear Mixed Effects Models for Longitudinal Data

The main objective in the analysis of data from a longitudinal study is to characterize the change in the repeated values of the response variable and to determine the explanatory variables most associated with any change. The distinguishing feature of a repeated-measures study is that the response variable of interest is measured a number of times on each individual in the study; some explanatory variables may also be recorded more than once, with others being recorded only at the beginning of the study. The repeated measurements of the response variable are very likely to be correlated rather than independent, so that models for such data need to include parameters linking the explanatory variables to the repeated measurements, parameters analogous to the regression coefficients in the usual multiple regression model (see Chapter 4), and, in addition, parameters that account for the correlational structure of the repeated measurements. It is the former parameters that are generally of most interest, with the latter often being regarded as nuisance parameters. However, providing an adequate model for the correlation between the repeated measures is usually necessary to avoid misleading inferences about those parameters that are of primary relevance to the researcher's questions of most interest about the data.

Linear mixed-effects models for repeated-measures data formalize the sensible idea that an individual's pattern of responses is likely to depend on many characteristics of that individual, including some that are unobserved. These unobserved variables are then included in the model as random variables, that is, random effects. The essential feature of the model is that correlation among the repeated measurements on the same unit arises from shared, unobserved variables. Conditional on the values of the random effects, the repeated measurements are assumed to be independent—the conditional independence assumption.

Linear mixed-effects models are introduced in Technical Section 8.1 by looking at two relatively simple examples of such models, namely, the random intercept, and random intercept and slope models. (In this chapter and in this book we only deal with repeated-measures data sets for which it can be assumed that, conditional on the explanatory variables, the response variable has a normal distribution or is at least approximately normally distributed; for details of the analysis of nonnormal longitudinal data, for example, where a binary response variable is observed on several occasions, see Rabe-Hesketh and Skrondal, 2008.)

Technical Section 8.1: Introducing Linear Mixed Effects Models

Consider a simple set of longitudinal data in which a number of individuals each has values of a response variable recorded at times t_1, t_2, \dots, t_r . (We assume the same set of time points for each individual to make the

description of linear mixed-effect models simpler, but longitudinal data in which each individual is observed at a different set of time points present no problems for such models.)

Let y_{ij} represent the value of the response for individual i at time t_j , with $j = 1, 2, \dots, r$ and $i = 1, 2, \dots, n$. If the repeated measurements of the response variable are independent of one another, then the fact that sets of r observations come from the same individual could be ignored, and the data might be described by a simple linear regression model of the form

$$y_{ij} = \beta_0 + \beta_1 t_j + \varepsilon_{ij}$$

But for repeated-measures data, independence is very unlikely, so this model is not appropriate. A possible model for the y_{ij} that does not assume independence is

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + \varepsilon_{ij}$$

which it is often helpful to write as

$$y_{ij} = (\beta_0 + u_i) + \beta_1 t_j + \varepsilon_{ij}$$

for reasons that will (hopefully) become clear later.

In the model above, the total residual that is present in the linear regression model has been partitioned into a subject-specific random component u_i , which is constant over time, plus a residual ε_{ij} , which varies randomly over time. The u_i is assumed to be normally distributed with 0 mean and variance σ_u^2 . Similarly, the ε_{ij} is, as always, assumed normally distributed with 0 mean and variance σ^2 . The u_i and ε_{ij} are assumed to be independent of each other and of t_j . This model is known as a random intercept model, u_i being the random intercepts. The repeated measurements made over time for an individual vary about that individual's own regression line, which differs in intercept but not in slope from the regression lines of other individuals. The random effects u_i model possible heterogeneity in the intercepts of the individuals' regression lines.

Let us now look at how the presence of the random effects introduces covariance between the repeated measurements over time. First, the variance of each repeated measurement implied by the random intercept model is given by

$$\text{Var}(u_i + \varepsilon_{ij}) = \sigma_u^2 + \sigma^2$$

Due to this decomposition of the total residual variance into a between-subject component σ_u^2 and a within subject component σ^2 , the model is sometimes referred to as a variance component model. The covariance between the total residuals at two time points t_j and $t_{j'}$ in the same individual i is given by

$$\text{Cov}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) = \sigma_u^2$$

Note that these covariances are induced by the shared random intercept; for individuals with $u_i > 0$, the total residuals will tend to be greater than the mean; for individuals with $u_i < 0$, they will tend to be less than the mean. It follows from the formulae for the variance and covariance that the correlations between residuals are given by

$$\text{Cor}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

This is an *intraclass correlation* interpreted as the proportion of the total residual variance due to residual variability between subjects.

The formulae given earlier for the variance of each repeated measurement and for the covariance of each pair of repeated measurements do not involve time, and demonstrate that the random intercept model constrains the variance of each repeated measurement to be the same and the correlation of each pair of measurements to be equal. This particular correlational structure is known as *compound symmetry*. Fitting a random intercept model to longitudinal data implies that the compound symmetry structure is considered appropriate for the data. However, this is very often not the case; for example, it is more common for measures taken closer to each other in time to be more highly correlated than those taken further apart. In addition, the variances of the measurements taken later in time are often greater than those measurements taken earlier. Consequently, for many longitudinal data sets, the random intercept model will not do justice to the observed pattern of variances and correlations between the repeated measurements, and will, therefore, not be the most appropriate model for the data. A model that allows a more realistic structure for covariances is one that allows heterogeneity in both intercepts and slopes, namely, the random slope and intercept model. In this model there are two types of random effects: the first modeling heterogeneity in intercepts, u_{i1} , and the second modeling heterogeneity in slopes, u_{i2} . Explicitly, the model is

$$y_{ij} = \beta_0 + \beta_1 t_j + u_{i1} + u_{i2} t_j + \varepsilon_{ij}$$

which may also be written as

$$y_{ij} = (\beta_0 + u_{i1}) + (\beta_1 + u_{i2})t_j + \varepsilon_{ij}$$

so as to show more clearly how the two subject random effects alter the intercept and the slope in the model. The two random effects are assumed to have a bivariate normal distribution with 0 means for both variables, variances $\sigma_{u_1}^2, \sigma_{u_2}^2$, and covariance $\sigma_{u_1 u_2}$. With this model, the total residual is $u_{i1} + u_{i2}t_j + \varepsilon_{ij}$ with variance

$$\text{Var}(u_{i1} + u_{i2}x_j + \varepsilon_{ij}) = \sigma_{u_1}^2 + 2\sigma_{u_1 u_2} t_j + \sigma_{u_2}^2 x_j^2 + \sigma^2$$

which is now no longer constant for different values of t_j . Similarly, the covariance between two total residuals of the same individual is

$$\text{Cov}(u_{i1} + u_{i2}t_j + \varepsilon_{ij}, u_{i1} + u_{i2}t_{j'} + \varepsilon_{ij'}) = \sigma_{u_1}^2 + \sigma_{u_1 u_2} (t_j + t_{j'}) + \sigma_{u_2}^2 t_j t_{j'}$$

and this is now not constrained to be the same for all pairs j and j' . The random intercept and slope model allows for both variances of the repeated measurements that change with time, and covariances of pairs of repeated measurements that are not all the same.

In the model we have been considering, time has a fixed effect measured by the parameter β_1 . It is this parameter that is likely to be of more interest to the investigator than the other parameters in the model, namely, the variance of the error terms and the variance (and possibly covariance) of the random effects. However, if the estimate of β_1 and its estimated standard error are derived from the simple regression model assuming independence, the standard error will be larger than it should be because of ignoring the likely within-subject dependences that will reduce the error variance in the model. Consequently, use of the simple regression model may give a misleading inference for β_1 . As we shall see later, the situation for between-subject fixed effects is the reverse of that for within-subject fixed effects, with the estimated standard error of the effect being smaller in the (usually) inappropriate independence model than in a linear mixed-effect model.

Linear mixed-effects models can be estimated by maximum likelihood (ML). However, this method tends to underestimate the variance components. A modified version of ML, known as restricted maximum likelihood (REML), is therefore often recommended. Details are given in Diggle et al. (2002) and Longford (1993). Competing linear mixed-effects models for a data set, for example, a random intercept model and a

random intercept and slope model, can be compared using a likelihood ratio test (although see later comments about this test). The distinction between ML and REML is relevant when using the likelihood ratio test to compare two nested models because, unlike ML, which places no restrictions on likelihood ratio tests involving fixed and random effects, when REML is used, such tests are only appropriate when both models have the same set of fixed effects (see Longford, 1993).

8.3 How Do Rats Grow?

To begin to see how linear mixed-effects models are applied in practice, we shall use some data from a nutrition study conducted in three groups of rats (Crowder and Hand, 1990). The three groups were put on different diets, and each animal's body weight (grams) was recorded repeatedly (approximately weekly, except in week seven when two recordings were taken) over a 9-week period. The question of most interest is whether the growth profiles of the three groups differ. The data are shown in Table 8.1.

TABLE 8.1
Body Weights of Rats Recorded Over a 9-Week Period

ID	Group	Day										
		1	8	15	22	29	36	43	44	50	57	64
1	1	240	250	255	260	262	258	266	266	265	272	278
2	1	225	230	230	232	240	240	243	244	238	247	245
3	1	245	250	250	255	262	265	267	267	264	268	269
4	1	260	255	255	265	265	268	270	272	274	273	275
5	1	255	260	255	270	270	273	274	273	276	278	280
6	1	260	265	270	275	275	277	278	278	284	279	281
7	1	275	275	260	270	273	274	276	271	282	281	284
8	1	245	255	260	268	270	265	265	267	273	274	278
9	2	410	415	425	428	438	443	442	446	456	468	478
10	2	405	420	430	440	448	460	458	464	475	484	496
11	2	445	445	450	452	455	455	451	450	462	466	472
12	2	555	560	565	580	590	597	595	595	612	618	628
13	3	470	465	475	485	487	493	493	504	507	518	525
14	3	535	525	530	533	535	540	525	530	543	544	559
15	3	520	525	530	540	543	546	538	544	553	555	548
16	3	510	510	520	515	530	538	535	542	550	553	569

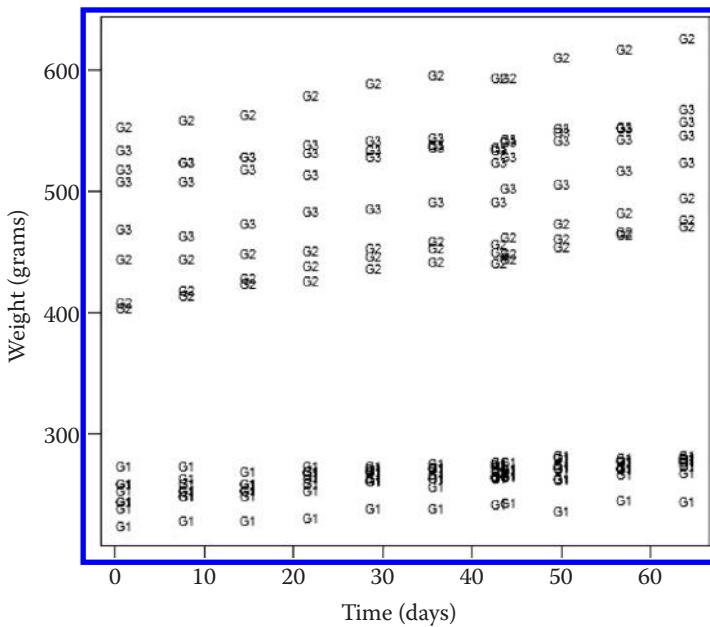
8.3.1 Fitting the Independence Model to the Rat Data

To begin, we shall ignore the repeated-measures structure of the data and assume that all the observations are independent of one another. It is easier to imagine this if we write the data in what is known as the long form (in [Table 8.1](#), the data are their wide form); the data for the first two rats in group 1 are shown in their long form in [Table 8.2](#). Now if we simply ignore that the sets of 11 weights come from the same rat, we have a data set consisting of 176 weights, times, and group memberships that we see can easily be analyzed using multiple linear regression. To begin, we will plot the data, identifying the observations in each group but ignoring the longitudinal nature of the data to give [Figure 8.1](#). Clearly, there is a difference between the weights of the group 1 rats and those in the other two groups. Continuing to ignore the repeated-measures structure of the data, we might fit a multiple linear regression model with weight as response and time and group (coded as two dummy variables D_1 and D_2 , with both D_1 and D_2 taking the value 0 for rats in group 1, D_1 being 1 and D_2 being 0 for rats in group 2, and D_1 being 0 and D_2 being 1 for rats in group 3) as explanatory variables. Fitting the model gives the results shown in [Table 8.3](#). As we might have anticipated from [Figure 8.1](#),

TABLE 8.2

Long form of Data for the First Two Rats in Group 1 in [Table 8.1](#)

ID	Group	Time	Weight
1	G1	1	240
2	G1	8	250
3	G1	15	255
4	G1	22	260
5	G1	29	262
6	G1	36	258
7	G1	43	266
8	G1	44	266
9	G1	50	265
10	G1	57	272
11	G1	64	278
12	G1	1	225
13	G1	8	230
14	G1	15	230
15	G1	22	232
16	G1	29	240
17	G1	36	240
18	G1	43	243
19	G1	44	244
20	G1	50	238
21	G1	57	247
22	G1	64	245

**FIGURE 8.1**

Plot of weight against time for rat data, ignoring the repeated-measures structure of the data but identifying the group to which each observation belongs.

both group 2 and group 3 differ significantly from group 1 conditional on time; the regression on time is also highly significant. We might go on to fit a model with a group \times time interaction, but we will not do this because we know from the structure of the data that the model considered here is wrong. The model assumes independence of the repeated measures of weight, and this assumption is highly unlikely. So, now we will move on to consider both some more appropriate graphics and appropriate models.

TABLE 8.3

Results from Fitting a Linear Regression Model to Rat Data with Weight as Response Variable, and Group and Time as Explanatory Variables, and Ignoring the Repeated-Measures Structure of the Data

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	244.0689	5.7725	42.281	<2e-16
Time	0.5857	0.1331	4.402	1.88e-05
D_1	220.9886	6.3402	34.855	<2e-16
D_2	262.0795	6.3402	41.336	<2e-16

Note: Multiple R-squared: 0.9283; F-statistic: 742.6 on 3 and 172 DF; p-value: <2.2e-16.

8.3.2 Fitting Linear Mixed Models to the Rat Data

We begin with a graphical display of the rat growth data that takes into account the longitudinal structure of the data by joining together the points belonging to each rat to show the weight growth profiles of individual rats; the plot appears in Figure 8.2. In Figure 8.3, a scatterplot matrix of the repeated measures of weight, although not a terribly helpful graphic, does demonstrate that the repeated measures are certainly not independent of one another.

To begin the more formal analysis of the rat growth data, we will first fit the random intercept model and include the two explanatory variables: time and group (coded as two dummy variables). If we represent the weight of the i th rat at time t_j by y_{ij} , the model can be written as

$$y_{ij} = (\beta_0 + u_i) + \beta_1 t_j + \beta_2 D_{i1} + \beta_3 D_{i2} + \varepsilon_{ijk}$$

where u_i is the random effect specific to the i th subject, with these random effects having a normal distribution with mean 0 and variance σ_u^2 ; the ε_{ijk} are the usual “error” terms with a normal distribution with mean 0 and variance σ^2 , and D_{i1} and D_{i2} are the same two dummy variables used to code the group membership of the i th rat as in the independence model fitted earlier. This model allows the linear regression fit for each rat to differ in intercept

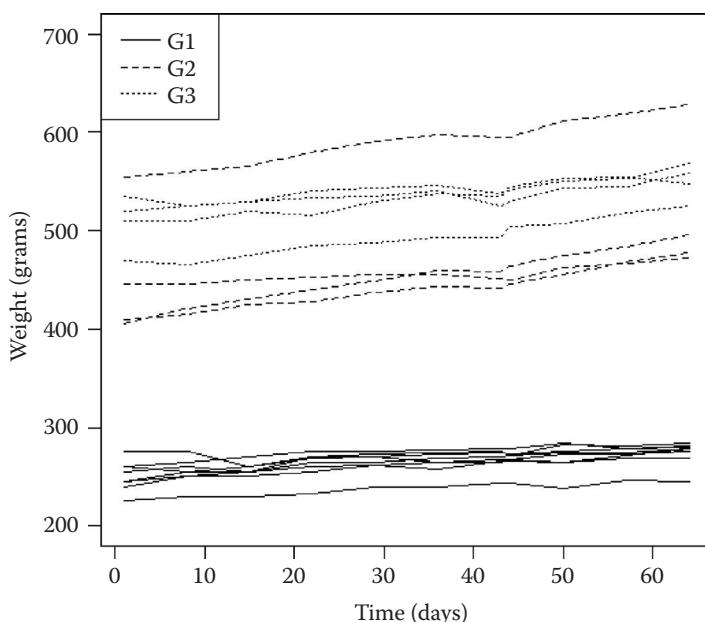
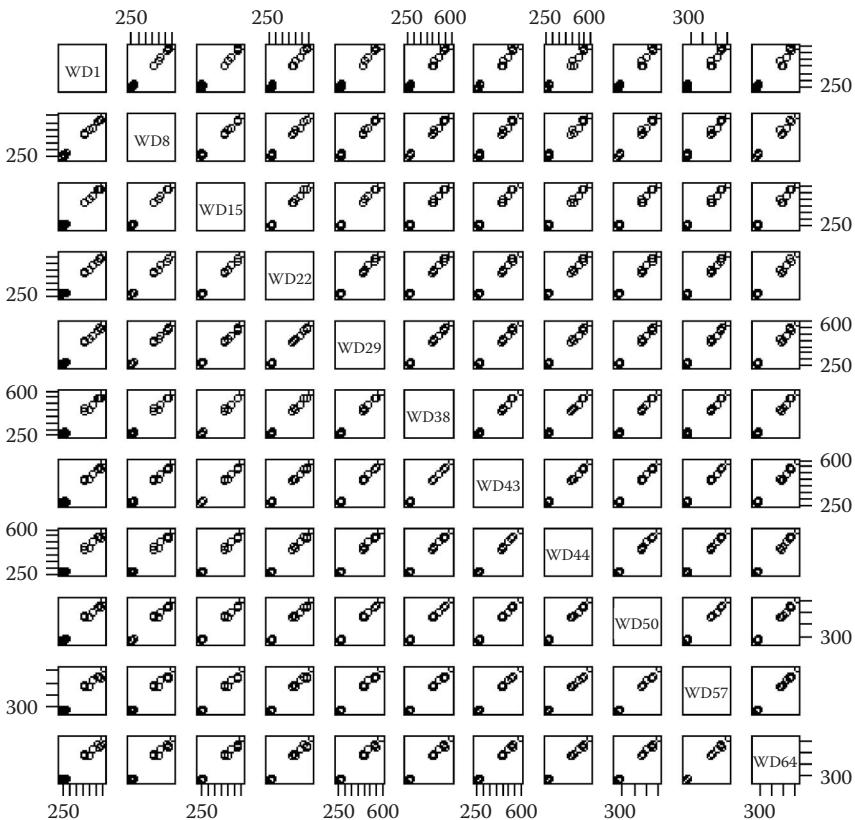


FIGURE 8.2
Plot of individual rat growth profiles.

**FIGURE 8.3**

Scatterplot matrix of repeated measures in rat growth data.

from other rats. Fitting this model gives the results shown in [Table 8.4](#). The estimated variance of the rat random effects is quite large, indicating the considerable variation in the intercepts of the regression fits of the individual rat growth profiles. The estimated regression parameters for time and the two dummy variables are very similar to those from fitting the independence model shown in [Table 8.3](#), and all are highly significant again as they were in [Table 8.3](#). However, the estimated standard error of time is much smaller in [Table 8.4](#) than it is in [Table 8.3](#), reflecting the point made in Technical Section 8.1 that assuming independence will lead to the standard error of a within-subject covariate such as time being larger than it should be because of ignoring the likely within-subject dependences, which will reduce the error variance in the model. In contrast, the standard errors of each dummy variable in [Table 8.4](#) are about three times the size of those in [Table 8.3](#). The dummy variables are between-subject effects, and the reason for the smaller standard errors with the independence model is that the effective sample size

for estimating these effects is less than the actual sample size because of the correlated nature of the data, and so the estimates for the independence model are unrealistically precise. In this example, the conclusions from the independence model and the random intercept model are the same, but in other examples this will not necessarily be so, as we shall see later.

Now we can move on to fit the random intercept and random slope model to the rat growth data; explicitly, the model is

$$y_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)t_j + \beta_2 D_{i1} + \beta_3 D_{i2} + \varepsilon_{ij}$$

where the extra term from the random intercept model is the random effect v_i that allows the linear regression fits for each individual to differ in slope; these random effects are assumed to have a normal distribution with mean 0 and variance σ_v^2 and are allowed to be correlated with the u random intercept effects (see Technical Section 8.1). The results from fitting the random intercept and slope model to the rat growth data are shown in Table 8.5. The results for the fixed effects are very similar to those in Table 8.4, but the likelihood ratio test for the random intercept model versus the random intercept and slope model gives a chi-squared statistic of 142.94 with 2 degrees of freedom (DF) (the two additional parameters in the latter model are the variance of the v random effects and the covariance of the u and v random effects), and the associated p-value is very small. The random intercept and slope model provides a better fit for these data. (There are some technical problems with this likelihood ratio test, which are discussed in detail in Rabe-Hesketh and Skrondal, 2008; fortunately, the correct p-value for testing which of the two models is to be preferred can be found simply by dividing the p-value from the flawed likelihood ratio test by 2.)

TABLE 8.4

Results from Fitting the Random Intercept Model, with Time and Group as Explanatory Variables, to Rat Growth Data

Random Effects			
$\hat{\sigma}_u^2 = 1085.92$ with estimated standard error 32.95			
$\hat{\sigma}^2 = 66.44$ with estimated standard error 8.15			
Fixed Effects			
	Estimate	Standard Error	t-Value
Intercept	244.06890	11.73023	20.81
Time	0.58568	0.03158	18.54
D_1	220.98864	20.23431	10.92
D_2	262.07955	20.23431	12.95

TABLE 8.5

Results from Fitting the Random Intercept and Slope Model, with Time and Group as Explanatory Variables, to Rat Growth Data

Random Effects			
Fixed Effects			
	Estimate	Standard Error	t-Value
Intercept	246.45465	11.81509	20.859
Time	0.58568	0.08548	6.852
D_1	214.59440	20.17923	10.634
D_2	258.93079	20.17923	12.832

^a Estimated correlation between the u and v random effects is -0.22.

Finally, we can fit a random intercept and slope model that allows for a group \times time interaction. Explicitly, this model can be written as

$$y_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)t_j + \beta_2 D_{i1} + \beta_3 D_{i2} + \beta_4 (D_{i1} \times t_j) + \beta_5 (D_{i2} \times t_j) + \varepsilon_{ij}$$

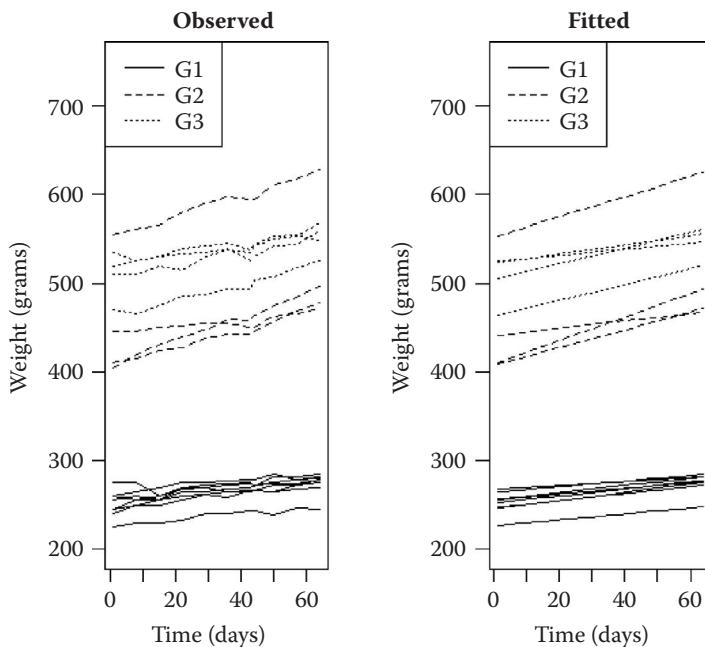
Fitting this model gives the results in Table 8.6. The likelihood ratio test of the interaction random intercept and slope model against the corresponding model without an interaction is 12.36 with 2 DF; the associated p-value is very small, and we can conclude that the interaction model provides a

TABLE 8.6

Results from Fitting the Random Intercept and Slope Model that Allows for a Group \times Time Interaction to Rat Growth Data

Random Effects ^a			
Fixed Effects			
	Estimate	Standard Error	t-Value
Intercept	251.65165	11.80280	21.321
Time	0.35964	0.08215	4.378
D_1	200.66549	20.44304	9.816
D_2	252.07168	20.44304	12.330
$D_1 \times$ time	0.60584	0.14229	4.258
$D_2 \times$ time	0.29834	0.14229	2.097

^a Estimated correlation between the u and v random effects is -0.15.

**FIGURE 8.4**

Fitted growth rate profiles from the interaction model and observed growth rate profiles.

better fit for the rat growth data. The estimated regression parameters for the interaction in Table 8.6 indicate that the growth rate slopes are considerably higher for rats in group 2 than for rats in group 1 (on average 0.61 higher with an approximate 95% confidence interval [CI] of 0.33,0.89) but less so when comparing group 3 rats with those in group 1 (on average 0.30 higher, CI 0.02,0.58). We can find the fitted values from the interaction model and plot the fitted growth rates for each rat; these are shown in Figure 8.4 alongside the observed values. This graphic underlines how well the interaction model fits the observed data. (The fitted values for each rat include “predicted” values of the u and v random effects for the rat; details of how these predicted values are calculated are given in Rabe-Hesketh and Skrondal, 2008.)

8.4 Computerized Delivery of Cognitive Behavioral Therapy—Beat the Blues

Depression is a major public health problem across the world. Antidepressants are the frontline treatment, but many patients either do not respond to them or do not like taking them. The main alternative is psychotherapy, and the

modern “talking treatments” such as cognitive behavioral therapy (CBT) have been shown to be as effective as drugs, and probably more so when it comes to relapse (Watkins and Williams, 1998). But there is a problem, namely, availability—there are simply nothing like enough skilled therapists to meet the demand and little prospect at all of this situation changing.

A number of alternative modes of delivery of CBT have been explored, including interactive systems making use of the new computer technologies. The principles of CBT lend themselves reasonably well to computerization, and perhaps surprisingly, patients adapt well to this procedure and do not seem to miss the physical presence of the therapist as much as one might expect. Workers at the Institute of Psychiatry in the United Kingdom have developed one particular program, known as Beating the Blues (BtB). Full details are given by Proudfoot et al. (2002), but in essence, BtB is an interactive program using multimedia techniques, in particular, video vignettes. The computer-based intervention consists of 9 sessions, followed by 8 therapy sessions, each lasting about 50 min. Nurses are used to explain how the program works but are instructed to spend no more than 5 min with each patient at the start of each session, and are there simply to assist with the technology. In a randomized controlled trial of the program, patients with depression recruited in primary care were randomized to either the BtB program or to “Treatment as Usual” (TAU). Patients randomized to BtB also received pharmacology and/or general practitioner (GP) support and practical/social help, offered as part of TAU, with the exception of any face-to-face counseling or psychological intervention. Patients allocated to TAU received whatever treatment their GP prescribed. The latter included, besides any medication, discussion of problems with GP, provision of practical/social help, referral to a counselor, referral to a practice nurse, referral to mental health professionals (psychologist, psychiatrist, community psychiatric nurse, counselor, etc.), or further physical examination.

A number of outcome measures were used in the trial, but here we concentrate on the Beck Depression Inventory II (BDI; Beck et al., 1996). Measurements on this variable were made on the following five occasions:

- Prior to treatment
- 2, 4, 6, and 8 months after treatment began

Data from 100 patients will be analyzed in this section; these data are a subset of the original and are used with the kind permission of the organizers of the study, in particular, Dr. Judy Proudfoot. Data for the first five patients from each treatment group are shown in [Table 8.7](#). Two additional explanatory variables are also available for each patient: the first, drug, is whether the patient was taking antidepressant drugs (yes or no), and the second, length, is the length of the current episode of depression categorized into less than 6 months or more than 6 months. The NAs (not available) in [Table 8.7](#) indicate where a protocol-specified measurement of the BDI was not made; here, all

TABLE 8.7

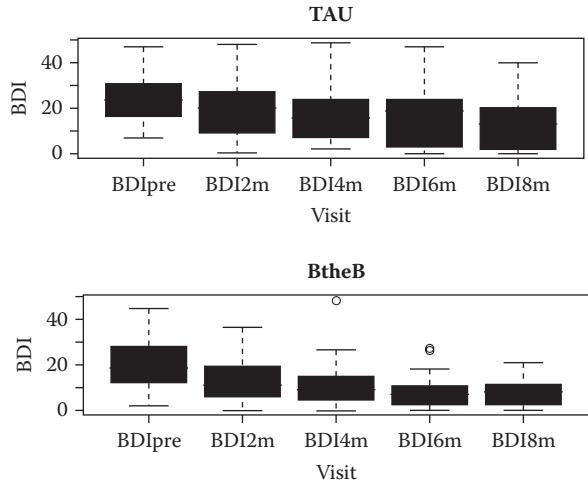
First Five Patients in Each Treatment Group of the “Beat the Blues” (BtB) Clinical Trial of CBT for Depression

Sub	Drug	Duration (month)	Treatment	BDIpre	BDI2m	BDI4m	BDI6m	BDI8m
1	No	>6	TAU	29	2	2	NA	NA
2	Yes	>6	BtheB	32	16	24	17	20
3	Yes	<6	TAU	25	20	NA	NA	NA
4	No	>6	BtheB	21	17	16	10	9
5	Yes	>6	BtheB	26	23	NA	NA	NA
6	Yes	<6	BtheB	7	0	0	0	0
7	Yes	<6	TAU	17	7	7	3	7
8	No	>6	TAU	20	20	21	19	13
9	Yes	<6	BtheB	18	13	14	20	11
10	No	>6	TAU	30	32	24	12	2

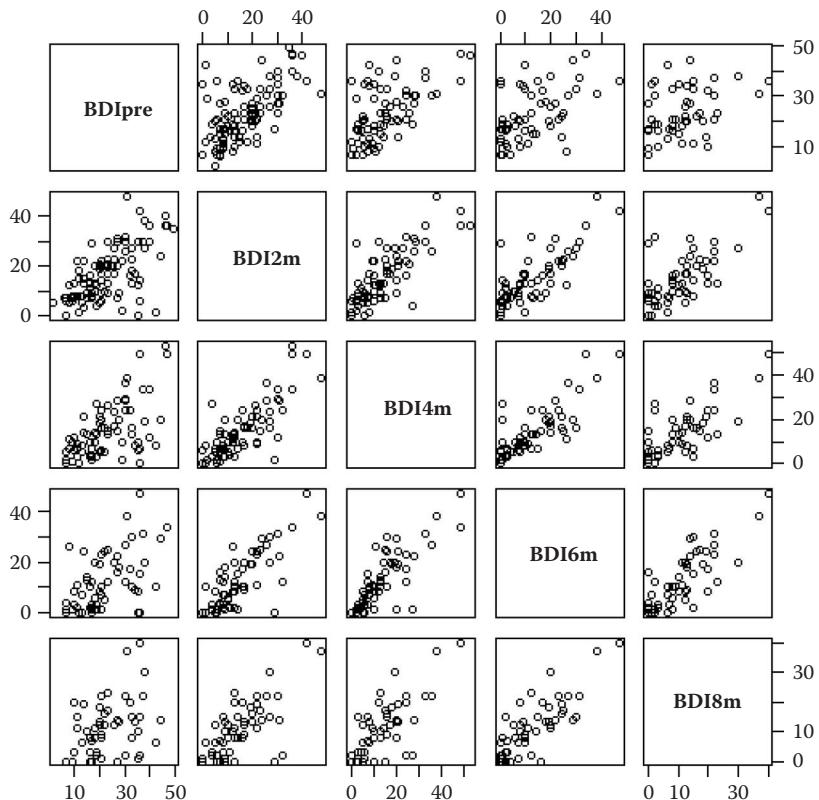
the NAs are due to patients dropping out of the study. How dropouts might affect the results obtained from the analysis of the data will be discussed later in the chapter. The main question of interest here is to estimate the treatment effect of the BtB program.

In the distant past, the analysis of the BtB data would have involved only those patients with a complete set of five BDI values. At best, a “completers only” analysis would have been inefficient because the subset of BDI values for patients who dropped out are not used, thus lowering the sample size on which the analysis is based. But using only the completers in an analysis could have more dire consequences such as giving rise to biased parameter estimates and thus incorrect inferences. By considering the data in the long form, however, we see that analyses that use all the available data, including the BDI values that are recorded for patients who eventually drop out of the study, are straightforward. However, before considering models for the data, we should, as with any data set, try to discover some features of the data from some graphical material. So, we will begin by looking at boxplots of the BDI scores at each occasion of recording for each treatment group; the plot is shown in [Figure 8.5](#). We see that the BDI scores decrease over time in each treatment group, but perhaps a little more in the BtB group, and the variance of the observations in the TAU group appears to be greater than those in the BtB group on each posttreatment time of recording. As a second graphic for these data, [Figure 8.6](#) shows the scatterplot matrix of the five BDI scores; clearly, the repeated BDI values are not independent of one another.

We now move on to considering models for the BtB data. Again, we begin with an unrealistic multiple linear regression model that assumes that the repeated measures of the BDI are independent and contains the explanatory variables, drug (coded 0 for no and 1 for yes), length (coded 0 for <6 months

**FIGURE 8.5**

Boxplots of BDI scores by occasion of recording and treatment group.

**FIGURE 8.6**

Scatterplot matrix of BDI scores.

and 1 for >6 months), treatment (coded 0 for TAU and 1 for BtB), time, and BDI prevalue. The results are shown in Table 8.8. As we might have expected, the regression parameters for time and pretreatment BDI (preBDI) are highly significant. The negative value for the time coefficient tells us what we have already surmised from the boxplots in Figure 8.5, namely, that the BDI scores decrease over time. The positive regression coefficient for BDIPre simply indicates that patients with, say, a higher-than-average BDI score before treatment begins will tend to have higher-than-average values posttreatment. The regression coefficient for drug is also very significant, and its negative value tells us that patients taking antidepressant drugs will tend to have lower BDI scores than those not taking such medication. The regression coefficient for length is not significant at the 5% level; there is no evidence that the length of the current episode of depression affects the BDI score. Finally, we see that the regression coefficient for treatment is also highly significant; treatment with BtB rather than TAU is estimated to lower depression scores on average by 3.36 BDI units conditional on the other explanatory variables with an approximate 95% CI of $[-3.36 - 2 \times 1.10, -3.36 + 2 \times 1.10]$, that is, $[-5.56, -1.16]$ (this estimated treatment difference applies to all posttreatment occasions because, in the model fitted, there is no allowance for a treatment \times time interaction). However, we know from Figure 8.6 that the independence assumption for the repeated BDI scores is almost certainly incorrect, and so we need to consider some linear mixed models for these data that do allow for departures from independence.

So, we start with a random intercept model including the same explanatory variables as the independence model. The results are shown in Table 8.9. The regression parameters for both time and preBDI remain highly significant, but those for treatment and drug are now not significant primarily because the associated estimated standard errors of these parameters have increased considerably. However, before using the estimates in Table 8.9 for interpretation, we should perhaps consider whether a random intercept and slope model gives a better fit. If we fit such a model, the likelihood ratio test

TABLE 8.8

Results from Fitting a Multiple Linear Regression Model to BtB Data Assuming Repeated Measurements of BDI are Independent

	Estimate	Standard Error	t-Value	Pr(> t)
Intercept	7.88307	1.78049	4.427	1.38e-05
preBDI	0.57237	0.05486	10.433	< 2e-16
Time	-0.96081	0.23263	-4.130	4.82e-05
Treatment	-3.35397	1.09832	-3.054	0.00248
Drug	-3.54601	1.14469	-3.098	0.00215
Length	1.75308	1.10850	1.581	0.11492

Note: Multiple R-squared: 0.3978; F-statistic: 36.2 on 5 and 274 DF; p-value: <2.2e-16.

TABLE 8.9

Results from Fitting a Random Intercept Model to BtB Data

Random Effects			
	Estimate	Standard Error	t-Value
$\hat{\sigma}_u^2 = 51.44$ with estimated standard error 7.17			
$\hat{\sigma}^2 = 25.27$ with estimated standard error 5.03			
Fixed Effects			
	Estimate	Standard Error	t-Value
Intercept	5.92153	2.30573	2.568
preBDI	0.63888	0.07961	8.026
Time	-0.71354	0.14664	-4.866
Treatment	-2.35904	1.70831	-1.381
Drug	-2.78887	1.76584	-1.579
Length	0.23815	1.67528	0.142

comparing the random intercept model with the random intercept and slope model has a value of 0.81 with 2 DF; it appears that the more complicated model is not needed for these data. So, returning to Table 8.9 and, in particular, the estimated regression parameter for treatment, we find that treatment with BtB is estimated to lower the average BDI score by 2.36 units conditional on the other covariates with 95% CI of [-5.78,1.06]; there is no compelling evidence that treatment with BtB is effective, a different conclusion from that produced by using the independence model.

8.5 The Problem of Dropouts in Longitudinal Studies

We now need to consider briefly how the dropouts may affect the analyses reported earlier. To understand the problems that patients dropping out can cause for the analysis of data from a longitudinal study, we need to consider a classification of dropout mechanisms because the type of mechanism involved has implications for which approaches to analysis are suitable and which are not. The following classification involves three types of dropout mechanism as suggested by Little (1995) and Diggle et al. (2002):

- Dropout completely at random (DCAR). Here, the probability that a subject drops out does not depend on either the observed or missing values of the response. Consequently, the observed (nonmissing) values effectively constitute a simple random sample of the values for all subjects. Possible examples include missing laboratory measurements because of a dropped test tube (if it was not

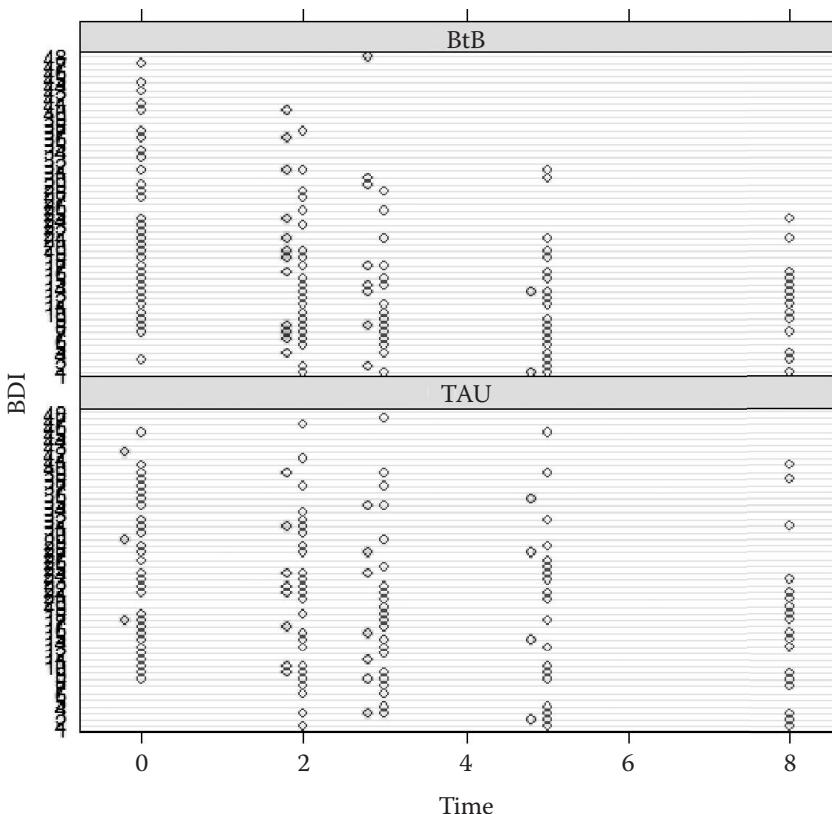
dropped because of the knowledge of any measurement), the accidental death of a participant in a study, or a participant moving to another area. Intermittent missing values in a longitudinal data set, whereby a patient misses a protocol-specified visit for transitory reasons (“went shopping instead” or the like), can reasonably be assumed to be DCAR. Completely random dropout causes the least problem for data analysis, but it is a strong assumption.

- Dropout at random (DAR). The DAR mechanism occurs when the probability of dropping out depends on the outcome measures that have been observed in the past but given that this information is conditionally independent of all the future (unrecorded) values of the outcome variable following dropout. Here, “missingness” depends only on the observed data with the distribution of future values for a subject who drops out at a particular time being the same as the distribution of the future values of a subject who remains in at that time, if they have the same covariates and the same past history of outcome up to and including the specific time point. Murray and Findlay (1988) provide an example of this type of missing value from a study of hypertensive drugs in which the outcome measure was the diastolic blood pressure. The protocol of the study specified that the participant was to be removed from the study when his or her blood pressure got too large. Here, blood pressure at the time of dropout was observed before the participant dropped out. So, although the dropout mechanism is not DCAR since it depends on the values of blood pressure, it is DAR because dropout depends only on the observed part of the data. A further example of a DAR mechanism is provided by Heitjan (1997), and involves a study in which the response measure is body mass index (BMI). Suppose that the measure is missing because subjects who had high BMI values at earlier visits avoided being measured at later visits out of embarrassment, regardless of whether they had gained or lost weight in the intervening period. The missing values here are DAR but not DCAR; consequently, methods applied to the data that assumed the latter might give misleading results.
- Nonignorable (sometimes referred to as informative). The final type of dropout mechanism is one where the probability of dropping out depends on the unrecorded missing values—observations are likely to be missing when the outcome values that would have been observed had the patient not dropped out are systematically higher or lower than usual (corresponding perhaps to their condition becoming worse or improving). An example is when individuals with lower income levels or very high incomes are less likely to provide their personal income in an interview. In a behavioral

context, another example might be a subject dropping out of a longitudinal study of pain when his or her pain becomes intolerable, and the associated pain value was not recorded. For the BDI example introduced earlier, if subjects were more likely to avoid being measured because they had put on extra weight since the last visit, then the data are nonignorably missing. Dealing with data containing missing values that result from this type of dropout mechanism is difficult. The correct analyses for such data must estimate the dependence of the missingness probability on the missing values. Models and software that attempt this are available (see, for example, Diggle and Kenward, 1994), but their use is not routine and, in addition, it must be remembered that the associated parameter estimates can be unreliable.

Under what type of dropout mechanism are the mixed-effects models considered in this chapter valid? The good news is that such models can be shown to give valid results under the relatively weak assumption that the dropout mechanism is DAR (see Carpenter et al., 2002). When the missing values are thought to be informative, any analysis is potentially problematic. But Diggle and Kenward (1994) have developed a modeling framework for longitudinal data with informative dropouts, in which random or completely random dropout mechanisms are also included as explicit models. The essential feature of the procedure is a logistic regression model for the probability of dropping out, in which the explanatory variables can include previous values of the response variable and, in addition, the unobserved value at dropout as a latent variable (i.e., an unobserved variable). In other words, the dropout probability is allowed to depend on both the observed measurement history and the unobserved value at dropout. This allows both a formal assessment of the type of dropout mechanism in the data and the estimation of effects of interest (e.g., treatment effects under different assumption about the dropout mechanism). A full technical account of the model is given in Diggle and Kenward (1994), and a detailed example that uses the approach is described in Carpenter et al. (2002).

One of the problems for an investigator struggling to identify the dropout mechanism in a data set is that there are no routine methods to help, although a number of largely ad hoc graphical procedures can be used as described by Diggle (1998), Everitt (2002), and Carpenter et al., (2002). One of these is illustrated in [Figure 8.7](#); here, the observations from each treatment group in the BtB data are plotted, differentiating between two categories of patients, namely, those who do and those who do not attend their next scheduled visit. Any clear difference between the distribution of BDI scores in these two categories indicates that dropout is not completely at random. In [Figure 8.7](#), there appears to be no very clear difference in the distribution of BDI scores for those patients who attend their next scheduled visit and those who have dropped out before this visit.

**FIGURE 8.7**

BtB data by treatment group and time, identifying patients who do and do not attend their next scheduled visit.

8.6 Summary

- Repeated measurements of a response under different experimental conditions or over a period of time occur often in behavioral research.
- The analysis of such data requires special techniques because of the likely nonindependence of the repeated measurements.
- Linear mixed models allow for correlations between the repeated measurements by introducing random effects for subjects.
- The essential feature of such models is that there is natural heterogeneity across individuals in their responses over time and that this heterogeneity can be represented by an appropriate probability distribution. Correlation between observations from the same

individual arises from unobserved or unmeasured characteristics of the individual that remain the same over time, for example, an increased propensity to the condition under investigation, or perhaps a predisposition to exaggerate symptoms.

- Conditional on the values of the random effects, the repeated measurements of the response variable are assumed independent—the local independence assumption.
 - Linear mixed-effects models can be fitted by maximum likelihood, and competing models can be assessed by a likelihood ratio test.
 - Subjects who drop out of longitudinal studies can cause problems for the analysis of a data set, and if there are a large number of dropouts, then conclusions drawn from any analysis must be treated with caution.
 - In this chapter, only responses that can be assumed to have a normal distribution conditional on the explanatory variables have been considered. For nonnormal longitudinal data, see, for example, Rabe-Hesketh and Skrondal (2008).
-

8.7 Exercises

- 8.1 For the BtB data, construct a plot that shows the mean profiles over time for each treatment group and has appropriate error bars at each time point.
- 8.2 Investigate whether there is any evidence of a treatment \times time interaction in the BtB data.
- 8.3 The data in exer_83.txt arise from a trial of estrogen patches in the treatment of postnatal depression. Women who had suffered an episode of postnatal depression were randomly allocated to two groups: the members of one group received an estrogen patch, and the members of the other group received a “dummy” patch—the placebo. The dependent variable was a composite measure of depression, which was recorded on two occasions prior to randomization and for each of 6 months posttreatment. A number of observations are missing (indicated by -9) because some women dropped out of the study. Begin by fitting an independence model to the data using multiple linear regression with the depression score as dependent variable, and treatment time and the two baseline measurements as explanatory variables. Next, fit both a random intercept and random intercept and slope model, and test which model is best. Use the estimated treatment effect from the model you choose to find a confidence interval for the treatment effect. What do you conclude from this confidence interval?

8.4 The data in exer_84.txt give the plasma inorganic phosphate levels for 33 subjects, 20 of whom are controls and 13 of whom have been classified as obese (Davis, 2002). Produce separate plots of the profiles of the individuals in each group and, guided by these plots, fit what you think might be a sensible linear mixed model to the data.

9

Multivariate Data and Multivariate Analysis

9.1 Introduction

In this short chapter and the following three longer chapters (Chapters 10, 11, and 12) we will be concerned with multivariate data. Such data arise when researchers measure several variables on each individual in their study, and where all variables need to be examined simultaneously in order both to uncover whatever “patterns” or “structure” the data may contain and understand the key features of the data. All the variables in a multivariate data set are random variables, unlike in the regression models of Chapter 3 to 8, where only the response variable is considered to be a random variable. The analysis of multivariate data, multivariate analysis, is essentially a collection of techniques, many largely descriptive rather than inferential, that have in common the aim to display or extract any “signal” in the data in the presence of noise and, in a very general sense, to discover what the data may be trying to tell us.

Multivariate data sets have a common form, and consist of a data matrix, the rows of which contain the units in the sample and the columns of which refer to the variables measured on each unit. Symbolically, a set of multivariate data can be represented by the matrix \mathbf{X} given by

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{1q} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix}$$

where n is the number of units in the sample, q is the number of variables measured on each unit, and x_{ij} denotes the value of the j th variable for the i th individual; an individual in this context may be something other than a human being. The variables will often be at different levels in the measurement hierarchy described in Chapter 1, for example, some categorical, some interval, and some ratio measurements. Further, often some variable values will be missing, and so, the problems associated with missing data mentioned in Chapter 1 and again in Chapter 8 may assume importance.

9.2 The Initial Analysis of Multivariate Data

The main techniques for analyzing multivariate data are those to be described in Chapters 10, 11, and 12, but an initial graphical and numerical description can often be very helpful in gaining some insight into the data and in interpreting the results from later analyses. We will illustrate what can be done with an initial analysis using first the very simple multivariate data set shown in Table 9.1. The data consists of chest, hip, and waist measurements (in inches) of 20 individuals.

9.2.1 Summary Statistics for Multivariate Data

In order to provide a numerical summary for a multivariate data set, we need to produce summary statistics for each of the variables separately and also calculate appropriate statistics that summarize the relationships between the variables. For the former, we generally use means and variances (assuming that we are dealing with continuous variables), and for the latter, we usually take pairs of variables at a time and look at their covariances or correlations. Population and sample versions of all these quantities are defined in Technical [Section 9.1](#).

TABLE 9.1
Chest, Waist, and Hip Measurements of 20 Individuals

Individual	Chest	Waist	Hips
1	34	30	32
2	37	32	37
3	38	30	36
4	36	33	39
5	38	29	33
6	43	32	38
7	40	33	42
8	38	30	40
9	40	30	37
10	41	32	39
11	36	24	35
12	36	25	37
13	34	24	37
14	33	22	34
15	36	26	38
16	37	26	37
17	34	25	38
18	36	26	37
19	38	28	40
20	35	23	35

Technical Section 9.1: Numerical Summary Statistics for Multivariate Data

For q variables, the population mean vector is usually represented as $\boldsymbol{\mu}' = [\mu_1, \mu_2, \dots, \mu_q]$, where $\mu_i = E(x_i)$ is the population mean (or expected value as denoted by the E operator here; see Chapter 6) of the i th variable. An estimate of $\boldsymbol{\mu}'$ based on n q -dimensional observations is $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_q]$, where \bar{x}_i is the sample mean of the variable x_i . The vector of population variances can be represented by $\boldsymbol{\sigma}' = [\sigma_1^2, \sigma_2^2, \dots, \sigma_q^2]$, where $\sigma_i^2 = E(x_i - \mu_i)^2$. An estimate of $\boldsymbol{\sigma}'$ based on n q -dimensional observations is $\mathbf{s}' = [s_1^2, s_2^2, \dots, s_q^2]$, where s_i^2 is the sample variance of x_i .

The population covariance of two variables x_i and x_j is defined by

$$\text{Cov}(x_i, x_j) = E(x_i - \mu_i)(x_j - \mu_j)$$

If $i = j$, we note that the covariance of the variable with itself is simply its variance, and therefore, there is no real need to define variances and covariances independently in the multivariate case. The covariance of x_i and x_j is usually denoted by σ_{ij} (so, the variance of the variable x_i is often denoted by σ_{ii} rather than σ_i^2).

With q variables, x_1, x_2, \dots, x_q , there are q variances and $q(q-1)/2$ covariances. In general, these quantities are arranged in a $q \times q$ symmetric matrix Σ where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq} \end{pmatrix}$$

Note that $\sigma_{ij} = \sigma_{ji}$. This matrix is generally known as the variance-covariance matrix or simply the covariance matrix. The matrix Σ is estimated by the matrix \mathbf{S} , given by

$$\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' / (n-1)$$

where $\mathbf{x}'_i = [x_{i1}, x_{i2}, \dots, x_{iq}]$ is the vector of observations for the i th individual. The diagonal of \mathbf{S} contains the variances of each variable. The covariance is often difficult to interpret because it depends on the units in which the two variables are measured; consequently, it is often standardized by

dividing by the product of the standard deviations of the two variables to give a quantity called the correlation coefficient, ρ_{ij} , where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

The correlation coefficient lies between -1 and $+1$ and gives a measure of the linear relationship between the variables x_i and x_j . It is positive if high values of x_i are associated with high values of x_j , and negative if high values of x_i are associated with low values of x_j . With q variables, there are $q(q-1)/2$ distinct correlations, which may be arranged in a $q \times q$ matrix whose diagonal elements are unity.

For sample data, the correlation matrix contains the usual estimates of the ρ values, namely, Pearson's correlation coefficient, and is generally denoted by \mathbf{R} . The matrix may be written in terms of the sample covariance matrix \mathbf{S} as follows:

$$\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}$$

where $\mathbf{D}^{-1/2} = \text{diag}(1/s_i)$.

In most situations, we will be dealing with covariance and correlation matrices of full rank q , so that both matrices will be nonsingular (i.e., invertible).

For the body measurements data, the numerical summary statistics described earlier are as follows:

Means:

Chest	Waist	Hips
37.00	28.00	37.05

Variances:

Chest	Waist	Hips
6.63	12.53	5.94

Covariance matrix:

	Chest	Waist	Hips
Chest	6.63	6.37	3.00
Waist	6.37	12.53	3.58
Hips	3.00	3.58	5.94

Correlation matrix:

	Chest	Waist	Hips
Chest	1.00	0.70	0.48
Waist	0.70	1.00	0.41
Hips	0.48	0.41	1.00

We see that the waist measurements have the largest variance, with chest and hip measurements having very similar variances. Waist and hip measurements have the largest correlation with a value of 0.70.

9.2.2 Graphical Descriptions of the Body Measurement Data

As always, numerical summaries need to be interpreted alongside appropriate graphics, and so, in Figure 9.1, we give the boxplots of each of the three body measurements, and then, in [Figure 9.2](#), a scatterplot matrix of the three variables is shown with a histogram of each measurement placed on the diagonal. The boxplot for chest measurements shows a mild degree of skewness and one potential outlier. The scatterplot matrix is more interesting,

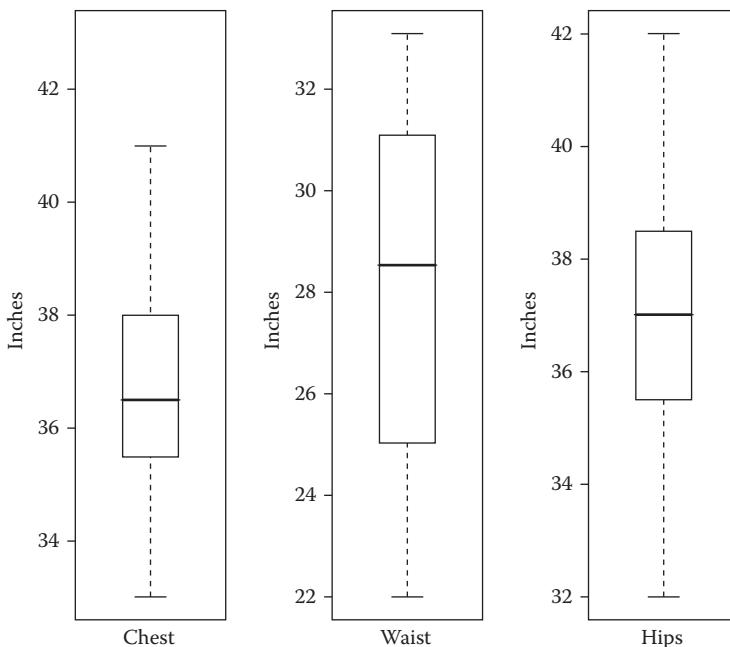
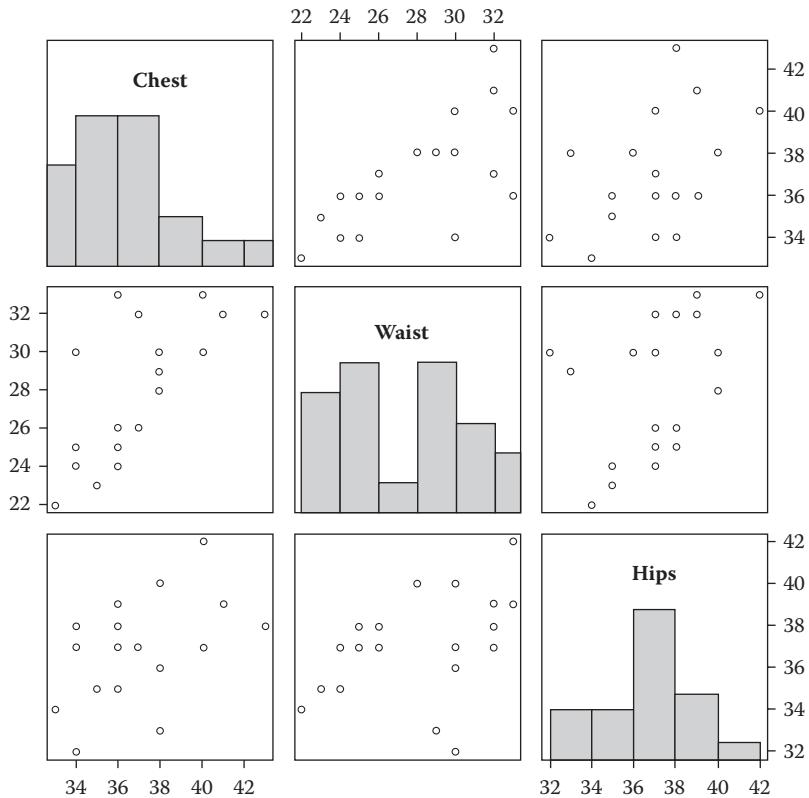


FIGURE 9.1
Boxplots of body measurements.

**FIGURE 9.2**

Scatterplot matrix of body measurements.

with the panel showing the scatterplot of waist and hip measurements suggesting the possibility that the data may consist of two separate groups of observations, a possibility underlined by the bimodality of the histogram for waist measurements. The two-group possibility could be investigated further by applying cluster analysis (see Chapter 12) to the data, but here, it is not too taxing to come up with an explanation for the possible two-group structure, which is that there are men and women in the sample. (If there really are distinct groups of observations in the data, the previously calculated summary statistics for the whole sample may not give an accurate description of the separate groups.)

9.3 The Multivariate Normal Probability Density Function

As we have seen in earlier chapters, the normal probability density function is the basis of the inferences derived from most multivariable techniques. In multivariate analysis, it is the multivariate normal density function that has

a similar role, although many multivariate analyses are carried out in the spirit of data exploration (what would now be called data mining perhaps), where questions of statistical significance are of minor or no importance. Nevertheless, researchers in behavioral sciences dealing with the complexities of multivariate data may on occasion need to know a little about the multivariate density function and, in particular, how to assess whether or not a set of multivariate data can be assumed to have this probability density function. So, in Technical [Section 9.2](#), we define the multivariate normal density and describe some of its properties.

[Technical Section 9.2: Multivariate Normal Density Function](#)

For a vector of q random variables $\mathbf{x}' = [x_1, x_2, \dots, x_q]$, the multivariate normal density function takes the form

$$f(\mathbf{x}) = (2\pi)^{-q/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\} - \infty < x_i < \infty$$

where Σ is the population covariance matrix of the variables, and $\boldsymbol{\mu}$ is the vector of population mean values of the variables. The simplest example of the multivariate normal density function is the bivariate normal density with $q = 2$; this can be written explicitly as

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \\ &\times \exp\left\{\frac{-1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right]\right\} \end{aligned}$$

where μ_1 and μ_2 are the population means of the two variables, σ_1^2 and σ_2^2 are the population variances, and ρ is the population correlation between the two variables. [Figure 9.3](#) shows an example of a bivariate normal density function with both means equal to 0, both variances equal to 1, and correlation equal to 0.5.

The population mean vector and the population covariance matrix of a multivariate density function are estimated from a sample of multivariate observations as described in Technical [Section 9.1](#).

One property of a multivariate normal density function that is worth mentioning here is that linear combinations of the variables, that is, $y = a_1x_1 + a_2x_2 + \dots + a_qx_q$ where a_1, a_2, \dots, a_q is a set of scalars, are themselves normally distributed with mean $\mathbf{a}'\boldsymbol{\mu}$ and variance $\mathbf{a}'\Sigma\mathbf{a}$ where $\mathbf{a}' = [a_1, a_2, \dots, a_q]$. Linear combinations of variables will be of importance in later chapters.

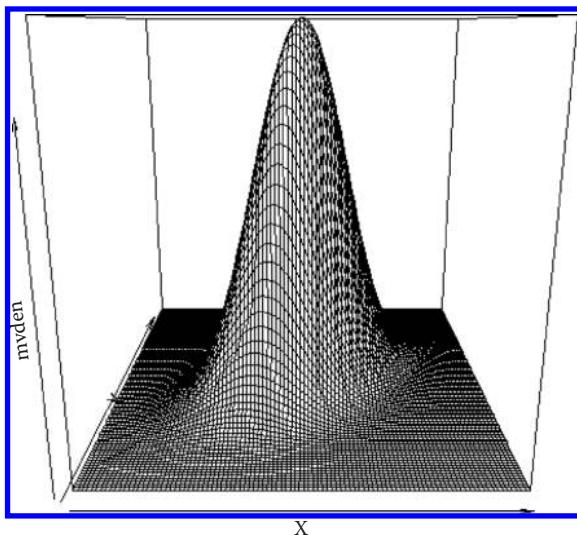


FIGURE 9.3
Bivariate normal density function.

For many multivariate methods to be described in later chapters, the assumption of multivariate normality is often not critical to the results of the analysis. But there may be occasions when testing for multivariate normality may be of interest. A start can be made perhaps by testing each separate variable for univariate normality using, say, a probability plot, as described in Chapter 2. Such plots can be helpful, but unfortunately, marginal multivariate normality does not necessarily imply multivariate normality. An alternative (additional) approach is to convert the multivariate observations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ into a set of generalized distances d_i^2 , giving a measure of the distance of each particular observation from the mean vector of the complete sample $\bar{\mathbf{x}}$; d_i^2 is calculated as

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

where \mathbf{S} is the sample covariance matrix. This distance measure takes into account the different variances of the variables and the covariances of pairs of variables. If the observations do arise from a multivariate normal distribution, then the generalized distances have, approximately, a chi-squared distribution with q degrees of freedom. So, plotting the ordered distances against the corresponding quantiles of the appropriate chi-squared distribution should lead to a straight line through the origin.

We will now assess the body measurements data for normality, although, because there are only 20 observations in the sample, there is really too little

information to come to any convincing conclusion. Figure 9.4 shows separate probability plots for each measurement; in the plots, there appears to be no evidence of any departures from linearity.

The chi-square plot of the 20 generalized distances in Figure 9.5 does seem to deviate a little from linearity, but with so few observations, it is hard to be certain.

Now let us look at a larger example of a multivariate data set collected in a health survey of 103 paint sprayers in a car assembly plant. Six variables were recorded on each individual:

- Hemo: Hemoglobin concentration
- PCV: Packed cell volume
- WBC: White blood cell count
- Lympho: Lymphocyte count
- Neutro: Neutrophil count
- Lead: Serum lead concentration

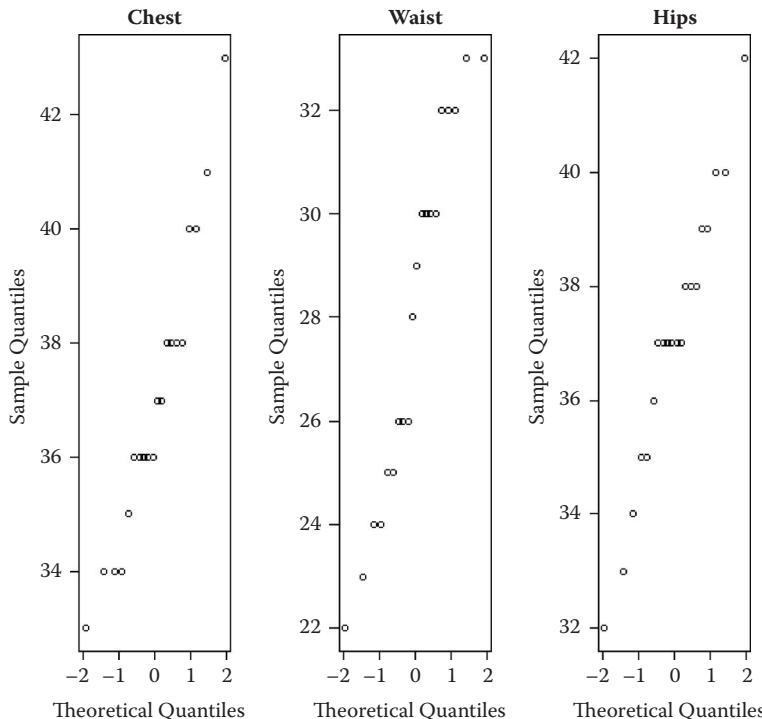
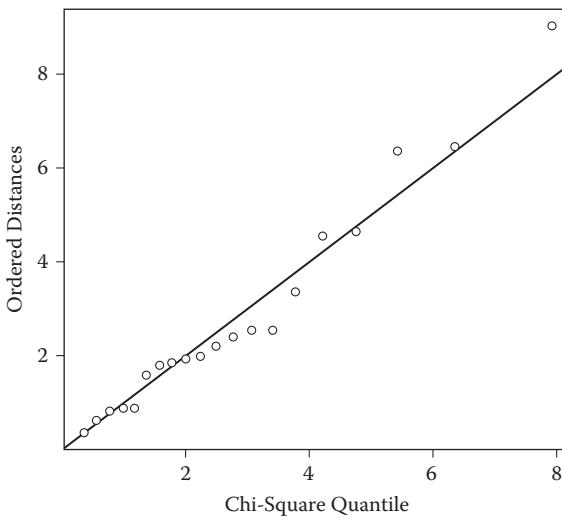


FIGURE 9.4
Normal probability plots of chest, waist, and hip measurements.

**FIGURE 9.5**

Chi-square plot of generalized distances for the body measurements data.

Data for the first five paint sprayers are shown in Table 9.2.

We begin by looking at separate normal probability plots of each of the six variables; the plots are given in Figure 9.6. The plots for WBC, Lympho, and Lead show some deviation for linearity that would suggest that the six variables do not have a multivariate normal density. The chi-square plot of generalized distances in Figure 9.7 appears to confirm this because there is considerable departure from linearity in this plot although this is primarily due to a relatively few observations. Here, it is of interest to see what happens when we take a log transformation of all the variables and then look at the chi-square plot of the transformed data, given in Figure 9.8. In this plot, it looks like just six observations deviate from the linearity required to give the data a multivariate normal density seal of approval. Identifying these outliers and creating a chi-square plot for the remaining log-transformed observations might be useful (see Exercise 9.2).

TABLE 9.2
Part of the Data on Paint Sprayers

Case	Hemo	PCV	WBC	Lympho	Neutro	Lead
1	13.4	39	4100	14	25	17
2	14.6	46	5000	15	30	20
3	13.5	42	4500	19	21	18
4	15.0	46	4600	23	16	18
5	14.6	44	5100	17	31	19

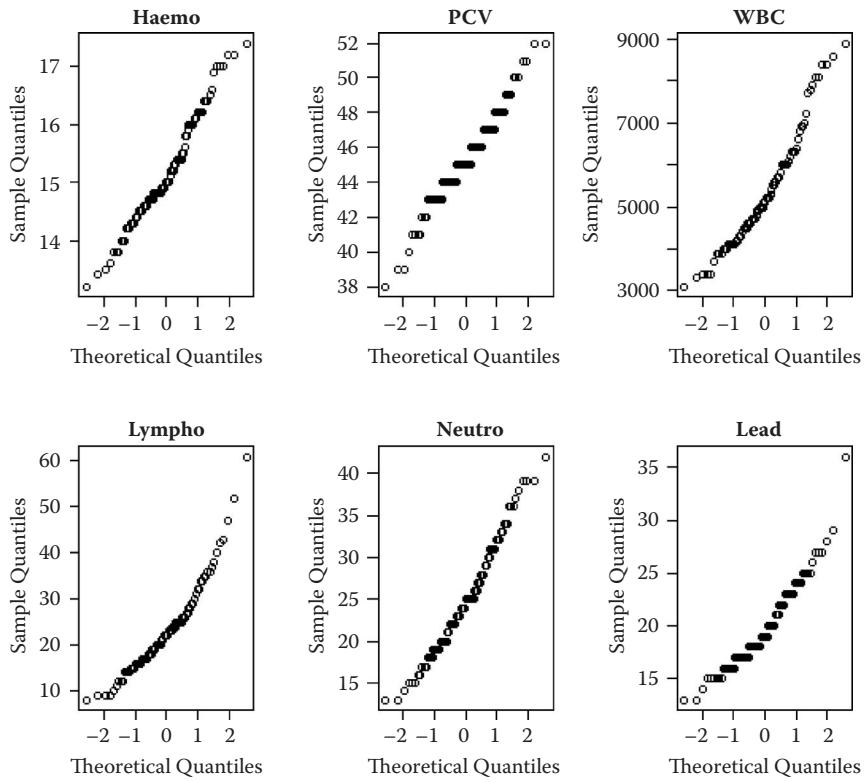


FIGURE 9.6
Normal probability plots for the six variables in the data on paint sprayers.

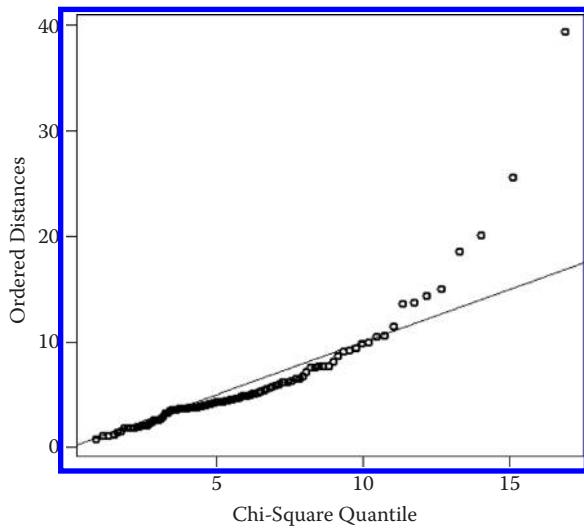
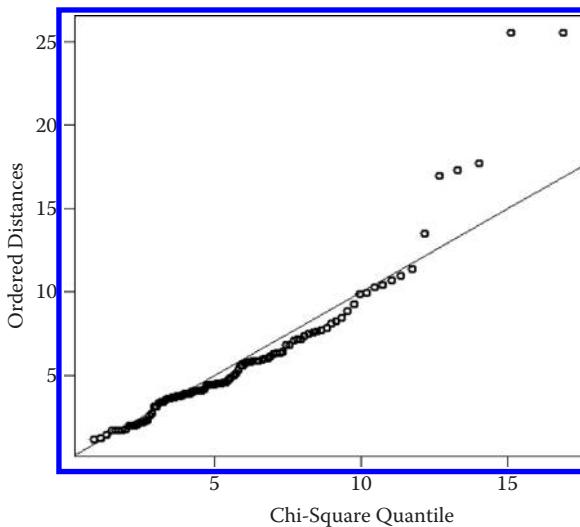


FIGURE 9.7
Chi-square plot of generalized distances for paint-sprayer data.

**FIGURE 9.8**

Chi-square plot of paint sprayers data after a log transformation of all the six variables.

9.4 Summary

- Multivariate data arise when researchers measure several variables on each individual in their sample and there is no response variable.
- Although in some cases it may make sense to isolate each variable and study it separately, in the main it does not. In most instances the variables are related in such a way that, when analyzed in isolation, they may often fail to reveal the full structure of the data. With the great majority of multivariate data sets, all the variables need to be examined simultaneously in order to uncover the patterns and key features in the data.
- Several multivariate techniques are largely “exploratory” in nature and, in many cases, a set of multivariate data may not arise as a sample from some populations. Consequently, questions of inference become less important.
- Where inference for multivariate data is an issue, it is usually based on the assumption that the sample data arise from a population in which the variables have a multivariate normal density function. In such cases, it may be worth assessing the assumption.

9.5 Exercises

- 9.1 For the body measurements data, using the scatterplot matrix of the data as a guide, try to identify the men and the women in the sample and then find the means, variances, and covariance and correlation matrices of the groups you identify.
- 9.2 For the paint-sprayer data, use some suitable graphics to identify any observations that you think are outliers. Construct chi-square plots of the generalized distances after removing the outlying observations both for the raw data and log-transformed data.
- 9.3 The data in exer_93.txt give the life expectancies in different countries by age and by sex. Find numerical summaries separately for men and women, and construct suitable graphics for an initial examination of the data.

10

Principal Components Analysis

10.1 Introduction

One of the problems with many sets of multivariate data is that there are simply too many variables to make the application of, say, some of the graphical techniques described in Chapter 2 successful in providing an informative initial assessment of the data. Further, having too many variables may cause problems for other statistical techniques that the researcher may want to apply to the data. The possible problem of too many variables is sometimes known as the curse of dimensionality. Clearly, the scatterplots, scatterplot matrices, and other graphics that might be applied to multivariate data for an initial assessment are likely to be more useful when the number of variables in the data, the dimensionality of the data, is relatively small rather than large. This brings us to principal components analysis (PCA), a multivariate technique with the central aim of reducing the dimensionality of a multivariate data set while retaining as much as possible of the variation present in it. This aim is achieved by transforming to a new set of variables the principal components that are uncorrelated and that are ordered, so that the first few of them account for most of the variation in all the original variables. In the best of all possible worlds, the result of a PCA would be the creation of a small number of new variables that can be used as surrogates for the originally large number of variables and, consequently, that provide a simpler basis for, say, graphing or summarizing the data and also, perhaps, when undertaking further multivariate analyses of the data.

10.2 Principal Components Analysis (PCA)

The basic goal of PCA is to describe variation in a set of correlated variables, x_1, x_2, \dots, x_q , in terms of a new set of uncorrelated variables, y_1, y_2, \dots, y_q , each of which is a linear combination of the x variables. The new variables are derived in decreasing order of “importance” in the sense that y_1 accounts for as much as possible of the variation in the original data among all linear

combinations of x_1, x_2, \dots, x_q . Then, y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1 , and so on. The new variables defined by this process, y_1, y_2, \dots, y_q , are the principal components. (Principal components analysis was first suggested by Pearson [1901] and independently by Hotelling [1933].)

The general hope of PCA is that the first few components will account for a substantial proportion of the variation in the original variables x_1, x_2, \dots, x_q , and can, consequently, be used to provide a convenient lower-dimensional summary of these variables that might prove useful for a variety of reasons. Consider, for example, a set of data consisting of examination scores for several different subjects for each of a number of students. One question of interest might be how best to construct an informative index of overall examination performance. One obvious possibility would be to take the mean score for each student, although, if the possible or observed range of examination scores varied from subject to subject, it might be more sensible to weight the scores in some way before calculating the average or, alternatively, standardize the results for the separate examinations before attempting to combine them. In this way, it might be possible to spread the students out further and so obtain a better ranking. The same result could often be achieved by applying the principal components to the observed examination results and using the students' scores on the first principal component to provide a measure of examination success that maximally discriminated between them.

A further possible application for PCA arises in the field of economics, where complex data are often summarized by some kind of index number, for example, indices of prices, wage rates, cost of living, and so on. When assessing changes in prices over time, the economist will wish to allow for the fact that prices of some commodities are more variable than others or that the prices of some of the commodities are considered more important than others; in each case, the index will need to be weighted accordingly. In such examples, the first principal component can often satisfy the investigator's requirements.

However, it is not always the first principal component that is of most interest to a researcher. A taxonomist, for example, when investigating variation in morphological measurements on animals for which all the pairwise correlations are likely to be positive, will often be more concerned with the second and subsequent components since these might provide a convenient description of aspects of an animal's "shape"; this will often be of more interest to the researcher than aspects of an animal's "size," which here, because of the positive correlations, will be reflected in the first principal component. For essentially the same reasons, the first principal component derived from, say, clinical psychiatric scores on patients may only provide an index of the severity of symptoms, and it is the remaining components that will give the psychiatrist important information about the "pattern" of symptoms.

The principal components are most commonly (and properly) used as a means of constructing an informative graphical representation of the data (see

[Section 10.10.2](#)) or as inputs to some other analysis. One example of the latter is provided by regression analysis; principal components may be useful here when

- There are too many explanatory variables relative to the number of observations.
- The explanatory variables are highly correlated.

Both situations lead to problems when applying regression techniques—problems that may be overcome by replacing the original explanatory variables with the first few principal component variables derived from them. An example will be given later, and other applications of the technique are described in Rencher (2002).

A further example when the results from a PCA may be useful is in the application of multivariate analysis of variance (see Chapter 13), when there are too many original variables to ensure that the technique can be used with reasonable power. In such cases, the first few principal components might be used to provide a smaller number of variables for analysis.

In the behavioral sciences, particularly psychology, the principal components are often considered an end in themselves, and researchers may then try to interpret them in a similar fashion to the factors in an exploratory factor analysis (see Chapter 11). We shall make some comments about this practice later in the chapter.

10.3 Finding the Sample Principal Components

PCA is overwhelmingly an exploratory technique for multivariate data. Although there are inferential methods available for using the sample principal components derived from a random sample of individuals from a population to test hypotheses about population principal components (see Jolliffe, 2002), they are very rarely to be seen in the accounts of PCA analysis that appear in the literature. Quintessentially, PCA is an aid in helping us understand the sample data. We use this observation as the rationale for describing only sample principal components in this chapter.

The first principal component of the observations is that linear combination of the original variables whose sample variance is the greatest among all possible such linear combinations. The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variances, subject to being uncorrelated with the first principal component. Subsequent components are defined similarly. The question now arises as to how the coefficients specifying the linear combinations of the original variables defining each component are found. This question is answered in Technical [Section 10.1](#).

Technical Section 10.1: Extracting Principal Components

The first principal component of the observations, y_1 , is the linear combination

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$$

whose sample variance is the greatest among all such linear combinations. Because the variance of y_1 could be increased without limit simply by increasing the values of the coefficients $a_{11}, a_{12}, \dots, a_{1q}$ (which we will write as the vector \mathbf{a}_1), a restriction must be placed on these coefficients. As we shall see later, a sensible constraint is to require that the sum of squares of the coefficients should take the value 1, although other constraints are possible, and any multiple of the vector \mathbf{a}_1 produces basically the same component. To find the coefficients defining the first principal component, we need to choose the elements of the vector \mathbf{a}_1 so as to maximize the variance of y_1 subject to the sum-of-squares constraint, which can be written as $\mathbf{a}_1' \mathbf{a}_1 = 1$. The sample variance of y_1 , which is a linear function of the x variables, is given by (see Chapter 9)

$$\text{var}(y_1) = \mathbf{a}_1' \mathbf{S} \mathbf{a}_1$$

where \mathbf{S} is the $q \times q$ sample covariance matrix of the x variables. To maximize a function of several variables subject to one or more constraints, the method of Lagrange multipliers is used. Full algebraic details are given in Morrison (1990) and Jolliffe (2002), and we will not give them here. (The algebra of an example with $q = 2$ is, however, given in [Section 10.5](#).) We simply state that the Lagrange multiplier approach leads to the solution that \mathbf{a}_1 is what is called an eigenvector or characteristic vector of the sample covariance matrix \mathbf{S} , and that it is the eigenvector corresponding to the largest of what are called the eigenvalues or characteristic roots of \mathbf{S} . (Eigenvalues of \mathbf{S} and the corresponding eigenvectors are found by numerical algorithms, the details of which are not necessary to know to understand PCA.)

The second principal component y_2 is defined to be the linear combination

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q$$

(that is, $y_2 = \mathbf{a}_2' \mathbf{x}$, where $\mathbf{a}_2' = [a_{21}, a_{22}, \dots, a_{2q}]$ and $\mathbf{x}' = [x_1, x_2, \dots, x_q]$)

which has the greatest variance subject to the following two conditions:

$$\mathbf{a}_2' \mathbf{a}_2 = 1$$

$$\mathbf{a}_2' \mathbf{a}_1 = 0$$

(The second condition specifies that y_1 and y_2 are uncorrelated).

Similarly, the j th principal component is the linear combination $y_1 = \mathbf{a}'\mathbf{x}$, which has the greatest variance subject to the conditions

$$\mathbf{a}'\mathbf{a}_j = 1$$

$$\mathbf{a}'\mathbf{a}_i = 0 \quad (i < j)$$

Application of the Lagrange multiplier technique demonstrates that the vector of coefficients defining the j th principal component, that is, \mathbf{a}_j , is the eigenvector of \mathbf{S} associated with its j th largest eigenvalue. If the q eigenvalues of \mathbf{S} are denoted by $\lambda_1, \lambda_2, \dots, \lambda_q$, then by requiring that $\mathbf{a}'\mathbf{a}_i = 1$, it can be shown that the variance of the i th principal component is given by λ_i . The total variance of the q principal components will equal the total variance of the original variables so that

$$\sum_{i=1}^q \lambda_i = s_1^2 + s_2^2 + \dots + s_q^2$$

where s_i^2 is the sample variance of x_i . We can write this more concisely as

$$\sum_{i=1}^q \lambda_i = \text{trace}(\mathbf{S})$$

Consequently, the j th principal component accounts for a proportion of the total variation of the original data, where

$$P_j = \frac{\lambda_j}{\text{trace}(\mathbf{S})}.$$

The first m principal components, where $m < q$, account for a proportion of the total variation in the original data, where

$$P^{(m)} = \frac{\sum_{i=1}^m \lambda_i}{\text{trace}(\mathbf{S})}$$

In geometrical terms, it is easy to show that the first principal component defines the line of best fit (in the least-squares sense) to the q -dimensional observations in the sample. These observations may therefore be represented in one dimension by taking their projection onto this line, that is, finding their first principal component score. If the observations happen to be collinear in q dimensions, this representation would completely account for the variation in the data, and the sample covariance matrix would have only one nonzero eigenvalue. In practice, of course, such collinearity is extremely unlikely, and an improved representation would be given by

projecting the q -dimensional observations onto the space of the best fit, this being defined by the first two principal components. Similarly, the first m components give the best fit in m dimensions. If the observations fit exactly into a space of m dimensions, it would be indicated by the presence of $q - m$ zero eigenvalues of the covariance matrix. This would imply the presence of $q - m$ linear relationships between the variables. Such constraints are sometimes referred to as structural relationships. In practice, in the vast majority of applications of PCA, all the eigenvalues of the covariance matrix will be nonzero.

10.4 Should Principal Components Be Extracted from the Covariance or the Correlation Matrix?

The account of principal components given above has them extracted from the covariance matrix of the data. However, imagine a set of multivariate data in which the variables x_1, x_2, \dots, x_p are of completely different types, for example, length, temperature, blood pressure, anxiety rating, etc. With such a data set, the structure of the principal components derived from the covariance matrix will depend upon the essentially arbitrary choice of units of measurement; for example, changing lengths from centimeters to inches will alter the derived components. Additionally, if there are large differences between the variances of the original variables, then the ones whose variances are the largest will tend to dominate the early components. This difficulty is overcome in practice by extracting the components from the correlation matrix \mathbf{R} . Extracting the components as the eigenvectors of \mathbf{R} is equivalent to calculating the principal components from the original variables after each has been standardized to have unit variance. It should be noted, however, that there is rarely any simple correspondence between the components derived from \mathbf{S} and those derived from \mathbf{R} . In addition, choosing to work with \mathbf{R} rather than \mathbf{S} involves a definite, but possibly arbitrary, decision to make the variables “equally important.”

To demonstrate how the principal components of the covariance matrix of a data set can differ from the components extracted from the data’s correlation matrix, we will use the example given in Jolliffe (2002). The data in this example consist of eight blood chemistry variables measured on 72 patients in a clinical trial. The correlation matrix of the data, together with the standard deviations of each of the eight variables, is given in [Table 10.1](#); there are considerable differences between these standard deviations. We can apply PCA to both the covariance and correlation matrix of the data (the covariance matrix is not given, but it can be easily calculated from the correlation matrix and the standard deviations—see Chapter 9). The details of the principal components of the covariance matrix are given in [Table 10.2](#), and those of the correlation matrix in [Table 10.3](#) (in both tables, very small values have been set to 0).

TABLE 10.1

Correlations of Blood Chemistry Variables and Their Standard Deviations

	rBlood	Plate	wBlood	Neut.	Lymph	Bilir.	Sodium	Potass.
rBlood	1.000	0.290	0.202	-0.055	-0.105	-0.252	-0.229	0.058
Plate	0.290	1.000	0.415	0.285	-0.376	-0.349	-0.164	-0.129
wBlood	0.202	0.415	1.000	0.419	-0.521	-0.441	-0.145	-0.076
Neut.	-0.055	0.285	0.419	1.000	-0.877	-0.076	0.023	-0.131
Lymph	-0.105	-0.376	-0.521	-0.877	1.000	0.206	0.034	0.151
Bilir.	-0.252	-0.349	-0.441	-0.076	0.206	1.000	0.192	0.077
Sodium	-0.229	-0.164	-0.145	0.023	0.034	0.192	1.000	0.423
Potass.	0.058	-0.129	-0.076	-0.131	0.151	0.077	0.423	1.000
Standard deviations	0.371	41.253	1.935	0.077	0.071	4.037	2.732	0.297

Examining the results in Tables 10.2 and 10.3, we see that each of the principal components of the covariance matrix is largely dominated by a single variable, whereas those for the correlation matrix have moderate-sized coefficients on several of the variables. In addition, the first component of the covariance matrix accounts for almost 99% of the total variance of the observed variables. The components of the covariance matrix are completely dominated by the fact that the variance of the plate variable is 100 times larger than the variance of any of the other seven variables. Consequently, the principal components from the covariance matrix simply reflect the order of sizes of the variances of

TABLE 10.2

Principal Components of the Covariance Matrix of the Blood Chemistry Data

	Variances, etc. of Components							
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Variance	1704.68	15.06	6.98	2.64	0.13	0.07	0.00	0.00
Proportion of variance	0.986	0.0087	0.00404	0.00153	0.000	0.000	0.000	0.000
Cumulative proportion	0.986	0.9943	0.99836	0.99989	1.00	1.00	1.00	1.000
Component Loadings								
rBlood	0.000	0.000	0.000	0.000	0.943	0.329	0.000	0.000-
Plate	-0.999	0.000	0.000	0.000	0.000	0.000	0.000	0.000
wBlood	0.000	-0.192	0.000	-0.981	0.000	0.000	0.000	0.000
Neut.	0.000	0.000	0.000	0.000	0.000	0.000	0.758	0.650
Lymph	0.000	0.000	0.000	0.000	0.000	0.000	-0.649	0.760
Bilir.	0.000	0.961	0.195	-0.191	0.000	0.000	0.000	0.000
Sodium	0.000	0.193	-0.979	0.000	0.000	0.000	0.000	0.000
Potass.	0.000	0.000	0.000	0.000	0.329	-0.942	0.000	0.000

TABLE 10.3

Principal Components of the Correlation Matrix of the Blood Chemistry Data

	Variances, etc. of the Components							
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Variances	2.792	1.532	1.249	0.778	0.622	0.489	0.436	0.102
Proportion of variance	0.349	0.191	0.156	0.0973	0.0777	0.0611	0.0545	0.0128
Cumulative proportion	0.349	0.540	0.697	0.7939	0.8716	0.9327	0.9872	1.0000
Component Loadings								
rBlood	-0.194	0.417	0.400	0.652	0.175	-0.363	0.176	0.102
Plate	-0.400	0.154	0.168	0.000	-0.848	0.230	-0.110	0.000
wBlood	-0.459	0.000	0.168	-0.274	0.251	0.403	0.677	0.000
Neut.	-0.430	-0.472	-0.171	0.169	0.118	0.000	-0.237	0.678
Lymph	0.494	0.360	0.000	-0.180	-0.139	0.136	0.157	0.724
Bilir.	0.319	-0.320	-0.277	0.633	-0.162	0.384	0.377	0.000
Sodium	0.177	-0.535	0.410	-0.163	-0.299	-0.513	0.367	0.000
Potass.	0.171	-0.245	0.709	0.000	0.198	0.469	-0.376	0.000

the observed variables. The results from the correlation matrix tell us, in particular, that a weighted contrast of the first four and last four variables is the linear function with the largest variance. This example illustrates that, when variables are on very different scales or have very different variances, a PCA of the data should be performed on the correlation matrix, not on the covariance matrix.

10.5 Principal Components of Bivariate Data with Correlation Coefficient r

Before we move on to look at some practical examples of the application of PCA, it will be helpful to look in a little more detail at the mathematics of the method in one very simple case. We will do this in Technical Section 10.2, using bivariate data in which the two variables x_1 and x_2 have correlation coefficient r .

Technical Section 10.2: Principal Components of Bivariate Data

Suppose we have just two variables x_1 and x_2 , measured on a sample of individuals, with sample correlation matrix given by

$$\mathbf{R} = \begin{pmatrix} 1.0 & r \\ r & 1.0 \end{pmatrix}$$

In order to find the principal components of the data r , we need to find the eigenvalues and eigenvectors of \mathbf{R} . The eigenvalues are roots of the equation

$$|\mathbf{R} - \lambda\mathbf{I}| = 0$$

where the vertical lines indicate the determinant of the matrix enclosed by them. This leads to the following quadratic equation in λ :

$$(1 - \lambda)^2 - r^2 = 0$$

which has roots (eigenvalues) $\lambda_1 = 1 + r$, $\lambda_2 = 1 - r$. The first component has variance $1 + r$, and the second has variance $1 - r$. Note that the sum of the eigenvalues is 2, equal to trace (\mathbf{R}), that is, the sum of the elements on the main diagonal. The eigenvector corresponding to λ_1 is obtained by solving the equation

$$\mathbf{R}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$$

This leads to the equations

$$a_{11} + ra_{12} = (1 + r)a_{11}$$

$$ra_{11} + a_{12} = (1 + r)a_{12}$$

The two equations are identical, and both reduce to $a_{11} = a_{12}$. If we now introduce the normalization constraint $\mathbf{a}_1' \mathbf{a}_1 = 1$, we find that

$$a_{11} = a_{12} = \frac{1}{\sqrt{2}}$$

Similarly, we find the elements of the second eigenvector as $a_{21} = 1/\sqrt{2}$ and $a_{22} = -1/\sqrt{2}$. The two principal components are then given by

$$y_1 = \frac{1}{\sqrt{2}}(x_1 + x_2)$$

$$y_2 = \frac{1}{\sqrt{2}}(x_1 - x_2)$$

Notice that if $r < 0$, the order of the eigenvalues, and hence that of the principal components, is reversed; if $r = 0$, the eigenvalues are both equal to 1, and any two solutions at right angles could be chosen to represent the two components.

Three further points:

- There is an arbitrary sign in the choice of the elements of \mathbf{a}_i ; it is customary to choose a_{i1} to be positive.

- The components do not depend on r , although the proportion of variance explained by each does change with r . As r tends to 1, the proportion of variance accounted for by y_1 , namely, $(1+r)/2$, also tends to 1.
 - When $r = 1$ the points all lie on a straight line, and the variation in the data is unidimensional.
-
-

10.6 Rescaling the Principal Components

The coefficients defining the principal components derived as described in [Section 10.5](#) are often rescaled so that they are correlations or covariances between the original variables and the derived components. The rescaled coefficients are often more useful in interpreting a PCA. The covariance of variable i with component j is given by

$$\text{Cov}(x_i, y_j) = \lambda_j a_{ji}$$

The correlation of variable x_i with component y_j is therefore

$$\begin{aligned} r_{x_i, y_j} &= \frac{\lambda_j a_{ji}}{\sqrt{\text{Var}(x_i)\text{Var}(y_j)}} \\ &= \frac{\lambda_j a_{ji}}{s_i \sqrt{\lambda_j}} = \frac{a_{ji} \sqrt{\lambda_j}}{s_i} \end{aligned}$$

If the components are extracted from the correlation matrix rather than the covariance matrix, the correlation between variable and component becomes

$$r_{x_i, y_i} = a_{ji} \sqrt{\lambda_j}$$

because in this case the standard deviation s_i is 1. (Although for convenience we have used the same nomenclature for the eigenvalues and the eigenvectors extracted from the covariance matrix or the correlation matrix, they will, of course, not be equal.)

The rescaled coefficients from a PCA of a correlation matrix are analogous to factor loadings as we shall see in Chapter 11. It is often these rescaled coefficients that are presented as the results of a PCA and used in interpretation.

10.7 How the Principal Components Predict the Observed Covariance Matrix

In Technical Section 10.3, we will look at how the principal components reproduce the observed covariance or correlation matrix from which they were extracted.

Technical Section 10.3: How Principal Components Reproduce the Sample Covariance Matrix

To begin, let the initial vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_q$ that define the principal components be used to form a $q \times q$ matrix, $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_q]$; these are vectors extracted from the covariance matrix \mathbf{S} and scaled so that $\mathbf{a}_i' \mathbf{a}_i = 1$. Arrange the eigenvalues $\lambda_1, \dots, \lambda_q$ along the main diagonal of a diagonal matrix, $\mathbf{\Lambda}$. Then, it can be shown that the covariance matrix of the observed variables x_1, x_2, \dots, x_q is given by

$$\mathbf{S} = \mathbf{A} \mathbf{\Lambda} \mathbf{A}'$$

Rescaling the vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_q$ so that the sum of squares of their elements is equal to the corresponding eigenvalue, that is, calculating $\mathbf{a}_i^* = \lambda_i^{1/2} \mathbf{a}_i$, allows \mathbf{S} to be written more simply as

$$\mathbf{S} = \mathbf{A}^* (\mathbf{A}^*)'$$

where $\mathbf{A}^* = [\mathbf{a}_1^* \cdots \mathbf{a}_q^*]$

If the matrix \mathbf{A}_m^* is formed from, say, the first m components rather than from all q , then $\mathbf{A}_m^* (\mathbf{A}_m^*)'$ gives the predicted value of \mathbf{S} based on these m components. It is often useful to calculate the predicted value based on the number of components considered to adequately describe the data. How this number might be chosen is considered in Section 10.8

10.8 Choosing the Number of Components

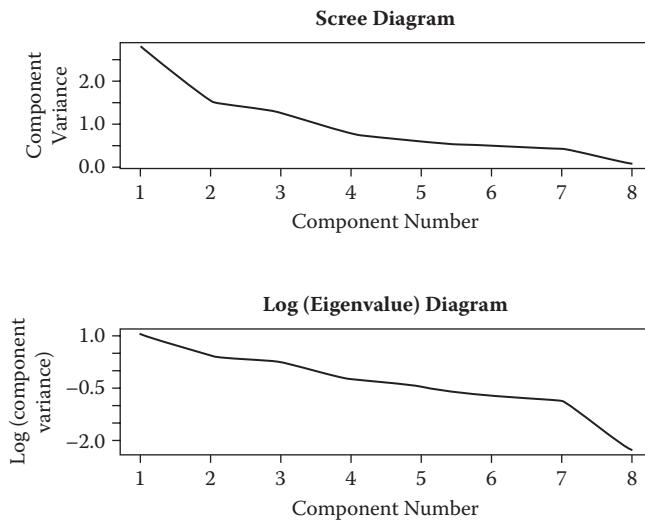
As described earlier, PCA is seen to be a technique for transforming a set of observed variables into a new set of variables that are uncorrelated with one another. The variation in the original q variables is only completely accounted for by all q principal components. The usefulness of these transformed variables, however, stems from their property of accounting for the variance in decreasing proportions. The first component, for example, accounts for the maximum amount of variation possible for any linear combination of the original variables. But, how useful is this artificial variate constructed from

the observed variables? To answer this question, we would first need to know the proportion of the total variance of the original variables for which it accounted. If, for example, 80% of the variation in a multivariate data set involving six variables could be accounted for by a simple weighted average of the variable values, then almost all the variation can be expressed along a single continuum rather than in six-dimensional space. The PCA would have provided a highly parsimonious summary (reducing the dimensionality of the data from six to one) that might be useful in later analysis.

So, the question we need to ask is how many components are needed to provide an adequate summary of a given data set? A number of informal and more formal techniques are available. Here we shall concentrate on the former; examples of the use of formal inferential methods are given in Jolliffe (2002) and Rencher (2002).

The most common of the relatively ad hoc procedures that have been suggested for deciding on the number of components to retain are the following:

- Retain just enough components to explain some specified, large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as q or n , the sample size, increases.
- Exclude those principal components whose eigenvalues are less than the average, that is, $\sum_{i=1}^q \lambda_i / q$. Since $\sum_{i=1}^q \lambda_i = \text{trace}(\mathbf{S})$ the average eigenvalue is also the average variance of the original variables. This method then retains those components that account for more variance than the average for the observed variables.
- When the components are extracted from the correlation matrix, $\text{trace}(\mathbf{R}) = q$, and the average variance is, therefore, 1; so, applying the rule in the previous bullet point, components with eigenvalues less than 1 will be excluded. This rule was originally suggested by Kaiser (1958), but Jolliffe (1972), on the basis of a number of simulation studies, proposed that a more appropriate procedure would be to exclude components extracted from a correlation matrix whose associated eigenvalues are less than 0.7.
- Cattell (1965) suggested examination of the plot of λ_i against i , the so-called *scree diagram*. The number of components selected is the value of i corresponding to an “elbow” in the curve, that is, a change of slope from “steep” to “shallow.” In fact, Cattell was more specific than this; he recommended looking for a point on the plot beyond which the scree diagram defines a more or less straight line, not necessarily horizontal. The first point on the straight line is then taken as the last component to be retained. Further, it should also be remembered that Cattell suggested the scree diagram in the context of factor analysis rather than as applied to PCA.

**FIGURE 10.1**

Scree diagram and log-eigenvalue diagram for the principal components of the correlation matrix of the blood chemistry data.

- A modification of the scree diagram described by Jolliffe (1989) is the log-eigenvalue diagram consisting of a plot of $\log(\lambda_i)$ against i .

Returning to the results of the PCA of the correlation matrix of the blood chemistry data given in [Section 10.4](#), we find that the first four components account for nearly 80% of the total variance, but it takes a further two components to push this figure up to 90%. A cutoff of 1 for the eigenvalues leads to retaining three components, and with a cutoff of 0.7, four components are kept. Figure 10.1 shows the scree diagram and the log-eigenvalue diagram for the data. The former plot may suggest four components, although this is fairly subjective, and the latter seems to be of little help here because it appears to indicate retaining seven components, which is hardly much of a dimensionality reduction. The example illustrates that the proposed methods for deciding how many components to keep can (and often do) lead to different conclusions.

10.9 Calculating Principal Component Scores

If we decide that we need, say, m principal components to adequately represent our data (using one or other of the methods described in [Section 10.8](#)), then we will generally wish to calculate the scores on each of these components for each individual in our sample. How we do this is described in Technical [Section 10.4](#).

Technical Section 10.4: Calculating Principal Component Scores

First, let us assume that we have derived the components from the covariance matrix \mathbf{S} . The m principal component scores for individual i with original $q \times 1$ vector of variable values \mathbf{x}_i are obtained as

$$y_{i1} = \mathbf{a}'_1 \mathbf{x}_i$$

$$y_{i2} = \mathbf{a}'_2 \mathbf{x}_i$$

⋮

$$y_{im} = \mathbf{a}'_m \mathbf{x}_i$$

If the components are derived from the correlation matrix, then \mathbf{x}_i would contain the i th individual's standardized scores for each variable.

The principal component scores calculated as shown here have variances equal to λ_j for $j=1, \dots, m$. Many investigators might prefer to have scores with means equal to 0 and variances equal to 1. Such scores can be found as follows:

$$\mathbf{z} = \mathbf{\Lambda}_m^{-1} \mathbf{A}'_m \mathbf{x}$$

where $\mathbf{\Lambda}_m$ is an $m \times m$ diagonal matrix with $\lambda_1, \lambda_2, \dots, \lambda_m$ on the main diagonal, $\mathbf{A}_m = [\mathbf{a}_1, \dots, \mathbf{a}_m]$, and \mathbf{x} is the $q \times 1$ vector of standardized scores. We should note here that the first m principal component scores are the same whether we retain all possible q components or just the first m . As we shall see in Chapter 11, this is not the case with the calculation of factor scores.

10.10 Some Examples of the Application of PCA

In this section, we will look at the application of PCA to a number of data sets, beginning with one involving only two variables as this allows us to illustrate graphically an important point about this type of analysis.

10.10.1 Head Size of Brothers

The data in [Table 10.4](#) give the head lengths (in millimeters) for each of the first two adult sons in 25 families. The mean vector and covariance matrix of the data are

$$\bar{\mathbf{x}}' = [185.72, 183.84] \quad \mathbf{S} = \begin{bmatrix} 95.29 & 69.66 \\ 69.66 & 100.81 \end{bmatrix}$$

TABLE 10.4
Head Lengths (in Millimeters) of First
and Second Sons of 25 Families

Family	First Son	Second Son
1	191	179
2	195	201
3	181	185
4	183	188
5	176	171
6	208	192
7	189	190
8	197	189
9	188	197
10	192	187
11	179	186
12	183	174
13	174	185
14	190	195
15	188	187
16	163	161
17	195	183
18	186	173
19	181	182
20	175	165
21	192	185
22	174	178
23	176	176
24	197	200
25	190	187

The principal components of these data extracted from their covariance matrix are

$$y_1 = 0.693x_1 + 0.721x_2 \quad y_2 = -0.721x_1 + 0.693x_2$$

with variances 167.77 and 28.33. The first principal component accounts for a proportion $167.77/(167.77 + 28.33) = 0.86$ of the total variance in the original variables. Note that the total variance of the principal components is 196.10, which, as expected, is equal to the total variance of the original variables found by adding the relevant terms in the covariance matrix above, that is, $95.29 + 100.81 = 196.10$.

How should the two derived components be interpreted? The first component is essentially the sum of the head lengths of the two sons, and the second component is the difference in head lengths. Perhaps we can label the first component “size” and the second component “shape,” but see [Section 10.10.2](#) for some comments about trying to give principal components such labels.

To calculate an individual’s score on a component, we simply multiply the variable values, subtract the appropriate mean by the loading for the variable, and add these values over all variables. We can illustrate this calculation using the data for the first family, in which the head length of the first son is 191 mm and the second son, 179 mm. The score for this family on the first principal component is calculated as

$$0.693 \times (191 - 185.72) + 0.721 \times (179 - 183.84) = 0.169$$

and on the second component the score is

$$-0.721 \times (191 - 185.72) + 0.693 \times (179 - 183.84) = -7.61$$

The variance of the first principal component scores will be 167.77, and that of the second principal component scores will be 28.33.

We can plot the data showing the axes corresponding to the principal components. The first axis passes through the mean of the data and has a slope 0.721/0.693, and the second axis also passes through the mean and has a slope $-0.693/0.721$. The plot is shown in Figure 10.2. This example illustrates

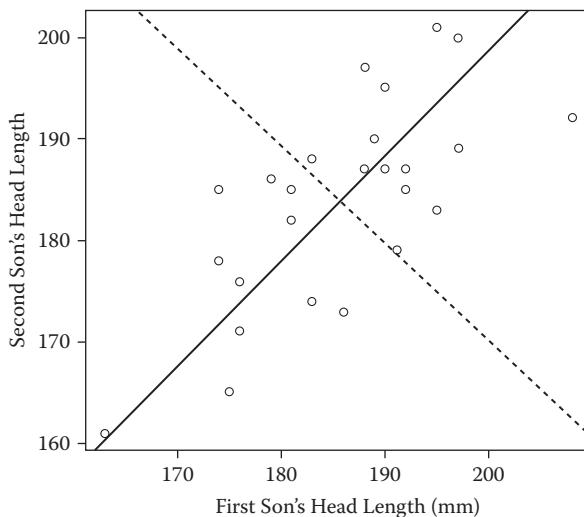
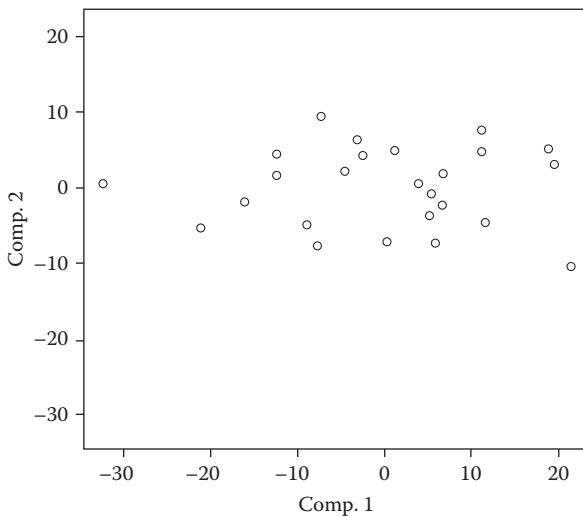


FIGURE 10.2

Head lengths of the first and second sons of 25 families, showing axes corresponding to the principal components of the sample covariance matrix of the data.

**FIGURE 10.3**

Plot of the first two principal component scores for the head size data.

that a PCA is essentially simply a rotation of the axes of the multivariate data scatter. Further, we can also plot the principal component scores to give Figure 10.3. (Note that in this figure, the range of the x-axis and the range of the y-axis have been made the same so that the larger variance of the first principal component is clearly shown.)

We can use the PCA of the head size data to demonstrate how the principal components reproduce the observed covariance matrix. We first need to rescale the principal components we have at this point by multiplying them by the square roots of their respective variances to give the new components:

$$y_1 = 12.952[0.693 x_1 + 0.721 x_2], \text{ that is, } y_1 = 8.976 x_1 + 9.338 x_2$$

and

$$y_2 = 5.323[-0.721 x_1 + 0.693 x_2], \text{ that is, } y_2 = -3.837 x_1 + 3.688 x_2$$

leading to the matrix \mathbf{A}_2^* as defined in [Section 10.7](#):

$$\mathbf{A}_2^* = \begin{bmatrix} 8.976 & -3.837 \\ 9.338 & 3.688 \end{bmatrix}$$

TABLE 10.5

Crime Rates in the United States

State	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
ME	2.0	14.8	28	102	803	2347	164
NH	2.2	21.5	24	92	755	2208	228
VT	2.0	21.8	22	103	949	2697	181
MA	3.6	29.7	193	331	1071	2189	906
RI	3.5	21.4	119	192	1294	2568	705

Multiplying this matrix by its transpose should recreate the covariance matrix of the head length data; performing the matrix multiplication shows that it does recreate \mathbf{S} :

$$\mathbf{A}_2^*(\mathbf{A}_2^*)' = \begin{bmatrix} 95.29 & 69.66 \\ 69.66 & 100.81 \end{bmatrix}$$

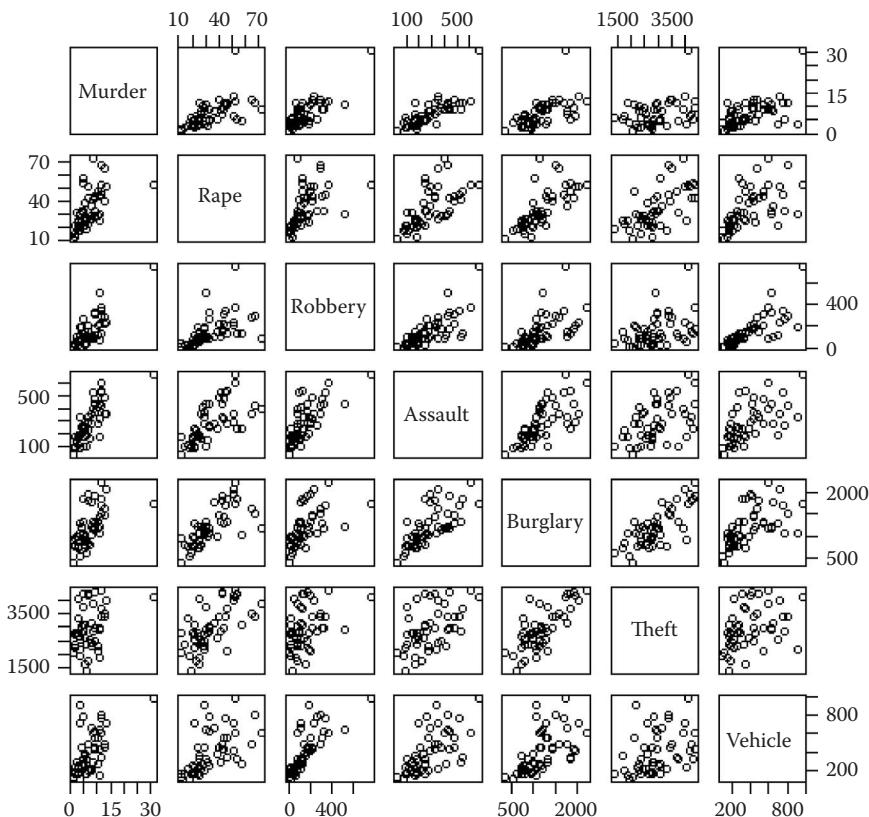
As an exercise, readers might like to find the predicted covariance matrix using only the first component.

The head size example has been useful for discussing some aspects of PCA, but it is not, of course, typical of multivariate data sets encountered in behavioral research, where many more than two variables will be recorded for each individual in a study. In the following two subsections, we consider some more realistic examples.

10.10.2 Crime Rates in the United States

The *Statistical Abstract of the USA* (1988) gives rates of different types of crime per 100,000 residents of 50 states of the United States plus the District of Columbia for the year 1986. The data for the first five states are given in Table 10.5. We shall use PCA to explore these data, but to start, it will be useful to look at the scatterplot matrix of the seven types of crime, and this is given in Figure 10.4. The plot shows that the relationships between crime rates are of varying strengths and that there are a number of outlying states in some of the panels, for example, the one for rape and murder rates. We shall ignore the outlier problem in the following analysis, but readers are encouraged to see how the results that follow are altered if the outliers are removed (see Exercise 10.1). In this example, the variables are all on the same scale, crimes per 100,000 of the resident population of a state, but if we look at the variances of each crime rate,

Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
23.20	212.31	18993.37	22004.31	177912.83	582812.84	50007.37

**FIGURE 10.4**

Scatterplot matrix of crime rate data.

we see that they are very different, and the results from a PCA on the unstandardized variables would be swamped by those variables with the largest variances. Consequently, we will apply the analysis to the standardized variables, that is, components will be extracted from the correlation rather than the covariance matrix.

The results of the PCA on the crime rate data are shown in [Table 10.6](#). A scree plot of variances of principal components is shown in [Figure 10.5](#). Only the variance of the first component is greater than 1, although that of the second component is very close to 1. The scree plot suggests that perhaps two components might be adequate to describe these data. Many users of PCA search for an interpretation of the derived components that allow them to be “labeled” in some sense. This requires examining the coefficients defining each component; in [Table 10.6](#), the coefficients are scaled so that their sums of squares equal 1; “—” indicates near-zero values. (Remember also that the signs of the coefficients are arbitrary in the sense that the minus signs and

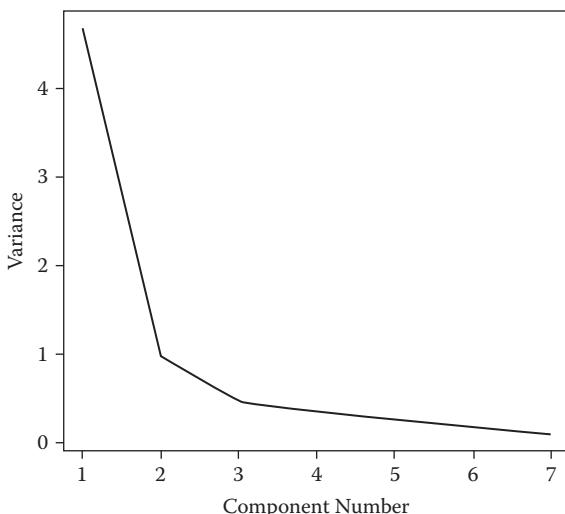
TABLE 10.6

Results from a PCA of the Correlation Matrix of Crime Rate Data

	Importance of Components						
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Variance	4.69	0.99	0.46	0.34	0.24	0.18	0.09
Proportion of variance	0.67	0.14	0.07	0.05	0.03	0.03	0.01
Cumulative proportion	0.67	0.81	0.88	0.93	0.96	0.99	1.00
Coefficients Defining Components							
Murder	-0.381	-0.350	-0.538	—	-0.274	0.370	0.480
Rape	-0.377	0.279	—	-0.830	-0.250	—	-0.151
Robbery	-0.391	-0.420	0.131	0.275	-0.387	—	-0.651
Assault	-0.410	-0.124	-0.335	—	0.564	-0.620	—
Burglary	-0.394	0.367	—	0.162	0.466	0.622	-0.283
Theft	-0.321	0.628	—	0.449	-0.388	-0.282	0.256
Vehicle	-0.366	-0.282	0.758	—	0.163	—	0.422

Note: — indicates a coefficient that is almost zero.

positive signs could be reversed without altering the structure or the interpretation of the components.) Examining the coefficients defining the principal components in Table 10.6, we see that the first component might be regarded as some index of overall crime rate in a state, with states that have

**FIGURE 10.5**

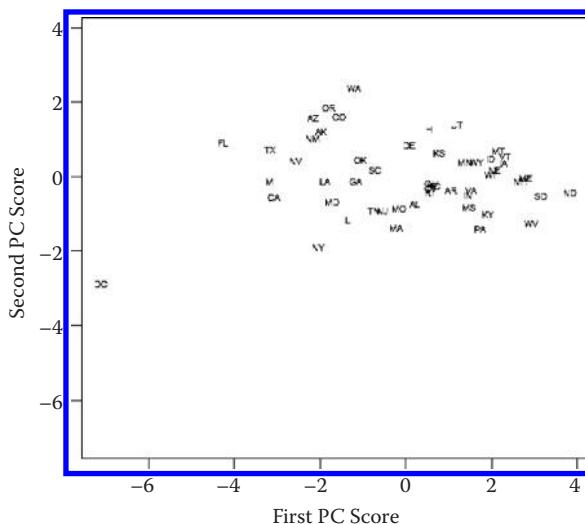
Scree plots of variances of principal components in the correlation matrix of crime rate data.

larger crime rates having larger negative scores on this component (negative because of the minus signs attached to each loading); perhaps we could label this component “dangerousness.” Labeling the second component is more difficult; in this component, the coefficients have a mixture of positive and negative signs, and the component appears to contrast “property crimes,” that is, larceny and burglary, with crimes against the person, that is, robbery and murder. A not very inspired label might be “property versus person.” The other components have very small variances, and we shall not try to interpret them.

Attempting to label components in this way is not without its critics; the following quotation from Marriott (1974) should act as a salutary warning about the dangers of overinterpretation.

It must be emphasized that no mathematical method is, or could be, designed to give physically meaningful results. If a mathematical expression of this sort has an obvious physical meaning, it must be attributed to a lucky chance, or to the fact that the data have a strongly marked structure that shows up in analysis. Even in the latter case, quite small sampling fluctuations can upset the interpretation; for example, the first two principal components may appear in reverse order or may become confused altogether. Reification then requires considerable skill and experience if it is to give a true picture of the physical meaning of the data.

Perhaps a more suitable use for the principal components of the crime rate data is as the basis of various graphical displays of cities. In fact, this is often the most useful aspect of a PCA and as a means to providing informative “views” of multivariate data, PCA has the advantage of making it less urgent or tempting to try to interpret and label the components. The first few component scores provide a low-dimensional “map” of the observations in which the Euclidean distances (see Chapter 12) between the points representing the individuals best approximate in some sense the Euclidean distances between the individuals based on the original variables (see Everitt, 2005, for more details). A plot of the crime rate data in the space of the first two principal components showing state labels is given in [Figure 10.6](#). Clearly, DC is not a place to live in by choice! On the second component, WA has a high score because it has a low murder and robbery rate and relatively high burglary and larceny rates; contrast NY, which has relatively high murder and robbery rates and relatively low burglary and larceny rates and has a low score on the second component. Because the first two components account for over 80% of the variance in the crime rates, the two-dimensional plot in [Figure 10.6](#) gives a very accurate description of the original seven-variable data. This claim is backed up by comparing the observed correlation matrix with the correlation matrix “predicted” by the two-component solution after it has been rescaled, as described in [Section 10.7](#). The two matrices are shown in [Table 10.7](#). Corresponding elements of the two matrices are quite similar.

**FIGURE 10.6**

Plot of crime rate data in the space of the first two principal components showing state labels.

TABLE 10.7

Observed Correlation Matrix for Crime Rate Data Compared to Matrix Constructed from the Two-Component Solution

	Observed						
	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
Murder	1.000	0.578	0.804	0.781	0.581	0.361	0.573
Rape	0.578	1.000	0.530	0.659	0.721	0.635	0.569
Robbery	0.804	0.530	1.000	0.740	0.551	0.400	0.786
Assault	0.781	0.659	0.740	1.000	0.710	0.512	0.638
Burglary	0.581	0.721	0.551	0.710	1.000	0.764	0.579
Theft	0.361	0.635	0.400	0.512	0.764	1.000	0.386
Vehicle	0.573	0.569	0.786	0.638	0.579	0.386	1.000
Predicted							
Murder	0.801	0.578	0.843	0.775	0.578	0.357	0.752
Rape	0.578	0.745	0.576	0.691	0.799	0.741	0.571
Robbery	0.843	0.576	0.889	0.802	0.571	0.328	0.788
Assault	0.775	0.691	0.802	0.803	0.713	0.540	0.739
Burglary	0.578	0.799	0.571	0.713	0.862	0.821	0.576
Theft	0.357	0.741	0.328	0.540	0.821	0.872	0.377
Vehicle	0.752	0.571	0.788	0.739	0.576	0.377	0.708

10.10.3 Drug Usage by American College Students

The majority of adult and adolescent Americans regularly use psychoactive substances and often do so for a substantial proportion of their lifetime. Various forms of licit and illicit psychoactive substance use are prevalent, suggesting that patterns of psychoactive substance taking are a major part of the individual's behavioral repertory and have pervasive implications on the performance of other behaviors. In an investigation of these phenomena, Huba et al. (1981) collected data on drug usage rates for 1634 students in the seventh to ninth grades in 11 schools in the greater metropolitan area of Los Angeles. Each participant completed a questionnaire about the number of times a particular substance had ever been used. The substances asked about were as follows:

1. Cigarettes
2. Beer
3. Wine
4. Liquor
5. Cocaine
6. Tranquillizers
7. Drugstore medications used to get high
8. Heroin and other opiates
9. Marijuana
10. Hashish
11. Inhalants (glue, gasoline, etc.)
12. Hallucinogenics (LSD, mescaline, etc.)
13. Amphetamine stimulants

Responses were recorded on a five-point scale:

1. Never tried
2. Only once
3. A few times
4. Many times
5. Regularly

The correlations between the usage rates of the 13 substances are shown in [Table 10.8](#).

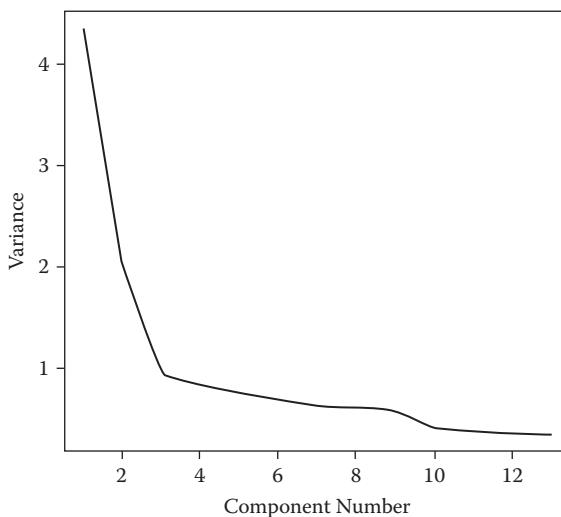
Applying a PCA to this correlation matrix gives the results shown in [Table 10.9](#). A scree plot of the variances of the components is shown in [Figure 10.7](#). The first two components have variances greater than 1, and the third has a variance slightly lower than 1. The scree plot suggests that

TABLE 10.8
Correlation Matrix for Drug Usage Data

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	0.447	1											
3	0.442	0.619	1										
4	0.435	0.604	0.583	1									
5	0.114	0.068	0.053	0.115	1								
6	0.203	0.146	0.139	0.258	0.349	1							
7	0.091	0.103	0.110	0.122	0.209	0.221	1						
8	0.082	0.063	0.066	0.097	0.321	0.355	0.201	1					
9	0.513	0.445	0.365	0.482	0.186	0.315	0.150	0.154	1				
10	0.304	0.318	0.240	0.368	0.303	0.377	0.163	0.219	0.534	1			
11	0.245	0.203	0.183	0.255	0.272	0.323	0.310	0.288	0.301	0.302	1		
12	0.101	0.088	0.074	0.139	0.279	0.367	0.232	0.320	0.204	0.368	0.304	1	
13	0.245	0.199	0.184	0.293	0.278	0.545	0.232	0.314	0.394	0.467	0.392	0.511	1

TABLE 10.9
Results of Applying PCA to the Correlation Matrix of Drug Usage Rates

	Variances of Components												
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
Variance	4.38	2.05	0.96	0.81	0.76	0.69	0.63	0.62	0.57	0.40	0.39	0.38	0.36
Proportion of variance	0.34	0.16	0.07	0.06	0.059	0.053	0.050	0.048	0.044	0.031	0.030	0.029	0.027
Cumulative proportion	0.34	0.49	0.57	0.63	0.69	0.74	0.79	0.84	0.88	0.91	0.94	0.97	1.00
Coefficients Defining Components													
Var	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
1	-0.280	-0.283	—	—	-0.300	-0.387	-0.124	0.137	0.655	-0.139	-0.136	-0.169	-0.263
2	-0.287	-0.394	0.120	—	0.187	0.161	0.114	—	—	—	—	0.695	-0.410
3	-0.267	-0.393	0.207	-0.139	0.309	0.141	—	—	0.107	-0.421	0.210	-0.188	0.564
4	-0.318	-0.322	—	—	0.181	0.142	—	-0.164	-0.214	0.563	-0.181	-0.519	-0.219
5	-0.208	0.290	—	-0.582	-0.432	0.416	0.185	-0.244	0.204	—	0.154	—	—
6	-0.293	0.262	-0.165	—	0.122	—	-0.629	-0.399	—	-0.124	-0.421	0.170	0.138
7	-0.176	0.190	0.723	0.372	-0.178	0.277	-0.309	0.253	—	—	—	—	—
8	-0.201	0.317	0.153	-0.534	0.327	-0.359	—	0.525	-0.169	—	—	—	—
9	-0.340	-0.160	-0.228	0.112	-0.365	-0.129	—	0.285	-0.149	0.418	0.154	0.285	0.502
10	-0.329	—	-0.352	0.125	-0.256	0.243	0.167	0.274	-0.400	-0.496	-0.187	-0.240	-0.152
11	-0.274	0.163	0.330	0.159	-0.152	-0.531	0.466	-0.417	-0.228	—	—	—	—
12	-0.245	0.327	-0.144	0.272	0.379	0.210	0.413	0.162	0.440	0.179	-0.308	—	0.159
13	-0.328	0.235	-0.235	0.267	0.203	—	-0.132	-0.177	—	—	0.733	—	-0.269

**FIGURE 10.7**

Scree plot of component variances for the drug usage data.

perhaps three components need to be used to describe the correlations for these data, although these three components account for only 57% of the variance in drug usage rates. The first component is clearly a measure of overall drug usage, as might be expected since all the coefficients have relatively similar numerical values and all have the same sign. The second component contrasts “legal” with “illegal” substances (with the exception of marijuana, which has the same sign as the legal substances). This component might be seen as contrasting “soft” and “hard” drug usage. So, after overall usage has been accounted for, the main source of variation is between the different patterns of consumption of soft and hard drugs. The third component is essentially a contrast of drugstore and inhalant substance usage on the one hand, with marijuana, hashish, and amphetamine usage on the other.

We will return to the drug usage data in Chapter 11.

10.11 Using PCA to Select a Subset of the Variables

Although the first few principal component scores for each individual may provide a very useful summary for a set of multivariate data, all the original variables are needed in their computation. In many cases, an investigator might be happier with determining a subset of the original variables that contains, in some sense, virtually all the information contained in the complete set of these variables. In a series of papers, Jolliffe (1970, 1972, 1973) discusses a number of approaches to selecting subsets of variables, several of

which are based on PCA. One such method is to first use one or other of the criteria for choosing the number of components described in [Section 10.8](#); this number, say, m , is taken to indicate the effective dimensionality of the data and will be the size of the subset of the original variables to be retained. The variables are then chosen, one associated with each of the first m components and having the largest absolute coefficient value on the component but not already having been selected. If we use the eigenvalue criterion suggested by Jolliffe, namely, retain components with eigenvalues greater than 0.7 (assuming the correlation matrix is being used) for the drug usage data, then we keep the first five components. Examining the coefficients defining these components, we find that we are led to the following five variables:

1. Marijuana usage
2. Beer usage
3. Drugstore usage
4. Cocaine usage
5. Hallucinogenic usage

These variables could be used in future analyses of the data with little loss of information, compared to using all 13 of the original variables.

10.12 Summary

- PCA provides a way of reducing the complexity of multivariate data by reducing their dimensionality.
- The reduction in dimensionality that can often be achieved by a PCA is possible only if the original variables are correlated; if the original variables are independent of one another, a PCA cannot lead to any simplification.
- In most applications, variables will be on different scales, so components will need to be extracted from the correlation matrix of the data.
- In essence, PCA is simply a rotation of the axes of multivariate data scatter.
- The first few principal component scores can often be used to provide a convenient summary of a multivariate data set, particularly for looking at the data via simple scatterplots.
- Trying to give meaningful labels to components is often confusing and a waste of time.
- Two techniques that are related to principal components but that are not covered in this book are correspondence analysis and multidimensional scaling (for details, see Everitt, 2005).

10.13 Exercises

- 10.1 The crime rate data considered in the text contains a number of possible outliers. Reanalyze the data using principal components after removing the observations you consider to be outliers, and compare your results with those given in the text.
- 10.2 Find the principal components of the following correlation matrix and compare how the one- and two-component solutions reproduce the matrix.

$$\mathbf{R} = \begin{bmatrix} 1.0000 & & \\ 0.6579 & 1.0000 & \\ 0.0034 & -0.0738 & 1.0000 \end{bmatrix}$$

- 10.3 MacDonnell (1902) obtained measurements on seven physical characteristics for each of 300 criminals. The seven variables measured were (1) head length, (2) head breadth, (3) face breadth, (4) left finger length, (5) left forearm length, (6) left foot length, and (7) height. The correlation matrix calculated by MacDonnell is

	1	2	3	4	5	6	7
1	1.000						
2	0.402	1.000					
3	0.396	0.618	1.000				
4	0.301	0.150	0.321	1.000			
5	0.305	0.135	0.289	0.846	1.000		
6	0.339	0.206	0.363	0.759	0.797	1.000	
7	0.340	0.183	0.345	0.661	0.800	0.736	1.000

Find the principal components of this correlation matrix and interpret the results.

- 10.4 The data in exer_104.txt give prestige, income, education, and suicide rates for 36 occupations, originally given in Labovitz (1970). Undertake a PCA of the data and use the results to try to answer the question of whether there are several distinct types of jobs.
- 10.5 Rescale the coefficients defining the principal components of the crime rate data so that they represent correlations between the components and crime rates.

11

Factor Analysis

11.1 Introduction

In many areas of psychology, and other disciplines in the behavioral sciences, it is often not possible to measure directly the concepts of primary interest. Two obvious examples are intelligence and social class. In such cases, the researcher is forced to examine the concepts indirectly by collecting information on variables that can be measured or observed directly, and which may also realistically be assumed to be indicators, in some sense, of the concepts of real interest. The psychologist who is interested in an individual's "intelligence," for example, may record examination scores in a variety of different subjects in the expectation that these scores are dependent in some way on what is widely regarded as intelligence but are also subject to random errors. Further, a sociologist, say, concerned with people's "social class," might pose questions about a person's occupation, educational background, home ownership, etc., on the assumption that these do reflect the concept in which he or she is really interested.

Both intelligence and social class are what are generally referred to as latent variables, that is, concepts that cannot be measured directly but can be assumed to relate to a number of measurable or manifest variables. The method of analysis most generally used to help uncover the relationships between the assumed latent variables and the manifest variables is factor analysis. The model on which the method is based is essentially that of multiple linear regression, except that now the manifest variables are regressed on the unobservable latent variables (often referred to in this context as common factors), so that direct estimation of the corresponding regression coefficients (known now as factor loadings) is not possible.

A point to be made at the outset is that factor analysis comes in two distinct varieties; the first is exploratory factor analysis, which is used to investigate the relationship between manifest variables and factors without making any assumptions about which manifest variables are related to which factors, and the second is confirmatory factor analysis, which is used to test whether a specific factor model postulated *a priori* provides an adequate fit for the covariances or correlations between the manifest variables. In this chapter,

we shall consider both approaches, although the discussion of confirmatory factor analysis will be somewhat brief.

11.2 The Factor Analysis Model

The basis of factor analysis is a regression model, linking the manifest variables to a set of unobserved (and unobservable) latent variables. In essence, the model assumes that the observed relationships between the manifest variables (as measured by their covariances or correlations) are the results of the relationships of these variables to the latent variables. A relatively brief account of the factor analysis model is given in Technical [Section 11.1](#).

Technical [Section 11.1](#): Factor Analysis Model

The q observed or manifest variables are represented as the vector $\mathbf{x}' = [x_1, x_2, \dots, x_q]$ and are all assumed, for convenience, to have zero mean (it is only information about the relationships between the manifest variables as contained in their covariance or correlation matrix that is of interest in factor analysis so the zero means assumption is of no consequence). The manifest variables are assumed to be related by a regression-type model to a smaller number of unobserved latent variables, the common factors, represented as the vector $\mathbf{f}' = [f_1, f_2, \dots, f_k]$ where $k < q$. We can write the assumed model as a series of regression-like equations:

$$x_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + \dots + \lambda_{1k}f_k + u_1$$

$$x_2 = \lambda_{21}f_1 + \lambda_{22}f_2 + \dots + \lambda_{2k}f_k + u_2$$

⋮

$$x_q = \lambda_{q1}f_1 + \lambda_{q2}f_2 + \dots + \lambda_{qk}f_k + u_q$$

The λ_{ij} values are essentially regression coefficients showing how each x_i depends on the k common factors; in this context, they are known as factor loadings. When estimated from a sample correlation matrix, the factor loadings are the estimated correlations between factors and manifest variables. The factor loadings are used in the interpretation of the factors, that is, larger values relate a factor to the corresponding observed variables, and from looking at which variables load highly on a factor, we can try to come up with a meaningful description or label for each factor. The u_i values are analogous to the residual terms in the usual multiple linear regression model (see Chapter 4), but as they are specific to each x_i in the factor analysis context, they are more commonly known as specific variates. The series of regression equations above may be written

more concisely as

$$\mathbf{x} = \boldsymbol{\Lambda}\mathbf{f} + \mathbf{u}$$

where

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1k} \\ \vdots & \vdots & \vdots \\ \lambda_{q1} & \dots & \lambda_{qk} \end{pmatrix} \quad \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix}$$

We assume that the residual terms u_1, \dots, u_q are uncorrelated with each other and with the common factors f_1, \dots, f_k . The two assumptions imply that, given the values of the factors, the manifest variables are independent, that is, the correlations of the observed variables arise from their relationships with the factors. Since the factors are unobserved, we can fix their location and scale arbitrarily. We will assume that they occur in standardized form with mean 0 and standard deviation 1. We will also assume, initially at least, that the factors are uncorrelated with one another, in which case the factor loadings are the correlations of the manifest variables and the factors. With these additional assumptions about the factors, the factor analysis model implies that the variance of variable x_i, σ_i^2 , is given by

$$\sigma_i^2 = \sum_{j=1}^k \lambda_{ij}^2 + \psi_i$$

where ψ_i is the variance of u_i . So, the factor analysis model implies that the variance of each observed variable can be split into two parts. The first part, h_i^2 , given by $h_i^2 = \sum_{j=1}^k \lambda_{ij}^2$, is known as the communality of the variable and represents the variance shared with the other variables via the common factors. The second part, ψ_i , is called the specific or unique variance and relates to the variability in x_i not shared with the other variables. In addition, the factor model leads to the following expression for the covariance of variables x_i and x_j :

$$\sigma_{ij} = \sum_{l=1}^k \lambda_{il} \lambda_{jl}$$

The covariances are not dependent on the specific variates in any way; it is the relationships of the manifest variables to the common factors that account for the relationships between the manifest variables. Collecting together the equations above relating observed variances and covariances to the factor loadings and the variances of the specific variates, we find that the factor analysis model implies that the population covariance matrix of the observed variables, Σ , has the form

$$\Sigma = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

where

$$\Psi = \text{diag}(\psi_i)$$

The converse also holds: if Σ can be decomposed into the form given here, then the k -factor model holds for \mathbf{x} . In practice, of course, Σ will need to be estimated by the sample covariance matrix \mathbf{S} (alternatively, the model will be applied to the correlation matrix \mathbf{R}), and we will need to obtain estimates of Λ and Ψ so that the observed covariance matrix takes the form required by the model (see later in the chapter for an account of estimation methods). We will also need to determine the value of k , the number of factors, so that the model provides the most parsimonious but adequate fit for \mathbf{S} or \mathbf{R} .

To apply the factor analysis model outlined in Technical [Section 11.1](#) to a sample of multivariate observations, we need to estimate the parameters of the model, factor loadings, and specific variances in some way. The estimation problem in factor analysis is essentially that of finding the estimates $\hat{\Lambda}$ and $\hat{\Psi}$ for which

$$\mathbf{S} \approx \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$$

(If the x_i values are standardized, then \mathbf{S} is replaced by \mathbf{R} .)

In some very simple cases, an exact solution is possible, that is, one in which the approximately equal sign in this equation becomes an equals sign. Considering such an example may be helpful before moving on to consider estimation in more realistic situations. The example we shall use here is one originally discussed by Spearman (1904), and concerns children's examination marks in three subjects: Classics (x_1), French (x_2), and English (x_3). The sample correlation matrix calculated by Spearman is as follows:

$$\mathbf{R} = \begin{bmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{bmatrix}$$

If we assume a single factor, then the appropriate factor analysis model is

$$x_1 = \lambda_1 f + u_1$$

$$x_2 = \lambda_2 f + u_2$$

$$x_3 = \lambda_3 f + u_3$$

In this example, the common factor f might be equated with intelligence or general intellectual ability, and the specific variates u_1, u_2, u_3 will have small variances if their associated observed variable is closely related to f . Here, the number of parameters in the model (six) is equal to the number of

independent elements in \mathbf{R} , and so, by equating elements of the observed correlation matrix to the corresponding values predicted by the single-factor model, we will be able to find estimates of $\lambda_1, \lambda_2, \lambda_3, \psi_1, \psi_2$, and ψ_3 such that the model fits exactly. The six equations derived from the matrix equality implied by the factor analysis model, that is,

$$\mathbf{R} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} [\lambda_1 \quad \lambda_2 \quad \lambda_3] + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

are

$$\hat{\lambda}_1\lambda_2 = 0.83$$

$$\hat{\lambda}_1\lambda_3 = 0.78$$

$$\hat{\lambda}_1\lambda_4 = 0.67$$

$$\hat{\psi}_1 = 1.0 - \hat{\lambda}_1^2$$

$$\hat{\psi}_2 = 1.0 - \hat{\lambda}_2^2$$

$$\hat{\psi}_3 = 1.0 - \hat{\lambda}_3^2$$

The solutions of these equations are

$$\begin{aligned} \hat{\lambda}_1 &= 0.99 & \hat{\lambda}_2 &= 0.84 & \hat{\lambda}_3 &= 0.79 \\ \hat{\psi}_1 &= 0.02 & \hat{\psi}_2 &= 0.30 & \hat{\psi}_3 &= 0.38 \end{aligned}$$

These values, when plugged in to the formula for the correlation matrix implied by the factor model, will reproduce the observed correlation matrix, but here, the factor model is not of any use because it uses six parameters to model the same number of independent elements in \mathbf{R} . The model does not provide a simplified description of the relationships between the three examination scores.

11.3 Estimating the Parameters in the Factor Analysis Model

We now have to consider how to estimate the parameters in the factor analysis model in those situations of practical interest in which the number of parameters in the model is less (and hopefully, considerably less) than the number of independent elements in the covariance or correlation matrix of the manifest

variables, so that the factor analysis model provides a genuinely more parsimonious description of the relationships between the manifest variables.

There are two main methods of estimation leading to what are known as principal factor analysis and maximum likelihood factor analysis, both of which are briefly described in the following technical Section.

Technical Section 11.2: Estimating the Parameters in the k -Factor Analysis Model

1. Principal Factor Analysis

Principal factor analysis is similar in many respects to principal components analysis (see Chapter 10); however, it does not operate directly on \mathbf{S} , the covariance matrix of the observed variables (or directly on \mathbf{R} , the correlation matrix), but on what is known as the reduced covariance matrix \mathbf{S}^* , defined as

$$\mathbf{S}^* = \mathbf{S} - \hat{\boldsymbol{\Psi}}$$

where $\hat{\boldsymbol{\Psi}}$ is a diagonal matrix with entries $\hat{\psi}_i$ that are estimates of the specific variances ψ_i . We remember from Technical Section 11.1 that, given a set of estimated loadings, the variance of variable x_i , s_i^2 , implied by the model is

$$s_i^2 = \sum_{j=1}^k \hat{\lambda}_{ij}^2 + \hat{\psi}_i$$

So, the diagonal elements of \mathbf{S}^* are given by $\sum_{j=1}^k \hat{\lambda}_{ij}^2$ for $i = 1, \dots, q$; these values are the estimated communalities—the parts of the variance of each observed variable that can be explained by the common factors. Unlike principal components analysis, factor analysis does not try to account for all observed variance; only that which is shared through the common factors. A matter of more concern in factor analysis is accounting for the covariances or correlations between the manifest variables rather than their variances.

To calculate the reduced covariance matrix \mathbf{S}^* (or with \mathbf{R} replacing \mathbf{S} , \mathbf{R}^*), we need values for the estimated communalities that are calculated from estimated factor loadings. But initially, we have no estimates of factor loadings. To get round this seemingly hen-or-egg situation, we need to find a sensible way of calculating initial values for the communalities that does not depend on having estimated factor loadings. When factor analysis is based on the correlation matrix of the manifest variables, there are two frequently used approaches:

- Take the communality of a variable x_i as the square of the multiple correlation coefficient of x_i with the other observed variables.
- Take the communality of x_i as the largest of the absolute values of the correlation coefficients between x_i and one of the other variables.

Each of these possibilities will lead to higher values for the initial communality when x_i is highly correlated with at least some of the other manifest variables, which is essentially what is required. Given initial communality values, a principal components analysis is performed on \mathbf{S}^* , and the first k eigenvectors are used to provide the estimates of the loadings in the k -factor model. The estimation process can stop here, or the loadings obtained at this stage can provide revised communality estimates calculated as $\sum_{j=1}^k \hat{\lambda}_{ij}^2$, where the $\hat{\lambda}_{ij}$ values are the loadings estimated in the previous step. The procedure is then repeated until some convergence criterion is satisfied. Difficulties can sometimes arise with this iterative approach if at any time a communality estimate exceeds the variance of the corresponding manifest variable, resulting in a negative estimate of the variable's specific variance. Such a result is known as a Heywood case (Heywood, 1931) and is clearly unacceptable since we cannot have a negative specific variance.

2. Maximum Likelihood Factor Analysis

Maximum likelihood is regarded, by statisticians at least, as perhaps the most respectable method of estimating the parameters in the factor analysis model. The essence of this approach is to define a type of "distance" measure, F , between the observed covariance matrix and the covariance matrix implied by the factor analysis model. The measure F is defined as follows:

$$F = \ln |\Lambda\Lambda' + \Psi| + \text{trace}(\mathbf{S} | \Lambda\Lambda' + \Psi |^{-1}) - \ln |\mathbf{S}| - q$$

The function F takes the value 0 if $\Lambda\Lambda' + \Psi$ is equal to \mathbf{S} , and values greater than 0 otherwise. Estimates of factor loadings and specific variances are found by minimizing F using an iterative procedure that begins with initial estimates of the parameters found from a principal factor analysis; details are given in Lawley and Maxwell (1971), Mardia et al. (1979), and Everitt (1984, 1987). Minimizing F is essentially equivalent to maximizing L , the likelihood function for the k -factor model, under the assumption of multivariate normality of data because it can be shown that $L = -\frac{1}{2}nF$ plus a function of the observations. As with iterated principal factor analysis, the maximum likelihood approach can also experience difficulties with Heywood cases.

Nowadays, maximum likelihood factor analysis is the recommended method for most applications.

11.4 Estimating the Numbers of Factors

Determining how many factors, k , are needed to give an adequate representation of the observed covariances or correlations is generally critical when fitting an exploratory factor analysis model. A k and $k + 1$ factor solution

will often produce quite different factors and factor loadings for all factors, unlike a principal component analysis in which the first k components will be identical in each solution. Further, as pointed out by Jolliffe (1998), with too few factors, there will be too many high loadings, and with too many factors, factors may be fragmented and difficult to interpret convincingly. Choosing k might be done by examining solutions corresponding to different values of k and deciding subjectively which can be given the most convincing interpretation—not an entirely convincing method in many circumstances. An advantage of the maximum likelihood approach is that it has an associated formal hypothesis-testing procedure for the number of factors.

The test statistic is

$$U = n' \min(F)$$

where $n' = n + 1 - \frac{1}{6}(2q + 5) - \frac{2}{3}k$. If k common factors are adequate to account for the observed covariances or correlations of the manifest variables, then U has, asymptotically, a chi-squared distribution with v degrees of freedom, where

$$v = \frac{1}{2}(q - k)^2 - \frac{1}{2}(q + k)$$

In most exploratory studies, k cannot be specified in advance, and so, a sequential procedure is used. Starting with some small value for k (usually $k = 1$), the parameters in the corresponding factor analysis model are estimated by maximum likelihood. If U is not significant, the current value of k is accepted; otherwise, k is increased by 1 and the process repeated. If at any stage the degrees of freedom of the test become 0, then either no nontrivial solution is appropriate or, alternatively, the factor model itself with its assumption of linearity between observed and latent variables is questionable.

11.5 Fitting the Factor Analysis Model: An Example

As our first example of fitting a factor analysis model, we shall return to the crime rate data introduced in Chapter 10. Applying the test for the number of factors described previously in [Section 11.4](#) to one-, two-, and three-factor models gives the following results:

Test for number of factors

Test of the hypothesis that one factor is sufficient.

The chi-square statistic is 64.63 on 14 degrees of freedom.

The p-value is 1.78e-08.

Test of the hypothesis that two factors are sufficient.

The chi-square statistic is 20.02 on 8 degrees of freedom.

The p-value is 0.0102.

Test of the hypothesis that three factors are sufficient.

The chi-square statistic is 4.9 on 3 degrees of freedom.

The p-value is 0.179.

The results indicate that three factors are needed to account for the relationships between the seven crime rates. The maximum likelihood estimated parameters of the three-factor model are shown in Table 11.1. Interpretation of the three factors will be left until later in the chapter for reasons that will hopefully become clear after [Section 11.6](#). However, we can use the estimated loadings and specific variances to calculate the correlation matrix of the seven crime rates implied by the fitted three-factor model using the formula

$$\hat{\mathbf{R}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$$

The result is shown in [Table 11.2](#); also shown in this table is the correlation matrix implied by the first three principal components of the data and the observed correlation matrix. Comparing only the off-diagonal elements, we see that the factor analysis solution with three factors reproduces the correlations between the observed variables rather better than the first three principal components. The diagonal values from the factor analysis and principal component solutions cannot be compared because the specific variance

TABLE 11.1

Estimated Parameters for the Three-Factor Model Fitted to Crime Rate Data by Maximum Likelihood

Specific Variance Estimates						
Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
0.030	0.360	0.185	0.253	0.136	0.279	0.005
Estimated Factor Loadings						
	Factor 1		Factor 2		Factor 3	
Murder	0.654		0.727		-0.115	
Rape	0.611		0.307		0.415	
Robbery	0.828		0.344		-0.103	
Assault	0.697		0.479		0.181	
Burglary	0.625		0.330		0.604	
Theft	0.422		0.231		0.700	
Vehicle	0.992		-0.106		—	
SS loadings	3.523		1.145		1.083	
Prop. var	0.503		0.164		0.155	
Cum. var	0.503		0.667		0.822	

TABLE 11.2

Observed Correlation Matrix of Crime Rate Data, and Correlation Matrices Implied by the Three-Factor Model and First Three Principal Components

Observed Correlation Matrix							
	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
Murder	1.000	0.578	0.804	0.781	0.581	0.361	0.573
Rape	0.578	1.000	0.530	0.659	0.721	0.635	0.569
Robbery	0.804	0.530	1.000	0.740	0.551	0.400	0.786
Assault	0.781	0.659	0.740	1.000	0.710	0.512	0.638
Burglary	0.581	0.721	0.551	0.710	1.000	0.764	0.579
Theft	0.361	0.635	0.400	0.512	0.764	1.000	0.386
Vehicle	0.573	0.569	0.786	0.638	0.579	0.386	1.000
Correlation Matrix Implied by the Three-Factor Model							
	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
Murder	1.000	0.575	0.804	0.783	0.579	0.364	0.573
Rape	0.575	1.000	0.568	0.648	0.734	0.619	0.569
Robbery	0.804	0.568	1.000	0.723	0.568	0.356	0.786
Assault	0.783	0.648	0.723	1.000	0.702	0.531	0.638
Burglary	0.579	0.734	0.568	0.702	1.000	0.762	0.578
Theft	0.364	0.619	0.356	0.531	0.762	1.000	0.386
Vehicle	0.573	0.569	0.786	0.638	0.578	0.386	1.000
Correlation Matrix Implied by the First Three Principal Components							
	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
Murder	0.936	0.583	0.810	0.859	0.581	0.338	0.563
Rape	0.583	0.745	0.575	0.694	0.799	0.740	0.565
Robbery	0.810	0.575	0.897	0.782	0.570	0.333	0.834
Assault	0.859	0.694	0.782	0.855	0.715	0.528	0.621
Burglary	0.581	0.799	0.570	0.715	0.863	0.820	0.572
Theft	0.338	0.740	0.333	0.528	0.820	0.874	0.404
Vehicle	0.563	0.565	0.834	0.621	0.572	0.404	0.975

parameters in the factor analysis model enable these values to be reproduced exactly. However, subtracting the specific variances from the correlation matrix implied by the factor model will show that the diagonal terms of the observed correlation matrix are fitted better by the principal components than by the common factors, as we would expect, because the common factors only attempt to account for the “shared variance” of the observed variables.

11.6 Rotation of Factors

Up to now, we have conveniently ignored a problem with the factor analysis model—that the factor loadings are not uniquely determined by this model. What this means is explained in Technical Section 11.3.

Technical Section 11.3: The Lack of Uniqueness of Factor Loadings

As we have seen previously, the k -factor model can be written in terms of a $q \times k$ matrix of factor loadings Λ , a vector of k -common factors f , and a vector of q residuals u , as $x = \Lambda f + u$. Now let us introduce a $k \times k$ orthogonal matrix M , such that $MM' = I$, and rewrite the basic regression equation linking the observed variables to the common factors as $x = (\Lambda M)(M'f) + u$. This satisfies all the requirements of a k -factor model as outlined previously with new factors $f' = M'f$, and new factor loadings ΛM . A model with these factor loadings and factors implies that the population covariance matrix of x -variables is given by

$$\Sigma = (\Lambda M)(\Lambda M)' + \Psi$$

because $MM' = I$ reduces to the original form of $\Sigma = \Lambda\Lambda' + \Psi$. This implies that factors f with loadings Λ , and factors f^* with loadings ΛM are, for any orthogonal matrix M , completely equivalent for explaining the covariance matrix of the observed variables.

The result of the lack of uniqueness of factor loadings is that, essentially, there are an infinite number of solutions to the factor analysis model, as previously formulated. Consequently, to define a unique solution, it becomes necessary to introduce some constraints on the parameters in the original factor analysis model. In general, what is done is to require the first factor to make maximal contribution to the common variance of the observed variables, the second to make maximal contribution to this variance subject to being uncorrelated to the first, etc. (Compare the derivation of principal components in Chapter 10.) Such constraints ensure a unique solution and lead to uncorrelated (orthogonal) factors that are arranged in descending order of "importance." However, these properties are not inherent in the factor analysis model, and merely considering such a solution may hinder interpretation; two consequences of the constraints, for example, are

- The factorial complexity of variables is likely to be greater than 1 regardless of the underlying true model; consequently, variables may have substantial loadings on more than one factor.
- Except for the first factor, the remaining factors are often bipolar, that is, they have a mixture of positive and negative loadings.

It may be possible that a more interpretable solution can be achieved using the equivalent model with loadings $\Lambda^* = \Lambda M$ for a particular orthogonal matrix M . Such a process is generally known as factor rotation, but before we consider how to choose M , that is, how to "rotate" the factors, we need to address the question "is factor rotation an acceptable process?" Certainly in the past, factor analysis has been the subject of severe criticism because of the possibility of rotating factors. Critics have suggested that this apparently

allows the investigator to impose on the data whatever type of solution he or she is looking for; some have even gone so far as to suggest that factor analysis has become popular in some areas precisely because it does enable users to impose their preconceived ideas of the structure behind the observed correlations (Blackith and Reyment, 1971). But, on the whole, such suspicions are not justified, and factor rotation can be a useful procedure for simplifying an exploratory factor analysis.

Factor rotation merely allows the fitted factor analysis model to be described as simply as possible; rotation does not alter the overall structure of a solution but only how the solution is described. Rotation is a process by which a solution is made more interpretable without changing its underlying mathematical properties. Initial factor solutions with variables loading on several factors and with bipolar factors can be difficult to interpret. Interpretation is more straightforward if each variable is highly loaded on at most one factor, and if all factor loadings are either large and positive, or near 0, with few intermediate values. The variables are thus split into disjoint sets, each of which is associated with a single factor. This aim is essentially what Thurstone (1931) referred to as simple structure. In more detail, such structure has the following properties:

- Each row of the factor-loading matrix should contain at least one zero.
- Each column of the loading matrix should contain at least k zeros.
- Every pair of columns of the loading matrix should contain several variables whose loadings vanish in one column but not in the other.
- If the number of factors is four or more, every pair of columns should contain a large number of variables with zero loadings in both columns.
- Conversely, for every pair of columns of the loading matrix, only a small number of variables should have nonzero loadings in both columns.

When simple structure is achieved, the observed variables will fall into mutually exclusive groups whose loadings are high on single factors, perhaps moderate to low on a few factors, and of negligible size on the remaining factors.

The search for simple structure or something close to it begins after an initial factoring has determined the number of common factors necessary and the communalities of each observed variable. The factor loadings are then transformed by post multiplication by a suitably chosen orthogonal matrix. Such a transformation is equivalent to a rigid rotation of the axes of the originally identified factor space.

11.6.1 A Simple Example of Graphical Rotation

For a two-factor model, the process of rotation can be performed graphically. As an example, consider the following correlation matrix for six

school subjects:

$$\mathbf{R} = \begin{matrix} \text{French} & 1.00 \\ \text{English} & 0.44 & 1.00 \\ \text{History} & 0.41 & 0.35 & 1.00 \\ \text{Arithmetic} & 0.29 & 0.35 & 0.16 & 1.00 \\ \text{Algebra} & 0.33 & 0.32 & 0.19 & 0.59 & 1.00 \\ \text{Geometry} & 0.25 & 0.33 & 0.18 & 0.47 & 0.46 & 1.00 \end{matrix}$$

The initial, unrotated maximum likelihood estimated factor loadings are shown in Table 11.3. The loadings on the second factor have a mixture of positive and negative signs, which can make interpretation difficult. The loadings are plotted in Figure 11.1. By referring each variable to the new axes shown by bold lines on Figure 11.1, which correspond to a rotation of the original axes through about 40°, a new set of loadings can be obtained, and these are also given in Table 11.3. The rotated factors are far easier to interpret—the first being perhaps a “mathematical” factor, and the second a “verbal” factor. Note that the communalities of each observed variable remain the same in the unrotated and the rotated solution.

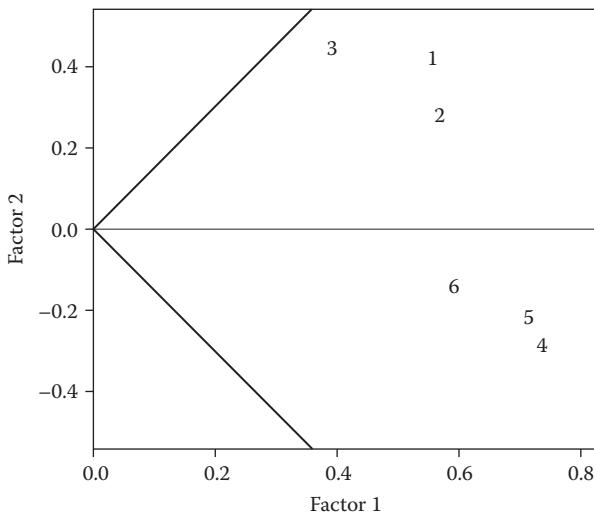
11.6.2 Numerical Rotation Methods

When there are more than two factors, more formal methods of rotation are needed. Further, during the rotation phase, we might choose to abandon one of the assumptions made previously, namely, that factors are orthogonal, that is, independent (the condition was assumed initially simply for convenience in describing the factor analysis model). Consequently, two types of rotation are possible:

- Orthogonal rotation: Methods restrict the rotated factors to being uncorrelated.
- Oblique rotation: Methods allow correlated factors.

TABLE 11.3
Maximum Likelihood Two-Factor Solution for
Correlations of Six School Subjects

	Unrotated		Rotated	
	Factor 1	Factor 2	Factor 1	Factor 2
1	0.56	0.42	0.23	0.66
2	0.57	0.29	0.32	0.55
3	0.39	0.45	0.08	0.59
4	0.74	-0.28	0.77	0.17
5	0.72	-0.21	0.72	0.22
6	0.59	-0.13	0.57	0.22

**FIGURE 11.1**

Plot of factor loadings for the correlation matrix of six school subjects showing the rotated axes that lead to a simpler interpretation of the two factors.

So, the first question that needs to be considered when rotating factors is: should we use an orthogonal or oblique rotation? As for many questions posed in data analysis, there is no single answer to this question. There are advantages and disadvantages to using either type of rotation procedures. As a general rule, if a researcher is primarily concerned with getting results that “best fit” his or her data, then the researcher should rotate the factors obliquely. If, on the other hand, the researcher is more interested in the generalizability of his or her results, then orthogonal rotation is probably to be preferred.

One major advantage of an orthogonal rotation is simplicity since the loadings represent correlations between factors and manifest variables. This is not the case with an oblique rotation because of correlations between the factors. Here, there are two parts of the solution to consider:

- Factor pattern coefficients: Regression coefficients that multiply with factors to produce measured variables according to the common factor model
- Factor structure coefficients: Correlation coefficients between manifest variables and the factors

Additionally, there is a matrix of factor correlations to consider. In many cases in which these correlations are relatively small, researchers may prefer to return to an orthogonal solution.

There are a variety of rotation techniques, although only relatively few are in general use. For orthogonal rotation, the two most commonly used techniques are known as *varimax* and *quartimax*:

- Varimax rotation: Originally proposed by Kaiser (1958), this has as its rationale the aim of factors with a few large loadings and as many near-zero loadings as possible. This is achieved by iterative maximization of a quadratic function of the loadings—details are given in Mardia et al. (1979). It produces factors that have high correlations with one small set of variables and little or no correlation with other sets. There is the tendency for any general factor to disappear because the factor variance is redistributed.
- Quartimax rotation: Originally suggested by Carroll (1953), this approach forces a given variable to correlate highly on one factor and either not at all or very low on other factors. This is far less popular than varimax.

For oblique rotation, the two methods most often used are *oblimin* and *promax*.

- Oblimin rotation: Invented by Jennrich and Sampson (1966), this method attempts to find simple structure with regard to the factor pattern matrix through a parameter that is used to control the degree of correlation between the factors. Fixing a value for this parameter is not straightforward, but Lackey and Sullivan (2003) suggest that values between about -0.5 and 0.5 are sensible for many applications.
- Promax rotation: This is a method suggested by Hendrickson and White (1964) that operates by raising the loadings in an orthogonal solution (generally, a varimax rotation) to some power. The goal is to obtain a solution that provides the best structure using the lowest possible power loadings and the lowest correlation between the factors.

Factor rotation is often regarded as controversial since it apparently allows the investigator to impose on the data whatever type of solution is required. However, this is clearly not the case since, although the axes may be rotated about their origin or allowed to become oblique, the distribution of the points will remain invariant. Rotation is simply a procedure that allows new axes to be chosen so that the positions of the points can be described as simply as possible.

It should be noted that rotation techniques are also often applied to the results from a principal components analysis in the hope that it will aid in their interpretability. Although in some cases this may be acceptable, it does have several disadvantages, which are listed by Jolliffe (1989). The main problem is that the defining property of principal components, namely, that

of accounting for maximal proportions of the total variation in the observed variables, is lost after rotation.

11.6.3 Rotating the Crime Rate Factors

As an illustration of the rotation of factors, we will use varimax rotation on the initial three-factor solution found for the crime rate data, given in [Table 11.1](#). The details of the rotated solution are given in Table 11.4. Note first that the estimated communalities and specific variances do not change from those given in [Table 11.1](#). The variances of the factors do change, but the total variance for all three factors is the same for both the unrotated and rotated solutions. The rotated solution in Table 11.4 will lead to exactly the same implied correlation matrix as that given in [Table 11.2](#) for the unrotated solution (readers might like to confirm that this is the case by carrying out the appropriate calculations themselves). The acid test of the rotated solution is whether or not it is easier to interpret than the unrotated solution. Looking at the loadings on the first factor of the rotated solution in Table 11.4, we see that burglary, theft and, to a slightly lesser extent, rape, are highly correlated with this factor. The second factor has a large correlation with the murder rate, and also with robbery and assault. The third factor is dominated by its correlation with vehicle crime. If we ignore the loading for rape on the first factor, it could perhaps be labeled property crime, and the second factor might be crime against the person. However, this type of exercise needs more knowledge of the subject matter, so I leave it to readers to amuse themselves with

TABLE 11.4
Varimax-Rotated Three-Factor Solution for Crime Rate Data

Estimates of Specific Variances						
Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
0.030	0.360	0.185	0.253	0.136	0.279	0.005
Loadings						
	Factor 1		Factor 2		Factor 3	
Murder	0.259		0.922		0.229	
Rape	0.645		0.369		0.297	
Robbery	0.243		0.664		0.561	
Assault	0.492		0.629		0.331	
Burglary	0.828		0.330		0.263	
Theft	0.831		0.132		0.120	
Vehicle	0.285		0.317		0.902	
SS loadings	2.24		2.049		1.463	
Prop. var	0.32		0.293		0.209	
Cum. var	0.32		0.613		0.822	

devising other labels. The example does demonstrate that even rotated solutions are not always open to easy interpretation.

11.7 Estimating Factor Scores

In most applications, an exploratory factor analysis will consist of the estimation of the parameters in the model and the rotation of the factors, followed by an (often heroic) attempt to interpret the fitted model. There are occasions, however, when the investigator would like to find factor scores for each individual in the sample. Such scores, similar to those derived in a principal components analysis (see Chapter 3), might be useful in a variety of ways. But the calculation of factor scores is not as straightforward as the calculation of principal components scores. In the original equation defining the factor analysis model, the variables are expressed in terms of factors, whereas to calculate scores we require the relationship to be in the opposite direction. Bartholomew (1987) makes the point that to talk about “estimating” factor scores is essentially misleading since they are random variables, and the issue is really one of prediction.

But if we make the assumption of normality, the conditional distribution of \mathbf{f} given \mathbf{x} can be found. It is

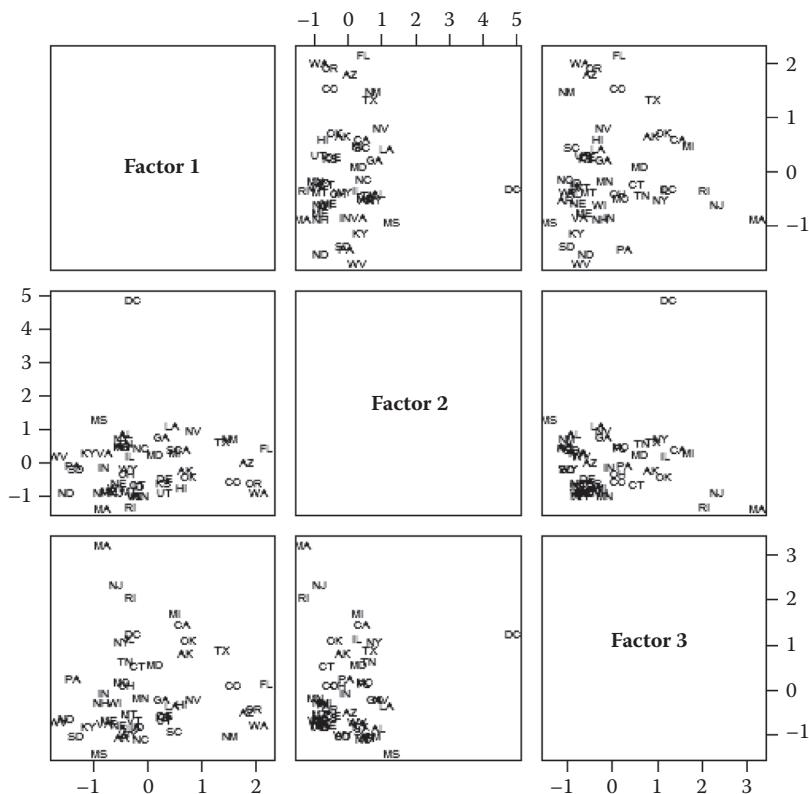
$$N[\Lambda' \Sigma^{-1} \mathbf{x}, (\Lambda' \Psi^{-1} \Lambda + \mathbf{I})^{-1}]$$

Consequently, one plausible way of calculating factor scores would be to use the sample version of the mean of this distribution, that is,

$$\hat{\mathbf{f}} = \hat{\Lambda}' \mathbf{S}^{-1} \mathbf{x}$$

where the vector of scores for an individual, \mathbf{x} , is assumed to have mean 0, that is, sample means for each variable have already been subtracted. Other possible methods for deriving factor scores are described in Rencher (2002). In many respects, the most damaging problem with factor analysis is not the rotational indeterminacy of the loadings but the indeterminacy of the factor scores.

We can illustrate the estimation (prediction) of factor scores, again using the crime rate data. A scatterplot matrix of factor scores for the three-factor model found using the aforementioned method is given in [Figure 11.2](#); points are labeled by state. The most notable feature of this plot is again the position of DC, which has a very high score on Factor 2, crime against the person; perhaps better to steer clear of DC!

**FIGURE 11.2**

Scatterplot matrix of factor scores from the three-factor model fitted to crime rate data.

11.8 Exploratory Factor Analysis and Principal Component Analysis Compared

Factor analysis, like principal components analysis, is an attempt to describe a set of multivariate data using a smaller number of dimensions than one begins with, but quite different approaches are used by each technique to achieve this goal. Some differences between principal components analysis and exploratory factor analysis are

- Factor analysis tries to explain the covariances or correlations of the observed variables by postulating a small number of underlying latent variables, the common factors in this context, to which the manifest variables are related. Principal components analysis

constructs linear functions of the observed variables that account for decreasing proportions of the variance of these variables, so that the first few components may account for a substantial proportion of the variance and thus provide a useful summary of the data.

- If the number of retained components is increased, say, from m to $m + 1$, the first m components are unchanged. This is not the case in factor analysis, where there can be substantial changes in all factors if the number of factors is changed.
- The calculation of principal component scores is straightforward; the calculation of factor scores is more complex, and a variety of methods have been suggested.
- There is usually no relationship between the principal components extracted from the sample correlation matrix and those based on the sample covariance matrix. For maximum likelihood factor analysis, however, the results of analyzing either matrix are essentially equivalent (this is not true of principal factor analysis).

Despite these differences, the results from both types of analysis are frequently very similar. Certainly, if the specific variances are small, we would expect both forms of analysis to give similar results. However, if the specific variances are large, they will be absorbed into all the principal components, both retained and rejected, whereas factor analysis makes special provision for them.

Lastly, it should be remembered that both principal components analysis and factor analysis are similar in one important respect—they are both pointless if the observed variables are almost uncorrelated. In this case, factor analysis has nothing to explain, and principal components analysis will simply lead to components that are similar to the original variables.

11.9 Confirmatory Factor Analysis

An exploratory factor analysis is used in the early investigation of a set of multivariate data to determine whether the factor analysis model is useful in finding a parsimonious way of describing the relationships between the observed variables—which observed variables are most highly correlated with the common factors and how many common factors are needed. In an exploratory factor analysis, no constraints are placed on which variables load on which factors. Based on the results of an exploratory factor analysis, the investigator might wish to postulate a specific model for a new set of similar data, one in which the loadings of some variables on some factors are fixed at 0 because they were “small” in the exploratory

analysis, and perhaps to allow some pairs of factors, but not others, to be correlated. We are now in the realm of confirmatory factor analysis. Before going into details of this technique, it is important to emphasize that, while it is perfectly appropriate to arrive at a factor model to submit to a confirmatory analysis from an exploratory factor analysis, the model must be tested on a fresh set of data. Models must not be generated and tested on the same data.

In a confirmatory factor model, the loadings for some observed variables on some of the postulated common factors will be set a priori to 0. Additionally, some correlations between factors might also be fixed at 0. Such a model is fitted to a set of data by estimating its free parameters, that is, those not fixed at 0 by the investigator, so that the variances and covariances of the manifest variables implied by the model are as close as possible in some sense to the corresponding observed values. The most commonly used method of estimation for confirmatory factor analysis models is maximum likelihood (details are given in Everitt and Dunn, 2001), but the quintessential part of the method is that the parameters of the postulated model, which we assume are set out in a vector Θ , are estimated by minimizing the function F given by

$$F = \log |\Sigma(\Theta)| - \log |\mathbf{S}| + \text{trace}[\mathbf{S}\Sigma(\Theta)^{-1}] - q$$

where \mathbf{S} is the covariance matrix of the manifest variables, and $\Sigma(\Theta)$ is the covariance matrix implied by the fitted model, that is, a matrix the elements of which are functions of the parameters contained in the vector Θ . We see by varying the parameters, so that $\Sigma(\Theta)$ becomes more like \mathbf{S} , that F approaches 0.

We will now illustrate the application of confirmatory factor analysis with two examples.

11.9.1 Ability and Aspiration

Caslyn and Kenny (1977) recorded the values of the following six variables for 556 white eighth-grade students:

1. Self-concept of ability (x_1)
2. Perceived parental evaluation (x_2)
3. Perceived teacher evaluation (x_3)
4. Perceived friend's evaluation (x_4)
5. Educational aspiration (x_5)
6. College plans (x_6)

Caslyn and Kenny postulated that two underlying latent variables generated the relationships between the observed variables, namely, ability and

aspiration. The first four of the manifest variables were assumed to be indicators of ability, and the last two indicators of aspiration, and ability and aspiration, were allowed to be correlated. So, the regression-like equations specifying the postulated model are

$$x_1 = \lambda_1 f_1 + 0f_2 + u_1$$

$$x_2 = \lambda_2 f_1 + 0f_2 + u_2$$

$$x_3 = \lambda_3 f_1 + 0f_2 + u_3$$

$$x_4 = \lambda_4 f_1 + 0f_2 + u_4$$

$$x_5 = 0f_1 + \lambda_5 f_2 + u_5$$

$$x_6 = 0f_1 + \lambda_6 f_2 + u_6$$

where f_1 represents the ability latent variable, and f_2 represents the aspiration latent variable. Note that unlike the exploratory factor analysis, a number of factor loadings are fixed at 0 and play no part in the estimation process. The model has a total of 13 parameters to estimate: 6 factor loadings (λ_1 to λ_6), 6 specific variances (ψ_1 to ψ_6), and 1 correlation between ability and aspiration (ρ). The observed correlation matrix given in Table 11.5 has 6 variances and 15 correlations, a total of 21 terms. Consequently, the postulated model has $21 - 13 = 8$ degrees of freedom. A path diagram (see Everitt and Dunn, 2001) for the correlated, two-factor model is shown in [Figure 11.3](#).

The results from fitting the ability and aspiration model to the observed correlations are shown in [Table 11.6](#) (note that the two latent variables are assumed to have variance 1 each; these two variances cannot, of course, be estimated). Of particular note is the estimate of the correlation between the two latent variables; this estimate (0.66 with a standard error of 0.03) is a disattenuated correlation, that is, the correlation between “true” ability and “true” aspiration uncontaminated by measurement error in the observed indicators

TABLE 11.5
Observed Correlations for the Ability and
Aspiration Example

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.00					
x_2	0.73	1.00				
x_3	0.70	0.68	1.00			
x_4	0.58	0.61	0.57	1.00		
x_5	0.46	0.43	0.40	0.37	1.00	
x_6	0.56	0.52	0.48	0.41	0.72	1.00

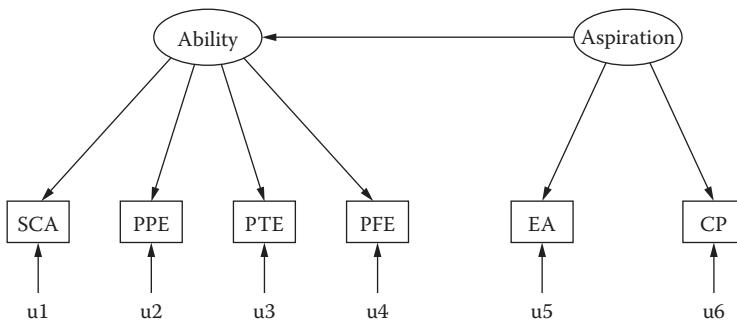


FIGURE 11.3
Path diagram for the ability and aspiration model.

of these concepts. An approximate 95% confidence interval for the disattenuated correlation is [0.60,0.72]. The fit of the model can be partially judged by a chi-square statistic (for details, see Everitt and Dunn, 2001), which in this case takes the value 9.26 with 8 degrees of freedom and an associated p-value of 0.32, suggesting that the postulated model fits the data very well. Note that, with confirmatory factor models, the standard errors of parameters assume importance because they allow the investigator to assess whether parameters might be dropped from the model to find a more parsimonious model

TABLE 11.6
Results of Fitting the Ability and Aspiration Correlated
Two-Factor Model to the Correlations in [Table 11.5](#)

Parameter	Estimate	Standard Error (SE)	Estimate/SE
λ_1	0.863	0.035	24.558
λ_2	0.849	0.035	23.961
λ_3	0.805	0.035	22.115
λ_4	0.695	0.039	18.000
λ_5	0.775	0.040	19.206
λ_6	0.929	0.039	23.569
ψ_1	0.255	0.023	19.911
ψ_2	0.279	0.024	11.546
ψ_3	0.352	0.027	13.070
ψ_4	0.516	0.035	14.876
ψ_5	0.399	0.038	10.450
ψ_6	0.137	0.044	3.152
ρ	0.667	0.032	21.521

that still provides an adequate fit to the data. In exploratory factor analysis, standard errors of factor loadings can be calculated, but they are hardly ever used; instead, an informal interpretation of factors is made.

11.9.2 A Confirmatory Factor Analysis Model for Drug Usage

For our second example of fitting a confirmatory factor analysis model, we return to the drug usage among students data introduced in Chapter 10. In the original investigation of these data reported by Huba et al. (1981), a confirmatory factor analysis model was postulated for the model arising from consideration of previously reported research on student drug usage. The model consisted of the following three latent variables:

- f_1 : Alcohol use—with nonzero loadings on beer, wine, spirits, and cigarette usage.
- f_2 : Cannabis use—with nonzero loadings on marijuana, hashish, cigarette, and wine usage. The cigarette variable is assumed to load on both the first and second latent variables because it sometimes occurs with both alcohol and marijuana use and, at other times, does not. The nonzero loading on wine was allowed because of reports that wine is frequently used with marijuana and, consequently, some of the use of wine might be an indicator of tendencies toward cannabis use.
- f_3 : Hard drug use—with nonzero loadings on amphetamines, tranquilizers, hallucinogenics, hashish, cocaine, heroin, drugstore medication, inhalants, and spirits. The use of each of these substances was considered to suggest a strong commitment to the notion of psychoactive drug use.

(Each of the three latent variables is assumed to have a variance of 1.)

Correlations between each pair of the three factors are allowed to be free parameters to be estimated. So, the proposed model can be specified by the following series of equations:

$$\text{cigarettes} = \lambda_1 f_1 + \lambda_2 f_2 + 0 f_3 + u_1$$

$$\text{beer} = \lambda_3 f_1 + 0 f_2 + 0 f_3 + u_2$$

$$\text{wine} = \lambda_4 f_1 + \lambda_5 f_2 + 0 f_3 + u_3$$

$$\text{spirits} = \lambda_6 f_1 + 0 f_2 + \lambda_7 f_3 + u_4$$

$$\text{cocaine} = 0 f_1 + 0 f_2 + \lambda_8 f_3 + u_5$$

$$\text{tranquillizers} = 0 f_1 + 0 f_2 + \lambda_9 f_3 + u_6$$

TABLE 11.7

Results of Fitting a Correlated Three-Factor Model to Drug Usage Data

Parameter	Estimate	Standard Error (SE)	Estimate/SE
λ_1	0.358	0.035	10.371
λ_2	0.332	0.035	9.401
λ_3	0.792	0.023	35.021
λ_4	0.875	0.038	23.285
λ_5	-0.152	0.037	-4.158
λ_6	0.722	0.024	30.673
λ_7	0.123	0.023	5.439
λ_8	0.465	0.026	18.079
λ_9	0.676	0.024	28.182
λ_{10}	0.359	0.025	13.602
λ_{11}	0.476	0.026	18.571
λ_{12}	0.912	0.030	29.958
λ_{13}	0.396	0.030	13.379
λ_{14}	0.381	0.029	13.050
λ_{15}	0.543	0.025	21.602
λ_{16}	0.618	0.025	25.233
λ_{17}	0.763	0.023	32.980
ψ_1	0.611	0.024	25.823
ψ_2	0.374	0.020	18.743
ψ_3	0.379	0.024	16.052
ψ_4	0.408	0.019	21.337
ψ_5	0.784	0.029	26.845
ψ_6	0.544	0.023	23.222
ψ_7	0.871	0.032	27.653
ψ_8	0.773	0.029	26.735
ψ_9	0.169	0.044	3.846
ψ_{10}	0.547	0.022	24.593
ψ_{11}	0.705	0.027	25.941
ψ_{12}	0.618	0.025	24.655
ψ_{13}	0.418	0.021	19.713
ρ_1	0.634	0.027	23.369
ρ_2	0.313	0.029	10.674
ρ_3	0.499	0.027	18.412

$$\text{drugstore medication} = 0f_1 + 0f_2 + \lambda_{10}f_3 + u_7$$

$$\text{heroin} = 0f_1 + 0f_2 + \lambda_{11}f_3 + u_8$$

$$\text{marijuana} = 0f_1 + \lambda_{12}f_2 + 0f_3 + u_9$$

$$\text{hashish} = 0f_1 + \lambda_{13}f_2 + \lambda_{14}f_3 + u_{10}$$

$$\text{inhalants} = 0f_1 + 0f_2 + \lambda_{15}f_3 + u_{11}$$

$$\text{hallucinogenics} = 0f_1 + 0f_2 + \lambda_{16}f_3 + u_{12}$$

$$\text{amphetamines} = 0f_1 + 0f_2 + \lambda_{17}f_3 + u_{13}$$

The proposed model also allows for nonzero correlations between each pair of latent variables, and so, has a total of 33 parameters to estimate—17 loadings (λ_1 to λ_{17}), 13 specific variances (ψ_1 to ψ_{13}), and 3 between latent variables correlations (ρ_1 to ρ_3). Consequently, the model has $91 - 33 = 58$ degrees of freedom.

The results of fitting the proposed model are given in [Table 11.7](#). Here, the chi-square test for goodness of fit takes the value 323.96, which with 58 degrees of freedom has an associated p-value that is very small; the model does not appear to fit very well, and readers are referred to the original Huba et al. (1981) paper for details of how the model was changed to try to achieve a better fit.

The description of confirmatory factor analysis models in this section has been necessarily brief. For detailed accounts of such models and for more complex models involving latent variables (structural equation models), readers are referred to Kline (2004) and Bollen (1989).

11.10 Summary

- Exploratory factor analysis models attempt to explain the relationships between a set of manifest variables, as measured by the variables covariance or correlation matrix in terms of the relationships of the observed variables to a small number of underlying latent variables—the common factors.
- Initial factor solutions are almost always “rotated” before any attempt is made to interpret the factors. Rotation methods attempt to achieve “simple structure,” and rotated solutions may be allowed to be orthogonal (uncorrelated common factors) or oblique (correlated common factors). Orthogonal solutions are more commonly used because they are easier to interpret.

- Principal components analysis and exploratory factor analysis both attempt to simplify multivariate data by reducing their dimensionality. Principal components analysis involves a straightforward mathematical transformation of the observed variables, and has the primary aim of accounting for the variance of these variables. Exploratory factor analysis postulates the existence of underlying latent variables and aims to account for the correlations or covariances of the observed variables.
 - The fitting of confirmatory factor analysis models becomes possible when the investigator has a specific factor model in mind. This model may have arisen from an earlier exploratory analysis or from theoretical considerations. In the former case, the derived model must be tested on new data.
-

11.11 Exercises

11.1 Returning to the small exploratory factor analysis example described in [Section 11.2](#), suppose now that the observed correlations had been

$$\mathbf{R} = \begin{matrix} & \text{Classics} & & \\ \text{French} & 1.00 & & \\ & 0.84 & 1.00 & \\ \text{English} & 0.60 & 0.35 & 1.00 \end{matrix}$$

Find the values of the parameters in a one-factor model fitted to these correlations. Are there any problems with the solution?

- 11.2 Apply principal factor analysis to the crime rate data, and compare the varimax-rotated solution to that given in the text and found using the maximum likelihood estimation. Also compare the predicted correlation matrices from the principal factor analysis and maximum likelihood factor analysis solutions.
- 11.3 Investigate the use of alternative rotation methods to varimax on the crime rate data.
- 11.4 The following matrix gives the correlations between ratings on pain made by 123 people suffering from extreme pain. The nine statements are
1. Whether or not I am in pain in the future depends on the skills of the doctors.
 2. Whenever I am in pain, it is usually because of something I have done or not done.

3. Whether or not I am in pain depends on what the doctors do for me.
4. I cannot get any help for my pain unless I go to seek medical advice.
5. When I am in pain, I know that it is because I have not been taking proper exercise or eating the right food.
6. People's pain results from their own carelessness.
7. I am directly responsible for my pain.
8. Relief from pain is chiefly controlled by the doctors.
9. People who are never in pain are just plain lucky.

Each statement was scored on a scale from 1 (total agreement) to 6 (total disagreement).

$$\mathbf{R} = \begin{pmatrix} 1.00 & & & & & & & & \\ -0.04 & 1.00 & & & & & & & \\ 0.61 & -0.07 & 1.00 & & & & & & \\ 0.45 & -0.12 & 0.59 & 1.00 & & & & & \\ 0.03 & 0.49 & 0.03 & -0.08 & 1.00 & & & & \\ -0.29 & 0.43 & -0.13 & -0.21 & 0.47 & 1.00 & & & \\ -0.30 & 0.30 & -0.24 & -0.19 & 0.41 & 0.63 & 1.00 & & \\ 0.45 & -0.31 & 0.59 & 0.63 & -0.14 & -0.13 & -0.26 & 1.00 & \\ 0.30 & -0.17 & 0.32 & 0.37 & -0.24 & -0.15 & -0.29 & 0.40 & 1.00 \end{pmatrix}$$

- a. Perform a principal components analysis on these data and examine the associated scree plot to decide on the appropriate number of components.
- b. Apply maximum likelihood factor analysis and use the test described in the chapter to select the necessary number of common factors.
- c. How do the principal components and factor analysis solutions compare?
- d. Rotate the factor solution selected, using both an orthogonal and an oblique procedure, and interpret the results.

- 11.5 Using the correlations between the first eight statements of the previous exercise, fit a correlated two-factor model in which statements 1, 3, 4, and 8 are assumed to be indicators of a latent variable, doctor's responsibility, and statements 2, 5, 6, and 7 are assumed to be indicators of a latent variable, patient's responsibility. Find a 95% confidence interval for the correlation between the two latent variables.

12

Cluster Analysis

12.1 Introduction

An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.

Steven Pinker, *How the Mind Works*, 1997

One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. The idea of sorting similar objects into categories is clearly a primitive one because early humans, for example, must have been able to realize that many individual objects shared certain properties such as being edible, or poisonous, or ferocious, and so on. Further, classification in its widest sense is needed for the development of language, which consists of words that help us recognize and discuss the different types of events, objects, and people we encounter. Each noun in a language, for example, is essentially a label used to describe a class of objects that have striking features in common; thus, animals are called cats, dogs, horses, etc., and each name collects individuals into groups. Naming and classifying are essentially synonymous.

As well as being a basic human conceptual activity, classification of the phenomena being studied is an important component of virtually all scientific research. In the behavioral sciences, for example, these “phenomena” may be individuals or societies, or even patterns of behavior or perception. The investigator is usually interested in finding a classification in which the items of interest are sorted into a small number of homogeneous groups or clusters, the terms being synonymous. Most commonly, the required classification is one in which the groups are mutually exclusive (an item belongs to a single group) rather than overlapping (items can be members of more than one group). At the very least, any derived classification scheme should provide a convenient method of organizing a large, complex set of multivariate data with the class labels providing a parsimonious way of describing the patterns of similarities and differences in the data. In market research, for example, it might be useful to group a large number of potential customers according to their needs in a particular product area. Advertising campaigns

might then be tailored to the different types of consumers as represented by the different groups.

But often, a classification may seek to serve a more fundamental purpose. In psychiatry, for example, the classification of psychiatric patients with different symptom profiles into clusters might help in the search for the causes of mental illnesses and even perhaps to improved therapeutic methods. These twin aims of prediction (separating diseases that require different treatments) and etiology (searching for the causes of disease) for classifications will be the same in other branches of medicine.

Clearly, a variety of classifications will always be possible for whatever is being classified. Human beings could, for example, be classified with respect to economic status into groups labeled lower class, middle class, and upper class, or they might be classified by the annual consumption of alcohol into low, medium, and high. Clearly, different classifications may not collect the same set of individuals into groups, but some classifications will be more useful than others—a point made clearly by the following extract from Needham (1967), in which he considers the classification of human beings into men and women.

The usefulness of this classification does not begin and end with all that can, in one sense, be strictly inferred from it, namely, a statement about sexual organs. It is a very useful classification because classifying a person as man or woman conveys a great deal more information about probable relative size, strength, certain types of dexterity, and so on. When we say that persons in class man are more suitable than persons in class woman for certain tasks, and conversely, we are only incidentally making a remark about sex, our primary concern being strength, endurance, etc. The point is that we have been able to use a classification of persons that conveys information on many properties. On the contrary, a classification of persons into those with hairs on their forearms between $\frac{3}{16}$ and $\frac{1}{4}$ in. long and those without, though it may serve some particular use, is certainly of no general use, for imputing membership in the former class to a person conveys information on this property alone. Put another way, there are no known properties that divide up a set of people in a similar way.

In a similar vein, a classification of books based on subject matter into classes such as dictionaries, novels, biographies, and so on is likely to be far more useful than one based on, say, the color of the book's binding. Such examples illustrate that any classification of a set of multivariate data is likely to be judged on its usefulness.

It should be noted here that this chapter is concerned only with the problems of classifying previously unclassified material, and so begins with both the number of groups and their composition as unknowns. The situation when groups are known a priori, and the aims are to assess whether these groups differ on a set of variables or to derive a rule for classifying new individuals on the basis of their scores on these variables, is taken up in Chapter 13.

12.2 Cluster Analysis

Cluster analysis is a generic term for a wide range of numerical methods with the common goal of uncovering or discovering groups or clusters of observations that are homogeneous and separated from other groups. Clustering techniques essentially try to formalize what human observers do so well in two or three dimensions. Consider, for example, the scatterplot shown in Figure 12.1. The conclusion that there are two natural groups or clusters of dots is reached with no conscious effort or thought. Clusters are identified by the assessment of the relative distances between points, and, in this example, the relative homogeneity of each cluster and the degree of separation between the clusters makes the task very simple. The examination of scatterplots based either on the original data or perhaps on the first few principal component scores of the data is often a very helpful initial phase when intending to apply some form of cluster analysis to a set of multivariate data.

Cluster analysis techniques are described in detail in Gordon (1987, 1999) and Everitt et al. (2001). In this chapter, we give a relatively brief account of three types of clustering methods: agglomerative hierarchical techniques, k -means clustering, and model-based clustering.

12.3 Agglomerative Hierarchical Clustering

This class of clustering methods produces a hierarchical classification of data. In a hierarchical classification, the data are not partitioned into a particular number of classes or groups at a single step. Instead, the classification consists

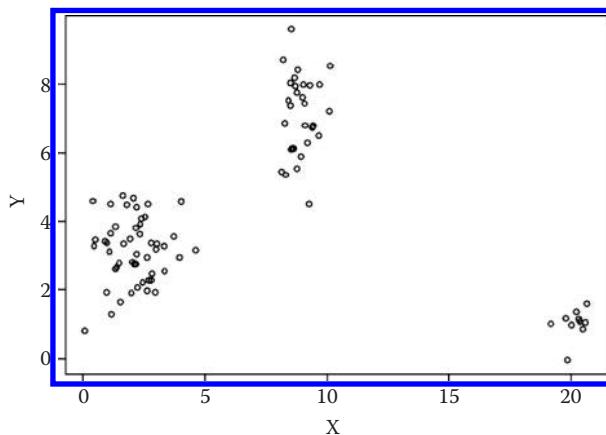


FIGURE 12.1

Bivariate data showing the presence of three clusters.

of a series of partitions that may run from a single “cluster” containing all individuals to n clusters, each containing a single individual. Agglomerative hierarchical clustering techniques produce partitions by a series of successive fusions of the n individuals into groups. With such methods, fusions, once made, are irreversible, so that when an agglomerative algorithm has placed two individuals in the same group, they cannot subsequently appear in different groups. Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, the investigator seeking the solution with the “best” fitting number of clusters will need to decide which division to choose. The problem of deciding on the “correct” number of clusters will be taken up later.

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, P_n, P_{n-1}, \dots, P_1 . The first, P_n , consists of n single-member clusters, and the last, P_1 , consists of a single group containing all n individuals. The basis operation of all methods is similar:

(START) Clusters C_1, C_2, \dots, C_n each containing a single individual.

1. Find the nearest pair of distinct clusters, say, C_i and C_j , merge C_i and C_j , delete C_j , and decrease the number of clusters by 1.
2. If number of clusters equals 1, then stop, else return to 1.

However, before the process can begin, an interindividual distance matrix or similarity matrix needs to be calculated. There are many ways to calculate distances or similarities between pairs of individuals, but here we only deal with a commonly used distance measure, namely, Euclidean distance, defined as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^q (x_{ik} - x_{jk})^2}$$

where d_{ij} is the Euclidean distance between individual i with variable values $x_{i1}, x_{i2}, \dots, x_{iq}$, and individual j with variable values $x_{j1}, x_{j2}, \dots, x_{jq}$. (Details of other possible distance measures and similarity measures are given in Everitt et al., 2001.) The Euclidean distances between each pair of individuals can be arranged in a matrix that is symmetric because $d_{ij} = d_{ji}$ and has zeros on the main diagonal. Such a matrix is the starting point of many clustering examples, although the calculation of Euclidean distances from the raw data may not be sensible when the variables are on very different scales. In such cases, the variables can be standardized in the usual way before calculating the distance matrix, although this can be unsatisfactory in some cases (see Everitt et al., 2001, for details).

Given an interindividual distance matrix, the hierarchical clustering can begin and, at each stage in the process, the methods fuse individuals or

groups of individuals formed earlier who are closest (or most similar). So, as groups are formed, the distance between an individual and a group containing several individuals, and the distance between two groups of individuals will need to be calculated. How such distances are defined leads to a variety of different techniques. Two simple intergroup measures are

$$d_{AB} = \min_{\substack{i \in A \\ i \in B}} (d_{ij})$$

$$d_{AB} = \max_{\substack{i \in A \\ i \in B}} (d_{ij})$$

where d_{AB} is the distance between two clusters A and B , and d_{ij} is the distance between individuals i and j found from the initial interindividual distance matrix.

The first intergroup distance measure above is the basis of single linkage clustering, the second that of complete linkage clustering. Both these techniques have the desirable property that they are invariant under monotone transformations of the original interindividual distances, that is, they only depend on the ranking on these distances, not their actual values.

A further possibility for measuring intercluster distance or dissimilarity is

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$

where n_A and n_B are the number of individuals in clusters A and B . This measure is the basis of a commonly used procedure known as group average clustering. All three intergroup measures described here are illustrated in [Figure 12.2](#).

Hierarchic classifications may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions made at each stage of the analysis. An example of such a diagram is given in [Figure 12.3](#). The structure of [Figure 12.3](#) resembles an evolutionary tree (see [Figure 12.4](#)), and it is in biological applications that hierarchical classifications are most relevant and most justified (although this type of clustering has also been used in many other areas).

12.3.1 Clustering Individuals Based on Body Measurements

As a first example of the application of the three clustering methods (single linkage, complete linkage, and group average), each will be applied to the chest, waist, and hip measurements of 20 individuals given in Chapter 9, Table 9.1. First Euclidean distances are calculated on the unstandardized measurements;

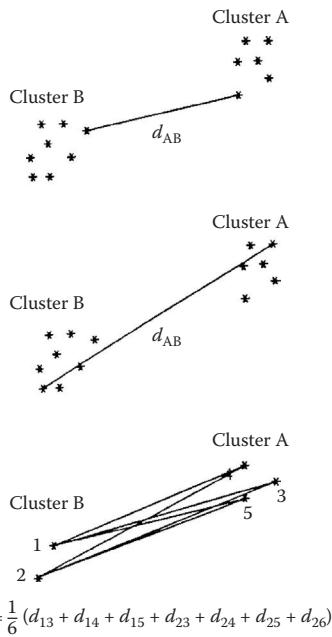


FIGURE 12.2
Intercluster distance measures.

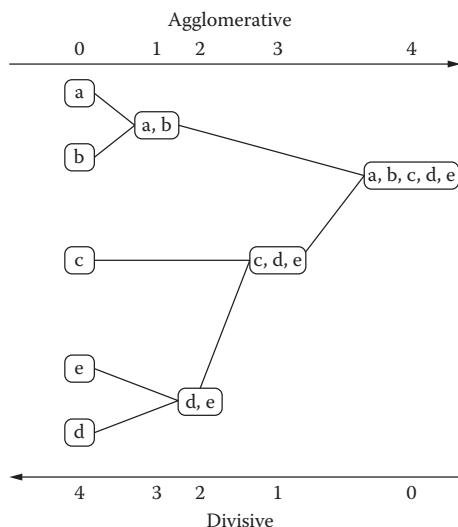
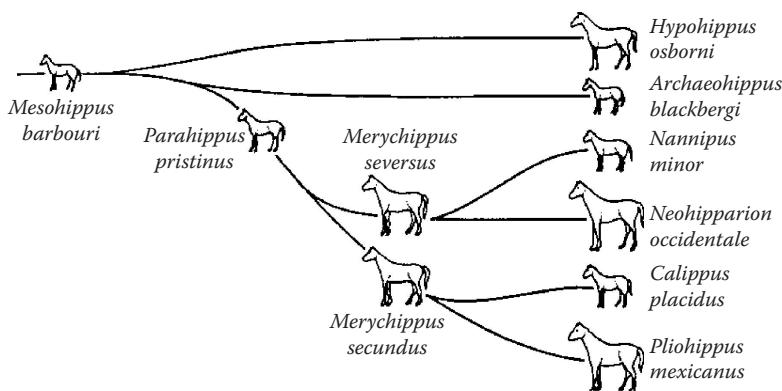
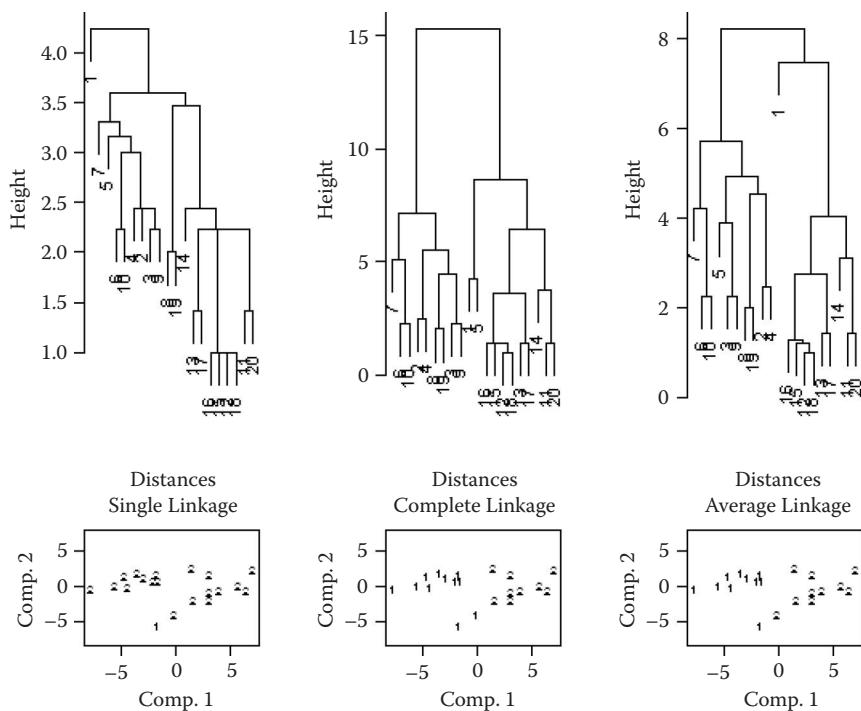


FIGURE 12.3
Example of a dendrogram.

**FIGURE 12.4**

Evolutionary tree. (From Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York. Used with permission of John Wiley & Sons.)

application of each of the three methods to this distance matrix gives the three dendograms shown in [Figure 12.5](#). How do we select specific partitions of the data from the complete dendograms? The answer is that we “cut” the dendrogram at some height, and this will give a partition with a particular number of groups. How do we choose where to cut or, in other words, how do we decide on a particular number of groups that is, in some sense, optimal for the data? One informal approach is to examine the sizes of the changes in height in the dendrogram and take a “large” change to indicate the appropriate number of clusters for the data. (More formal approaches are described in Everitt et al., 2001.) Even using this informal approach on the dendograms in [Figure 12.5](#), it is not easy to decide where to “cut.” So, instead, because we know that these data consist of measurements on 10 men and 10 women, we will look at the two-group solutions from each method that are obtained by cutting the dendograms at suitable heights. We can display and compare the three solutions graphically by plotting the first two principal component scores of the data, labeling the points to identify the cluster solution of one of the methods. Such plots are also shown in [Figure 12.5](#). The plot associated with the single linkage solution immediately demonstrates one of the problems with using this method in practice, and that is a phenomenon known as chaining, which refers to the tendency to incorporate intermediate points between clusters into an existing cluster rather than initiating a new one. As a result, single linkage solutions often contain long “straggly” clusters that do not give a useful description of the data. The two-group solutions from complete linkage and average linkage, also shown in [Figure 12.5](#), are similar and, in essence, place the men (observations 1 to 10) together in one cluster and women (observations 11 to 20) in the other.

**FIGURE 12.5**

Dendrograms for single linkage, complete linkage, and average linkage applied to body measurements data.

12.3.2 Clustering Countries on the Basis of Life Expectancy Data

Keyfitz and Flieger (1971) list life expectancies by age and by sex for a number of countries. In this section, we shall use the data for men, and Table 12.1 shows life expectancies for the first 5 of the 27 countries in the data set.

TABLE 12.1

Life Expectancies at Different Ages for Men in Five Countries

	Birth	Aged 25	Aged 50	Aged 75
Algeria	63	51	30	13
Cameroon	34	29	13	5
Madagascar	38	30	17	7
Mauritius	59	42	20	6
Reunion	56	38	18	7

The variances of the life expectancies at the four different ages are

$$\text{Birth variance} = 66.08$$

$$\text{Aged 25 variance} = 25.26$$

$$\text{Aged 50 variance} = 13.07$$

$$\text{Aged 75 variance} = 4.56$$

As might have been predicted, the variances are quite different, so calculating the initial intercountry Euclidean distance matrix on the life expectancies as they are in [Table 12.1](#) would not appear to be very sensible; instead, we will standardize each of the four life expectancies to have the variance of one and then calculate the required Euclidean distances on the standardized data. Here, we shall apply only complete linkage to the data, and the resulting dendrogram is shown in [Figure 12.6](#). Again, there is no completely “best” place to cut the dendrogram, but the four-group cluster solution produced by cutting at a height of 3 is shown in [Table 12.2](#). The countries grouped together in each cluster are perhaps different from what might have been expected from intuition. The three countries in group 1 have relatively high life expectancies at each age, and the two countries in group 2 have generally low life expectancies, particularly at birth, probably due to high infant mortality in these countries. Groups 3 and 4 have a similar pattern for their mean profiles, but countries in group 4 tend to have slightly greater life

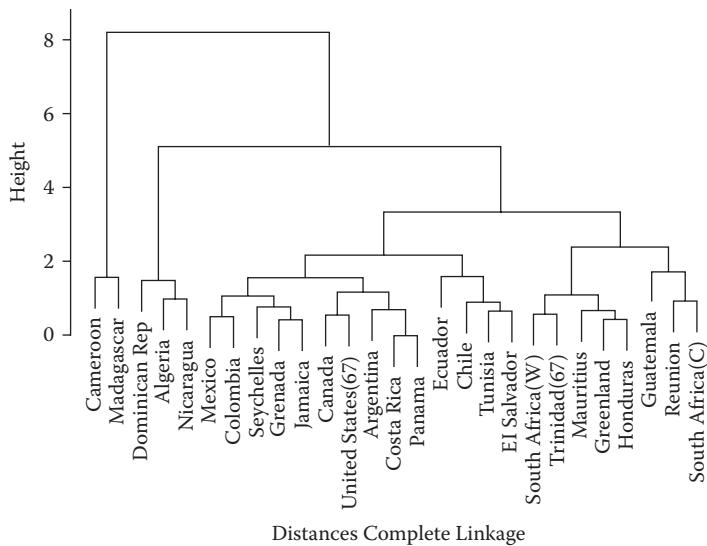


FIGURE 12.6

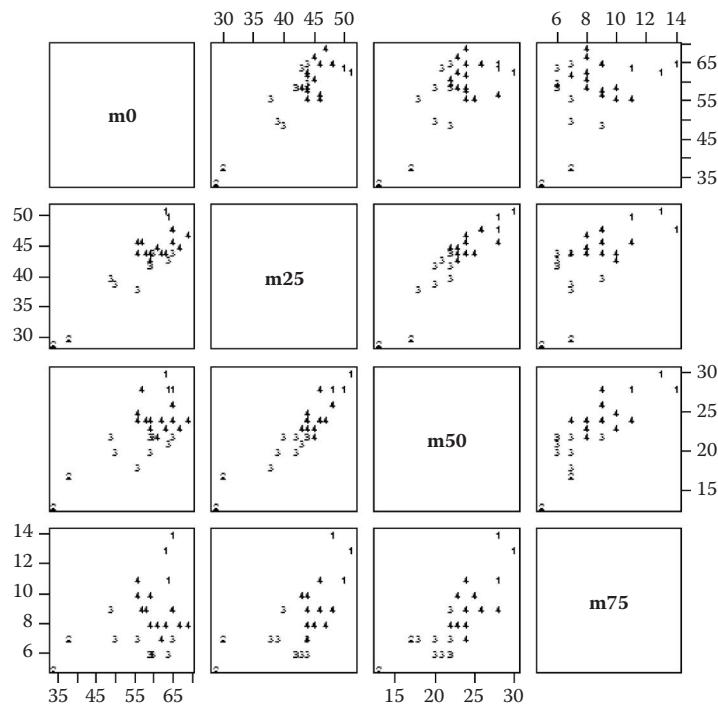
Average-linkage dendrogram for life expectancy data.

TABLE 12.2

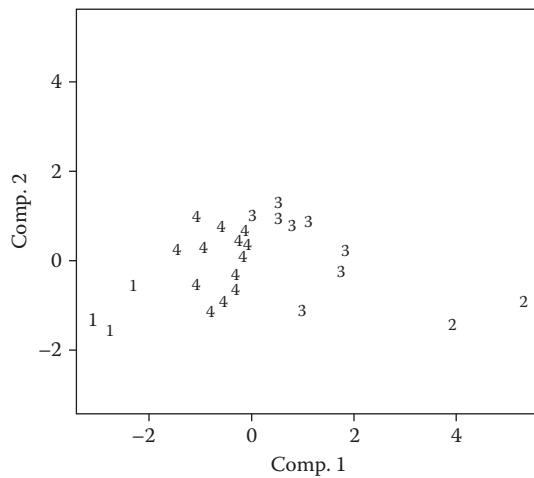
Four-Group Solution Produced by Complete Linkage Clustering Applied to the Life Expectancy Data

Countries and Mean Life Expectancies in Each Group				
Mean				
m0	m25	m50	m75	
64.0	49.7	28.7	12.7	
Group 2				
Algeria, Dominican Republic, Nicaragua				
Mean				
m0	m25	m50	m75	
36.0	29.5	15.0	6.0	
Group 3				
Mauritius, Reunion, South Africa (C), South Africa (W), Greenland, Guatemala, Honduras, Trinidad (67)				
Mean				
m0	m25	m50	m75	
57.75	41.50	20.88	6.75	
Group 4				
Seychelles, Tunisia, Canada, Costa Rica, El Salvador, Grenada, Jamaica, Mexico, Panama, United States (67), Argentina, Chile, Columbia, Ecuador				
Mean				
m0	m25	m50	m75	
61.57	45.29	24.29	8.79	

expectancies at each age. We can illustrate the cluster solution graphically in a number of ways. Here, we shall first plot the four-group cluster solution on the scatterplot matrix of the four life expectancies to give [Figure 12.7](#). Groups 1 and 2 are seen to be clearly separate, but groups 3 and 4 tend to merge into one another on some panels in the plot. A further graphic of the cluster solution is shown in [Figure 12.8](#), where the data are plotted in the space of the first two components and labeled by cluster membership. The first component is essentially the average of the four life expectancies, and the second contrasts life expectancy at birth with life expectancy at age 75. The first two components account for 66% of the variance of the four life expectancies. [Figure 12.8](#) demonstrates that the four clusters differ largely on their average life expectancy over the four ages.

**FIGURE 12.7**

Scatterplot matrix of life expectancy data showing the four-cluster solution from complete linkage.

**FIGURE 12.8**

Four-cluster solution from complete linkage applied to life expectancy data plotted in the space of the first two principal components.

12.4 *k*-Means Clustering

The *k*-means clustering technique seeks to partition the n individuals in a set of multivariate data into k groups or clusters, (G_1, G_2, \dots, G_k) , where G_i denotes the set of n_i individuals in the i th group, and k is given (or a possible range is specified by the researcher; the problem of choosing the “true” value of k will be taken up later) by minimizing some numerical criterion, low values of which are considered indicative of a “good” solution. The most commonly used implementation of *k*-means clustering is one that tries to find the partition of the n individuals into k groups that minimizes the within-group sum of squares (WGSS) over all variables; explicitly this criterion is

$$\text{WGSS} = \sum_{j=1}^q \sum_{l=1}^k \sum_{i \in G_l} (x_{ij} - \bar{x}_j^{(l)})^2 \quad \text{where } \bar{x}_j^{(l)} = \frac{1}{n_l} \sum_{i \in G_l} x_{ij}$$

is the mean of the individuals in Group G_l on variable j .

The problem then appears relatively simple; consider every possible partition of the n individuals into k groups, and select the one with the lowest within-group sum of squares. Unfortunately, the problem in practice is not so straightforward. The numbers involved are so vast that complete enumeration of every possible partition remains impossible even with the fastest computer. The scale of the problem is illustrated in the following table:

n	k	Number of Possible Partitions
15	3	2,375,101
20	4	45,232,115,901
25	8	690,223,721,118,368,580
100	5	10^{68}

The impracticability of examining every possible partition has led to the development of algorithms designed to search for the minimum values of the clustering criterion by rearranging existing partitions and keeping the new one only if it provides an improvement. Such algorithms do not, of course, guarantee finding the global minimum of the criterion. The essential steps in these algorithms are as follows:

1. Find some initial partition of the individuals into the required number of groups. (Such an initial partition could be provided by a solution from one of the hierarchical clustering techniques described in [Section 12.3](#).)

2. Calculate the change in the clustering criterion produced by “moving” each individual from its own to another cluster.
3. Make the change that leads to the greatest improvement in the value of the clustering criterion.
4. Repeat steps (2) and (3) until no move of an individual causes the clustering criterion to improve.

For a more detailed account of the typical k -means algorithm, see Steinley (2008).

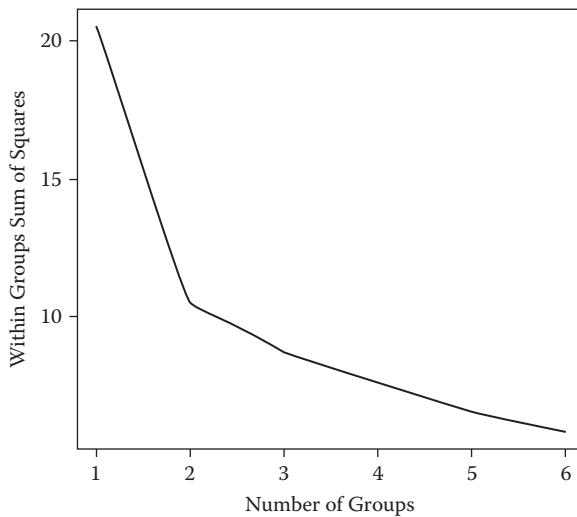
The k -means approach to clustering using the minimization of the WGSS over all the variables is widely used, but it suffers from two problems: (1) not being scale invariant, that is, different solutions may result from clustering the raw data and the data standardized in some way, and (2) imposing a spherical structure on the data, that is, it will find clusters shaped like hyper- footballs even if the “true” clusters in the data are of some other shape (see Everitt et al., 2001, for some examples of the latter phenomenon). Nevertheless, the k -means method remains very popular. With k -means clustering, the investigator can choose to partition the data into a specified number of groups. In practice, solutions for a range of values for number of groups are found and, in some way, the optimal or “true” number of groups for the data must be chosen. Several suggestions have been made as to how to answer the number of groups question, but none is completely satisfactory. The method we shall use in the forthcoming example is to plot the WGSS associated with the k -means solution for each number of groups. As the number of groups increases, the sum of squares will necessarily decrease, but an obvious “elbow” in the plot may be indicative of the most useful solution for the investigator to look at in detail. (Compare the scree plot described in Chapter 11.)

We shall illustrate the application of k -means clustering using the crime rate data introduced in Chapter 10 after removing the outlier, DC, identified in Chapter 10. If we first calculate the variances of the crime rates for the different types of crime we find the following:

	Murder	Rape	Robbery	Assault	Burglary	Theft
Variances	11.93	209.76	1889.53	19373.54	175895.00	565276.59

The variances are very different, and using k -means on the raw data would not be sensible; we must standardize the data in some way, and here we standardize each variable by its range. After such standardization, the variances become

	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
Variances	0.076	0.056	0.046	0.059	0.052	0.062	0.068

**FIGURE 12.9**

Plot of within-groups sum of squares against number of clusters.

The variances of the standardized data are very similar, and we can now progress with clustering the data. First, we plot the WGSS for one- to six-group solutions to see if we can get any indication of number of groups. The plot is shown in Figure 12.9. The only “elbow” in the plot occurs for two groups, and so we will now look at the two-group solution. In Table 12.3, the group membership and means are given. Everything is worse in group 1! A plot of the two-group solution in the space of the first two principal components of the correlation matrix of the data is shown in [Figure 12.10](#). The

TABLE 12.3

Group Membership and Means for the Two-Group Solution from k -Means Applied to Crime Rate Data

Group 1

MA, NY, NJ, IL, MI, MO, MD, NC, SC, GA, KY, AR, LA, OK, WY, CO, NM, UT, NV, WA, OR, CA

Mean Crime Rates

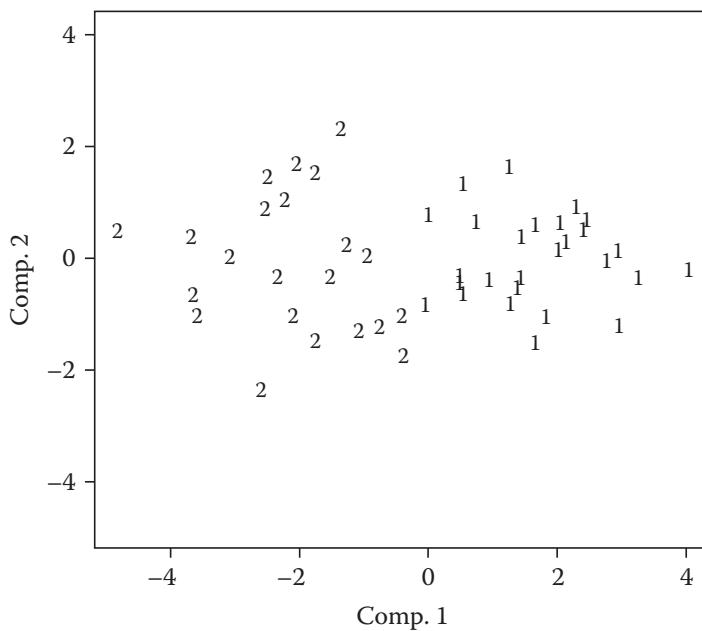
Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
9.37	45.37	229.00	394.77	1543.41	3368.05	554.27

Group 2

ME, NH, VT, RI, CT, PA, OH, IN, WI, MN, IA, ND, SD, NE, KS, DE, DC, VA, WV, FL, TN, AL, MS, TX, MT, ID, AZ, AK, HI

Mean Crime Rates

Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
4.74	24.80	73.82	182.07	924.21	2564.71	247.04

**FIGURE 12.10**

Plot of k -means two-group solution for standardized crime rate data.

two groups are created essentially on the basis of the first principal component score, which is, as we have seen in Chapter 10, a weighted average of the crime rates. Perhaps all that cluster analysis is doing here is dividing into two parts a homogenous set of data? This is always a possibility as discussed in some detail in Everitt et al. (2001).

12.5 Model-Based Clustering

The agglomerative hierarchical and k -means clustering methods described in Sections 12.3 and 12.4 are based largely on heuristic but intuitively reasonable procedures. However, they are not based on formal models for cluster structure in the data, making problems such as deciding between methods, estimating the number of clusters, etc., particularly difficult, and, of course, without a reasonable model, formal inference is precluded. In practice, these may not be insurmountable objections to the use of either the agglomerative methods or k -means clustering because cluster analysis is most often used as an “exploratory” tool for data analysis. However, if an acceptable model for cluster structure could be found, then cluster analysis based on the model might give more persuasive solutions (more persuasive to statisticians at least). A variety of possibilities have been proposed, but perhaps the

most successful approach is that proposed by Scott and Symon (1971) and extended by Banfield and Raftery (1993) and Fraley and Raftery (2002). In that approach, it is assumed that the population from which the observations arise consists of c subpopulations, each corresponding to a cluster, and that the probability density of a q -dimensional observation from the j th subpopulation is $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ for some unknown vector of parameters $\boldsymbol{\theta}_j$. Next, a vector $\boldsymbol{\gamma}' = [\gamma_1, \dots, \gamma_n]$ is introduced, where n is the sample size and $\gamma_i = k$ if the i th individual is from the k th subpopulation; the γ_i values label the subpopulation of each observation. The clustering problem now becomes that of choosing $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_c)$ and $\boldsymbol{\gamma}$ to maximize the likelihood function associated with the set of assumptions. The technique, which is known as the classification maximum likelihood procedure, is described briefly in Technical Section 12.1.

Technical Section 12.1: Classification Maximum Likelihood Clustering

Assume that the population consists of c subpopulations, each corresponding to a cluster of observations, and that the probability density function of a q -dimensional observation from the j th subpopulation is $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ for some unknown vector of parameters, $\boldsymbol{\theta}_j$. Also, assume that $\boldsymbol{\gamma}' = [\gamma_1, \dots, \gamma_n]$ gives the labels of the subpopulation to which each observation belongs, so $\gamma_i = j$ if \mathbf{x}_i is from the j th population. The clustering problem becomes that of choosing $\boldsymbol{\Theta}' = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_c]$ and $\boldsymbol{\gamma}$ to maximize the likelihood

$$L(\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \prod_{i=1}^n f_{\gamma_i}(\mathbf{x}_i; \boldsymbol{\theta}_{\gamma_i})$$

If $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ is assumed to be a multivariate normal density with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, then likelihood has the form

$$L(\boldsymbol{\Theta}; \boldsymbol{\gamma}) = \text{const} \prod_{k=1}^c \prod_{i \in E_k} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}$$

where $E_j = \{i : \gamma_i = j\}$ labels the set of observations in cluster j . The maximum likelihood estimator of $\boldsymbol{\mu}_j$ is $\hat{\mathbf{x}}_j = n_j^{-1} \sum_{i \in E_j} \mathbf{x}_i$, where \mathbf{x}_i and n_j are the number of elements in E_j . Replacing $\boldsymbol{\mu}_j$ in the likelihood with its maximum likelihood estimator yields the following log-likelihood:

$$\log L(\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \text{const} - \frac{1}{2} \sum_{i=1}^c \text{trace}(\mathbf{W}_j \boldsymbol{\Sigma}_j^{-1} + n_j \log |\boldsymbol{\Sigma}_j|)$$

where \mathbf{W}_j is the $q \times q$ matrix of sums of squares and cross products of the variables for subpopulation j . Banfield and Raftery (1993) demonstrate the following:

- If $\Sigma_k = \sigma^2 \mathbf{I}$ for $k = 1, 2, \dots, c$, that is, the clusters are spherical, then the log-likelihood is maximized by choosing γ to minimize $\text{trace}(\mathbf{W})$, where $\mathbf{W} = \sum_{k=1}^c \mathbf{W}_k$, that is, minimization of the MGSS, essentially equivalent to k -means clustering. Use of this criterion in a cluster analysis will tend to produce spherical clusters of largely equal sizes, matching those in the population. If, of course, the population clusters are not of this type, then a cluster structure may be imposed on the data and the real clusters in the data not found.
- If $\mathbf{S}_k = \mathbf{S}$ for $k = 1, 2, \dots, c$, that is, if all the clusters in the population have the same “shape” but not necessarily spherical, then the likelihood is maximized by choosing γ to minimize $|\mathbf{W}|$ (the determinant of \mathbf{W}), a clustering criterion discussed by Friedman and Rubin (1967) and Mariott (1982). The use of this criterion in a cluster analysis will tend to produce clusters with the same elliptical shape, again matching those in the population. Once again, if the population clusters are not of this type, then a cluster structure may be imposed on the data and the real clusters in the data not found.
- If the clusters in the population have different shapes that is, different covariance matrices, the likelihood is maximized by choosing γ to minimize $\sum_{k=1}^c n_k \log |\mathbf{W}_k| / n_k$.

Banfield and Raftery (1993) also consider criteria that allow the shape of clusters to be less constrained than the minimization of $\text{trace}(\mathbf{W})$ and $|\mathbf{W}|$ criteria, but that are more parsimonious than the model in which all the population clusters are allowed to have different shapes. For example, constraining clusters to be spherical but not to have the same volume, or constraining clusters to have diagonal covariance matrices but allowing their shapes, sizes, and orientations to vary. Details of the maximum likelihood estimation are given in Fraley and Raftery (2002).

Model selection is a combination of choosing both the appropriate clustering model for the population from which the n observations have been taken, that is, are all clusters spherical, all elliptical, all different shapes, or somewhere in between, and the optimal number of clusters. A Bayesian approach is adopted (see Fraley and Raftery, 1998, 2002), using what is known as the Bayesian information criterion (BIC). The result is a cluster solution that “fits” the observed data as well as possible, and this can include a solution that has only one “cluster,” implying that cluster analysis is not really a useful technique for the data.

To illustrate the use of the classification likelihood method, we will apply it to the data shown in [Table 12.4](#). These data arise from a study of what gastroenterologists

TABLE 12.4

Proportion of Respondents Answering Yes to Each of the Questions
in the Survey of Gastroenterologists

	Q1	Q2	Q3	Q4	Q5	Q6
Iceland	1.0000	1.000	1.0000	1.000	1.000	1.000
Norway	0.8571	0.833	1.0000	1.000	1.000	0.800
Sweden	1.0000	0.636	1.0000	1.000	0.500	0.667
Finland	1.0000	0.667	1.0000	1.000	0.833	0.667
Denmark	0.9231	0.692	1.0000	0.750	0.364	0.538
U.K.	0.6316	0.889	1.0000	0.950	0.526	1.000
Ireland	1.0000	0.667	1.0000	0.000	0.000	1.000
Germany	1.0000	1.000	1.0000	0.857	0.154	0.929
Netherlands	1.0000	1.000	1.0000	0.875	0.714	0.875
Belgium	0.0000	1.000	1.0000	0.500	0.000	1.000
Switzerland	1.0000	1.000	1.0000	0.500	0.000	1.000
France	0.3000	0.875	0.6250	0.200	0.000	0.875
Spain	0.0833	1.000	0.8000	0.545	0.000	1.000
Portugal	0.1667	1.000	0.6667	0.500	0.000	1.000
Italy	0.4667	1.000	0.9286	0.400	0.133	1.000
Greece	0.1250	1.000	0.6250	0.125	0.000	1.000
Yugoslavia	0.2667	1.000	0.5333	0.267	0.000	1.000
Albania	0.4000	0.600	0.4000	0.400	0.600	0.600
Bulgaria	0.0000	1.000	0.3333	0.000	0.000	1.000
Romania	0.0000	1.000	0.1429	0.143	0.143	1.000
Hungary	0.2000	1.000	0.8000	0.000	0.000	1.000
Czechoslovakia	0.0606	0.971	0.0882	0.000	0.000	0.571
Poland	0.0000	1.000	0.2632	0.105	0.000	0.947
Russia	0.0000	0.857	0.2857	0.000	0.000	0.857
Lithuania	0.0000	1.000	0.0000	0.000	0.000	1.000
Latvia	0.0000	1.000	0.0000	0.000	0.000	1.000
Estonia	0.6667	1.000	1.0000	0.000	0.000	1.000

in Europe tell their cancer patients (Thomsen et al., 1993). A questionnaire was sent to about 600 gastroenterologists in 27 European countries (the study took place before the recent changes in the political map of the continent) asking what they would tell a patient with newly diagnosed cancer of the colon, and his or her spouse, about the diagnosis. The respondent gastroenterologists were asked to read a brief case history and then to answer six questions with a yes or no answer. The questions were as follows:

- Q1: Would you tell this patient that he or she has cancer if he or she asks no questions?
- Q2: Would you tell the wife or husband that the patient has cancer?
(In the patient's absence.)

Q3: Would you tell the patient that he or she has a cancer if he or she directly asks you to disclose the diagnosis?

(During surgery, the surgeon notices several small metastases in the liver.)

Q4: Would you tell the patient about the metastases (supposing the patient asks to be told the results of the operation)?

Q5: Would you tell the patient that the condition is incurable?

Q6: Would you tell the wife or husband that the operation revealed metastases?

The data in [Table 12.4](#) give the proportion of respondents that answered each question "yes."

Applying the classification likelihood clustering approach to these data (see Fraley and Raftery, 1999), we can first examine the resulting plot of BIC values shown in Figure 12.11. In this diagram, the numbers refer to different model assumptions about the shape of clusters:

1. Spherical, equal volume
2. Spherical, unequal volume
3. Diagonal, equal volume and shape

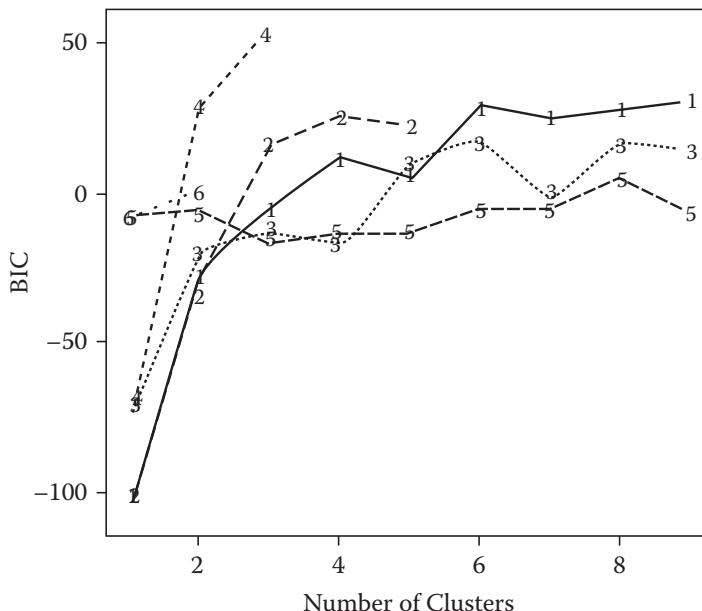


FIGURE 12.11

Plot of BIC values for a variety of cluster structures assumed by the classification likelihood clustering approach and a range of number of clusters.

TABLE 12.5

Results of the Classification-Likelihood Clustering Approach Applied to Cancer Questionnaire Data

Cluster Means						
	Q1	Q2	Q3	Q4	Q5	Q6
C1	0.88059	0.799	0.940	0.7823	0.5680	0.808
C2	0.32684	0.988	0.798	0.3035	0.0133	0.988
C3	0.00866	0.975	0.159	0.0355	0.0204	0.911
Cluster Membership						
<i>Cluster 1</i>						
Iceland, Norway, Sweden, Finland, Denmark, U.K., Ireland, Germany, Netherlands, Albania						
<i>Cluster 2</i>						
Belgium, Switzerland, France, Spain, Portugal, Italy, Greece, Yugoslavia, Hungary, Estonia						
<i>Cluster 3</i>						
Bulgaria, Romania, Czechoslovakia, Poland, Russia, Lithuania, Latvia						

4. Diagonal, varying volume and shape
5. Ellipsoidal, equal volume, shape, and orientation
6. Ellipsoidal, varying volume, shape, and orientation

The BIC criterion selects model 4 and three clusters as the optimal solution. Details of this solution are given in Table 12.5. The first cluster consists of countries in which the large majority of respondents gave yes answers to questions 1, 2, 3, 4, and 6, and about half also gave a yes answer to question 5. This cluster includes all the Scandinavian countries: the United Kingdom, Ireland, Germany, the Netherlands, and Albania. Ireland and Albania do not perhaps seem to be “natural” members of this group, but in both countries the number of respondents was small. In the second cluster, the majority of respondents answer “no” to questions 1, 4, 5 and “yes” to questions 2, 3, and 6; in these countries, it appears that the clinicians do not mind giving bad news to the spouses of patients but not to the patients themselves unless they are directly asked by the patient about his or her condition. This cluster contains Catholic countries such as Spain, Portugal, and Italy. In cluster three, the large majority of respondents answer no to questions 1, 3, 4, and 5, and again, a large majority answer yes to questions 2 and 6. In these countries very few clinicians appear to be willing to give the patient bad news even if asked directly by the patient about his or her condition.

12.6 Summary

- Cluster analysis techniques are used to search for clusters/groups in a priori unclassified multivariate data.

- Although clustering techniques are potentially very useful for the exploration of multivariate data, they require care in their application if misleading solutions are to be avoided.
 - Many methods of cluster analysis have been developed, and most studies have shown that no one method is best for all types of data. However, the more statistical techniques covered briefly in [Section 12.5](#) and in more detail in Everitt et al. (2001) have definite statistical advantages because the clustering is based on sensible models for the data.
 - Cluster analysis is a large area and has been covered only briefly in this chapter. The many problems that need to be considered when using clustering in practice have barely been touched upon. For a detailed discussion of these problems, see Everitt et al. (2001).
-

12.7 Exercises

- 12.1 Reanalyze the data on life expectancies without standardizing the variables. How do the results compare with those given in the text for the standardized data?
- 12.2 Apply k -means to the crime rate data after standardizing each variable by its standard deviation. Compare the results with those given in the text found by standardizing by a variable's range.
- 12.3 The data in exer_123.txt give the lowest temperatures in degrees Fahrenheit recorded in various months for cities in the United States. Plot the data in any way that you think might be helpful and explore whether there may be clusters of cities using some method of cluster analysis. Display graphically whatever cluster solutions you produce.
- 12.4 Apply the model-based clustering approach to the data on life expectancies and compare the results with those from the k -means clustering given in the text. Do the same for the crime rate data.
- 12.5 The data in exer_124.txt give the protein consumption in 25 European countries for 9 food groups. Is there any evidence that the countries cluster in some way?

13

Grouped Multivariate Data

13.1 Introduction

The importance of classification in science and in general and behavioral science in particular has already been remarked upon in Chapter 12, in which techniques were described for examining multivariate data to discover whether the data consisted of a number of relatively distinct groups or clusters of observations. In this chapter, a further aspect of classification will be discussed, namely that when the groups are known *a priori*. Such data arises when investigators collect samples of multivariate observations from several different populations, for example, observations on a number of symptoms for patients from different diagnostic categories.

A variety of questions might be asked about grouped multivariate data, and so, there are a variety of (overlapping) approaches to their analysis. In some cases, the investigator will simply be interested in testing whether the groups differ on the variables that have been recorded. When there are two groups, the multivariate analog of Student's t -test, Hotelling's T^2 , can be used, and when there are more than two groups, multivariate analysis of variance (MANOVA) is available. Both methods will be described later in the chapter.

A further question that is often of interest for grouped multivariate data is whether or not it is possible to use the measurements made to construct a classification rule derived from the original observations (the training set) that will allow new individuals having the same set of measurements (the test sample), but no group label, to be allocated to a group in such a way that misclassifications are minimized. The relevant technique is now some form of discriminant function analysis, which is the subject of [Section 13.2.2](#). A question that might be posed about constructing such an allocation rule is, "if group labels can be allocated *a priori* in some definitive fashion, why would we want to use the recorded variables for classification?" The answer might simply be "convenience" if definitive group labeling is costly or lengthy, or it might be "necessity," for example, in medicine if definitive group labeling can only be made by postmortem examination.

We begin by looking at the simplest case of grouped multivariate data, namely, when there are only two groups.

13.2 Two-Group Multivariate Data

13.2.1 Hotelling's T^2 Test

Willerman et al. (1991) collected data on 20 male and 20 female right-handed Anglo psychology students at a large university in the United States. The subjects took three subtests of the Wechsler Adult Intelligence Scale-Revised test. The scores recorded were full-scale IQ (FSIQ), verbal IQ (VIQ), and performance IQ (PIQ). The data for the first five men and the first five women are given in [Table 13.1](#).

Here, interest lies in testing the hypothesis that the three-dimensional mean vectors of IQ scores are the same for men and women. The appropriate test is Hotelling's T^2 , the multivariate analog of the independent samples t -test. The test and the assumptions on which it is based are described in Technical [Section 13.1](#).

Technical [Section 13.1](#): Hotelling's T^2

If there are q variables, the null hypothesis is that the means of the variables in the first population equal the means of the variables in the second population.

If $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of the two populations, the null hypothesis can be written as

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

The test statistic T^2 is defined as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

where n_1 and n_2 are the sample sizes in each group, and D^2 is the generalized distance introduced in Chapter 9, namely,

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the two sample mean vectors, and \mathbf{S} is the estimate of the assumed common covariance matrix of the two populations, calculated from the two sample covariance matrices \mathbf{S}_1 and \mathbf{S}_2 as

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

Note that the form of the test statistic in the multivariate case is very similar to that for the univariate independent samples t -test, involving a difference between "means" (here, mean vectors), and an assumed common "variance" (here, a covariance matrix). Under H_0 (and when the

TABLE 13.1
Wechsler Adult Intelligence IQ Scores
for Five Men and Five Women

Subject	FSIQ	VIQ	PIQ
1	140	150	124
2	139	123	150
3	133	129	128
4	89	93	84
5	133	114	147
6	133	132	124
7	137	132	134
8	99	90	110
9	138	136	131
10	92	90	98

Note: FSIQ = Full-scale IQ; VIQ = verbal IQ;
PIQ = performance IQ.

assumptions given below hold), the statistic F given by

$$F = \frac{(n_1 + n_2 - q - 1)T^2}{(n_1 + n_2 - 2)q}$$

has a Fisher's F -distribution with q and $n_1 + n_2 - q - 1$ degrees of freedom.

The T^2 test is based on the following assumptions:

- In each population, the variables have a multivariate normal distribution.
- The two populations have the same covariance matrix.
- The observations are independent.

Hotelling's T^2 takes the value 0.27, with the corresponding F -statistic being 0.09, having 3 and 36 degrees of freedom; the associated p-value is 0.97. There is no evidence of a sex difference on the three measures of IQ.

It might be thought that the results of Hotelling's T^2 test would simply reflect those that would be obtained using a series of univariate t -tests, in the sense that if no significant differences are found by the separate t -tests, then the T^2 test will inevitably lead to acceptance of the null hypothesis that the population mean vectors are equal. On the other hand, if any significant difference is found when using the t -tests on the individual variables, then the T^2 statistic must also lead to a significant result. But, these speculations are not correct (if they were, the T^2 test would be a waste of time). It is entirely possible to find no significant difference for each separate t -test but a significant result for the T^2 test, and vice versa. An explanation of how this can happen in the case of two variables is provided in Technical Section 13.2.

Technical Section 13.2: Univariate and Multivariate Tests for Equality of Means of Two Variables

Suppose we have a sample of n observations on two variables x_1 and x_2 , and we wish to test whether the population means of the two variables μ_1 and μ_2 are both 0. Assume that the mean and standard deviation of the x_1 observations are \bar{x}_1 and s_1 , respectively, and of the x_2 observations, \bar{x}_2 and s_2 . If we test separately whether each mean takes the value 0, then we would use two t -tests. For example, to test $\mu_1 = 0$ against $\mu_1 \neq 0$, the appropriate test statistic is

$$t = \frac{\bar{x}_1 - 0}{s_1 \sqrt{n}}$$

The hypothesis $\mu_1 = 0$ would be rejected at the α percent level of significance, if

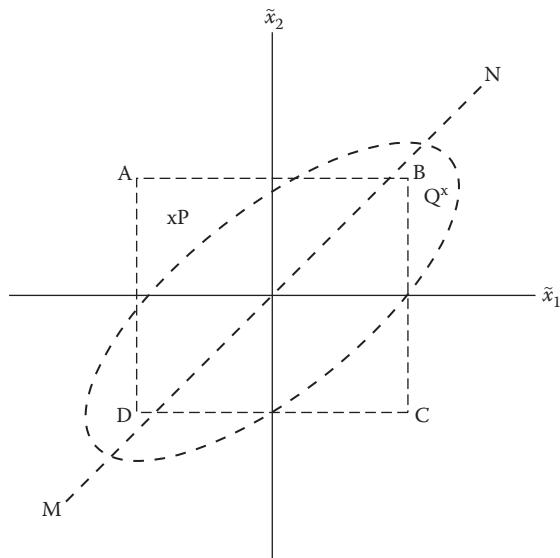
$$t < -t_{100(1-\frac{1}{2}\alpha)} \text{ or } t > t_{100(1-\frac{1}{2}\alpha)}$$

that is, if \bar{x}_1 fell outside the interval $[-s_1 t_{100(1-\frac{1}{2}\alpha)} / \sqrt{n}, s_1 t_{100(1-\frac{1}{2}\alpha)} / \sqrt{n}]$ where $t_{100(1-\frac{1}{2}\alpha)}$ is the $100(1-\frac{1}{2}\alpha)$ percent point of the t distribution with $n-1$ degrees of freedom. Thus, the hypothesis would not be rejected if \bar{x}_1 fell within this interval. Similarly, the hypothesis $\mu_2 = 0$ for the variable x_2 would not be rejected if the mean \bar{x}_2 of the x_2 observations fell within a corresponding interval with s_2 substituted for s_1 .

The multivariate hypothesis $[\mu_1, \mu_2] = [0, 0]$ would therefore not be rejected if both these conditions were satisfied. If we were to plot the point (\bar{x}_1, \bar{x}_2) against rectangular axes, the area within which the point could lie and the multivariate hypothesis not rejected is given by the rectangle ABCD of Figure 13.1, where AB and DC are of length $2s_1 t_{100(1-\frac{1}{2}\alpha)} \sqrt{n}$, while AD and BC are of length $2s_2 t_{100(1-\frac{1}{2}\alpha)} \sqrt{n}$.

Thus, a sample that gave the means (\bar{x}_1, \bar{x}_2) represented by the point P would lead to acceptance of the multivariate hypothesis. Suppose, however, that the variables x_1 and x_2 are moderately highly correlated. Then all points (x_1, x_2) , and hence, (\bar{x}_1, \bar{x}_2) , should lie reasonably close to the straight line MN through the origin marked on the diagram. Therefore, samples consistent with the multivariate hypothesis should be represented by points (\bar{x}_1, \bar{x}_2) that lie within a region encompassing the line MN. When we take account of the nature of the variation of bivariate normal samples that include correlation, this region can be shown to be an ellipse such as that marked on the diagram. The point P is not consistent with this region and, in fact, should be rejected for this sample. Thus, the inference drawn from the two separate univariate tests conflicts with the one drawn from a single multivariate test, and it is the wrong inference.

A sample giving the (\bar{x}_1, \bar{x}_2) values represented by point Q would give the other type of mistake, where the application of two separate

**FIGURE 13.1**

Why the results of univariate and multivariate tests can differ.

univariate tests leads to the rejection of the null hypothesis, but the correct multivariate inference is that the hypothesis should not be rejected. This explanation is taken with permission from Krzanowski (1991).

13.2.2 Fisher's Linear Discriminant Function

Spicer et al. (1987), in an investigation of sudden infant death (SID) syndrome, recorded four variables for each of 16 babies who were victims of SID and for 49 control babies. The babies who died and the control babies all had a gestational age of 37 weeks or more. Part of the data is shown in [Table 13.2](#). The factor 68 variable arises from a particular aspect of 24-h recordings or electrocardiograms and respiratory movements made for each child; the SID victims and the controls were matched for age at which these recordings were made. Here, interest lies in deriving a classification rule that could use measurements of the four variables on babies to be able to identify children at risk of SID and, if possible, take appropriate action to prevent the death of the baby.

The required classification rule can be constructed using Fisher's linear discriminant function, and this is described in Technical [Section 13.3](#).

TABLE 13.2
Part of the SIDs Data

Group	HR	BW	F68	GA
1	108.2	3000	0.321	37
1	131.1	4310	0.450	40
1	129.7	3975	0.244	40
1	142.0	3000	0.173	40
1	145.5	3940	0.304	41
2	139.7	3740	0.409	40
2	121.3	3005	0.626	38
2	131.4	4790	0.383	40
2	152.8	1890	0.432	38
2	125.6	2920	0.347	40

Note: Group = 1 for controls and 2 for SID victims; HR = heart rate (bpm); BW = birth weight (g); F68 = factor 68; GA = gestational age (weeks).

Technical Section 13.3: Fisher's Linear Discriminant Function

The aim is to find a way of classifying observations into one of two known groups using a set of variables, x_1, x_2, \dots, x_q . Fisher's idea was to find a linear function of the variables $z = a_1x_1 + a_2x_2 + \dots + a_qx_q$ such that the ratio of the between-group variance of z to its within-group variance is maximized. Therefore, the coefficients $\mathbf{a}' = [a_1, \dots, a_q]$ have to be chosen so that V , given by

$$V = \frac{\mathbf{a}' \mathbf{B} \mathbf{a}}{\mathbf{a}' \mathbf{S} \mathbf{a}}$$

is maximized, where \mathbf{S} is the pooled within-group covariance matrix, and \mathbf{B} the covariance matrix of group means defined as follows:

$$\begin{aligned}\mathbf{S} &= \frac{1}{n-2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)' \\ \mathbf{B} &= \sum_{i=1}^2 n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'\end{aligned}$$

where $\mathbf{x}'_{ij} = [x_{ij1}, x_{ij2}, \dots, x_{ijq}]$ represents the set of q variable values for the j th individual in group i , $\bar{\mathbf{x}}_j$ is the mean vector of the j th group, and $\bar{\mathbf{x}}$ is the mean vector of all observations. The number of observations in group 1 is n_1 , and in group 2 is n_2 , with $n = n_1 + n_2$. The vector \mathbf{a} that maximizes V is given by the solution of the following equation:

$$(\mathbf{B} - \lambda \mathbf{S})\mathbf{a} = 0$$

In the two-group situation, the single solution can be shown to be

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

The allocation rule is now to allocate an individual with discriminant score z to group 1 if $z > \frac{\bar{z}_1 + \bar{z}_2}{2}$, where \bar{z}_1 and \bar{z}_2 are the mean discriminant scores in each group. (We are assuming that the groups are labeled such that $\bar{z}_1 > \bar{z}_2$.)

Fisher's discriminant function also arises from assuming that, in the population, the observations in group 1 have a multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$, and those in group 2 have a multivariate distribution with mean vector $\boldsymbol{\mu}_2$ and, again, covariance matrix $\boldsymbol{\Sigma}$. Misclassifications are minimized if an individual with vector of scores \mathbf{x} is allocated to group 1 if $MVN(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) > MVN(\mathbf{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where MVN is shorthand for the multivariate normal density function. Substituting sample mean vectors for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the matrix \mathbf{S} defined earlier for $\boldsymbol{\Sigma}$, we are led to the same allocation rule as that given previously. But, the derived classification rule is only valid if the prior probabilities of being in each group are assumed to be the same. If the prior probability of group 1 is π_1 and that of group 2 is π_2 , then the new allocation rule becomes allocated to group 1 if $z > \frac{\bar{z}_1 + \bar{z}_2}{2} + \log \frac{\pi_2}{\pi_1}$.

To begin, we will use only the factor 68 variable and birth weight in applying Fisher's linear discriminant function to the SID data. We can first construct a scatterplot of the data (see Figure 13.2). The numerical details of the

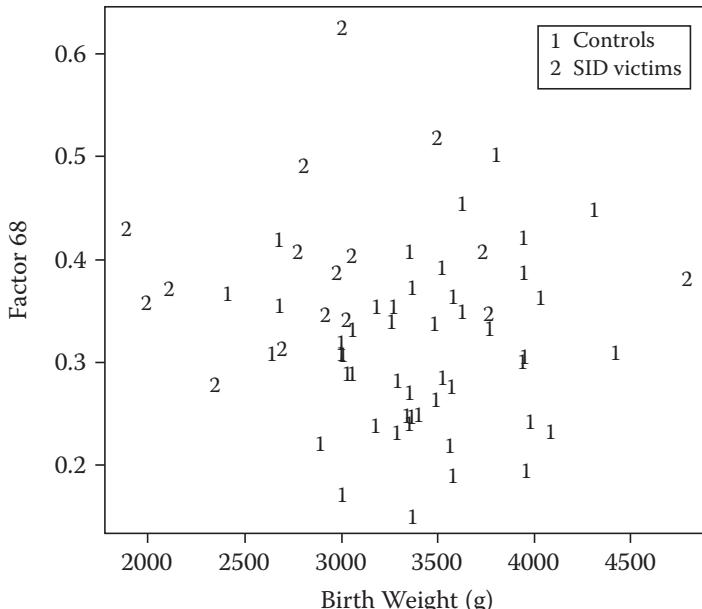


FIGURE 13.2

Scatterplot of the variables birth weight and factor 68 from the SID data.

TABLE 13.3

Calculating Fisher's Linear Discriminant Function on SID Data Using Birth Weight and Factor 68 Variables

Group	Controls		SID	
Means	BW	F68	BW	F68
	3437.88	0.31	2964.69	0.41
Covariance	BW	F68	BW	F68
Matrix	BW	1.95e+05	3.24	5.45e+05
	F68	3.24	0.006	7.76
	Pooled Covariance Matrix			
	BW	F68		
BW	278612.28	4.32		
F68	4.32	0.006		
	Coefficients of Discriminant Function			
BW	0.00195			
F68	−16.07705			
Discriminant function mean in controls				
1.6984				
Discriminant function mean in SID victims				
−0.6860				
Cutoff value				
0.5062				

Note: BW = Birth weight; F68 = factor 68.

calculations involved are shown in Table 13.3. The discriminant function is

$$z = 0.00195 \times \text{birth weight} - 16.077 \times \text{factor 68}$$

Assuming equal prior probabilities (unrealistic, but a point we shall return to later), the allocation rule for a new infant becomes "allocate to group 1 (little SID risk) if $z > 0.506$ (the "cutoff" value halfway between the discriminant function means of each group), otherwise allocate to group 2 (SID risk)." The discriminant function can be shown on the scatterplot of the birth weight and factor 68 variables simply by plotting the line $z - 0.506 = 0$, that is, a line with intercept given by $0.506/a_2$ and a slope of $-a_1/a_2$, where a_1 and a_2 are the discriminant function coefficients, to give Figure 13.3. In terms of this plot, a new infant with values of the two variables leading to a position on the plot above the line would be allocated to the SID risk group, and an infant with a position below the line to the little SID risk group.

A deficiency of the derived allocation rule is that it takes no account of the prior probabilities of class membership in the population under study. Therefore, if used as a screening device for babies at risk of SID in the simple form suggested here many more infants would be considered at risk than is genuinely merited because, fortunately, SID is known to be a relatively rare condition.

A question of some importance about a discriminant function allocation rule is “how well does it perform?” One way this question could be answered is to see how many of the original sample of observations (the training set) it misclassifies. In the case of the discriminant function for the SID data derived here based on birth weight and factor 68, the results of applying it to the data from which it was calculated are

	Actual Group	Allocation Rule Group
	Controls	SID
Controls	41	8
SID	3	13

So, the percentage of misclassifications is 14.67. This method of estimating the misclassification rate is known to be optimistic in many cases. Other more realistic methods for estimating the misclassification rate are described in Everitt and Dunn (2001). (Finding Fisher’s linear discriminant function based on all four variables recorded in the SID data is left as an exercise for the reader—see Exercise 13.3.)

Fisher’s linear discriminant function is optimal when the data arise from populations having multivariate normal distributions with the same covariance matrices. When the distributions are clearly nonnormal, an alternative

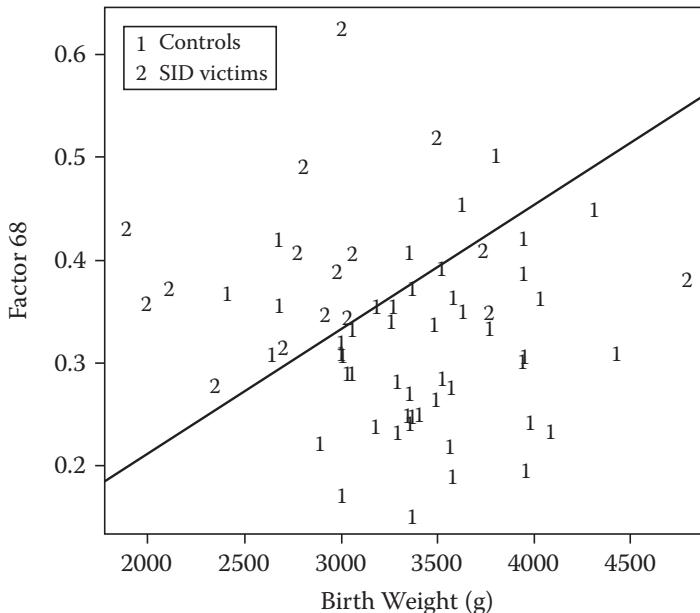


FIGURE 13.3

Scatterplot of factor 68 against birth weight for SID data, showing Fisher’s linear discriminant function based on the two variables.

approach is logistic discrimination (see, for example, Anderson, 1972), although the results of both this and Fisher's method are likely to be very similar in most cases. When the two covariance matrices are thought to be unequal, then the linear discriminant function is no longer optimal, and a quadratic version may be needed. Details are given in Everitt and Dunn (2001). The quadratic discriminant function has the advantage of increased flexibility compared to the linear version. There is, however, a penalty involved in the form of potential overfitting, making the derived function poor at classifying new observations. Friedman (1998) attempts to find a compromise between the data variability of quadratic discrimination and the possible bias of linear discrimination by adopting a weighted sum of the two, called regularized discriminant analysis.

13.3 More Than Two Groups

13.3.1 Multivariate Analysis of Variance (MANOVA)

Timm (2002) reports the data collected in a large study by Dr. Stanley Jacobs and Mr. Ronald Hritz at the University of Pittsburgh to investigate risk-taking behavior. Students were randomly assigned to three different direction treatments known as Arnold and Arnold (AA), Coombs (C), and Coombs with no penalty (NC) in the direction. Using the three treatment conditions, students were administered two parallel forms of a test given under low and high penalty. Part of the data is shown in Table 13.4. The question of interest here is whether the two-dimensional population mean vectors for the three groups are the same. The technique to be used is MANOVA, which is an extension of univariate analysis of variance to multivariate observations. A short account of one-way MANOVA is given in Technical Section 13.4, but MANOVA can, of course, be used with more complex designs when the response is multidimensional.

TABLE 13.4
Part of Data from Investigation of Risk Taking

AA		C		NC	
Low	High	Low	High	Low	High
8	28	46	13	50	55
18	28	26	10	57	51
8	23	47	22	62	52

Note: AA = Arnold and Arnold; C = Coombs; NC = Coombs with no penalty.

Technical Section 13.4: One-Way MANOVA

We assume that we have multivariate observations of a sample of individuals from m different populations, where $m \geq 2$, and there are n_i observations sampled from population i . The linear model for observation x_{ijk} , the j th observation on variable k in group i , $k = 1 \cdots q, j = 1 \cdots n_i, i = 1 \cdots m$ is

$x_{ijk} = \mu_k + \alpha_{ik} + \varepsilon_{ijk}$, where μ_k is a general effect for the k th variable, α_{ik} is the effect of group i on the k th variable, and ε_{ijk} is a random disturbance term. The vector $\boldsymbol{\varepsilon}_{ij} = [\varepsilon_{ij1} \cdots \varepsilon_{ijq}]$ is assumed to have a multivariate normal distribution with null mean vector and covariance matrix $\boldsymbol{\Sigma}$, assumed to be the same in all m populations. The error terms of different individuals are assumed independent of one another.

The hypothesis of equal mean vectors in the m populations can be written as

$$H_0 : \alpha_{ik} = 0, i = 1 \cdots m, k = 1 \cdots q$$

MANOVA is based on two matrices \mathbf{H} and \mathbf{E} , the elements of which are defined as follows:

$$h_{rs} = \sum_{i=1}^k n_i (\bar{x}_{ir} - \bar{x}_r)(\bar{x}_{is} - \bar{x}_s), r, s = 1, \cdots q$$

$$e_{rs} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ijr} - \bar{x}_{ir})(\bar{x}_{ijs} - \bar{x}_{is}), r, s = 1, \cdots q$$

where \bar{x}_{ir} is the mean of variable r in group i , and \bar{x}_r is the grand mean of variable r . The diagonal elements of \mathbf{H} and \mathbf{E} are, respectively, the between-groups sum of squares for each variable and the within-group sum of squares for the variable. The off-diagonal elements of \mathbf{H} and \mathbf{E} are the corresponding sums of cross products for pairs of variables. In the multivariate situation when $m > 2$, there is no single test statistic that is always the most powerful one for detecting all types of departures from the null hypothesis of the mean vectors of the populations. A number of different test statistics have been proposed that may lead to different conclusions when used in the same data set, although on most occasions they will not. The following are the principal test statistics for MANOVA:

a. Wilks' determinantal ratio

$$\Lambda = |\mathbf{E}| / |\mathbf{H} + \mathbf{E}|$$

b. Roy's greatest root: Here, the criterion is the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$.

c. Lawley–Hotelling trace

$$t = \text{trace}(\mathbf{E}^{-1}\mathbf{H})$$

d. Pillai trace

$$v = \text{trace}[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}]$$

Each test statistic can be converted into an approximate F -statistic that allows associated p -values to be calculated. For details, see Tabachnick and Fidell (1989).

When there are only two groups, all four test criteria are equivalent and lead to the same F -value as Hotelling's T^2 described in Technical Section 13.1.

Prior to any formal analysis of the data from the risk-taking investigation, it is useful to look at some boxplots, and these are given in Figure 13.4. The "Low" scores appear to increase across the three groups, with the group differences on the "High" score being rather smaller. Applying MANOVA to the data from the investigation of risk taking, we get the results shown in Table 13.5. Clearly, the two-dimensional mean vectors of low and high scores differ in the three groups.

The tests applied in MANOVA assume multivariate normality for the error terms in the corresponding model. An informal assessment of this assumption can be made using the chi-square plot described in Chapter 2, applied

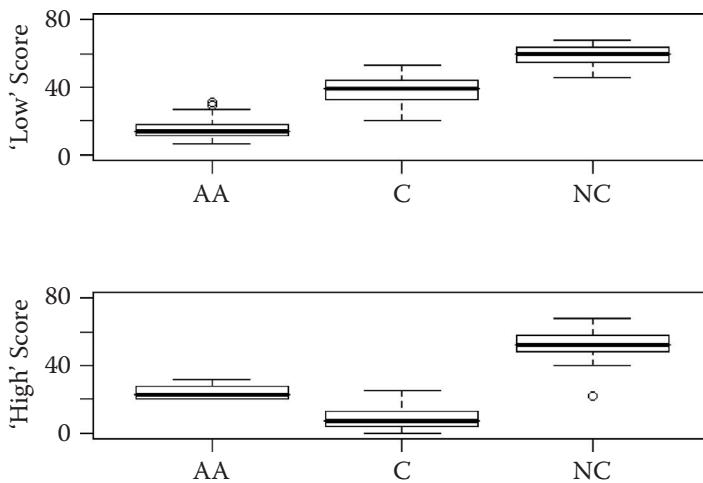
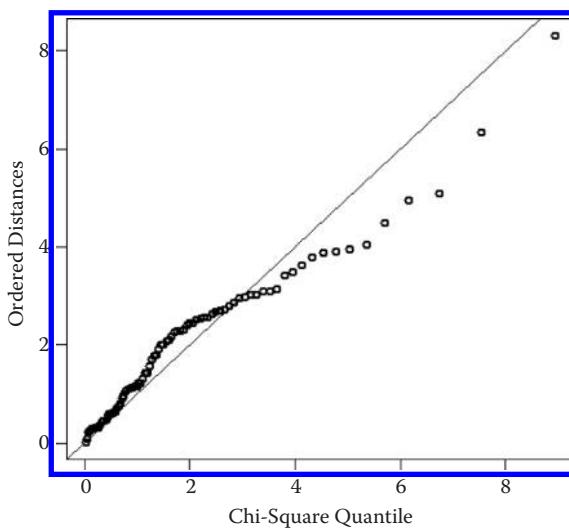


FIGURE 13.4
Boxplots for data from risk-taking experiment.

TABLE 13.5
MANOVA on Data from the Risk-Taking Investigation

	Df	Pillai	approx. F	num DF	den Df	Pr(>F)
Group	1	0.866	268.323	2	83	<2.2e-16
		Wilks				
Group	1	0.134	268.323	2	83	<2.2e-16
		Hotelling-Lawley				
Group	1	6.466	268.323	2	83	<2.2e-16
		Roy				
Group	1	6.466	268.323	2	83	<2.2e-16

**FIGURE 13.5**

Chi-square plot of residuals from fitting one-way MANOVA to data from risk-taking experiment.

to the residuals from fitting the one-way MANOVA model; note that the residuals in this case are each two-dimensional vectors. The plot is shown in Figure 13.5. There is some evidence of departure from the multivariate normal, but the p-values in Table 13.5 are so small that minor departures from the distributional assumption are unlikely to change the conclusions.

13.3.2 Classification Functions

In this section we will use data generated during a functional magnetic resonance imaging (fMRI) investigation. Two measures of intensity of each voxel in an image were recorded—PD and T_2 . Part of the data is shown in Table 13.6. One aim of the investigation was to derive a rule for allocating each voxel in an image into one of three classes: grey matter, white matter, or cerebrospinal fluid (CSF). When more than two groups are involved, we can again derive

TABLE 13.6
Part of fMRI Data

Class	PD	T_2
Grey	124	58
Grey	107	44
White	142	122
White	144	148
CSF	98	45
CSF	87	34

Note: CSF = Cerebrospinal fluid.

classification functions by comparing the assumed multivariate normal densities for each group. Technical Section 13.5 explains how.

Technical Section 13.5: Discriminant Analysis for Three Groups

Assuming that the observations in the three groups have multivariate normal densities with different means, μ_1 , μ_2 and μ_3 but a common covariance matrix, S , the allocation rule for an individual with vector of scores x becomes;

Allocate to group 1 if

$$MVN(x, \mu_1, S) > MVN(x, \mu_2, S)$$

and

$$MVN(x, \mu_1, S) > MVN(x, \mu_3, S)$$

Allocate to group 2 if

$$MVN(x, \mu_2, S) > MVN(x, \mu_1, S)$$

and

$$MVN(x, \mu_2, S) > MVN(x, \mu_3, S)$$

Allocate to group 3 if

$$MVN(x, \mu_3, S) > MVN(x, \mu_1, S)$$

and

$$MVN(x, \mu_3, S) > MVN(x, \mu_2, S)$$

This leads to sample-based allocation rules as follows:

Allocate to group 1 if $h_{12}(x) > 0$ and $h_{13}(x) > 0$

Allocate to group 2 if $h_{12}(x) < 0$ and $h_{23}(x) > 0$

Allocate to group 3 if $h_{13}(x) < 0$ and $h_{23}(x) < 0$

where

$$h_{ij}(x) = (\bar{x}_i - \bar{x}_j)'S^{-1} \left[x - \frac{1}{2}(\bar{x}_i + \bar{x}_j) \right]$$

and \bar{x}_i and \bar{x}_j are group mean vectors, S is the sample estimate of the assumed common covariance matrix of the three groups and is given by

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3}{n_1 + n_2 + n_3 - 3}$$

and \mathbf{S}_1 , \mathbf{S}_2 , and \mathbf{S}_3 are the estimates of the covariance matrices of each group.

To begin, we can plot the data labeling the three classes. The plot is shown in Figure 13.6. To find the three discriminant functions, we first need to find the mean vectors and covariance matrices of each class and

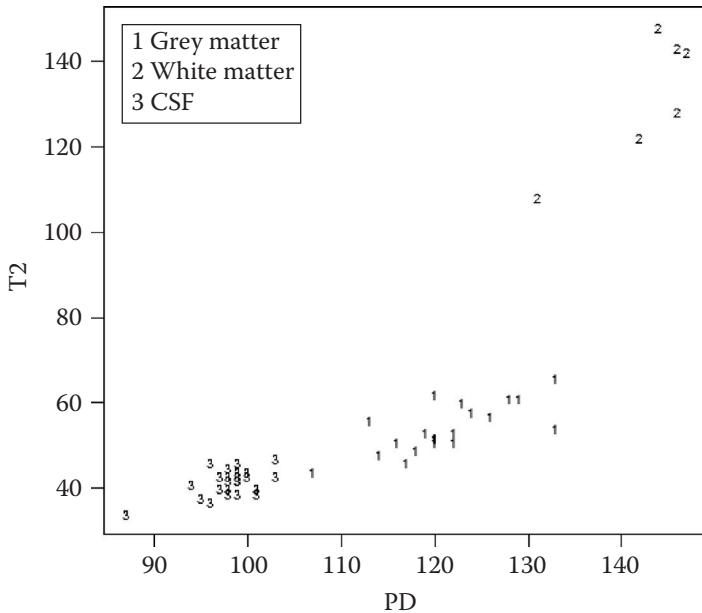


FIGURE 13.6

Scatterplot of imaging data with three classes labeled.

TABLE 13.7

Means and Covariance Matrices of Each Class in Imaging Data

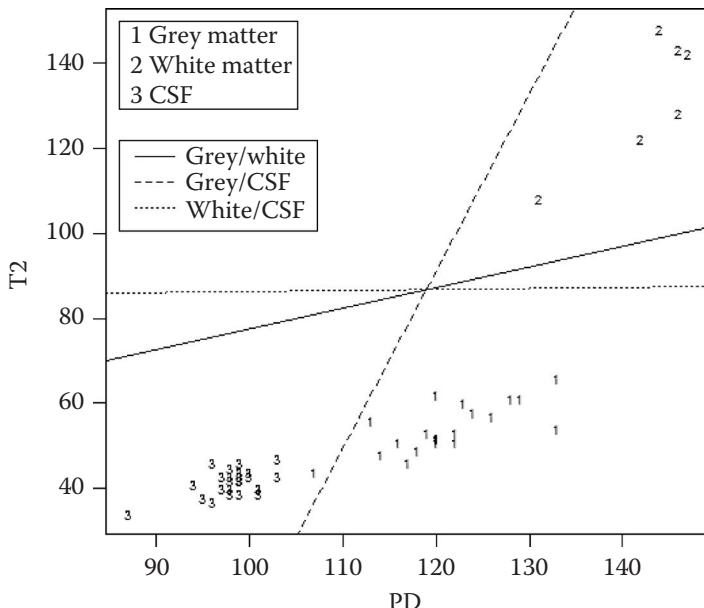
Grey Matter ($n = 20$)		White Matter ($n = 6$)		CSF ($n = 24$)	
	PD		PD		T ₂
Mean	121.20	54.25	142.67	131.83	98.08
Covariance Matrix					
PD	42.48	26.95	35.87	74.93	10.34
T ₂	26.95	33.25	74.93	233.77	5.77
Pooled Covariance Matrix					
PD		26.05		21.69	
T ₂		21.69		43.02	

TABLE 13.8

Discriminant Functions and Thresholds for Imaging Data

	Grey Matter	White Matter	CSF			
	PD	T ₂	PD	T ₂	PD	T ₂
Discriminant coefficients	1.17	-2.39	1.11	-0.27	-0.06	2.13
Thresholds						
Grey v White		Grey v CSF		White v CSF		
-68.54		108.86		177.39		

then the pooled covariance matrix; all are shown in [Table 13.7](#). The coefficients of each linear discriminant function and the thresholds calculated from the information in [Table 13.7](#) are shown in Table 13.8. Each of the discriminant functions can now be shown on the scatterplot of the data using the same approach as that used for a discriminant function for two groups. The scatterplot showing the three discriminant functions is shown in Figure 13.7. This plot would allow an investigator to classify new, unlabeled voxels, although in practice the three discriminant functions would need to be calculated from a much larger sample of previously labeled voxels.

**FIGURE 13.7**

Scatterplot of imaging data showing the three discriminant functions.

13.4 Summary

- Grouped multivariate data frequently occur in practice.
 - The appropriate method of analysis depends on the question of most interest to the investigator.
 - Hotelling's T^2 and MANOVA can be used to assess hypotheses about population mean vectors.
 - Fisher's linear discriminant function can be used to construct formal classification rules for allocating new individuals into one of two a priori known groups. And similar rules can be found when there are more than two groups.
 - Full details of discriminant function methods are given in Hand (1998).
-

13.5 Exercises

- 13.1 Return to the body measurement data introduced in Chapter 9, and find Fisher's linear discriminant function for allocating individuals to be men and women (I am aware that there is a foolproof method). (The first 10 observations in the data set are men, the remaining 10 are women.) Construct a scatterplot of the data showing the group membership and the derived linear discriminant function.
- 13.2 For the data from the risk-taking investigation, produce a scatterplot of "High" versus "Low" scores showing the three groups. As an exercise, construct discriminant functions for each pair of groups, and show these on the scatterplot.
- 13.3 In the SID data by coding a variable, group, as controls = 1 and SID victims = -1, show the equivalence of a multiple regression model for group as the response variable and explanatory variables HR, BW, F68, and GA with the discriminant function for the data derived in the text.
- 13.4 The data in exer_134.txt show the chemical composition for nine oxides of 48 specimens of Romano-British pottery determined by atomic absorption spectra. Also, given in the data file is the label of the kiln site at which the pot was found. Use MANOVA to test whether the pots found at different kiln sites differ in their chemical compositions. The five kiln sites are actually from three different

regions with kiln 1 from region 1, kilns 2 and 3 from region 2, and kilns 4 and 5 from region 3. Find the allocation rule for allocating a new pot to one of the three regions, and use the rule on a pot with the following chemical composition:

Al ₂ O ₃	Fe ₂ O ₃	MgO	CaO	Na ₂ O	K ₂ O	TiO ₂	MnO	BaO
15.5	5.71	2.07	0.98	0.65	3.01	0.76	0.09	0.012

References

- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York.
- Aitkin, M. (1978). The analysis of unbalanced cross-classification. *Journal of the Royal Statistical Society, Series A*, 141, 195–223.
- Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19–35.
- Banfield, J. D. and Raftery, A. E. (1993). Model based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.
- Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Oxford University Press, Oxford.
- Beck, A. T., Steer, A. and Brown, G. K. (1996). *Beck Depression Inventory Manual*. The Psychological Corporation, San Antonio, TX.
- Becker, R. A. and Cleveland, W. S. (1994). *S-PLUS Trellis Graphics User's Manual*. Mathsoft, Seattle, WA.
- Bertin, J. (1981). *Semiology of Graphics*. University of Wisconsin Press, WI.
- Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187, 398–404.
- Blackith, R. E. and Reymont, R. A. (1971). *Multivariate Morphometrics*. Academic Press, London.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, New York.
- Caslyn, J. R. and Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology*, 69, 136–145.
- Cattell, R. B. (1965). Factor analysis: An introduction to essentials. *Biometrics*, 21, 190–215.
- Carpenter, J., Pocock, S. and Lamm, C. J. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine*, 21, 1043–1066.
- Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Chapman and Hall/CRC, London.
- Chatterjee, S., Hadji, A. S. and Price, B. (1999). *Regression Analysis by Example*, 3rd edition. John Wiley & Sons, New York.
- Cleveland, W. S. (1985). *The Elements of Graphing Data*. Hobart Press, Summit, NJ.
- Cleveland, W. S. (1994). *The Elements of Graphing Data*. Hobart Press, Summit, NJ.
- Cleveland, W. S. and McGill, M. E. (1987). *Dynamic Graphics for Statistics*. Wadsworth, Belmont, CA.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, 128, 134–155.
- Collett, D. (2003a). *Modelling Binary Data*. 2nd edition. Chapman and Hall/CRC, London.

- Collett, D. (2003b). *The Analysis of Survival Data*, 2nd edition. Chapman and Hall/CRC, London.
- Colman, A. M. (Ed.). (1994). *The Companion Encyclopedia of Psychology*. Routledge, London.
- Cook, T.D. and Campbell, D.T. (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings, Houghton-Mifflin, Boston.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall/CRC, London.
- Cox, D. R. (1972). Regression models and life table. *Journal of the Royal Statistical Society, Series B*, 34, 187–200.
- Crowder, M. J. and Hand, D. J. (1990). *Analysis of Repeated Measurements*. Chapman and Hall/CRC, London.
- Davis, C.S. (2002) *Statistical Methods for the Analysis of Repeated Measurements*, Springer, New York.
- Diggle, P. J. and Kenward, M. G. (1994). Informative dropout in longitudinal studies. *Applied Statistics*, 43, 49–93.
- Diggle, P. J., Heagerty, K., Liang, K., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edition. Oxford University Press, Oxford.
- Dizney, H. and Gromen, L. (1967). Predictive validity and differential achievement on three MLA comparative foreign language tests. *Educational and Psychological Measurement*, 27, 1127–1130.
- Dobson, A. J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, London.
- Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall, London.
- Everitt, B. S. (1987). *An Introduction to Optimization Methods and their Applications in Statistics*. Chapman and Hall, London.
- Everitt, B. S. (2005). *An R and S-PLUS Companion to Multivariate Analysis*. Springer, New York.
- Everitt, B. S. and Dunn, G. (2001). *Applied Multivariate Data Analysis*. Edward Arnold, London.
- Everitt, B. S. and Hothorn, T. (2008). *A Handbook of Statistical Analyses Using R*. 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.
- Everitt, B. S. and Rabe-Hesketh, S. (2001). *Analysing Medical Data Using S-PLUS*. Springer, New York.
- Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis*, 4th edition. Edward Arnold, London.
- Everitt, B. S. and Wessely, S. (2008). *Clinical Trials in Psychiatry*. John Wiley & Sons, Chichester, U.K.
- Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, New York.
- Fraley, C. and Raftery, A. E. (1998). How many clusters? Which cluster method? Answers via model-based cluster analysis. *Computer Journal*, 41, 578–588.
- Fraley, C. and Raftery, A. E. (1999). MCLUS: Software for the model-based cluster analysis. *Journal of Classification*, 16, 297–306.
- Fraley, C. and Raftery, A. E. (2002). Model based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97, 611–631.
- Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159–1178.

- Friedman, J.H. (1989) Regularized discriminant analysis, *Journal of the American Statistical Association*, 84, 165–175.
- Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than P-values: Estimation rather than hypothesis testing. *British Medical Journal*, 292, 746–750.
- Goldberg, B. P. (1972). *The Detection of Psychiatric Illness by Questionnaire*. Oxford University Press, Oxford.
- Goldberg, K. M. and Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, 34, 307–320.
- Gordon, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A*, 150, 119–137.
- Gordon, A. D. (1999). *Classification*, 2nd edition. Chapman and Hall/CRC, London.
- Hand, D.J. (2005) Discriminant analysis, linear, In *Encyclopedia of Biostatistics*, 2nd edition, (eds. P. Armitage and T. Colton), Wiley, Chichester.
- Heitjan, D.F. (1997) Bayesian interim analysis of phase II cancer clinical trials, *Statistics in Medicine*, 16, 1791–1802.
- Hendrickson, A. E. and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, 17, 65–70.
- Heywood, H.B. (1931) On finite sequences of real numbers, *Proceedings of the Royal Statistical Society, Series A*, 134, 486–501.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441.
- Howell, D. C. (2002). *Statistical Methods for Psychology*. Duxbury Press, Belmont, CA.
- Howell, D. C. and Huessy, H. R. (1981). Hyperkinetic behavior followed from 7 to 21 years of age. In *Intervention Strategies with Hyperactive Children*, Ed. M. M. Gittleman. M. E. Sharp, Armonk, NY.
- Howell, D. C. and Huessy, H. R. (1985). A fifteen year follow-up of a behavioral history of attention deficit disorder (ADD). *Pediatrics*, 76, 185–190.
- Huba, G. J., Wingard, J. A. and Bentler, P. M. (1981). A comparison of two latent variable causal models for adolescent drug use. *Journal of Personality and Social Psychology*, 40, 180–193.
- Huck, S.W. and Sandler, H. M (1979) *Rival Hypotheses*, Harper and Row, New York.
- Hutcheson, G.D., Baxter, J. S., Telfer, K., and Warden, D. (1995) Child witness statement quality: question type and errors of omission, *Law and Human Behavior*, 19, 631–648.
- Jacobson, G. C. and Dimock, M. (1994) Checking Out: The effects of bank overdrafts on the 1992 House election, *American Journal of Political Science*, 38, 601–624.
- Jennrich, R. J. and Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika*, 31, 313–323.
- Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer-Verlag, New York.
- Jolliffe, I. T. (1970). Redundant variables in multivariate analysis. D.Phil. thesis, University of Sussex.
- Jolliffe, I. T. (1972). Discarding variables in a principal components analysis. I: Artificial data. *Applied Statistics*, 21, 160–173.
- Jolliffe, I. T. (1973). Discarding variables in a principal components analysis. II: Real data. *Applied Statistics*, 22, 21–31.
- Jolliffe, I. T. (1989). Rotation of ill-defined components. *Applied Statistics*, 38, 139–148.

- Jolliffe, I. T. (2002). Principal Components Analysis. Springer, New York.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.
- Kalbfleisch, J.D. and Prentice, R.L. (1980) The Statistical Analysis of Failure Time Data, Wiley, New York.
- Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York.
- Keele, L. (2008). Semiparametric Regression for the Social Sciences. John Wiley & Sons, Chichester, U.K.
- Keyfitz, N. and Flieger, W., (1971). Population: The Facts and Methods of Demography. W. H. Freeman, San Francisco, CA.
- Kinsey, A. C., Wardell, B. P. and Martin, C. E. (1948). Sexual Behavior in the Human Male. W. B. Saunders, Philadelphia, PA.
- Kinsey, A. C., Wardell, B. P., Martin, C. E. and Gebhard, P. H. (1953). Sexual Behavior in the Human Female. W. B. Saunders, Philadelphia, PA.
- Kleinbaum, D.G., Kupper, L.L., and Muller, K.E. (1988) Applied Regression Analysis and Other Multivariate Methods, 2nd edition, PWS-Kent Publishing, Boston.
- Kline, R. B. (2004). Principles and Practice of Structural Equation Modeling, 2nd edition. Guilford Press, New York.
- Krzanowski, W. J. (1988). Principles of Multivariate Analysis. Oxford University Press, Oxford.
- Labonitz, S. (1970) The assignments of numbers to rank order categories, *American Sociological Review*, 35, 515–524.
- Lackey, N. R. and Sullivan, J. (2003). Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research. Sage Publications, London.
- Lawley, D. N. and Maxwell, A. E. (1971). Factor Analysis as a Statistical Method, 2nd edition. Butterworths, London.
- Lee, Y. J. (1983). Quick and simple approximations of sample sizes for comparing two independent binomial distributions. *Biometrics*, 40, 239–242.
- Lehman, D., Wortman, C. and Williams, A. (1987). Long term effects of losing a spouse or a child in a motor vehicle crash. *Journal of Personality and Social Psychology*, 52, 218–231.
- Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measure studies. *Journal of the American Statistical Association*, 90, 1112–1121.
- Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. John Wiley & Sons, New York.
- Longford, N. T. (1993). Random Coefficient Models. Oxford University Press, Oxford.
- McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd edition. Chapman and Hall, London.
- McHugh, R. B. and Lee, C. T. (1984). Confidence estimation and the size of a clinical trial. *Controlled Clinical Trials*, 5, 157–164.
- McKay, R. J. and Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis. I: Description. *British Journal of Mathematical and Statistical Psychology*, 35, 1–29.
- McKay, R. J. and Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis. II: Allocation. *British Journal of Mathematical and Statistical Psychology*, 35, 30–41.

- MacDonnell, W.R. (1902) On criminal anthropometry and the identification of criminals, *Biometrika*, 1, 177–227.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- Marriott, F. H. C. (1974). *The Interpretation of Multiple Observations*. Academic Press, London.
- Marriott, F. H. C. (1982). Optimization methods of cluster analysis. *Biometrika*, 69, 417–421.
- Matthews, D. E. (2005). Multiple linear regression. In *Encyclopedia of Biostatistics*, 2nd edition, Ed. P. Armitage and T. Colton. John Wiley & Sons, Chichester, U.K.
- Maxwell, S. E. and Delaney, H. D. (2003). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 2nd edition. Lawrence Erlbaum, Mahwah, NJ.
- Miles, J. and Shevlin, M (2001). *Applying Regression and Correlation*. Sage, London.
- Morrison, D. F. (1990). *Multivariate Statistical Methods*, 3rd edition. McGraw-Hill, New York.
- Murray, G. and Findlay, J. (1988) Correcting for bias caused by dropouts in hypertension trials, *Statistics in Medicine*, 7, 941–946.
- Needham, R. M. (1967). Automatic classification in linguistics. *The Statistician*, 17, 45–54.
- Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society, Series A*, 140, 48–63.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 155, 370–384.
- Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. John Wiley & Sons, Chichester, U.K.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 559–572.
- Pocock, S. J. (1996). Clinical trials: A statistician's perspective, in *Advances in Biometry*, Ed. P. Armitage and H. A. David. John Wiley & Sons, Chichester, U.K.
- Proudfoot, J., Ryden, C., and Everitt, B.S (2004) Clinical efficacy of computerized-cognitive behavioural therapy for anxiety and depression; Randomized controlled trial, *British Journal of Psychiatry*, 185, 146–154.
- Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*, 2nd edition. STATA Press, College Station, TX.
- Rawlings, J. O., Sastri, G. P. and Dickey, D. A. (2001). *Applied Regression Analysis: A Research Tool*. Springer, New York.
- Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, New York.
- Rosenbaum, P. R. (2002). *Observational Studies*, 2nd edition. Springer, New York.
- Rosenbaum, P. R. (2005) Observational study, in *Encyclopedia of Statistics in Behavioral Science*, Ed. B. S. Everitt and D. C. Howell. John Wiley & Sons, Chichester, U.K.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.
- Sarkar, D. (2008). *Lattice: Multivariate Visualization with R*. Springer, New York.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3–15.
- Schmid, C. F. (1954). *Handbook of Graphic Presentation*. Ronald Press, New York.
- Schmidt, U., Evans, K., Tiller, J. and Treasure, J. (1995). Puberty, sexual milestones abuse: How are they related in eating disorder patients? *Psychological Medicine*, 25, 413–417.

- Schoenfield, D. A. (1983). Sample size formulae for the proportional hazards regression model. *Biometrika*, 70, 499–503.
- Schuman, H. and Kalton, G. (1985). Survey methods. In *Handbook of Social Psychology*, Vol. 1, Ed. G. Lindzey and E. Aronson. Addison-Wesley, Reading, MA, p. 635.
- Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 387–398.
- Senn, S. J. (1997). *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, U.K.
- Sieh, F. Y. (1987). A simple method for sample-size calculations in equal sample-size designs that use the logrank or t-test. *Statistics in Medicine*, 6, 577–582.
- Simon, R. (1991) A decade of progress in statistical methodology for clinical trials, *Statistics in Medicine*, 10, 1789–1817.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Spicer, C. C., Laurence, G. J. and Southall, D. P. (1987). Statistical analysis of heart rates and subsequent victims of sudden infant death syndrome. *Statistics in Medicine*, 6, 159–166.
- Sudman, S. and Bradburn, N. (1982). *Asking Questions*. Jossey-Bass, San Francisco, CA.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.
- Thomsen, O. Ö., Wulff, H.R., Martin, A., and Springer, P.A. (1993) What do gastroenterologists in Europe tell cancer patients? *The Lancet*, 341, 473–476.
- Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, 39, 406–427.
- Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer, New York.
- Tourangeau, R., Rips, L. J. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, New York.
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.
- Velleman, P. F. and Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47, 65–72.
- Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *Applied Statistics*, 48, 269–312.
- Vetter, B. M. (1980). Working women scientists and engineers. *Science*, 207, 28–34.
- Wainer, H. (1997). *Visual Revelations*. Springer, New York.
- Watkins, E. and Williams, R.M. (1998) The efficacy of cognitive-behavioural therapy, In *The Management of Depression* (ed. S. Checkley), Blackwell Science, Oxford.
- Willerman, L., Schultz, R., Rutledge, J. N. and Bigler, E. (1991). In vivo brain size and intelligence. *Intelligence*, 15, 223–228.
- Wilkinson, L. (1992). Graphical displays. *Statistical Methods in Medical Research*, 1, 3–25.
- Wittes, J. and Wallenstein, S. (1987) The power of the Mantel-Haenszel tests, *Journal of the American Statistical Association*, 82, 1104–1109.
- Wright, M. A., Wintemute, G. J. and Rivara, F. P. (1999) Effectiveness of denial of handgun purchase to persons believed to be at high risk for firearm violence, *American Journal of Public Health*, 89, 88–90.

Appendix: Solutions to Selected Exercises

Chapter 1

- 1.1 One alternative explanation is the systematic bias that may be produced by always using the letter Q for Coke and the letter M for Pepsi. In fact, when the Coca-Cola Company conducted another study in which Coke was put into both glasses, one labeled M and the other Q, the results showed that a majority of people chose the glass labeled M in preference to the glass labeled Q.
- 1.3 Clearly, you cannot carry out an experiment in which you assign guns at random to one group of convicted felons and not to another group. So, you need an observational study, although such a study also faces substantial difficulties. It would not be reasonable to estimate the effects of such a law by comparing the rate of criminal violence among convicted felons barred from handgun purchases to the rate among all other individuals permitted to purchase handguns; convicted felons may be more prone to criminal violence and may have greater access to illegally purchased guns than typical purchasers of guns without felony convictions. For some ideas about how the study could be carried out, see Wright, Wintemute, and Rivara (1999). The answer given here is taken from Rosenbaum (2005).
- 1.4 Be suspicious—very, very suspicious!
- 1.5 (a) Florence Nightingale, (b) Lloyd George, (c) Joseph Stalin, (d) W. H. Auden, (e) Mr. Justice Streatfield, and (f) Logan Pearsall Smith.
- 1.6 You need to convert both temperatures to the Kelvin scale before taking the ratio. The Kelvin scale is simply $273 + \text{centigrade}$, and $\text{centigrade} = 5/9 (\text{Fahrenheit} - 32)$. So, 115°F is 46°C , and so, the required ratio is $(273 + 46)/(273 + 25) = 1.07$.

Chapter 2

2.2 R code to produce side-by-side boxplots and probability plots

```
#Exercise 2.2
#read in data
length_dat<-source("c:\\mvmvanswers\\exer_22.txt")$value
#
#print data
length_dat
#
#produce side-by-side boxplots and normal probability plots
# after converting guess in meters to feet
#
attach(lengths_dat)
feet <- group == "Feet"
convert<-ifelse(feet,1,3.28)
layout(matrix(c(1,2,1,3),nrow=2,ncol=2,byrow=FALSE))
boxplot(I(guesses*convert)~group,ylab= "Width guess(feet)",
var.width=TRUE,names=c("Width guesses in metres (after
conversion to
feet)",
"Width guesses in feet"))
qqnorm(guesses[!feet],ylab= "Width guesses in metres")
qqline(guesses[!feet])
qqnorm(guesses[feet],ylab="Width guesses in feet")
qqline(guesses[feet])
```

2.3 The graph commits the cardinal sin of quoting data out of context; remember that graphics often lie by omission, leaving out data sufficient for comparisons. Here, a few more data points for other years in the area would be helpful, as would similar data for other areas in which stricter enforcement of speeding had not been implemented.

2.5 R code for boxplots

```
#read in data
suicide_dat<-source("c:\\mvmvanswers\\exer_25.txt")$value
#get boxplots
boxplot(suicide_dat[,1],suicide_dat[,2],suicide_
dat[,3],suicide_dat[,4],
suicide_dat[,5],names=c("25-34","35-44","45-54","55-64","65-74"),
ylab="Male suicide rates per 100,000")
#
```

A set of boxplots for the different countries might also be interesting.

Chapter 3

3.2 R code for plots and fitting regression

```
#ex3.2
#read in data
exam_dat<-source("c:\\mvmvanswers\\exer_32.txt")$value
exam_dat
#
attach(exam_dat)
#plot data
#
layout(matrix(c(2,0,1,3),2,2,byrow=TRUE),c(2,1),c(1,2),TRUE)
plot(marks,times)
abline(lm(times~marks))
hist(times)
boxplot(marks)
#
exam_reg<-lm(times~marks)
pred<-predict(exam_reg)
resd<-residuals(exam_reg)
par(mfrow=c(1,2))
plot(pred,resd,xlab="Fitted value",ylab="Residual")
plot(marks,resd,ylab="Residual")
```

The residual plots show some large positive residuals. A probability plot of residuals may be helpful, and then, perhaps a log transform of the response might be worth investigating.

3.4 R code for plot and fitting regression

```
#ex3.4
mardiv_rates<-source("c:\\mvmvanswers\\exer_34.txt")$value
#
attach(mardiv_rates)
mardiv_reg<-lm(divrate~marrate)
summary(mardiv_reg)
plot(divrate~marrate,xlab="Marriage rate",ylab="Divorce rate")
abline(mardiv_reg)
#
divpred8<-0.6646+0.4808*8
divpred14<-0.6646+0.4808*14
```

The prediction for a marriage rate of 14 is extrapolating outside the observed range of marriage rates—a procedure fraught with danger! Find the standard errors of both predictions.

Chapter 4

4.2 R code for regression, etc; you need to add some graphics.

```
#ex4.2
quality_dat<-source("c:\\mvmvanswers\\exer_42.txt")$value
#
attach(quality_dat)
cor(coherence,maturity)
#
quality_reg<-lm(qualityct~age+location+maturity+delay+prosecute)
#
#show structure of dummy variables
contrasts(age)
contrasts(sex)
contrasts(location)
contrasts(prosecute)
#
summary(quality_reg)
#
step(quality_reg,method="backwards")
#
quality1_reg<-lm(qualityct~age+delay+prosecute)
#
pred<-predict(quality1_reg)
resd<-residuals(quality1_reg)
qqnorm(resd)
plot(pred,resd)
#etc,etc
```

Maturity and coherence are highly correlated, so coherence is dropped from regression. Backwards search also drops maturity; residual plots look okay. You need to interpret the estimated regression coefficients.

4.4 R code for fitting regression and plotting

```
#ex4.4
fat_dat<-source("c:\\mvmvanswers\\exer_44.txt")$value
attach(fat_dat)
fat_reg<-lm(Pcfat~Age)
#
plot(Age,Pcfat,xlab="Age",ylab="%fat",type="n")
text(Age,Pcfat,labels=Sex)
abline(fat_reg)
#
```

```

summary(lm(Pcfat~Age+Sex))
#
plot(Age,Pcfat,xlab="Age",ylab="%fat",type="n")
text(Age,Pcfat,labels=Sex)
#use figures from summary to find slope and intercepts of
lines for
#men and women
abline(a=5.93,b=0.29)
abline(a=17.49,b=0.29,lty=2)
legend("topleft",c("Females","Males"),lty=1:2)
#
summary(lm(Pcfat~Age*Sex))
plot(Age,Pcfat,xlab="Age",ylab="%fat",type="n")
text(Age,Pcfat,labels=Sex)
abline(a=3.47,b=0.35)
abline(a=20.01,b=0.24,lty=2)
legend("topleft",c("Females","Males"),lty=1:2)

```

Chapter 5

5.1 R code for analysis and to plot the residuals

```

#ex5.1
bloodpress_dat<-source("c:\\mvmvanswers\\exer_51.txt")$value
#
attach(bloodpress_dat)
bloodpress_reg<-lm(Bloodp~History*Smoking)
summary(bloodpress_reg)
qqnorm(residuals(bloodpress_reg))

```

5.2 R code for regression and finding confidence interval (CI)

```

#ex5.2
oestrogen_dat<-source("c:\\mvmvanswers\\ex_52.txt")$value
#
attach(oestrogen_dat)
oestrogen_reg<-lm(Depression~Treatment+BL1+BL2)
summary(oestrogen_reg)
#
treatci<-c(-2.477-2*1.707,-2.477+2*1.707)

```

CI contains the value 0, so there is no evidence of a treatment effect.

Chapter 6

6.1 R code for fitting logistic and plotting various things

```
#ex6.1
womensrole_dat<-source("c:\\\\mvmvanswers\\\\ex_61.txt")$value
#
attach(womensrole_dat)
#main effects model
womensrole_glm<-glm(cbind(agree,disagree)~sex+education,
family=binomial)
summary(womensrole_glm)
#interaction model
womensrole_glm1<-glm(cbind(agree,disagree)~sex*education,
family=binomial)
summary(womensrole_glm1)
#interaction significant
#
fitted<-predict(womensrole_glm1,type="response")
fittedF<-fitted[sex=="Female"]
fittedM<-fitted[sex!="Female"]
pobsv<-agree/(agree+disagree)
plot(education,pobsv,type="n",xlab="Education",
ylab="Probability of agreeing")
text(education,pobsv,ifelse(sex=="Female","\\VE",
"\\MA"),vfont=c("serif","plain"),
cex=1.25)
lines(education[sex=="Female"],fittedF)
lines(education[sex!="Female"],fittedM,lty=2)
legend("topright",c("Fitted(Female)","Fitted(Male)"),lty=1:2)
```

The interaction shows that, for fewer years of education, women have a higher probability of agreeing with the statement than men, but when the years of education exceed about 10, then this situation reverses.

6.5 R code for fitting model and plotting observed and predicted probabilities

```
#ex6.5
menstruation_dat<-source("c:\\\\mvmvanswers\\\\exer_65.txt")$value
attach(menstruation_dat)
menstruation_reg<-glm(cbind(bmens,n-
bmens)~age,family=binomial)
summary(menstruation_reg)
#
```

```

plot(age,bmens/n,xlab="Age (years)",ylab="Probability of
menstruating")
abline(lm(bmens/n~age),lty=2)
lines(age,predict(menstruation_reg,type="response"))
legend("topleft",c("Logistic","Linear"),lty=1:2)

```

This shows why the linear model is useless and the logistic model is not.

Chapter 7

7.2 R code to plot survival curves and perform log-rank test

```

#ex7.2
breastcan_dat<-source("c:\\\\mvmvanswers\\\\exer_72.txt")$value
attach(breastcan_dat)
#
library("survival")
#plot of survival function
plot(survfit(Surv(time,event)~metastized),
xlab="Time",
ylab="Proportion",lty=1:2,legend.text=c("No","Yes"))
#
#logrank test
library("coin")
survdiff(Surv(time,event)~metastized)

```

7.3 R code for fitting Cox model

```

#ex7.3
glioma_dat<-source("c:\\\\mvmvanswers\\\\exer_73.txt")$value
#
library("survival")
attach(glioma_dat)
glioma_cox<-coxph(Surv(time,event)~age+sex+histology+group)
#required CI given in summary-hazard on RIT is between about
8% and 58%
#of hazard on standard therapy
summary(glioma_cox)

```

You should plot the survival curves of radioimmunotherapy (RIT) and standard treatments.

Chapter 8

8.4 R code to plot data and fit models that include a quadratic effect for time; this can be seen in the plots of the data.

```
#ex8.4
phosphate_dat<-source("c:\\mvmvanswers\\exer_84.txt")$value
#
Group<-phosphate_dat[,1]
#plot individual profiles separately for the two groups
par(mfrow=c(1,2))
matplot(c(0,0.5,1,1.5,2,3,4,5),t(phosphate_dat[Group=="control",
2:9]),type="l",lty=1,
axes=F,xlab="Time after glucose challenge (hours)",
ylab="Plasma inorganic phosphate level",ylim=c(0,7))
axis(1,at=c(0,0.5,1,1.5,2,3,4,5),labels=c("Pre","30 mins",
"1 hour",
"1.5 hours","2 hours","3 hours","4 hours","5 hours"))
axis(2)
title("Controls")
matplot(c(0,0.5,1,1.5,2,3,4,5),t(phosphate_dat[Group=="obese",
2:9]),type="l",lty=1,
axes=F,xlab="Time after glucose challenge (hours)",
ylab="Plasma inorganic phosphate level",ylim=c(0,7))
axis(1,at=c(0,0.5,1,1.5,2,3,4,5),labels=c("Pre","30 mins",
"1 hour",
"1.5 hours","2 hours","3 hours","4 hours","5 hours"))
axis(2)
title("Obese")
#
library("lme4")
#
#put data into long form for analysis
#
group<-rep(c(0,1),c(104,160))
#
time<-c(0.0,0.5,1.0,1.5,2.0,3.0,4.0,5.0)
time<-rep(time,33)
#
subject<-rep(1:33,rep(8,33))
phosphatel_dat<-cbind(subject,time,group,as.
vector(t(phosphate_dat[,2:9])))
dimnames(phosphatel_dat)<-list(NULL,c("Subject","Time","Group",
"Plasma"))
#
phosphatel_dat<-data.frame(phosphatel_dat)
phosphatel_dat$Group<-factor(phosphatel_dat$Group,
levels=c(0,1),labels=c("Control","Obese"))
```

```
#  
attach(phosphate_l_dat)  
#  
#fit independence model allowing a quadratic effect for time  
summary(lm(Plasma~Time+I(Time*Time)+Group))  
#random intercept model  
phosphate_lme1<-lmer(Plasma~Time+I(Time*Time)+Group+(1|  
Subject))  
summary(phosphate_lme1)  
#random intercept and slope model  
phosphate_lme2<-lmer(Plasma~Time+Group+I(Time*Time)+(Time|  
Subject))  
summary(phosphate_lme2)  
anova(phosphate_lme1,phosphate_lme2)  
#
```

Chapter 9

9.3 R code for summary statistics and some plots

```
#ex9.3  
lifeex_dat<-source("c:\\mvmvanswers\\exer_93.txt")$value  
#  
get summary statistics for men and for women  
Rmen<-cor(lifeex_dat[,1:4])  
Smen<-var(lifeex_dat[,1:4])  
mean_men<-apply(lifeex_dat[,1:4],2,mean)  
#  
Rwomen<-cor(lifeex_dat[,5:8])  
Swomen<-var(lifeex_dat[,5:8])  
mean_women<-apply(lifeex_dat[,5:8],2,mean)  
#  
pairs(lifeex_dat[,1:4])  
pairs(lifeex_dat[,5:8])  
#  
pairs(lifeex_dat[,1:4],panel=function(x,y)  
text(x,y,abbreviate(row.names(lifeex_dat)),cex=0.5))  
#  
pairs(lifeex_dat[,5:8],panel=function(x,y)  
text(x,y,abbreviate(row.names(lifeex_dat)),cex=0.5))  
#  
attach(lifeex_dat)  
par(mfrow=c(1,2))  
plot(m0,m75,xlab="Life expectation at birth",  
ylab="Life expectation at age 75",type="n",ylim=c(5,16),xlim=c  
(30,80))
```

```

text(m0,m75,abbreviate(row.names(lifeex_dat)),cex=0.6)
title("Men")
#
plot(w0,w75,xlab="Life expectation at birth",
ylab="Life expectation at age 75",type="n",ylim=c(5,16),xlim=c
(30,80))
text(m0,m75,abbreviate(row.names(lifeex_dat)),cex=0.6)
title("Women")
#

```

Chapter 10

10.2 R code for principal components and matrix multiplications

```

#ex10.2
#
R<-matrix (c(1,0.6579,0.0034,0.6579,1.0,-0.0738,0.0034,-
0.0738,1.0),
ncol=3,byrow=T)
#
R_pc<-princomp(covmat=R)
#get component variances
varpc<-R_pc$sd^2
R_pc$loadings[,1] %*% diag(varpc[1]) %*% t(R_pc$loadings[,1])
R_pc$loadings[,1:2] %*% diag(varpc[1:2]) %*% t(R_
pc$loadings[,1:2])
#just to show the three components reproduce R
R_pc$loadings[,1:3] %*% diag(varpc[1:3]) %*% t(R_
pc$loadings[,1:3])

```

10.4 R code for principal components of correlation matrix and plot

```

#ex10.4
prestige_dat<-source("c:\\\\mvmvanswers\\\\exer_104.txt")$value
#
prestige_pc<-princomp(prestige_dat,cor=T)
summary(prestige_pc,loadings=T)
#
#use first two component scores to plot the data
xlim<-range(prestige_pc$scores[,1])
plot(prestige_pc$scores[,1:2],ylim=xlim,type="n")
text(prestige_pc$scores[,1:2],abbreviate(row.names(prestige_
dat)),cex=0.5)
#

```

```
options(digits=3)
cbind(prestige_pc$scores[,1:2])
#the above is helpful in interpreting the plot
```

Division into professional and nonprofessional perhaps?

Chapter 11

11.1 Now, the equations given in the text become

$$\hat{\lambda}_1 \hat{\lambda}_2 = 0.84$$

$$\hat{\lambda}_1 \hat{\lambda}_3 = 0.60$$

$$\hat{\lambda}_2 \hat{\lambda}_3 = 0.35$$

$$\hat{\psi}_1 = 1.0 - \hat{\lambda}_1^2$$

$$\hat{\psi}_2 = 1.0 - \hat{\lambda}_2^2$$

$$\hat{\psi}_3 = 1.0 - \hat{\lambda}_3^2$$

In this case, the solution for the parameters of a single-factor model is

$$\hat{\lambda}_1 = 1.2, \hat{\lambda}_2 = 0.7, \hat{\lambda}_3 = 0.5$$

$$\hat{\psi}_1 = -0.44, \hat{\psi}_2 = 0.51, \hat{\psi}_3 = 0.75$$

Clearly, this solution is unacceptable because of the negative estimate for the first specific variance.

11.5 R code to fit two-factor model to correlations between first eight questions about pain in Chapter 11, Exercise 11.4.

```
#ex11.5
#use correlation matrix from ex11.4
#remove Q9
R<-R[-9,-9]
model.test<-specify.model()
  Doctor -> pfuture,lambda1,NA
  Doctor -> pdoc,lambda2,NA
  Doctor -> pseek,lambda3,NA
```

```

Doctor -> pcontdoc,lambda4,NA
Patient -> pme,lambda5,NA
Patient -> pexerf,lambda6,NA
Patient -> pcareless,lambda7, NA
Patient -> pmyresp,lambda8, NA
Doctor <-> Patient, rho,NA
pfuture <-> pfuture,theta1,NA
pme <-> pme,theta2,NA
pdoc <-> pdoc,theta3,NA
pseek <-> pseek,theta4,NA
pexerf <-> pexerf,theta5,NA
pcareless <-> pcareless,theta6,NA
pmyresp <-> pmyresp,theta7,NA
pcontdoc <-> pcontdoc, theta8,NA
Doctor <-> Doctor,NA,1
Patient <-> Patient,NA,1
model.test
sem.test<-sem(model.test,R,123)
summary(sem.test)
#

```

Chapter 12

12.3 R code for plots and k -means; try other methods

```

#ex12.3
lowtemp_dat<-source("c:\\mvmvanswers\\exer_123.txt")$value
#
pairs(lowtemp_dat)
lowtemp_pc<-princomp(lowtemp_dat)
xlim<-range(lowtemp_pc$scores[,1])
plot(lowtemp_pc$scores[,1:2],ylim=xlim)
#possible 2 or three clusters?
#
lowtemp_km2<-kmeans(lowtemp_dat,2)
lowtemp_km2
lowtemp_km3<-kmeans(lowtemp_dat,3)
lowtemp_km3
#
par(mfrow=c(1,2))
plot(lowtemp_pc$scores[,1:2],ylim=xlim,type="n")
text(lowtemp_pc$scores[,1:2],labels=as.numeric(lowtemp_
km2$cluster),cex=0.6)
plot(lowtemp_pc$scores[,1:2],ylim=xlim,type="n")
text(lowtemp_pc$scores[,1:2],labels=as.numeric(lowtemp_
km3$cluster),cex=0.6)

```

12.4 R code to read in data and do principal components analysis (PCA)—left for you to do CA.

```
#ex12.4
protein_dat<-source("c:\\\\mvmvanswers\\\\exer_124.txt")$value
#
protein_pc<-princomp(protein_dat,cor=T)
xlim<-range(protein_pc$scores[,1])
plot(protein_pc$scores[,1:2],ylim=xlim)
#possible 2 clusters?
#try some agglomerative methdos and plot solutions in PC space
```

Chapter 13

13.4 R code for MANOVA and discriminant functions

```
#ex13.4
pottery_dat<-source("c:\\\\mvmvanswers\\\\exer_134.txt")$value
pottery_manova<-manova(pottery_dat[,2:9]~pottery_dat[,1])
summary(pottery_manova,test="Pillai")
summary(pottery_manova,test="Wilks")
summary(pottery_manova,test="Hotelling")
summary(pottery_manova,test="Roy")
#all highly significant
#
region<-rep(1,length(pottery_dat[,1]))
region[pottery_dat[,1]==2|pottery_dat[,1]==3]<-2
region[pottery_dat[,1]==4|pottery_dat[,1]==5]<-3
#
m1<-apply(pottery_dat[region==1,-1],2,mean)
m2<-apply(pottery_dat[region==2,-1],2,mean)
m3<-apply(pottery_dat[region==3,-1],2,mean)
#
#find numbers in each class
n1<-length(pottery_dat[region==1,1])
n2<-length(pottery_dat[region==2,1])
n3<-length(pottery_dat[region==3,1])
#
#find pooled covariance matrix
S123<-((n1-1)*var(pottery_dat[region==1,-1])+(n2-1)
*var(pottery_dat[region==2,-1])+
(n3-1)*var(pottery_dat[region==3,-1]))/(n1+n2+n3-3)
#
#find coefficients for each classification class
invS<-solve(S123)
```

```
a1<-invS%*%(m1-m2)
a2<-invS%*%(m1-m3)
a3<-invS%*%(m2-m3)
#
#find thresholds
z12<- (m1%*%a1+m2%*%a1)/2
z13<- (m1%*%a2+m3%*%a2)/2
z23<- (m2%*%a3+m3%*%a3)/2
#
newvalues<-c(15.5,5.71,2.07,0.98,0.65,3.01,0.76,0.09,0.012)
(newvalues-z12)%*%a1
(newvalues-z13)%*%a2
(newvalues-z23)%*%a3
#allocate to region 2
```

Index

A

- Agglomerative hierarchical clustering, 241–249
Agresti, A., 93–94
Akaike's information criterion (AIC), 94–95, 126–128
All subsets regression, 92
Altman, D. G., 18
Analysis of variance (ANOVA)
 equivalence of multiple regression and, 103–110
 origin of, 103

B

- Banfield, J. D., 255
Bar charts, 23, 24, 27–29
Bartholomew, D. J., 227
Beating the Blues (BtB) program, 158–162
Beck Depression Inventory II, 158
Bertin, J., 24
Between-subject variation, 145
BIC criterion, 257–258
Bivariate boxplots, 40–42, 43, 46
Bivariate normal density function, 175–176
Blood chemistry data, 188–190, 195
Body measurement data, 173–174, 176–178, 243–245, 246
Box, George, 13
Boxplots, 31–34, 173
 bivariate, 40–42, 43, 46
Bubbleplots, 38–40

C

- Calculation of principal component scores, 195–196, 198
Case-control investigations, 5–6
Caslyn, J. R., 230
Categorical measurements, 7, 22–30
Cattell, R. B., 194
Chambers, J. M., 21
Chi-square plots, 177–178
Classifications, 239–240
 agglomerative hierarchical clustering, 241–249
 functions, 273–276

- maximum likelihood clustering, 254–255
Cleveland, W. S., 24, 26, 31
Cluster analysis
 agglomerative hierarchical clustering, 241–249
 body measurement data, 243–245, 246
 classifications and, 239–240
 defined, 241
 dendograms in, 243–245, 246
 k-means clustering, 250–253
 life expectancy data, 246–248, 249
 model-based clustering, 253–258
Cochran, W. G., 5
Cognitive behavioral therapy (CBT), 157–162
Companion Encyclopedia of Psychology, 3
Compound symmetry, 148
Conditioning plots, 45–48
Confidence intervals, 16–18, 116–117, 232
Confirmatory factor analysis, 229–235
Correlation
 coefficient and bivariate data, 190–192
 intraclass, 148
 matrices, 172–173, 188–190, 201–202, 204, 205–208, 215, 219–220
 observed, 231–232
Covariance
 correlation matrices and, 188–190, 193, 201–202, 204, 205–208
 matrices and discriminant analysis, 275–276
 matrix S^* , 216–217
 population, 171–172
Cox, D., 132, 138
Cox regression, 132, 138–143
Crime rate data, 200–203, 204, 226–227
Cubic spline, 75

D

- Data. *See also* Graphical displays
 BIC criterion, 257–258
 bivariate, 190–192
 blood chemistry, 188–190, 195
 body measurement, 173–174, 176–178, 243–245, 246
 categorical, 22–30
 classification of, 239–240, 241–249
 crime rate, 200–203, 204, 226–227

- drug usage by American college students, 205–208, 233–235
 head size of brothers, 196–200
 interval/quasi-interval, 30–35
 k -means clustering of, 250–253
 life expectancy, 246–248, 249
 longitudinal, 146–150
 measurement types, 7–10
 missing values, 10–11
 multivariate, 169–180, 250–253
 outliers, 32
 repeated-measures, 147
 role of models in analysis of, 11–13
 sample size determination, 14–16
 subsets, 208–209
 sudden infant death (SID), 265–270
- Deception, graphical, 52–57
 Degrees of freedom (DF), 63
 Dendograms, 243–245, 246
 Density function, normal, 174–178, 179–180
 Diagnostics, regression, 68–71, 72, 96–97, 98–99
 Diggle, P. J., 162, 164
 Dimmock, M., 76–78
 Discriminant analysis, 274–276
 Dot plots, 24, 25
 Dropout at random (DAR), 163
 Dropout completely at random (DCAR), 162–163
 Dropouts in longitudinal studies, 162–164, 165
 Drug usage data, 205–208, 233–235
 Dunn, G., 269–270
- E**
- Equality of means of two variables, 264–265
 Estimation
 factor scores, 227, 228
 least-squares, 82–83
 number of factors, 217–218
 Everitt, B. S., 269–270
 Evolutionary trees, 243, 245
 Experiments, 4–5
 quasi-, 6
 Explanatory and response variables, 9–10
 logistic regression and, 118–119
 multicollinearity and, 90–91
 random intercept model and, 155–157
 Exploratory factor analysis, 228–229
- F**
- Factor analysis
 confirmatory, 229–235
- drug usage, 233–235
 estimating factor scores in, 227, 228
 estimating number of factors in, 217–218
 exploratory, 228–229
 k -, 214, 216–217
 lack of uniqueness in factor loadings, 221–227
 model fitting, 218–220
 parameters, 214, 215–217, 219–220
 principal, 216–217
 rotation of factors in, 220–227
 types of, 211–212
 variables in, 212–215
- Fisher, R. A., 4, 5, 103
 Fisher's linear discriminant function, 265–270
- Fleiss, J. L., 7
 Flieger, W., 246
 Formulation, model, 13
 Friedman, H. P., 270
 F-statistics, 83, 263
 F-test, 84
- G**
- Gardner, M. J., 18
 Generalized linear model (GLM), 110–112
 Graphical deception, 52–57
 Graphical displays. *See also Statistics*
 advantages of, 21–22
 bar charts, 23, 24, 27–29
 bivariate boxplot, 40–42, 43, 46
 boxplot, 31–34, 40–42, 43, 173
 bubbleplot, 38–40
 categorical data, 22–30
 conditioning plot, 45–48
 deception, 52–57
 dot plots, 24, 25
 histograms, 31
 interval/quasi-interval data in, 30–35
 multivariate data, 173–174
 normal probability plots, 174–178, 179–180
 pie charts, 22–24
 probability plot, 34–35
 scatterplot, 35–43, 68–71, 72, 72–79, 89–91, 174, 200–201
 scatterplot matrices, 44–45
 simple, 22–35
 trellis graphic, 48–52
- Graphical rotation, 222–223
 Grouped multivariate data
 Fisher's linear discriminant function, 265–270

- Hotelling's T^2 test, 262–263
tests for equality of means of, 264–265
two-group, 262–270
- H**
- Hazard function, 136–138
Cox regression, 138–143
- Histograms, 31
- Hotelling's T^2 test, 262–263
- Howell, D. C., 89
- Huba, G. J., 205, 233, 235
- I**
- Imputation, multiple, 11
- Independence model, 151–152
- Intercluster distance measures, 243, 244
- Interval/quasi-interval data, 30–35
- Interval scales, 8
- Intraclass correlation, 148
- J**
- Jacobson, G. C., 76–78, 77
- Jolliffe, I. T., 188–189, 194, 195, 208–209
- Journal of the American Statistical Association*, 1
- K**
- Kaplan-Meier estimated survival function, 133–136
- Keele, L., 76, 78–79
- Kenny, D. A., 230
- Kenward, M. G., 164
- Keyfitz, N., 246
- K-factor analysis model, 214, 216–217
- Kinsey, Alfred Charles, 3
- K-means clustering, 250–253
- L**
- Lagrange multiplier, 187
- Latent variables, 230–232
- Least-squares estimation process, 82–83
- Life expectancy data, 246–248, 249
- Likelihood ratio test, 156–157, 161–162
- Linear combinations, 185–187
- Linear discriminant function, 265–270
- Linear mixed effects models
 applied to rat growth data, 150–157
- cognitive behavioral therapy delivery, 157–162
- dropouts problems in, 162–164, 165
- fitting independence model to data, 151–152
- fitting to data, 153–157
- for longitudinal data, 146–150, 162–164, 165
- Linear regression, 35–38, 61–62. *See also* Regression
 Multiple linear regression
 generalized, 110–112
 simple, 62–64
- Little, R. J. A., 162
- Loadings, factor, 221–227
- Locally weighted regression, 72–79
- Logistic regression. *See also* Regression
 Akaike's information criterion (AIC) in, 126–127
 and linear regression fit to data, 120–124
 model, 117–119
 odds and odds ratios and, 115–117, 121, 127–128
 selecting the most parsimonious model in, 124–128
- Longitudinal data and linear mixed effects models, 146–150, 162–164, 165
- Lowess fit, 73–74
- M**
- Marriott, F. H. C., 203
- Matrices, scatterplot, 44–45
- Maximum likelihood (ML), 149–150, 223
 clustering classification, 254–255
- Measurements
 interval scale, 8
 missing data and, 10
 nominal or categorical, 7
 ordinal scale, 8
 ratio scale, 9
 response and explanatory variables, 9–10
- Missing values, 10–11
- Model-based clustering, 253–258
- Models, data analysis, 11–13
- Multicollinearity, 90–91
- Multiple imputation, 11
- Multiple linear regression. *See also* Linear regression
 choosing the most parsimonious model when applying, 89–96
 equivalence of ANOVA and, 103–110
 estimated regression coefficients, 85–86
 explanatory variables and, 84–89
 fitted model, 86–89

- model details, 81–83
 multicollinearity and, 90–91
 regression diagnostics, 96–97, 98–99
 scatterplots, 89–91
- Multivariate analysis of variance (MANOVA), 261, 270–273
- Multivariate data
 discriminant analysis, 274–276
 graphical descriptions of, 173–174
 initial analysis of, 170–174
 k -means clustering of, 250–253
 normal probability density function, 174–178, 179–180
 sets, 169
 summary statistics for, 170–173
 tests for equality of means of two variables, 264–265
- N**
- Needham, R. M., 240
- Nominal measurements, 7
- Nonignorable dropouts, 163–164
- Normal probability density function, 174–178, 179–180
- Numerical rotation methods, 223–226
- O**
- Oakes, M., 17, 18
- Oblimin rotation, 225
- Oblique rotation, 223–224
- Observational studies, 5–6
 interval/quasi-interval data in, 30–35
- Odds and odds ratios, 115–117, 121, 127–128
- One-way MANOVA, 270–273
- Ordinal scale measurements, 8
- Orthogonal rotation, 223–224
- Outliers, 32, 97
- P**
- Parameters, factor analysis, 214, 215–217, 219–220
- Partial likelihood, 140
- Percentages, 26–27
- Pie charts, 22–24
- Pinker, S., 239
- Pocock, S. J., 15
- Population covariance, 171–172
- Population mean vector, 171
- Principal components analysis (PCA)
 applications of, 183–185, 196–208
 bivariate data, 190–192
 calculating principal component scores in, 195–196, 198
 choosing number of components for, 193–195
 compared to exploratory factor analysis, 228–229
 covariance and correlation matrix in, 188–190, 201–202, 204, 205–208
 crime rates, 200–203, 204
 defined, 183
 drug usage by American college students, 205–208
 finding sample principal components for, 185–188
 head size of brothers, 196–200
 linear combinations in, 185–187
 predicting observed covariance matrix using, 193
 rescaling principal components and, 192
 selecting subset of variables using, 208–209
- Principal factor analysis, 216–217
- Probability plots, 34–35
- Promax rotation, 225
- Proportional hazards model, 139, 143
- Proudfoot, J., 158
- P-values, 16–18, 84
- Q**
- Quartimax rotation, 225
- Quasi-experiments, 6
- Quasi-interval data, 30–35
- R**
- Raftery, A. E., 255
- Random intercept model, 153–157, 161–162
- Randomization, 4–5
- Random slope model, 155–157
- Ratio(s)
 odds, 115–117, 121, 127–128
 scales, 9
- Regression. *See also* Logistic regression; Multiple linear regression
 all subsets, 92
 coefficients, estimated, 85–86, 125–126
 Cox, 132, 138–143
 diagnostics, 68–71, 72, 96–97, 98–99
 generalized linear model, 110–112
 linear, 35–38, 61–67, 68
 locally weighted, 72–79

- lowess fit, 73–74
mean square (RGMS), 63
random intercept model, 153–157, 161–162
scatterplot smoothers, 73–79
spline smoothers, 74–79
sum of squares, 106–110
- Rencher, A. C., 194
- Repeated-measures data, 147
- Rescaling principal components, 192
- Residual mean square (RMS), 63
- Residual value plots, 68–71, 72
- Response and explanatory variables, 9–10
- Restricted maximum likelihood (REML), 149–150
- Rotation of factors in factor analysis, 220–227
- S**
- Sagan, C., 21
- Sample size determination, 14–16
- Sarkar, D., 29
- Scales
- interval, 8
 - ratio, 9
- Scatterplots, 35–38, 200–201
- bubbleplot, 38–40
 - factor analysis, 227, 228
 - locally weighted regression, 72–79
 - matrices, 44–45
 - multiple linear regression, 89–91
 - multivariate data, 174
 - residual value, 68–71, 72
 - simple linear regression, 65–67, 68
 - smoothers, 73–79
- Schmid, C. F., 21
- Scientific research
- experiments in, 4–5
 - observational studies in, 5–6
 - quasi-experiments in, 6
 - statistics in, 1–2
 - surveys in, 3
 - types of measurement in, 7–10
 - types of study in, 2–6
- Scores
- factor, 227, 228
 - principal component, 195–196, 198
- Senn, S. J., 15, 16
- Sexual Behavior in the Human Female*, 3
- Sexual Behavior in the Human Male*, 3
- Significance tests, 16–18
- Simon, R., 16
- Simple linear regression, 62–64
 - fitting to data, 64–65
 - kinesiology example, 65–67, 68
- Smoothers
- scatterplot, 73–79
 - spline, 74–79
- Spline smoothers, 74–79
- Statistics, 1–2. *See also* Graphical displays; Multiple linear regression; Regression
- confidence intervals in, 16–18, 116–117
 - models, 11–13
 - multivariate data summary, 170–173
 - p-values in, 16–18, 84
 - sample size determination, 14–16
 - significance tests in, 16–18
 - t-values in, 17, 85–86
- Stratified proportional hazards model, 140
- Study, types of, 2–6
- Subjects, experiment, 4
- Subsets of variables, 208–209
- Sudden infant death (SID) syndrome, 265–270
- Summary statistics for multivariate data, 170–173
- Sum of squares, 106–110
- Surveys, 3
- Survival analysis
- Cox regression in, 132, 138–143
 - data used in, 131–132
 - hazard function in, 136–138
 - Kaplan-Meier estimator in, 133–136
 - survival function in, 132–136
- Symmetry, compound, 148
- T**
- Treatment as Usual program, 158–162
- Trellis graphics, 48–52
- Tufte, E. R., 21, 24, 35
- T-values, 17, 85–86, 263
- U**
- Univariate tests for equality of means of two variables, 264–265
- V**
- Variables
- equality of means of two, 264–265
 - factor analysis model, 212–215
 - latent, 230–232

logistic regression and, 118–119
multicollinearity and, 90–91
random intercept model and,
155–157
response and explanatory, 9–10
selecting subset of, 208–209
Variation, within-subject and between
subject, 145

Varimax rotation, 225, 226–227
Vetter, B. M., 26

W

Wainer, H., 24, 27
Wilkinson, L., 31
Within-subject variation, 145