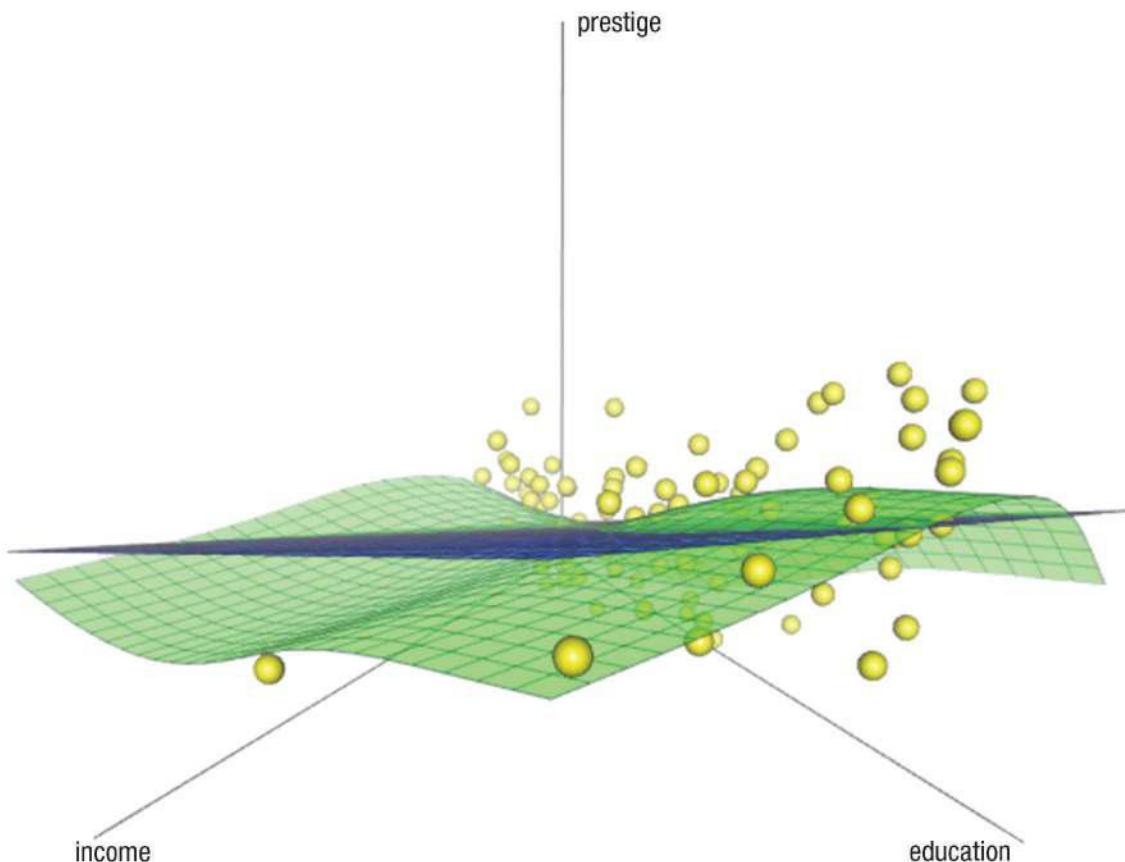


3^{EDITION}

APPLIED REGRESSION ANALYSIS *&* GENERALIZED LINEAR MODELS



John Fox



THIRD EDITION

APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS

For Bonnie and Jesse (yet again)

THIRD EDITION

APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS

John Fox

McMaster University



Los Angeles | London | New Delhi
Singapore | Washington DC | Boston



Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Copyright © 2016 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Cataloging-in-Publication Data is available for this title from the Library of Congress.

ISBN 978-1-4522-0566-3

Printed in the United States of America

Acquisitions Editor: Vicki Knight
Associate Digital Content Editor: Katie Bierach
Editorial Assistant: Yvonne McDuffee
Production Editor: Kelly DeRosa
Copy Editor: Gillian Dickens
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Jennifer Grubba
Cover Designer: Anupama Krishnan
Marketing Manager: Nicole Elliott

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

Brief Contents

Preface	xv
About the Author	xxiv
1. Statistical Models and Social Science	1
I. DATA CRAFT	12
2. What Is Regression Analysis?	13
3. Examining Data	28
4. Transforming Data	55
II. LINEAR MODELS AND LEAST SQUARES	81
5. Linear Least-Squares Regression	82
6. Statistical Inference for Regression	106
7. Dummy-Variable Regression	128
8. Analysis of Variance	153
9. Statistical Theory for Linear Models*	202
10. The Vector Geometry of Linear Models*	245
III. LINEAR-MODEL DIAGNOSTICS	265
11. Unusual and Influential Data	266
12. Diagnosing Non-Normality, Nonconstant Error Variance, and Nonlinearity	296
13. Collinearity and Its Purported Remedies	341
IV. GENERALIZED LINEAR MODELS	369
14. Logit and Probit Models for Categorical Response Variables	370
15. Generalized Linear Models	418

V. EXTENDING LINEAR AND GENERALIZED LINEAR MODELS	473
16. Time-Series Regression and Generalized Least Squares*	474
17. Nonlinear Regression	502
18. Nonparametric Regression	528
19. Robust Regression*	586
20. Missing Data in Regression Models	605
21. Bootstrapping Regression Models	647
22. Model Selection, Averaging, and Validation	669
VI. MIXED-EFFECTS MODELS	699
23. Linear Mixed-Effects Models for Hierarchical and Longitudinal Data	700
24. Generalized Linear and Nonlinear Mixed-Effects Models	743
Appendix A	759
References	762
Author Index	773
Subject Index	777
Data Set Index	791

Contents

Preface	xv
About the Author	xxiv
1. Statistical Models and Social Science	1
1.1 Statistical Models and Social Reality	1
1.2 Observation and Experiment	4
1.3 Populations and Samples	8
Exercise	10
Summary	10
Recommended Reading	11
I. DATA CRAFT	12
2. What Is Regression Analysis?	13
2.1 Preliminaries	15
2.2 Naive Nonparametric Regression	18
2.3 Local Averaging	22
Exercise	25
Summary	26
3. Examining Data	28
3.1 Univariate Displays	30
3.1.1 Histograms	30
3.1.2 Nonparametric Density Estimation	33
3.1.3 Quantile-Comparison Plots	37
3.1.4 Boxplots	41
3.2 Plotting Bivariate Data	44
3.3 Plotting Multivariate Data	47
3.3.1 Scatterplot Matrices	48
3.3.2 Coded Scatterplots	50
3.3.3 Three-Dimensional Scatterplots	50
3.3.4 Conditioning Plots	51
Exercises	53
Summary	53
Recommended Reading	54
4. Transforming Data	55
4.1 The Family of Powers and Roots	55
4.2 Transforming Skewness	59
4.3 Transforming Nonlinearity	63

4.4 Transforming Nonconstant Spread	70
4.5 Transforming Proportions	72
4.6 Estimating Transformations as Parameters*	76
Exercises	78
Summary	79
Recommended Reading	80
II. LINEAR MODELS AND LEAST SQUARES	81
5. Linear Least-Squares Regression	82
5.1 Simple Regression	83
5.1.1 Least-Squares Fit	83
5.1.2 Simple Correlation	87
5.2 Multiple Regression	92
5.2.1 Two Explanatory Variables	92
5.2.2 Several Explanatory Variables	96
5.2.3 Multiple Correlation	98
5.2.4 Standardized Regression Coefficients	100
Exercises	102
Summary	105
6. Statistical Inference for Regression	106
6.1 Simple Regression	106
6.1.1 The Simple-Regression Model	106
6.1.2 Properties of the Least-Squares Estimator	109
6.1.3 Confidence Intervals and Hypothesis Tests	111
6.2 Multiple Regression	112
6.2.1 The Multiple-Regression Model	112
6.2.2 Confidence Intervals and Hypothesis Tests	113
6.3 Empirical Versus Structural Relations	117
6.4 Measurement Error in Explanatory Variables*	120
Exercises	123
Summary	126
7. Dummy-Variable Regression	128
7.1 A Dichotomous Factor	128
7.2 Polytomous Factors	133
7.2.1 Coefficient Quasi-Variances*	138
7.3 Modeling Interactions	140
7.3.1 Constructing Interaction Regressors	141
7.3.2 The Principle of Marginality	144
7.3.3 Interactions With Polytomous Factors	145
7.3.4 Interpreting Dummy-Regression Models With Interactions	145
7.3.5 Hypothesis Tests for Main Effects and Interactions	146
7.4 A Caution Concerning Standardized Coefficients	149
Exercises	150
Summary	151
8. Analysis of Variance	153
8.1 One-Way Analysis of Variance	153
8.1.1 Example: Duncan's Data on Occupational Prestige	155
8.1.2 The One-Way ANOVA Model	156

8.2 Two-Way Analysis of Variance	159
8.2.1 Patterns of Means in the Two-Way Classification	160
8.2.2 Two-Way ANOVA by Dummy Regression	166
8.2.3 The Two-Way ANOVA Model	168
8.2.4 Fitting the Two-Way ANOVA Model to Data	170
8.2.5 Testing Hypotheses in Two-Way ANOVA	172
8.2.6 Equal Cell Frequencies	174
8.2.7 Some Cautionary Remarks	175
8.3 Higher-Way Analysis of Variance	177
8.3.1 The Three-Way Classification	177
8.3.2 Higher-Order Classifications	180
8.3.3 Empty Cells in ANOVA	186
8.4 Analysis of Covariance	187
8.5 Linear Contrasts of Means	190
Exercises	194
Summary	200
9. Statistical Theory for Linear Models*	202
9.1 Linear Models in Matrix Form	202
9.1.1 Dummy Regression and Analysis of Variance	203
9.1.2 Linear Contrasts	206
9.2 Least-Squares Fit	208
9.2.1 Deficient-Rank Parametrization of Linear Models	210
9.3 Properties of the Least-Squares Estimator	211
9.3.1 The Distribution of the Least-Squares Estimator	211
9.3.2 The Gauss-Markov Theorem	212
9.3.3 Maximum-Likelihood Estimation	214
9.4 Statistical Inference for Linear Models	215
9.4.1 Inference for Individual Coefficients	215
9.4.2 Inference for Several Coefficients	216
9.4.3 General Linear Hypotheses	219
9.4.4 Joint Confidence Regions	220
9.5 Multivariate Linear Models	225
9.6 Random Regressors	227
9.7 Specification Error	229
9.8 Instrumental Variables and Two-Stage Least Squares	231
9.8.1 Instrumental-Variables Estimation in Simple Regression	231
9.8.2 Instrumental-Variables Estimation in Multiple Regression	232
9.8.3 Two-Stage Least Squares	234
Exercises	236
Summary	241
Recommended Reading	243
10. The Vector Geometry of Linear Models*	245
10.1 Simple Regression	245
10.1.1 Variables in Mean Deviation Form	247
10.1.2 Degrees of Freedom	250
10.2 Multiple Regression	252
10.3 Estimating the Error Variance	256
10.4 Analysis-of-Variance Models	258
Exercises	260

Summary	262
Recommended Reading	264
III. LINEAR-MODEL DIAGNOSTICS	265
11. Unusual and Influential Data	266
11.1 Outliers, Leverage, and Influence	266
11.2 Assessing Leverage: Hat-Values	270
11.3 Detecting Outliers: Studentized Residuals	272
11.3.1 Testing for Outliers in Linear Models	273
11.3.2 Anscombe's Insurance Analogy	274
11.4 Measuring Influence	276
11.4.1 Influence on Standard Errors	277
11.4.2 Influence on Collinearity	280
11.5 Numerical Cutoffs for Diagnostic Statistics	280
11.5.1 Hat-Values	281
11.5.2 Studentized Residuals	281
11.5.3 Measures of Influence	281
11.6 Joint Influence	282
11.6.1 Added-Variable Plots	282
11.6.2 Forward Search	286
11.7 Should Unusual Data Be Discarded?	288
11.8 Some Statistical Details*	289
11.8.1 Hat-Values and the Hat-Matrix	289
11.8.2 The Distribution of the Least-Squares Residuals	290
11.8.3 Deletion Diagnostics	290
11.8.4 Added-Variable Plots and Leverage Plots	291
Exercises	293
Summary	294
Recommended Reading	294
12. Diagnosing Non-Normality, Nonconstant Error Variance, and Nonlinearity	296
12.1 Non-Normally Distributed Errors	297
12.1.1 Confidence Envelopes by Simulated Sampling*	300
12.2 Nonconstant Error Variance	301
12.2.1 Residual Plots	301
12.2.2 Weighted-Least-Squares Estimation*	304
12.2.3 Correcting OLS Standard Errors for Nonconstant Variance*	305
12.2.4 How Nonconstant Error Variance Affects the OLS Estimator*	306
12.3 Nonlinearity	307
12.3.1 Component-Plus-Residual Plots	308
12.3.2 Component-Plus-Residual Plots for Models With Interactions	313
12.3.3 When Do Component-Plus-Residual Plots Work?	314
12.4 Discrete Data	318
12.4.1 Testing for Nonlinearity ("Lack of Fit")	318
12.4.2 Testing for Nonconstant Error Variance	322
12.5 Maximum-Likelihood Methods*	323
12.5.1 Box-Cox Transformation of Y	324
12.5.2 Box-Tidwell Transformation of the X s	326
12.5.3 Nonconstant Error Variance Revisited	329
12.6 Structural Dimension	331

Exercises	334
Summary	338
Recommended Reading	339
13. Collinearity and Its Purported Remedies	341
13.1 Detecting Collinearity	342
13.1.1 Principal Components*	348
13.1.2 Generalized Variance Inflation*	357
13.2 Coping With Collinearity: No Quick Fix	358
13.2.1 Model Respecification	359
13.2.2 Variable Selection	359
13.2.3 Biased Estimation	361
13.2.4 Prior Information About the Regression Coefficients	364
13.2.5 Some Comparisons	365
Exercises	366
Summary	368
IV. GENERALIZED LINEAR MODELS	369
14. Logit and Probit Models for Categorical Response Variables	370
14.1 Models for Dichotomous Data	370
14.1.1 The Linear-Probability Model	372
14.1.2 Transformations of π : Logit and Probit Models	375
14.1.3 An Unobserved-Variable Formulation	379
14.1.4 Logit and Probit Models for Multiple Regression	380
14.1.5 Estimating the Linear Logit Model*	389
14.2 Models for Polytomous Data	392
14.2.1 The Polytomous Logit Model	392
14.2.2 Nested Dichotomies	399
14.2.3 Ordered Logit and Probit Models	400
14.2.4 Comparison of the Three Approaches	407
14.3 Discrete Explanatory Variables and Contingency Tables	408
14.3.1 The Binomial Logit Model*	411
Exercises	413
Summary	415
Recommended Reading	416
15. Generalized Linear Models	418
15.1 The Structure of Generalized Linear Models	418
15.1.1 Estimating and Testing GLMs	425
15.2 Generalized Linear Models for Counts	427
15.2.1 Models for Overdispersed Count Data	431
15.2.2 Loglinear Models for Contingency Tables	434
15.3 Statistical Theory for Generalized Linear Models*	443
15.3.1 Exponential Families	443
15.3.2 Maximum-Likelihood Estimation of Generalized Linear Models	445
15.3.3 Hypothesis Tests	449
15.3.4 Effect Displays	453
15.4 Diagnostics for Generalized Linear Models	453
15.4.1 Outlier, Leverage, and Influence Diagnostics	454
15.4.2 Nonlinearity Diagnostics	456

15.4.3 Collinearity Diagnostics*	459
15.5 Analyzing Data From Complex Sample Surveys	460
Exercises	464
Summary	468
Recommended Reading	471
V. EXTENDING LINEAR AND GENERALIZED LINEAR MODELS	473
16. Time-Series Regression and Generalized Least Squares*	474
16.1 Generalized Least-Squares Estimation	475
16.2 Serially Correlated Errors	476
16.2.1 The First-Order Autoregressive Process	477
16.2.2 Higher-Order Autoregressive Processes	481
16.2.3 Moving-Average and Autoregressive-Moving-Average Processes	482
16.2.4 Partial Autocorrelations	485
16.3 GLS Estimation With Autocorrelated Errors	485
16.3.1 Empirical GLS Estimation	487
16.3.2 Maximum-Likelihood Estimation	487
16.4 Correcting OLS Inference for Autocorrelated Errors	488
16.5 Diagnosing Serially Correlated Errors	489
16.6 Concluding Remarks	494
Exercises	496
Summary	499
Recommended Reading	500
17. Nonlinear Regression	502
17.1 Polynomial Regression	503
17.1.1 A Closer Look at Quadratic Surfaces*	506
17.2 Piece-wise Polynomials and Regression Splines	507
17.3 Transformable Nonlinearity	512
17.4 Nonlinear Least Squares*	515
17.4.1 Minimizing the Residual Sum of Squares	516
17.4.2 An Illustration: U.S. Population Growth	519
Exercises	521
Summary	526
Recommended Reading	527
18. Nonparametric Regression	528
18.1 Nonparametric Simple Regression: Scatterplot Smoothing	528
18.1.1 Kernel Regression	528
18.1.2 Local-Polynomial Regression	532
18.1.3 Smoothing Splines*	549
18.2 Nonparametric Multiple Regression	550
18.2.1 Local-Polynomial Multiple Regression	550
18.2.2 Additive Regression Models	563
18.3 Generalized Nonparametric Regression	572
18.3.1 Local Likelihood Estimation*	572
18.3.2 Generalized Additive Models	575
Exercises	578
Summary	580
Recommended Reading	585

19. Robust Regression*	586
19.1 <i>M</i> Estimation	586
19.1.1 Estimating Location	586
19.1.2 <i>M</i> Estimation in Regression	592
19.2 Bounded-Influence Regression	595
19.3 Quantile Regression	597
19.4 Robust Estimation of Generalized Linear Models	600
19.5 Concluding Remarks	601
Exercises	601
Summary	603
Recommended Reading	604
20. Missing Data in Regression Models	605
20.1 Missing Data Basics	606
20.1.1 An Illustration	607
20.2 Traditional Approaches to Missing Data	609
20.3 Maximum-Likelihood Estimation for Data Missing at Random*	613
20.3.1 The EM Algorithm	616
20.4 Bayesian Multiple Imputation	619
20.4.1 Inference for Individual Coefficients	621
20.4.2 Inference for Several Coefficients*	624
20.4.3 Practical Considerations	625
20.4.4 Example: A Regression Model for Infant Mortality	626
20.5 Selection Bias and Censoring	629
20.5.1 Truncated- and Censored-Normal Distributions	629
20.5.2 Heckman's Selection-Regression Model	632
20.5.3 Censored-Regression Models	637
Exercises	639
Summary	643
Recommended Reading	646
21. Bootstrapping Regression Models	647
21.1 Bootstrapping Basics	647
21.2 Bootstrap Confidence Intervals	655
21.2.1 Normal-Theory Intervals	655
21.2.2 Percentile Intervals	655
21.2.3 Improved Bootstrap Intervals	656
21.3 Bootstrapping Regression Models	658
21.4 Bootstrap Hypothesis Tests*	660
21.5 Bootstrapping Complex Sampling Designs	662
21.6 Concluding Remarks	663
Exercises	664
Summary	667
Recommended Reading	668
22. Model Selection, Averaging, and Validation	669
22.1 Model Selection	669
22.1.1 Model Selection Criteria	671
22.1.2 An Illustration: Baseball Salaries	681
22.1.3 Comments on Model Selection	683

22.2 Model Averaging*	685
22.2.1 Application to the Baseball Salary Data	687
22.2.2 Comments on Model Averaging	687
22.3 Model Validation	690
22.3.1 An Illustration: Refugee Appeals	691
22.3.2 Comments on Model Validation	693
Exercises	694
Summary	696
Recommended Reading	698
VI. MIXED-EFFECTS MODELS	699
23. Linear Mixed-Effects Models for Hierarchical and Longitudinal Data	700
23.1 Hierarchical and Longitudinal Data	701
23.2 The Linear Mixed-Effects Model	702
23.3 Modeling Hierarchical Data	704
23.3.1 Formulating a Mixed Model	708
23.3.2 Random-Effects One-Way Analysis of Variance	710
23.3.3 Random-Coefficients Regression Model	712
23.3.4 Coefficients-as-Outcomes Model	714
23.4 Modeling Longitudinal Data	717
23.5 Wald Tests for Fixed Effects	724
23.6 Likelihood-Ratio Tests of Variance and Covariance Components	726
23.7 Centering Explanatory Variables, Contextual Effects, and Fixed-Effects Models	727
23.7.1 Fixed Versus Random Effects	730
23.8 BLUPs	733
23.9 Statistical Details*	734
23.9.1 The Laird-Ware Model in Matrix Form	734
23.9.2 Wald Tests Revisited	737
Exercises	738
Summary	740
Recommended Reading	741
24. Generalized Linear and Nonlinear Mixed-Effects Models	743
24.1 Generalized Linear Mixed Models	743
24.1.1 Example: Migraine Headaches	745
24.1.2 Statistical Details*	748
24.2 Nonlinear Mixed Models*	749
24.2.1 Example: Recovery From Coma	751
24.2.2 Estimating Nonlinear Mixed Models	755
Exercises	757
Summary	757
Recommended Reading	758
Appendix A	759
References	762
Author Index	773
Subject Index	777
Data Set Index	791

Preface

Linear models, their variants, and extensions—the most important of which are *generalized* linear models—are among the most useful and widely used statistical tools for social research. This book aims to provide an accessible, in-depth, modern treatment of regression analysis, linear models, generalized linear models, and closely related methods.

The book should be of interest to students and researchers in the social sciences. Although the specific choice of methods and examples reflects this readership, I expect that the book will prove useful in other disciplines that employ regression models for data analysis and in courses on applied regression and generalized linear models where the subject matter of applications is not of special concern.

I have endeavored to make the text as accessible as possible (but no more accessible than possible—i.e., I have resisted watering down the material unduly). With the exception of four chapters, several sections, and a few shorter passages, the prerequisite for reading the book is a course in basic applied statistics that covers the elements of statistical data analysis and inference. To the extent that I could without doing violence to the material, I have tried to present even relatively advanced topics (such as methods for handling missing data and bootstrapping) in a manner consistent with this prerequisite.

Many topics (e.g., logistic regression in Chapter 14) are introduced with an example that motivates the statistics or (as in the case of bootstrapping, in Chapter 21) by appealing to familiar material. The general mode of presentation is from the specific to the general: Consequently, simple and multiple linear regression are introduced before the general linear model, and linear, logit, and probit models are introduced before generalized linear models, which subsume all the previous topics. Indeed, I could start with the generalized linear mixed-effects model (GLMM), described in the final chapter of the book, and develop all these other topics as special cases of the GLMM, but that would produce a much more abstract and difficult treatment (cf., e.g., Stroup, 2013).

The exposition of regression analysis starts (in Chapter 2) with an elementary discussion of nonparametric regression, developing the notion of regression as a conditional average—in the absence of restrictive assumptions about the nature of the relationship between the response and explanatory variables. This approach begins closer to the data than the traditional starting point of linear least-squares regression and should make readers skeptical about glib assumptions of linearity, constant variance, and so on.

More difficult chapters and sections are marked with asterisks. These parts of the text can be omitted without loss of continuity, but they provide greater understanding and depth, along with coverage of some topics that depend on more extensive mathematical or statistical background. I do not, however, wish to exaggerate the background that is required for this “more difficult” material: All that is necessary is some exposure to matrices, elementary linear algebra, elementary differential calculus, and some basic ideas from probability and mathematical statistics. Appendices to the text provide the background required for understanding the more advanced material.

All chapters include summary information in boxes interspersed with the text and at the end of the chapter, and most conclude with recommendations for additional reading. You will find theoretically focused exercises at the end of most chapters, some extending the material in the text. More difficult, and occasionally challenging, exercises are marked with asterisks. In addition, data-analytic exercises for each chapter are available on the website for the book, along with the associated data sets.

What Is New in the Third Edition?

The first edition of this book, published by Sage in 1997 and entitled *Applied Regression, Linear Models, and Related Methods*, originated in my 1984 text *Linear Statistical Models and Related Methods* and my 1991 monograph *Regression Diagnostics*. The title of the 1997 edition reflected a change in organization and emphasis: I thoroughly reworked the book, removing some topics and adding a variety of new material. But even more fundamentally, the book was extensively rewritten. It was a new and different book from my 1984 text.

The second edition had a (slightly) revised title, making reference to “generalized linear models” rather than to “linear models” (and dropping the reference to “related methods” as unnecessary), reflecting another change in emphasis. I retain that title for the third edition. There was quite a bit of new material in the second edition, and some of the existing material was reworked and rewritten, but the general level and approach of the book was similar to the first edition, and most of the material in the first edition, especially in Parts I through III (see below), was preserved in the second edition. I was gratified by the reception of the first and second editions of this book by reviewers and other readers; although I felt the need to bring the book up to date and to improve it in some respects, I also didn’t want to “fix what ain’t broke.”

The second edition introduced a new chapter on generalized linear models, greatly augmenting a very brief section on this topic in the first edition. What were previously sections on time-series regression, nonlinear regression, nonparametric regression, robust regression, and bootstrapping became separate chapters, many with extended treatments of their topics. I added a chapter on missing data and another on model selection, averaging, and validation (incorporating and expanding material on model validation from the previous edition).

Although I have made small changes throughout the text, the principal innovation in the third edition is the introduction of a new section on mixed-effects models for hierarchical and longitudinal data, with chapters on linear mixed-effects models and on nonlinear and generalized linear mixed-effects models (Chapters 23 and 24). These models are used increasingly in social research and I thought that it was important to incorporate them in this text. There is also a revised presentation of analysis of variance models in Chapter 8, which includes a simplified treatment, allowing readers to skip the more complex aspects of the topic, if they wish; an introduction to instrumental-variables estimation and two-stage least squares in Chapter 9; and a brief consideration of design-based inference for statistical models fit to data from complex survey samples in Chapter 15.

As in the second edition, the appendices to the book (with the exception of Appendix A on notation) are on the website for the book. In addition, data-analytic exercises and data sets from the book are on the website.

Synopsis

Chapter 1 discusses the role of statistical data analysis in social science, expressing the point of view that statistical models are essentially descriptive, not direct (if abstract)

representations of social processes. This perspective provides the foundation for the data-analytic focus of the text.

Part I: Data Craft

The first part of the book consists of preliminary material:¹

Chapter 2 introduces the notion of regression analysis as tracing the conditional distribution of a response variable as a function of one or several explanatory variables. This idea is initially explored “nonparametrically,” in the absence of a restrictive statistical model for the data (a topic developed more extensively in Chapter 18).

Chapter 3 describes a variety of graphical tools for examining data. These methods are useful both as a preliminary to statistical modeling and to assist in the diagnostic checking of a model that has been fit to data (as discussed, e.g., in Part III).

Chapter 4 discusses variable transformation as a solution to several sorts of problems commonly encountered in data analysis, including skewness, nonlinearity, and nonconstant spread.

Part II: Linear Models and Least Squares

The second part, on linear models fit by the method of least squares, along with Part III on diagnostics and Part IV on generalized linear models, comprises the heart of the book:

Chapter 5 discusses linear least-squares regression. Linear regression is the prototypical linear model, and its direct extension is the subject of Chapters 7 to 10.

Chapter 6, on statistical inference in regression, develops tools for testing hypotheses and constructing confidence intervals that apply generally to linear models. This chapter also introduces the basic methodological distinction between empirical and structural relationships—a distinction central to understanding causal inference in nonexperimental research.

Chapter 7 shows how “dummy variables” can be employed to extend the regression model to qualitative explanatory variables (or “factors”). Interactions among explanatory variables are introduced in this context.

Chapter 8, on analysis of variance models, deals with linear models in which all the explanatory variables are factors.

*Chapter 9** develops the statistical theory of linear models, providing the foundation for much of the material in Chapters 5 to 8 along with some additional, and more general, results. This chapter also includes an introduction to instrumental-variables estimation and two-stage least squares.

*Chapter 10** applies vector geometry to linear models, allowing us literally to visualize the structure and properties of these models. Many topics are revisited from the geometric perspective, and central concepts—such as “degrees of freedom”—are given a natural and compelling interpretation.

¹I believe that it was Michael Friendly of York University who introduced me to the term *data craft*, a term that aptly characterizes the content of this section and, indeed, of the book more generally.

Part III: Linear-Model Diagnostics

The third part of the book describes “diagnostic” methods for discovering whether a linear model fit to data adequately represents the data. Methods are also presented for correcting problems that are revealed:

Chapter 11 deals with the detection of unusual and influential data in linear models.

Chapter 12 describes methods for diagnosing a variety of problems, including non-normally distributed errors, nonconstant error variance, and nonlinearity. Some more advanced material in this chapter shows how the method of maximum likelihood can be employed for selecting transformations.

Chapter 13 takes up the problem of collinearity—the difficulties for estimation that ensue when the explanatory variables in a linear model are highly correlated.

Part IV: Generalized Linear Models

The fourth part of the book is devoted to generalized linear models, a grand synthesis that incorporates the linear models described earlier in the text along with many of their most important extensions:

Chapter 14 takes up linear-like logit and probit models for qualitative and ordinal categorical response variables. This is an important topic because of the ubiquity of categorical data in the social sciences (and elsewhere).

Chapter 15 describes the generalized linear model, showing how it encompasses linear, logit, and probit models along with statistical models (such as Poisson and gamma regression models) not previously encountered in the text. The chapter includes a treatment of diagnostic methods for generalized linear models, extending much of the material in Part III, and ends with an introduction to inference for linear and generalized linear models in complex survey samples.

Part V: Extending Linear and Generalized Linear Models

The fifth part of the book discusses important extensions of linear and generalized linear models. In selecting topics, I was guided by the proximity of the methods to linear and generalized linear models and by the promise that these methods hold for data analysis in the social sciences. The methods described in this part of the text are given introductory—rather than extensive—treatments. My aim in introducing these relatively advanced topics is to provide (1) enough information so that readers can begin to use these methods in their research and (2) sufficient background to support further work in these areas should readers choose to pursue them. To the extent possible, I have tried to limit the level of difficulty of the exposition, and only Chapter 19 on robust regression is starred in its entirety (because of its essential reliance on basic calculus).

Chapter 16 describes time-series regression, where the observations are ordered in time and hence cannot usually be treated as statistically independent. The chapter introduces the method of generalized least squares, which can take account of serially correlated errors in regression.

Chapter 17 takes up nonlinear regression models, showing how some nonlinear models can be fit by linear least squares after transforming the model to linearity, while other, fundamentally nonlinear, models require the method of nonlinear least squares. The chapter includes treatments of polynomial regression and regression splines, the latter closely related to the topic of the subsequent chapter.

Chapter 18 introduces nonparametric regression analysis, which traces the dependence of the response on the explanatory variables in a regression without assuming a particular functional form for their relationship. This chapter contains a discussion of generalized nonparametric regression, including generalized additive models.

Chapter 19 describes methods of robust regression analysis, which are capable of automatically discounting unusual data.

Chapter 20 discusses missing data, explaining the potential pitfalls lurking in common approaches to missing data, such as complete-case analysis, and describing more sophisticated methods, such as multiple imputation of missing values. This is an important topic because social science data sets are often characterized by a large proportion of missing data.

Chapter 21 introduces the “bootstrap,” a computationally intensive simulation method for constructing confidence intervals and hypothesis tests. In its most common nonparametric form, the bootstrap does not make strong distributional assumptions about the data, and it can be made to reflect the manner in which the data were collected (e.g., in complex survey sampling designs).

Chapter 22 describes methods for model selection, model averaging in the face of model uncertainty, and model validation. Automatic methods of model selection and model averaging, I argue, are most useful when a statistical model is to be employed for prediction, less so when the emphasis is on interpretation. Validation is a simple method for drawing honest statistical inferences when—as is commonly the case—the data are employed both to select a statistical model and to estimate its parameters.

Part VI: Mixed-Effects Models

Part VI, new to the third edition of the book, develops linear, generalized linear, and nonlinear mixed-effects models for clustered data, extending regression models for independent observations covered earlier in the text. As in Part V, my aim is to introduce readers to the topic, providing a basis for applying these models in practice as well as for reading more extensive treatments of the subject. As mentioned earlier in this preface, mixed-effects models are in wide use in the social sciences, where they are principally applied to hierarchical and longitudinal data.

Chapter 23 introduces linear mixed-effects models and describes the fundamental issues that arise in the analysis of clustered data through models that incorporate random effects. Illustrative applications include both hierarchical and longitudinal data.

Chapter 24 describes generalized linear mixed-effects models for non-normally distributed response variables, such as logistic regression for a dichotomous response, and Poisson and related regression models for count data. The chapter also introduces nonlinear mixed-effects models for fitting fundamentally nonlinear equations to clustered data.

Appendices

Several appendices provide background, principally—but not exclusively—for the starred portions of the text. With the exception of Appendix A, which is printed at the back of the book, all the appendices are on the website for the book.

Appendix A describes the notational conventions employed in the text.

Appendix B provides a basic introduction to matrices, linear algebra, and vector geometry, developing these topics from first principles. Matrices are used extensively in statistics, including in the starred portions of this book. Vector geometry provides the basis for the material in Chapter 10 on the geometry of linear models.

Appendix C reviews powers and logs and the geometry of lines and planes, introduces elementary differential and integral calculus, and shows how, employing matrices, differential calculus can be extended to several independent variables. Calculus is required for some starred portions of the text—for example, the derivation of least-squares and maximum-likelihood estimators. More generally in statistics, calculus figures prominently in probability theory and in optimization problems.

Appendix D provides an introduction to the elements of probability theory and to basic concepts of statistical estimation and inference, including the essential ideas of Bayesian statistical inference. The background developed in this appendix is required for some of the material on statistical inference in the text and for certain other topics, such as multiple imputation of missing data and model averaging.

Computing

Nearly all the examples in this text employ real data from the social sciences, many of them previously analyzed and published. The online exercises that involve data analysis also almost all use real data drawn from various areas of application. I encourage readers to analyze their own data as well.

The data sets for examples and exercises can be downloaded free of charge via the World Wide Web; point your web browser at www.sagepub.com/fox3e. Appendices and exercises are distributed as portable document format (PDF) files.

I occasionally comment in passing on computational matters, but the book generally ignores the finer points of statistical computing in favor of methods that are computationally simple. I feel that this approach facilitates learning. Thus, for example, linear least-squares coefficients are obtained by solving the normal equations formed from sums of squares and products of the variables rather than by a more numerically stable method. Once basic techniques are absorbed, the data analyst has recourse to carefully designed programs for statistical computations.

I think that it is a mistake to tie a general discussion of linear and related statistical models too closely to particular software. Any reasonably capable statistical software will do almost everything described in this book. My current personal choice of statistical software, both for research and for teaching, is R—a free, open-source implementation of the S statistical programming language and computing environment (Ihaka & Gentleman, 1996; R Core Team, 2014). R is now the dominant statistical software among statisticians; it is used increasingly in the social sciences but is by no means dominant there. I have coauthored a separate book (Fox & Weisberg, 2011) that provides a general introduction to R and that describes its use in applied regression analysis.

Reporting Numbers in Examples

A note on how numbers are reported in the data analysis examples: I typically show numbers to four or five significant digits in tables and in the text. This is greater precision than is usually desirable in research reports, where showing two or three significant digits makes the results more digestible by readers. But my goal in this book is generally to allow the reader to reproduce the results shown in examples. In many instances, numbers in examples are computed from each other, rather than being taken directly from computer output; in these instances, a reader comparing the results in the text to those in computer output may encounter small differences, usually of one unit in the last decimal place.

To Readers, Students, and Instructors

I have used the material in this book and its predecessors for two types of courses (along with a variety of short courses and lectures):

- I cover the unstarred sections of Chapters 1 to 8, 11 to 15, 20, and 22 in a one-semester course for social science graduate students (at McMaster University in Hamilton, Ontario, Canada) who have had (at least) a one-semester introduction to statistics at the level of Moore, Notz, and Fligner (2013). The outline of this course is as follows:

Week	Topic	Reading (Chapter)
1	Introduction to the course and to regression	1, 2
2	Examining and transforming data	3, 4
3	Linear least-squares regression	5
4	Statistical inference for regression	6
5	Dummy-variable regression and analysis of variance	7, 8
6	Review: Through dummy regression	
7	Diagnostics I: Unusual and influential data	11
8	Diagnostics II: Nonlinearity and other ills	12
9	Diagnostics III: Collinearity and model selection	13, 22
10	Logit and probit models for dichotomous data	14
11	Logit and probit models for polychoric data	14
12	Generalized linear models	15
13	Missing data	20
14	Review: From analysis of variance	

These readings are supplemented by selections from *An R Companion to Applied Regression, Second Edition* (Fox & Weisberg, 2011). Students complete required weekly homework assignments, which focus primarily on data analysis. Homework is collected and corrected but not

graded. I distribute answers after the homework is collected. There are midterm and final take-home exams (after the review classes), also focused on data analysis.

- I used the material in the predecessors of Chapters 1 to 15 and the several appendices for a two-semester course for social science graduate students (at York University in Toronto) with similar statistical preparation. For this second, more intensive, course, background topics (such as linear algebra) were introduced as required and constituted about one fifth of the course. The organization of the course was similar to the first one.

Both courses include some treatment of statistical computing, with more information on programming in the second course. For students with the requisite mathematical and statistical background, it should be possible to cover almost all the text in a reasonably paced two-semester course.

In learning statistics, it is important for the reader to participate actively, both by working through the arguments presented in the book and—even more important—by applying methods to data. Statistical data analysis is a *craft*, and, like any craft, developing proficiency requires effort and practice. Reworking examples is a good place to start, and I have presented illustrations in such a manner as to facilitate reanalysis and further analysis of the data.

Where possible, I have relegated formal “proofs” and derivations to exercises, which nevertheless typically provide some guidance to the reader. I believe that this type of material is best learned constructively. As well, including too much algebraic detail in the body of the text invites readers to lose the statistical forest for the mathematical trees. You can decide for yourself (or your students) whether or not to work the theoretical exercises. It is my experience that some people feel that the process of working through derivations cements their understanding of the statistical material, while others find this activity tedious and pointless. Some of the theoretical exercises, marked with asterisks, are comparatively difficult. (Difficulty is assessed relative to the material in the text, so the threshold is higher in starred sections and chapters.)

In preparing the data-analytic exercises, I have tried to find data sets of some intrinsic interest that embody a variety of characteristics. In many instances, I try to supply some direction in the data-analytic exercises, but—like all real-data analysis—these exercises are fundamentally open-ended. It is therefore important for instructors to set aside time to discuss data-analytic exercises in class, both before and after students tackle them. Although students often miss important features of the data in their initial analyses, this experience—properly approached and integrated—is an unavoidable part of learning the craft of data analysis.

A few exercises, marked with pound-signs (#) are meant for “hand” computation. Hand computation (i.e., with a calculator) is tedious, and is practical only for unrealistically small problems, but it sometimes serves to make statistical procedures more concrete (and increases our admiration for our pre-computer-era predecessors). Similarly, despite the emphasis in the text on analyzing real data, a small number of exercises generate simulated data to clarify certain properties of statistical methods.

I struggled with the placement of cross-references to exercises and to other parts of the text, trying brackets [too distracting!], marginal boxes (too imprecise), and finally settling on traditional footnotes.² I suggest that you ignore both the cross-references and the other footnotes on first reading of the text.³

Finally, a word about style: I try to use the first person singular—“I”—when I express opinions. “We” is reserved for you—the reader—and I.

²Footnotes are a bit awkward, but you don’t have to read them.

³Footnotes other than cross-references generally develop small points and elaborations.

Acknowledgments

Many individuals have helped me in the preparation of this book.

I am grateful to the York University Statistical Consulting Service study group, which read, commented on, and corrected errors in the manuscript, both of the previous edition of the book and of the new section on mixed-effects models introduced in this edition.

A number of friends and colleagues donated their data for illustrations and exercises—implicitly subjecting their research to scrutiny and criticism.

Several individuals contributed to this book by making helpful comments on it and its predecessors (Fox, 1984, 1997, 2008): Patricia Ahmed, University of Kentucky; Robert Andersen; A. Alexander Beaujean, Baylor University; Ken Bollen; John Brehm, University of Chicago; Gene Denzel; Shirley Dowdy; Michael Friendly; E. C. Hedberg, NORC at the University of Chicago; Paul Herzberg; Paul Johnston; Michael S. Lynch, University of Georgia; Vida Maralani, Yale University; William Mason; Georges Monette; A. M. Parkhurst, University of Nebraska-Lincoln; Doug Rivers; Paul D. Sampson, University of Washington; Corey S. Sparks, The University of Texas at San Antonio; Robert Stine; and Sanford Weisberg. I am also in debt to Paul Johnson's students at the University of Kansas, to William Mason's students at UCLA, to Georges Monette's students at York University, to participants at the Inter-University Consortium for Political and Social Research Summer Program in Robert Andersen's advanced regression course, and to my students at McMaster University, all of whom were exposed to various versions of the second edition of this text prior to publication and who improved the book through their criticism, suggestions, and—occasionally—informative incomprehension.

Edward Ng capably assisted in the preparation of some of the figures that appear in the book.

C. Deborah Laughton, Lisa Cuevas, Sean Connelly, and—most recently—Vicki Knight, my editors at Sage Publications, were patient and supportive throughout the several years that I worked on the various editions of the book.

I have been very lucky to have colleagues and collaborators who have been a constant source of ideas and inspiration—in particular, Michael Friendly and Georges Monette at York University in Toronto and Sanford Weisberg at the University of Minnesota. I am sure that they will recognize their influence on this book. I owe a special debt to Georges Monette for his contributions, both direct and indirect, to the new chapters on mixed-effects models in this edition. Georges generously shared his materials on mixed-effects models with me, and I have benefited from his insights on the subject (and others) over a period of many years.

Finally, a number of readers have contributed corrections to earlier editions of the text, and I thank them individually in the posted errata to these editions. Paul Laumans deserves particular mention for his assiduous pursuit of typographical errors. No doubt I'll have occasion in due course to thank readers for corrections to the current edition.

If, after all this help and the opportunity to prepare a new edition of the book, deficiencies remain, then I alone am at fault.

John Fox
Toronto, Canada
August 2014

About the Author

John Fox is Professor of Sociology at McMaster University in Hamilton, Ontario, Canada. Professor Fox earned a PhD in Sociology from the University of Michigan in 1972, and prior to arriving at McMaster, he taught at the University of Alberta and at York University in Toronto, where he was cross-appointed in the Sociology and Mathematics and Statistics departments and directed the university's Statistical Consulting Service. He has delivered numerous lectures and workshops on statistical topics in North and South America, Europe, and Asia, at such places as the Summer Program of the Inter-University Consortium for Political and Social Research, the Oxford University Spring School in Quantitative Methods for Social Research, and the annual meetings of the American Sociological Association. Much of his recent work has been on formulating methods for visualizing complex statistical models and on developing software in the R statistical computing environment. He is the author and coauthor of many articles, in such journals as *Sociological Methodology*, *Sociological Methods and Research*, *The Journal of the American Statistical Association*, *The Journal of Statistical Software*, *The Journal of Computational and Graphical Statistics*, *Statistical Science*, *Social Psychology Quarterly*, *The Canadian Review of Sociology and Anthropology*, and *The Canadian Journal of Sociology*. He has written a number of other books, including *Regression Diagnostics* (1991), *Nonparametric Simple Regression* (2000), *Multiple and Generalized Nonparametric Regression* (2000), *A Mathematical Primer for Social Statistics* (2008), and, with Sanford Weisberg, *An R Companion to Applied Regression* (2nd ed., 2010). Professor Fox also edits the Sage Quantitative Applications in the Social Sciences (“QASS”) monograph series.

1

Statistical Models and Social Science

The social world is exquisitely complex and rich. From the improbable moment of birth, each of our lives is governed by chance and contingency. The statistical models typically used to analyze social data—and, in particular, the models considered in this book—are, in contrast, ludicrously simple. How can simple statistical models help us to understand a complex social reality? As the statistician George Box famously remarked (e.g., in Box, 1979), “All models are wrong but some are useful” (p. 202). Can statistical models be useful in the social sciences?

This is a book on data analysis and statistics, not on the philosophy of the social sciences. I will, therefore, address this question, and related issues, very briefly here. Nevertheless, I feel that it is useful to begin with a consideration of the role of data analysis in the larger process of social research. You need not agree with the point of view that I express in this chapter to make productive use of the statistical tools presented in the remainder of the book, but the emphasis and specific choice of methods in the text partly reflect the ideas in this chapter. You may wish to reread this material after you study the methods described in the sequel.

1.1 Statistical Models and Social Reality

As I said, social reality is complex: Consider how my income is “determined.” I am a relatively well-paid professor in the sociology department of a Canadian university. That the billiard ball of my life would fall into this particular pocket was, however, hardly predictable a half-century ago, when I was attending a science high school in New York City. My subsequent decision to study sociology at New York’s City College (after several other majors), my interest in statistics (the consequence of a course taken without careful consideration in my senior year), my decision to attend graduate school in sociology at the University of Michigan (one of several more or less equally attractive possibilities), and the opportunity and desire to move to Canada (the vote to hire me at the University of Alberta was, I later learned, very close) are all events that could easily have occurred differently.

I do not mean to imply that personal histories are completely capricious, unaffected by social structures of race, ethnicity, class, gender, and so on, just that they are not *in detail* determined by these structures. That social structures—and other sorts of systematic factors—condition, limit, and encourage specific events is clear from each of the illustrations in the previous paragraph and in fact makes sense of the argument for the statistical analysis of social data presented below. To take a particularly gross example: The public high school that I attended admitted its students by competitive examination, but no young women could apply (a policy that has happily changed).

Each of these precarious occurrences clearly affected my income, as have other events—some significant, some small—too numerous and too tedious to mention, even if I were aware of them all. If, for some perverse reason, you were truly interested in my income (and, perhaps, in other matters more private), you could study my biography and through that study arrive at a detailed (if inevitably incomplete) understanding. It is clearly impossible, however, to pursue this strategy for many individuals or, more to the point, for individuals in general.

Nor is an understanding of income in general inconsequential, because income inequality is an (increasingly, as it turns out) important feature of our society. If such an understanding hinges on a literal description of the process by which each of us receives an income, then the enterprise is clearly hopeless. We might, alternatively, try to capture significant features of the process in general without attempting to predict the outcome for specific individuals. One could draw formal analogies (largely unproductively, I expect, although some have tried) to chaotic physical processes, such as the determination of weather and earthquakes.

Concrete mathematical theories purporting to describe social processes sometimes appear in the social sciences (e.g., in economics and in some areas of psychology), but they are relatively rare.¹ If a theory, like Newton's laws of motion, is mathematically concrete, then, to be sure, there are difficulties in applying and testing it; but, with some ingenuity, experiments and observations can be devised to estimate the free parameters of the theory (a gravitational constant, for example) and to assess the fit of the theory to the resulting data.

In the social sciences, *verbal* theories abound. These social theories tend to be vague, elliptical, and highly qualified. Often, they are, at least partially, a codification of “common sense.” I believe that vague social theories are potentially useful abstractions for understanding an intrinsically complex social reality, but how can such theories be linked empirically to that reality?

A vague social theory may lead us to expect, for example, that racial prejudice is the partial consequence of an “authoritarian personality,” which, in turn, is a product of rigid childrearing. Each of these terms requires elaboration and procedures of assessment or measurement. Other social theories may lead us to expect that higher levels of education should be associated with higher levels of income, perhaps because the value of labor power is enhanced by training, because occupations requiring higher levels of education are of greater functional importance, because those with higher levels of education are in relatively short supply, or because people with high educational attainment are more capable in the first place. In any event, we need to consider how to assess income and education, how to examine their relationship, and what other variables need to be included.²

Statistical models of the type considered in this book are grossly *simplified* descriptions of complex social reality. Imagine that we have data from a social survey of a large sample of employed individuals. Imagine further, anticipating the statistical methods described in subsequent chapters, that we regress these individuals' income on a variety of putatively relevant characteristics, such as their level of education, gender, race, region of residence, and so on. We recognize that a model of this sort will fail to account perfectly for individuals' incomes, so our model includes a “residual,” meant to capture the component of income unaccounted

¹The methods for fitting nonlinear models described in Chapter 17 are sometimes appropriate to the rare theories in social science that are mathematically concrete.

²See Section 1.2.

for by the systematic part of the model, which incorporates the “effects” on income of education, gender, and so forth.

The residuals for our model are likely very large. Even if the residuals were small, however, we would still need to consider the relationships among our social “theory,” the statistical model that we have fit to the data, and the social “reality” that we seek to understand. Social reality, along with our methods of observation, produces the data; our theory aims to explain the data, and the model to describe them. That, I think, is the key point: Statistical models are almost always fundamentally *descriptive*.

I believe that a statistical model cannot, and is not literally meant, to capture the social process by which incomes are “determined.” As I argued above, individuals receive their incomes as a result of their almost unimaginably complex personal histories. No regression model, not even one including a residual, can reproduce this process: It is not as if my income is partly determined by my education, gender, race, and so on, and partly by the detailed trajectory of my life. It is, therefore, not sensible, at the level of real social processes, to relegate chance and contingency to a random term that is simply added to the systematic part of a statistical model. The unfortunate tendency to *reify* statistical models—to forget that they are descriptive summaries, not literal accounts of social processes—can only serve to discredit quantitative data analysis in the social sciences.

Nevertheless, and despite the rich chaos of individuals’ lives, social theories imply a structure to income inequality. Statistical models are capable of capturing and describing that structure or at least significant aspects of it. Moreover, social research is often motivated by questions rather than by hypotheses: Has income inequality between men and women changed recently? Is there a relationship between public concern over crime and the level of crime? Data analysis can help to answer these questions, which frequently are of practical—as well as theoretical—concern. Finally, if we proceed carefully, data analysis can assist us in the discovery of social facts that initially escape our hypotheses and questions.

It is, in my view, a paradox that the statistical models that are at the heart of most modern quantitative social science are at once taken too seriously and not seriously enough by many practitioners of social science. On one hand, social scientists write about simple statistical models as if they were direct representations of the social processes that they purport to describe. On the other hand, there is frequently a failure to attend to the descriptive accuracy of these models.

As a shorthand, reference to the “effect” of education on income is innocuous. That the shorthand often comes to dominate the interpretation of statistical models is reflected, for example, in much of the social science literature that employs structural-equation models (once commonly termed “causal models,” a usage that has thankfully declined). There is, I believe, a valid sense in which income is “affected” by education, because the complex real process by which individuals’ incomes are determined is partly conditioned by their levels of education, but—as I have argued above—one should not mistake the model for the process.³

Although statistical models are very simple in comparison to social reality, they typically incorporate strong claims about the descriptive pattern of data. These claims rarely reflect the

³There is the danger here of simply substituting one term (“conditioned by”) for another (“affected by”), but the point is deeper than that: Education affects income because the choices and constraints that partly structure individuals’ lives change systematically with their level of education. Many highly paid occupations in our society are closed to individuals who lack a university education, for example. To recognize this fact, and to examine its descriptive reflection in a statistical summary, is different from claiming that a university education literally adds an increment to individuals’ incomes.

substantive social theories, hypotheses, or questions that motivate the use of the statistical models, and they are very often wrong. For example, it is common in social research to assume *a priori*, and without reflection, that the relationship between two variables, such as income and education, is linear. Now, we may well have good reason to believe that income tends to be *higher* at higher levels of education, but there is no reason to suppose that this relationship is *linear*. Our practice of data analysis should reflect our ignorance as well as our knowledge.

A statistical model is of no practical use if it is an inaccurate description of the data, and we will, therefore, pay close attention to the descriptive accuracy of statistical models. Unhappily, the converse is not true, for a statistical model may be descriptively accurate but of little practical use; it may even be descriptively accurate but substantively misleading. We will explore these issues briefly in the next two sections, which tie the interpretation of statistical models to the manner in which data are collected.

With few exceptions, statistical data analysis describes the outcomes of real social processes and not the processes themselves. It is therefore important to attend to the descriptive accuracy of statistical models and to refrain from reifying them.

1.2 Observation and Experiment

It is common for (careful) introductory accounts of statistical methods (e.g., Freedman, Pisani, & Purves, 2007; Moore, Notz, & Fligner, 2013) to distinguish strongly between observational and experimental data. According to the standard distinction, causal inferences are justified (or, at least, more certain) in experiments, where the explanatory variables (i.e., the possible “causes”) are under the direct control of the researcher; causal inferences are especially compelling in a randomized experiment, in which the values of explanatory variables are assigned by some chance mechanism to experimental units. In nonexperimental research, in contrast, the values of the explanatory variables are observed—not assigned—by the researcher, along with the value of the response variable (the “effect”), and causal inferences are not justified (or, at least, are less certain). I believe that this account, although essentially correct, requires qualification and elaboration.

To fix ideas, let us consider the data summarized in Table 1.1, drawn from a paper by Greene and Shaffer (1992) on Canada’s refugee determination process. This table shows the outcome of 608 cases, filed in 1990, in which refugee claimants who were turned down by the Immigration and Refugee Board asked the Federal Court of Appeal for leave to appeal the board’s determination. In each case, the decision to grant or deny leave to appeal was made by a single judge. It is clear from the table that the 12 judges who heard these cases differed widely in the percentages of cases that they granted leave to appeal. Employing a standard significance test for a contingency table (a chi-square test of independence), Greene and Shaffer calculated that a relationship as strong as the one in the table will occur by chance alone about two times in 100,000. These data became the basis for a court case contesting the fairness of the Canadian refugee determination process.

If the 608 cases had been assigned at random to the judges, then the data would constitute a natural experiment, and we could unambiguously conclude that the large differences among

Table 1.1 Percentages of Refugee Claimants in 1990 Who Were Granted or Denied Leave to Appeal a Negative Decision of the Canadian Immigration and Refugee Board, Classified by the Judge Who Heard the Case

Judge	Leave Granted?			Number of cases
	Yes	No	Total	
Pratte	9	91	100	57
Linden	9	91	100	32
Stone	12	88	100	43
Iacobucci	12	88	100	33
Décary	20	80	100	80
Hugessen	26	74	100	65
Urie	29	71	100	21
MacGuigan	30	70	100	90
Heald	30	70	100	46
Mahoney	34	66	100	44
Marceau	36	64	100	50
Desjardins	49	51	100	47
All judges	25	75	100	608

SOURCE: Adapted from Table 1 in Greene and Shaffer, "Leave to Appeal and Leave to Commence Judicial Review in Canada's Refugee-Determination System: Is the Process Fair?" *International Journal of Refugee Law*, 1992, Vol. 4, No. 1, p. 77, by permission of Oxford University Press.

the judges reflect differences in their propensities to grant leave to appeal.⁴ The cases were, however, assigned to the judges not randomly but on a rotating basis, with a single judge hearing all of the cases that arrived at the court in a particular week. In defending the current refugee determination process, expert witnesses for the Crown argued that the observed differences among the judges might therefore be due to characteristics that systematically differentiated the cases that different judges happened to hear.

It is possible, in practice, to "control" statistically for such extraneous "confounding" variables as may explicitly be identified, but it is not, in principle, possible to control for *all* relevant explanatory variables, because we can never be certain that all relevant variables have been identified.⁵ Nevertheless, I would argue, the data in Table 1.1 establish a *prima facie* case for systematic differences in the judges' propensities to grant leave to appeal to refugee claimants. Careful researchers control statistically for potentially relevant variables that they can identify; cogent critics demonstrate that an omitted confounding variable accounts for the observed association between judges and decisions or at least argue persuasively that a specific omitted variable *may* be responsible for this association—they do not simply maintain the abstract possibility that such a variable *may* exist.

⁴Even so, this inference is not reasonably construed as a representation of the *cognitive process* by which judges arrive at their determinations. Following the argument in the previous section, it is unlikely that we could ever trace out that process in detail; it is quite possible, for example, that a specific judge would make different decisions faced with the same case on different occasions.

⁵See the further discussion of the refugee data in Section 22.3.1.

What makes an omitted variable “relevant” in this context?⁶

1. The omitted variable must influence the response. For example, if the gender of the refugee applicant has no impact on the judges’ decisions, then it is irrelevant to control statistically for gender.
2. The omitted variable must be related as well to the explanatory variable that is the focus of the research. Even if the judges’ decisions are influenced by the gender of the applicants, the relationship between outcome and judge will be unchanged by controlling for gender (e.g., by looking separately at male and female applicants) unless the gender of the applicants is also related to judges—that is, unless the different judges heard cases with substantially different proportions of male and female applicants.

The strength of randomized experimentation derives from the second point: If cases were randomly assigned to judges, then there would be no systematic tendency for them to hear cases with differing proportions of men and women—or, for that matter, with systematic differences of *any* kind.

It is, however, misleading to conclude that causal inferences are completely unambiguous in experimental research, even within the bounds of statistical uncertainty (expressed, for example, in the *p*-value of a statistical test). Although we can unambiguously ascribe an observed difference to an experimental *manipulation*, we cannot unambiguously identify that manipulation with the explanatory variable that is the focus of our research.

In a randomized drug study, for example, in which patients are prescribed a new drug or an inactive placebo, we may establish with virtual certainty that there was greater average improvement among those receiving the drug, but we cannot be sure that this difference is due (or solely due) to the putative active ingredient in the drug. Perhaps the experimenters inadvertently conveyed their enthusiasm for the drug to the patients who received it, influencing the patients’ responses, or perhaps the bitter taste of the drug subconsciously convinced these patients of its potency.

Experimenters try to rule out alternative interpretations of this kind by following careful experimental practices, such as “double-blind” delivery of treatments (neither the subject nor the experimenter knows whether the subject is administered the drug or the placebo), and by holding constant potentially influential characteristics deemed to be extraneous to the research (the taste, color, shape, etc., of the drug and placebo are carefully matched). One can never be certain, however, that *all* relevant variables are held constant in this manner. Although the degree of certainty achieved is typically much greater in a randomized experiment than in an observational study, the distinction is less clear-cut than it at first appears.

Causal inferences are most certain—if not completely definitive—in randomized experiments, but observational data can also be reasonably marshaled as evidence of causation. Good experimental practice seeks to avoid confounding experimentally manipulated explanatory variables with other variables that can influence the response variable. Sound analysis of observational data seeks to control statistically for potentially confounding variables.

⁶These points are developed more formally in Sections 6.3 and 9.7.

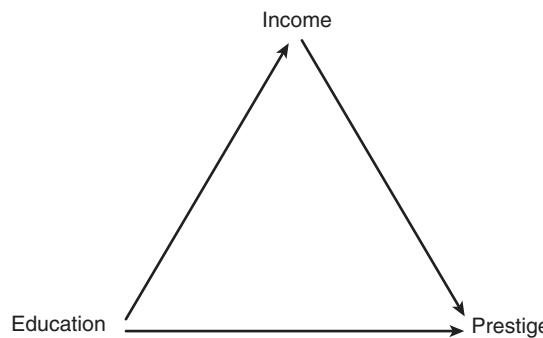


Figure 1.1 Simple “causal model” relating education, income, and prestige of occupations. Education is a common prior cause of both income and prestige; income intervenes causally between education and prestige.

In subsequent chapters, we will have occasion to examine observational data on the prestige, educational level, and income level of occupations. It will materialize that occupations with higher levels of education tend to have higher prestige and that occupations with higher levels of income also tend to have higher prestige. The income and educational levels of occupations are themselves positively related. As a consequence, when education is controlled statistically, the relationship between prestige and income grows smaller; likewise, when income is controlled, the relationship between prestige and education grows smaller. In neither case, however, does the relationship disappear.

How are we to understand the pattern of statistical associations among the three variables? It is helpful in this context to entertain an informal “causal model” for the data, as in Figure 1.1. That is, the educational level of occupations influences (potentially) both their income level and their prestige, while income potentially influences prestige. The association between prestige and income is “spurious” (i.e., not causal) to the degree that it is a consequence of the mutual dependence of these two variables on education; the reduction in this association when education is controlled represents the removal of the spurious component. In contrast, the causal relationship between education and prestige is partly mediated by the “intervening variable” income; the reduction in this association when income is controlled represents the articulation of an “indirect” effect of education on prestige (i.e., through income).

In the former case, we partly *explain away* the association between income and prestige: Part of the relationship is “really” due to education. In the latter case, we partly *explain* the association between education and prestige: Part of the relationship is mediated by income.

In analyzing observational data, it is important to distinguish between a variable that is a common prior cause of an explanatory and response variable and a variable that intervenes causally between the two.

Causal interpretation of observational data is always risky, especially—as here—when the data are cross-sectional (i.e., collected at one point in time) rather than longitudinal (where the data

are collected over time). Nevertheless, it is usually impossible, impractical, or immoral to collect experimental data in the social sciences, and longitudinal data are often hard to come by.⁷ Moreover, the essential difficulty of causal interpretation in nonexperimental investigations—due to potentially confounding variables that are left uncontrolled—applies to longitudinal as well as to cross-sectional observational data.⁸

The notion of “cause” and its relationship to statistical data analysis are notoriously difficult ideas. A relatively strict view requires an experimentally manipulable explanatory variable, at least one that is manipulable in principle.⁹ This is a particularly sticky point because, in social science, many explanatory variables are intrinsically not subject to direct manipulation, even in principle. Thus, for example, according to the strict view, gender cannot be considered a cause of income, even if it can be shown (perhaps after controlling for other determinants of income) that men and women systematically differ in their incomes, because an individual’s gender cannot be changed.¹⁰

I believe that treating nonmanipulable explanatory variables, such as gender, as potential causes is, at the very least, a useful shorthand. Men earn higher incomes than women because women are (by one account) concentrated into lower paying jobs, work fewer hours, are directly discriminated against, and so on (see, e.g., Ornstein, 1983). Explanations of this sort are perfectly reasonable and are subject to statistical examination; the sense of “cause” here may be weaker than the narrow one, but it is nevertheless useful.

It is overly restrictive to limit the notion of statistical causation to explanatory variables that are manipulated experimentally, to explanatory variables that are manipulable in principle, or to data that are collected over time.

1.3 Populations and Samples

Statistical inference is typically introduced in the context of random sampling from an identifiable population. There are good reasons for stressing this interpretation of inference—not the least of which are its relative concreteness and clarity—but the application of statistical inference is, at least arguably, much broader, and it is certainly broader in practice.

Take, for example, a prototypical experiment, in which subjects are assigned values of the explanatory variables at random: Inferences may properly be made to the hypothetical

⁷Experiments with human beings also frequently distort the processes that they purport to study: Although it might well be possible, for example, to recruit judges to an experimental study of judicial decision making, the artificiality of the situation could easily affect their simulated decisions. Even if the study entailed real judicial judgments, the mere act of observation might influence the judges’ decisions—they might become more careful, for example.

⁸We will take up the analysis of longitudinal data in Chapters 23 and 24 on mixed-effects models.

⁹For clear presentations of this point of view, see, for example, Holland (1986) and Berk (2004).

¹⁰This statement is, of course, arguable: There are historically many instances in which individuals have changed their gender, for example by disguise, not to mention surgery. Despite some fuzziness, however, I believe that the essential point—that some explanatory variables are not (normally) subject to manipulation—is valid. A more subtle point is that in certain circumstances, we could imagine experimentally manipulating the *apparent* gender of an individual, for example, on a job application.

population of random rearrangements of the subjects, even when these subjects are not sampled from some larger population. If, for example, we find a highly “statistically significant” difference between two experimental groups of subjects in a randomized experiment, then we can be sure, with practical certainty, that the difference was due to the experimental manipulation. The rub here is that our interest almost surely extends *beyond* this specific group of subjects to some larger—often ill-defined—population.

Even when subjects in an experimental or observational investigation are literally sampled at random from a real population, we usually wish to generalize beyond that population. There are exceptions—election polling comes immediately to mind—but our interest is seldom confined to the population that is directly sampled. This point is perhaps clearest when *no* sampling is involved—that is, when we have data on every individual in a real population.

Suppose, for example, that we examine data on population density and crime rates for all large U.S. cities and find only a weak association between the two variables. Suppose further that a standard test of statistical significance indicates that this association is so weak that it easily could have been the product of “chance.”¹¹ Is there any sense in which this information is interpretable? After all, we have before us data on the *entire* population of large U.S. cities at a particular historical juncture.

Because our interest inheres not directly—at least not exclusively—in these *specific* cities but in the complex social processes by which density and crime are determined, we can reasonably imagine a different outcome. Were we to replay history conceptually, we would not observe precisely the same crime rates and population density statistics, dependent as these are on a myriad of contingent and chancy events; indeed, if the ambit of our conceptual replay of history is sufficiently broad, then the identities of the cities themselves might change. (Imagine, for example, that Henry Hudson had not survived his trip to the New World or, if he survived it, that the capital of the United States had remained in New York. Less momentously, imagine that Fred Smith had not gotten drunk and killed a friend in a brawl, reducing the number of homicides in New York by one.) It is, in this context, reasonable to draw statistical inferences to the process that produced the currently existing populations of cities. Similar considerations arise in the analysis of historical statistics, for example, of time-series data.¹²

Much interesting data in the social sciences—and elsewhere—are collected haphazardly. The data constitute neither a sample drawn at random from a larger population nor a coherently defined population. Experimental randomization provides a basis for making statistical inferences to the population of rearrangements of a haphazardly selected group of subjects, but that is in itself cold comfort. For example, an educational experiment is conducted with students recruited from a school that is conveniently available. We are interested in drawing conclusions about the efficacy of teaching methods for students in general, however, not just for the students who participated in the study.

Haphazard data are also employed in many observational studies—for example, volunteers are recruited from among university students to study the association between eating disorders and overexercise. Once more, our interest transcends this specific group of volunteers.

To rule out haphazardly collected data would be a terrible waste; it is, instead, prudent to be careful and critical in the interpretation of the data. We should try, for example, to satisfy ourselves that our haphazard group does not differ in presumably important ways from the larger

¹¹Cf. the critical discussion of crime and population density in Freedman (1975).

¹²See Chapter 16 for a discussion of regression analysis with time-series data.

population of interest, or to control statistically for variables thought to be relevant to the phenomena under study.

Statistical inference can speak to the *internal stability* of patterns in haphazardly collected data and—most clearly in experimental data—to causation. *Generalization* from haphazardly collected data to a broader population, however, is inherently a matter of judgment.

Randomization and good sampling design are desirable in social research, but they are not prerequisites for drawing statistical inferences. Even when randomization or random sampling is employed, we typically want to generalize beyond the strict bounds of statistical inference.

Exercise

Exercise 1.1. Imagine that students in an introductory statistics course complete 20 assignments during two semesters. Each assignment is worth 1% of a student's final grade, and students get credit for assignments that are turned in on time and that show reasonable effort. The instructor of the course is interested in whether doing the homework contributes to learning, and (anticipating material to be taken up in Chapters 5 and 6), she observes a linear, moderately strong, and highly statistically significant relationship between the students' grades on the final exam in the course and the number of homework assignments that they completed. For concreteness, imagine that for each additional assignment completed, the students' grades on average were 1.5 higher (so that, e.g., students completing all of the assignments on average scored 30 points higher on the exam than those who completed none of the assignments).

- (a) Can this result be taken as evidence that completing homework assignments *causes* higher grades on the final exam? Why or why not?
- (b) Is it possible to design an experimental study that could provide more convincing evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how might such an experiment be designed?
- (c) Is it possible to marshal stronger observational evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how?

Summary

- With few exceptions, statistical data analysis describes the outcomes of real social processes and not the processes themselves. It is therefore important to attend to the descriptive accuracy of statistical models and to refrain from reifying them.
- Causal inferences are most certain—if not completely definitive—in randomized experiments, but observational data can also be reasonably marshaled as evidence of causation. Good experimental practice seeks to avoid confounding experimentally manipulated explanatory variables with other variables that can influence the response variable.

Sound analysis of observational data seeks to control statistically for potentially confounding variables.

- In analyzing observational data, it is important to distinguish between a variable that is a common prior cause of an explanatory and response variable and a variable that intervenes causally between the two.
- It is overly restrictive to limit the notion of statistical causation to explanatory variables that are manipulated experimentally, to explanatory variables that are manipulable in principle, or to data that are collected over time.
- Randomization and good sampling design are desirable in social research, but they are not prerequisites for drawing statistical inferences. Even when randomization or random sampling is employed, we typically want to generalize beyond the strict bounds of statistical inference.

Recommended Reading

- Chance and contingency are recurrent themes in Stephen Gould's fine essays on natural history; see, in particular, Gould (1989). I believe that these themes are relevant to the social sciences as well, and Gould's work has strongly influenced the presentation in Section 1.1.
- The legitimacy of causal inferences in nonexperimental research is and has been a hotly debated topic. Sir R. A. Fisher, for example, famously argued in the 1950s that there was no good evidence that smoking causes lung cancer, because the epidemiological evidence for the relationship between the two was, at that time, based on observational data (see, e.g., the review of Fisher's work on lung cancer and smoking in Stolley, 1991). Perhaps the most vocal recent critic of the use of observational data was David Freedman. See, for example, Freedman's (1987) critique of structural-equation modeling in the social sciences and the commentary that follows it.
- A great deal of recent work on causal inference in statistics has been motivated by "Rubin's causal model." For a summary and many references, see Rubin (2004). A very clear presentation of Rubin's model, followed by interesting commentary, appears in Holland (1986). Pearl (2009) develops a different account of causal inference from nonexperimental data using directed graphs. For an accessible, book-length treatment of these ideas, combining Rubin's "counterfactual" approach with Pearl's, see Morgan and Winship (2007). Also see Murnane and Willett (2011), who focus their discussion on research in education.
- Berk (2004) provides an extended, careful discussion, from a point of view different from mine, of many of the issues raised in this chapter.
- The place of sampling and randomization in statistical investigations has also been widely discussed and debated in the literature on research design. The classic presentation of the issues in Campbell and Stanley (1963) is still worth reading, as is Kish (1987). In statistics, these themes are reflected in the distinction between model-based and design-based inference (see, e.g., Koch & Gillings, 1983) and in the notion of superpopulation inference (see, e.g., Thompson, 1988).
- Achen (1982) argues eloquently for the descriptive interpretation of statistical models, illustrating his argument with effective examples.

PART I

Data Craft

2

What Is Regression Analysis?

As mentioned in Chapter 1, statistical data analysis is a *craft*, part art (in the sense of a skill developed through practice) and part science (in the sense of systematic, formal knowledge). Introductions to applied statistics typically convey some of the craft of data analysis but tend to focus on basic concepts and the logic of statistical inference. This and the next two chapters develop some of the elements of statistical data analysis:

- The current chapter introduces regression analysis in a general context, tracing the conditional distribution of a response variable as a function of one or several explanatory variables. There is also some discussion of practical methods for looking at regressions with a minimum of prespecified assumptions about the data.
- Chapter 3 describes graphical methods for looking at data, including methods for examining the distributions of individual variables, relationships between pairs of variables, and relationships among several variables.
- Chapter 4 takes up methods for transforming variables to make them better behaved—for example, to render the distribution of a variable more symmetric or to make the relationship between two variables more nearly linear.

Figure 2.1 is a *scatterplot* showing the relationship between hourly wages (in dollars) and formal education (in years) for a sample of 14,601 employed Canadians. The line in the plot shows the mean value of wages for each level of education and represents (in one sense) the *regression of wages on education*.¹ Although there are many observations in this scatterplot, few individuals in the sample have education below, say, 5 years, and so the mean wages at low levels of education cannot be precisely estimated from the sample, despite its large overall size. Discounting, therefore, variation in average wages at very low levels of education, it appears as if average wages are relatively flat until about 10 years of education, at which point they rise gradually and steadily with education.

Figure 2.1 raises several issues that we will take up in subsequent chapters.² Because of the large number of points in the plot and the discreteness of education (which is represented as number of years completed), the plot is difficult to examine. It is, however, reasonably clear that the distribution of wages at fixed levels of education is positively skewed. One such *conditional distribution* is shown in the histogram in Figure 2.2. The mean is a problematic measure of the center of a skewed distribution, and so basing the regression on the mean is not a good idea for such data. It is also clear that the relationship between hourly wages and education is

¹See Exercise 5.2 for the original statistical meaning of the term “*regression*.”

²See, in particular, Chapter 3 on examining data and Chapter 4 on transforming data.

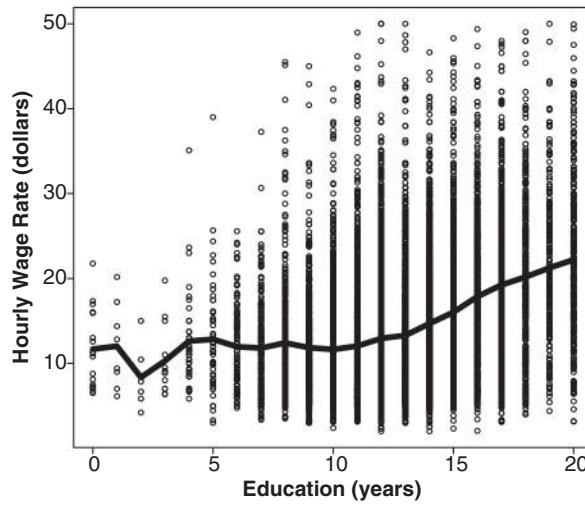


Figure 2.1 A scatterplot showing the relationship between hourly wages (in dollars) and education (in years) for a sample of 14,601 employed Canadians. The line connects the mean wages at the various levels of education. The data are drawn from the 1994 Survey of Labour and Income Dynamics (SLID).

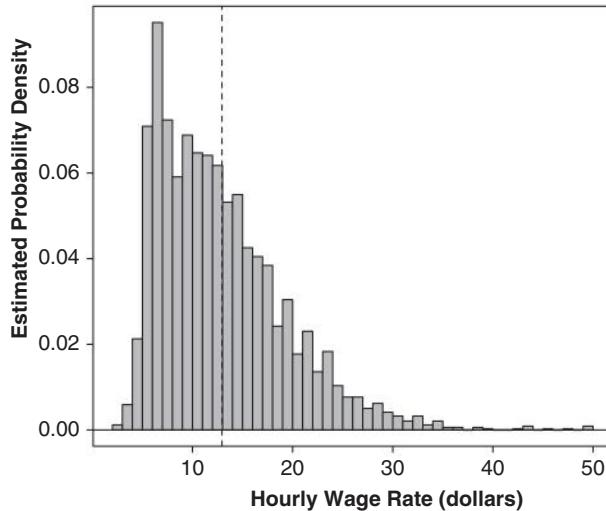


Figure 2.2 The conditional distribution of hourly wages for the 3,384 employed Canadians in the SLID, who had 12 years of education. The vertical axis is scaled as *density*, which means that the total area of the bars of the histogram is 1. Moreover, because each bar of the histogram has a width of 1, the height of the bar also (and coincidentally) represents the proportion of the sample in the corresponding interval of wage rates. The vertical broken line is at the mean wage rate for those with 12 years of education: \$12.94.

not linear—that is, not reasonably summarized by a straight line—and so the common reflex to summarize relationships between quantitative variables with lines is also not a good idea here.

Thinking more abstractly, *regression analysis*, broadly construed, traces the distribution of a response variable (denoted by Y)—or some characteristic of this distribution (such as its mean)—as a function of one or more explanatory variables (X_1, \dots, X_k):³

$$p(y|x_1, \dots, x_k) = f(x_1, \dots, x_k) \quad (2.1)$$

Here, $p(y|x_1, \dots, x_k)$ represents the probability (or, for continuous Y , the probability density) of observing the specific value y of the response variable, *conditional* on a set of specific values (x_1, \dots, x_k) of the explanatory variables, and $p(Y|x_1, \dots, x_k)$ is the *probability distribution* of Y (or the density function of Y) for these specific values of the X s.⁴ In the relationship between the response variable wages (Y) and the single explanatory variable education (X), for example, $p(Y|x)$ represents the population distribution of wages for all individuals who share the specific value x of education (e.g., 12 years). Figure 2.1 is therefore the sample analog of the population conditional distribution of Y .

The relationship of Y to the X s is of particular interest when we entertain the possibility that the X s affect Y or—more weakly—when we wish to use the X s to predict the value of Y . Primarily for convenience of exposition, I will initially use the term *regression analysis* to refer to those cases in which both Y and the X s are quantitative (as opposed to qualitative) variables.⁵ This chapter introduces basic concepts of regression analysis in a very general setting and explores some simple methods of regression analysis that make very weak assumptions about the structure of the data.

Regression analysis examines the relationship between a quantitative response variable, Y , and one or more explanatory variables, X_1, \dots, X_k . Regression analysis traces the conditional distribution of Y —or some aspect of this distribution, such as its mean—as a function of the X s.

2.1 Preliminaries

Figure 2.3 illustrates the regression of a continuous Y on a single, discrete X , which takes on several values, labeled x_1, x_2, \dots, x_5 . Alternatively, you can think of X as a continuous variable for which x_1, x_2, \dots, x_5 are specific representative values. As illustrated in the figure, the values of X need not be evenly spaced. For concreteness, imagine (as in Figure 2.1) that Y represents

³The response variable is often called the *dependent variable*, and the explanatory variables are often called *independent variables* or *predictors*.

⁴If the concept of (or notation for) a conditional distribution is unfamiliar, you should consult online Appendix D on probability and estimation. Please keep in mind more generally that background information is located in the appendices, available on the website for the book.

⁵Later in the book, we will have occasion to consider statistical models in which the explanatory variables (Chapters 7 and 8) and the response variable (Chapter 14) are qualitative/categorical variables. This material is centrally important because categorical variables are very common in the social sciences.

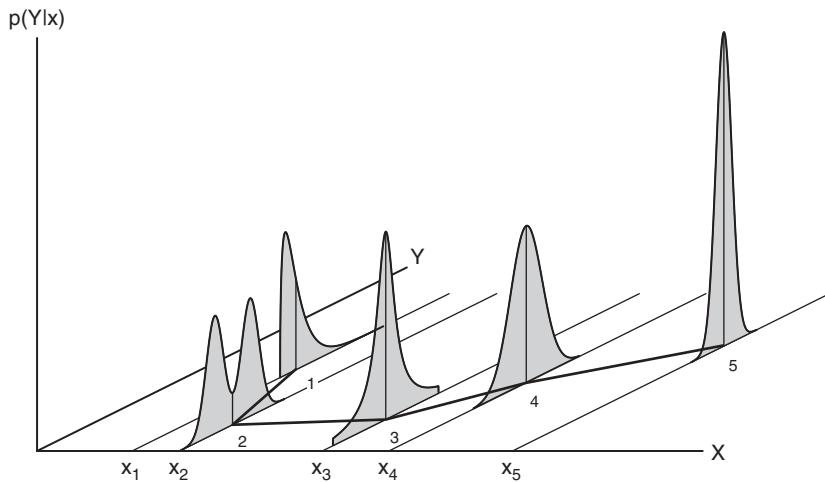


Figure 2.3 Population regression of Y on X . The conditional distribution of Y , $p(Y|x)$, is shown for each of a few values of X . The distribution of Y at $X = x_1$ is positively skewed; at $X = x_2$, it is bimodal; at $X = x_3$, it is heavy tailed; at $X = x_4$, it has greater spread than at $X = x_5$. The conditional means of Y given X , that is, μ_1, \dots, μ_5 , are not a linear function of X .

wages, that X represents years of formal education, and that the graph shows the conditional distribution $p(Y|x)$ of wages for some of the values of education.

Most discussions of regression analysis begin by assuming that the conditional distribution of the response variable, $p(Y|x_1, \dots, x_k)$, is a normal distribution; that the variance of Y conditional on the X 's is everywhere the same regardless of the specific values of x_1, \dots, x_k ; and that the expected value (the mean) of Y is a linear function of the X 's:

$$\mu \equiv E(Y|x_1, \dots, x_k) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k \quad (2.2)$$

This utopian situation is depicted for a single X in Figure 2.4. As we will see,⁶ the assumptions of normality, common variance, and linearity, along with independent random sampling, lead to linear least-squares estimation of the model in Equation 2.2. In this chapter, in contrast, we will pursue the notion of regression with as few assumptions as possible.

Figure 2.3 illustrates why we should not be too hasty to make the assumptions of normality, equal variance, and linearity:

- *Skewness.* If the conditional distribution of Y is skewed, as is $p(Y|x_1)$, then the mean will not be a good summary of its center. This is the case as well in Figure 2.1, where the (sample) conditional distributions of wages given education are all positively skewed.
- *Multiple modes.* If the conditional distribution of Y is multimodal, as is $p(Y|x_2)$, then it is intrinsically unreasonable to summarize its center by a single number.

⁶Chapters 6 and 9.

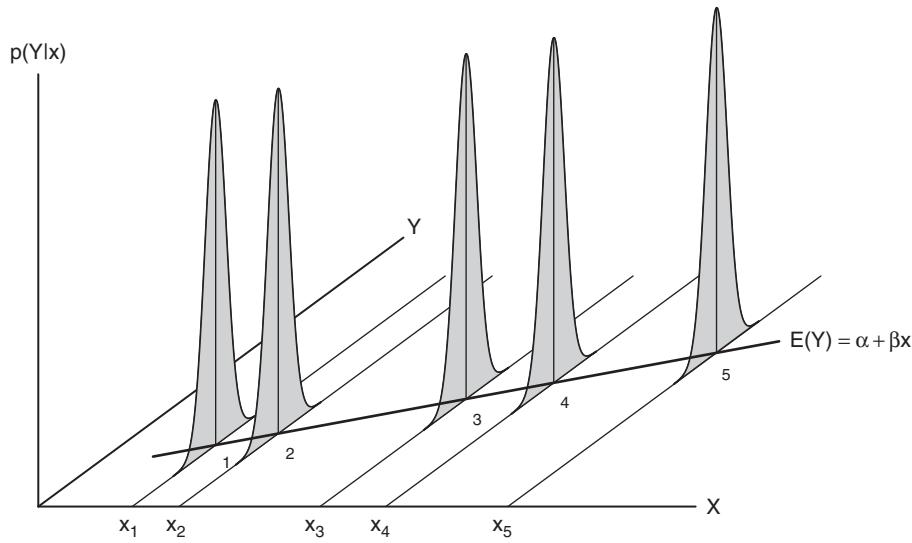


Figure 2.4 Common assumptions in regression analysis: The conditional distributions $p(Y|x)$ are all normal distributions with the same variance, and the conditional means of Y (here μ_1, \dots, μ_5) are all on a straight line.

- *Heavy tails.* If the conditional distribution of Y is substantially non-normal—for example, heavy-tailed, as is $p(Y|x_3)$ —then the sample mean will not be an efficient estimator of the center of the Y -distribution even when this distribution is symmetric.
- *Unequal spread.* If the conditional variance of Y changes with the values of the X s—compare, for example, $p(Y|x_4)$ and $p(Y|x_5)$ —then the efficiency of the usual least-squares estimates may be compromised; moreover, the nature of the dependence of the variance on the X s may itself be of interest.
- *Nonlinearity.* Although we are often in a position to suppose that the values of Y will increase or decrease with some X , there is almost never good reason to assume a priori that the relationship between Y and X is linear; this problem is compounded when there are several X s. In Figure 2.3, for example, the conditional means of Y , the μ_i , do not lie on a line in the X, Y plane (as they do in Figure 2.4).

This is not to say, of course, that linear regression analysis or, more generally, linear statistical models, are of little practical use. Much of this book is devoted to the exposition of linear models. It is, however, prudent to begin with an appreciation of the limitations of linear models because their effective use in data analysis frequently depends on adapting to these limitations: We may, for example, transform data to make the assumptions of normality, equal variance, and linearity more nearly correct.⁷

There are two additional advantages to approaching regression analysis from a general perspective: First, an appreciation of the practical difficulties of fitting the very general model in Equation 2.1 to data motivates the specification of more restrictive models, such as the usual

⁷See Chapters 4 and 12.

linear regression model. Second, modern methods of *nonparametric regression*, while not quite as general as the model in Equation 2.1, are emerging as practical alternatives to the more traditional linear models.

The balance of the present chapter is devoted to an initial foray into the territory of nonparametric regression. I will begin by taking a direct or “naïve” approach to the problem and then will extend this approach by local averaging. In the process, we will encounter for the first time a number of recurring themes in this book, including the direct examination of data by graphical displays, smoothing to clarify patterns in data, and the detection and treatment of unusual data.⁸

2.2 Naive Nonparametric Regression

Imagine once more that we are interested in the relationship between wages and education. We do not have data for the whole population, but we have a very large sample—say, of 1 million employed Canadians. We could easily display the conditional distribution of income for each of the values of education (0, 1, 2, . . . , 25) that occur in our data because (I assume) each value of education occurs many times. The example in the previous section, illustrated in Figures 2.1 and 2.2, approaches this situation for some of the more common levels of education.

Although wages is (for practical purposes) a continuous variable, the large quantity of data makes it practical to display its conditional distribution using a histogram with narrow bars (each, say, \$1.00 wide, as in Figure 2.2).⁹ If, as is often the case, our interest is in the average or typical value of wages conditional on education, we could—in light of the large size of our data set—estimate these conditional averages very accurately. The distribution of wages given education is likely positively skewed, so it would be better to use conditional medians rather than conditional means as typical values; nevertheless, we will, for simplicity, focus initially on the conditional means, $\bar{Y}|x$.¹⁰

Imagine now that X , along with Y , is a continuous variable. For example, X is the reported weight in kilograms for each of a sample of individuals, and Y is their measured weight, again in kilograms. We want to use reported weight to predict actual (i.e., measured) weight, and so we are interested in the mean value of Y as a function of X in the population of individuals from among whom the sample was randomly drawn:¹¹

$$\mu = E(Y|x) = f(x) \quad (2.3)$$

⁸More sophisticated methods for nonparametric regression are discussed in Chapter 18.

⁹We will explore other approaches to displaying distributions in the next chapter.

¹⁰Think of a graph like Figure 2.1 that shows the *population* conditional means, $\mu|x$ —the values that we now want to estimate from our sample.

¹¹This is an interesting—and unusual—problem in several respects: First, although it is more reasonable to suppose that actual weight “affects” the report than vice versa, our desire to use the report to predict actual weight (presumably because it is easier to elicit a verbal report than actually to weigh people) motivates treating measured weight as the response variable. Second, this is one of those comparatively rare instances in which a linear-regression equation is a natural specification, because if people are *unbiased* reporters of their weight, then we should have $\mu = x$ (i.e., expected reported weight equal to actual weight). Finally, if people are *accurate* as well as unbiased reporters of their weight, then the conditional variance of Y given x should be very small.

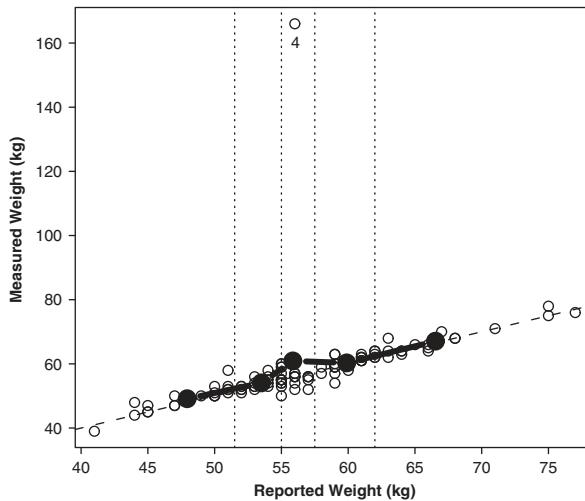


Figure 2.5 Naive nonparametric regression of measured weight on reported weight, each in kilograms. The range of reported weight has been dissected into five bins (separated by broken lines), each containing about 20 observations. The solid line connects the averages of measured weight and reported weight in the five bins, shown as filled circles. The dotted line around which the points cluster is $Y = X$. The fourth observation is an outlier. Because of the very different ranges of measured and reported weight (due to the outlier), the scales for the axes are different, and the line $Y = X$ is not at 45 degrees.

Even if the sample is large, replicated values of X will be rare because X is continuous.¹² In the absence of replicated X s, we cannot directly examine the conditional distribution of Y given X , and we cannot directly calculate conditional means. If we indeed have a large sample of individuals at our disposal, however, then we can dissect the range of X into many narrow intervals, or *bins*, of reported weight, each bin containing many observations; within each such bin, we can display the conditional distribution of measured weight and estimate the conditional mean of Y with great precision.

In very large samples, and when the explanatory variables are discrete, it is possible to estimate a regression by directly examining the conditional distribution of Y given the X s. When the explanatory variables are continuous, we can proceed similarly by dissecting the X s into a large number of narrow bins.

If, as is more typical, we have only a relatively small sample, then we have to make do with fewer bins, each containing relatively few observations. This situation is illustrated in Figure 2.5, using

¹²No numerical data are literally continuous, of course, because data are always recorded to some finite number of digits, and in the current example, people would be unlikely to report their weights in fractions of a kilogram. This is why tied values are possible. Individuals' measured weights (Y , in the example), however, could well be measured to greater precision. The philosophical issues surrounding continuity are subtle but essentially irrelevant to us: For practical purposes, a variable is continuous when it takes on many different values.

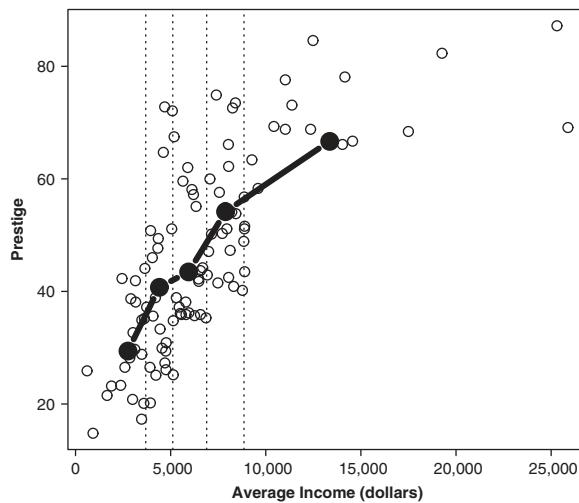


Figure 2.6 Naive nonparametric regression of occupational prestige on average income for 102 Canadian occupations in 1971. The range of income has been dissected into five bins, each containing about 20 observations. The line connects the average prestige and income scores in the five bins, shown as filled circles.

data on reported and measured weight for each of 101 Canadian women engaged in regular exercise.¹³ A partially contrasting example, using the prestige and income levels of 102 Canadian occupations in 1971, is shown in Figure 2.6.¹⁴

The X -axes in Figures 2.5 and 2.6 are carved into five unequal-width bins, each bin containing approximately 20 observations (the middle bin contains the extra observations). The nonparametric regression line displayed on each plot is calculated by connecting the points defined by the conditional response variable means \bar{Y} and the explanatory variable means \bar{X} in the five intervals.

Recalling our purpose, which is to estimate the model in Equation 2.3, there are two sources of error in this simple procedure of binning and averaging:

- *Sampling error (variance).* The conditional sample means \bar{Y} will, of course, change if we select a new sample (even if we could retain the same selection of x s). Sampling error is minimized by using a small number of relatively wide bins, each with many observations.

¹³These data were generously made available to me by Caroline Davis of York University, who used them as part of a larger study; see Davis (1990). The error in the data described below was located by Professor Davis. The 101 women were volunteers for the study, not a true sample from a larger population.

The observant reader will have noticed that there are apparently fewer than 101 points in Figure 2.5: Because both measured and reported weight are given to the nearest kilogram, many points are overplotted (i.e., lie on top of one another). We will learn to deal with overplotting in Chapter 3.

¹⁴The Canadian occupational prestige data are described in Fox and Suschnigg (1989). Although there are many more occupations in the Canadian census, these 102 do not constitute a random sample from the larger population of occupations. Justification for treating the 102 occupations as a sample implicitly rests on the claim that they are “typical” of the population, at least with respect to the relationship between prestige and income—a problematic, if arguable, claim.

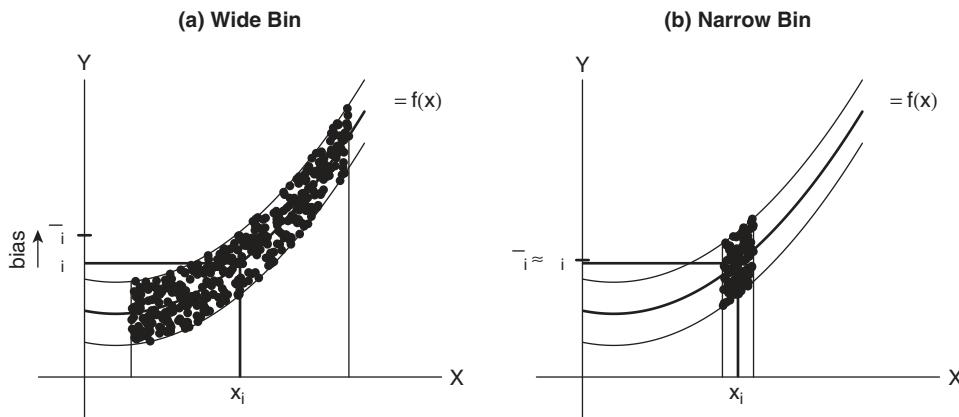


Figure 2.7 When the regression of Y on X is nonlinear in a bin centered at x_i , the average value of Y in the interval $\bar{\mu}_i$ can be different from the regression curve at the center of the interval $\mu_i = f(x_i)$. The bias, $\bar{\mu}_i - \mu_i$, will tend to be larger in a wide bin (a) than in a narrow bin (b).

- **Bias.** Let x_i denote the center of the i th bin (here, $i = 1, \dots, 5$), and suppose that the X -values are evenly spread in the bin. If the population regression curve $f(x)$ is nonlinear within the bin, then the average population value of Y in the bin, say $\bar{\mu}_i$, is usually different from the value of the regression curve at the center of the bin, $\mu_i = f(x_i)$. This situation is illustrated in Figure 2.7. Bias—that is, $\bar{\mu}_i - \mu_i$ —is therefore minimized by making the bins as numerous and as narrow as possible.

As is typically the case in statistical estimation, reducing bias and reducing sampling variance work at cross-purposes. Only if we select a very large sample can we have our cake and eat it, too—by constructing a very large number of narrow bins, each with many observations. This situation was, of course, our starting point.

The nonparametric regression lines in Figures 2.5 and 2.6 are also very crude. Although reported weights vary from about 40 to about 80 kg, we have evaluated the regression at only five points in this substantial range; likewise, income values for the 102 occupations vary from about \$600 to about \$26,000. Nevertheless, it is clear from Figure 2.5 that, except for one very discrepant data point (Observation 4),¹⁵ the data are very close to the line $Y = X$, and it is clear from Figure 2.6 that while prestige appears to increase with income, the increase is nonlinear, with prestige values leveling off at relatively high income.

The opportunity that a very large sample presents to reduce both bias and variance suggests that naive nonparametric regression is, under very broad conditions, a consistent estimator of the population regression curve.¹⁶ For, as the sample size gets larger (i.e., as $n \rightarrow \infty$), we can

¹⁵It seems difficult to comprehend how a 166-kg woman could have reported a weight of only 56 kg, but the solution to this mystery is simple: The woman's weight in kilograms and height in centimeters were accidentally switched when the data were entered into the computer.

¹⁶For example, we need to assume that the regression curve $\mu = f(X)$ is reasonably smooth and that the distribution of Y given x has finite variance (see Manski, 1991, for some details). We should also remember that the reasonableness of focusing on the mean μ depends on the symmetry—and unimodality—of the conditional distribution of Y .

ensure that the intervals grow successively narrower and yet have each contain more data (e.g., by employing \sqrt{n} intervals, each containing on average \sqrt{n} observations). In the limit—never, of course, attained—we have an infinite number of intervals, each of zero width and each containing an infinite number of observations. In this statistical nirvana, the naive nonparametric regression and the population regression curve coincide.

It may appear as if naive nonparametric regression—that is, binning and averaging—is a practical procedure in large data sets or when explanatory variables are discrete. Although this conclusion is essentially correct, it is instructive—and sobering—to consider what happens when there is more than one explanatory variable.

Suppose, for example, that we have three discrete explanatory variables, each with 10 values. There are, then, $10^3 = 1,000$ combinations of values of the three variables, and within each such combination, there is a conditional distribution of Y (i.e., $p(Y|x_1, x_2, x_3)$). Even if the X s are uniformly and independently distributed—implying equal expected numbers of observations for each of the 1,000 combinations—we would require a very large sample to calculate the conditional means of Y with sufficient precision. The situation is even worse when the X s are continuous, because dissecting the range of each X into as few as 10 bins might introduce nonnegligible bias into the estimation.

The problem of dividing the data into too many parts grows exponentially more serious as the number of X s increases. Statisticians, therefore, often refer to the intrinsic sparseness of multivariate data as the “curse of dimensionality.” Moreover, the imaginary calculation on which the consistency of naive nonparametric regression is based—in which the number of explanatory variables remains the same as the sample size grows—is itself unrealistic because we are apt, in large samples, to entertain more complex statistical models than in small samples.¹⁷

2.3 Local Averaging

Let us return to Figure 2.6, showing the naive nonparametric regression of occupational prestige on income. One problem with this procedure is that we have estimated the regression at only five points—a consequence of our desire to have relatively stable conditional averages, each based on a sufficiently large number of observations (here, 20). There is no intrinsic reason, however, why we should restrict ourselves to partitioning the data by X -values into nonoverlapping bins.

We can allow X to vary continuously across the range of observed values, calculating the average value of Y within a moving bin or *window* of fixed width centered at the current focal value x . Alternatively, we can employ a window of varying width, constructed to accommodate a fixed number of data values (say, m) that are the nearest X -neighbors to the focal x -value. The fraction of the data included in each window, $s \equiv m/n$, is called the *span* of the local-average regression. As a practical matter, of course, we cannot perform these calculations at the uncountably infinite number of points produced by allowing X to vary continuously, but using a computer, we can quickly calculate averages at a large number of focal values spanning the range of X . One attractive procedure, if the sample size n is not very large, is to evaluate the

¹⁷I am indebted to Robert Stine, of the University of Pennsylvania, for this insight.

local average of Y in a window around each of the X -values observed in the data: x_1, x_2, \dots, x_n .

In smaller samples, local averages of Y can be calculated in a neighborhood surrounding each x -value in the data. In larger samples, we can calculate local averages of Y at representative x -values spanning the range of X .

Figure 2.8 illustrates the process of local averaging for the Canadian occupational prestige data, employing $m = 20$ of the 102 observations for each local average, representing a span of $s = 20/102 \approx 0.2$. Figure 2.9 shows the result of applying local averaging to Davis's data on reported and measured weight, again using $m = 20$ observations for each local average. Three defects of the procedure are apparent from these examples:

1. The first few local averages are identical to one another, as are the last few, flattening the estimated regression line at extreme X -values.¹⁸ This artificial flattening at the edges of the data is called *boundary bias*.
2. The line connecting the averages tends to be rough because the average “jumps” up or down as observations enter and exit the moving window. (This roughness is more apparent in Figure 2.8, where the relationship between the two variables is weaker, than in Figure 2.9.)
3. Unusual data values, called *outliers*, unduly influence the average when they fall in the window (as is the case for Observation 4 in Figure 2.9). In regression analysis, an outlier is a value of Y that is very different from other response variable values associated with similar X s.

More adequate methods of nonparametric regression are the subject of Chapter 18. Nevertheless, because we will often want to smooth scatterplots in examining data, I anticipate that treatment by applying one such method, called *lowess*, to the two examples.¹⁹ Lowess is in many respects similar to the local-averaging smoother that I just described, except that instead of computing an average Y -value within the neighborhood of a focal x , the lowess smoother computes a fitted value based on a locally weighted least-squares line, giving more weight to observations in the neighborhood that are close to the focal x than to those relatively far away.²⁰ The lowess smoother also makes provision for discounting outliers.

¹⁸Imagine, for example, that the x -values are evenly spaced and that m is 19. Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ represent these x -values ordered from smallest to largest. Then, the first 19 observations—the Y -values associated with $x_{(1)}, x_{(2)}, \dots, x_{(19)}$ —would be used for the first 10 local averages, making these averages identical to one another. One solution is to employ “symmetric” neighborhoods around each $x_{(i)}$, with the same number of observations below and above the focal $x_{(i)}$, but this procedure implies using smaller and smaller spans as we approach the extreme values $x_{(1)}$ and $x_{(n)}$. For each extreme, for example, the symmetric neighborhood only includes the observation itself.

¹⁹Lowess is an acronym for *locally weighted scatterplot smoother* and is sometimes rendered as *loess*, for *local regression*.

²⁰Weighted least-squares regression is described (in a different context) in Section 12.2.2. The details of local regression are deferred to Chapter 18.

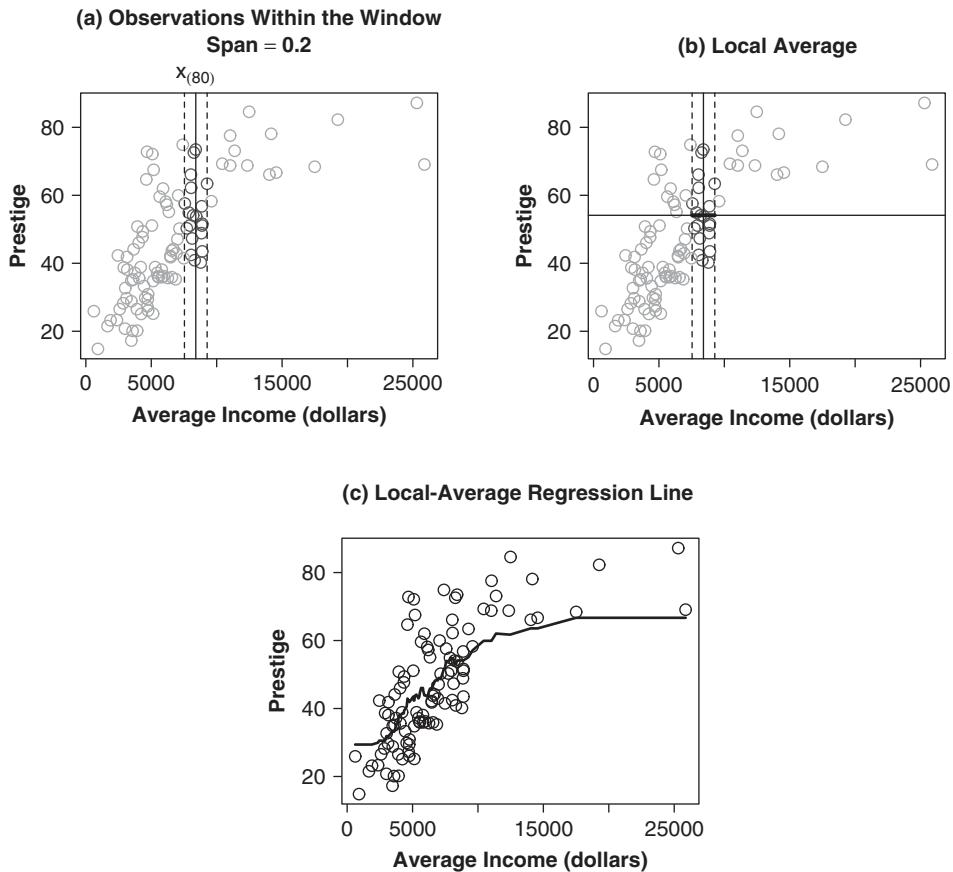


Figure 2.8 Nonparametric regression of occupational prestige on income, using local averages. Each average includes 20 of the 102 observations (i.e., a span of 0.2). Panel (a) shows the window encompassing the 20 nearest neighbors of $x_{(80)}$, the 80th ordered X -value. The mean of the Y -values for these 20 observations is represented by the horizontal line in panel (b). In panel (c), the local-average Y -values for all 102 observations are connected by a line. Note the roughness of the regression line and the flattening of the regression at the far left and right of the plot.

As with local averaging, we have to decide how many observations to include in each local regression; this is usually expressed by the span of the lowess smoother—the fraction of the data used to compute each fitted value. As was true of local averaging, larger spans reduce variance but may increase bias; smaller spans can reduce bias but increase variance. Put alternatively, larger spans produce smoother regressions.

One way to determine the span is by visual trial and error: Select the smallest span that yields a reasonably smooth result. Applying this procedure to the Canadian occupational prestige data and to Davis's data on reported and measured weight led me to the plots in Figures 2.10 and 2.11. Lowess produces smoother results than local averaging, shows no evidence of boundary bias, and ignores the outlier in the Davis data.

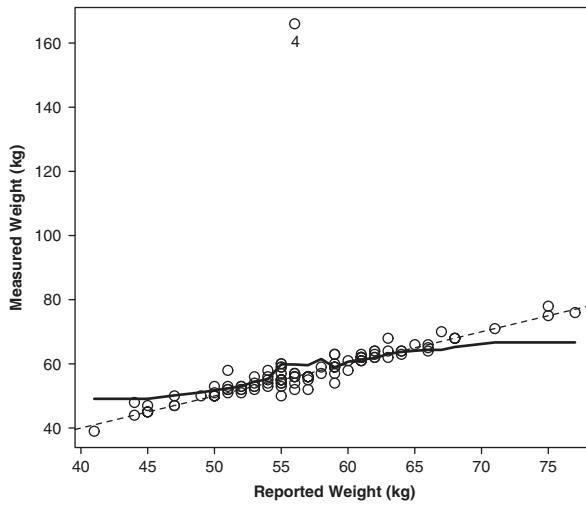


Figure 2.9 Nonparametric regression by local averaging for Davis's data on reported and measured weight. The local averages at each x -value are given by the line traced through the plot. Each local average is based on 20 of the 101 observations. Note the impact of the outlying observation on the averages that include it and the flattening of the regression at the lowest and highest reported weights. The broken line is the line of unbiased reporting, $Y = X$.

Lowess (locally weighted regression) produces smoother results than local averaging, reduces boundary bias, and can discount outliers. The degree of smoothness is controlled by the span of the lowess smoother: Larger spans yield smoother results.

Exercise

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 2.1.²¹ *Figure 2.7 illustrates how, when the relationship between Y and X is nonlinear in an interval, the average value of Y in the interval can be a biased estimate of $E(Y|x)$ at the center of the interval. Imagine that X -values are evenly distributed in an interval centered at x_i , and let $\mu_i \equiv E(Y|x_i)$.

- If the relationship between Y and X is linear in the interval, is the average value of Y a biased or an unbiased estimator of μ_i ?
- Are there any circumstances under which the average Y in the interval is an unbiased estimator of μ_i if the relationship between Y and X is nonlinear in the interval?
- What happens when the distribution of X -values in the interval is not uniform?

²¹Relatively difficult exercises are marked with an asterisk (*).

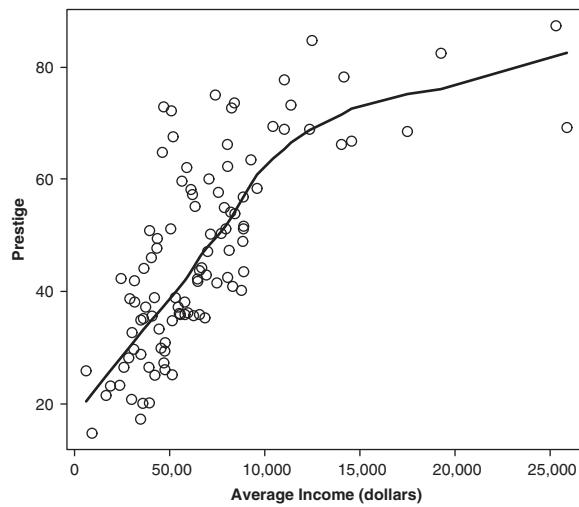


Figure 2.10 Lowess smooth of the relationship between occupational prestige and income. The span of the lowess smoother, 0.6, was determined by visual trial and error.

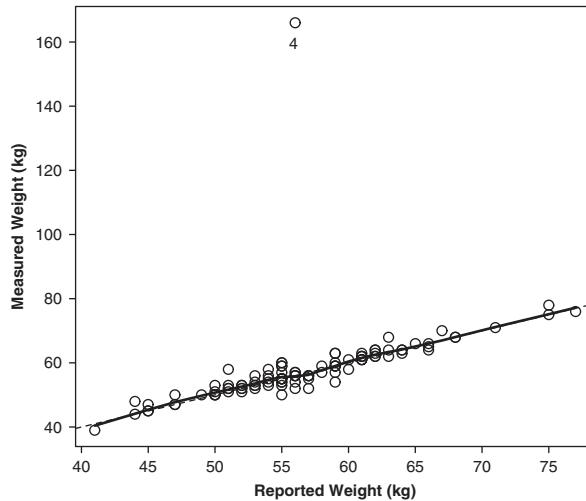


Figure 2.11 Lowess smooth of the relationship between reported and measured weight. The span of the smoother is 0.3. The broken line, almost entirely obscured by the lowess smooth, is the line of unbiased reporting of weight, $Y = X$.

Summary

- Regression analysis examines the relationship between a quantitative response variable, Y , and one or more explanatory variables, X_1, \dots, X_k . Regression analysis traces the

conditional distribution of Y —or some aspect of this distribution, such as its mean—as a function of the X s.

- In very large samples, and when the explanatory variables are discrete, it is possible to estimate a regression by directly examining the conditional distribution of Y given the X s. When the explanatory variables are continuous, we can proceed similarly in large samples by dissecting the X s into many narrow bins.
- In smaller samples, local averages of Y can be calculated in a neighborhood or window surrounding each x -value. There is a trade-off in local averaging between the bias and the variance of the estimates: Narrow windows reduce bias but, because they include fewer observations, increase variance.
- Lowess (locally weighted regression) produces smoother results than local averaging, reduces boundary bias, and can discount outliers. The degree of smoothness is controlled by the span of the lowess smoother: Larger spans yield smoother lowess regressions.

3

Examining Data

This chapter, on graphical methods for examining data, and the next, on transformations, represent a digression from the principal focus of the book. Nevertheless, the material here is important to us for two reasons: First, careful data analysis should begin with inspection of the data.¹ You will find in this chapter simple methods for graphing univariate, bivariate, and multivariate data. Second, the techniques for examining and transforming data that are discussed in Chapters 3 and 4 will find direct application to the analysis of data using linear models.² Feel free, of course, to pass lightly over topics that are familiar.

To motivate the material in the chapter, and to demonstrate its relevance to the study of linear models, consider the four scatterplots shown in Figure 3.1.³ The data for these plots, given in Table 3.1, were cleverly contrived by Anscombe (1973) to illustrate the central role of graphical methods in data analysis: Anticipating the material in Chapters 5 and 6, the least-squares regression line and all other common regression “outputs”—such as the correlation coefficient, standard deviation of the residuals, and standard errors of the regression coefficients—are identical in the four data sets.

It is clear, however, that each graph tells a different story about the data. Of course, the data are simply made up, so we have to allow our imagination some latitude:

- In Figure 3.1(a), the least-squares line is a reasonable descriptive summary of the tendency of Y to increase with X .
- In Figure 3.1(b), the linear regression fails to capture the clearly curvilinear relationship between the two variables; we would do much better to fit a quadratic function here,⁴ that is, $Y = a + bX + cX^2$.
- In Figure 3.1(c), there is a perfect linear relationship between Y and X for all but one outlying data point. The least-squares line is pulled toward the outlier, distorting the relationship between the two variables for the rest of the data. Perhaps the outlier represents an error in data entry or an observation that differs in some fundamental respect from the others. When we encounter an outlier in real data, we should look for an explanation.⁵

¹An eminent statistician who has engaged in frequent consulting (and who will remain nameless for fear of embarrassing him) told me that his clients routinely extract only about 30% of the information in their data relevant to their research. He attributed this inefficiency largely to failure to examine the data carefully at an early stage in statistical analysis. Many problems can be detected and dealt with effectively prior to engaging in statistical modeling of the data.

²See, for example, the treatments of graphical regression “diagnostics” and transformations in Chapters 11 and 12.

³See Section 3.2 for a general discussion of scatterplots.

⁴Quadratic and other polynomial regression models are discussed in Section 17.1.

⁵Outlier detection in linear models is taken up in Chapter 11.

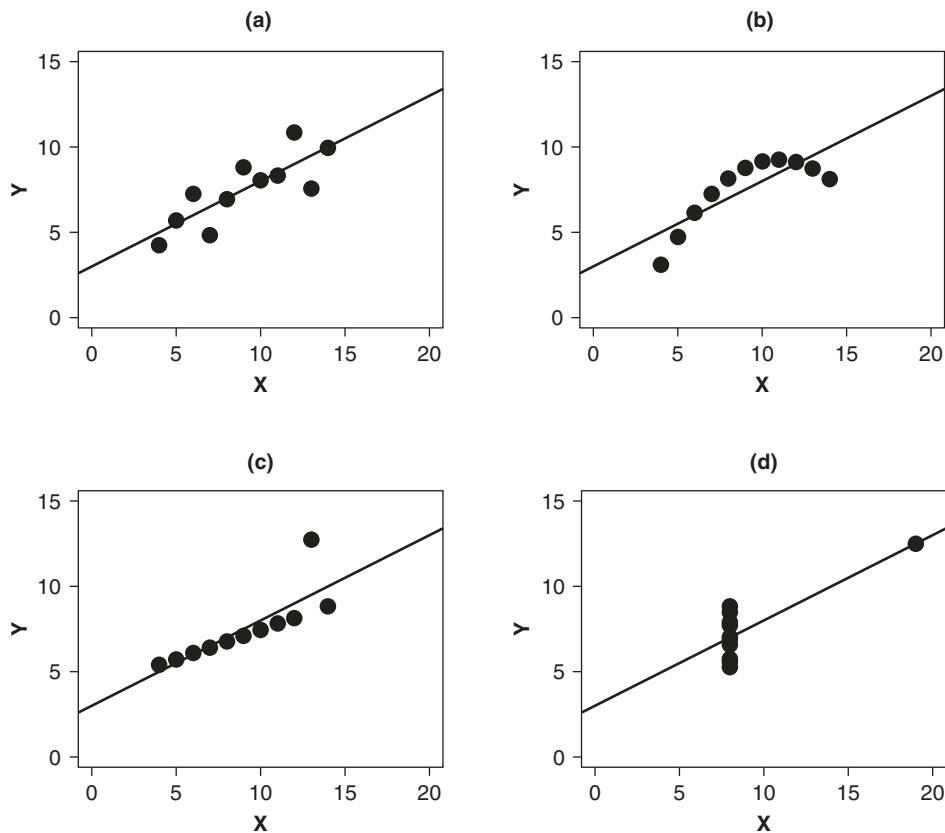


Figure 3.1 Four data sets, due to Anscombe (1973), with identical linear least-squares regressions. In (a), the linear regression is an accurate summary; in (b), the linear regression distorts the curvilinear relationship between Y and X ; in (c), the linear regression is drawn toward an outlier; in (d), the linear regression “chases” the influential observation at the right. The least-squares line is shown on each plot.

SOURCE: Adapted from Figures 1 & 2, p. 19 and Figures 3 & 4 p. 20 in F. J. Anscombe, “Graphs in Statistical Analysis” pp. 17–21, Vol. 27, No. 1, Feb., 1973.

- Finally, in Figure 3.1(d), the values of X are invariant (all are equal to 8), with the exception of one point (which has an X -value of 19); the least-squares line would be undefined but for this point—the line necessarily goes through the mean of the 10 Y s that share the value $X = 8$ and through the point for which $X = 19$. Furthermore, if this point were moved, then the regression line would chase it. We are usually uncomfortable having the result of a data analysis depend so centrally on a single influential observation.⁶

The essential point to be derived from Anscombe’s “quartet” (so dubbed by Tufte, 1983) is that it is frequently helpful to examine data graphically. Important characteristics of data are often disguised by numerical summaries and—worse—the summaries can be fundamentally

⁶Influential data are discussed in Chapter 11.

Table 3.1 Four Contrived Regression Data Sets From Anscombe (1973)

$X_{a, b, c}$	Y_a	Y_b	Y_c	X_d	Y_d
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

misleading. Moreover, directly examining the numerical data is often uninformative: Only in the fourth of Anscombe's data sets is a problem immediately apparent upon inspection of the numbers.

Statistical graphs are central to effective data analysis, both in the early stages of an investigation and in statistical modeling.

3.1 Univariate Displays

3.1.1 Histograms

Figure 3.2 shows a *histogram* for the distribution of infant mortality among 193 countries, as reported in 1998 by the United Nations. The infant mortality rate is expressed as number of deaths of children aged less than 1 year per 1,000 live births. I assume that the histogram is a familiar graphical display, so I will offer only a brief description: To construct a histogram for infant mortality, dissect the range of the variable into equal-width intervals (called “bins”), count the number of observations falling in each bin, and display the frequency counts in a bar graph. The histogram in Figure 3.2 uses bins of width 10, starting at 0 (i.e., 0 to 10, 10 to 20, etc.).

Figure 3.3 shows an alternative form of histogram, called a *stem-and-leaf display*. The stem-and-leaf plot, introduced by John Tukey (1972, 1977), ingeniously employs the numerical data to form the bars of the histogram. As Tukey suggests, it is simple to construct a stem-and-leaf display by hand to “scratch down” a small data set.

You may be familiar with the stem-and-leaf display. Here is a terse explanation:

- Each data value is broken between two adjacent digits into a “stem” and a “leaf”: In Figure 3.3, the break takes place between the tens and units digits. For example, the infant mortality rate in Albania was 32, which translates into the stem 3 and leaf 2.

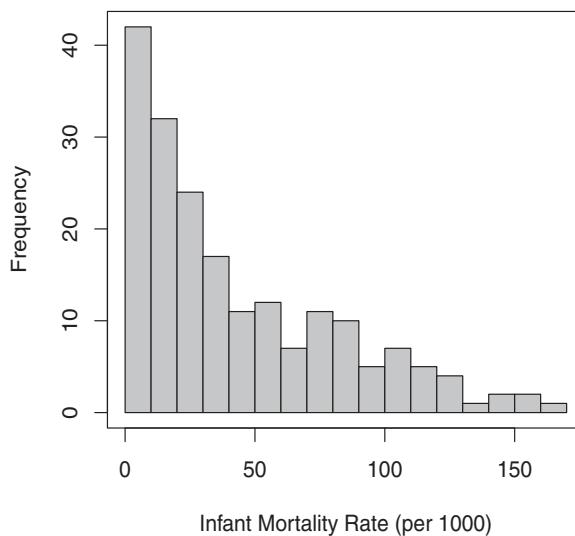


Figure 3.2 Histogram of infant mortality rates for 193 nations.

SOURCE: United Nations (1998).

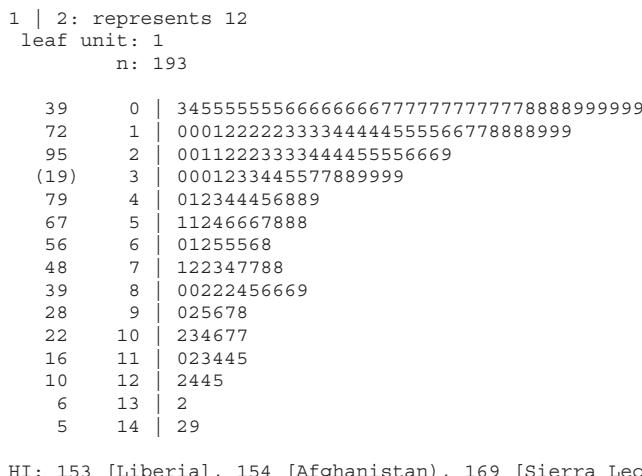


Figure 3.3 Stem-and-leaf display for infant mortality.

- Stems (here, 0, 1, . . . , 14) are constructed to cover the data, implicitly defining a system of bins, each of width 10. Each leaf is placed to the right of its stem, and the leaves on each stem are then sorted into ascending order. We can produce a finer system of bins by dividing each stem into two parts (taking, respectively, leaves 0–4 and 5–9), or five parts (0–1, 2–3, 4–5, 6–7, 8–9); for the infant mortality data, two-part stems would

correspond to bins of width 5 and five-part stems to bins of width 2. We could employ still finer bins by dividing stems from leaves between the ones and tenths digits, but for infant mortality, that would produce a display with almost as many bins as observations. Similarly, a coarser division between the hundreds and tens digits would yield only two stems—0 and 1 representing hundreds (each of which could be divided into two or five parts, to get bins of width 50 or 20, respectively).

- Unusually large values—*outliers*—are collected on a special “HI” stem and displayed individually. Here, there are three countries with unusually large infant mortality rates. Were there countries with unusually small infant mortality rates, then these would be collected and displayed individually on a “LO” stem.⁷
- The column of *depths* to the left of the stems counts in toward the median from both ends of the distribution. The median is the observation at depth $(n + 1)/2$, where (as usual) n is the number of observations. For the infant mortality data, the median is at depth $(193 + 1)/2 = 97$. In Figure 3.3, there are 39 observations at stem 0, 72 at and below stem 1, and so on; there are five observations (including the outliers) at and above stem 14, six at and above stem 13, and so forth. The count at the stem containing the median is shown in parentheses—here, 19 at stem 3. Note that $95 + 19 + 79 = 193$.

In constructing histograms (including stem-and-leaf displays), we want enough bins to preserve some detail but not so many that the display is too rough and dominated by sampling variation. Let n^* represent the number of nonoutlying observations. Then, for $n^* \leq 100$, it usually works well to use no more than about $2\sqrt{n^*}$ bins; likewise, for $n^* > 100$, we can use a maximum of about $10 \times \log_{10} n^*$ bins. Of course, in constructing a histogram, we also want bins that start and end at “nice” numbers (e.g., 10 to 20 rather than 9.5843 to 21.0457); in a stem-and-leaf display, we are limited to bins that correspond to breaks between digits of the data values. Computer programs that construct histograms incorporate rules such as these.⁸

For the distribution of infant mortality, $n^* = 193 - 3 = 190$, so we should aim for no more than $10 \times \log_{10}(190) \approx 23$ bins. The stem-and-leaf display in Figure 3.3 uses 15 stems (plus the “HI” stem).

Histograms, including stem-and-leaf displays, are very useful graphs, but they suffer from several problems:

- The visual impression of the data conveyed by a histogram can depend on the arbitrary origin of the bin system—that is, the lower boundary of the first bin. Consider, for example, the two histograms in Figure 3.4, showing the distribution of prestige for the 102 occupations in the Canadian occupational prestige data set.⁹ Both histograms use bins of width 10, but the bins in Figure 3.4(a) start at 0, while those in Figure 3.4(b) start at 10.

⁷The rule for identifying outliers is explained in Section 3.1.4 on boxplots.

⁸More sophisticated rules for the number of bins take into account information beyond n . For example, Freedman and Diaconis (1981) suggest

$$\text{number of bins} \approx \left\lceil \frac{n^{1/3} (x_{(n)} - x_{(1)})}{2(Q_3 - Q_1)} \right\rceil$$

where $x_{(n)} - x_{(1)}$ is the range of the data, $Q_3 - Q_1$ is the interquartile range, and the “ceiling” brackets indicate rounding up to the next integer.

⁹This data set was introduced in Chapter 2.

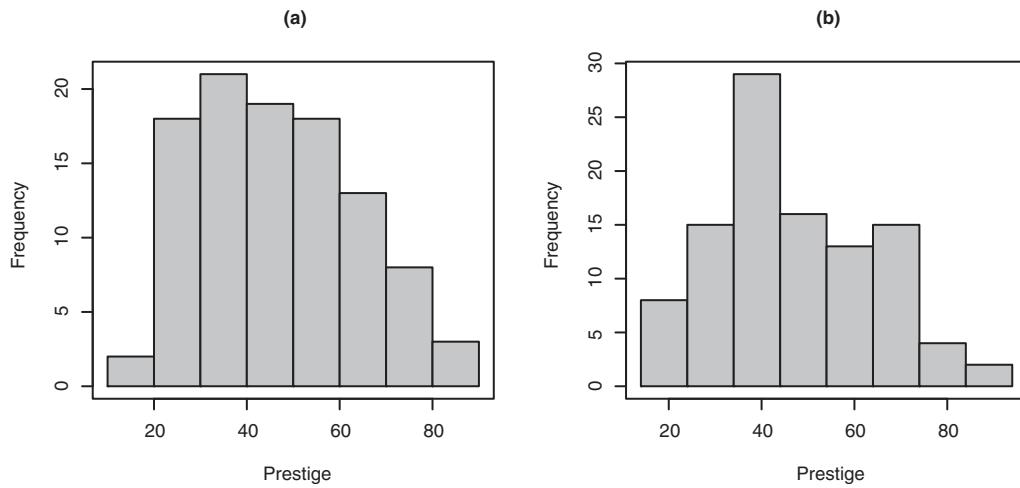


Figure 3.4 Alternative histograms for the prestige of 102 Canadian occupations: (a) with bins of width 10 starting at 0 and (b) with bins of width 10 starting at 15.

- Because the bin system dissects the range of the variable into class intervals, the histogram is discontinuous (i.e., rough) even if, as in the case of infant mortality, the variable is continuous.¹⁰
- The form of the histogram depends on the arbitrary width of the bins.
- Moreover, if we use bins that are narrow enough to capture detail where data are plentiful—usually near the center of the distribution—then they may be too narrow to avoid “noise” where data are sparse—usually in the tails of the distribution.

3.1.2 Nonparametric Density Estimation

Nonparametric density estimation addresses the deficiencies of traditional histograms by averaging and smoothing. As the term implies, *density estimation* can be construed formally as an attempt to estimate the probability density function of a variable based on a sample, but it can also be thought of informally as a descriptive technique for smoothing histograms.

In fact, the histogram—suitably rescaled—is a simple density estimator.¹¹ Imagine that the origin of the bin system is at x_0 and that each of the m bins has width $2h$; the end points of the bins are then at $x_0, x_0 + 2h, x_0 + 4h, \dots, x_0 + 2mh$. An observation X_i falls in the j th bin if (by convention)

¹⁰That is, infant mortality rates are continuous for practical purposes in that they can take on many different values. Actually, infant mortality rates are ratios of integers and hence are rational numbers, and the rates in the UN data set are rounded to the nearest whole number.

¹¹Rescaling is required because a density function encloses a total area of 1. Histograms are typically scaled so that the height of each bar represents frequency (or percent), and thus the heights of the bars sum to the sample size n (or 100). If each bar spans a bin of width $2h$ (anticipating the notation below), then the total area enclosed by the bars is $n \times 2h$. Dividing the height of each bar by $2nh$ therefore produces the requisite density rescaling.

$$x_0 + 2(j-1)h \leq X_i < x_0 + 2jh$$

The histogram estimator of the density at any x -value located in the j th bin is based on the number of observations that fall in that bin:

$$\hat{p}(x) = \frac{\#_{i=1}^n [x_0 + 2(j-1)h \leq X_i < x_0 + 2jh]}{2nh}$$

where $\#$ is the counting operator.

We can dispense with the arbitrary origin x_0 of the bin system by counting locally within a continuously moving window of half-width h centered at x :

$$\hat{p}(x) = \frac{\#_{i=1}^n (x - h \leq X_i < x + h)}{2nh}$$

In practice, of course, we would use a computer program to evaluate $\hat{p}(x)$ at a large number of x -values covering the range of X . This “naive density estimator” (so named by Silverman, 1986) is equivalent to locally weighted averaging, using a rectangular weight function:

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^n W\left(\frac{x - X_i}{h}\right) \quad (3.1)$$

where

$$W(z) = \begin{cases} \frac{1}{2} & \text{for } |z| < 1 \\ 0 & \text{otherwise} \end{cases}$$

a formulation that will be useful below when we consider alternative weight functions to smooth the density. Here z is a “stand-in” for the argument to the $W(\cdot)$ weight function—that is, $z = (x - X_i)/h$. The naive estimator is like a histogram that uses bins of width $2h$ but has no fixed origin and is similar in spirit to the local-averaging nonparametric-regression estimator introduced in Chapter 2.

An illustration, using the UN infant mortality data, appears in Figure 3.5 and reveals the principal problem with the naive estimator: Because the estimated density jumps up and down as observations enter and leave the window, the naive density estimator is intrinsically rough.

The rectangular weight function $W(z)$ in Equation 3.1 is defined to enclose an area of $2 \times \frac{1}{2} = 1$, producing a density estimate that (as required) also encloses an area of 1. Any function that has this property—probability density functions are obvious choices—may be used as a weight function, called a *kernel*. Choosing a kernel that is smooth, symmetric, and unimodal smooths out the rough edges of the naive density estimator. This is the essential insight of *kernel density estimation*.

The general kernel density estimator, then, is given by

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

There are many reasonable choices of the kernel function $K(z)$, including the familiar standard normal density function, $\phi(z)$, which is what I will use here. While the naive density estimator in effect sums suitably scaled rectangles centered at the observations, the more general

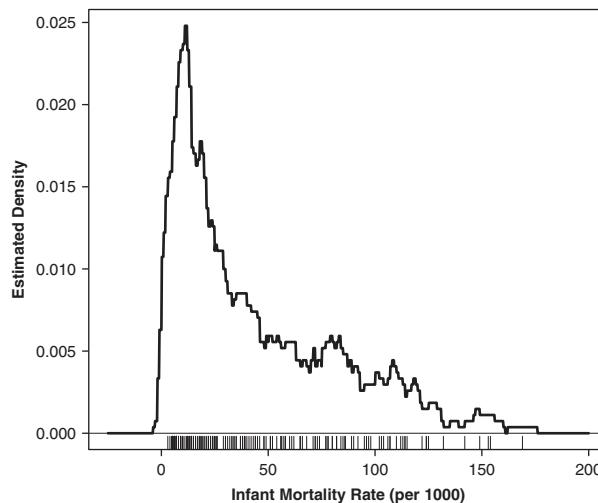


Figure 3.5 Naive density estimator for infant mortality, using a window half-width of $h = 7$. Note the roughness of the estimator. A *rug-plot* (or “one-dimensional scatterplot”) appears at the bottom of the graph, showing the location of the data values.

kernel estimator sums smooth lumps. An example is shown in Figure 3.6, in which the kernel density estimator is given by the broken line.¹²

Selecting the window width for the kernel estimator is primarily a matter of trial and error—we want a value small enough to reveal detail but large enough to suppress random noise. We can, however, look to statistical theory for rough guidance:¹³ If the underlying density that we are trying to estimate is normal with standard deviation σ , then (for the normal kernel) estimation is most efficient with the window half-width

$$h = 0.9\sigma n^{-1/5} \quad (3.2)$$

As is intuitively reasonable, the optimal window grows gradually narrower as the sample size is increased, permitting finer detail in large samples than in small ones.¹⁴

Although we might, by reflex, be tempted to replace the unknown σ in Equation 3.2 with the sample standard deviation S , it is prudent to be more cautious, for if the underlying density is sufficiently non-normal, then the sample standard deviation may be seriously inflated. A common compromise is to use an “adaptive” estimator of spread:

¹²Notice that there is nonzero estimated density in Figure 3.6 below an infant mortality rate of 0. Of course, this does not make sense, and although I will not pursue it here, it is possible to constrain the lower and upper limits of the kernel estimator.

¹³See, for example, Silverman (1986, chap. 3) for a detailed discussion of these issues.

¹⁴If we really knew that the density were normal, then it would be even more efficient to estimate it parametrically by substituting the sample mean \bar{X} and standard deviation S for μ and σ in the formula for the normal density, $p(x) = (2\pi\sigma^2)^{-1/2}\exp[-(x - \mu)^2/2\sigma^2]$.

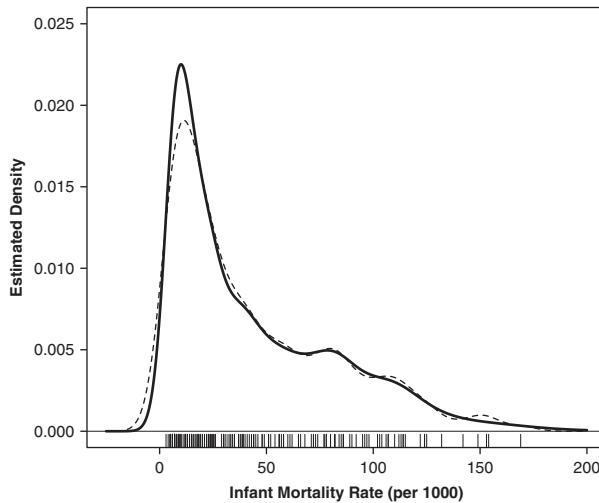


Figure 3.6 Kernel (broken line) and adaptive-kernel (solid line) density estimates for the distribution of infant mortality, using a normal kernel and a window half-width of $h = 7$. Note the relative “lumpiness” of the kernel estimator at the right, where data are sparse.

$$A = \min\left(S, \frac{\text{interquartile range}}{1.349}\right) \quad (3.3)$$

The factor 1.349 is the interquartile range of the standard normal distribution, making (interquartile range)/1.349 a robust estimator of σ in the normal setting.

One further caveat: If the underlying density is substantially non-normal—in particular, if it is skewed or multimodal—then basing h on the adaptive estimator A generally produces a window that is too wide. A good procedure, then, is to start with

$$h = 0.9An^{-1/5}$$

and to adjust this value downwards until the resulting density plot becomes too rough. This is the procedure that was used to find the window width in Figure 3.6, where $S = 38.55$ and $(\text{interquartile range})/1.349 = (68 - 13)/1.349 = 40.77$. Here, the “optimal” window width is $h = 0.9 \times 38.55 \times 197^{-1/5} = 12.061$.

The kernel density estimator usually does a pretty good job, but the window half-width h remains a compromise: We would prefer a narrower window where data are plentiful (to preserve detail) and a wider one where data are sparse (to suppress noise). Because *plentiful* and *sparse* refer implicitly to the underlying density that we are trying to estimate, it is natural to begin with an initial estimate of the density and to adjust the window half-width on the basis of the initial estimate.¹⁵ The result is the *adaptive-kernel estimator* (not to be confused with the adaptive estimator of spread in Equation 3.3).

¹⁵An alternative is to use a *nearest-neighbor* approach, as in the nonparametric-regression methods discussed in Chapter 2.

1. Calculate an initial density estimate, $\tilde{p}(x)$ —for example, by the kernel method.
2. Using the initial estimate, compute local window factors by evaluating the estimated density at the observations:

$$f_i = \left[\frac{\tilde{p}(X_i)}{\tilde{p}} \right]^{-1/2}$$

In this formula, \tilde{p} is the geometric mean of the initial density estimates at the observations—that is,

$$\tilde{p} = \left[\prod_{i=1}^n \tilde{p}(X_i) \right]^{1/n}$$

(where the operator \prod indicates continued multiplication). As a consequence of this definition, the f_i s have a product of 1, and hence a geometric mean of 1, ensuring that the area under the density estimate remains equal to 1.

3. Calculate the adaptive-kernel density estimator using the local window factors to adjust the width of the kernels centered at the observations:

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{f_i} K\left(\frac{x - X_i}{f_i h}\right)$$

Applying the adaptive-kernel estimator to the infant mortality distribution produces the solid line in Figure 3.6: For this distribution, the kernel and adaptive-kernel estimates are very similar, although the adaptive kernel more sharply defines the principal mode of the distribution near 20 and produces a smoother long right tail.

3.1.3 Quantile-Comparison Plots

Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution—something that is more commonly of interest for derived quantities such as test statistics or residuals than for observed variables. A strength of the display is that it does not require the use of arbitrary bins or windows.

Let $P(x)$ represent the theoretical *cumulative distribution function* (CDF) with which we want to compare the data; that is, $P(x) = \Pr(X \leq x)$. A simple (but not terribly useful) procedure is to graph the *empirical cumulative distribution function* (ECDF) for the observed data, which is simply the proportion of data below each value of x , as x moves continuously from left to right:

$$\hat{P}(x) = \frac{\#_{i=1}^n (X_i \leq x)}{n}$$

As illustrated in Figure 3.7, however, the ECDF is a “stair-step” function (where each step occurs at an observation and is of height $1/n$), while the CDF is typically smooth, making the comparison difficult.

The quantile-comparison plot avoids this problem by never constructing the ECDF explicitly:

1. Order the data values from smallest to largest, $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. The $X_{(i)}$ are called the *order statistics* of the sample.

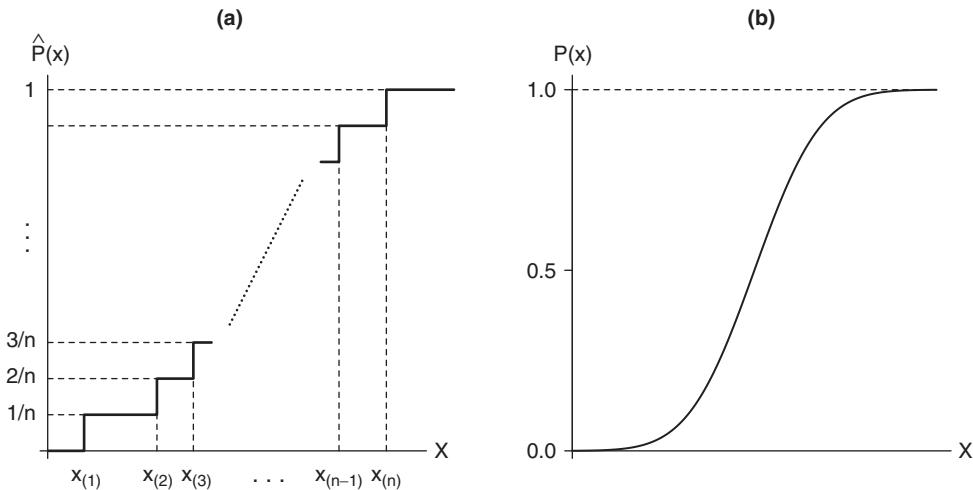


Figure 3.7 A “typical” empirical cumulative distribution function (ECDF) is shown in (a), a “typical” theoretical cumulative distribution function (CDF) in (b). $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ represent the data values ordered from smallest to largest. Note that the ordered data values are not, in general, equally spaced.

2. By convention, the cumulative proportion of the data “below” $X_{(i)}$ is given by¹⁶

$$P_i = \frac{i - \frac{1}{2}}{n}$$

3. Use the inverse of the CDF (i.e., the *quantile function*) to find the value z_i corresponding to the cumulative probability P_i ; that is,¹⁷

$$z_i = P^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$$

4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates. If X is sampled from the distribution P , then $X_{(i)} \approx z_i$. That is, the plot should be approximately linear, with an intercept of 0 and slope of 1. This relationship is only approximate because of sampling error (see Step 6). If the distributions are identical except for location, then the plot is approximately linear with a nonzero intercept, $X_{(i)} \approx \mu + z_i$; if the distributions are identical except for scale, then the plot is approximately linear with a slope different from 1, $X_{(i)} \approx \sigma z_i$; finally, if the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$.

¹⁶This definition avoids cumulative proportions of 0 or 1, which would be an embarrassment in Step 3 for distributions, like the normal, that never quite reach cumulative probabilities of 0 or 1. In effect, we count half of each observation below its exact value and half above. Another common convention is to use $P_i = (i - \frac{1}{3}) / (n + \frac{1}{3})$.

¹⁷This operation assumes that the CDF has an inverse—that is, that P is a strictly increasing function (one that never quite levels off). The common continuous probability distributions in statistics—for example, the normal, t -, F -, and χ^2 distributions—all have this property. These and other distributions are reviewed in online Appendix D on probability and estimation.

5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity. The line can be plotted by eye, attending to the central part of the data, or we can draw a line connecting the quartiles. For a normal quantile-comparison plot—comparing the distribution of the data with the standard normal distribution—we can alternatively use the median as a robust estimator of μ and the interquartile range/1.349 as a robust estimator of σ . (The more conventional estimates $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = S$ will not work well when the data are substantially non-normal.)
6. We expect some departure from linearity because of sampling variation; it therefore assists interpretation to display the expected degree of sampling error in the plot. The standard error of the order statistic $X_{(i)}$ is

$$\text{SE}(X_{(i)}) = \frac{\hat{\sigma}}{p(z_i)} \sqrt{\frac{P_i(1 - P_i)}{n}} \quad (3.4)$$

where $p(z)$ is the probability density function corresponding to the CDF $P(z)$. The values along the fitted line are given by $\hat{X}_{(i)} = \hat{\mu} + \hat{\sigma}z_i$. An approximate 95% confidence “envelope” around the fitted line is, therefore,¹⁸

$$\hat{X}_{(i)} \pm 2 \times \text{SE}(X_{(i)})$$

Figures 3.8 to 3.11 display normal quantile-comparison plots for several illustrative distributions:

- Figure 3.8 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$. The plotted points are reasonably linear and stay within the rough 95% confidence envelope.
- Figure 3.9 plots a sample of $n = 100$ observations from the positively skewed chi-square distribution with 2 degrees of freedom. The positive skew of the data is reflected in points that lie *above* the comparison line in both tails of the distribution. (In contrast, the tails of negatively skewed data would lie *below* the comparison line.)
- Figure 3.10 plots a sample of $n = 100$ observations from the heavy-tailed t -distribution with 2 degrees of freedom. In this case, values in the upper tail lie above the corresponding normal quantiles, and values in the lower tail below the corresponding normal quantiles.
- Figure 3.11 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent. The possibly bimodal character of the data, however, is not easily discerned in this display.

Quantile-comparison plots highlight the tails of distributions. This is important, because the behavior of the tails is often problematic for standard estimation methods like least squares, but it is useful to supplement quantile-comparison plots with other displays—such as histograms

¹⁸By the method of construction, the 95% confidence level applies (pointwise) to each $\hat{X}_{(i)}$, not to the whole envelope: There is a greater probability that *at least one* point strays outside the envelope even if the data are sampled from the comparison distribution. Determining a *simultaneous* 95% confidence envelope would be a formidable task, because the order statistics are not independent.

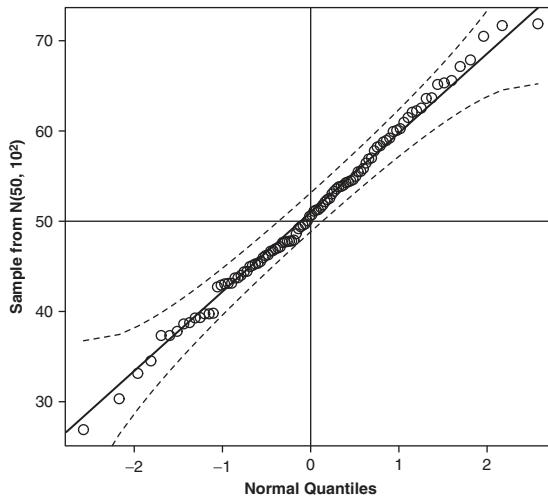


Figure 3.8 Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, and the broken lines give a pointwise 95% confidence interval around the fit.

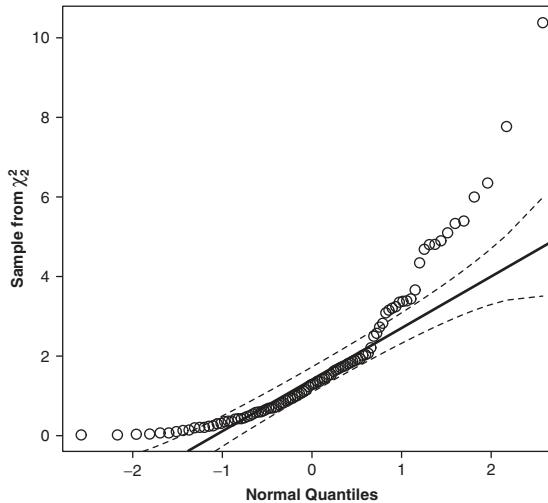


Figure 3.9 Normal quantile-comparison plot for a sample of 100 observations from the positively skewed chi-square distribution with 2 degrees of freedom.

or kernel-density estimates—that provide more intuitive representations of distributions. A key point is that there is no reason to limit ourselves to a single picture of a distribution when different pictures bring different aspects of the distribution into relief.

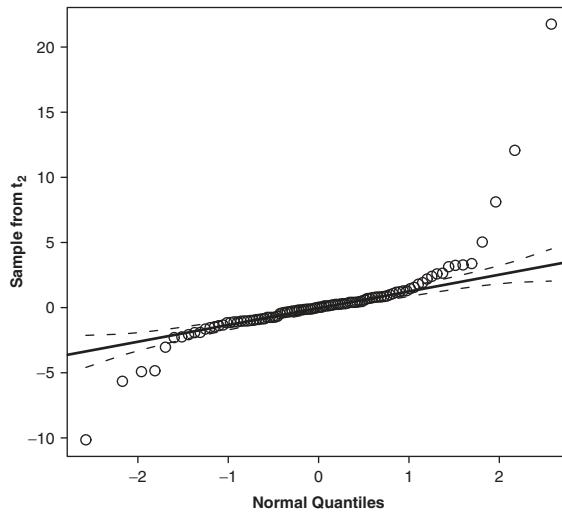


Figure 3.10 Normal quantile-comparison plot for a sample of 100 observations from the heavy-tailed t -distribution with 2 degrees of freedom.

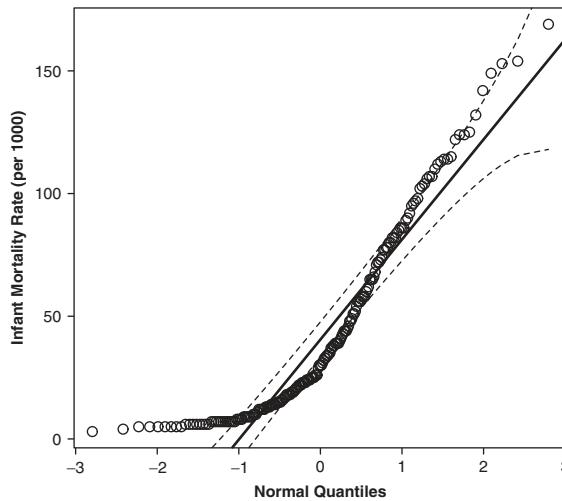


Figure 3.11 Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew.

3.1.4 Boxplots

Unlike histograms, density plots, and quantile-comparison plots, *boxplots* (due to Tukey, 1977) present only summary information on center, spread, and skewness, along with individual outlying observations. Boxplots are constructed from the *five-number summary* of a

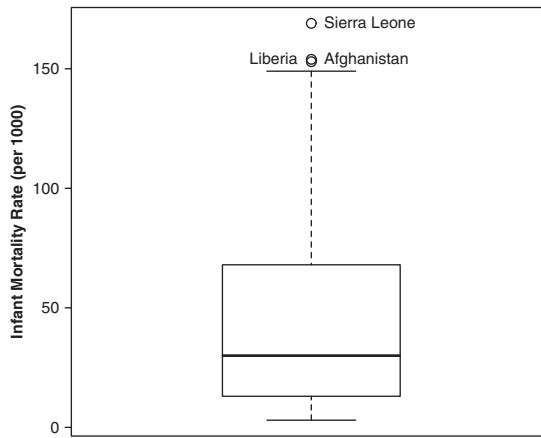


Figure 3.12 Boxplot for infant mortality. The central box is drawn between the hinges, the position of the median is marked in the box, and outlying observations are displayed individually.

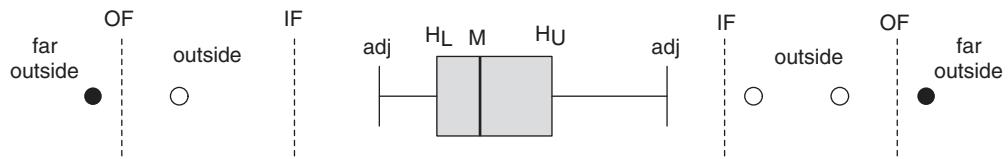


Figure 3.13 Schematic boxplot, showing the median (M), hinges (H_L and H_U), adjacent values (adj), inner and outer fences (IF and OF), and outside and far-outside observations.

distribution—the minimum, first quartile, median, third quartile, and maximum—and outliers, if they are present. Boxplots, therefore, are useful when we require a compact representation of a distribution (as, for example, in the margins of a scatterplot), when we wish to compare the principal characteristics of several distributions,¹⁹ or when we want to select a transformation that makes a distribution more symmetric.²⁰

An illustrative boxplot for infant mortality appears in Figure 3.12. This plot is constructed according to the following conventions (illustrated in the schematic horizontal boxplot in Figure 3.13):

1. A scale is laid off to accommodate the extremes of the data. The infant mortality data, for example, range between 3 and 169.
2. The central box is drawn between the *hinges*, which are simple definitions of the first and third quartiles, and therefore encompasses the middle half of the data. The line in the central box represents the median. Recall that the depth of the median is

¹⁹See Section 3.2.

²⁰Transformations to symmetry are discussed in Chapter 4.

$$\text{depth}(M) = \frac{n+1}{2}$$

giving the position of the middle observation after the data are ordered from smallest to largest: $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. When n is even, the depth of the median has a fractional part; using “floor” brackets to represent truncation to an integer, we count in from either end to average the two observations at depth $\lfloor(n+1)/2\rfloor$. For the infant mortality data, $\text{depth}(M) = (193+1)/2 = 97$, and $M = X_{(97)} = 30$.

Likewise, the depth of the hinges is

$$\text{depth}(H) = \frac{\lfloor \text{depth}(M) \rfloor + 1}{2}$$

If $\text{depth}(H)$ has a fractional part, then, for each hinge, we average the two observations at the adjacent positions, that is, at $\lfloor \text{depth}(H) \rfloor$ and $\lfloor \text{depth}(H) \rfloor + 1$. For the infant mortality distribution, $\text{depth}(H) = (97+1)/2 = 49$. The lower hinge is, therefore, $H_L = X_{(49)} = 13$, and the upper hinge is $H_U = X_{(145)} = 68$. (Counting down 97 observations from the top yields the subscript $193 - 49 + 1 = 145$.)

3. The following rules are used to identify outliers, which are shown individually in the boxplot:

- The *hinge-spread* (roughly the interquartile range) is the difference between the hinges:

$$H\text{-spread} = H_U - H_L$$

- The lower and upper “inner fences” are located 1.5 hinge-spreads beyond the hinges:

$$\text{IF}_L = H_L - 1.5 \times H\text{-spread}$$

$$\text{IF}_U = H_U + 1.5 \times H\text{-spread}$$

Observations beyond the inner fences (but within the outer fences, defined below) are termed “outside” and are represented by open circles. The fences themselves are not shown in the display.

- The “outer fences” are located three hinge-spreads beyond the hinges:²¹

$$\text{OF}_L = H_L - 3 \times H\text{-spread}$$

$$\text{OF}_U = H_U + 3 \times H\text{-spread}$$

Observations beyond the outer fences are termed “far outside” and are represented by filled circles. There are no far-outside observations in the infant mortality data.

²¹Here is a rough justification for the fences: In a normal population, the hinge-spread is 1.349 standard deviations, and so $1.5 \times H\text{-spread} = 1.5 \times 1.349 \times \sigma \approx 2\sigma$. The hinges are located $1.349/2 \approx 0.7$ standard deviations above and below the mean. The inner fences are, therefore, approximately at $\mu \pm 2.7\sigma$ and the outer fences at $\mu \pm 4.7\sigma$. From the standard normal table, $\Pr(Z > 2.7) \approx .003$, so we expect slightly less than 1% of the observations beyond the inner fences ($2 \times .003 = .006$); likewise, because $\Pr(Z > 4.7) \approx 1.3 \times 10^{-6}$, we expect less than one observation in 100,000 beyond the outer fences.

- The “whisker” growing from each end of the central box extends either to the extreme observation on its side of the distribution (as at the low end of the infant mortality data) or to the most extreme nonoutlying observation, called the “adjacent value” (as at the high end of the infant mortality distribution).²²

The boxplot of infant mortality in Figure 3.12 clearly reveals the skewness of the distribution: The lower whisker is much shorter than the upper whisker, the median is closer to the lower hinge than to the upper hinge, and there are several outside observations at the upper end of the infant mortality distribution but none at the lower end. The apparent bimodality of the infant mortality data is not captured by the boxplot, however.

There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile comparison plots are useful for comparing data with a theoretical probability distribution. Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and outliers.

3.2 Plotting Bivariate Data

The *scatterplot*—a direct geometric representation of observations on two quantitative variables (generically, Y and X)—is the most useful of all statistical graphs. The scatterplot is a natural representation of data partly because the media on which we draw plots—paper, computer screens—are intrinsically two-dimensional. Scatterplots are as familiar and essentially simple as they are useful; I will therefore limit this presentation to a few points. There are many examples of bivariate scatterplots in this book, including in the preceding chapter.

- In analyzing data, it is convenient to work in a computing environment that permits the interactive identification of observations in a scatterplot.
- Because relationships between variables in the social sciences are often weak, scatterplots can be dominated visually by “noise.” It often helps, therefore, to plot a nonparametric regression of Y on X .
- Scatterplots in which one or both variables are highly skewed are difficult to examine, because the bulk of the data congregate in a small part of the display. Consider, for example, the scatterplot for infant mortality and gross domestic product (GDP) per capita in Figure 3.14. It often helps to “correct” substantial skews prior to examining the relationship between Y and X .²³
- Scatterplots in which the variables are discrete can also be difficult to examine. An extreme instance of this phenomenon is shown in Figure 3.15, which plots scores on a

²²All of the folksy terminology—*hinges*, *fences*, *whiskers*, and so on—originates with Tukey (1977).

²³See Chapter 4.

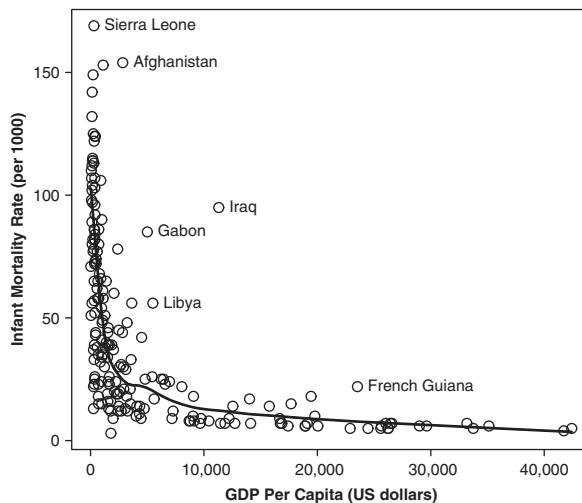


Figure 3.14 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

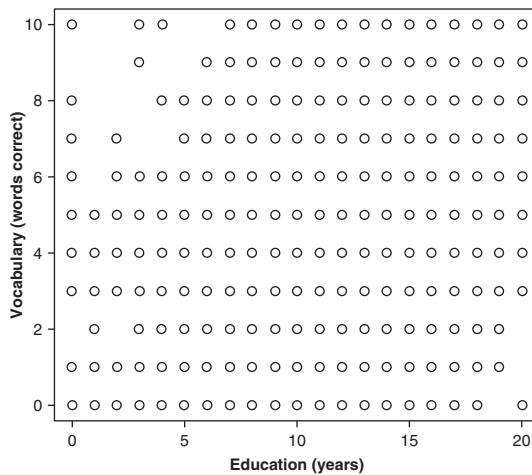


Figure 3.15 Scatterplot of scores on a 10-item vocabulary test versus years of education. Although there are nearly 22,000 observations in the data set, most of the plotted points fall on top of one another.

SOURCE: National Opinion Research Center (2005).

10-item vocabulary test against years of education. The data are from 16 of the U.S. General Social Surveys conducted by the National Opinion Research Center between 1974 and 2004 and include in total 21,638 observations. One solution—especially useful when only X is discrete—is to focus on the conditional distribution of Y for each value

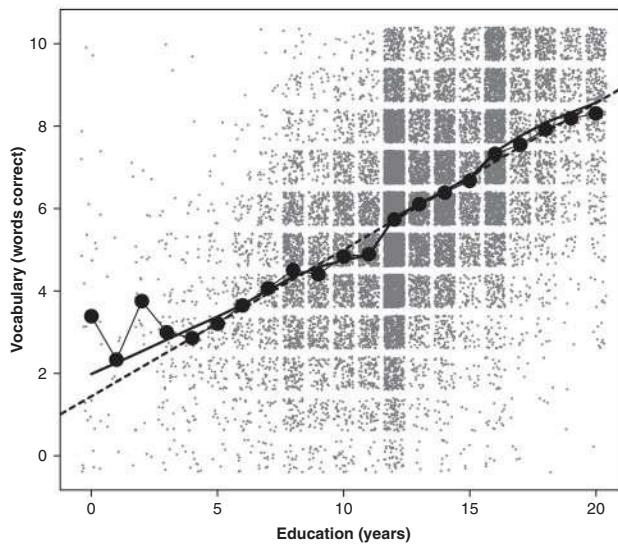


Figure 3.16 Jittered scatterplot for vocabulary score versus years of education. A uniformly distributed random quantity between -0.4 and $+0.4$ was added to each score for both variables. The heavier solid line is for a lowess fit to the data, with a span of 0.2; the broken line is the linear least-squares fit; the conditional means for vocabulary given education are represented by the dots, connected by the lighter solid line.

of X . Boxplots, for example, can be employed to represent the conditional distributions (see Figure 3.17, discussed below). Another solution is to separate overlapping points by adding a small random quantity to the discrete scores. In Figure 3.16, for example, I have added a uniform random variable on the interval $[-0.4, +0.4]$ to each value of vocabulary and education. Paradoxically, the tendency for vocabulary to increase with education is much clearer in the randomly “jittered” display.²⁴

The bivariate scatterplot is a natural graphical display of the relationship between two quantitative variables. Interpretation of a scatterplot can often be assisted by graphing a nonparametric regression, which summarizes the relationship between the two variables. Scatterplots of the relationship between discrete variables can be enhanced by randomly jittering the data.

As mentioned, when the explanatory variable is discrete, parallel boxplots can be used to display the conditional distributions of Y . One common case occurs when the explanatory variable is a qualitative/categorical variable. An example is shown in Figure 3.17, using data collected by Michael Ornstein (1976) on interlocking directorates among the 248 largest Canadian firms.

²⁴The idea of jittering a scatterplot, as well as the terminology, is due to Cleveland (1994).

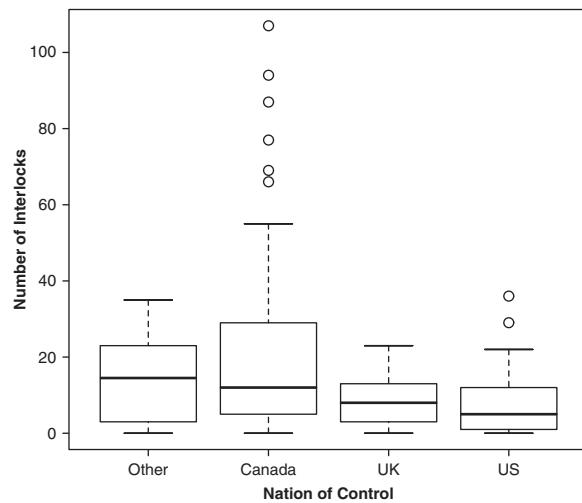


Figure 3.17 Number of interlocking directorate and executive positions by nation of control for 248 dominant Canadian firms.

SOURCE: Personal communication from Michael Ornstein.

The response variable in this graph is the number of interlocking directorships and executive positions maintained by each firm with others in the group of 248. The explanatory variable is the nation in which the corporation is controlled, coded as Canada, the United Kingdom, the United States, and other foreign.

It is apparent from the graph that the average level of interlocking is greater among other-foreign and Canadian corporations than among corporations controlled in the United Kingdom and the United States. It is relatively difficult to discern detail in this display: first, because the conditional distributions of interlocks are positively skewed and, second, because there is an association between level and spread—variation is also greater among other-foreign and Canadian firms than among U.K. and U.S. firms.²⁵

Parallel boxplots display the relationship between a quantitative response variable and a discrete (categorical or quantitative) explanatory variable.

3.3 Plotting Multivariate Data

Because paper and computer screens are two-dimensional, graphical display of multivariate data is intrinsically difficult. Multivariate displays for quantitative data often project the

²⁵We will revisit this example in Section 4.4. Because the names of the firms are unavailable, I have not identified the outliers in the plot.

higher-dimensional “point cloud” of the data onto a two-dimensional space. It is, of course, impossible to view a higher-dimensional scatterplot directly (but see the discussion of the three-dimensional case below). The essential trick of effective multidimensional display is to select projections that reveal important characteristics of the data. In certain circumstances, projections can be selected on the basis of a statistical model fit to the data or on the basis of explicitly stated criteria.²⁶

3.3.1 Scatterplot Matrices

A simple approach to multivariate data, which does not require a statistical model, is to examine bivariate scatterplots for all pairs of variables. Arraying these plots in a *scatterplot matrix* produces a graphical analog to the correlation matrix.

An illustrative scatterplot matrix, for data on the prestige, education, and income levels of 45 U.S. occupations, appears in Figure 3.18. In this data set, first analyzed by Duncan (1961), “prestige” represents the percentage of respondents in a survey who rated an occupation as “good” or “excellent” in prestige, “education” represents the percentage of incumbents in the occupation in the 1950 U.S. Census who were high school graduates, and “income” represents the percentage of occupational incumbents who earned incomes in excess of \$3,500. Duncan’s purpose was to use a regression analysis of prestige on income and education to predict the prestige levels of other occupations, for which data on income and education were available but for which there were no direct prestige ratings.²⁷

The variable names on the diagonal of the scatterplot matrix in Figure 3.18 label the rows and columns of the display: For example, the vertical axis for the two plots in the first row of the display is “prestige”; the horizontal axis for the two plots in the second column is “education.” Thus, the scatterplot in the first row, second column is for prestige (on the vertical axis) versus education (on the horizontal axis).

It is important to understand an essential limitation of the scatterplot matrix as a device for analyzing multivariate data: By projecting the multidimensional point cloud onto pairs of axes, the plot focuses on the *marginal* relationships between the corresponding pairs of variables. The object of data analysis for several variables, however, is typically to investigate *partial* relationships (between pairs of variables, “controlling” statistically for other variables), not marginal associations. For example, in the Duncan data set, we are more interested in the partial relationship of prestige to education holding income constant than in the marginal relationship between prestige and education ignoring income.

The response variable Y can be related marginally to a particular X , even when there is no partial relationship between the two variables controlling for other X s. It is also possible for there to be a partial association between Y and an X but no marginal association. Furthermore, if the X s themselves are nonlinearly related, then the marginal relationship between Y and a specific X can be nonlinear even when their partial relationship is linear.²⁸

Despite this intrinsic limitation, scatterplot matrices often uncover interesting features of the data, and this is indeed the case in Figure 3.18, where the display reveals three

²⁶We will apply these powerful ideas in Chapters 11 and 12.

²⁷We will return to this regression problem in Chapter 5.

²⁸These ideas are explored in Chapter 12.

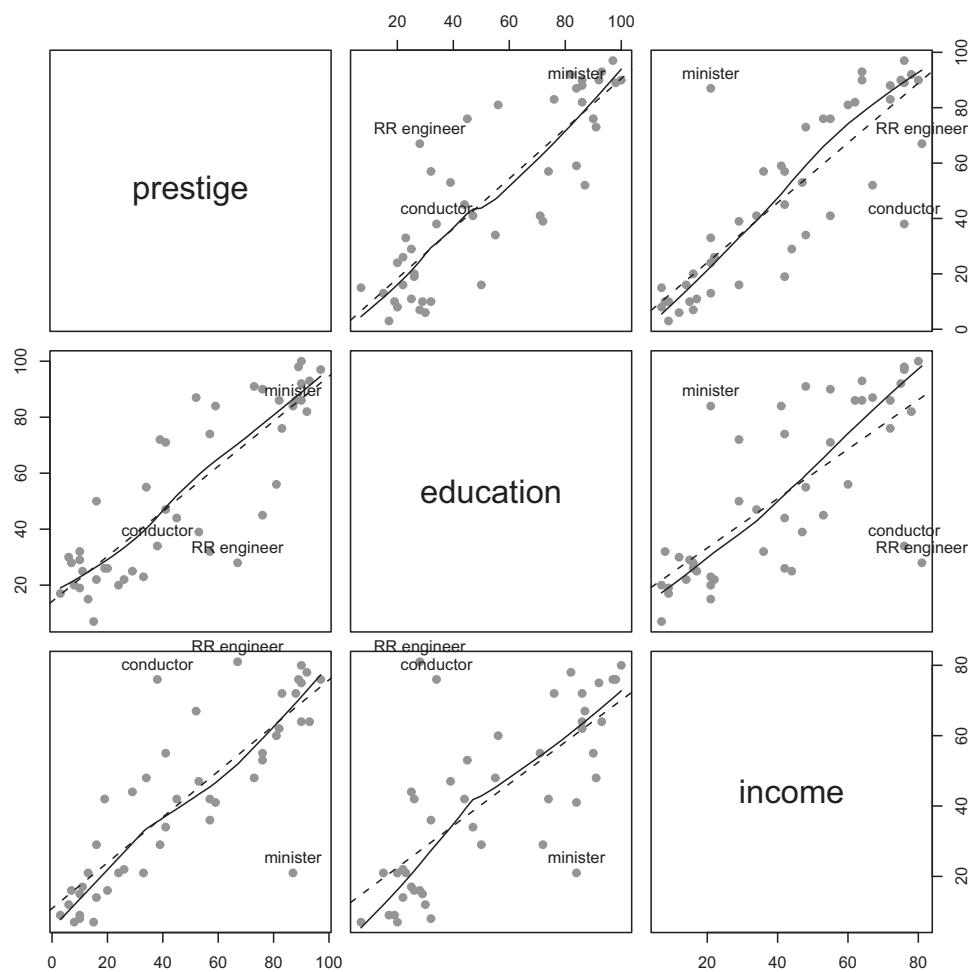


Figure 3.18 Scatterplot matrix for occupational prestige, level of education, and level of income for 45 U.S. occupations in 1950. The least-squares regression line (broken line) and lowess smooth (for a span of 0.6, solid line) are shown on each plot. Three unusual observations are identified.

SOURCE: Duncan (1961).

unusual observations: *Ministers* have relatively low income for their relatively high level of education and relatively high prestige for their relatively low income; *railroad conductors* and *railroad engineers* have relatively high incomes for their more-or-less average levels of education; *railroad conductors* also have relatively low prestige for their relatively high incomes. This pattern bodes ill for the least-squares linear regression of prestige on income and education.²⁹

²⁹See the discussion of Duncan's occupational-prestige regression in Chapter 11.

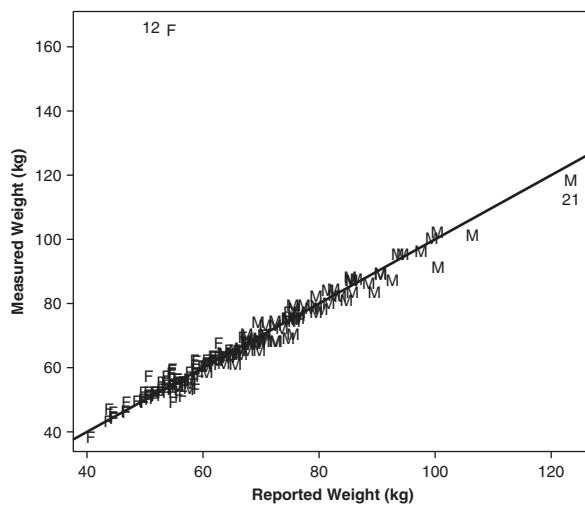


Figure 3.19 Davis’s data on measured and reported weight, by gender. Data points are represented by Ms for men and Fs for women and are jittered slightly to reduce overplotting. The line on the graph is $Y = X$. In the combined data set for men and women, the outlying observation is number 12.

3.3.2 Coded Scatterplots

Information about a categorical third variable can be entered on a bivariate scatterplot by coding the plotting symbols. The most effective codes use different colors to represent categories, but degrees of fill, distinguishable shapes, and distinguishable letters can also be effective.³⁰

Figure 3.19 shows a scatterplot of Davis’s (1990) data on measured and reported weight.³¹ Observations are displayed as Ms for men and Fs for women. Except for the outlying point (number 12—which, recall, represents an error in the data), the points both for men and for women cluster near the line $Y = X$; it is also clear from the display that most men are heavier than most women, as one would expect, and that, discounting the bad data point, one man (number 21) is quite a bit heavier than everyone else.

3.3.3 Three-Dimensional Scatterplots

Another useful multivariate display, directly applicable to three variables at a time, is the *three-dimensional scatterplot*. Moreover, just as data can be projected onto a judiciously chosen plane in a two-dimensional plot, higher-dimensional data can be projected onto a three-dimensional space, expanding the range of application of three-dimensional scatterplots.³²

³⁰See Spence and Lewandowsky (1990) for a fine review of the literature on graphical perception, including information on coded scatterplots.

³¹Davis’s data were introduced in Chapter 2, where only the data for women were presented.

³²For example, there are three-dimensional versions of the added-variable and component-plus-residual plots discussed in Chapters 11 and 12. See, for example, Cook and Weisberg (1989).

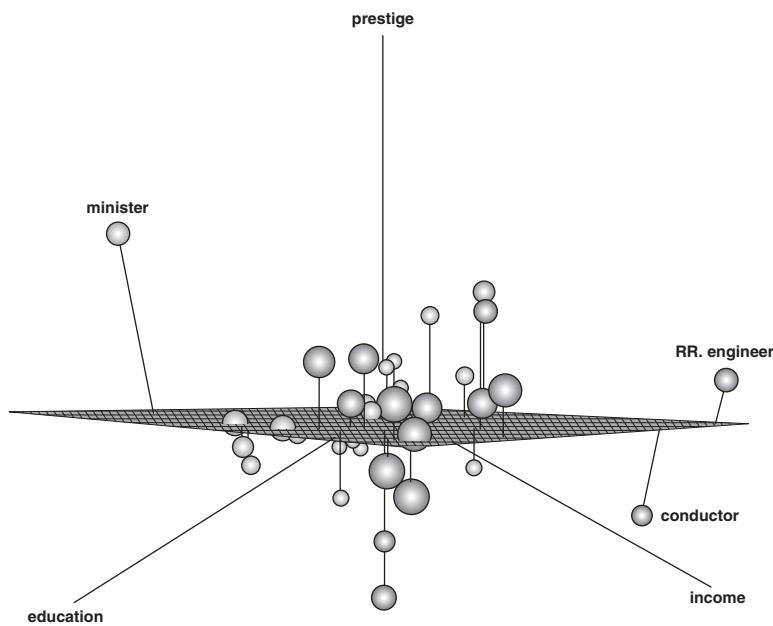


Figure 3.20 Three-dimensional scatterplot for Duncan's occupational prestige data, rotated into an orientation that reveals three unusual observations. From this orientation, the least-squares regression plane, also shown in the plot, is viewed nearly edge on.

Barring the use of a true stereoscopic display, the three-dimensional scatterplot is an illusion produced by modern statistical software: The graph represents a projection of a three-dimensional space onto a two-dimensional computer screen. Nevertheless, motion (e.g., rotation) and the ability to interact with the display—possibly combined with the effective use of perspective, color, depth cueing, and other visual devices—can produce a vivid impression of directly examining objects in three-dimensional space.

It is literally impossible to convey this impression adequately on the static, two-dimensional page of a book, but Figure 3.20 shows Duncan's (1961) prestige data rotated interactively into a revealing orientation: Looking down the cigar-shaped scatter of most of the data, the three unusual observations stand out very clearly.

3.3.4 Conditioning Plots

Conditioning plots (or *coplots*), described in Cleveland (1993), are another graphical device for examining multidimensional data. The essential idea of the coplot is to focus on the relationship between the response variable and a particular explanatory variable, dividing the data into groups based on the values of other explanatory variables—the *conditioning variables*. If the conditioning variables are discrete, then this division is straightforward and natural. If a conditioning variable is continuous, it can be binned: Cleveland suggests using overlapping bins, which are called “shingles.”

An illustrative coplot, for the General Social Survey vocabulary data, is shown in Figure 3.21. This graph displays the relationship between vocabulary score and education,

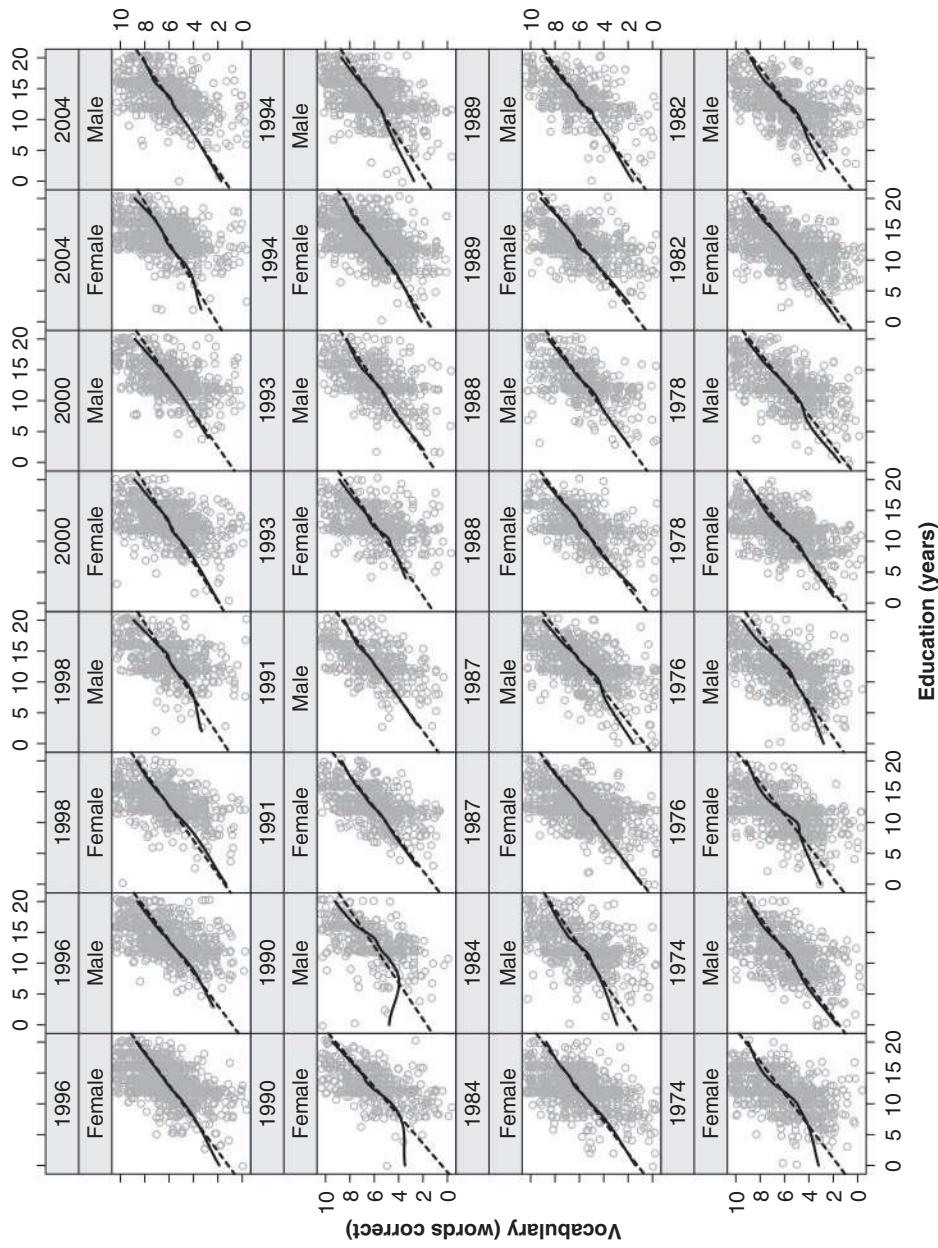


Figure 3.21 Coplot showing the relationship between vocabulary score and education controlling for year and gender. The points in each panel are jittered to reduce overplotting. The broken line shows the linear-squares fit, while the solid line gives the loess fit for a span of 0.6.

“controlling for” gender and the year of the survey. The partial relationships are remarkably similar in the different panels of the coplot; that is, gender and year appear to make little difference to the relationship between vocabulary score and education. The relationships also appear to be very close to linear: In a few panels, the lowess line departs from the linear least-squares line at the far left, but data in this region are quite sparse.

Although they can be effective graphs, coplots have limitations: First, if there are more than two, or perhaps three, conditioning variables, it becomes difficult to perceive how the partial relationship between the response and the focal explanatory variable changes with the conditioning variables. Second, because coplots require the division of the data into groups, they are most useful for large data sets, an issue that grows more acute as the number of conditioning variables increases.

Visualizing multivariate data is intrinsically difficult because we cannot directly examine higher-dimensional scatterplots. Effective displays project the higher-dimensional point cloud onto two or three dimensions; these displays include the scatterplot matrix, the dynamic three-dimensional scatterplot, and the conditioning plot.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Summary

- Statistical graphs are central to effective data analysis, both in the early stages of an investigation and in statistical modeling.
- There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile-comparison plots are useful for comparing data with a theoretical probability distribution. Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and outliers.
- The bivariate scatterplot is a natural graphical display of the relationship between two quantitative variables. Interpretation of a scatterplot can often be assisted by graphing a nonparametric regression, which summarizes the relationship between the two variables. Scatterplots of the relationship between discrete variables can be enhanced by randomly jittering the data.
- Parallel boxplots display the relationship between a quantitative response variable and a discrete explanatory variable.
- Visualizing multivariate data is intrinsically difficult because we cannot directly examine higher-dimensional scatterplots. Effective displays project the higher-dimensional point cloud onto two or three dimensions; these displays include the scatterplot matrix, the dynamic three-dimensional scatterplot, and the conditioning plot.

Recommended Reading

The literature—especially the recent literature—on statistical graphics is truly voluminous. I will furnish only the briefest of bibliographies:

- Fox (2000c) presents a brief overview of statistical graphics, including information on the history of the subject. Jacoby (1997, 1998) gives a more extended overview addressed to social scientists.
- Tufte's (1983) influential book on graphical presentation of quantitative information is opinionated but well worth reading. (Tufte has since published several other books on graphics, broadly construed, but I prefer his first book.)
- Modern interest in statistical graphics is the direct result of John Tukey's work on exploratory data analysis; unfortunately, Tukey's idiosyncratic writing style makes his seminal book (Tukey, 1977) difficult to read. Velleman and Hoaglin (1981) provide a more digestible introduction to the topic. There is interesting information on the statistical theory underlying exploratory data analysis in two volumes edited by Hoaglin, Mosteller, and Tukey (1983, 1985).
- Tukey's influence made Bell Labs a center of work on statistical graphics, much of which is described in two accessible and interesting books by William Cleveland (1993, 1994) and in Chambers, Cleveland, Kleiner, and Tukey (1983). Cleveland (1994) is a good place to start.
- Modern statistical graphics is closely associated with advances in statistical computing: The S statistical computing environment (Becker, Chambers, & Wilks, 1988; Chambers, 1998; Chambers & Hastie, 1992), also a product of Bell Labs, is particularly strong in its graphical capabilities. R, a free, open-source implementation of S, was mentioned in the preface. Cook and Weisberg (1994, 1999) use the Lisp-Stat statistical computing environment (Tierney, 1990) to produce an impressive statistical package, called Arc, which incorporates a variety of statistical graphics of particular relevance to regression analysis (including many of the methods described later in this text). Friendly (1991) describes how to construct modern statistical graphs using the SAS/Graph system. Brief presentations of these and other statistical computing environments appear in a book edited by Stine and Fox (1996).
- Atkinson (1985) presents a variety of innovative graphs in support of regression analysis, as do Cook (1998) and Cook and Weisberg (1994, 1999).

4

Transforming Data

“Classical” statistical models, for example, linear least-squares regression, make strong assumptions about the structure of data—assumptions that, more often than not, fail to hold in practice. One solution is to abandon classical methods in favor of more flexible alternatives, such as nonparametric regression analysis. These newer methods are valuable, and I expect that they will be used with increasing frequency, but they are more complex and have their own limitations, as we saw in Chapter 2.¹

It is, alternatively, often feasible to transform the data so that they conform more closely to the restrictive assumptions of classical statistical models. In addition, and as we will discover in this chapter, transformations can often assist in the examination of data, even in the absence of a statistical model. The chapter introduces two general families of transformations² and shows how they can be used to make distributions symmetric, to make the relationship between two variables linear, and to equalize variation across groups.

Transformations can often facilitate the examination and statistical modeling of data.

4.1 The Family of Powers and Roots

There is literally an infinite variety of functions $f(x)$ that could be used to transform a quantitative variable X . In practice, of course, it helps to be more restrictive, and a particularly useful group of transformations is the “family” of powers and roots:

$$X \rightarrow X^p \quad (4.1)$$

where the arrow indicates that we intend to replace X with the transformed variable X^p . If p is negative, then the transformation is an inverse power: For example, $X^{-1} = 1/X$ (i.e., inverse), and $X^{-2} = 1/X^2$ (inverse square). If p is a fraction, then the transformation represents a root: For example, $X^{1/3} = \sqrt[3]{X}$ (cube root) and $X^{-1/2} = 1/\sqrt{X}$ (inverse square root).

For some purposes, it is convenient to define the family of power transformations in a slightly more complex manner, called the *Box-Cox family* of transformations (introduced in a seminal paper on transformations by Box & Cox, 1964):³

¹Also see Chapter 18.

²A third family of transformations is described in Exercise 4.4.

³In addition to revealing the relative effect of different power transformations, the Box-Cox formulation is useful for estimating a transformation as a parameter, as in Section 4.6.

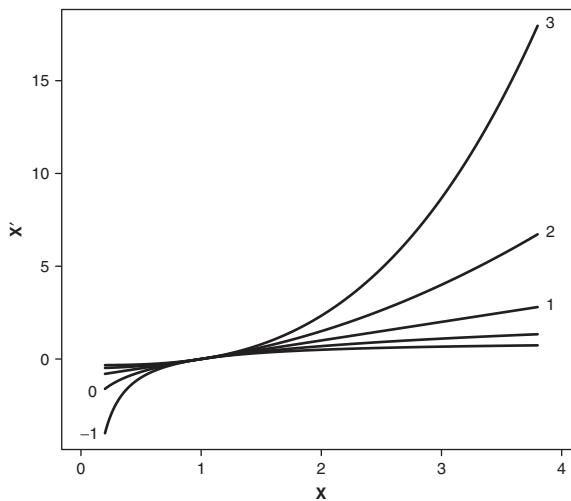


Figure 4.1 The Box-Cox family of power transformations X' of X . The curve labeled p is the transformation $X^{(p)}$, that is, $(X^p - 1)/p$; $X^{(0)}$ is $\log_e(p)$.

$$X \rightarrow X^{(p)} \equiv \frac{X^p - 1}{p} \quad (4.2)$$

We use the parenthetical superscript (p) to distinguish this definition from the more straightforward one in Equation 4.1. Because $X^{(p)}$ is a linear function of X^p , the two transformations have the same essential effect on the data, but, as is apparent in Figure 4.1, the definition in Equation 4.2 reveals more transparently the essential unity of the family of powers and roots:⁴

- Dividing by p preserves the direction of X , which otherwise would be reversed when p is negative, as illustrated in the following example:

X	X^{-1}	$\frac{X^{-1}}{-1}$	$\frac{X^{-1} - 1}{-1}$
1	1	-1	0
2	1/2	-1/2	1/2
3	1/3	-1/3	2/3
4	1/4	-1/4	3/4

Note that subtracting 1 from the numerator does not affect *differences* between adjacent transformed values in the table.

- The transformations $X^{(p)}$ are “matched” above $X = 1$ both in level and in slope: (1) $1^{(p)} = 0$, for all values of p , and (2) each transformation has a slope of 1 at $X = 1$.⁵
- Matching the transformations facilitates comparisons among them and highlights their relative effects on the data. In particular, descending the “ladder” of powers and roots

⁴See Exercise 4.1.

⁵*That is, the derivative of $X^{(p)}$ at $X = 1$ is 1; see Exercise 4.2.

toward $X^{(-1)}$ compresses the large values of X and spreads out the small ones; ascending the ladder of powers and roots toward $X^{(2)}$ has the opposite effect.⁶ As p moves further from $p = 1$ (i.e., no transformation) in either direction, the transformation grows more powerful, increasingly “bending” the data.

- The power transformation X^0 is useless because it changes all values to 1, but we can think of the log (i.e., logarithm) transformation as a kind of “zeroth” power: As p gets very close to 0, the log function more and more closely approximates $X^{(p)}$.⁷ Because the log transformation is so useful, we will, by convention, take $X^{(0)} \equiv \log_e X$, where $e \approx 2.718$ is the base of the natural logarithms.⁸

In practice, it is generally more convenient to use logs to the base 10 or base 2, which are more easily interpreted than logs to the base e : For example, increasing $\log_{10} X$ by 1 is equivalent to multiplying X by 10; increasing $\log_2 X$ by 1 is equivalent to doubling X . Selection of a base for the log transformation is essentially arbitrary and inconsequential, however, because changing bases is equivalent to multiplying by a constant; for example,

$$\log_{10} X = \log_{10} e \times \log_e X \approx 0.4343 \times \log_e X$$

Likewise, because of its relative simplicity, we usually use X^p in applied work in preference to $X^{(p)}$ when $p \neq 0$. Transformations such as log, square root, square, and inverse have a long history of use in data analysis, often without reference to each other; thinking about these transformations as members of a family facilitates their systematic application, as illustrated later in this chapter.

The powers and roots are a particularly useful family of transformations: $X \rightarrow X^p$. When $p = 0$, we employ the log transformation in place of X^0 .

The effects of the various power transformations are apparent in Figure 4.1 and in the following simple examples (in which the numbers by the braces give *differences* between adjacent values):

$-1/X$	$\log_2 X$	X	X^2	X^3
-1	0	1	1	1
$\frac{1}{2}\{$ -1/2	1 { 1}	2 } 1	4 } 3	8 } 7
$\frac{1}{6}\{$ -1/3	0.59 { 1.59}	3 } 1	9 } 5	27 } 19
$\frac{1}{12}\{$ -1/4	0.41 { 2}	4 } 1	16 } 7	64 } 37

⁶The heuristic characterization of the family of powers and roots as a “ladder” follows Tukey (1977).

⁷More formally,

$$\lim_{p \rightarrow 0} \frac{X^p - 1}{p} = \log_e X$$

⁸Powers and logarithms are reviewed in online Appendix C.

Power transformations are sensible only when all the values of X are positive. First of all, some of the transformations, such as square root, log, and inverse, are undefined for negative or zero values (or both). Second, even when they are defined, the power transformations are not monotone—that is, not order preserving—if there are both positive and negative values in the data; for example,

X	X^2
−2	4
−1	1
0	0
1	1
2	4

This is not, however, a practical limitation, because we can always add a positive constant (called a “start”) to each data value to make all the values positive, calculating the transformation $X \rightarrow (X + s)^p$;⁹ in the preceding example,

X	$(X+3)^2$
−2	1
−1	4
0	9
1	16
2	25

It is, finally, worth pointing out that power transformations are effective only when the ratio of the largest data values to the smallest ones is sufficiently large; if, in contrast, this ratio is close to 1, then power transformations are nearly linear and, hence, ineffective at bending the data. Consider the following example, where the ratio of the largest to the smallest data value is only $2015/2011 = 1.002 \approx 1$:

X	$\log_{10} X$
2011	3.30341
1 { 2012	3.30363 } 0.00022
1 { 2013	3.30384 } 0.00021
1 { 2014	3.30406 } 0.00022
1 { 2015	3.30428 } 0.00022

⁹An alternative family of power transformations that can handle positive and negative data is described in Exercise 4.4.

Using a negative start produces the desired effect:

X	$\log_{10}(X - 2010)$
2011	0
1 { 2012}	0.301 } 0.301
1 { 2013}	0.176 } 0.176
1 { 2014}	0.477 } 0.477
1 { 2015}	0.602 } 0.602
	0.097 } 0.097

This strategy should be considered whenever the ratio of the largest to the smallest data value is less than about 5. When the ratio is sufficiently large—either initially or after subtracting a suitable start—an adequate power transformation can typically be found in the range $-2 \leq p \leq 3$. We usually select integer values of p or simple fractions such as $\frac{1}{2}$ or $\frac{1}{3}$.

Power transformations preserve the order of the data only when all values are positive and are effective only when the ratio of the largest to the smallest data values is itself large. When these conditions do not hold, we can impose them by adding a positive or negative start to all the data values.

4.2 Transforming Skewness

Power transformations can make a skewed distribution more symmetric. But why should we bother?

- Highly skewed distributions are difficult to examine because most of the observations are confined to a small part of the range of the data. Recall from the previous chapter, for example, the distribution of infant mortality rates, redisplayed in Figure 4.2.¹⁰
- Apparently, outlying values in the direction of the skew are brought in toward the main body of the data when the distribution is made more symmetric. In contrast, unusual values in the direction opposite to the skew can be hidden prior to transforming the data.
- Some of the most common statistical methods summarize distributions using means. Least-squares regression, which traces the mean of Y conditional on X , comes immediately to mind.¹¹ The mean of a skewed distribution is not, however, a good summary of its center.

¹⁰Adapting Figure 3.6 on page 46.

¹¹See Chapter 5.

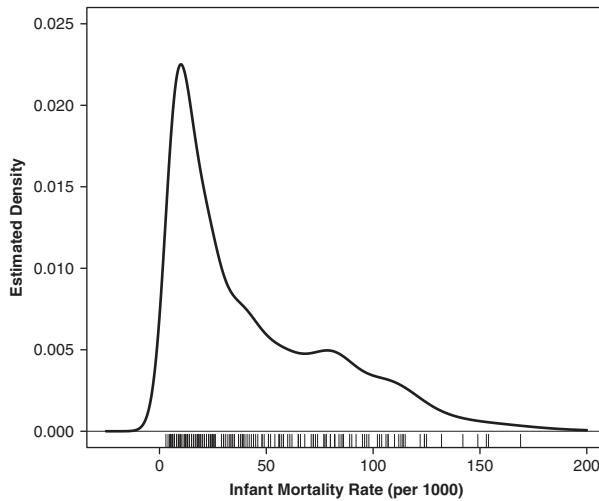


Figure 4.2 Adaptive-kernel density estimate for the distribution of infant mortality rates of 193 nations of the world. The data values are displayed in the rug-plot at the bottom of the figure.

The following simple example illustrates how a power transformation can eliminate a positive skew:

X	$\log_{10} X$
9	0
{ 10	1 }
{ 90	1 }
{ 100	2 }
{ 900	1 }
1000	3

Descending the ladder of powers to $\log X$ makes the distribution more symmetric by pulling in the right tail. Ascending the ladder of powers (toward X^2 and X^3) can, similarly, “correct” a negative skew.

An effective transformation can be selected analytically or by trial and error.¹² Examining the median and the hinges, moreover, can provide some guidance to trial and error. A convenient property of order statistics—including the median and hinges—is that they are preserved under nonlinear monotone transformations of the data, such as powers and roots; that is, if

¹²See Sections 4.6 and 12.5 for analytic methods for selecting transformations.

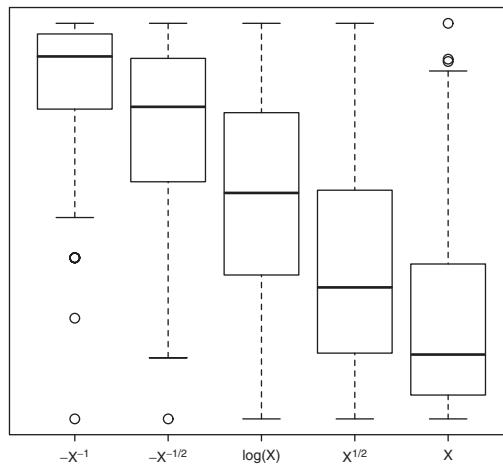


Figure 4.3 Boxplots for various power transformations of infant mortality; because the distribution of infant mortality is positively skewed, only transformations “down” the ladder of powers and roots are considered.

$X' = X^{(p)}$, then $X'_{(i)} = [X_{(i)}]^{(p)}$, and thus $\text{median}(X') = [\text{median}(X)]^{(p)}$.¹³ This is not the case for the mean and standard deviation.

In a symmetric distribution, the median is midway between the hinges, and consequently, the ratio

$$\frac{\text{Upper hinge} - \text{Median}}{\text{Median} - \text{Lower hinge}}$$

is approximately 1. In contrast, a positive skew is reflected in a ratio that exceeds 1 and a negative skew in a ratio that is smaller than 1. Trial and error can begin, therefore, with a transformation that makes this ratio close to 1.

Some statistical software allows the transformation p to be selected interactively using a “slider,” while a graph of the distribution—for example, a density plot—is updated when the value of p changes. This is an especially convenient and effective approach. A static alternative is to show parallel boxplots for various transformations, as in Figure 4.3 for the infant mortality data.

For the distribution of infant mortality rates, we have

Transformation	H_U	Median	H_L	$\frac{H_U - \text{Median}}{\text{Median} - H_L}$
X	68	30	13	2.23
\sqrt{X}	8.246	5.477	3.605	1.48
$\log_{10} X$	1.833	1.477	1.114	0.98
$-1/\sqrt{X}$	-0.1213	-0.1825	-0.2773	0.65
$-1/X$	-0.01471	-0.03333	-0.07692	0.43

¹³There is some slippage here because the median and hinges sometimes require averaging adjacent order statistics. The two averaged values are seldom very far apart, however, and therefore the distinction between the median of the transformed values and the transformation of the median is almost always trivial. The same is true for the hinges. The results presented for the example give the median and hinges of the transformed data.

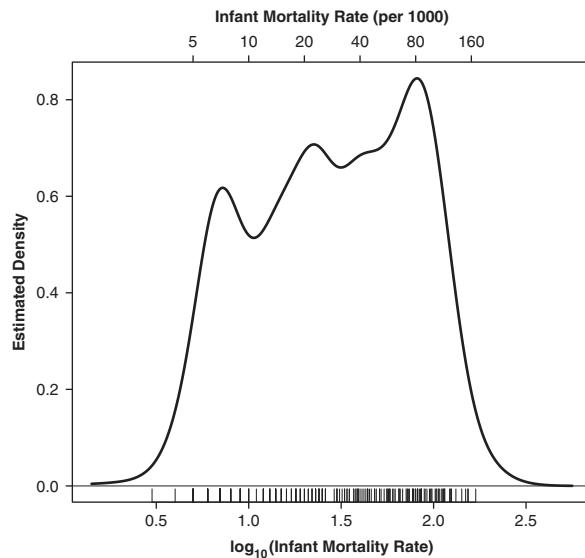


Figure 4.4 Adaptive-kernel density estimate for the distribution of \log_{10} infant mortality. The window half-width for the adaptive-kernel estimator is $h = 0.1$ (on the \log_{10} infant mortality scale). A rug-plot of the data values appears at the bottom of the graph and the original infant mortality scale at the top.

This table and the boxplots in Figure 4.3 suggest the log transformation of infant mortality, and the result of transforming the data is shown in Figure 4.4. Not only is the distribution much more symmetric than before, but three modes are clearly resolved (and there is the suggestion of a fourth); the modes at infant mortality rates of about 7 and 20 were not distinguishable in the untransformed, positively skewed data.

Note the untransformed scale for infant mortality at the top of the graph: These values, which are equally spaced on the log scale, represent doubling of infant mortality rates. This is, in my experience, an effective device for presenting the results of a statistical analysis in which the familiar scale of a variable is lost through a transformation.

Although it is not the case here, where the log transformation is clearly indicated, we often have a choice between transformations that perform roughly equally well. Although we should try to avoid distorting the data, we may prefer one transformation to another because of interpretability. I have already mentioned that the log transformation has a convenient multiplicative interpretation. In certain contexts, other transformations may have specific substantive meanings. Here are a few common examples: The inverse of the time (say, in hours) required to travel a given distance (a kilometer) is speed (kilometers per hour); the inverse of response latency (say, in milliseconds, as in a psychophysical experiment) is response frequency (responses per 1,000 seconds); the square root of a measure of area (say, in square meters) is a linear measure of size (in meters); and the cube of a linear measure of size (say in centimeters) can be interpreted as a volume (cubic centimeters).

We generally prefer interpretable transformations when variables are measured on familiar and meaningful scales. Conversely, because the rating “scales” that are ubiquitous in social

research are not really measurements, there is typically no reason to prefer the original scores to a better-behaved monotone transformation of them.¹⁴

Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.

4.3 Transforming Nonlinearity

Power transformations can also be used to make many nonlinear relationships more nearly linear. Again, we ask, why bother?

- Linear relationships—expressible in the form $\hat{Y} = A + BX$ —are particularly simple. Recall that this equation specifies that the average value of the response variable Y is a linear function of the explanatory variable X , with intercept A and slope B . Linearity implies that a unit *increase* in X —regardless of the *level* of X —is associated, on average, with a change of B units in Y .¹⁵ Fitting a linear equation to data makes it relatively easy to answer certain questions about the data: If B is positive, for example, then Y tends to increase with X .
- Especially when there are several explanatory variables, the alternative of nonparametric regression may not be feasible because of the sparseness of the data. Even if we can fit a nonparametric regression with several X s, it may be difficult to visualize the multidimensional result.¹⁶
- There is a simple and elegant statistical theory for linear models, which we explore in subsequent chapters. If these models are reasonable for the data, then their use is convenient.
- There are certain technical advantages to having linear relationships among the *explanatory* variables in a regression analysis.¹⁷

The following simple example suggests how a power transformation can serve to straighten a nonlinear relationship: Suppose that $Y = \frac{1}{3}X^2$ (with no residual) and that X takes on successive integer values between 1 and 5:

¹⁴Rating scales are composed, for example, of items with response categories labeled *strongly agree*, *agree*, *disagree*, and *strongly disagree*. A scale is constructed by assigning arbitrary numbers to the categories (e.g., 1–4) and adding or averaging the items. See Coombs, Dawes, and Tversky (1970, Chapters 2 and 3) for an elementary treatment of measurement issues in the social sciences and Duncan (1984) for an interesting account of the history and practice of social measurement. I believe that social scientists should pay more attention to measurement issues (employing, e.g., the methods of item response theory; e.g., Baker & Kim, 2004). It is unproductive, however, simply to discard rating scales and similar “measurements by fiat” (a felicitous term borrowed from Torgerson, 1958): There is a *prima facie* reasonableness to many rating scales, and to refuse to use them without adequate substitutes would be foolish.

¹⁵I use the terms “increase” and “change” loosely here as a shorthand for static comparisons between average values of Y for X -values that differ by one unit: *Literal* change is not necessarily implied.

¹⁶See, however, the additive regression models discussed in Section 18.2.2, which overcome this deficiency.

¹⁷This point is developed in Section 12.3.3.

X	Y
1	0.2
2	0.8
3	1.8
4	3.2
5	5.0

These “data” are graphed in panel (a) of Figure 4.5, where the nonlinearity of the relationship between Y and X is apparent. Because of the manner in which the example was constructed, it is obvious that there are two simple ways to transform the data to achieve linearity:

1. We could replace Y by $Y' = \sqrt{Y}$, in which case $Y' = \sqrt{\frac{1}{5}X}$.
2. We could replace X by $X' = X^2$, in which case $Y = \frac{1}{5}X'$.

In either event, the relationship is rendered perfectly linear, as shown graphically in panels (b) and (c) of Figure 4.5. To achieve an intuitive understanding of this process, imagine that the original plot in panel (a) is drawn on a rubber sheet: Transforming Y “down” the ladder of powers to square root differentially stretches the rubber sheet vertically so that small values are spread out relative to large ones, stretching the curve in (a) into the straight line in (b). Likewise, transforming X “up” the ladder of powers spreads out the large values relative to the small ones, stretching the curve into the straight line in (c).

A power transformation works here because the relationship between Y and X is smooth, monotone (in this instance, strictly increasing), and simple. What I mean by “simple” in this context is that the direction of curvature of the function relating Y to X does not change (i.e., there is no point of inflection). Figure 4.6 seeks to clarify these distinctions: The relationship in panel (a) is simple and monotone; the relationship in panel (b) is monotone but not simple; and the relationship in panel (c) is simple but not monotone. I like to use the term “curvilinear” for cases such as (c), to distinguish nonmonotone from monotone nonlinearity, but this is not standard terminology. In panel (c), no power transformation of Y or X can straighten the relationship between them, but we could capture this relationship with a quadratic model of the form $\hat{Y} = A + B_1X + B_2X^2$.¹⁸

Like transformations to reduce skewness, a transformation to correct nonlinearity can be selected analytically or by guided trial and error.¹⁹ Figure 4.7 introduces Mosteller and Tukey’s (1977) “bulging rule” for selecting a transformation: If the “bulge” points *down* and to the *right*, for example, we need to transform Y *down* the ladder of powers or X *up* (or both). This case corresponds to the example in Figure 4.5, and the general justification of the rule follows from the need to stretch an axis differentially to transform the curve into a straight line. Trial and error is simplest with software that provides “sliders” for the power transformations of X and Y , immediately displaying the effect of a change in either power on the scatterplot relating the two variables, but we can in any event examine a series of scatterplots for different transformations.

¹⁸Quadratic and other polynomial regression models are discussed in Section 17.1.

¹⁹See Sections 4.6 and 12.5 for analytic methods of selecting linearizing transformations.

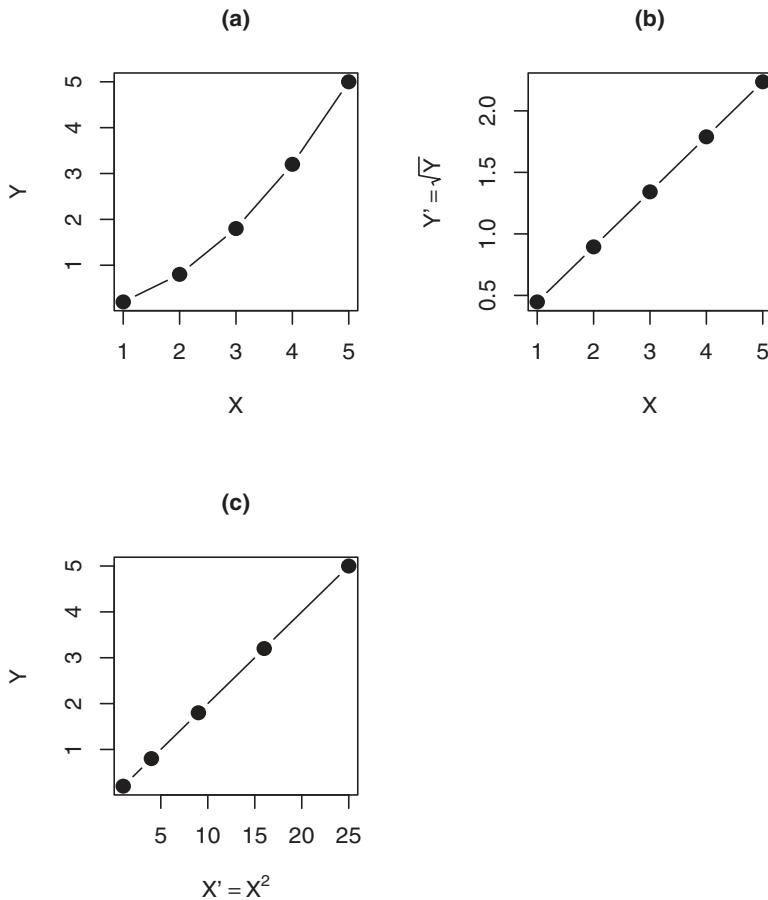


Figure 4.5 How a power transformation of Y or X can make a simple monotone nonlinear relationship linear. Panel (a) shows the relationship $Y = \frac{1}{5}X^2$. In panel (b), Y is replaced by the transformed value $Y' = Y^{1/2}$. In panel (c), X is replaced by the transformed value $X' = X^2$.

Simple monotone nonlinearity can often be corrected by a power transformation of X , of Y , or of both variables. Mosteller and Tukey's bulging rule assists in selecting linearizing transformations.

Let us reexamine, in the light of this discussion, the relationship between prestige and income for the 102 Canadian occupations first encountered in Chapter 2 and shown in Figure 4.8.²⁰ The relationship between prestige and income is clearly monotone and nonlinear: Prestige rises

²⁰Repeating Figure 2.10 on page 26.

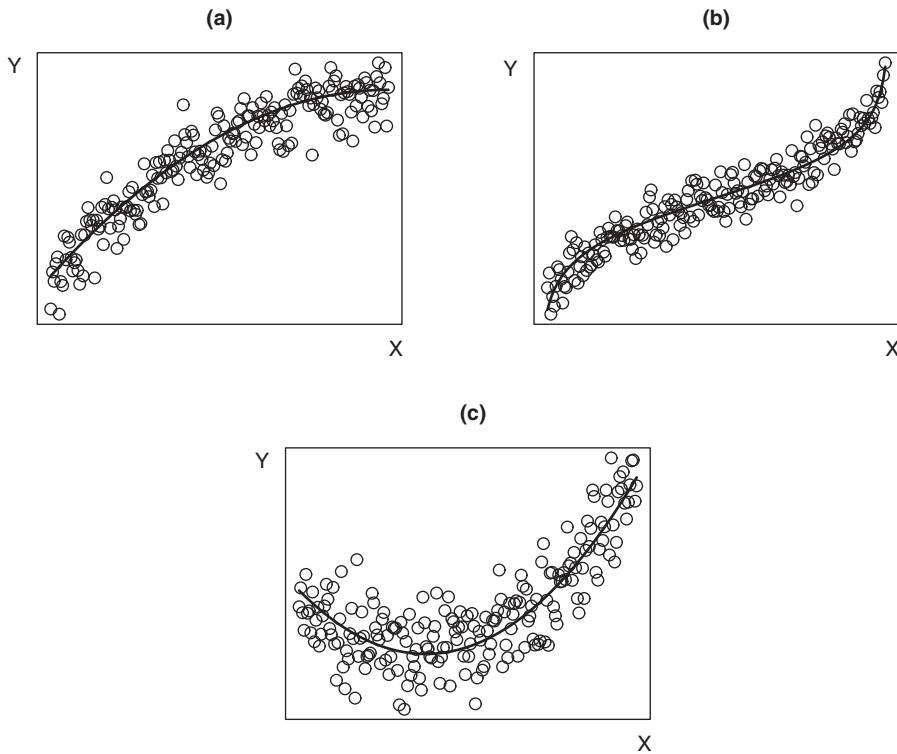


Figure 4.6 (a) A simple monotone relationship between Y and X ; (b) a monotone relationship that is not simple; (c) a relationship that is simple but not monotone. A power transformation of Y or X can straighten (a) but not (b) or (c).

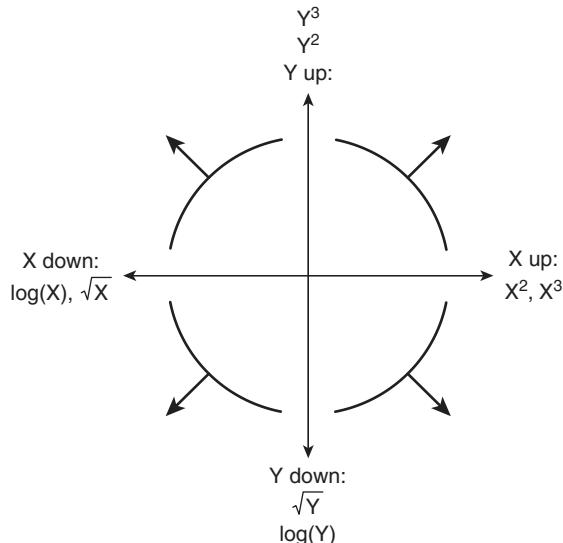


Figure 4.7 Tukey and Mosteller's bulging rule: The direction of the bulge indicates the direction of the power transformation of Y and/or X to straighten the relationship between them.

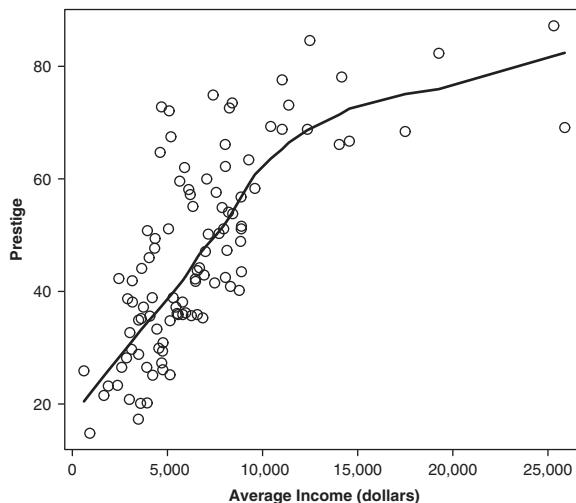


Figure 4.8 The relationship between prestige and income for the Canadian occupational prestige data. The nonparametric regression line on the plot is computed by lowess, with a span of 0.6.

with income, but the slope is steeper at the left of the plot, where income is low, than at the right, where it is high. The change in slope appears fairly abrupt rather than smooth, however, and we might do better to model the relationship with two straight lines (one for relatively small values of income, one for relatively large ones) than simply to transform prestige or income.²¹

Nevertheless, the bulge points up and to the left, and so we can try transforming prestige up the ladder of powers or income down. Because the income distribution is positively skewed, I prefer to transform income rather than prestige, which is more symmetrically distributed. As shown in Figure 4.9, the cube-root transformation of income works reasonably well here. Some nonlinearity remains, but it is not simple, and the linear regression of prestige on income no longer *grossly* distorts the relationship between the two variables. I would have preferred to use the log transformation, which makes the income distribution more symmetric and is simpler to interpret, but this transformation “overcorrects” the nonlinearity in the relationship between prestige and income.

For a more extreme and ultimately more successful example, consider the relationship between infant mortality and gross domestic product (GDP) per capita, shown in Figure 4.10 and first discussed in Chapter 3.²² As I pointed out previously, both variables are highly positively skewed and, consequently, most of the data are confined to a small region at the lower left of the plot.

²¹For an alternative interpretation of the relationship between prestige and income, plot the data using different symbols for different types of occupations. (The data set distinguishes among blue-collar, white-collar, and professional and managerial occupations.)

²²Repeating Figure 3.14 on page 45. This example is motivated by a discussion of similar data in Leinhardt and Wasserman (1979).

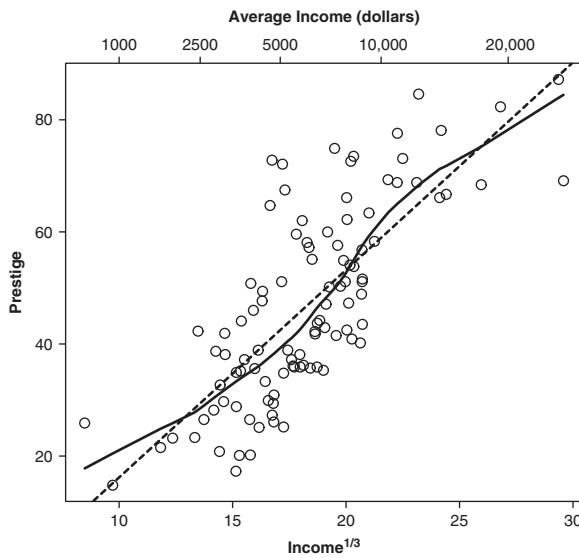


Figure 4.9 Scatterplot of prestige versus income^{1/3}. The broken line shows the linear least-squares regression, while the solid line shows the lowess smooth, with a span of 0.6. The original income scale is shown at the top of the graph.

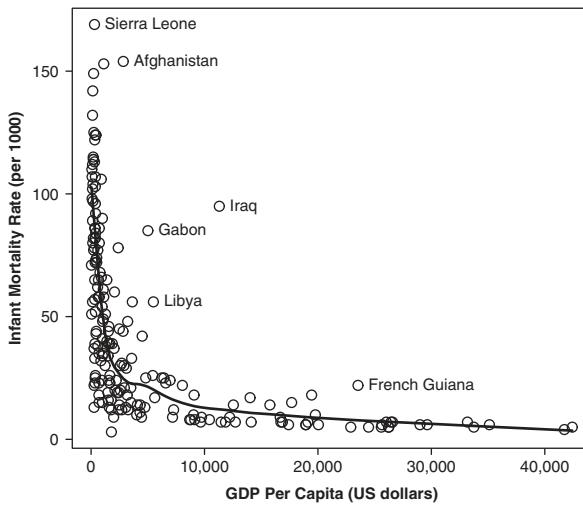


Figure 4.10 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

The skewness of infant mortality and income in Figure 4.10 makes the scatterplot difficult to interpret; despite this fact, the nonparametric regression shown on the plot reveals a highly

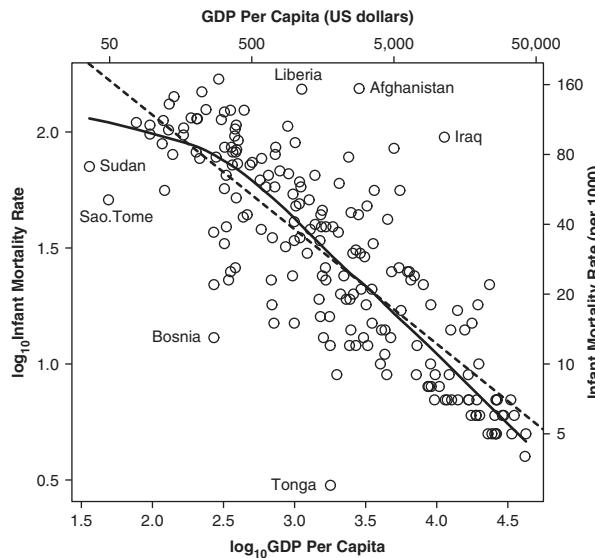


Figure 4.11 Scatterplot of \log_{10} infant mortality rate versus \log_{10} per-capita GDP. The broken line was calculated by linear least-squares regression and the solid line by lowess with a span of 1/2. The original scales of the variables appear at the top and to the right.

nonlinear but monotone relationship between infant mortality and income. The bulging rule suggests that infant mortality or income should be transformed down the ladder of powers and roots. In this case, transforming both variables by taking logs makes the relationship nearly linear (as shown in Figure 4.11). Moreover, although several countries still stand out as having relatively high infant mortality for their GDP, others now are revealed to have relatively *low* infant mortality in comparison to countries with similar GDP.

The least-squares regression line in Figure 4.11 has the equation

$$\widehat{\log_{10} \text{Infant mortality}} = 3.06 - 0.493 \times \log_{10} \text{GDP}$$

Because both variables are expressed on log scales to the same base, the slope of this relationship has a simple interpretation: A 1% increase in per-capita income is associated, on average, with an approximate 0.49% decline in the infant mortality rate. Economists call this type of coefficient an “elasticity.”²³

²³Increasing X by 1% is equivalent to multiplying it by 1.01, which in turn implies that the log of X increases by $\log_{10} 1.01 = 0.00432$. The corresponding change in $\log Y$ is then $B \times 0.00432 = -0.493 \times 0.00432 = -0.00213$. Subtracting 0.00213 from $\log Y$ is equivalent to multiplying Y by $10^{-0.00213} = 0.99511$, that is, decreasing Y by $100 \times (1 - 0.99511) = 0.489 \approx B$. The approximation holds because the log function is nearly linear across the small domain of X -values between $\log 1$ and $\log 1.01$.

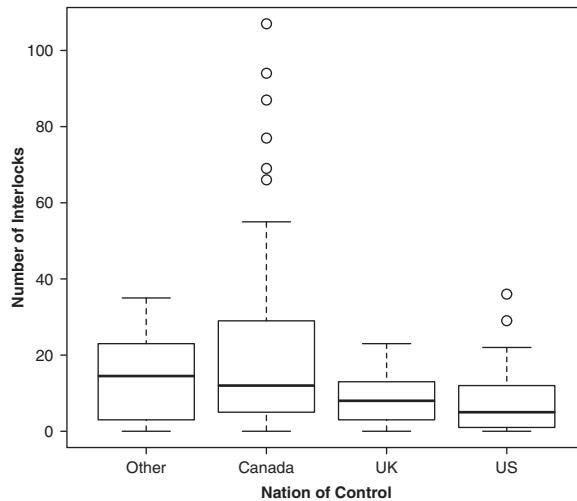


Figure 4.12 Number of interlocking directorate and executive positions by nation of control, for 248 dominant Canadian firms.

4.4 Transforming Nonconstant Spread

When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in level across the groups. We encountered this problem in the preceding chapter, where we compared the distribution of the number of interlocking directorships by nation of control, employing Ornstein's data on 248 dominant Canadian corporations, shown in Figure 4.12.²⁴

Differences in spread are often systematically related to differences in level: Groups with higher levels tend to have higher spreads. Using the median and hinge-spread as indices of level and spread, respectively, the following table shows that there is indeed an association, if only an imperfect one, between spread and level for Ornstein's data:

Nation of Control	Lower Hinge	Median	Upper Hinge	Hinge Spread
Other	3	14.5	23	20
Canada	5	12.0	29	24
United Kingdom	3	8.0	13	10
United States	1	5.0	12	11

Tukey (1977) suggests graphing the log hinge-spread against the log median, as shown in Figure 4.13. Because some firms maintained 0 interlocks, I used a start of 1 to construct this graph, which has the effect of adding 1 to each median but leaves the hinge-spreads unchanged.

²⁴Repeating Figure 3.17 on page 47.

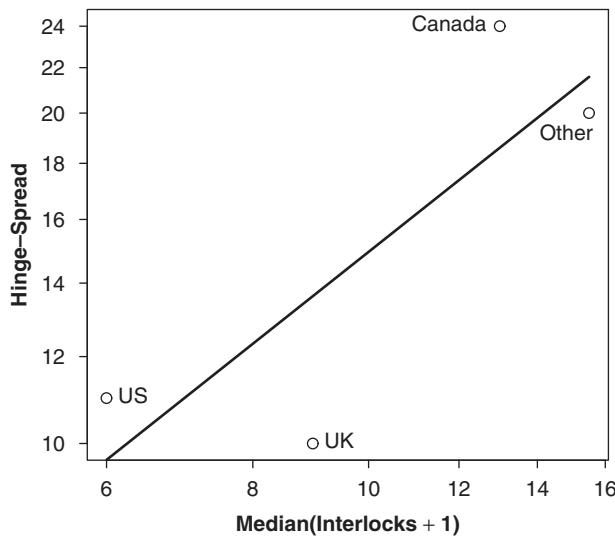


Figure 4.13 Spread (log hinge-spread) versus level [$\log(\text{median} + 1)$]. The plot is for Ornstein's interlocking-directorate data, with groups defined by nation of control. The line on the plot was fit by least squares.

The slope of the linear “trend,” if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data: Express the linear fit as

$$\log \text{spread} \approx a + b \log \text{level}$$

Then the corresponding spread-stabilizing transformation uses the power $p = 1 - b$. When spread is positively related to level (i.e., $b > 0$), therefore, we select a transformation *down* the ladder of powers and roots.

When there is a positive association between the level of a variable in different groups and its spread, the spreads can be made more constant by descending the ladder of powers. A negative association between level and spread is less common but can be corrected by ascending the ladder of powers.

In Figure 4.13, a line was fit by least squares to the spread-level plot for the interlocking directorate data. The slope of this line, $b = 0.85$, suggests the power transformation $p = 1 - 0.85 = 0.15 \approx 0$. I decided, therefore, to try a log transformation. Figure 4.14 shows the result, employing logs to the base 2.²⁵ The spreads of the several groups are now much more similar, and differences in level are easier to discern. The within-group distributions are more symmetric as well, and there are no outliers.

²⁵Recall that increasing $\log_2 X$ by 1 represents doubling X (where, here, X is the number of interlocks plus 1).

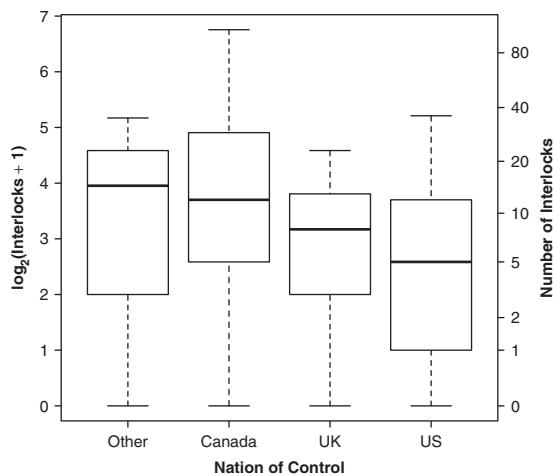


Figure 4.14 Parallel boxplots of number of interlocks by nation of control, transforming interlocks+1 to the \log_2 scale. Compare this plot with Figure 4.12, where number of interlocks is not transformed. The original scale for number of interlocks is shown at the right.

The problems of unequal spread and skewness commonly occur together because they often have a common origin. When, as here, the data represent frequency counts (*number of interlocks*), the impossibility of obtaining a negative count tends to produce positive skewness, together with a tendency for larger levels to be associated with larger spreads. The same is true of other types of variables that are bounded below (e.g., wage and salary income). Likewise, variables that are bounded above but not below (e.g., grades on a very simple exam) tend both to be negatively skewed and to show a negative association between spread and level. In the latter event, a transformation “up” the ladder of powers (e.g., to X^2) usually provides a remedy.²⁶

4.5 Transforming Proportions

Power transformations are often unhelpful for proportions because these quantities are bounded below by 0 and above by 1. Of course, if the data values do not approach the two boundaries, then proportions can be handled much like other sorts of data.

Percentages and many sorts of rates (e.g., infant mortality rate per 1,000 live births) are simply rescaled proportions and, therefore, are similarly affected. It is, moreover, common to encounter “disguised” proportions, such as the number of questions correct on an exam of fixed length or the number of affirmative responses to a series of dichotomous attitude questions.

²⁶Plotting log spread against log level to select a spread-stabilizing transformation is quite a general idea. In Section 12.2, for example, we will use a version of the spread-level plot to find a variance-stabilizing transformation in regression analysis.

1 2:	represents 12
	leaf unit: 1
	n: 102
32	0*
44	0.
(8)	1*
50	1.
43	2*
39	2.
37	3*
32	3.
	4*
30	4.
27	5*
24	5.
22	6*
21	6.
18	7*
15	7.
11	8*
	8.
8	9*
	012

Figure 4.15 Stem-and-leaf display of percentage of women in each of 102 Canadian occupations in 1970. Note how the data “stack up” against both boundaries.

An example, drawn from the Canadian occupational prestige data, is shown in the stem-and-leaf display in Figure 4.15. The distribution is for the percentage of women among the incumbents of each of 102 occupations. There are many occupations with no women or a very small percentage of women, but the distribution is not simply positively skewed, because there are also occupations that are predominantly female. In contrast, relatively few occupations are balanced with respect to their gender composition.

Several transformations are commonly employed for proportions, P , including the following:

- The *logit* transformation,

$$P \rightarrow \text{logit}(P) = \log_e \frac{P}{1 - P}$$

The logit transformation is the log of the “odds,” $P/(1 - P)$. The “trick” of the logit transformation is to remove the upper and lower boundaries of the scale, spreading out the tails of the distribution and making the resulting quantities symmetric about 0; for example,

P	$\frac{P}{1 - P}$	logit
.01	1/99	-4.59
.05	1/19	-2.94
.1	1/9	-2.20
.3	3/7	-0.85
.5	1	0
.7	7/3	0.85
.9	9/1	2.20
.95	19/1	2.94
.99	99/1	4.59

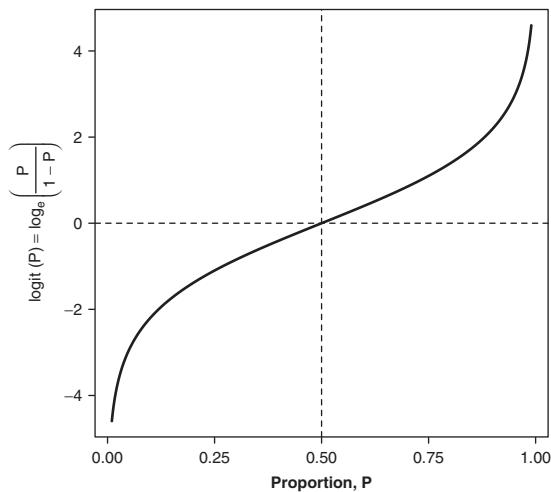


Figure 4.16 The logit transformation $\log_e[P/(1 - P)]$ of a proportion P .

A graph of the logit transformation, shown in Figure 4.16, reveals that the transformation is nearly linear in its center, between about $P = .2$ and $P = .8$.

- The *probit* transformation,

$$P \rightarrow \text{probit}(P) = \Phi^{-1}(P)$$

where Φ^{-1} is the inverse distribution function (i.e., the quantile function) for the standard normal distribution. Once their scales are equated, the logit and probit transformations are, for practical purposes, indistinguishable: $\text{logit} \approx (\pi/\sqrt{3}) \times \text{probit}$.²⁷

- The *arcsine-square-root* transformation also has a similar shape:

$$P \rightarrow \sin^{-1} \sqrt{P}$$

Tukey (1977) has embedded these common transformations for proportions into the family of “folded” powers and roots, indexed by the power q , which takes on values between 0 and 1:

$$P \rightarrow P^q - (1 - P)^q$$

When $q = 0$, we take the natural log, producing the logit transformation. Setting $q = 0.14$ yields (to a very close approximation) a multiple of the probit transformation. Setting $q = 0.41$ produces (again, to a close approximation) a multiple of the arcsine-square-root transformation. When $q = 1$, the transformation is just twice the “plurality” (i.e., the difference between P and $\frac{1}{2}$), leaving the shape of the distribution of P unaltered:

$$P \rightarrow P - (1 - P) = 2(P - \frac{1}{2})$$

²⁷We will encounter the logit and probit functions again in a different context when we take up the analysis of categorical data in Chapter 14.

1		2: represents 1.2
		leaf unit: 0.1
		n: 102
5	-5*	22222
8	-4.	555
16	-4*	44332111
21	-3.	98875
31	-3*	4432111000
39	-2.	98887655
48	-2*	443220000
(10)	-1.	9888666555
44	-1*	331110
38	-0.	987666
32	-0*	44110
27	0*	00122
22	0.	577889
16	1*	01111
11	1.	556
8	2*	23
6	2.	5
5	3*	00014

Figure 4.17 Stem-and-leaf display for the logit transformation of proportion of women in each of 102 Canadian occupations. Because some occupations have no women, the proportions were mapped to the interval .005 to .995 prior to calculating the logits.

Power transformations are ineffective for proportions P that simultaneously push the boundaries of 0 and 1 and for other variables (e.g., percentages, rates, disguised proportions) that are bounded both below and above. The folded powers $P \rightarrow P^q - (1 - P)^q$ are often effective in this context; for $q = 0$, we employ the logit transformation, $P \rightarrow \log_e[P/(1 - P)]$.

The logit and probit transformations cannot be applied to proportions of exactly 0 or 1. If, however, we have access to the original counts on which the proportions were based, then we can avoid this embarrassment by employing

$$P' = \frac{F + \frac{1}{2}}{N + 1}$$

in place of P . Here, F is the frequency count in the focal category (e.g., number of women) and N is the total count (total number of occupational incumbents, women plus men). If the original counts are not available, then we can use the expedient of mapping the proportions to an interval that excludes 0 and 1. For example, $P' = .005 + .99 \times P$ maps proportions to the interval [.005, .995].

Employing the latter strategy for the Canadian occupational data produces the distribution for $\text{logit}(P'_{\text{women}})$ that appears in Figure 4.17. Spreading out the tails of the distribution has improved its behavior considerably, although there is still some stacking up of low and high values.

4.6 Estimating Transformations as Parameters*

If we lived in a world in which the joint distribution of all quantitative data were multivariate-normal, then statistical analysis would be simple indeed: Outliers would be rare, all variables would be symmetrically distributed, all regressions would be linear, and least-squares regression would be a fine method of estimation. Making data as close to multivariate-normal as possible by transformation, therefore, can facilitate their analysis.

If the vector random variable $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ with population mean vector $\boldsymbol{\mu}_{(p \times 1)}$ and covariance matrix $\boldsymbol{\Sigma}_{(p \times p)}$ is multivariate-normal, then its probability density function is²⁸

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2} \sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]$$

In shorthand, $\mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For a sample of n observations, $\mathbf{X}_{(n \times p)}$, we have

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[\frac{1}{(2\pi)^{p/2} \sqrt{\det \boldsymbol{\Sigma}}} \right]^n \exp\left\{ \sum_{i=1}^n \left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right\}$$

where \mathbf{x}'_i is the i th row of \mathbf{X} . The log-likelihood for the parameters is, therefore,²⁹

$$\log_e L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = -\frac{np}{2} \log_e(2\pi) - \frac{n}{2} \log_e \det \boldsymbol{\Sigma} - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

The maximum-likelihood estimators (MLEs) of the mean and covariance matrix are, then,³⁰

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \bar{\mathbf{x}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)' \\ \hat{\boldsymbol{\Sigma}} &= \left\{ \hat{\sigma}_{jj'} \right\} = \left\{ \frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}_{j'})}{n} \right\}\end{aligned}$$

Now, suppose that \mathbf{x} is not multivariate-normal but that it can be made so by a power transformation of its elements.³¹ It is convenient to use the Box-Cox family of power transformations (Equation 4.2) because they are continuous at $p = 0$. Rather than thinking about these powers informally, let us instead consider them as additional parameters,³² $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)',$ one for each element of \mathbf{x} , so that

$$\mathbf{x}^{(\lambda)} = \left[x_1^{(\lambda_1)}, x_2^{(\lambda_2)}, \dots, x_p^{(\lambda_p)} \right]'$$

²⁸See online Appendix D on probability and estimation.

²⁹The likelihood function and maximum-likelihood estimation are described in online Appendix D on probability and estimation.

³⁰Note that the MLEs of the covariances have n rather than $n - 1$ in the denominator and consequently will be biased in small samples.

³¹This cannot be strictly correct, because Box-Cox transformations are only applicable when the elements of \mathbf{x} are positive and normal distributions are unbounded, but it may be true to a close-enough approximation. There is no guarantee, however, that \mathbf{x} can be made normal by a power transformation of its elements.

³²We will encounter this general approach again in Section 12.5 in the context of the linear regression model.

Then,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = \frac{1}{(2\pi)^{p/2} \sqrt{\det \boldsymbol{\Sigma}}} \exp \left[-\frac{1}{2} (\mathbf{x}^{(\lambda)} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(\lambda)} - \boldsymbol{\mu}) \right] \prod_{j=1}^p X_j^{\lambda_j - 1} \quad (4.3)$$

where now $\boldsymbol{\mu} = E[\mathbf{x}^{(\lambda)}]$ and $\boldsymbol{\Sigma} = V[\mathbf{x}^{(\lambda)}]$ are the mean vector and covariance matrix of the transformed variables, and $\prod_{j=1}^p X_j^{\lambda_j - 1}$ is the Jacobian of the transformation from $\mathbf{x}^{(\lambda)}$ to \mathbf{x} .³³

The log-likelihood for the model is

$$\begin{aligned} \log_e L(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}) &= -\frac{np}{2} \log_e(2\pi) - \frac{n}{2} \log_e \det \boldsymbol{\Sigma} - \frac{1}{2} \sum_{i=1}^n \left[(\mathbf{x}_i^{(\lambda)} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i^{(\lambda)} - \boldsymbol{\mu}) \right] \\ &\quad + \sum_{j=1}^p (\lambda_j - 1) \sum_{i=1}^n \log_e X_{ij} \end{aligned}$$

There is no closed-form solution for the MLEs of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, but we can find the MLEs by numerical methods. Standard errors for the estimated transformations are available in the usual manner from the inverse of the information matrix, and both Wald and likelihood-ratio tests can be formulated for the transformation parameters.

Moreover, because our real interest lies in the transformation parameters $\boldsymbol{\lambda}$, the means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$ are “nuisance” parameters; indeed, given $\hat{\boldsymbol{\lambda}}$, the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are just the sample mean vector and covariance matrix of $\mathbf{x}^{(\lambda)}$. Let us define the *modified Box-Cox family* of transformations as follows:

$$X^{[\lambda]} = \begin{cases} \tilde{X}^{1-\lambda} \frac{X^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \tilde{X} \log_e X & \text{for } \lambda = 0 \end{cases}$$

where

$$\tilde{X} = \left(\prod_{i=1}^n X_i \right)^{1/n}$$

is the *geometric mean* of X . Multiplication by $\tilde{X}^{1-\lambda}$ is a kind of standardization, equating the scales of different power transformations of X . Let $\mathbf{V}^{[\lambda]}$ represent the sample covariance matrix of

$$\mathbf{x}^{[\lambda]} \equiv [x_1^{[\lambda_1]}, x_2^{[\lambda_2]}, \dots, x_p^{[\lambda_p]}]'$$

Velilla (1993) shows that the MLEs of $\boldsymbol{\lambda}$ in Equation 4.3 are the values that minimize the determinant of $\mathbf{V}^{[\lambda]}$.

Applying this approach to the joint distribution of infant mortality and GDP per capita produces the following results:

³³See online Appendix D on probability and estimation.

	$\hat{\lambda}_j$	$SE(\hat{\lambda}_j)$	$z_0 = \frac{\hat{\lambda}_j - 1}{SE(\hat{\lambda}_j)}$	p
Infant mortality	-0.0009	0.0655	-15.28	<<.0001
GDP per capita	0.0456	0.0365	-26.14	<<.0001

The first column in this table gives the MLE of each transformation parameter, the second column gives the asymptotic standard error of the transformation, the third column gives the Wald statistic for testing the hypothesis $H_0: \lambda_j = 1$ (i.e., that no transformation is required), and the final column gives the two-sided p -value for this test. In this case, evidence for the need to transform the two variables is very strong. Moreover, both estimated transformations are very close to 0—that is, the log transformation. A likelihood-ratio test for the hypothesis $H_0: \lambda_1 = \lambda_2 = 1$ yields the chi-square test statistic $G_0^2 = 680.25$ on 2 degrees of freedom, which is also wildly statistically significant. In contrast, testing the hypothesis that $H_0: \lambda_1 = \lambda_2 = 0$ produces $G_0^2 = 1.649$ on 2 degrees of freedom, for which $p = .44$, supporting the use of the log transformations of infant mortality and GDP. We know from our previous work that these transformations make the distributions of the two variables symmetric and linearize their relationship.

Finally, we can also apply this method to individual variables to attempt to normalize their *univariate* distributions. For the current example, the individual MLEs of the power transformation parameters λ for infant mortality and GDP are similar to those reported above:

	$\hat{\lambda}$	$SE(\hat{\lambda})$	$z_0 = \frac{\hat{\lambda} - 1}{SE(\hat{\lambda})}$	p
Infant mortality	0.0984	0.0786	-11.47	<<.0001
GDP per capita	-0.0115	0.0440	-23.00	<<.0001

The method of maximum likelihood can be used to estimate normalizing power transformations of variables.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 4.1. Create a graph like Figure 4.1, but for the *ordinary* power transformations $X \rightarrow X^p$ for $p = -1, 0, 1, 2, 3$. (When $p = 0$, however, use the log transformation.) Compare your graph to Figure 4.1, and comment on the similarities and differences between the two families of transformations X^p and $X^{(p)}$.

Exercise 4.2. *Show that the derivative of $f(X) = (X^p - 1)/p$ is equal to 1 at $X = 1$ regardless of the value of p .

Exercise 4.3. *We considered starts for transformations informally to ensure that all data values are positive and that the ratio of the largest to the smallest data values is sufficiently large. An alternative is to think of the start as a parameter to be estimated along with the transformation power to make the distribution of the variable as normal as possible. This approach defines a *two-parameter Box-Cox family*:

$$X^{(\alpha, \lambda)} \equiv \frac{(X - \alpha)^\lambda}{\lambda}$$

- (a) Develop the MLEs of α and λ for the two-parameter Box-Cox family.
- (b) Attempt to apply the estimator to data. Do you encounter any obstacles? [Hint: Examine the correlation between the parameter estimates $\hat{\alpha}$ and $\hat{\lambda}$.]

Exercise 4.4. The Yeo-Johnson family of modified power transformations (Yeo & Johnson, 2000) is an alternative to using a start when both negative (or 0) and positive values are included in the data. The Yeo-Johnson family is defined as follows:

$$X \rightarrow X^{[p]} \equiv \begin{cases} (X + 1)^{(p)} & \text{for } X \geq 0 \\ (1 - X)^{(2-p)} & \text{for } X < 0 \end{cases}$$

where the parenthetical superscript (p) gives the Box-Cox power, as in Equation 4.2

- (a) Graph the transformations $X^{[p]}$ in the Yeo-Johnson family for values of X between -10 and $+10$ and powers p of $-1, -0.5, 0, 0.5, 1$, and 2 .
- (b) Now consider strictly positive X -values between 0.1 and 10 . Compare the Yeo-Johnson and Box-Cox transformations of X for powers p of $-1, -0.5, 0, 0.5, 1$, and 2 .
- (c) *As in Section 4.6 for Box-Cox transformations, derive the maximum-likelihood estimator of the Yeo-Johnson transformations to multivariate normality of a vector of X s.

Summary

- Transformations can often facilitate the examination and statistical modeling of data.
- The powers and roots are a particularly useful family of transformations: $X \rightarrow X^p$. When $p = 0$, we employ the log transformation in place of X^0 .
- Power transformations preserve the order of the data only when all values are positive and are effective only when the ratio of largest to smallest data values is itself large. When these conditions do not hold, we can impose them by adding a positive or negative start to all the data values.
- Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.
- Simple monotone nonlinearity can often be corrected by a power transformation of X , of Y , or of both variables. Mosteller and Tukey's bulging rule assists in selecting linearizing transformations.
- When there is a positive association between the level of a variable in different groups and its spread, the spreads can be made more constant by descending the ladder of

powers. A negative association between level and spread is less common but can be corrected by ascending the ladder of powers.

- Power transformations are ineffective for proportions, P , that simultaneously push the boundaries of 0 and 1 and for other variables (e.g., percentages, rates, disguised proportions) that are bounded both below and above. The folded powers $P \rightarrow P^q - (1 - P)^q$ are often effective in this context; for $q = 0$, we employ the logit transformation, $P \rightarrow \log_e[P/(1 - P)]$.
- The method of maximum likelihood can be used to estimate normalizing power transformations of variables.

Recommended Reading

Because examination and transformation of data are closely related topics, most of the readings here were also listed at the end of the previous chapter.

- Tukey's important text on exploratory data analysis (Tukey, 1977) and the companion volume by Mosteller and Tukey (1977) on regression analysis have a great deal of interesting information and many examples. As mentioned in the previous chapter, however, Tukey's writing style is opaque. Velleman and Hoaglin (1981) is easier to digest, but it is not as rich in material on transformations.
- Several papers in a volume edited by Hoaglin, Mosteller, and Tukey (1983) have valuable material on the family of power transformations, including a general paper by Emerson and Stoto, an extended discussion of the spread-versus-level plot in a paper on boxplots by Emerson and Strenio, and a more difficult paper by Emerson on the mathematics of transformations.
- The tools provided by the Lisp-Stat statistical computing environment (described in Tierney, 1990)—including the ability to associate a transformation with a slider and to link different plots—are especially helpful in selecting transformations. Cook and Weisberg (1994, 1999) have developed a system for data analysis and regression based on Lisp-Stat that includes these capabilities. Similar facilities are built into some statistical packages and can be implemented in other statistical computing environments (such as R).

PART II

Linear Models and Least Squares

5

Linear Least-Squares Regression

On several occasions in the first part of the text, I emphasized the limitations of linear least-squares regression. Despite these limitations, linear least squares lies at the very heart of applied statistics:¹

- Some data are adequately summarized by linear least-squares regression.
- The effective application of linear regression is considerably expanded through data transformations and techniques for diagnosing problems such as nonlinearity and overly influential data.
- As we will see, the general linear model—a direct extension of linear least-squares regression—is able to accommodate a very broad class of specifications, including, for example, qualitative explanatory variables and polynomial and other nonlinear functions of quantitative explanatory variables.
- Linear least-squares regression provides a computational basis for a variety of generalizations, including weighted least-squares regression, robust regression, nonparametric regression, and generalized linear models.

Linear least-squares regression and the closely related topic of linear statistical models are developed in this chapter and in Chapters 6 through 10:

- The current chapter describes the mechanics of linear least-squares regression. That is, I will explain how the method of least squares can be employed to fit a line to a bivariate scatterplot, a plane to a three-dimensional scatterplot, and a general linear surface to multivariate data (which, of course, cannot be directly visualized).
- Chapter 6 develops general and flexible methods of statistical inference for linear models.
- Chapters 7 and 8 extend linear models to situations in which some or all of the explanatory variables are qualitative and categorical rather than quantitative.
- Chapter 9 casts the linear model in matrix form and describes the statistical theory of linear models more formally and more generally.
- Chapter 10 introduces the vector geometry of linear models, a powerful tool for conceptualizing linear models and least-squares estimation.²

¹The extensions of linear least-squares regression mentioned here are the subject of subsequent chapters.

²Chapters 9 and 10 are “starred” (i.e., marked with asterisks) and therefore are more difficult; like all starred material in this book, these chapters can be skipped without loss of continuity, although some of the later starred material depends on earlier starred text.

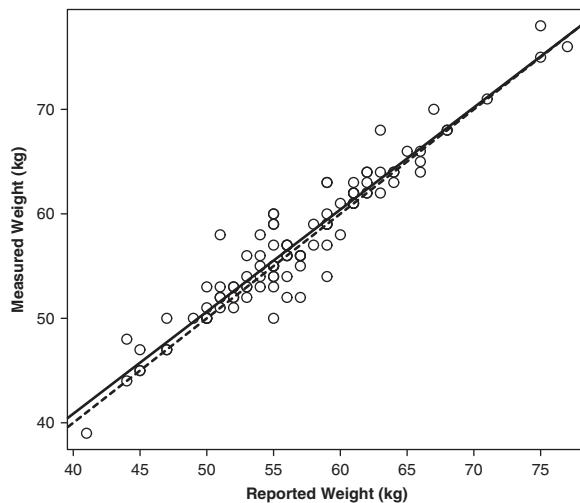


Figure 5.1 Scatterplot of Davis’s data on the measured and reported weight of 101 women. The solid line gives the least-squares fit; the broken line is $Y = X$. Because weight is given to the nearest kilogram, both variables are discrete, and some points are overplotted.

5.1 Simple Regression

5.1.1 Least-Squares Fit

Figure 5.1 shows Davis’s data, introduced in Chapter 2, on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.³ The relationship between measured and reported weight appears to be linear, so it is reasonable to fit a line to the plot. A line will help us determine whether the subjects in Davis’s study were accurate and unbiased reporters of their weights, and it can provide a basis for predicting the measured weight of similar women for whom only reported weight is available.

Denoting measured weight by Y and reported weight by X , a line relating the two variables has the equation $Y = A + BX$.⁴ It is obvious, however, that no line can pass perfectly through all the data points, despite the strong linear relationship between these two variables. We introduce a *residual*, E , into the regression equation to reflect this fact; writing the regression equation for the i th of the $n = 101$ observations:

$$\begin{aligned} Y_i &= A + BX_i + E_i \\ &= \hat{Y}_i + E_i \end{aligned} \tag{5.1}$$

where $\hat{Y}_i = A + BX_i$ is the *fitted value* for observation i . The essential geometry is shown in Figure 5.2, which reveals that the residual

³The misrecorded data value that produced an outlier in Figure 2.5 on page 19 has been corrected.

⁴See online Appendix C for a review of the geometry of lines and planes.

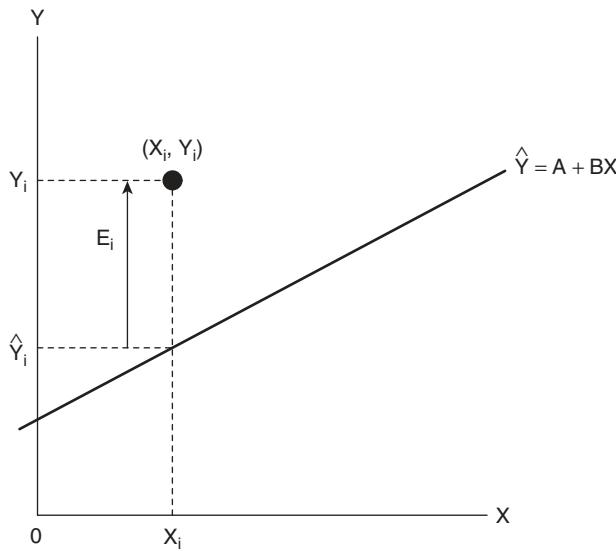


Figure 5.2 Linear regression of Y on X , showing the residual E_i for the i th observation.

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$$

is the signed vertical distance between the point and the line—that is, the residual is negative when the point lies below the line and positive when the point is above the line [as is the point (X_i, Y_i) in Figure 5.2].

A line that fits the data well therefore makes the residuals small, but to determine a line analytically, we need to be more precise about what we mean by “small.” First of all, we want residuals that are small in magnitude, because large negative residuals are as offensive as large positive ones. For example, simply requiring that the sum of residuals, $\sum_{i=1}^n E_i$, be small is futile, because large negative residuals can offset large positive ones.

Indeed, any line through the means of the variables—the point (\bar{X}, \bar{Y}) —has $\sum E_i = 0$. Such a line satisfies the equation $\bar{Y} = A + B\bar{X}$. Subtracting this equation from Equation 5.1 produces

$$Y_i - \bar{Y} = B(X_i - \bar{X}) + E_i$$

Then, summing over all observations,

$$\sum_{i=1}^n E_i = \sum(Y_i - \bar{Y}) - B \sum(X_i - \bar{X}) = 0 - B \times 0 = 0 \quad (5.2)$$

Two possibilities immediately present themselves: We can employ the unsigned vertical distances between the points and the line, that is, the absolute values of the residuals, or we can employ the squares of the residuals. The first possibility leads to *least-absolute-value (LAV) regression*:

Find A and B to minimize the sum of the absolute residuals, $\sum |E_i|$.

The second possibility leads to the *least-squares criterion*:

Find A and B to minimize the sum of squared residuals, $\sum E_i^2$.

Squares are more tractable mathematically than absolute values, so we will focus on least squares here, but LAV regression should not be rejected out of hand, because it provides greater resistance to outlying observations.⁵

We need to consider the residuals in the aggregate, because it is no trick to produce a 0 residual for an individual point simply by placing the line directly through the point. The least-squares criterion therefore minimizes the *sum* of squared residuals over all observations; that is, we seek the values of A and B that minimize

$$S(A, B) = \sum_{i=1}^n E_i^2 = \sum (Y_i - A - BX_i)^2$$

I have written this expression as a *function* $S(A, B)$ of the regression coefficients A and B to emphasize the dependence of the sum of squared residuals on the coefficients: For a fixed set of data $\{X_i, Y_i\}$, $i = 1, \dots, n$, each possible choice of values for A and B corresponds to a specific residual sum of squares, $\sum E_i^2$; we want the pair of values for the regression coefficients that makes this sum of squares as small as possible.

*The most direct approach to finding the least-squares coefficients is to take the partial derivatives of the sum-of-squares function with respect to the coefficients:⁶

$$\begin{aligned}\frac{\partial S(A, B)}{\partial A} &= \sum (-1)(2)(Y_i - A - BX_i) \\ \frac{\partial S(A, B)}{\partial B} &= \sum (-X_i)(2)(Y_i - A - BX_i)\end{aligned}$$

Setting the partial derivatives to 0 yields simultaneous linear equations for the least-squares coefficients, A and B .⁷

Simultaneous linear equations for the least-squares coefficients A and B , the so-called *normal equations*⁸ for simple regression, are

$$\begin{aligned}An + B \sum X_i &= \sum Y_i \\ A \sum X_i + B \sum X_i^2 &= \sum X_i Y_i\end{aligned}$$

where n is the number of observations. Solving the normal equations produces the least-squares coefficients:

$$\begin{aligned}A &= \bar{Y} - B\bar{X} \\ B &= \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\end{aligned}\tag{5.3}$$

The formula for A implies that the least-squares line passes through the point of means of the two variables. By Equation 5.2, therefore, the least-squares residuals sum to 0. The second normal equation implies that $\sum X_i E_i = 0$, for

⁵We will return to LAV regression in Chapter 19, which discusses robust regression.

⁶In Chapter 10, I will derive the least-squares solution by an alternative geometric approach.

⁷As a formal matter, it remains to be shown that the solution of the normal equations *minimizes* the least-squares function $S(A, B)$. See Section 9.2.

⁸The term *normal* here refers not to the normal distribution but to orthogonality (perpendicularity); see Chapter 10 on the vector geometry of regression.

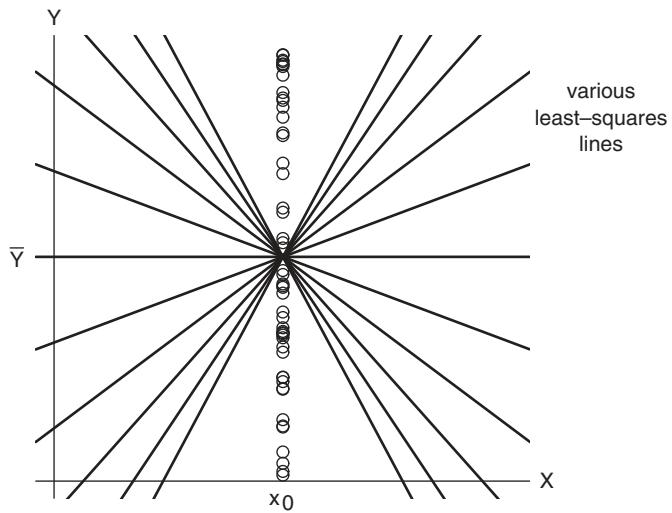


Figure 5.3 When all values of X are the same (x_0), any line through the point (x_0, \bar{Y}) is a least-squares line.

$$\sum X_i E_i = \sum X_i(Y_i - A - BX_i) = \sum X_i Y_i - A \sum X_i - B \sum X_i^2 = 0$$

Similarly, $\sum \hat{Y}_i E_i = 0$.⁹ These properties, which will be useful to us below, imply that the least-squares residuals are uncorrelated with both the explanatory variable X and the fitted values \hat{Y} .¹⁰

It is clear from Equations 5.3 that the least-squares coefficients are uniquely defined as long as the explanatory-variable values are not all identical, for when there is no variation in X , the denominator of B vanishes. This result is intuitively plausible: Only if the explanatory-variable scores are spread out can we hope to fit a (unique) line to the X, Y scatter; if, alternatively, all the X -values are the same (say, equal to x_0), then, as is shown in Figure 5.3, any line through the point (x_0, \bar{Y}) is a least-squares line.

I will illustrate the least-squares calculations using Davis's data on measured weight (Y) and reported weight (X), for which

$$\begin{aligned} n &= 101 \\ \bar{Y} &= \frac{5780}{101} = 57.228 \\ \bar{X} &= \frac{5731}{101} = 56.743 \\ \sum (X_i - \bar{X})(Y_i - \bar{Y}) &= 4435.9 \\ \sum (X_i - \bar{X})^2 &= 4539.3 \\ B &= \frac{4435.9}{4539.3} = 0.97722 \\ A &= 57.228 - 0.97722 \times 56.743 = 1.7776 \end{aligned}$$

⁹See Exercise 5.1.

¹⁰See the next section for a definition of correlation.

Thus, the least-squares regression equation is

$$\widehat{\text{Measured weight}} = 1.78 + 0.977 \times \text{Reported weight}$$

Interpretation of the least-squares slope coefficient is straightforward: $B = 0.977$ indicates that a 1-kg increase in reported weight is associated, on average, with just under a 1-kg increase in measured weight. Because the data are not longitudinal, the phrase “a unit increase” here implies not a literal change over time but rather a notional static comparison between two individuals who differ by 1 kg in their reported weights.

Ordinarily, we may interpret the intercept A as the fitted value associated with $X = 0$, but it is, of course, impossible for an individual to have a reported weight equal to 0. The intercept A is usually of little direct interest, because the fitted value above $X = 0$ is rarely important. Here, however, if individuals’ reports are unbiased predictions of their actual weights, then we should have the equation $\widehat{Y} = X$ —that is, an intercept of 0 and a slope of 1. The intercept $A = 1.78$ is indeed close to 0, and the slope $B = 0.977$ is close to 1.

In simple linear regression

$$Y_i = A + BX_i + E_i$$

the least-squares coefficients are given by $A = \bar{Y} - B\bar{X}$ and $B = \sum(X_i - \bar{X})(Y_i - \bar{Y}) / \sum(X_i - \bar{X})^2$. The slope coefficient B represents the average change in Y associated with a one-unit increase in X . The intercept A is the fitted value of Y when $X = 0$.

5.1.2 Simple Correlation

Having calculated the least-squares line, it is of interest to determine how closely the line fits the scatter of points. This is a vague question, which may be answered in a variety of ways. The standard deviation of the residuals, S_E , often called the *standard error of the regression* or the *residual standard error*, provides one sort of answer.¹¹ Because of estimation considerations, the variance of the residuals is defined using *degrees of freedom* $n - 2$, rather than the sample size n , in the denominator.¹²

$$S_E^2 = \frac{\sum E_i^2}{n - 2}$$

The residual standard error is, therefore,

$$S_E = \sqrt{\frac{\sum E_i^2}{n - 2}}$$

¹¹The term *standard error* is usually used for the estimated standard deviation of the sampling distribution of a statistic, and so the use here to denote the standard deviation of the residuals is potentially misleading. This usage is common, however, and I therefore adopt it.

¹²Estimation is discussed in the next chapter. Also see the discussion in Section 10.3.

Because it is measured in the units of the response variable and represents a type of “average” residual, the standard error is simple to interpret. For example, for Davis’s regression of measured weight on reported weight, the sum of squared residuals is $\sum E_i^2 = 418.87$, and thus the standard error of the regression is

$$S_E = \sqrt{\frac{418.87}{101 - 2}} = 2.0569\text{kg}$$

On average, then, using the least-squares regression line to predict measured weight from reported weight results in an error of about 2 kg, which is small but perhaps not negligible. Moreover, if the residuals are approximately normally distributed, then about 2/3 of them are in the range ± 2 , and about 95% are in the range ± 4 . I believe that social scientists overemphasize correlation (described immediately below) and pay insufficient attention to the standard error of the regression as an index of fit.

In contrast to the standard error of the regression, the *correlation coefficient* provides a *relative* measure of fit: To what degree do our predictions of Y improve when we base these predictions on the linear relationship between Y and X ? A relative index of fit requires a baseline—how well can Y be predicted if X is disregarded?

To disregard the explanatory variable is implicitly to fit the equation $\hat{Y}'_i = A'$ or, equivalently,

$$Y_i = A' + E'_i$$

By ignoring the explanatory variable, we lose our ability to differentiate among the observations; as a result, the fitted values are constant. The constant A' is generally different from the intercept A of the least-squares line, and the residuals E'_i are different from the least-squares residuals E_i .

How should we find the best constant A' ? An obvious approach is to employ a least-squares fit—that is, to minimize

$$S(A') = \sum E'^2_i = \sum (Y_i - A')^2$$

As you may be aware, the value of A' that minimizes this sum of squares is simply the response-variable mean, \bar{Y} .¹³

The residuals $E_i = Y_i - \hat{Y}_i$ from the linear regression of Y on X will mostly be smaller in magnitude than the residuals $E'_i = Y_i - \bar{Y}$, and it is necessarily the case that

$$\sum (Y_i - \hat{Y}_i)^2 \leq \sum (Y_i - \bar{Y})^2$$

This inequality holds because the “null model,” $Y_i = A' + E'_i$, specifying no relationship between Y and X , is a special case of the more general linear regression “model,” $Y_i = A + BX_i + E_i$: The two models are the same when $B = 0$.¹⁴ The null model therefore cannot have a smaller sum of squared residuals. After all, the least-squares coefficients A and B are selected precisely to minimize $\sum E_i^2$, so constraining $B = 0$ cannot improve the fit and will usually make it worse.

¹³See Exercise 5.3.

¹⁴A formal statistical model for linear regression is introduced in the next chapter.

We call

$$\sum E_i'^2 = \sum (Y_i - \bar{Y})^2$$

the *total sum of squares* for Y , abbreviated TSS, while

$$\sum E_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

is called the *residual sum of squares* and is abbreviated RSS. The difference between the two, termed the *regression sum of squares*,

$$\text{RegSS} \equiv \text{TSS} - \text{RSS}$$

gives the reduction in squared error due to the linear regression. The ratio of RegSS to TSS, the proportional reduction in squared error, defines the square of the correlation coefficient:

$$r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

To find the *correlation coefficient* r , we take the positive square root of r^2 when the simple-regression slope B is positive and the negative square root when B is negative.

Thus, if there is a perfect positive linear relationship between Y and X (i.e., if all of the residuals are 0 and $B > 0$), then $r = 1$. A perfect negative linear relationship corresponds to $r = -1$. If there is no linear relationship between Y and X , then $\text{RSS} = \text{TSS}$, $\text{RegSS} = 0$, and $r = 0$. Between these extremes, r gives the direction of the linear relationship between the two variables, and r^2 can be interpreted as the proportion of the total variation of Y that is “captured” by its linear regression on X . Figure 5.4 illustrates several levels of correlation. As is clear in Figure 5.4(b), where $r = 0$, the correlation can be small even when there is a strong *nonlinear* relationship between X and Y .

It is instructive to examine the three sums of squares more closely: Starting with an individual observation, we have the identity

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

This equation is interpreted geometrically in Figure 5.5. Squaring both sides of the equation and summing over observations produces

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 + 2 \sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

The last term in this equation is 0,¹⁵ and thus the regression sum of squares, which I previously defined as the difference $\text{TSS} - \text{RSS}$, may also be written directly as

$$\text{RegSS} = \sum (\hat{Y}_i - \bar{Y})^2$$

This decomposition of total variation into “explained” and “unexplained” components, paralleling the decomposition of each observation into a fitted value and a residual, is typical of linear models. The decomposition is called the *analysis of variance* for the regression: $\text{TSS} = \text{RegSS} + \text{RSS}$.

Although I have developed the correlation coefficient from the regression of Y on X , it is also possible to define r by analogy with the correlation $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ between two random

¹⁵See Exercise 5.1 and Section 10.1.

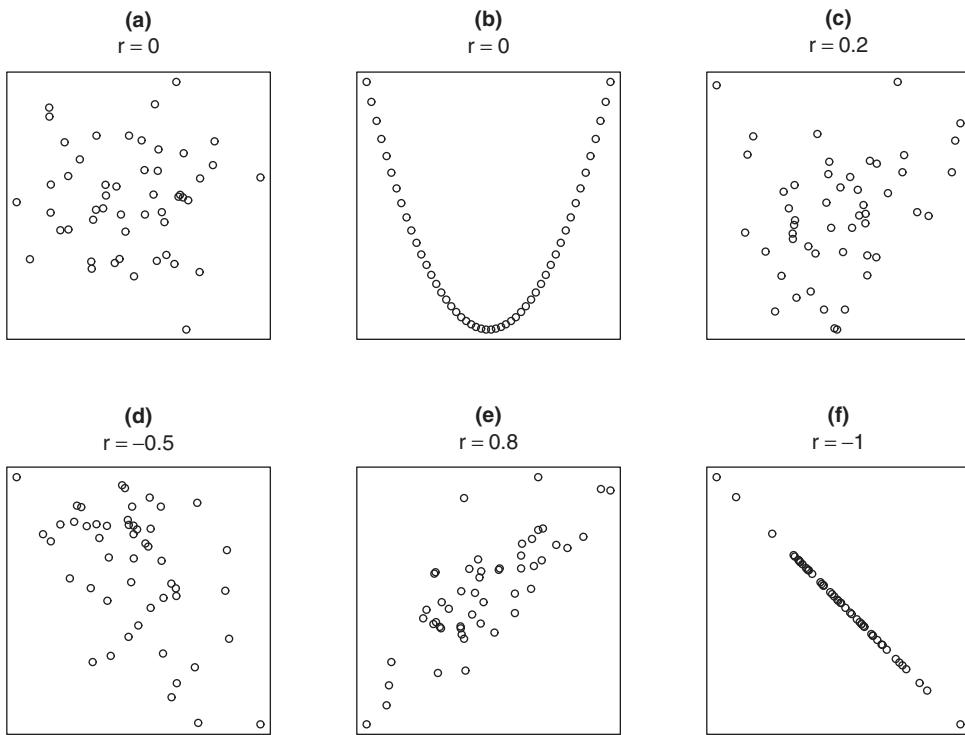


Figure 5.4 Scatterplots illustrating different levels of correlation: $r = 0$ in both (a) and (b), $r = .2$ in (c), $r = -.5$ in (d), $r = .8$ in (e), and $r = -1$ in (f). All the data sets have $n = 50$ observations. Except in panel (b), the data were generated by sampling from bivariate normal distributions.

variables (where σ_{XY} is the covariance of the random variables X and Y , σ_X is the standard deviation of X , and σ_Y is the standard deviation of Y).¹⁶ First defining the *sample covariance* between X and Y ,

$$S_{XY} \equiv \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

we may then write

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \quad (5.4)$$

where S_X and S_Y are, respectively, the sample standard deviations of X and Y .¹⁷

It is immediately apparent from the symmetry of Equation 5.4 that the correlation does not depend on which of the two variables is treated as the response variable. This property of r is surprising in light of the *asymmetry* of the regression equation used to define the sums of

¹⁶See online Appendix D on probability and estimation.

¹⁷The equivalence of the two formulas for r is established in Section 10.1 on the geometry of simple regression analysis.

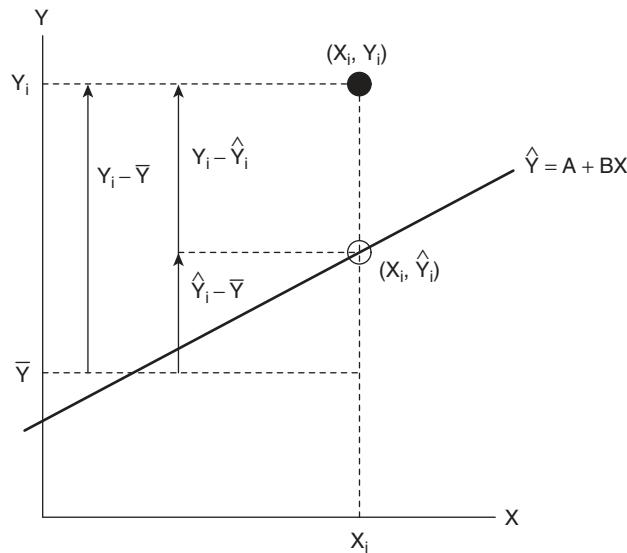


Figure 5.5 Decomposition of the total deviation $Y_i - \bar{Y}$ into components $Y_i - \hat{Y}_i$ and $\hat{Y}_i - \bar{Y}$.

squares: Unless there is a perfect correlation between the two variables, the least-squares line for the regression of Y on X differs from the line for the regression of X on Y .¹⁸

There is another central property, aside from symmetry, that distinguishes the correlation coefficient r from the regression slope B . The slope coefficient B is measured in the units of the response variable per unit of the explanatory variable. For example, if dollars of income are regressed on years of education, then the units of B are dollars/year. The correlation coefficient r , however, is unitless, as can be seen from either of its definitions. As a consequence, a change in scale of Y or X produces a compensating change in B but does not affect r . If, for example, income is measured in thousands of dollars rather than in dollars, the units of the slope become \$1,000s/year, and the value of the slope decreases by a factor of 1,000, but r remains the same.¹⁹

For Davis's regression of measured on reported weight,

$$\text{TSS} = 4753.8$$

$$\text{RSS} = 418.87$$

$$\text{RegSS} = 4334.9$$

Thus,

$$r^2 = \frac{4334.9}{4753.8} = .91188$$

¹⁸See Exercise 5.2.

¹⁹A peculiarity of the regression of measured on reported weight is that both Y and X are measured in kilograms. As a consequence, the units of B (kg/kg) cancel, and if both X and Y were rescaled to other units—such as pounds—the value of B would be unchanged. The general lesson remains the same, however: Interpret regression coefficients in relation to the units of measurement of the variables.

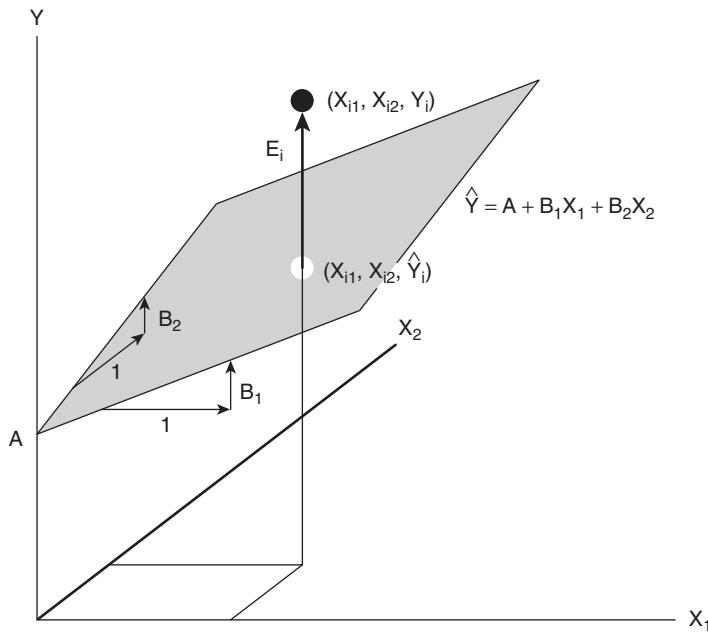


Figure 5.6 The multiple-regression plane, showing the partial slopes B_1 and B_2 and the residual E_i for the i th observation. The white dot in the regression plane represents the fitted value. Compare this graph with Figure 5.2 for simple regression.

and, because B is positive, $r = +\sqrt{.91188} = .9549$. The linear regression of measured on reported weight, therefore, captures 91% of the variation in measured weight. Equivalently,

$$\begin{aligned} S_{XY} &= \frac{4435.9}{101 - 1} = 44.359 \\ S_X^2 &= \frac{4539.3}{101 - 1} = 45.393 \\ S_Y^2 &= \frac{4753.8}{101 - 1} = 47.538 \\ r &= \frac{44.359}{\sqrt{45.393 \times 47.538}} = .9549 \end{aligned}$$

5.2 Multiple Regression

5.2.1 Two Explanatory Variables

The linear multiple-regression equation

$$\hat{Y} = A + B_1X_1 + B_2X_2$$

for two explanatory variables, X_1 and X_2 , describes a plane in the three-dimensional $\{X_1, X_2, Y\}$ space, as shown in Figure 5.6. As in simple regression, it is unreasonable to expect

that the multiple-regression regression plane will pass precisely through every point, so the fitted value \hat{Y}_i for observation i in general differs from the observed value Y_i . The residual is the signed vertical distance from the point to the plane:

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + B_1X_{i1} + B_2X_{i2})$$

To make the plane come as close as possible to the points in the aggregate, we want the values of A , B_1 , and B_2 that minimize the sum of squared residuals:

$$S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1X_{i1} - B_2X_{i2})^2$$

*As in simple regression, we can proceed by differentiating the sum-of-squares function with respect to the regression coefficients:

$$\frac{\partial S(A, B_1, B_2)}{\partial A} = \sum (-1)(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_1} = \sum (-X_{i1})(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_2} = \sum (-X_{i2})(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$

Setting the partial derivatives to 0 and rearranging terms produces the normal equations for the regression coefficients A , B_1 , and B_2 .

The normal equations for the regression coefficients A , B_1 , and B_2 are

$$\begin{aligned} An + B_1 \sum X_{i1} + B_2 \sum X_{i2} &= \sum Y_i \\ A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1}X_{i2} &= \sum X_{i1}Y_i \\ A \sum X_{i2} + B_1 \sum X_{i2}X_{i1} + B_2 \sum X_{i2}^2 &= \sum X_{i2}Y_i \end{aligned} \tag{5.5}$$

Because Equations 5.5 are a system of three linear equations in three unknowns, they usually provide a unique solution for the least-squares regression coefficients A , B_1 , and B_2 . We can write out the solution explicitly, if somewhat tediously: Dropping the subscript i for observations, and using asterisks to denote variables in mean deviation form (e.g., $Y^* \equiv Y_i - \bar{Y}$),

$$\begin{aligned} A &= \bar{Y} - B_1\bar{X}_1 - B_2\bar{X}_2 \\ B_1 &= \frac{\sum X_1^* Y^* \sum X_2^{*2} - \sum X_2^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2} \\ B_2 &= \frac{\sum X_2^* Y^* \sum X_1^{*2} - \sum X_1^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2} \end{aligned} \tag{5.6}$$

The denominator of B_1 and B_2 is nonzero—and, therefore, the least-squares coefficients are uniquely defined—as long as

$$\sum X_1^{*2} \sum X_2^{*2} \neq (\sum X_1^* X_2^*)^2$$

This condition is satisfied unless X_1 and X_2 are perfectly correlated or unless one or both of the explanatory variables are invariant.²⁰ If X_1 and X_2 are perfectly correlated, then they are said to be *collinear*.

To illustrate the computation of multiple-regression coefficients, I will employ Duncan's occupational prestige data, which were introduced in Chapter 3. For the time being, I will disregard the problems with these data that were revealed by graphical analysis. Recall that Duncan wished to predict the prestige of occupations (Y) from their educational and income levels (X_1 and X_2 , respectively). I calculated the following quantities from Duncan's data:

$$\begin{aligned} n &= 45 \\ \bar{Y} &= \frac{2146}{45} = 47.689 \\ \bar{X}_1 &= \frac{2365}{45} = 52.556 \\ \bar{X}_2 &= \frac{1884}{45} = 41.867 \\ \sum X_1^{*2} &= 38,971 \\ \sum X_2^{*2} &= 26,271 \\ \sum X_1^* X_2^* &= 23,182 \\ \sum X_1^* Y^* &= 35,152 \\ \sum X_2^* Y^* &= 28,383 \end{aligned}$$

Substituting these values into Equations 5.6 produces $A = -6.0647$, $B_1 = 0.54583$, and $B_2 = 0.59873$. The fitted least-squares regression equation is, therefore,

$$\widehat{\text{Prestige}} = -6.065 + 0.5458 \times \text{Education} + 0.5987 \times \text{Income}$$

Although the development of least-squares linear regression for two explanatory variables is very similar to the development for simple regression, there is this important difference in interpretation: The slope coefficients for the explanatory variables in multiple regression are *partial* coefficients, while the slope coefficient in simple regression gives the *marginal* relationship between the response variable and a single explanatory variable. That is, each slope in multiple regression represents the “effect” on the response variable of a one-unit increment in the corresponding explanatory variable *holding constant* the value of the other explanatory variable. The simple-regression slope effectively *ignores* the other explanatory variable.

This interpretation of the multiple-regression slope is apparent in Figure 5.7, which shows the multiple-regression plane for Duncan's regression of prestige on education and income (also see Figure 5.6). Because the regression plane is flat, its slope (B_1) in the direction of education, holding income constant, does not depend on the specific value at which income is fixed. Likewise, the slope in the direction of income, fixing the value of education, is always B_2 .

Algebraically, let us fix X_2 to the specific value x_2 and see how \widehat{Y} changes as X_1 is increased by 1, from some specific value x_1 to $x_1 + 1$:

²⁰The correlation between X_1 and X_2 is, in the current notation,

$$r_{12} = \frac{\sum X_1^* X_2^*}{\sqrt{\sum X_1^{*2} \sum X_2^{*2}}}$$

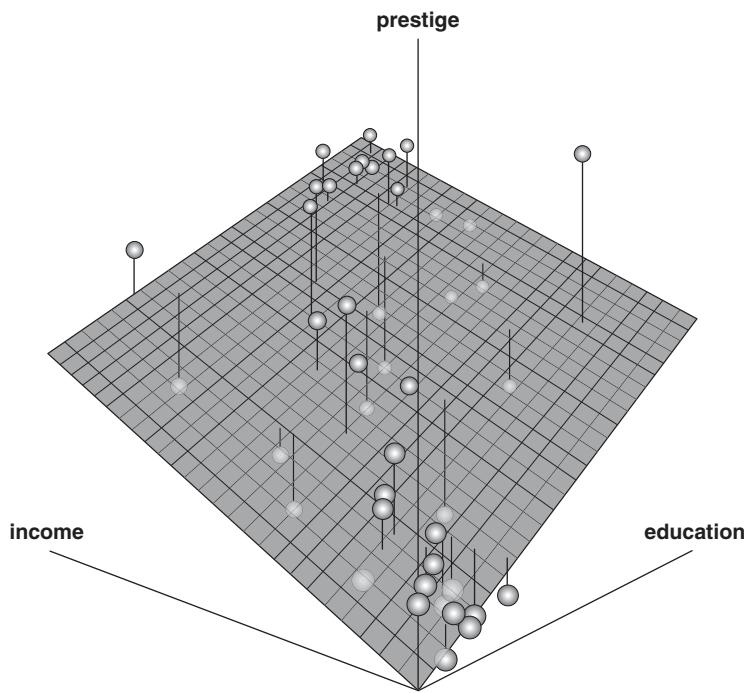


Figure 5.7 The multiple-regression plane in Duncan's regression of prestige on education and income. The two sets of parallel lines on the regression plane represent the partial relationship of prestige to each explanatory variable, holding the other explanatory variable at particular values.

$$[A + B_1(x_1 + 1) + B_2x_2] - (A + B_1x_1 + B_2x_2) = B_1$$

Similarly, increasing X_2 by 1, fixing X_1 to x_1 produces

$$[A + B_1x_1 + B_2(x_2 + 1)] - (A + B_1x_1 + B_2x_2) = B_2$$

*Because the regression surface

$$\hat{Y} = A + B_1X_1 + B_2X_2$$

is a plane, precisely the same results follow from differentiating the regression equation with respect to each of X_1 and X_2 :

$$\begin{aligned}\frac{\partial \hat{Y}}{\partial X_1} &= B_1 \\ \frac{\partial \hat{Y}}{\partial X_2} &= B_2\end{aligned}$$

Nothing new is learned here, but differentiation is often a useful approach for understanding *nonlinear* statistical models, for which the regression surface is not flat.²¹

²¹For example, see the discussion of quadratic surfaces in Section 17.1.1.

For Duncan's regression, then, a unit increase in education (i.e., in the percentage of high school graduates in an occupation), holding income constant, is associated, on average, with an increase of 0.55 units in prestige (which, recall, is the percentage of respondents rating the prestige of the occupation as good or excellent). A unit increase in income (i.e., in the percentage of relatively high-income earners), holding education constant, is associated, on average, with an increase of 0.60 units in prestige. Because Duncan's data are not longitudinal, this language of "increase" or "change" is a shorthand for hypothetical static comparisons (as was the case for the simple regression of measured on reported weight using Davis's data).

The regression intercept, $A = -6.1$, has the following literal interpretation: The fitted value of prestige is -6.1 for a hypothetical occupation with education and income levels both equal to 0. Literal interpretation of the intercept is problematic here, however. Although there are some observations in Duncan's data set with small education and income levels, no occupations have levels of 0. Moreover, the response variable cannot take on negative values.

5.2.2 Several Explanatory Variables

The extension of linear least-squares regression to several explanatory variables is straightforward. For the general case of k explanatory variables, the multiple-regression equation is

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + E_i \\ &= \hat{Y}_i + E_i \end{aligned}$$

It is, of course, not possible to visualize the point cloud of the data directly when $k > 2$, but it is a relatively simple matter to find the values of A and the B s that minimize the sum of squared residuals:

$$S(A, B_1, B_2, \dots, B_k) = \sum_{i=1}^n [Y_i - (A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik})]^2$$

Minimization of the sum-of-squares function produces the normal equations for general multiple regression:²²

$$\begin{aligned} An + B_1 \sum X_{i1} + B_2 \sum X_{i2} + \cdots + B_k \sum X_{ik} &= \sum Y_i \\ A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1} X_{i2} + \cdots + B_k \sum X_{i1} X_{ik} &= \sum X_{i1} Y_i \\ A \sum X_{i2} + B_1 \sum X_{i2} X_{i1} + B_2 \sum X_{i2}^2 + \cdots + B_k \sum X_{i2} X_{ik} &= \sum X_{i2} Y_i \\ &\vdots \\ A \sum X_{ik} + B_1 \sum X_{ik} X_{i1} + B_2 \sum X_{ik} X_{i2} + \cdots + B_k \sum X_{ik}^2 &= \sum X_{ik} Y_i \end{aligned} \tag{5.7}$$

We cannot write out a general solution to the normal equations without specifying the number of explanatory variables k , and even for k as small as 3, an explicit solution would be very complicated.²³ Nevertheless, because the normal equations are linear, and because there are as many equations as unknown regression coefficients ($k + 1$), there is usually a unique solution

²²See Exercise 5.5.

²³As I will show in Section 9.2, however, it is simple to write out a general solution to the normal equations using matrices.

Table 5.1 Sums of Squares (Diagonal), Sums of Products (Off Diagonal), and Sums (Last Row) for the Canadian Occupational Prestige Data

Variable	Prestige	Education	Income	Percentage of women
Prestige	253,618.	55,326.	37,748,108.	131,909.
Education	55,326.	12,513.	8,121,410.	32,281.
Income	37,748,108.	8,121,410.	6,534,383,460.	14,093,097.
Percentage of women	131,909.	32,281.	14,093,097.	187,312.
Sum	4777.	1095.	693,386.	2956.

for the coefficients A, B_1, B_2, \dots, B_k . Only when one explanatory variable is a perfect linear function of others, or when one or more explanatory variables are invariant, will the normal equations not have a unique solution. Dividing the first normal equation through by n reveals that the least-squares surface passes through the point of means $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k, \bar{Y})$.

The least-squares coefficients in multiple linear regression

$$Y_i = A + B_1 X_{i1} + B_2 X_{i2} + \dots + B_k X_{ik} + E_i$$

are found by solving the normal equations for the intercept A and the slope coefficients B_1, B_2, \dots, B_k . The slope coefficient B_1 represents the average change in Y associated with a one-unit increase in X_1 when the other X s are held constant.

To illustrate the solution of the normal equations, let us return to the Canadian occupational prestige data, regressing the prestige of the occupations on average education, average income, and the percentage of women in each occupation. Recall that our graphical analysis of the data in Chapter 2 cast doubt on the appropriateness of the linear regression, but I will disregard this problem for now.

The various sums, sums of squares, and sums of products that are required are given in Table 5.1. Notice that the sums of squares and products are very large, especially for income, which is scaled in small units (dollars of annual income). Substituting these values into the four normal equations and solving for the regression coefficients produces

$$\begin{aligned} A &= -6.7943 \\ B_1 &= 4.1866 \\ B_2 &= 0.0013136 \\ B_3 &= -0.0089052 \end{aligned}$$

The fitted regression equation is, therefore,

$$\widehat{\text{Prestige}} = -6.794 + 4.187 \times \text{Education} + 0.001314 \times \text{Income} \\ - 0.008905 \times \text{Percent women}$$

In interpreting the regression coefficients, we need to keep in mind the units of each variable. Prestige scores are arbitrarily scaled and range from a minimum of 14.8 to a maximum of 87.2

for these 102 occupations; the interquartile range of prestige is 24.1 points. Education is measured in years, and hence the impact of education on prestige is considerable—a little more than 4 points, on average, for each year of education, holding income and gender composition constant. Likewise, despite the small absolute size of its coefficient, the partial effect of income is also fairly large—more than 0.001 points, on average, for an additional dollar of income, or more than 1 point for each \$1,000. In contrast, the impact of gender composition, holding education and income constant, is very small—an average decline of about 0.01 points for each 1% increase in the percentage of women in an occupation.

5.2.3 Multiple Correlation

As in simple regression, the residual standard error in multiple regression measures the “average” size of the residuals. As before, we divide by degrees of freedom, here $n - (k + 1) = n - k - 1$, rather than by the sample size n , to calculate the variance of the residuals; thus, the standard error of the regression is

$$S_E = \sqrt{\frac{\sum E_i^2}{n - k - 1}}$$

Heuristically, we “lose” $k + 1$ degrees of freedom by calculating the $k + 1$ regression coefficients, A, B_1, \dots, B_k .²⁴

For Duncan’s regression of occupational prestige on the income and educational levels of occupations, the standard error is

$$S_E = \sqrt{\frac{7506.7}{45 - 2 - 1}} = 13.37$$

Recall that the response variable here is the percentage of raters classifying the occupation as good or excellent in prestige; an average prediction error of 13 is substantial given Duncan’s purpose, which was to use the regression equation to calculate substitute prestige scores for occupations for which direct ratings were unavailable. For the Canadian occupational prestige data, regressing prestige scores on average education, average income, and gender composition, the standard error is

$$S_E = \sqrt{\frac{6033.6}{102 - 3 - 1}} = 7.846$$

which is also a substantial figure.

The sums of squares in multiple regression are defined in the same manner as in simple regression:

$$\begin{aligned} \text{TSS} &= \sum (Y_i - \bar{Y})^2 \\ \text{RegSS} &= \sum (\hat{Y}_i - \bar{Y})^2 \\ \text{RSS} &= \sum (Y_i - \hat{Y}_i)^2 = \sum E_i^2 \end{aligned}$$

²⁴A deeper understanding of the central concept of degrees of freedom and its relationship to estimating the error variance is developed in Chapter 10.

Of course, the fitted values \hat{Y}_i and residuals E_i now come from the multiple-regression equation. Moreover, we have a similar analysis of variance for the regression:

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

The least-squares residuals are uncorrelated with the fitted values and with each of the X s.²⁵

The linear regression decomposes the variation in Y into “explained” and “unexplained” components: $\text{TSS} = \text{RegSS} + \text{RSS}$. The least-squares residuals, E , are uncorrelated with the fitted values, \hat{Y} , and with the explanatory variables, X_1, \dots, X_k .

The squared multiple correlation R^2 , representing the proportion of variation in the response variable captured by the regression, is defined in terms of the sums of squares:

$$R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

Because there are now several slope coefficients, potentially with different signs, the *multiple correlation coefficient* is, by convention, the positive square root of R^2 . The multiple correlation is also interpretable as the simple correlation between the fitted and observed Y values—that is, $r_{\hat{Y}Y}$.

The standard error of the regression, $S_E = \sqrt{\sum E_i^2 / (n - k - 1)}$, gives the “average” size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS}/\text{TSS}$, indicates the proportion of the variation in Y that is captured by its linear regression on the X s.

For Duncan’s regression, we have the following sums of squares:

$$\text{TSS} = 43,688.$$

$$\text{RegSS} = 36,181.$$

$$\text{RSS} = 7506.7$$

The squared multiple correlation,

$$R^2 = \frac{36,181}{43,688} = .8282$$

indicates that more than 80% of the variation in prestige among the 45 occupations is accounted for by the linear regression of prestige on the income and educational levels of the occupations. For the Canadian prestige regression, the sums of squares and R^2 are as follows:

²⁵These and other properties of the least-squares fit are derived in Chapters 9 and 10.

$$\begin{aligned} \text{TSS} &= 29,895 \\ \text{RegSS} &= 23,862 \\ \text{RSS} &= 6033.6 \\ R^2 &= \frac{23,862}{29,895} = .7982 \end{aligned}$$

Because the multiple correlation can only rise, never decline, when explanatory variables are added to the regression equation,²⁶ investigators sometimes penalize the value of R^2 by a “correction” for degrees of freedom. The corrected (or “adjusted”) R^2 is defined as

$$\tilde{R}^2 \equiv 1 - \frac{S_E^2}{S_Y^2} = 1 - \frac{\frac{\text{RSS}}{n-k-1}}{\frac{\text{TSS}}{n-1}}$$

Unless the sample size is very small, however, \tilde{R}^2 will differ little from R^2 . For Duncan’s regression, for example,

$$\tilde{R}^2 = 1 - \frac{\frac{7506.7}{45-2-1}}{\frac{43,688}{45-1}} = .8200$$

5.2.4 Standardized Regression Coefficients

Social researchers often wish to compare the coefficients of different explanatory variables in a regression analysis. When the explanatory variables are commensurable (i.e., measured in the same units on the same scale), or when they can be reduced to a common standard, comparison is straightforward. In most instances, however, explanatory variables are not commensurable. Standardized regression coefficients permit a limited assessment of the relative effects of incommensurable explanatory variables.

To place standardized coefficients in perspective, let us first consider an example in which the explanatory variables are measured in the same units. Imagine that the annual dollar income of wage workers is regressed on their years of education, years of labor force experience, and some other explanatory variables, producing the fitted regression equation

$$\widehat{\text{Income}} = A + B_1 \times \text{Education} + B_2 \times \text{Experience} + \dots$$

Because education and experience are each measured in years, the coefficients B_1 and B_2 are both expressed in dollars/year and, consequently, can be directly compared. If, for example, B_1 is larger than B_2 , then (disregarding issues arising from sampling variation) a year’s increment in education yields a greater average return in income than a year’s increment in labor force experience, holding constant the other explanatory variables in the regression equation.

It is, as I have mentioned, much more common for explanatory variables to be measured in different units. In the Canadian occupational prestige regression, for example, the coefficient for education is expressed in points (of prestige) per year, the coefficient for income is expressed in points per dollar, and the coefficient of gender composition is expressed in points

²⁶See Exercise 5.6.

per percentage of women. I have already pointed out that the income coefficient (0.001314) is much smaller than the education coefficient (4.187) not because income is a much less important determinant of prestige but because the unit of income (the dollar) is small, while the unit of education (the year) is relatively large. If we were to reexpress income in \$1,000s, then we would multiply the income coefficient by 1,000.

By the literal meaning of the term, *incommensurable* quantities cannot be directly compared. Still, in certain circumstances, incommensurables can be reduced to a common (e.g., monetary) standard. In most cases, however—as in the prestige regression—there is no obvious basis for this sort of reduction.

In the absence of a theoretically meaningful basis for comparison, an empirical comparison can be drawn by rescaling regression coefficients according to a measure of explanatory-variable spread. We can, for example, multiply each regression coefficient by the interquartile range of the corresponding explanatory variable. For the Canadian prestige data, the interquartile range of education is 4.2025 years; of income, 4081.3 dollars; and of gender composition, 48.610%. When each explanatory variable is varied over this range, holding the other explanatory variables constant, the corresponding average changes in prestige are

$$\begin{aligned} \text{Education: } 4.2025 \times 4.1866 &= 17.59 \\ \text{Income: } 4081.3 \times 0.0013136 &= 5.361 \\ \text{Gender: } 48.610 \times -0.0089052 &= -0.4329 \end{aligned}$$

Thus, education has a larger effect than income over the central half of scores observed in the data, and the effect of gender is very small. This conclusion is distinctly circumscribed: For other data, where the variation in education and income may be different, the relative impact of the variables may also differ, even if the regression coefficients are unchanged.

There is no profound justification for equating the interquartile range of one explanatory variable to that of another, as we have done here implicitly in calculating the relative “effect” of each. Indeed, the following observation should give you pause: If two explanatory variables are commensurable, and if their interquartile ranges differ, then performing this calculation is, in effect, to adopt a rubber ruler. If expressing coefficients relative to a measure of spread potentially distorts their comparison when explanatory variables are commensurable, then why should the procedure magically allow us to compare coefficients that are measured in different units?

It is much more common to standardize regression coefficients using the standard deviations of the explanatory variables rather than their interquartile ranges. Although I will proceed to explain this procedure, keep in mind that the standard deviation is not a good measure of spread when the distributions of the explanatory variables depart considerably from normality. The usual procedure standardizes the response variable as well, but this is an inessential element of the computation of standardized coefficients, because the *relative* size of the slope coefficients does not change when Y is rescaled.

Beginning with the fitted multiple-regression equation

$$Y_i = A + B_1 X_{i1} + \cdots + B_k X_{ik} + E_i$$

let us eliminate the regression constant A , expressing all the variables in mean deviation form by subtracting²⁷

²⁷Recall that the least-squares regression surface passes through the point of means for the $k + 1$ variables.

$$\bar{Y} = A + B_1 \bar{X}_1 + \cdots + B_k \bar{X}_k$$

which produces

$$Y_i - \bar{Y} = B_1(X_{i1} - \bar{X}_1) + \cdots + B_k(X_{ik} - \bar{X}_k) + E_i$$

Then divide both sides of the equation by the standard deviation of the response variable S_Y and simultaneously multiply and divide the j th term on the right-hand side of the equation by the standard deviation S_j of X_j . These operations serve to standardize each variable in the regression equation:

$$\frac{Y_i - \bar{Y}}{S_Y} = \left(B_1 \frac{S_1}{S_Y} \right) \frac{X_{i1} - \bar{X}_1}{S_1} + \cdots + \left(B_k \frac{S_k}{S_Y} \right) \frac{X_{ik} - \bar{X}_k}{S_k} + \frac{E_i}{S_Y}$$

$$Z_{iY} = B_1^* Z_{i1} + \cdots + B_k^* Z_{ik} + E_i^*$$

In this equation, $Z_{iY} \equiv (Y_i - \bar{Y})/S_Y$ is the standardized response variable, linearly transformed to a mean of 0 and a standard deviation of 1; Z_{i1}, \dots, Z_{ik} are the explanatory variables, similarly standardized; $E_i^* \equiv E_i/S_Y$ is the transformed residual, which, note, *does not* have a standard deviation of 1; and $B_j^* \equiv B_j(S_j/S_Y)$ is the *standardized partial regression coefficient* for the j th explanatory variable. The standardized coefficient is interpretable as the average change in Y , in standard deviation units, for a one standard deviation increase in X_j , holding constant the other explanatory variables.

By rescaling regression coefficients in relation to a measure of variation—such as the interquartile range or the standard deviation—standardized regression coefficients permit a limited comparison of the relative impact of incommensurable explanatory variables.

For the Canadian prestige regression, we have the following calculations:

Education:	$4.1866 \times 2.7284 / 17.204$	= 0.6640
Income:	$0.0013136 \times 4245.9 / 17.204$	= 0.3242
Gender:	$-0.0089052 \times 31.725 / 17.204$	= -0.01642

Because both income and gender composition have substantially non-normal distributions, however, the use of standard deviations here is difficult to justify.

I have stressed the restricted extent to which standardization permits the comparison of coefficients for incommensurable explanatory variables. A common misuse of standardized coefficients is to employ them to make comparisons of the effects of the *same* explanatory variable in two or more samples drawn from different populations. If the explanatory variable in question has different spreads in these samples, then spurious differences between coefficients may result, even when *unstandardized* coefficients are similar; on the other hand, differences in unstandardized coefficients can be masked by compensating differences in dispersion.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 5.1. *Prove that the least-squares fit in simple-regression analysis has the following properties:

- (a) $\sum \hat{Y}_i E_i = 0$.
- (b) $\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum E_i(\hat{Y}_i - \bar{Y}) = 0$.

Exercise 5.2. *Suppose that the means and standard deviations of Y and X are the same: $\bar{Y} = \bar{X}$ and $S_Y = S_X$.

- (a) Show that, under these circumstances,

$$B_{Y|X} = B_{X|Y} = r_{XY}$$

where $B_{Y|X}$ is the least-squares slope for the simple regression of Y on X , $B_{X|Y}$ is the least-squares slope for the simple regression of X on Y , and r_{XY} is the correlation between the two variables. Show that the intercepts are also the same, $A_{Y|X} = A_{X|Y}$.

- (b) Why, if $A_{Y|X} = A_{X|Y}$ and $B_{Y|X} = B_{X|Y}$, is the least-squares line for the regression of Y on X different from the line for the regression of X on Y (as long as $r^2 < 1$)?
- (c) “Regression toward the mean” (the original sense of the term *regression*): Imagine that X is father’s height and Y is son’s height for a sample of father-son pairs. Suppose, as above, that $S_Y = S_X$, that $\bar{Y} = \bar{X}$, and that the regression of sons’ heights on fathers’ heights is linear. Finally, suppose that $0 < r_{XY} < 1$ (i.e., fathers’ and sons’ heights are positively correlated, but not perfectly so). Show that the expected height of a son whose father is shorter than average is also less than average, but to a smaller extent; likewise, the expected height of a son whose father is taller than average is also greater than average, but to a smaller extent. Does this result imply a contradiction—that the standard deviation of a son’s height is in fact less than that of a father’s height?
- (d) What is the expected height for a father whose son is shorter than average? Of a father whose son is taller than average?
- (e) Regression effects in research design: Imagine that educational researchers wish to assess the efficacy of a new program to improve the reading performance of children. To test the program, they recruit a group of children who are reading substantially below grade level; after a year in the program, the researchers observe that the children, on average, have improved their reading performance. Why is this a weak research design? How could it be improved?

Exercise 5.3. *Show that $A' = \bar{Y}$ minimizes the sum of squares

$$S(A') = \sum_{i=1}^n (Y_i - A')^2$$

Exercise 5.4. Linear transformation of X and Y :

- (a) Suppose that the explanatory-variable values in Davis’s regression are transformed according to the equation $X' = X - 10$ and that Y is regressed on X' . Without redoing

the regression calculations in detail, find A' , B' , S'_E , and r' . What happens to these quantities when $X' = 10X$? When $X' = 10(X - 1) = 10X - 10$?

- (b) Now suppose that the response variable scores are transformed according to the formula $Y'' = Y + 10$ and that Y'' is regressed on X . Find A'' , B'' , S''_E , and r'' . What happens to these quantities when $Y'' = 5Y$? When $Y'' = 5(Y + 2) = 5Y + 10$?
- (c) In general, how are the results of a simple-regression analysis affected by linear transformations of X and Y ?

Exercise 5.5. *Derive the normal equations (Equations 5.7) for the least-squares coefficients of the general multiple-regression model with k explanatory variables. [Hint: Differentiate the sum-of-squares function $S(A, B_1, \dots, B_k)$ with respect to the regression coefficients, and set the partial derivatives to 0.]

Exercise 5.6. Why is it the case that the multiple-correlation coefficient R^2 can never get smaller when an explanatory variable is added to the regression equation? [Hint: Recall that the regression equation is fit by minimizing the residual sum of squares, which is equivalent to maximizing R^2 (why?).]

Exercise 5.7. Consider the general multiple-regression equation

$$Y = A + B_1X_1 + B_2X_2 + \dots + B_kX_k + E$$

An alternative procedure for calculating the least-squares coefficient B_1 is as follows:

1. Regress Y on X_2 through X_k , obtaining residuals $E_{Y|2\dots k}$.
 2. Regress X_1 on X_2 through X_k , obtaining residuals $E_{1|2\dots k}$.
 3. Regress the residuals $E_{Y|2\dots k}$ on the residuals $E_{1|2\dots k}$. The slope for this simple regression is the multiple-regression slope for X_1 , that is, B_1 .
- (a) Apply this procedure to the multiple regression of prestige on education, income, and percentage of women in the Canadian occupational prestige data, confirming that the coefficient for education is properly recovered.
 - (b) The intercept for the simple regression in Step 3 is 0. Why is this the case?
 - (c) In light of this procedure, is it reasonable to describe B_1 as the “effect of X_1 on Y when the influence of X_2, \dots, X_k is removed from both X_1 and Y ”?
 - (d) The procedure in this problem reduces the multiple regression to a series of simple regressions (in Step 3). Can you see any practical application for this procedure? (See the discussion of added-variable plots in Section 11.6.1.)

Exercise 5.8. Partial correlation: The *partial correlation* between X_1 and Y “controlling for” X_2 through X_k is defined as the simple correlation between the residuals $E_{Y|2\dots k}$ and $E_{1|2\dots k}$, given in the previous exercise. The partial correlation is denoted $r_{Y1|2\dots k}$.

- (a) Using the Canadian occupational prestige data, calculate the partial correlation between prestige and education, controlling for income and percentage women (see the previous exercise).

- (b) In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{Y1|2\dots k} = 0$ if and only if B_1 (from the multiple regression of Y on X_1 through X_k) is 0?

Exercise 5.9. *Show that in simple-regression analysis, the standardized slope coefficient B^* is equal to the correlation coefficient r . (In general, however, standardized slope coefficients *are not* correlations and can be outside of the range $[0, 1]$.)

Summary

- In simple linear regression

$$Y_i = A + BX_i + E_i$$

the least-squares coefficients are given by

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

$$A = \bar{Y} - B\bar{X}$$

The slope coefficient B represents the average change in Y associated with a one-unit increase in X . The intercept A is the fitted value of Y when $X = 0$.

- The least-squares coefficients in multiple linear regression

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \dots + B_kX_{ik} + E_i$$

are found by solving the normal equations for the intercept A and the slope coefficients B_1, B_2, \dots, B_k . The slope coefficient B_1 represents the average change in Y associated with a one-unit increase in X_1 when the other X s are held constant.

- The least-squares residuals, E , are uncorrelated with the fitted values, \hat{Y} , and with the explanatory variables, X_1, \dots, X_k .
- The linear regression decomposes the variation in Y into “explained” and “unexplained” components: $TSS = \text{RegSS} + \text{RSS}$. This decomposition is called the analysis of variance for the regression.
- The standard error of the regression, $S_E = \sqrt{\sum E_i^2 / (n - k - 1)}$, gives the “average” size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS}/TSS$, indicates the proportion of the variation in Y that is captured by its linear regression on the X s.
- By rescaling regression coefficients in relation to a measure of variation—such as the interquartile range or the standard deviation—standardized regression coefficients permit a limited comparison of the relative impact of incommensurable explanatory variables.

6

Statistical Inference for Regression

The previous chapter developed linear least-squares regression as a descriptive technique for fitting a linear surface to data. The subject of the present chapter, in contrast, is statistical inference. I will discuss point estimation of regression coefficients, along with elementary but powerful procedures for constructing confidence intervals and performing hypothesis tests in simple and multiple regression.¹ I will also develop two topics related to inference in regression: the distinction between empirical and structural relationships and the consequences of random measurement error in regression.

6.1 Simple Regression

6.1.1 The Simple-Regression Model

Standard statistical inference in simple regression is based on a *statistical model*, assumed to be descriptive of the population or process that is sampled:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where Y_i is the value of the response variable Y and x_i is the value of the explanatory variable X for the i th of n observations. The coefficients α and β are the *population regression parameters*; the central object of simple-regression analysis is to estimate these coefficients. The *error* ε_i represents the aggregated omitted causes of Y (i.e., the causes of Y beyond the explanatory variable X), other explanatory variables that could have been included in the regression model (at least in principle), measurement error in Y , and whatever component of Y is inherently random. A Greek letter, epsilon, is used for the errors because, without knowledge of the values of α and β , the errors are not directly observable. The key assumptions of the simple-regression model concern the behavior of the errors—or, equivalently, of the distribution of Y conditional on X :

- *Linearity.* The expectation of the error—that is, the average value of ε given the value of X —is 0: $E(\varepsilon_i) \equiv E(\varepsilon|x_i) = 0$. Equivalently, the expected value of the response variable is a linear function of the explanatory variable:

¹The focus here is on the procedures themselves: The statistical theory underlying these methods and some extensions of them are developed in Chapters 9 and 10.

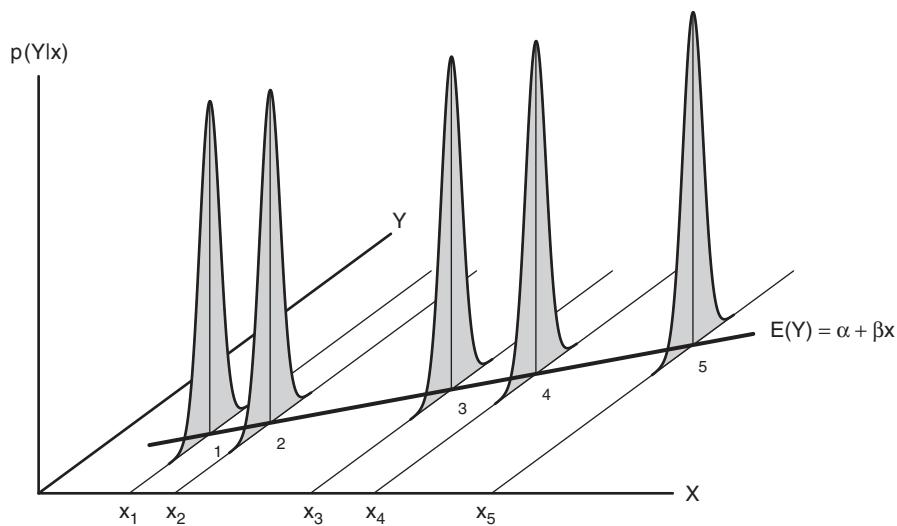


Figure 6.1 The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions $p(Y|x)$ of Y for several values of the explanatory variable X , labeled x_1, \dots, x_5 . The conditional means of Y given x are denoted μ_1, \dots, μ_5 . (Repeating Figure 2.4 on page 17.)

$$\begin{aligned}\mu_i &\equiv E(Y_i) \equiv E(Y|x_i) = E(\alpha + \beta x_i + \varepsilon_i) \\ &= \alpha + \beta x_i + E(\varepsilon_i) \\ &= \alpha + \beta x_i + 0 \\ &= \alpha + \beta x_i\end{aligned}$$

We can remove $\alpha + \beta x_i$ from the expectation operator because α and β are fixed parameters, while the value of X is conditionally fixed to x_i .²

- *Constant variance.* The variance of the errors is the same regardless of the value of X : $V(\varepsilon|x_i) = \sigma_\varepsilon^2$. Because the distribution of the errors is the same as the distribution of the response variable around the population regression line, constant error variance implies constant conditional variance of Y given X :

$$V(Y|x_i) = E[(Y_i - \mu_i)^2] = E[(Y_i - \alpha - \beta x_i)^2] = E(\varepsilon_i^2) = \sigma_\varepsilon^2$$

Note that because the mean of ε_i is 0, its variance is simply $E(\varepsilon_i^2)$.

- *Normality.* The errors are normally distributed: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Equivalently, the conditional distribution of the response variable is normal: $Y_i \sim N(\alpha + \beta x_i, \sigma_\varepsilon^2)$. The assumptions of linearity, constant variance, and normality are illustrated in Figure 6.1. It should be abundantly clear from the graph that these assumptions place very strong constraints on the structure of the data.

²I use a lowercase x here to stress that the value x_i is fixed—either literally, as in experimental research (see below), or by conditioning on the observed value x_i of X_i .

- *Independence.* The observations are sampled independently: Any pair of errors ε_i and ε_j (or, equivalently, of conditional response-variable values Y_i and Y_j) are independent for $i \neq j$. The assumption of independence needs to be justified by the procedures of data collection. For example, if the data constitute a simple random sample drawn from a large population, then the assumption of independence will be met to a close approximation. In contrast, if the data comprise a time series, then the assumption of independence may be very wrong.³
- *Fixed X , or X measured without error and independent of the error.* Depending on the design of a study, the values of the explanatory variable may be fixed in advance of data collection or they may be sampled along with the response variable. Fixed X corresponds almost exclusively to experimental research, in which the value of the explanatory variable is under the direct control of the researcher; if the experiment were replicated, then—at least in principle—the values of X (i.e., the x_i s) would remain the same.

Most social research, however, is observational, and therefore, X -values are sampled, not fixed by design (in which case, we should represent the value of X for the i th observation as X_i , as is appropriate for a random variable). Under these circumstances, we assume that the explanatory variable is measured without error and that the explanatory variable and the error are independent in the population from which the sample is drawn. That is, the error has the same distribution, $N(0, \sigma_\varepsilon^2)$, for every value of X in the population. This is in an important respect the most problematic of the assumptions underlying least-squares estimation because causal inference in nonexperimental research hinges on this assumption and because the assumption cannot be checked directly from the observed data.⁴

- *X is not invariant.* If the explanatory variable is fixed, then its values cannot all be the same, and if it is random, then there must be variation in X in the population. It is not possible to fit a line to data in which the explanatory variable is invariant.⁵

Standard statistical inference for least-squares simple-regression analysis is based on the statistical model

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

The key assumptions of the model concern the behavior of the errors ε_i : (1) Linearity, $E(\varepsilon_i) = 0$; (2) constant variance, $V(\varepsilon_i) = \sigma_\varepsilon^2$; (3) normality, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$; (4) independence, $\varepsilon_i, \varepsilon_j$ are independent for $i \neq j$; (5) the X values are fixed or, if random, are measured without error and are independent of the errors; and (6) X is not invariant.

³Chapter 16 discusses regression analysis with time-series data, while Section 15.5 takes up inference for regression in complex sample surveys. Also see Chapters 23 and 24 on mixed-effects models for hierarchical and longitudinal data.

⁴See Sections 1.2, 6.3, and 9.7 for further discussion of causal inference from observational data.

⁵See Figure 5.3 on page 86.

6.1.2 Properties of the Least-Squares Estimator

Under the strong assumptions of the simple-regression model, the sample least-squares coefficients A and B have several desirable properties as estimators of the population regression coefficients α and β :⁶

- The least-squares intercept and slope are *linear estimators*, in the sense that they are linear functions of the observations Y_i . For example, for fixed explanatory-variable values x_i ,

$$B = \sum_{i=1}^n m_i Y_i$$

where

$$m_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

While unimportant in itself, this property makes it simple to derive the sampling distributions of A and B .

- The sample least-squares coefficients are *unbiased estimators* of the population regression coefficients:

$$\begin{aligned} E(A) &= \alpha \\ E(B) &= \beta \end{aligned}$$

Only the assumption of linearity is required to establish this result.⁷

- Both A and B have simple sampling variances:

$$\begin{aligned} V(A) &= \frac{\sigma_\varepsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \\ V(B) &= \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

The assumptions of linearity, constant variance, and independence are employed in the derivation of these formulas.⁸

It is instructive to examine the formula for $V(B)$ more closely to understand the conditions under which least-squares estimation is precise. Rewriting the formula,

$$V(B) = \frac{\sigma_\varepsilon^2}{(n - 1)S_x^2}$$

Thus, the sampling variance of the slope estimate will be small when (1) the error variance σ_ε^2 is small, (2) the sample size n is large, and (3) the explanatory-variable values

⁶I will simply state and briefly explain these properties here; derivations can be found in the exercises to this chapter and in Chapter 9.

⁷See Exercise 6.1.

⁸See Exercise 6.2.

are spread out (i.e., have a large variance, S_X^2). The estimate of the intercept has small sampling variance under similar circumstances and, in addition, when the X -values are centered near 0 and, hence, $\sum x_i^2$ is not much larger than $\sum (x_i - \bar{x})^2$.⁹

- Of all the linear unbiased estimators, the least-squares estimators are the most efficient—that is, they have the smallest sampling variance and hence the smallest mean-squared error. This result, called the Gauss-Markov theorem, requires the assumptions of linearity, constant variance, and independence but not the assumption of normality.¹⁰ Under normality, moreover, the least-squares estimators are the most efficient among *all* unbiased estimators, not just among linear estimators. This is a much more compelling result, because the restriction to linear estimators is merely a matter of convenience. When the error distribution is heavier tailed than normal, for example, the least-squares estimators may be much less efficient than certain robust-regression estimators, which are *not* linear functions of the data.¹¹
- Under the full suite of assumptions, the least-squares coefficients A and B are the maximum-likelihood estimators of α and β .¹²
- Under the assumption of normality, the least-squares coefficients are themselves normally distributed. Summing up,

$$\begin{aligned} A &\sim N\left[\alpha, \frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}\right] \\ B &\sim N\left[\beta, \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}\right] \end{aligned} \quad (6.1)$$

Even if the errors are not normally distributed, the distributions of A and B are approximately normal, under very broad conditions, with the approximation improving as the sample size grows.¹³

Under the assumptions of the regression model, the least-squares coefficients have certain desirable properties as estimators of the population regression coefficients. The least-squares coefficients are linear functions of the data and therefore have simple sampling distributions, unbiased estimators of the population regression coefficients, the most efficient unbiased estimators of the population regression coefficients, maximum-likelihood estimators, and normally distributed.

⁹See Exercise 6.3.

¹⁰The theorem is named after the 19th-century German mathematical genius Carl Friedrich Gauss and the 20th-century Russian mathematician A. A. Markov. Although Gauss worked in the context of measurement error in the physical sciences, much of the general statistical theory of linear models is due to him.

¹¹See Chapter 19.

¹²See Exercise 6.5. For an explanation of maximum-likelihood estimation, see online Appendix D on probability and estimation.

¹³The asymptotic normality of A and B follows from the central limit theorem, because the least-squares coefficients are linear functions of the Y_i s.

6.1.3 Confidence Intervals and Hypothesis Tests

The distributions of A and B , given in Equations 6.1, cannot be directly employed for statistical inference because the error variance, σ_ε^2 , is never known in practice. The variance of the residuals provides an unbiased estimator of σ_ε^2 .¹⁴

$$S_E^2 = \frac{\sum E_i^2}{n - 2}$$

With the estimated error variance in hand, we can estimate the sampling variances of A and B :

$$\begin{aligned}\hat{V}(A) &= \frac{S_E^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2} \\ \hat{V}(B) &= \frac{S_E^2}{\sum (x_i - \bar{x})^2}\end{aligned}$$

As in statistical inference for the mean, the added uncertainty induced by estimating the error variance is reflected in the use of the t -distribution, in place of the normal distribution, for confidence intervals and hypothesis tests.

To construct a $100(1 - \alpha)\%$ confidence interval for the slope, we take

$$\beta = B \pm t_{\alpha/2} \text{SE}(B)$$

where $t_{\alpha/2}$ is the critical value of t with $n - 2$ degrees of freedom and a probability of $\alpha/2$ to the right, and $\text{SE}(B)$ is the standard error of B [i.e., the square root of $\hat{V}(B)$]. For a 95% confidence interval, $t_{0.025} \approx 2$, unless n is very small. Similarly, to test the hypothesis, $H_0: \beta = \beta_0$, that the population slope is equal to a specific value (most commonly, the null hypothesis $H_0: \beta = 0$), calculate the test statistic

$$t_0 = \frac{B - \beta_0}{\text{SE}(B)}$$

which is distributed as t with $n - 2$ degrees of freedom under the hypothesis H_0 . Confidence intervals and hypothesis tests for α are usually of less interest, but they follow the same pattern.

The standard error of the slope coefficient B in simple regression is $\text{SE}(B) = S_E / \sqrt{\sum (x_i - \bar{x})^2}$, which can be used to construct t -tests and t -intervals for β .

For Davis's regression of measured on reported weight (described in the preceding chapter), for example, we have the following results:

¹⁴See Section 10.3.

$$S_E = \sqrt{\frac{418.87}{101 - 2}} = 2.0569$$

$$\text{SE}(A) = \frac{2.0569 \times \sqrt{329,731}}{\sqrt{101 \times 4539.3}} = 1.7444$$

$$\text{SE}(B) = \frac{2.0569}{\sqrt{4539.3}} = 0.030529$$

Because $t_{0.025}$ for $101 - 2 = 99$ degrees of freedom is 1.9842, the 95% confidence intervals for α and β are

$$\alpha = 1.7775 \pm 1.9842 \times 1.7444 = 1.777 \pm 3.461$$

$$\beta = 0.97722 \pm 1.9842 \times 0.030529 = 0.9772 \pm 0.0606$$

The estimates of α and β are therefore quite precise. Furthermore, the confidence intervals include the values $\alpha = 0$ and $\beta = 1$, which, recall, imply unbiased prediction of measured weight from reported weight.¹⁵

6.2 Multiple Regression

Most of the results for multiple-regression analysis parallel those for simple regression.

6.2.1 The Multiple-Regression Model

The statistical model for multiple regression is

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

The assumptions underlying the model concern the errors, $\varepsilon_i \equiv \varepsilon|x_{i1}, \dots, x_{ik}|$, and are identical to the assumptions in simple regression:

- *Linearity*: $E(\varepsilon_i) = 0$.
- *Constant variance*: $V(\varepsilon_i) = \sigma_\varepsilon^2$.
- *Normality*: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.
- *Independence*: $\varepsilon_i, \varepsilon_j$ are independent for $i \neq j$.
- *Fixed Xs or Xs measured without error and independent of ε* .

In addition, we assume that the X s are not invariant in the population and that no X is a perfect linear function of the others.¹⁶

Under these assumptions (or particular subsets of them), the least-squares estimators A , B_1, \dots, B_k of $\alpha, \beta_1, \dots, \beta_k$ are

¹⁵There is, however, a subtlety here: To construct *separate* confidence intervals for α and β is not quite the same as constructing a *joint confidence region* for both coefficients simultaneously. See Section 9.4.4 for a discussion of confidence regions in regression.

¹⁶We saw in Section 5.2.1 that when explanatory variables in regression are invariant or perfectly collinear, the least-squares coefficients are not uniquely defined.

- linear functions of the data and hence relatively simple,
- unbiased,
- maximally efficient among unbiased estimators,
- maximum-likelihood estimators, and
- normally distributed.

The slope coefficient B_j in multiple regression has sampling variance¹⁷

$$\begin{aligned} V(B_j) &= \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2} \end{aligned} \quad (6.2)$$

where R_j^2 is the squared multiple correlation from the regression of X_j on all the other X s, and the \hat{x}_{ij} are the fitted values from this auxiliary regression. The second factor in the first line of Equation 6.2 is similar to the sampling variance of the slope in simple regression, although the error variance σ_ε^2 is smaller than before because some of the explanatory variables that were implicitly in the error in simple regression are now incorporated into the systematic part of the model. The first factor—called the *variance-inflation factor*—is new, however. The variance-inflation factor $1/(1 - R_j^2)$ is large when the explanatory variable X_j is strongly correlated with other explanatory variables. The denominator in the second line of Equation 6.2 is the residual sum of squares from the regression of X_j on the other X s, and it makes a similar point: When the *conditional* variation in X_j given the other X s is small, the sampling variance of B_j is large.¹⁸

We saw in Chapter 5 that when one explanatory variable is perfectly collinear with others, the least-squares regression coefficients are not uniquely determined; in this case, the variance-inflation factor is infinite. The variance-inflation factor tells us that strong, although less-than-perfect, collinearity presents a problem for estimation, for although we can calculate least-squares estimates under these circumstances, their sampling variances may be very large. Equation 6.2 reveals that the other sources of imprecision of estimation in multiple regression are the same as in simple regression: large error variance, a small sample, and explanatory variables with little variation.¹⁹

6.2.2 Confidence Intervals and Hypothesis Tests

Individual Slope Coefficients

Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of simple-regression analysis: To find the standard error of a slope coefficient, we need to

¹⁷Although we are usually less interested in inference about α , it is also possible to find the sampling variance of the intercept A . See Section 9.3.

¹⁸I am grateful to Georges Monette of York University for pointing this out to me.

¹⁹Collinearity is discussed further in Chapter 13.

substitute an estimate of the error variance for the unknown σ_ε^2 in Equation 6.2. The variance of the residuals provides an unbiased estimator of σ_ε^2 :

$$S_E^2 = \frac{\sum E_i^2}{n - k - 1}$$

Then, the standard error of B_j is

$$\text{SE}(B_j) = \frac{1}{\sqrt{1 - R_j^2}} \times \frac{S_E}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2}}$$

Confidence intervals and tests, based on the t -distribution with $n - k - 1$ degrees of freedom, follow straightforwardly.

The standard error of the slope coefficient B_j in multiple regression is $\text{SE}(B_j) = S_E / \sqrt{(1 - R_j^2) \sum (x_{ij} - \bar{x}_j)^2}$. The coefficient standard error can be used in t -intervals and t -tests for β_j .

For example, for Duncan's regression of occupational prestige on education and income (from the previous chapter), we have

$$S_E^2 = \frac{7506.7}{45 - 2 - 1} = 178.73$$

$$r_{12} = .72451$$

$$\text{SE}(B_1) = \frac{1}{\sqrt{1 - .72451^2}} \times \frac{\sqrt{178.73}}{\sqrt{38,971}} = 0.098252$$

$$\text{SE}(B_2) = \frac{1}{\sqrt{1 - .72451^2}} \times \frac{\sqrt{178.73}}{\sqrt{26,271}} = 0.11967$$

With only two explanatory variables, $R_1^2 = R_2^2 = r_{12}^2$; this simplicity and symmetry are peculiar to the two-explanatory-variable case. To construct 95% confidence intervals for the slope coefficients, we use $t_{.025} = 2.0181$ from the t -distribution with $45 - 2 - 1 = 42$ degrees of freedom. Then,

$$\begin{aligned} \text{Education: } \beta_1 &= 0.54583 \pm 2.0181 \times 0.098252 = 0.5459 \pm 0.1983 \\ \text{Income: } \beta_2 &= 0.59873 \pm 2.0181 \times 0.11967 = 0.5987 \pm 0.2415 \end{aligned}$$

Although they are far from 0, these confidence intervals are quite broad, indicating that the estimates of the education and income coefficients are imprecise—as is to be expected in a sample of only 45 occupations.

All Slopes

We can also test the null hypothesis that all the population regression slopes are 0:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{6.3}$$

Testing this global or “omnibus” null hypothesis is not quite the same as testing the k separate hypotheses

$$H_0^{(1)}: \beta_1 = 0; H_0^{(2)}: \beta_2 = 0; \dots; H_0^{(k)}: \beta_k = 0$$

If the explanatory variables are very highly correlated, for example, we might be able to reject the omnibus hypothesis (Equation 6.3) without being able to reject *any* of the individual hypotheses.

An F -test for the omnibus null hypothesis is given by

$$\begin{aligned} F_0 &= \frac{\text{RegSS}/k}{\text{RSS}/(n - k - 1)} \\ &= \frac{n - k - 1}{k} \times \frac{R^2}{1 - R^2} \end{aligned}$$

Under the omnibus null hypothesis, this test statistic has an F -distribution with k and $n - k - 1$ degrees of freedom. The omnibus F -test follows from the analysis of variance for the regression, and the calculation of the test statistic can be organized in an *analysis-of-variance table*, which shows the partition of the total variation of Y into its components:

Source	Sum of Squares	df	Mean Square	F
Regression	RegSS	k	$\frac{\text{RegSS}}{k}$	$\frac{\text{RegMS}}{\text{RMS}}$
Residuals	RSS	$n - k - 1$	$\frac{\text{RSS}}{n - k - 1}$	
Total	TSS	$n - 1$		

Note that the degrees of freedom (df) add in the same manner as the sums of squares and that the *residual mean square*, RMS , is simply the estimated error variance, S_E^2 .

It turns out that when the null hypothesis is true, RMS and the *regression mean square*, $RegMS$, provide independent estimates of the error variance, so the ratio of the two mean squares should be close to 1. When, alternatively, the null hypothesis is false, $RegMS$ estimates the error variance plus a positive quantity that depends on the β s, tending to make the numerator of F_0 larger than the denominator:

$$E(F_0) \approx \frac{E(\text{RegMS})}{E(\text{RMS})} = \frac{\sigma_\varepsilon^2 + \text{positive quantity}}{\sigma_\varepsilon^2} > 1$$

We consequently reject the omnibus null hypothesis for values of F_0 that are sufficiently larger than 1.²⁰

²⁰The reasoning here is only approximate because the expectation of the ratio of two independent random variables is not the ratio of their expectations. Nevertheless, when the sample size is large, the null distribution of the F -statistic has an expectation very close to 1. See online Appendix D on probability and estimation for information about the F -distribution.

An omnibus F -test for the null hypothesis that all the slopes are 0 can be calculated from the analysis of variance for the regression.

For Duncan's regression, we have the following analysis-of-variance table:

Source	Sum of Squares	df	Mean Square	F	p
Regression	36181.	2	18090.	101.2	<<.0001
Residuals	7506.7	42	178.73		
Total	43688.	44			

The p -value for the omnibus null hypothesis—that is, $\Pr(F > 101.2)$ for an F -distribution with 2 and 42 degrees of freedom—is very close to 0.

A Subset of Slopes

It is, finally, possible to test a null hypothesis about a *subset* of the regression slopes

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \quad (6.4)$$

where $1 \leq q \leq k$. Purely for notational convenience, I have specified a hypothesis on the *first* q coefficients; we can, of course, equally easily test a hypothesis for *any* q slopes. The “full” regression model, including all the explanatory variables, can be written as

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \beta_{q+1} x_{i,q+1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

If the null hypothesis is correct, then the first q of the β s are 0, yielding the “null” model

$$\begin{aligned} Y_i &= \alpha + 0x_{i1} + \cdots + 0x_{iq} + \beta_{q+1} x_{i,q+1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \alpha + \beta_{q+1} x_{i,q+1} + \cdots + \beta_k x_{ik} + \varepsilon_i \end{aligned}$$

In effect, then, the null model omits the first q explanatory variables, regressing Y on the remaining $k - q$ explanatory variables.

An F -test of the null hypothesis in Equation 6.4 is based on a comparison of these two models. Let RSS_1 and RegSS_1 represent, respectively, the residual and regression sums of squares for the full model; similarly, RSS_0 and RegSS_0 are the residual and regression sums of squares for the null model. Because the null model is *nested within* (i.e., is a special case of) the full model, constraining the first q slopes to 0, $\text{RSS}_0 \geq \text{RSS}_1$. The residual and regression sums of squares in the two models add to the same total sum of squares; it follows that $\text{RegSS}_0 \leq \text{RegSS}_1$. If the null hypothesis is wrong and (some of) β_1, \dots, β_q are nonzero, then the *incremental* (or “extra”) *sum of squares* due to fitting the additional explanatory variables

$$\text{RSS}_0 - \text{RSS}_1 = \text{RegSS}_1 - \text{RegSS}_0$$

should be large.

The F -statistic for testing the null hypothesis in Equation 6.4 is

$$\begin{aligned} F_0 &= \frac{(\text{RegSS}_1 - \text{RegSS}_0)/q}{\text{RSS}_1/(n - k - 1)} \\ &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \end{aligned}$$

where R_1^2 and R_0^2 are the squared multiple correlations from the full and null models, respectively. Under the null hypothesis, this test statistic has an F -distribution with q and $n - k - 1$ degrees of freedom.

The denominator of the incremental F -statistic is the estimated error variance for the full model, which provides an unbiased estimator of σ_e^2 whether or not H_0 is true (reader: why?). More generally, in computing incremental F -tests, we will always estimate the error variance from the most complete model that we fit to the data.

An F -test for the null hypothesis that a subset of slope coefficients is 0 is based on a comparison of the regression sums of squares for two models: the full regression model and a null model that deletes the explanatory variables in the null hypothesis.

The motivation for testing a subset of coefficients will become clear in the next chapter, which takes up regression models that incorporate qualitative explanatory variables. I will, for the present, illustrate the incremental F -test by applying it to the trivial case in which $q = 1$ (i.e., a single coefficient).

In Duncan's data set, the regression of prestige on income alone produces $\text{RegSS}_0 = 30,665$, while the regression of prestige on both income and education produces $\text{RegSS}_1 = 36,181$ and $\text{RSS}_1 = 7506.7$. Consequently, the incremental sum of squares due to education is $36,181 - 30,665 = 5516$. The F -statistic for testing $H_0: \beta_{\text{Education}} = 0$ is, then,

$$F_0 = \frac{5516/1}{7506.7/(45 - 2 - 1)} = 30.86$$

with 1 and 42 degrees of freedom, for which $p < .0001$.

When, as here, $q = 1$, the incremental F -test is equivalent to the t -test obtained by dividing the regression coefficient by its estimated standard error: $F_0 = t_0^2$. For the current example,

$$\begin{aligned} t_0 &= \frac{0.54583}{0.098252} = 5.5554 \\ t_0^2 &= 5.5554^2 = 30.86 \end{aligned}$$

(which is the same as F_0).

6.3 Empirical Versus Structural Relations

There are two fundamentally different interpretations of regression coefficients, and failure to distinguish clearly between them is the source of much confusion. Borrowing Goldberger's

(1973) terminology, we may interpret a regression descriptively, as an *empirical association* among variables, or causally, as a *structural relation* among variables.

I will deal first with empirical associations because the notion is simpler. Suppose that, in a population of interest, the relationship between two variables, Y and X_1 , is well described by the simple-regression model:²¹

$$Y = \alpha' + \beta'_1 X_1 + \varepsilon'$$

That is to say, the conditional mean of Y is a linear function of X . We do not assume that X_1 necessarily causes Y or, if it does, that the omitted causes of Y , incorporated in ε' , are independent of X_1 . There is, quite simply, a linear empirical relationship between Y and X_1 in the population. If we proceed to draw a random sample from this population, then the least-squares sample slope B'_1 is an unbiased estimator of β'_1 .

Suppose, now, that we introduce a second explanatory variable, X_2 , and that, in the same sense as before, the population relationship between Y and the two X s is linear:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

That is, the conditional mean of Y is a linear function of X_1 and X_2 . The slope β_1 of the population regression plane can, and generally will, differ from β'_1 , the simple-regression slope (see below). The sample least-squares coefficients for the multiple regression, B_1 and B_2 , are unbiased estimators of the corresponding population coefficients, β_1 and β_2 .

That the simple-regression slope β'_1 differs from the multiple-regression slope β_1 and that, therefore, the sample *simple*-regression coefficient B'_1 is a *biased* estimator of the population *multiple*-regression slope β_1 is not problematic, for these are simply empirical relationships, and we do not, in this context, interpret a regression coefficient as the *effect* of an explanatory variable on the response variable. The issue of *specification error*—fitting a false model to the data—does not arise, as long as the linear regression model adequately describes the empirical relationship between the response variable and the explanatory variables in the population. This would not be the case, for example, if the relationship in the population were nonlinear.

The situation is different, however, if we view the regression equation as representing a structural relation—that is, a model of how response-variable scores are determined.²² Imagine now that response-variable scores are *constructed* according to the multiple-regression model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{6.5}$$

where the error ε satisfies the usual regression assumptions; in particular, $E(\varepsilon) = 0$, and ε is independent of X_1 and X_2 .

If we use least squares to fit this model to sample data, we obtain unbiased estimators of β_1 and β_2 . Suppose, however, that instead we fit the simple-regression model

$$Y = \alpha + \beta_1 X_1 + \varepsilon' \tag{6.6}$$

where, implicitly, the effect of X_2 on Y is absorbed by the error $\varepsilon' \equiv \varepsilon + \beta_2 X_2$ because X_2 is now among the omitted causes of Y . In the event that X_1 and X_2 are correlated, there is a correlation induced between X_1 and ε' . If we proceed to assume wrongly that X_1 and ε' are

²¹Because this discussion applies to observational data, where the explanatory variables are random, I use uppercase X s.

²²In the interest of clarity, I am making this distinction more categorically than I believe is justified. I argued in Chapter 1 that it is unreasonable to treat statistical models as literal representations of social processes. Nevertheless, it is useful to distinguish between purely empirical descriptions and descriptions from which we intend to infer causation.

uncorrelated, as we do if we fit the model in Equation 6.6 by least squares, then we make an error of specification. The consequence of this error is that our simple-regression estimator of β_1 is biased: Because X_1 and X_2 are correlated and because X_2 is omitted from the model, part of the effect of X_2 is mistakenly attributed to X_1 .

To make the nature of this specification error more precise, let us take the expectation of both sides of Equation 6.5, obtaining

$$\mu_Y = \alpha + \beta_1\mu_1 + \beta_2\mu_2 + 0 \quad (6.7)$$

where, for example, μ_Y is the population mean of Y ; to obtain Equation 6.7, we use the fact that $E(\varepsilon)$ is 0. Subtracting this equation from Equation 6.5 has the effect of eliminating the constant α and expressing the variables as deviations from their population means:

$$Y - \mu_Y = \beta_1(X_1 - \mu_1) + \beta_2(X_2 - \mu_2) + \varepsilon$$

Next, multiply this equation through by $X_1 - \mu_1$:

$$(X_1 - \mu_1)(Y - \mu_Y) = \beta_1(X_1 - \mu_1)^2 + \beta_2(X_1 - \mu_1)(X_2 - \mu_2) + (X_1 - \mu_1)\varepsilon$$

Taking the expectation of both sides of the equation produces

$$\sigma_{1Y} = \beta_1\sigma_1^2 + \beta_2\sigma_{12}$$

where σ_{1Y} is the covariance between X_1 and Y , σ_1^2 is the variance of X_1 , and σ_{12} is the covariance of X_1 and X_2 .²³ Solving for β_1 , we get

$$\beta_1 = \frac{\sigma_{1Y}}{\sigma_1^2} - \beta_2 \frac{\sigma_{12}}{\sigma_1^2} \quad (6.8)$$

Recall that the least-squares coefficient for the simple regression of Y on X_1 is $B = S_{1Y}/S_1^2$. The simple regression therefore estimates not β_1 but rather $\sigma_{1Y}/\sigma_1^2 \equiv \beta'_1$. Solving Equation 6.8 for β'_1 produces $\beta'_1 = \beta_1 + \text{bias}$, where $\text{bias} = \beta_2\sigma_{12}/\sigma_1^2$.

It is instructive to take a closer look at the bias in the simple-regression estimator. For the bias to be nonzero, two conditions must be met: (1) X_2 must be a *relevant* explanatory variable—that is, $\beta_2 \neq 0$ —and (2) X_1 and X_2 must be *correlated*—that is, $\sigma_{12} \neq 0$. Moreover, depending on the signs of β_2 and σ_{12} , the bias in the simple-regression estimator may be either positive or negative.

It is important to distinguish between interpreting a regression descriptively, as an empirical association among variables, and structurally, as specifying causal relations among variables. In the latter event, but not in the former, it is sensible to speak of bias produced by omitting an explanatory variable that (1) is a cause of Y and (2) is correlated with an explanatory variable in the regression equation. Bias in least-squares estimation results from the correlation that is induced between the included explanatory variable and the error by incorporating the omitted explanatory variable in the error.

²³This result follows from the observation that the expectation of a mean deviation product is a covariance, and the expectation of a mean deviation square is a variance (see online Appendix D on probability and estimation). $E[(X_1 - \mu_1)\varepsilon] = \sigma_{1\varepsilon}$ is 0 because of the independence of X_1 and the error.

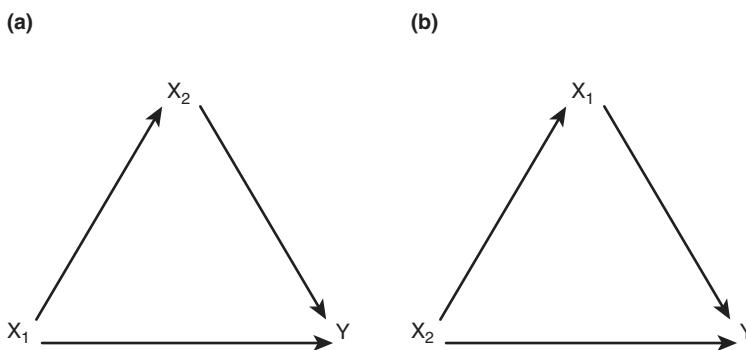


Figure 6.2 Two causal schemes relating a response variable to two explanatory variables: In (a) X_2 intervenes causally between X_1 and Y , while in (b) X_2 is a common prior cause of both X_1 and Y . In the second case, but not in the first, it is important to control for X_2 in examining the effect of X_1 on Y .

There is one final subtlety: The proper interpretation of the “bias” in the simple-regression estimator depends on the nature of the causal relationship between X_1 and X_2 . Consider the situation depicted in Figure 6.2(a), where X_2 *intervenes causally* (or *mediates* the relationship) between X_1 and Y . Here, the bias term $\beta_2 \sigma_{12} / \sigma_1^2$ is simply the *indirect effect* of X_1 on Y transmitted through X_2 , because σ_{12} / σ_1^2 is the population slope for the regression of X_2 on X_1 . If, however, as in Figure 6.2(b), X_2 is a *common prior cause* of both X_1 and Y , then the bias term represents a *spurious*—that is, noncausal—component of the empirical association between X_1 and Y . In the latter event, but not in the former, it is critical to control for X_2 in examining the relationship between Y and X_1 .²⁴ Indeed, if our goal is to estimate the effect of X_1 on Y , then we *should not* control statistically for intervening causes, such as X_2 in 6.2(a), except to articulate the causal mechanism through which X_1 affects Y .

An omitted common prior cause that accounts (or partially accounts) for the association between two variables is sometimes called a “lurking variable.” It is the always possible existence of lurking variables that makes it difficult to infer causation from observational data.

6.4 Measurement Error in Explanatory Variables*

Variables are rarely—if ever—measured without error.²⁵ Even relatively straightforward characteristics, such as education, income, height, and weight, are imperfectly measured, especially when we rely on individuals’ verbal reports. Measures of “subjective” characteristics, such as racial prejudice and conservatism, almost surely have substantial components of error. Measurement error affects not only characteristics of individuals: As you are likely aware,

²⁴Note that panels (a) and (b) in Figure 6.2 simply exchange the roles of X_1 and X_2 .

²⁵Indeed, one of the historical sources of statistical theory in the 18th and 19th centuries was the investigation of measurement errors in the physical sciences by great mathematicians like Gauss (mentioned previously) and Pierre Simon Laplace.

official statistics relating to crime, the economy, and so on are also subject to a variety of measurement errors.

The regression model accommodates measurement error in the *response* variable, because measurement error can be conceptualized as a component of the general error term ε , but the explanatory variables in regression analysis are assumed to be measured without error. In this section, I will explain the consequences of violating this assumption. To do so, we will examine the multiple-regression equation

$$Y = \beta_1\tau + \beta_2X_2 + \varepsilon \quad (6.9)$$

To keep the notation as simple as possible, all the variables in Equation 6.9 are expressed as deviations from their expectations, so the intercept α disappears from the regression equation.²⁶ One of the explanatory variables, X_2 , is measured without error, but the other, τ , is not directly observable. Instead, we have a fallible *indicator* X_1 of τ :

$$X_1 = \tau + \delta \quad (6.10)$$

where δ represents measurement error.

In addition to the usual assumptions about the regression errors ε , I will assume that the measurement errors δ are “random” and “well behaved”:

- $E(\delta) = 0$, so there is no systematic tendency for measurements to be too large or too small.
- The measurement errors δ are uncorrelated with the “true-score” variable τ . This assumption could easily be wrong. If, for example, individuals who are lighter than average tend to overreport their weights and individuals who are heavier than average tend to underreport their weights, then there will be a negative correlation between the measurement errors and true weight.
- The measurement errors δ are uncorrelated with the regression errors ε and with the other explanatory variable X_2 .

Because $\tau = X_1 - \delta$, we can rewrite Equation 6.9 as

$$\begin{aligned} Y &= \beta_1(X_1 - \delta) + \beta_2X_2 + \varepsilon \\ &= \beta_1X_1 + \beta_2X_2 + (\varepsilon - \beta_1\delta) \end{aligned} \quad (6.11)$$

As in the previous section, we can proceed by multiplying Equation 6.11 through by X_1 and X_2 and taking expectations; because all variables are in mean deviation form, expected products are covariances and expected squares are variances:²⁷

$$\begin{aligned} \sigma_{Y1}^2 &= \beta_1\sigma_1^2 + \beta_2\sigma_{12} - \beta_1\sigma_\delta^2 \\ \sigma_{Y2}^2 &= \beta_1\sigma_{12} + \beta_2\sigma_2^2 \end{aligned} \quad (6.12)$$

²⁶There is no loss of generality here, because we can always subtract the mean from each variable. See the previous section.

²⁷See Exercise 6.10.

Then, solving for the regression coefficients,

$$\begin{aligned}\beta_1 &= \frac{\sigma_{Y_1}\sigma_2^2 - \sigma_{12}\sigma_{Y_2}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2 - \sigma_\delta^2\sigma_2^2} \\ \beta_2 &= \frac{\sigma_{Y_2}\sigma_1^2 - \sigma_{12}\sigma_{Y_1}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} - \frac{\beta_1\sigma_{12}\sigma_\delta^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\end{aligned}\tag{6.13}$$

Suppose, now, that we (temporarily) ignore the measurement error in X_1 and proceed by least-squares regression of Y on X_1 and X_2 . The population analogs of the least-squares regression coefficients are as follows:²⁸

$$\begin{aligned}\beta'_1 &= \frac{\sigma_{Y_1}\sigma_2^2 - \sigma_{12}\sigma_{Y_2}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ \beta'_2 &= \frac{\sigma_{Y_2}\sigma_1^2 - \sigma_{12}\sigma_{Y_1}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\end{aligned}\tag{6.14}$$

Comparing Equations 6.13 and 6.14 reveals the consequences of ignoring the measurement error in X_1 . The denominator of β_1 in Equations 6.13 is necessarily positive, and its component $-\sigma_\delta^2\sigma_2^2$ is necessarily negative. Ignoring this component therefore inflates the denominator of β'_1 in Equations 6.14, driving the coefficient β'_1 toward 0. Put another way, ignoring measurement error in an explanatory variable tends to *attenuate* its coefficient, which makes intuitive sense.

The effect of measurement error in X_1 on the coefficient of X_2 is even more pernicious. Here, we can write $\beta'_2 = \beta_2 + \text{bias}$, where

$$\text{bias} = \frac{\beta_1\sigma_{12}\sigma_\delta^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

The bias term can be positive or negative, toward 0 or away from it. To get a better grasp on the bias in the least-squares *estimand* β'_2 , imagine that the measurement error variance σ_δ^2 grows larger and larger. Because σ_δ^2 is a component of σ_1^2 , this latter quantity also grows larger, but because the measurement errors δ are uncorrelated with variables other than X_1 , other variances and covariances are unaffected.²⁹

Using Equations 6.14,

$$\lim_{\sigma_\delta^2 \rightarrow \infty} \beta'_2 = \frac{\sigma_{Y_2}\sigma_1^2}{\sigma_1^2\sigma_2^2} = \frac{\sigma_{Y_2}}{\sigma_2^2}$$

which is the population analog of the least-squares slope for the *simple* regression of Y on X_2 alone. Once more, the result is simple and intuitively plausible: Substantial measurement error in X_1 renders it an ineffective statistical control, driving β'_2 toward the marginal relationship between X_2 and Y , and away from the partial relationship between these two variables.³⁰

Measurement error in an explanatory variable tends to attenuate its regression coefficient and to make the variable an imperfect statistical control.

²⁸See Exercise 6.11.

²⁹See Exercise 6.12.

³⁰I am grateful to Georges Monette, of York University, for this insight. See Exercise 6.13 for an illustration.

Although there are statistical methods that attempt to estimate regression equations taking account of measurement errors, these methods are beyond the scope of the presentation in this book and, in any event, involve assumptions that often are difficult to justify in practice.³¹ Perhaps the most important lessons to be drawn from the results of this section are (1) that large measurement errors in the X s can invalidate a regression analysis; (2) that, therefore, where measurement errors are likely to be substantial, we should not view the results of a regression as definitive; and (3) that it is worthwhile to expend effort to improve the quality of social measurements.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 6.1. *Demonstrate the unbias of the least-squares estimators A and B of α and β in simple regression:

- (a) Expressing the least-squares slope B as a linear function of the observations, $B = \sum m_i Y_i$ (as in the text), and using the assumption of linearity, $E(Y_i) = \alpha + \beta x_i$, show that $E(B) = \beta$. [Hint: $E(B) = \sum m_i E(Y_i)$.]
- (b) Show that A can also be written as a linear function of the Y s. Then, show that $E(A) = \alpha$.

Exercise 6.2. *Using the assumptions of linearity, constant variance, and independence, along with the fact that A and B can each be expressed as a linear function of the Y s, derive the sampling variances of A and B in simple regression. [Hint: $V(B) = \sum m_i^2 V(Y_i)$.]

Exercise 6.3. Examining the formula for the sampling variance of A in simple regression,

$$V(A) = \frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

why is it intuitively sensible that the variance of A is large when the mean of the xs is far from 0? Illustrate your explanation with a graph.

Exercise 6.4. The formula for the sampling variance of B in simple regression,

$$V(B) = \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}$$

shows that, to estimate β precisely, it helps to have spread out xs. Explain why this result is intuitively sensible, illustrating your explanation with a graph. What happens to $V(B)$ when there is *no* variation in X ?

Exercise 6.5. *Maximum-likelihood estimation of the simple-regression model: Deriving the maximum-likelihood estimators of α and β in simple regression is straightforward. Under the

³¹Measurement errors in explanatory variables are often discussed in the context of *structural equation models*, which are multiple-equation regression models in which the response variable in one equation can appear as an explanatory variable in others. Duncan (1975, Chapters 9 and 10) presents a fine elementary treatment of the topic, part of which I have adapted for the presentation in this section. A more advanced development may be found in Bollen (1989).

assumptions of the model, the Y_i s are independently and normally distributed random variables with expectations $\alpha + \beta x_i$ and common variance σ^2_ε . Show that if these assumptions hold, then the least-squares coefficients A and B are the maximum-likelihood estimators of α and β and that $\hat{\sigma}^2_\varepsilon = \sum E_i^2/n$ is the maximum-likelihood estimator of σ^2_ε . Note that the MLE of the error variance is biased. (*Hints:* Because of the assumption of independence, the joint probability density for the Y_i s is the product of their marginal probability densities

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2_\varepsilon}} \exp\left[-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2_\varepsilon}\right]$$

Find the log-likelihood function; take the partial derivatives of the log likelihood with respect to the parameters α , β , and σ^2_ε ; set these partial derivatives to 0; and solve for the maximum-likelihood estimators.) A more general result is proved in Section 9.3.3.

Exercise 6.6. Linear transformation of X and Y in simple regression (continuation of Exercise 5.4):

- (a) Suppose that the X -values in Davis's regression of measured on reported weight are transformed according to the equation $X' = 10(X - 1)$ and that Y is regressed on X' . Without redoing the regression calculations in detail, find $SE(B')$ and $t'_0 = B'/SE(B')$.
- (b) Now, suppose that the Y values are transformed according to the equation $Y'' = 5(Y + 2)$ and that Y'' is regressed on X . Find $SE(B'')$ and $t''_0 = B''/SE(B'')$.
- (c) In general, how are hypothesis tests and confidence intervals for β affected by linear transformations of X and Y ?

Exercise 6.7. Consider the regression model $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. How can the incremental sum-of-squares approach be used to test the hypothesis that the two population slopes are equal to each other, $H_0: \beta_1 = \beta_2$? [Hint: Under H_0 , the model becomes $Y = \alpha + \beta x_1 + \beta x_2 + \varepsilon = Y = \alpha + \beta(x_1 + x_2) + \varepsilon$, where β is the common value of β_1 and β_2 .] Under what circumstances would a hypothesis of this form be meaningful? (Hint: Consider the units of measurement of x_1 and x_2 .) Now, test the hypothesis that the "population" regression coefficients for education and income in Duncan's occupational prestige regression are equal to each other. Is this test sensible?

Exercise 6.8. Examples of specification error (also see the discussion in Section 9.7):

- (a) Describe a nonexperimental research situation—real or contrived—in which failure to control statistically for an omitted variable induces a correlation between the error and an explanatory variable, producing erroneous conclusions. (For example: An educational researcher discovers that university students who study more get lower grades on average; the researcher concludes that studying has an adverse effect on students' grades.)
- (b) Describe an experiment—real or contrived—in which faulty experimental practice induces an explanatory variable to become correlated with the error, compromising the validity of the results produced by the experiment. (For example: In an experimental study of a promising new therapy for depression, doctors administering the treatments tend to use the new therapy with patients for whom more traditional approaches have

failed; it is discovered that subjects receiving the new treatment tend to do worse, on average, than those receiving older treatments or a placebo; the researcher concludes that the new treatment is not effective.)

- (c) Is it fair to conclude that a researcher is *never* able absolutely to rule out the possibility that an explanatory variable of interest is correlated with the error? Is experimental research no better than observational research in this respect? Explain your answer.

Exercise 6.9. Suppose that the “true” model generating a set of data is $Y = \alpha + \beta_1 X_1 + \varepsilon$, where the error ε conforms to the usual linear-regression assumptions. A researcher fits the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, which includes the irrelevant explanatory variable X_2 —that is, the true value of β_2 is 0. Had the researcher fit the (correct) simple-regression model, the variance of B_1 would have been $V(B_1) = \sigma_\varepsilon^2 / \sum (X_{i1} - \bar{X}_1)^2$.

- (a) Is the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ wrong? Is B_1 for this model a biased estimator of β_1 ?
 (b) The variance of B_1 in the multiple-regression model is

$$V(B_1) = \frac{1}{1 - r_{12}^2} \times \frac{\sigma_\varepsilon^2}{\sum (X_{i1} - \bar{X}_1)^2}$$

What, then, is the cost of including the irrelevant explanatory variable X_2 ? How does this cost compare to that of *failing* to include a relevant explanatory variable?

Exercise 6.10. *Derive Equations 6.12 by multiplying Equation 6.11 through by each of X_1 and X_2 . (*Hints:* Both X_1 and X_2 are uncorrelated with the regression error ε . Likewise, X_2 is uncorrelated with the measurement error δ . Show that the covariance of X_1 and δ is simply the measurement error variance σ_δ^2 by multiplying $X_1 = \tau + \delta$ through by δ and taking expectations.)

Exercise 6.11. *Show that the population analogs of the regression coefficients can be written as in Equations 6.14. (*Hint:* Ignore the measurement errors, and derive the population analogs of the normal equations by multiplying the “model” $Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ through by each of X_1 and X_2 and taking expectations.)

Exercise 6.12. *Show that the variance of $X_1 = \tau + \delta$ can be written as the sum of “true-score variance,” σ_τ^2 , and measurement error variance, σ_δ^2 . (*Hint:* Square both sides of Equation 6.10 and take expectations.)

Exercise 6.13. Recall Duncan’s regression of occupational prestige on the educational and income levels of occupations. Following Duncan, regress prestige on education and income. Also, perform a simple regression of prestige on income alone. Then add random measurement errors to education. Sample these measurement errors from a normal distribution with mean 0, repeating the exercise for each of the following measurement error variances: $\sigma_\delta^2 = 10^2, 25^2, 50^2, 100^2$. In each case, recompute the regression of prestige on income and education. Then, treating the initial multiple regression as corresponding to $\sigma_\delta^2 = 0$, plot the coefficients of education and income as a function of σ_δ^2 . What happens to the education coefficient as measurement error in education grows? What happens to the income coefficient?

Exercise 6.14. *Instrumental-variable estimation: As explained, when we want to construe a regression causally, the most potentially problematic of the assumptions underlying the linear regression model is the assumption that the errors and explanatory variables are independent, because this assumption cannot be checked against the data. Consider the simple-regression model $Y^* = \beta X^* + \varepsilon$, where, for simplicity, both $Y^* \equiv Y - E(Y)$ and $X^* \equiv X - E(X)$ are expressed as deviations from their expectations, so that $E(Y^*) = E(X^*) = 0$ and the intercept α is eliminated from the model.

- (a) Suppose that X and ε are independent. Show that the ordinary least-squares estimator of β , $B_{OLS} = S_{XY}/S_X^2$ (where S_{XY} is the sample covariance of X and Y , and S_X^2 is the sample variance of X), can be derived by (1) multiplying the model through by X^* , (2) taking the expectation of both sides of the resulting equation, and (3) substituting the sample variance and covariance for their population analogs. Because the sample variance and covariance are consistent estimators of the population variance and covariance, B_{OLS} is a consistent estimator of β .
- (b) Now suppose that it is unreasonable to assume that X and ε are independent, but there is a third observed variable, Z , that is (1) independent of ε and (2) correlated with X . Z is called an *instrumental variable* (or an *instrument*). Proceeding in a manner similar to part (a), but multiplying the model through by $Z^* \equiv Z - E(Z)$ rather than X^* , show that the instrumental-variable estimator $B_{IV} = S_{ZY}/S_{ZX}$ is a consistent estimator of β . Why are both conditions (1) and (2) necessary for the instrumental variable Z to do its job?
- (c) Suggest a substantive application in which it is unreasonable to assume that X is independent of other, prior causes of Y but where there is a third variable Z that is both correlated with X and, arguably, independent of the error.

Instrumental-variables estimation is elaborated in Section 9.8.

Summary

- Standard statistical inference for least-squares regression analysis is based on the statistical model

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

The key assumptions of the model concern the behavior of the errors ε_i :

1. *Linearity:* $E(\varepsilon_i) = 0$.
2. *Constant variance:* $V(\varepsilon_i) = \sigma_\varepsilon^2$.
3. *Normality:* $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.
4. *Independence:* $\varepsilon_i, \varepsilon_j$ are independent for $i \neq j$.
5. The X values are fixed or, if random, are measured without error and are independent of the errors.

In addition, we assume that the X s are not invariant and that no X is a perfect linear function of the others.

- Under these assumptions, or particular subsets of them, the least-squares coefficients have certain desirable properties as estimators of the population regression coefficients. The least-squares coefficients are
 1. linear functions of the data and therefore have simple sampling distributions,
 2. unbiased estimators of the population regression coefficients,
 3. the most efficient unbiased estimators of the population regression coefficients,
 4. maximum-likelihood estimators, and
 5. normally distributed.
- The standard error of the slope coefficient B in simple regression is

$$\text{SE}(B) = \frac{S_E}{\sqrt{\sum (x_i - \bar{x})^2}}$$

The standard error of the slope coefficient B_j in multiple regression is

$$\text{SE}(B_j) = \frac{1}{\sqrt{1 - R_j^2}} \times \frac{S_E}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2}}$$

In both cases, these standard errors can be used in t -intervals and t -tests for the corresponding population slope coefficients.

- An F -test for the omnibus null hypothesis that all the slopes are 0 can be calculated from the analysis of variance for the regression:

$$F_0 = \frac{\text{RegSS}/k}{\text{RSS}/(n - k - 1)}$$

The omnibus F -statistic has k and $n - k - 1$ degrees of freedom.

- There is an incremental F -test for the null hypothesis that a subset of q slope coefficients is 0. This test is based on a comparison of the regression sums of squares for the full regression model (model 1) and for a null model (model 0) that deletes the explanatory variables in the null hypothesis:

$$F_0 = \frac{(\text{RegSS}_1 - \text{RegSS}_0)/q}{\text{RSS}_1/(n - k - 1)}$$

This F -statistic has q and $n - k - 1$ degrees of freedom.

- It is important to distinguish between interpreting a regression descriptively, as an empirical association among variables, and structurally, as specifying causal relations among variables. In the latter event, but not in the former, it is sensible to speak of bias produced by omitting an explanatory variable that (1) is a cause of Y and (2) is correlated with an explanatory variable in the regression equation. Bias in least-squares estimation results from the correlation that is induced between the included explanatory variable and the error by incorporating the omitted explanatory variable in the error.
- Measurement error in an explanatory variable tends to attenuate its regression coefficient and to make the variable an imperfect statistical control.

7

Dummy-Variable Regression

An obvious limitation of multiple-regression analysis, as presented in Chapters 5 and 6, is that it accommodates only quantitative response and explanatory variables. In this chapter and the next, I will show how qualitative (i.e., categorical) explanatory variables, called *factors*, can be incorporated into a linear model.¹

The current chapter begins by introducing a *dummy-variable regressor*, coded to represent a *dichotomous* (i.e., two-category) factor. I proceed to show how a set of dummy regressors can be employed to represent a *polytomous* (many-category) factor. I next describe how interactions between quantitative and qualitative explanatory variables can be included in dummy-regression models and how to summarize models that incorporate interactions. Finally, I explain why it does not make sense to standardize dummy-variable and interaction regressors.

7.1 A Dichotomous Factor

Let us consider the simplest case: one dichotomous factor and one quantitative explanatory variable. As in the two previous chapters, assume that relationships are *additive*—that is, that the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant. As well, suppose that the other assumptions of the regression model hold: The errors are independent and normally distributed, with zero means and constant variance.

The general motivation for including a factor in a regression is essentially the same as for including an additional quantitative explanatory variable: (1) to account more fully for the response variable, by making the errors smaller, and (2) even more important, to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting a causally prior explanatory variable that is related to it.

For concreteness, suppose that we are interested in investigating the relationship between education and income among women and men. Figure 7.1(a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the “effect” of gender is the vertical distance between the two regression lines, which—for parallel lines—is everywhere the same. Likewise, holding gender constant, the “effect” of education is captured by the within-gender education slope, which—for parallel lines—is the same for men and women.²

¹Chapter 14 deals with qualitative *response* variables.

²I will consider nonparallel within-group regressions in Section 7.3.

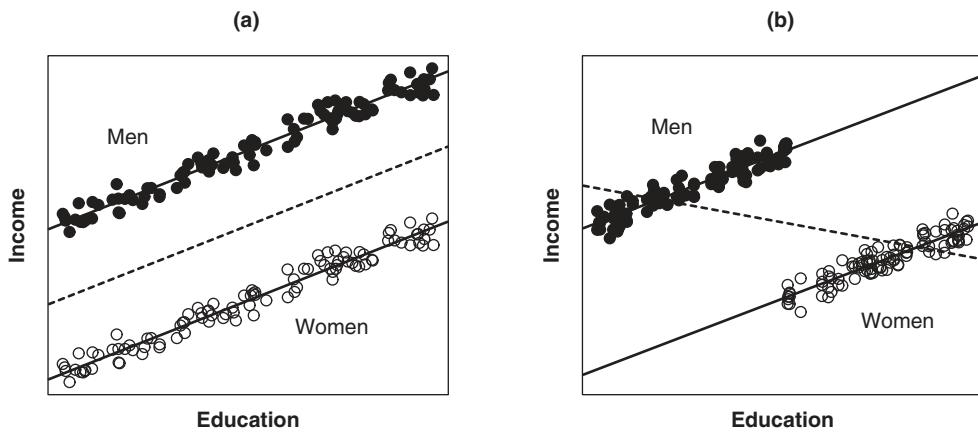


Figure 7.1 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e., marginal) regression of income on education (ignoring gender) is given by the broken line.

In Figure 7.1(a), the explanatory variables gender and education are unrelated to each other: Women and men have identical distributions of education scores (as can be seen by projecting the points onto the horizontal axis). In this circumstance, if we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender regressions. Because women have lower incomes than men of equal education, however, by ignoring gender, we inflate the size of the errors.

The situation depicted in Figure 7.1(b) is importantly different. Here, gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income: Because women have a higher average level of education than men, and because—for a given level of education—women's incomes are lower, on average, than men's, the overall regression of income on education has a *negative* slope even though the within-gender regressions have a *positive* slope.³

In light of these considerations, we might proceed to partition our sample by gender and perform separate regressions for women and men. This approach is reasonable, but it has its limitations: Fitting separate regressions makes it difficult to estimate and test for gender differences in income. Furthermore, if we can reasonably assume parallel regressions for women and men, we can more efficiently estimate the common education slope by pooling sample data drawn from both groups. In particular, if the usual assumptions of the regression model hold, then it is desirable to fit the common-slope model by least squares.

One way of formulating the common-slope model is

³That marginal and partial relationships can differ in sign is called *Simpson's paradox* (Simpson, 1951). Here, the marginal relationship between income and education is negative, while the partial relationship, controlling for gender, is positive.

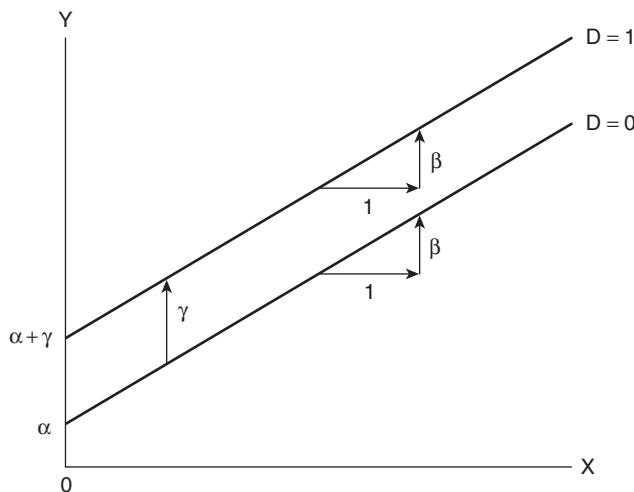


Figure 7.2 The additive dummy-variable regression model. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \quad (7.1)$$

where D , called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

Thus, for women, the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

and for men

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

These regression equations are graphed in Figure 7.2.

This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors*. Here, *gender* is a qualitative explanatory variable (i.e., a factor), with categories *male* and *female*. The dummy variable D is a regressor, representing the factor gender. In contrast, the quantitative explanatory variable *education* and the regressor X are one and the same. Were we to transform education, however, prior to entering it into the regression equation—say, by taking logs—then there would be a distinction between the explanatory variable (education) and the regressor (log education). In subsequent sections of this chapter, it will transpire that an explanatory variable can give rise to several regressors and that some regressors are functions of more than one explanatory variable.

Returning to Equation 7.1 and Figure 7.2, the coefficient γ for the dummy regressor gives the difference in intercepts for the two regression lines. Moreover, because the within-gender

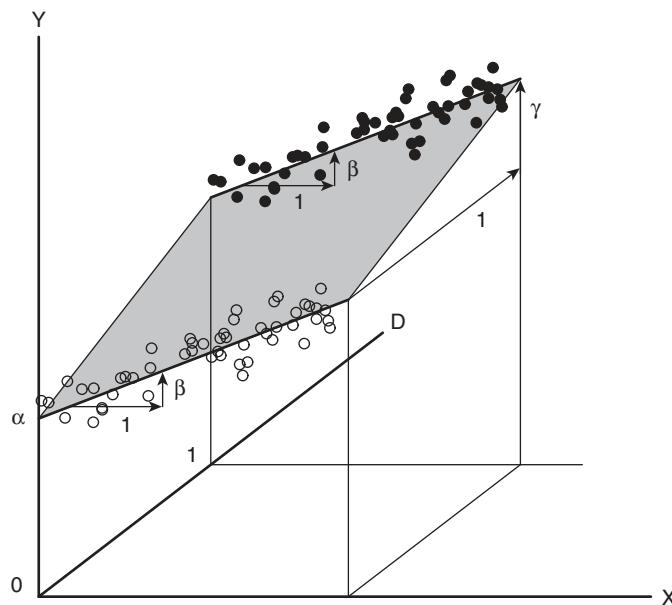


Figure 7.3 The geometric “trick” underlying dummy regression: The linear regression plane is defined only at $D=0$ and $D=1$, producing two regression lines with slope β and vertical separation γ . The hollow circles represent women, for whom $D=0$, and the solid circles men, for whom $D=1$.

regression lines are parallel, γ also represents the constant vertical separation between the lines, and it may, therefore, be interpreted as the expected income advantage accruing to men when education is held constant. If men were *disadvantaged* relative to women with the same level of education, then γ would be *negative*. The coefficient α gives the intercept for women, for whom $D=0$, and β is the common within-gender education slope.

Figure 7.3 reveals the fundamental geometric “trick” underlying the coding of a dummy regressor: We are fitting a regression plane to the data, but the dummy regressor D is defined only at the values 0 and 1. The regression plane intersects the planes $\{X, Y|D=0\}$ and $\{X, Y|D=1\}$ in two lines, each with slope β . Because the difference between $D=0$ and $D=1$ is one unit, the difference in the Y -intercepts of these two lines is the slope of the plane in the D direction, that is, γ . Indeed, Figure 7.2 is simply the projection of the two regression lines onto the $\{X, Y\}$ plane.

Essentially similar results are obtained if we instead code D equal to 0 for men and 1 for women, making men the *baseline* (or *reference*) category (see Figure 7.4): The *sign* of γ is reversed, because it now represents the difference in intercepts between women and men (rather than vice versa), but its *magnitude* remains the same. The coefficient α now gives the income intercept for men. It is therefore immaterial which group is coded 1 and which is coded 0, as long as we are careful to interpret the coefficients of the model—for example, the sign of γ —in a manner consistent with the coding scheme that is employed.

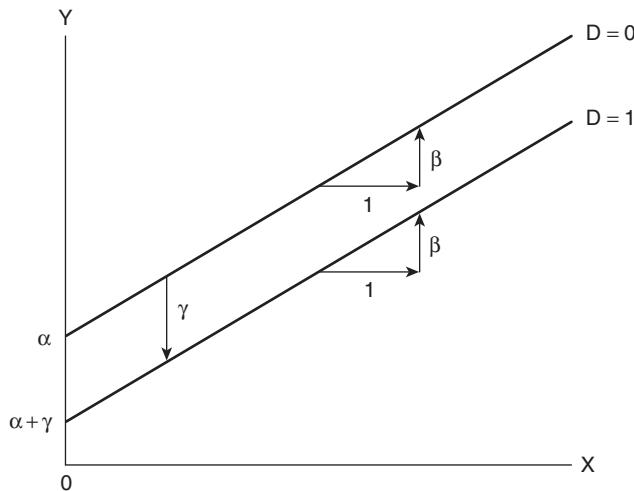


Figure 7.4 The additive dummy-regression model coding $D=0$ for men and $D=1$ for women (cf. Figure 7.2).

To determine whether gender affects income, controlling for education, we can test $H_0: \gamma = 0$, either by a t -test, dividing the estimate of γ by its standard error or, equivalently, by dropping D from the regression model and formulating an incremental F -test. In either event, the statistical-inference procedures of the previous chapter apply.

Although I have developed dummy-variable regression for a single quantitative regressor, the method can be applied to any number of quantitative explanatory variables, as long as we are willing to assume that the slopes are the same in the two categories of the factor—that is, that the regression surfaces are parallel in the two groups. In general, if we fit the model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

then, for $D = 0$, we have

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

and, for $D = 1$,

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the factor and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

7.2 Polytomous Factors

The coding method of the previous section generalizes straightforwardly to polytomous factors. By way of illustration, recall (from the previous chapter) the Canadian occupational prestige data. I have classified the occupations into three rough categories: (1) professional and managerial occupations, (2) “white-collar” occupations, and (3) “blue-collar” occupations.⁴

Figure 7.5 shows conditioning plots for the relationship between prestige and each of income and education within occupational types.⁵ The partial relationships between prestige and the explanatory variables appear reasonably linear, although there seems to be evidence that the income slope (and possibly the education slope) varies across the categories of type of occupation (a possibility that I will pursue in the next section of the chapter). Indeed, this change in slope is an explanation of the nonlinearity in the relationship between prestige and income that we noticed in Chapter 4. These conditioning plots do not tell the whole story, however, because the income and education levels of the occupations are correlated, but they give us a reasonable initial look at the data. Conditioning the plot for income by level of education (and vice versa) is out of the question here because of the small size of the data set.

The *three*-category occupational-type factor can be represented in the regression equation by introducing *two* dummy regressors, employing the following coding scheme:

Category	D_1	D_2	(7.2)
Professional and managerial	1	0	
White collar	0	1	
Blue collar	0	0	

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i \quad (7.3)$$

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{aligned} \text{Professional: } Y_i &= (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{White collar: } Y_i &= (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{Blue collar: } Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

The coefficient α , therefore, gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income). Assuming, for simplicity, that all coefficients are positive and that $\gamma_1 > \gamma_2$, the geometry of the model in Equation 7.3 is illustrated in Figure 7.6.

⁴Although there are 102 occupations in the full data set, several are difficult to classify and consequently were dropped from the analysis. The omitted occupations are athletes, babysitters, farmers, and “newsboys,” leaving us with 98 observations.

⁵In the preceding chapter, I also included the gender composition of the occupations as an explanatory variable, but I omit that variable here. Conditioning plots are described in Section 3.3.4.

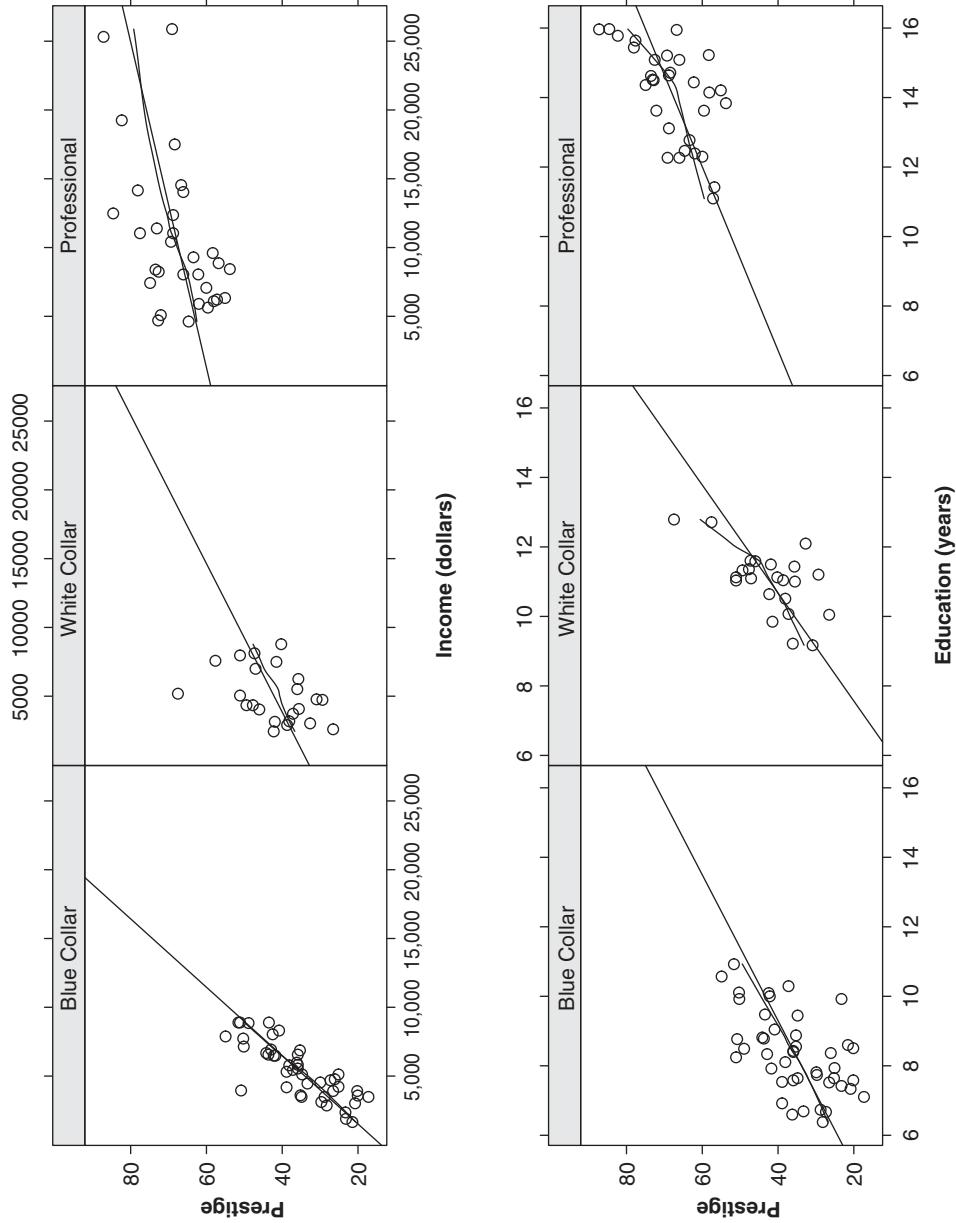


Figure 7.5 Conditioning plots for the relationship between prestige and each of income (top panel) and education (bottom panel) by type of occupation, for the Canadian occupational prestige data. Each panel shows the linear least-squares fit and a lowess smooth with a span of 0.9. The graphs labeled “Professional” are for professional and managerial occupations.

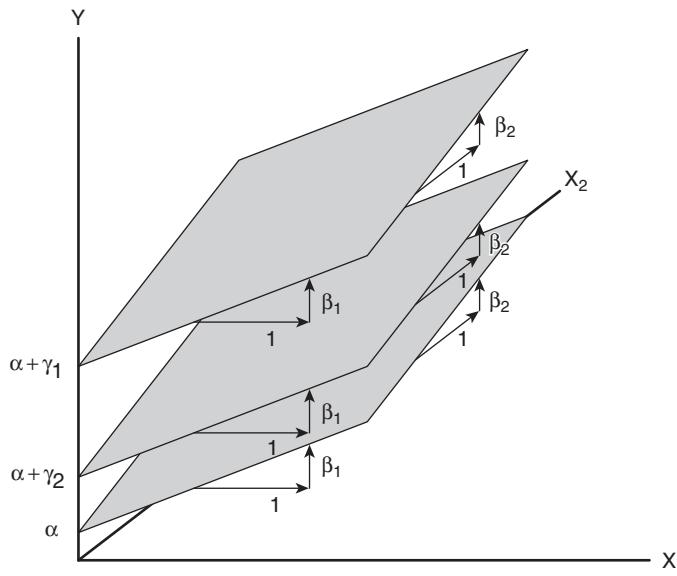


Figure 7.6 The additive dummy-regression model with two quantitative explanatory variables X_1 and X_2 represents parallel planes with potentially different intercepts in the $\{X_1, X_2, Y\}$ space.

Because blue-collar occupations are coded 0 for both dummy regressors, “blue collar” implicitly serves as the baseline category to which the other occupational-type categories are compared. The choice of a baseline category is essentially arbitrary, for we would fit precisely the same three regression planes regardless of which of the three occupational-type categories is selected for this role. The values (and meaning) of the individual dummy-variable coefficients γ_1 and γ_2 depend, however, on which category is chosen as the baseline.

It is sometimes natural to select a particular category as a basis for comparison—an experiment that includes a “control group” comes immediately to mind. In this instance, the individual dummy-variable coefficients are of interest, because they reflect differences between the “experimental” groups and the control group, holding other explanatory variables constant.

In most applications, however, the choice of a baseline category is entirely arbitrary, as it is for the occupational prestige regression. We are, therefore, most interested in testing the null hypothesis of no effect of occupational type, controlling for education and income,

$$H_0: \gamma_1 = \gamma_2 = 0 \quad (7.4)$$

but the individual hypotheses $H_0: \gamma_1 = 0$ and $H_0: \gamma_2 = 0$ —which test, respectively, for differences between professional and blue-collar occupations, as well as between white-collar and blue-collar occupations—are of less intrinsic interest.⁶ The null hypothesis in Equation 7.4 can

⁶The essential point here is not that the separate hypotheses are of *no* interest but that they are an arbitrary subset of the pairwise differences among the categories. In the present case, where there are three categories, the individual hypotheses represent two of the three pairwise group comparisons. The third comparison, between professional and white-collar occupations, is not *directly* represented in the model, although it is given indirectly by the difference $\gamma_1 - \gamma_2$. See Section 7.2.1 for an elaboration of this point.

be tested by the incremental sum-of-squares approach, dropping the two dummy variables for type of occupation from the model.

I have demonstrated how to model the effects of a three-category factor by coding two dummy regressors. It may seem more natural, however, to treat the three occupational categories symmetrically, coding *three* dummy regressors, rather than arbitrarily selecting one category as the baseline:

Category	D_1	D_2	D_3	
Professional and managerial	1	0	0	(7.5)
White collar	0	1	0	
Blue collar	0	0	1	

Then, for the j th occupational type, we would have

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

The problem with this procedure is that there are too many parameters: We have used four parameters ($\alpha, \gamma_1, \gamma_2, \gamma_3$) to represent only three group intercepts. As a consequence, we could not find unique values for these four parameters even if we knew the three population regression lines. Likewise, we cannot calculate unique least-squares estimates for the model because the set of three dummy variables is perfectly collinear; for example, as is apparent from the table in (7.5), $D_3 = 1 - D_1 - D_2$.

In general, then, for a polytomous factor with m categories, we need to code $m - 1$ dummy regressors. One simple scheme is to select the last category as the baseline and to code $D_{ij} = 1$ when observation i falls in category j , for $j = 1, \dots, m - 1$, and 0 otherwise:

Category	D_1	D_2	\dots	D_{m-1}	
1	1	0	\dots	0	
2	0	1	\dots	0	(7.6)
\vdots	\vdots	\vdots		\vdots	
$m-1$	0	0	\dots	1	
m	0	0	\dots	0	

A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded 0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

When there is more than one factor, and if we assume that the factors have additive effects, we can simply code a set of dummy regressors for each. To test the hypothesis that the effect of a

factor is nil, we delete its dummy regressors from the model and compute an incremental F -test of the hypothesis that all the associated coefficients are 0.

Regressing occupational prestige (Y) on income (X_1) and education (X_2) produces the fitted regression equation

$$\hat{Y} = -7.621 + 0.001241X_1 + 4.292X_2 \quad R^2 = .81400$$

$$(3.116) \quad (0.000219) \quad (0.336)$$

As is common practice, I have shown the estimated standard error of each regression coefficient in parentheses beneath the coefficient. The three occupational categories differ considerably in their average levels of prestige:

Category	Number of Cases	Mean Prestige
Professional and managerial	31	67.85
White collar	23	42.24
Blue collar	44	35.53
All occupations	98	47.33

Inserting dummy variables for type of occupation into the regression equation, employing the coding scheme shown in Equation 7.2, produces the following results:

$$\hat{Y} = -0.6229 + 0.001013X_1 + 3.673X_2 + 6.039D_1 - 2.737D_2$$

$$(5.2275) \quad (0.000221) \quad (0.641) \quad (3.867) \quad (2.514) \quad (7.7)$$

$$R^2 = .83486$$

The three fitted regression equations are, therefore,

$$\begin{aligned} \text{Professional: } \hat{Y} &= 5.416 + 0.001013X_1 + 3.673X_2 \\ \text{White collar: } \hat{Y} &= -3.360 + 0.001013X_1 + 3.673X_2 \\ \text{Blue collar: } \hat{Y} &= -0.623 + 0.001013X_1 + 3.673X_2 \end{aligned}$$

where the intercept for professional occupations is $-0.623 + 6.039 = 5.416$, and the intercept for white-collar occupations is $-0.623 - 2.737 = -3.360$.

Note that the coefficients for both income and education become slightly smaller when type of occupation is controlled. As well, the dummy-variable coefficients (or, equivalently, the category intercepts) reveal that when education and income levels are held constant statistically, the difference in average prestige between professional and blue-collar occupations declines greatly, from $67.85 - 35.53 = 32.32$ points to 6.04 points. The difference between white-collar and blue-collar occupations is reversed when income and education are held constant, changing from $42.24 - 35.53 = +6.71$ points to -2.74 points. That is, the greater prestige of professional occupations compared to blue-collar occupations appears to be due mostly to differences in education and income between these two classes of occupations. While white-collar occupations have greater prestige, on average, than blue-collar occupations, they have lower prestige than blue-collar occupations of the same educational and income levels.⁷

⁷These conclusions presuppose that the additive model that we have fit to the data is adequate, which, as we will see in Section 7.3.5, is not the case.

To test the null hypothesis of no partial effect of type of occupation,

$$H_0 : \gamma_1 = \gamma_2 = 0$$

we can calculate the incremental F -statistic

$$\begin{aligned} F_0 &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \\ &= \frac{98 - 4 - 1}{2} \times \frac{.83486 - .81400}{1 - .83486} = 5.874 \end{aligned} \quad (7.8)$$

with 2 and 93 degrees of freedom, for which $p = .0040$. The occupational-type effect is therefore statistically significant but (examining the coefficient standard errors) not very precisely estimated. The education and income coefficients are several times their respective standard errors and hence are highly statistically significant.⁸

7.2.1 Coefficient Quasi-Variances*

Consider a dummy-regression model with p quantitative explanatory variables and an m -category factor:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{m-1} D_{i,m-1} + \varepsilon_i$$

The dummy-variable coefficients $\gamma_1, \gamma_2, \dots, \gamma_{m-1}$ represent differences (or *contrasts*) between each of the other categories of the factor and the reference category m , holding constant X_1, \dots, X_p . If we are interested in a comparison between any other two categories, we can simply take the difference in their dummy-regressor coefficients. Thus, in the preceding example (letting $C_1 \equiv \hat{\gamma}_1$ and $C_2 \equiv \hat{\gamma}_2$),

$$C_1 - C_2 = 6.039 - (-2.737) = 8.776$$

is the estimated average difference in prestige between professional and white-collar occupations of equal income and education.

Suppose, however, that we want to know the standard error of $C_1 - C_2$. The standard errors of C_1 and C_2 are available directly in the regression “output” (Equation 7.7), but to compute the standard error of $C_1 - C_2$, we need in addition the estimated sampling covariance of these two coefficients. That is,⁹

$$\text{SE}(C_1 - C_2) = \sqrt{\hat{V}(C_1) + \hat{V}(C_2) - 2 \times \hat{C}(C_1, C_2)}$$

where $\hat{V}(C_j) = [\text{SE}(C_j)]^2$ is the estimated sampling variance of coefficient C_j , and $\hat{C}(C_1, C_2)$ is the estimated sampling covariance of C_1 and C_2 . For the occupational prestige regression, $\hat{C}(C_1, C_2) = 6.797$, and so

⁸In classical hypothesis testing, a result is “statistically significant” if the p -value for the null hypothesis is as small as or smaller than a preestablished level, typically .05. Strictly speaking, then, a result cannot be “highly statistically significant”—it is either statistically significant or not. I regard the phrase “highly statistically significant,” which appears commonly in research reports, as a harmless shorthand for a very small p -value, however, and will occasionally, as here, use it in this manner.

⁹See online Appendix D on probability and estimation. The computation of regression-coefficient covariances is taken up in Chapter 9.

$$\text{SE}(C_1 - C_2) = \sqrt{3.867^2 + 2.514^2 - 2 \times 6.797} = 2.771$$

We can use this standard error in the normal manner for a *t*-test of the difference between C_1 and C_2 .¹⁰ For example, noting that the difference exceeds twice its standard error suggests that it is statistically significant.

Although computer programs for regression analysis typically report the covariance matrix of the regression coefficients if asked to do so, it is not common to include coefficient covariances in published research along with estimated coefficients and standard errors, because with $k + 1$ coefficients in the model, there are $k(k + 1)/2$ variances and covariances among them—a potentially large number. Readers of a research report are therefore put at a disadvantage by the arbitrary choice of a reference category in dummy regression, because they are unable to calculate the standard errors of the differences between all pairs of categories of a factor.

Quasi-variances of dummy-regression coefficients (Firth, 2003; Firth & De Menezes, 2004) speak to this problem. Let $\tilde{V}(C_j)$ denote the quasi-variance of dummy coefficient C_j . Then,

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\tilde{V}(C_j) + \tilde{V}(C_{j'})}$$

The squared relative error of this approximation for the contrast $C_j - C_{j'}$ is

$$\text{RE}_{jj'} \equiv \frac{\tilde{V}(C_j - C_{j'})}{\widehat{V}(C_j - C_{j'})} = \frac{\tilde{V}(C_j) + \tilde{V}(C_{j'})}{\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})}$$

The approximation is accurate for this contrast when $\text{RE}_{jj'}$ is close to 1 or, equivalently, when

$$\log(\text{RE}_{jj'}) = \log[\tilde{V}(C_j) + \tilde{V}(C_{j'})] - \log[\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})]$$

is close to 0. The quasi-variances $\tilde{V}(C_j)$ are therefore selected to minimize the sum of squared log relative errors of approximation over all pairwise contrasts, $\sum_{j < j'} [\log(\text{RE}_{jj'})]^2$. The resulting errors of approximation are typically very small (Firth, 2003; Firth & De Menezes, 2004).

The following table gives dummy-variable coefficients, standard errors, and quasi-variances for type of occupation in the Canadian occupational prestige regression:

Category	C_j	$\text{SE}(C_j)$	$\tilde{V}(C_j)$
Professional	6.039	3.867	8.155
White collar	-2.737	2.514	-0.4772
Blue collar	0	0	6.797

I have set to 0 the coefficient (and its standard error) for the baseline category, blue collar. The negative quasi-variance for the white-collar coefficient is at first blush disconcerting (after all, ordinary variances cannot be negative), but it is not wrong: The quasi-variances are computed to provide accurate variance approximations for coefficient *differences*; they do not apply

¹⁰Testing all differences between pairs of factor categories raises an issue of simultaneous inference, however. See the discussion of Scheffé confidence intervals in Section 9.4.4.

directly to the coefficients themselves. For the contrast between professional and white-collar occupations, we have

$$\text{SE}(C_1 - C_2) \approx \sqrt{8.155 - 0.4772} = 2.771$$

Likewise, for the contrast between professional and blue-collar occupations,

$$C_1 - C_3 = 6.039 - 0 = 6.039$$

$$\text{SE}(C_1 - C_3) \approx \sqrt{8.155 + 6.797} = 3.867$$

Note that in this application, the quasi-variance “approximation” to the standard error proves to be exact, and indeed this is necessarily the case when there are just three factor categories, because there are then just three pairwise differences among the categories to capture.¹¹

7.3 Modeling Interactions

Two explanatory variables are said to *interact* in determining a response variable when the partial effect of one depends on the value of the other. The additive models that we have considered thus far therefore specify the *absence* of interactions. In this section, I will explain how the dummy-variable regression model can be modified to accommodate interactions between factors and quantitative explanatory variables.¹²

The treatment of dummy-variable regression in the preceding two sections has assumed parallel regressions across the several categories of a factor. If these regressions are *not* parallel, then the factor interacts with one or more of the quantitative explanatory variables. The dummy-regression model can easily be modified to reflect these interactions.

For simplicity, I return to the contrived example of Section 7.1, examining the regression of income on gender and education. Consider the hypothetical data shown in Figure 7.7 (and contrast these examples with those shown in Figure 7.1 on page 129, where the effects of gender and education are additive). In Figure 7.7(a) [as in Figure 7.1(a)], gender and education are independent, because women and men have identical education distributions; in Figure 7.7(b) [as in Figure 7.1(b)], gender and education are related, because women, on average, have higher levels of education than men.

It is apparent in both Figure 7.7(a) and Figure 7.7(b), however, that the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women. Because the effect of education varies by gender, education and gender interact in affecting income.

It is also the case, incidentally, that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes (indeed, grows) with education. Interaction, then, is a symmetric concept—that the effect of education varies by gender implies that the effect of gender varies by education (and, of course, vice versa).

The simple examples in Figures 7.1 and 7.7 illustrate an important and frequently misunderstood point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one

¹¹For the details of the computation of quasi-variances, see Chapter 15, Exercise 15.11.

¹²Interactions between factors are taken up in the next chapter on analysis of variance; interactions between quantitative explanatory variables are discussed in Section 17.1 on polynomial regression.

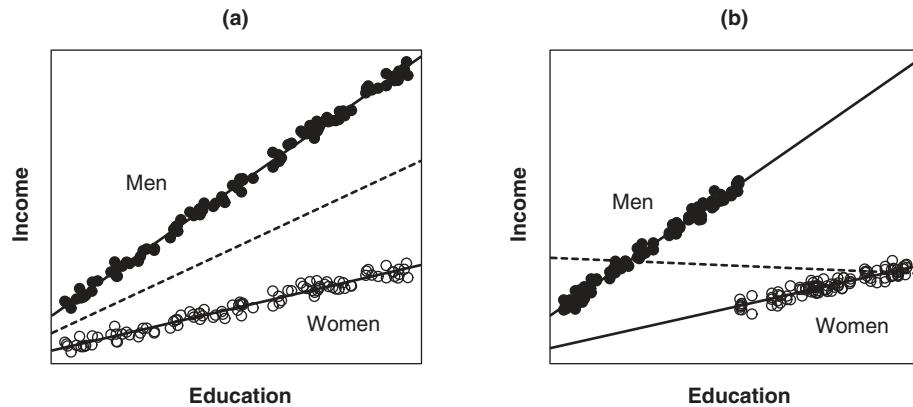


Figure 7.7 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both cases, the within-gender regressions (solid lines) are not parallel—the slope for men is greater than the slope for women—and, consequently, education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line.

another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves.

Two explanatory variables interact when the effect on the response variable of one depends on the value of the other. Interaction and correlation of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact whether or not they are related to one another statistically. Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

7.3.1 Constructing Interaction Regressors

We could model the data in Figure 7.7 by fitting separate regressions of income on education for women and men. As before, however, it is more convenient to fit a combined model, primarily because a combined model facilitates a test of the gender-by-education interaction. Moreover, a properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions: The full sample is composed of the two groups, and, consequently, the residual sum of squares for the full sample is minimized when the residual sum of squares is minimized in each group.¹³

¹³See Exercise 7.4.

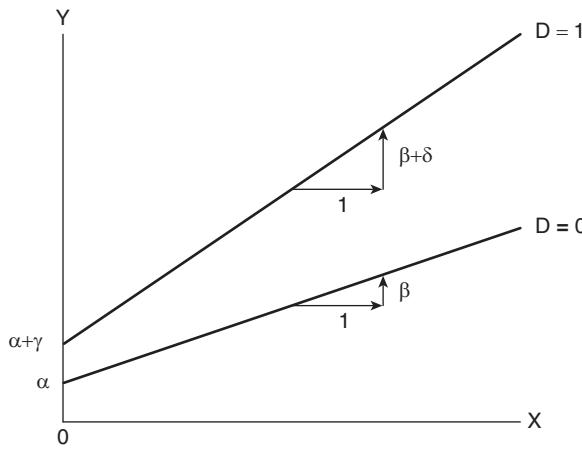


Figure 7.8 The dummy-variable regression model with an interaction regressor. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i \quad (7.9)$$

Along with the quantitative regressor X for education and the dummy regressor D for gender, I have introduced the *interaction regressor* XD into the regression equation. The interaction regressor is the *product* of the other two regressors; although XD is therefore a function of X and D , it is not a *linear* function, and perfect collinearity is avoided.¹⁴

For women, model (7.9) becomes

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)X_i + \varepsilon_i \end{aligned}$$

These regression equations are graphed in Figure 7.8: The parameters α and β are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); γ gives the difference in intercepts between the male and female groups; and δ gives the difference in slopes between the two groups. To test for interaction, therefore, we may simply test the hypothesis $H_0: \delta = 0$.

¹⁴If this procedure seems illegitimate, then think of the interaction regressor as a new variable, say $Z \equiv XD$. The model is linear in X , D , and Z . The “trick” of introducing an interaction regressor is similar to the trick of formulating dummy regressors to capture the effect of a factor: In both cases, there is a distinction between explanatory variables and regressors. Unlike a dummy regressor, however, the interaction regressor is a function of *both* explanatory variables.

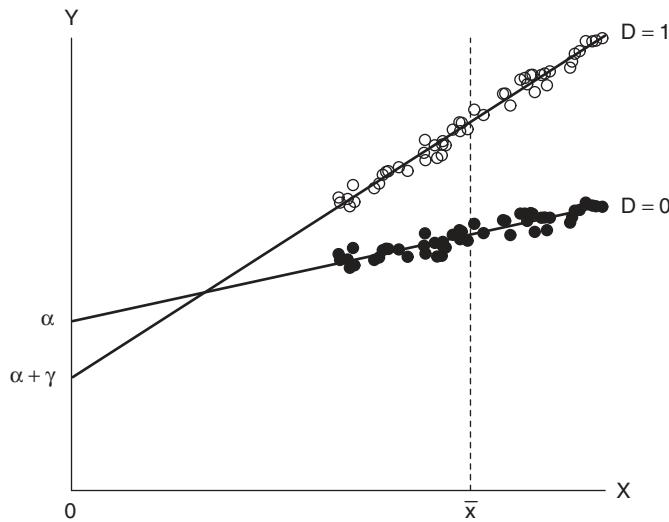


Figure 7.9 Why the difference in intercepts does not represent a meaningful partial effect for a factor when there is interaction: The difference-in-intercepts parameter γ is *negative* even though, within the range of the data, the regression line for the group coded $D = 1$ is *above* the line for the group coded $D = 0$.

Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The resulting model permits different slopes in different groups—that is, regression surfaces that are not parallel.

In the additive, no-interaction model of Equation 7.1 and Figure 7.2 (page 130), the dummy-regressor coefficient γ represents the *unique* partial effect of gender (i.e., the expected income difference between men and women of equal education, regardless of the value at which education is fixed), while the slope β represents the *unique* partial effect of education (i.e., the within-gender expected increment in income for a one-unit increase in education, for both women and men). In the interaction model of Equation 7.9 and Figure 7.8, in contrast, γ is no longer interpretable as the unqualified income difference between men and women of equal education.

Because the within-gender regressions are not parallel, the separation between the regression lines changes; here, γ is simply the separation at $X = 0$ —that is, above the origin. It is generally no more important to assess the expected income difference between men and women of 0 education than at other educational levels, and therefore the difference-in-intercepts parameter γ is not of special interest in the interaction model. Indeed, in many instances (although not here), the value $X = 0$ may not occur in the data or may be impossible (as, for example, if X is weight). In such cases, γ has no literal interpretation in the interaction model (see Figure 7.9).

Likewise, in the interaction model, β is not the unqualified partial effect of education but rather the effect of education among women. Although this coefficient *is* of interest, it is not

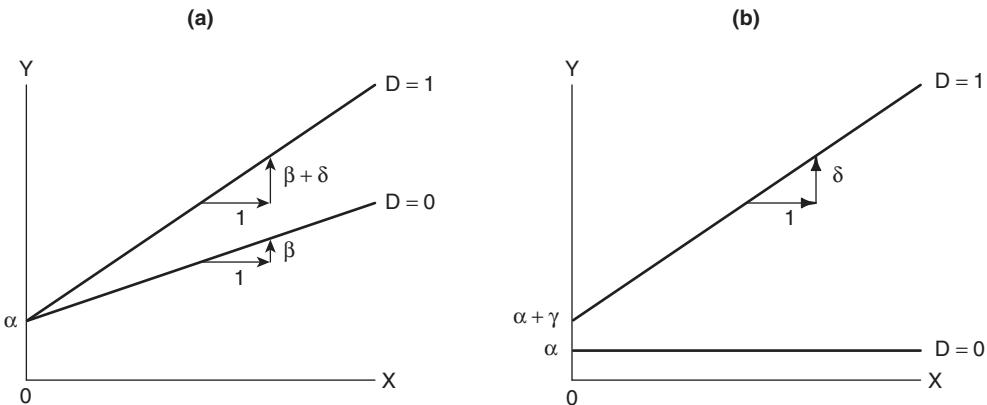


Figure 7.10 Two models that violate the principle of marginality: In (a), the dummy regressor D is omitted from the model $E(Y) = \alpha + \beta X + \delta(XD)$; in (b), the quantitative explanatory variable X is omitted from the model $E(Y) = \alpha + \gamma D + \delta(XD)$. These models violate the principle of marginality because they include the term XD , which is a higher-order relative of both X and D (one of which is omitted from each model).

necessarily more important than the effect of education among men ($\beta + \delta$), which does not appear *directly* in the model.

7.3.2 The Principle of Marginality

Following Nelder (1977), we say that the separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction. In general, we neither test nor interpret the main effects of explanatory variables that interact. If, however, we can rule out interaction either on theoretical or on empirical grounds, then we can proceed to test, estimate, and interpret the main effects.

As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction regressors but that omit main effects that are marginal to them. This is not to say that such models—which violate the *principle of marginality*—are uninterpretable: They are, rather, not broadly applicable.

The principle of marginality specifies that a model including a *high-order term* (such as an interaction) should normally also include the “lower-order relatives” of that term (the main effects that “compose” the interaction).

Suppose, for example, that we fit the model

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$

which omits the dummy regressor D but includes its “higher-order relative” XD . As shown in Figure 7.10(a), this model describes regression lines for women and men that have the same

intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest. Similarly, the model

$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

graphed in Figure 7.10(b), constrains the slope for women to 0, which is needlessly restrictive. Moreover, in this model, the choice of baseline category for D is consequential.

7.3.3 Interactions With Polytomous Factors

The method for modeling interactions by forming product regressors is easily extended to polytomous factors, to several factors, and to several quantitative explanatory variables. I will use the Canadian occupational prestige regression to illustrate the application of the method, entertaining the possibility that occupational type interacts both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = & \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ & + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i \end{aligned} \quad (7.10)$$

We require one interaction regressor for each product of a dummy regressor with a quantitative explanatory variable. The regressors $X_1 D_1$ and $X_1 D_2$ capture the interaction between income and occupational type; $X_2 D_1$ and $X_2 D_2$ capture the interaction between education and occupational type. The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{aligned} \text{Professional: } & Y_i = (\alpha + \gamma_1) + (\beta_1 + \delta_{11}) X_{i1} + (\beta_2 + \delta_{21}) X_{i2} + \varepsilon_i \\ \text{White collar: } & Y_i = (\alpha + \gamma_2) + (\beta_1 + \delta_{12}) X_{i1} + (\beta_2 + \delta_{22}) X_{i2} + \varepsilon_i \\ \text{Blue collar: } & Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned} \quad (7.11)$$

Blue-collar occupations, which are coded 0 for both dummy regressors, serve as the baseline for the intercepts and slopes of the other occupational types. As in the no-interaction model, the choice of baseline category is generally arbitrary, as it is here, and is inconsequential. Fitting the model in Equation 7.10 to the prestige data produces the following results:

$$\begin{aligned} \widehat{Y}_i = & 2.276 + 0.003522 X_1 + 1.713 X_2 + 15.35 D_1 - 33.54 D_2 \\ (7.057) & (0.000556) \quad (0.957) \quad (13.72) \quad (17.65) \\ & - 0.002903 X_1 D_1 - 0.002072 X_1 D_2 \\ & \quad (0.000599) \quad (0.000894) \\ & + 1.388 X_2 D_1 + 4.291 X_2 D_2 \\ & \quad (1.289) \quad (1.757) \\ R^2 = & .8747 \end{aligned} \quad (7.12)$$

This example is discussed further in the following section.

7.3.4 Interpreting Dummy-Regression Models With Interactions

It is difficult in dummy-regression models with interactions (and in other complex statistical models) to understand what the model is saying about the data simply by examining the

regression coefficients. One approach to interpretation, which works reasonably well in a relatively straightforward model such as Equation 7.12, is to write out the implied regression equation for each group (using Equations 7.11):

$$\begin{aligned}\text{Professional: } \widehat{\text{Prestige}} &= 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education} \\ \text{White collar: } \widehat{\text{Prestige}} &= -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education} \\ \text{Blue collar: } \widehat{\text{Prestige}} &= 2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education}\end{aligned}\quad (7.13)$$

From these equations, we can see, for example, that income appears to make much more difference to prestige in blue-collar occupations than in white-collar occupations and has even less impact on prestige in professional and managerial occupations. Education, in contrast, has the largest impact on prestige among white-collar occupations and has the smallest effect in blue-collar occupations.

An alternative approach (from Fox, 1987, 2003; Fox & Andersen, 2006) that generalizes readily to more complex models is to examine the high-order terms of the model. In the illustration, the high-order terms are the interactions between income and type and between education and type.

- Focusing in turn on each high-order term, we allow the variables in the term to range over their combinations of values in the data, fixing other variables to typical values. For example, for the interaction between type and income, we let type of occupation take on successively the categories blue collar, white collar, and professional [for which the dummy regressors D_1 and D_2 are set to the corresponding values given in the table (7.5) on page 136], in combination with income values between \$1500 and \$26,000 (the approximate range of income in the Canadian occupational prestige data set); education is fixed to its average value in the data, $\bar{X}_2 = 10.79$.
- We next compute the fitted value of prestige at each combination of values of income and type of occupation. These fitted values are graphed in the “effect display” shown in the upper panel of Figure 7.11; the lower panel of this figure shows a similar effect display for the interaction between education and type of occupation, holding income at its average value. The broken lines in Figure 7.11 give ± 2 standard errors around the fitted values—that is, approximate 95% pointwise confidence intervals for the effects.¹⁵ The nature of the interactions between income and type and between education and type is readily discerned from these graphs.

7.3.5 Hypothesis Tests for Main Effects and Interactions

To test the null hypothesis of no interaction between income and type, $H_0: \delta_{11} = \delta_{12} = 0$, we need to delete the interaction regressors $X_1 D_1$ and $X_1 D_2$ from the full model (Equation 7.10) and calculate an incremental F -test; likewise, to test the null hypothesis of no interaction between education and type, $H_0: \delta_{21} = \delta_{22} = 0$, we delete the interaction regressors $X_2 D_1$ and $X_2 D_2$ from the full model. These tests, and tests for the main effects of income, education, and occupational type, are detailed in Tables 7.1 and 7.2: Table 7.1 gives the regression sums of squares for several models, which, along with the residual sum of squares for the full model,

¹⁵For standard errors of fitted values, see Exercise 9.14.

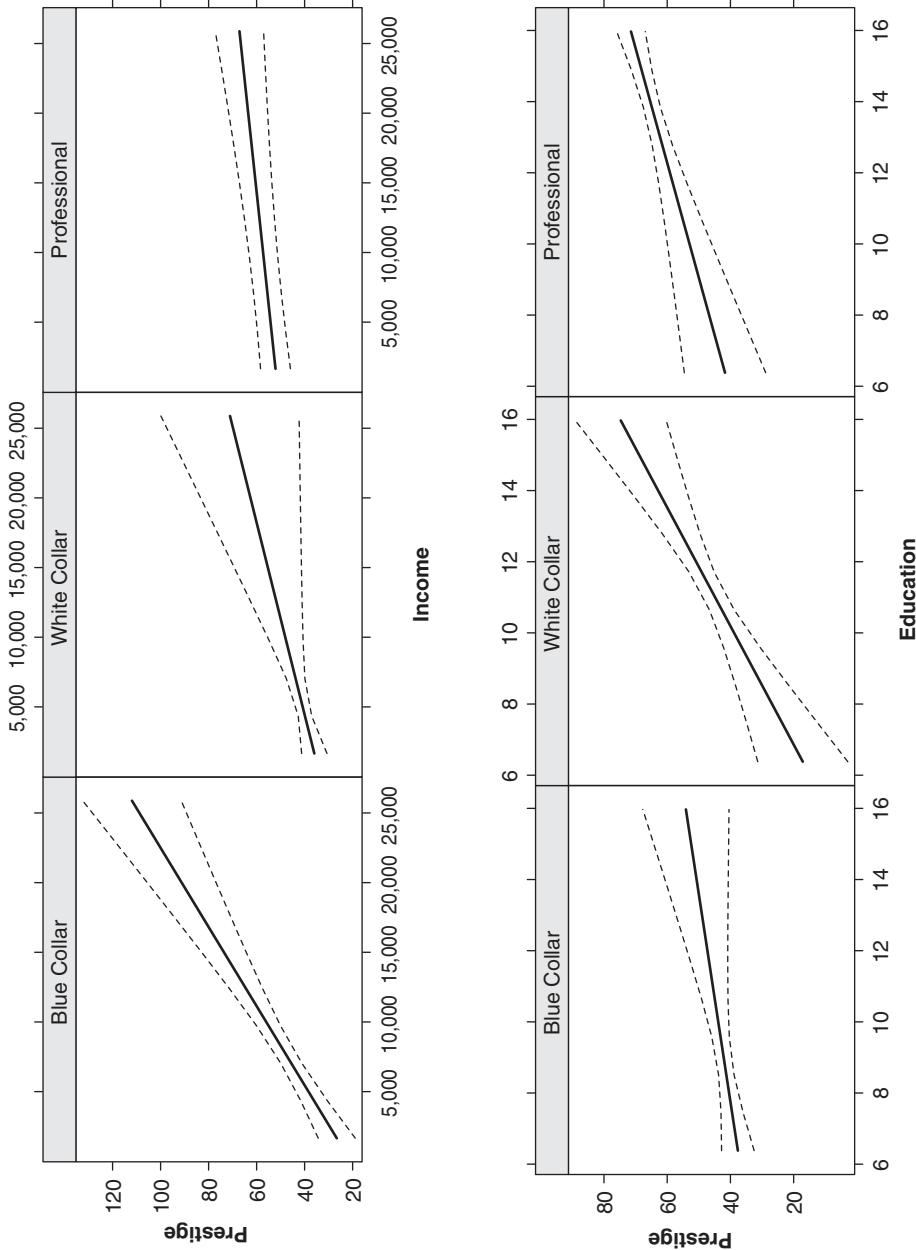


Figure 7.11 Income-by-type (upper panel) and education-by-type (lower panel) “effect displays” for the regression of prestige on income, education, and type of occupation. The solid lines give fitted values under the model, while the broken lines give 95% pointwise confidence intervals around the fit. To compute fitted values in the upper panel, education is set to its average value in the data; in the lower panel, income is set to its average value.

Table 7.1 Regression Sums of Squares for Several Models Fit to the Canadian Occupational Prestige Data

Model	Terms	Parameters	Regression	
			Sum of Squares	df
1	$I, E, T, I \times T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794.	8
2	$I, E, T, I \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	24,556.	6
3	$I, E, T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{22}	23,842.	6
4	I, E, T	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666.	4
5	I, E	α, β_1, β_2	23,074.	2
6	$I, T, I \times T$	$\alpha, \beta_1, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	23,488.	5
7	$E, T, E \times T$	$\alpha, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{22}	22,710.	5

NOTE: These sums of squares are the building blocks of incremental F -tests for the main and interaction effects of the explanatory variables. The following code is used for “terms” in the model: I , income; E , education; T , occupational type.

Table 7.2 Analysis-of-Variance Table, Showing Incremental F -Tests for the Terms in the Canadian Occupational Prestige Regression

Source	Models Contrasted	Sum of Squares	df	F	p
Income	3–7	1132.	1	28.35	<.0001
Education	2–6	1068.	1	26.75	<.0001
Type	4–5	592.	2	7.41	<.0011
Income \times Type	1–3	952.	2	11.92	<.0001
Education \times Type	1–2	238.	2	2.98	.056
Residuals		3553.	89		
Total		28,347.	97		

$RSS_1 = 3553$, are the building blocks of the incremental F -tests shown in Table 7.2. Table 7.3 shows the hypothesis tested by each of the incremental F -statistics in Table 7.2.

Although the analysis-of-variance table (Table 7.2) conventionally shows the tests for the main effects of education, income, and type before the education-by-type and income-by-type interactions, the structure of the model makes it sensible to examine the interactions first: Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of the main effect are 0 (as shown in Table 7.3). Thus, for example, the test for the income main effect assumes that the income-by-type

Table 7.3 Hypotheses Tested by the Incremental *F*-Tests in Table 7.2

Source	Models Contrasted	Null Hypothesis
Income	3–7	$\beta_1 = 0 \delta_{11} = \delta_{12} = 0$
Education	2–6	$\beta_2 = 0 \delta_{21} = \delta_{22} = 0$
Type	4–5	$\gamma_1 = \gamma_2 = 0 \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
Income × Type	1–3	$\delta_{11} = \delta_{12} = 0$
Education × Type	1–2	$\delta_{21} = \delta_{22} = 0$

interaction is absent (i.e., that $\delta_{11} = \delta_{12} = 0$), but not that the education-by-type interaction is absent ($\delta_{21} = \delta_{22} = 0$).¹⁶

The principle of marginality serves as a guide to constructing incremental *F*-tests for the terms in a model that includes interactions.

In this case, then, there is weak evidence of an interaction between education and type of occupation and much stronger evidence of an income-by-type interaction. Considering the small number of cases, we are squeezing the data quite hard, and it is apparent from the coefficient standard errors (in Equation 7.12) and from the effect displays in Figure 7.11 that the interactions are not precisely estimated. The tests for the main effects of income, education, and type, computed assuming that the higher-order relatives of each such term are absent, are all highly statistically significant. In light of the strong evidence for an interaction between income and type, however, the income and type main effects are not really of interest.¹⁷

The degrees of freedom for the several sources of variation add to the total degrees of freedom, but—because the regressors in different sets are correlated—the sums of squares do not add to the total sum of squares.¹⁸ What is important here (and more generally) is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

7.4 A Caution Concerning Standardized Coefficients

In Chapter 5, I explained the use—and limitations—of standardized regression coefficients. It is appropriate to sound another cautionary note here: Inexperienced researchers sometimes

¹⁶Tests constructed to conform to the principle of marginality are sometimes called “Type II” tests, terminology introduced by the SAS statistical software package. This terminology and alternative tests are described in the next chapter.

¹⁷We tested the occupational type main effect in Section 7.2 (Equation 7.8 on page 138), but using an estimate of error variance based on Model 4, which does not contain the interactions. In Table 7.2, the estimated error variance is based on the full model, Model 1. As mentioned in Chapter 6, sound general practice is to use the largest model fit to the data to estimate the error variance, even when, as is frequently the case, this model includes effects that are not statistically significant. The largest model necessarily has the smallest residual sum of squares, but it also has the fewest residual degrees of freedom. These two factors tend to offset one another, and it usually makes little difference whether the estimated error variance is based on the full model or on a model that deletes nonsignificant terms. Nevertheless, using the full model ensures an unbiased estimate of the error variance.

¹⁸See Section 10.2 for a detailed explanation of this phenomenon.

report standardized coefficients for dummy regressors. As I have explained, an *unstandardized* coefficient for a dummy regressor is interpretable as the expected response-variable difference between a particular category and the baseline category for the dummy-regressor set (controlling, of course, for the other explanatory variables in the model).

If a dummy-regressor coefficient is standardized, then this straightforward interpretation is lost. Furthermore, because a 0/1 dummy regressor cannot be increased by one standard deviation, the usual interpretation of a standardized regression coefficient also does not apply. Standardization is a linear transformation, so many characteristics of the regression model—the value of R^2 , for example—do not change, but the standardized coefficient itself is not directly interpretable. These difficulties can be avoided by standardizing only the response variable and *quantitative* explanatory variables in a regression, leaving dummy regressors in 0/1 form.

A similar point applies to interaction regressors. We may legitimately standardize a quantitative explanatory variable *prior* to taking its product with a dummy regressor, but to standardize the interaction regressor itself is not sensible: The interaction regressor cannot change independently of the main-effect regressors that compose it and are marginal to it.

It is not sensible to standardize dummy regressors or interaction regressors.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 7.1. Suppose that the values -1 and 1 are used for the dummy regressor D in Equation 7.1 instead of 0 and 1 . Write out the regression equations for men and women, and explain how the parameters of the model are to be interpreted. Does this alternative coding of the dummy regressor adequately capture the effect of gender? Is it fair to conclude that the dummy-regression model will “work” properly as long as two distinct values of the dummy regressor are employed, one each for women and men? Is there a reason to prefer one coding to another?

Exercise 7.2. Adjusted means (based on Section 7.2): Let \bar{Y}_1 represent the (“unadjusted”) mean prestige score of professional occupations in the Canadian occupational prestige data, \bar{Y}_2 that of white-collar occupations, and \bar{Y}_3 that of blue-collar occupations. Differences among the \bar{Y}_j may partly reflect differences among occupational types in their income and education levels. In the dummy-variable regression in Equation 7.7, type-of-occupation differences are “controlled” for income and education, producing the fitted regression equation

$$\hat{Y} = A + B_1X_1 + B_2X_2 + C_1D_1 + C_2D_2$$

Consequently, if we fix income and education at particular values—say, $X_1 = x_1$ and $X_2 = x_2$ —then the fitted prestige scores for the several occupational types are given by (treating “blue collar” as the baseline type):

$$\hat{Y}_1 = (A + C_1) + B_1x_1 + B_2x_2$$

$$\hat{Y}_2 = (A + C_2) + B_1x_1 + B_2x_2$$

$$\hat{Y}_3 = A + B_1x_1 + B_2x_2$$

- (a) Note that the *differences* among the \hat{Y}_j depend only on the dummy-variable coefficients C_1 and C_2 and not on the values of x_1 and x_2 . Why is this so?
- (b) When $x_1 = \bar{X}_1$ and $x_2 = \bar{X}_2$, the \hat{Y}_j are called *adjusted means* and are denoted \tilde{Y}_j . How can the adjusted means \tilde{Y}_j be interpreted? In what sense is \tilde{Y}_j an “adjusted” mean?
- (c) Locate the “unadjusted” and adjusted means for women and men in each of Figures 7.1(a) and (b) (on page 129). Construct a similar figure in which the difference between adjusted means is *smaller* than the difference in unadjusted means.
- (d) Using the results in the text, along with the mean income and education values for the three occupational types, compute adjusted mean prestige scores for each of the three types, controlling for income and education. Compare the adjusted with the unadjusted means for the three types of occupations and comment on the differences, if any, between them.

Exercise 7.3. Can the concept of an adjusted mean, introduced in Exercise 7.2, be extended to a model that includes interactions? If so, show how adjusted means can be found for the data in Figure 7.7(a) and (b) (on page 141).

Exercise 7.4. Verify that the regression equation for each occupational type given in Equations 7.13 (page 146) is identical to the results obtained by regressing prestige on income and education *separately* for each of the three types of occupations. Explain why this is the case.

Summary

- A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the variable and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant separation between the surfaces given by the coefficient of the dummy regressor.
- A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded 0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.
- Two explanatory variables interact when the effect on the response variable of one depends on the value of the other. Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The model permits different slopes in different groups—that is, regression surfaces that are not parallel.
- *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory

variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves.

- The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that “compose” the interaction). The principle of marginality also serves as a guide to constructing incremental F -tests for the terms in a model that includes interactions and for examining the effects of explanatory variables.
- It is not sensible to standardize dummy regressors or interaction regressors.

8

Analysis of Variance

I introduced the term *analysis of variance* in Chapter 5 to describe the partition of the response-variable sum of squares into “explained” and “unexplained” components, noting that this decomposition applies generally to linear models. For historical reasons, *analysis of variance* (abbreviated *ANOVA*) also refers to procedures for fitting and testing linear models in which the explanatory variables are categorical.¹

When there is a single factor (also termed a *classification*), these procedures are called *one-way ANOVA*, the subject of the first section of this chapter. Two factors produce *two-way* analysis of variance, three factors, *three-way ANOVA*, and so on. Two-way ANOVA is taken up in Section 8.2 and higher-way ANOVA in Section 8.3.

The dummy-variable regression model of the previous chapter incorporates both quantitative and categorical explanatory variables; in Section 8.4, we will examine an alternative formulation of this model called *analysis of covariance* (ANCOVA).

Finally, I will explain how *linear contrasts* can be used to “customize” hypothesis tests in ANOVA and in linear models more generally.

Readers desiring a basic introduction to analysis of variance, bypassing most of the subtleties and details, can, without loss of coherence, read Section 8.1 on one-way ANOVA through Subsection 8.1.1 and Section 8.2 on two-way ANOVA through Subsection 8.2.2. These parts of the chapter explain how to perform analysis of variance using dummy-variable regressors.

8.1 One-Way Analysis of Variance

In Chapter 7, we learned how to construct dummy regressors to represent the effects of factors alongside those of quantitative explanatory variables. Suppose, however, that there are *no* quantitative explanatory variables—only a single factor. For example, for a three-category classification, we have the model

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i \quad (8.1)$$

employing the following coding for the dummy regressors:

¹The methods and terminology of analysis of variance were introduced by the great British statistician R. A. Fisher (1925). Fisher’s many other seminal contributions to statistics include the technique of randomization in experimental design and the method of maximum likelihood.

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

The expectation of the response variable in each *group* (i.e., in each category or *level* of the factor) is the population group mean, denoted μ_j for the j th group. Because the error ε has a mean of 0 under the usual linear-model assumptions, taking the expectation of both sides of the model (Equation 8.1) produces the following relationships between group means and model parameters:

$$\text{Group 1: } \mu_1 = \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1$$

$$\text{Group 2: } \mu_2 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2$$

$$\text{Group 3: } \mu_3 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha$$

There are three parameters (α , γ_1 , and γ_2) and three group means, so we can solve uniquely for the parameters in terms of the group means:

$$\alpha = \mu_3$$

$$\gamma_1 = \mu_1 - \mu_3$$

$$\gamma_2 = \mu_2 - \mu_3$$

It is not surprising that α represents the mean of the baseline category (Group 3) and that γ_1 and γ_2 capture differences between the other group means and the mean of the baseline category.

One-way ANOVA focuses on testing for differences among group means. The omnibus F -statistic for the model (Equation 8.1) tests $H_0: \gamma_1 = \gamma_2 = 0$, which corresponds to $H_0: \mu_1 = \mu_2 = \mu_3$, the null hypothesis of no differences among the population group means. Our consideration of one-way ANOVA might well end here, but for a desire to develop methods that generalize easily to more complex situations in which there are several, potentially interacting, factors. Indeed, as I will explain,² we can employ 0/1 dummy regressors even when there are two (or more) factors.

One-way ANOVA examines the relationship between a quantitative response variable and a factor. The omnibus F -statistic for the regression of the response variable on 0/1 dummy regressors constructed from the factor tests for differences in the response means across levels of the factor.

²See Section 8.2.2.

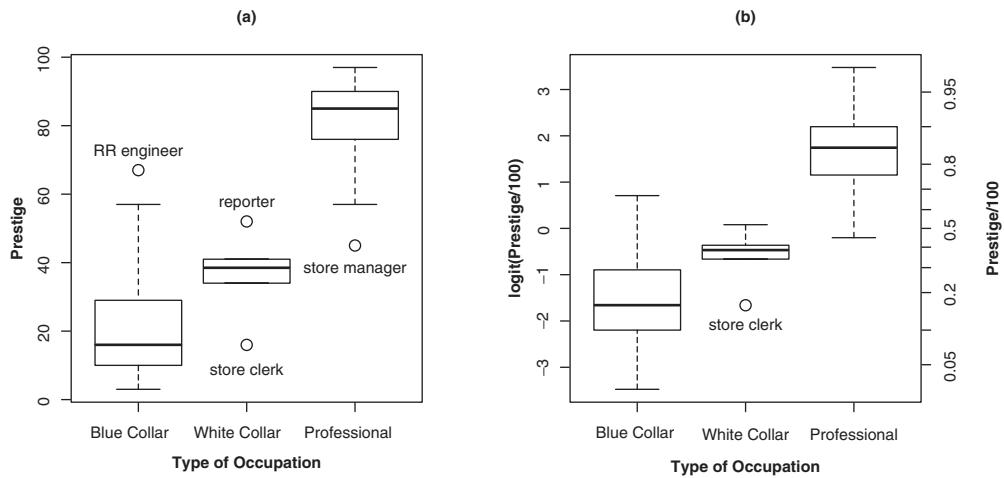


Figure 8.1 Parallel boxplots for (a) occupational prestige and (b) the logit of occupational prestige by type of occupation.

8.1.1 Example: Duncan's Data on Occupational Prestige

I will use Duncan's data on the prestige of 45 U.S. occupations to illustrate one-way ANOVA.³ Parallel boxplots for prestige in three types of occupations appear in Figure 8.1(a). Prestige, recall, is a percentage, and the data in Figure 8.1(a) push both the lower and upper boundaries of 0% and 100%, suggesting the logit transformation in Figure 8.1(b).⁴ The data are better behaved on the logit scale, which eliminates the skew in the blue-collar and professional groups and pulls in all the outlying observations, with the exception of store clerks in the white-collar category.

Means, standard deviations, and frequencies for prestige within occupational types are as follows:

Type of Occupation	Prestige		
	Mean	Standard Deviation	Frequency
Professional and managerial	80.44	14.11	18
White collar	36.67	11.79	6
Blue collar	22.76	18.05	21

Professional occupations therefore have the highest average level of prestige, followed by white-collar and blue-collar occupations. The order of the group means is the same on the logit scale:

³Duncan's data were introduced in Chapter 3.

⁴The logit transformation of proportions was introduced in Section 4.5.

Type of Occupation	<i>logit(Prestige/100)</i>	
	Mean	Standard Deviation
Professional and managerial	1.6321	0.9089
White collar	-0.5904	0.5791
Blue collar	-1.4821	1.0696

On both scales, the standard deviation is greatest among the blue-collar occupations and smallest among the white-collar occupations, but the differences are not very large, especially considering the small number of observations in the white-collar category.⁵

Using the logit of prestige as the response variable, the one-way ANOVA for the Duncan data is

Source	Sum of Squares	df	Mean Square	F	p
Type of Occupation	95.550	2	47.775	51.98	<.0001
Residuals	38.604	42	0.919		
Total	134.154	44			

We therefore have very strong evidence against the null hypothesis of no difference in average level of prestige across the occupational types. Occupational types account for nearly three quarters of the variation in the logit of prestige among these occupations ($R^2 = 95.550 / 134.154 = 0.712$).

8.1.2 The One-Way ANOVA Model

The first innovation is notational: Because observations are partitioned according to groups, it is convenient to let Y_{ij} denote the i th observation within the j th of m groups. The number of observations in the j th group is n_j , and therefore the total number of observations is $n = \sum_{j=1}^m n_j$. As above, $\mu_j \equiv E(Y_{ij})$ represents the population mean in group j .

The one-way ANOVA model is written in the following manner:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad (8.2)$$

where we would like μ to represent, in some reasonable sense, the general level of the response variable in the population; α_j should represent the effect on the response variable of membership in the j th group; and ε_{ij} is an error variable that follows the usual linear-model assumptions—that is, the ε_{ij} are independent and normally distributed with zero expectations and equal variances.

Upon taking expectations, Equation 8.2 becomes

⁵The assumption of constant error variance implies that the *population* variances should be the same in the several groups. See Section 12.4.2 for further discussion of this point and a test for nonconstant variance in ANOVA.

$$\mu_j = \mu + \alpha_j$$

The parameters of the model are, therefore, underdetermined, for there are $m + 1$ parameters (including μ) but only m population group means. For example, for $m = 3$, we have four parameters but only three equations:

$$\begin{aligned}\mu_1 &= \mu + \alpha_1 \\ \mu_2 &= \mu + \alpha_2 \\ \mu_3 &= \mu + \alpha_3\end{aligned}$$

Even if we knew the three population group means, we could not solve uniquely for the parameters.

Because the parameters of the model (Equation 8.2) are themselves underdetermined, they cannot be uniquely estimated. To estimate the model, we would need to code one dummy regressor for each group-effect parameter α_j , and—as we discovered in the previous chapter—the resulting dummy regressors would be perfectly collinear.

One convenient way out of this dilemma is to place a linear restriction on the parameters of the model, of the form

$$w_0\mu + \sum_{j=1}^m w_j\alpha_j = 0$$

where the w s are prespecified constants, not all equal to 0. It turns out that *any* such restriction will do, in the sense that all linear restrictions yield the same F -test for the null hypothesis of no differences in population group means.⁶ For example, if we employ the restriction $\alpha_m = 0$, we are, in effect, deleting the parameter for the last category, making it a baseline category. The result is the dummy-coding scheme of the previous chapter. Alternatively, we could use the restriction $\mu = 0$, which is equivalent to deleting the constant term from the linear model, in which case the “effect” parameters and group means are identical: $\alpha_j = \mu_j$ —an especially simple solution.

There is, however, an advantage in selecting a restriction that produces easily interpretable parameters and estimates and that generalizes usefully to more complex models. For these reasons, we will impose the constraint

$$\sum_{j=1}^m \alpha_j = 0 \tag{8.3}$$

Equation 8.3 is often called a *sigma constraint* or *sum-to-zero constraint*. Employing this restriction to solve for the parameters produces

$$\begin{aligned}\mu &= \frac{\sum \mu_j}{m} \equiv \mu. \\ \alpha_j &= \mu_j - \mu.\end{aligned}\tag{8.4}$$

The dot (in μ) indicates averaging over the range of a subscript, here over groups. The *grand* or *general mean* μ , then, is the average of the population group means, while α_j gives the

⁶See Section 10.4 for an explanation of this surprising result.

difference between the mean of group j and the grand mean.⁷ It is clear that, under the sigma constraint, the hypothesis of no differences in group means

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

is equivalent to the hypothesis that all of the effect parameters are 0:

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

All of this is well and good, but how can we estimate the one-way ANOVA model under the sigma constraint? One approach is to code *deviation regressors*, an alternative to the dummy-coding scheme, which (recall) implicitly imposes the constraint $\alpha_m = 0$. We require $m - 1$ deviation regressors, S_1, S_2, \dots, S_{m-1} , the j th of which is coded according to the following rule:⁸

$$S_j = \begin{cases} 1 & \text{for observations in group } j \\ -1 & \text{for observations in group } m \\ 0 & \text{for observations in all other groups} \end{cases}$$

For example, when $m = 3$,

Group	(α_1)	(α_2)
	S_1	S_2
1	1	0
2	0	1
3	-1	-1

For ease of reference, I have shown in parentheses the parameter associated with each deviation regressor.

Writing out the equations for the group means in terms of the deviation regressors demonstrates how these regressors capture the sigma constraint on the parameters of the model:

$$\begin{aligned} \text{Group 1: } \mu_1 &= \mu + 1 \times \alpha_1 + 0 \times \alpha_2 = \mu + \alpha_1 \\ \text{Group 2: } \mu_2 &= \mu + 0 \times \alpha_1 + 1 \times \alpha_2 = \mu + \alpha_2 \\ \text{Group 3: } \mu_3 &= \mu - 1 \times \alpha_1 - 1 \times \alpha_2 = \mu - \alpha_1 - \alpha_2 \end{aligned}$$

The equation for the third group incorporates the sigma constraint because $\alpha_3 = -\alpha_1 - \alpha_2$ is equivalent to $\alpha_1 + \alpha_2 + \alpha_3 = 0$.

The null hypothesis of no differences among population group means is tested by the omnibus F -statistic for the deviation-coded model: The omnibus F -statistic tests the hypothesis $H_0: \alpha_1 = \alpha_2 = 0$, which, under the sigma constraint, implies that α_3 is 0 as well.

⁷There is a subtle distinction between μ (the mean of the group means) and the overall (i.e., unconditional) mean of Y in the population. In a real population, μ and $E(Y)$ will generally differ if the groups have different numbers of observations. In an infinite or hypothetical population, we can speak of the grand mean but not of the overall (unconditional) mean $E(Y)$.

⁸I use S_j (for “sum-to-zero”) to distinguish these from the $(0, 1)$ dummy regressors D_j defined previously.

The one-way ANOVA model $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ is underdetermined because it uses $m + 1$ parameters to model m group means. This indeterminacy can be removed, however, by placing a restriction on its parameters. Setting one of the α_j s to 0 leads to $(0, 1)$ dummy-regressor coding. Constraining the α_j s to sum to 0 leads to $(1, 0, -1)$ deviation-regressor coding. The two coding schemes are equivalent in that they provide the same fit to the data, producing the same regression and residual sums of squares, and hence the same F -test for differences among group means.

Although it is often convenient to fit the one-way ANOVA model by least-squares regression, it is also possible to estimate the model and calculate sums of squares directly. The sample mean \bar{Y}_j in group j is the least-squares estimator of the corresponding population mean μ_j . Estimates of μ and the α_j may therefore be written as follows (substituting estimates into Equations 8.4):

$$\begin{aligned} M &\equiv \hat{\mu} = \frac{\sum \bar{Y}_j}{m} = \bar{Y}. \\ A_j &\equiv \hat{\alpha}_j = \bar{Y}_j - \bar{Y}. \end{aligned}$$

Furthermore, the fitted Y values are the group means:

$$\hat{Y}_{ij} = M + A_j = \bar{Y}. + (\bar{Y}_j - \bar{Y}.) = \bar{Y}_j$$

and the regression and residual sums of squares therefore take particularly simple forms in one-way ANOVA:⁹

$$\begin{aligned} \text{RegSS} &= \sum_{j=1}^m \sum_{i=1}^{n_j} (\hat{Y}_{ij} - \bar{Y})^2 = \sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2 \\ \text{RSS} &= \sum_{j=1}^m \sum_{i=1}^{n_j} (Y_{ij} - \hat{Y}_{ij})^2 = \sum \sum (Y_{ij} - \bar{Y}_j)^2 \end{aligned}$$

This information can be presented in an ANOVA table, as shown in Table 8.1.¹⁰

8.2 Two-Way Analysis of Variance

The inclusion of a second factor permits us to model and test partial relationships, as well as to introduce interactions. Most issues pertaining to ANOVA can be developed for the two-factor “design.” Before immersing ourselves in the details of model specification and hypothesis testing for two-way ANOVA, however, it is useful to step back and consider the patterns of relationship that can occur when a quantitative response variable is classified by two factors.

⁹If the n_j are unequal, as is usually the case in observational research, then the mean of the group means \bar{Y}_j generally differs from the overall sample mean \bar{Y} of the response variable, for $\bar{Y} = (\sum \sum Y_{ij})/n = (\sum n_j \bar{Y}_j)/n$, while $\bar{Y}_j = (\sum \bar{Y}_{ij})/m$. (See Footnote 3 for a similar point with respect to population means.)

¹⁰Although the notation may differ, this ANOVA table corresponds to the usual treatment of one-way ANOVA in introductory statistics texts. It is common to call the regression sum of squares in one-way ANOVA “the between-group sum of squares” and the residual sum of squares “the within-group sum of squares.”

Table 8.1 General One-Way Analysis-of-Variance Table

Source	Sum of Squares	df	Mean Square	F	H_0
Groups	$\sum n_j(\bar{Y}_j - \bar{Y})^2$	$m-1$	$\frac{\text{RegSS}}{m-1}$	$\frac{\text{RegMS}}{\text{RMS}}$	$\alpha_1 = \dots = \alpha_m = 0$ $(\mu_1 = \dots = \mu_m)$
Residuals	$\sum \sum (Y_{ij} - \bar{Y}_j)^2$	$n-m$	$\frac{\text{RSS}}{n-m}$		
Total	$\sum \sum (Y_{ij} - \bar{Y})^2$	$n-1$			

8.2.1 Patterns of Means in the Two-Way Classification

So as not to confuse ourselves with issues of estimation, we will imagine at the outset that we have access to population means. The notation for the two-way classification is shown in the following table:

	C_1	C_2	\dots	C_c	
R_1	μ_{11}	μ_{12}	\dots	μ_{1c}	$\mu_{1\cdot}$
R_2	μ_{21}	μ_{22}	\dots	μ_{2c}	$\mu_{2\cdot}$
\vdots	\vdots	\vdots		\vdots	\vdots
R_r	μ_{r1}	μ_{r2}	\dots	μ_{rc}	$\mu_{r\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	\dots	$\mu_{\cdot c}$	$\mu_{\cdot \cdot}$

The factors, R and C (for “rows” and “columns” of the table of means), have r and c categories, respectively. The factor categories are denoted R_j and C_k .

Within each *cell* of the design—that is, for each combination of categories $\{R_j, C_k\}$ of the two factors—there is a population cell mean μ_{jk} for the response variable. Extending the dot notation introduced in the previous section,

$$\mu_{j\cdot} \equiv \frac{\sum_{k=1}^c \mu_{jk}}{c}$$

is the *marginal mean* of the response variable in row j ;

$$\mu_{\cdot k} \equiv \frac{\sum_{j=1}^r \mu_{jk}}{r}$$

is the marginal mean in column k ; and

$$\mu_{\cdot \cdot} \equiv \frac{\sum_j \sum_k \mu_{jk}}{r \times c} = \frac{\sum_j \mu_{j\cdot}}{r} = \frac{\sum_k \mu_{\cdot k}}{c}$$

is the grand mean.

If R and C do not interact in determining the response variable, then the partial relationship between each factor and Y does not depend on the category at which the other factor is “held constant.” The difference in cell means $\mu_{jk} - \mu_{j'k}$ across two categories of R (i.e., categories R_j and $R_{j'}$) is constant across all the categories of C —that is, this difference is the same for all

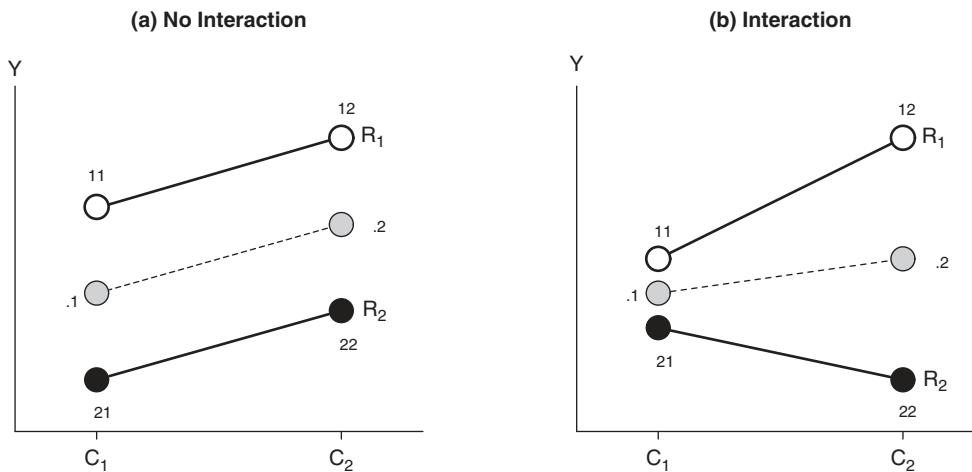


Figure 8.2 Interaction in the two-way classification. In (a), the parallel profiles of means (given by the white and black circles connected by solid lines) indicate that R and C do not interact in affecting Y . The R -effect—that is, the difference between the two profiles—is the same at both C_1 and C_2 . Likewise, the C -effect—that is, the rise in the line from C_1 to C_2 —is the same for both profiles. In (b), the R -effect differs at the two categories of C , and the C -effect differs at the two categories of R : R and C interact in affecting Y . In both graphs, the column marginal means $\mu_{.1}$ and $\mu_{.2}$ are shown as averages of the cell means in each column (represented by the gray circles connected by broken lines).

$k = 1, 2, \dots, c$. Consequently, the difference in cell means across rows is equal to the corresponding difference in the row marginal means:

$$\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'} = \mu_{j\cdot} - \mu_{j'\cdot} \text{ for all } j, j' \text{ and } k, k'$$

This pattern is illustrated in Figure 8.2(a) for the simple case where $r = c = 2$. Interaction—where the row difference $\mu_{1k} - \mu_{2k}$ changes across columns $k = 1, 2$ —is illustrated in Figure 8.2(b). Note that no interaction implies parallel “profiles” of cell means. Parallel profiles also imply that the column difference $\mu_{j1} - \mu_{j2}$ for categories C_1 and C_2 is constant across rows $j = 1, 2$ and is equal to the difference in column marginal means $\mu_{.1} - \mu_{.2}$. As we discovered in Chapter 7 on dummy regression, interaction is a symmetric concept: If R interacts with C , then C interacts with R . When interactions are absent, the partial effect of each factor—the factor’s *main effect*—is given by differences in the population marginal means.

Several patterns of relationship in the two-way classification, all showing no interaction, are graphed in Figure 8.3. Plots of means, incidentally, not only serve to clarify the ideas underlying ANOVA but are also a useful tool for summarizing and presenting data. Indeed, it is very difficult to inspect, understand, and interpret patterns of means in ANOVA *without* plotting the means. In the illustrations, factor C has three levels, which are marked off along the horizontal axis. Because C is a qualitative variable, the order of its categories and the spacing between

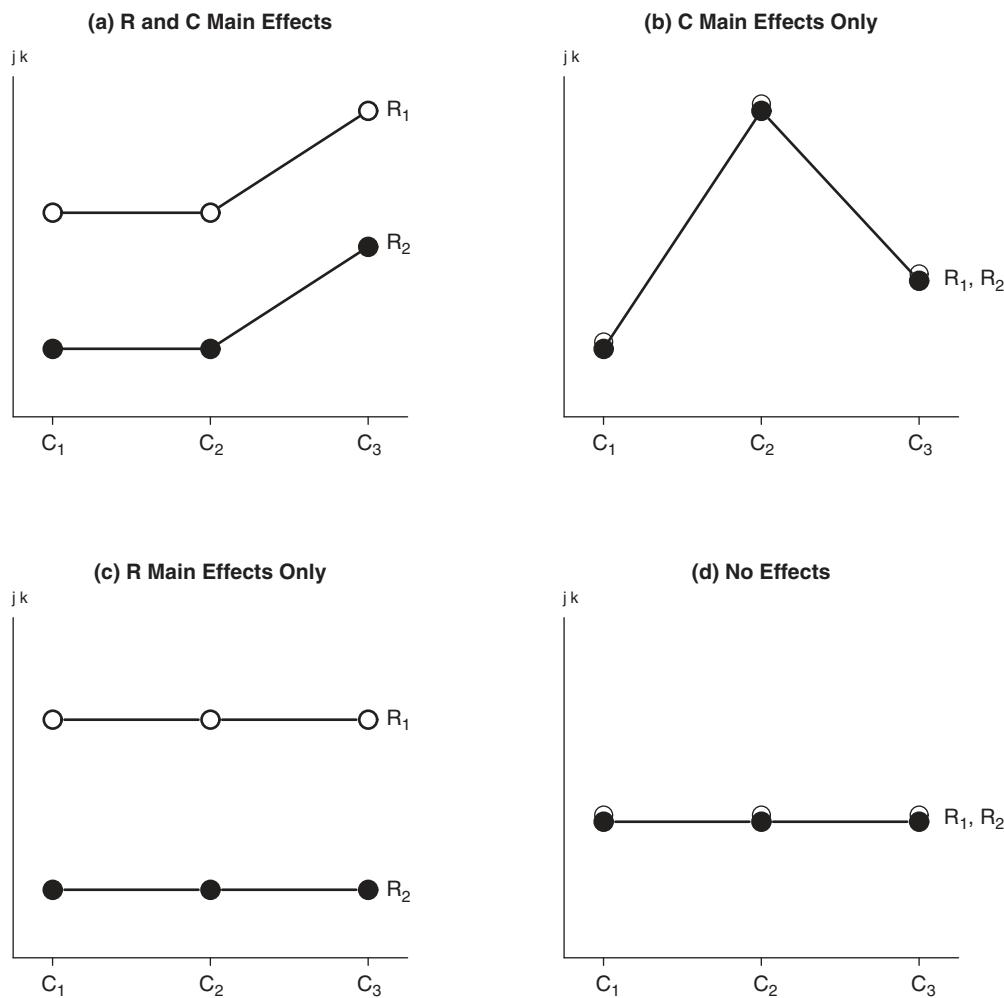


Figure 8.3 Several patterns of relationship in the two-way classification. In all these cases, R and C do not interact. (a) Both R and C main effects. (b) C main effects (R main effects nil). (c) R main effects (C main effects nil). (d) No effects (both R and C main effects nil).

them are arbitrary.¹¹ Factor R has two categories. The six cell means are plotted as points, connected by lines (profiles) according to the levels of factor R . The separation between the lines at level C_k (where k is 1, 2, or 3) represents the difference $\mu_{1k} - \mu_{2k}$. As noted above, when there is no interaction, therefore, the separation between the profiles is constant and the profiles themselves are parallel.

¹¹ANOVA is also useful when the levels of a factor are ordered (“low,” “medium,” and “high,” for example) or even discrete and quantitative (e.g., number of bedrooms for apartment dwellers—0, 1, 2, 3, 4), but, in general, I will assume that factors are simply nominal (i.e., qualitative) variables.

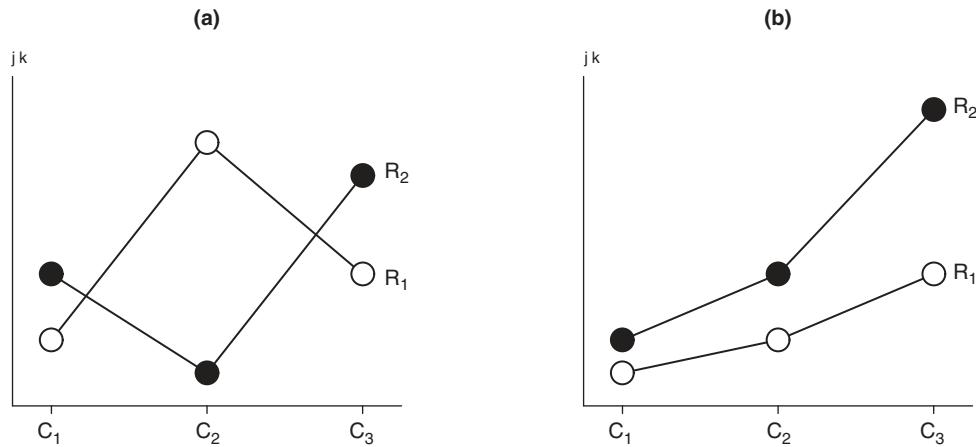


Figure 8.4 Two patterns of interaction in the two-way classification. In (a), the interaction is “disordinal” in that the order of means for one factor changes across the levels of the other factor. In (b), the profiles are not parallel, but the order of means does not change.

In Figure 8.3(a), both R and C have nonzero main effects. In Figure 8.3(b), the differences $\mu_{1k} - \mu_{2k} = \mu_1 - \mu_2$ are 0, and consequently, the R main effects are nil. In Figure 8.3(c), the C main effects are nil, because the differences $\mu_{jk} - \mu_{j'k} = \mu_{\cdot k} - \mu_{\cdot k'}$ are all 0. Finally, in Figure 8.3(d), both sets of main effects are nil.

Figure 8.4 shows two different patterns of interactions. It is clear from the previous discussion that R and C interact when the profiles of means are not parallel—that is, when the row differences $\mu_{jk} - \mu_{j'k}$ change across the categories of the column factor or, equivalently, when the column differences $\mu_{jk} - \mu_{jk'}$ change across the categories of the row factor. In Figure 8.4(a), the interaction is dramatic: The mean for level R_2 is above the mean for R_1 at levels C_1 and C_3 , but at level C_2 , the mean for R_1 is substantially above the mean for R_2 . Likewise, the means for the three categories of C are ordered differently within R_1 and R_2 . Interaction of this sort is sometimes called *disordinal*. In Figure 8.4(b), in contrast, the profile for R_2 is above that for R_1 across all three categories of C , although the separation between the profiles of means changes. This less dramatic form of interaction can sometimes be transformed away (e.g., by taking logs).

Even when interactions are absent in the population, we cannot expect perfectly parallel profiles of *sample* means: There is, of course, sampling error in sampled data. We have to determine whether departures from parallelism observed in a sample are sufficiently large to be statistically significant or whether they could easily be the product of chance. Moreover, in large samples, we also want to determine whether “statistically significant” interactions are of sufficient magnitude to be of substantive interest. We may well decide to ignore interactions that are statistically significant but trivially small.

In general, however, if we conclude that interactions are present and nonnegligible, then we do not interpret the main effects of the factors—after all, to conclude that two variables interact

Table 8.2 Conformity by Authoritarianism and Partner's Status, for Moore and Krupat's (1971) Experiment

Partner's Status		Authoritarianism		
		Low	Medium	High
Low	\bar{Y}_{jk}	8.900	7.250	12.63
	S_{jk}	2.644	3.948	7.347
	n_{jk}	10	4	8
High	\bar{Y}_{jk}	17.40	14.27	11.86
	S_{jk}	4.506	3.952	3.934
	n_{jk}	5	11	7

NOTE: Each cell shows (from top to bottom) the conformity mean and standard deviation, as well as the cell frequency.

is to deny that they have *unique* partial effects. This point is a reflection of the principle of marginality, introduced in Chapter 7 in the context of dummy-variable regression: Here, the R and C main effects are marginal to the RC interaction.¹²

Two factors interact when the profiles of population means are not parallel; when the profiles of means are parallel, the effects of the two factors are additive.

Example: Moore and Krupat's Conformity Experiment

Table 8.2 shows means, standard deviations, and cell frequencies for data from a social-psychological experiment reported by Moore and Krupat (1971).¹³ The experiment was designed to determine how the relationship between conformity and social status is influenced by “authoritarianism.” The subjects in the experiment were asked to make perceptual judgments of stimuli that were intrinsically ambiguous. On forming an initial judgment, the subjects were presented with the judgment of another individual (their “partner”) who was ostensibly participating in the experiment; the subjects were then asked for a final judgment. In fact, the partner’s judgments were manipulated by the experimenters so that subjects were faced with nearly continuous disagreement.

The measure of conformity employed in the study was the number of times in 40 critical trials that subjects altered their judgments in response to disagreement. This measure is a disguised proportion (but because it does not push the boundaries of 0 and 40, I leave the response variable untransformed in the analysis reported below). The 45 university student

¹²In cases of disordinal interaction, such as in Figure 8.4(a), interpreting main effects is clearly misleading because it makes no sense to average over levels of one factor to examine the effect of the other. In cases such as Figure 8.4(b), however, there may be some sense to examining the marginal means for one factor averaged over levels of the other, despite the interaction.

¹³The data were generously made available by James Moore, Department of Sociology, York University.

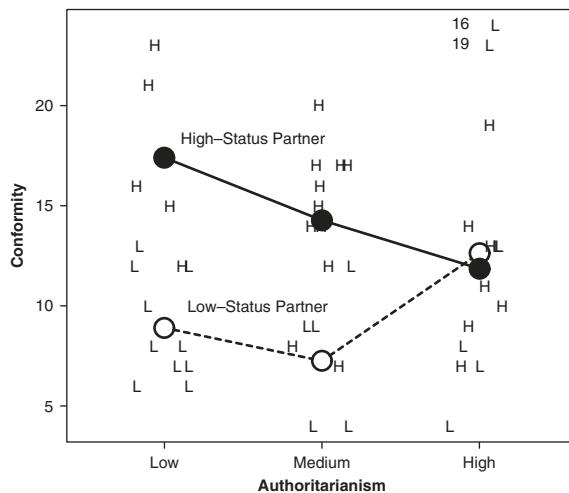


Figure 8.5 The data and cell means for Moore and Krupat's conformity experiment. The black circles connected by solid lines give the means for the high-status partner condition (with the data values represented by Hs); the white circles connected by broken lines give the means for the low-status partner condition (with the data values represented by Ls). The points are jittered horizontally to reduce overplotting. There are two outlying subjects (Numbers 16 and 19) in the high-authoritarianism, lower-partner status group.

subjects in the study were randomly assigned to two experimental conditions: In one condition, the partner was described as of relatively high social status (a “physician”); in the other condition, the partner was described as of relatively low status (a “postal clerk”).

A standard authoritarianism scale (the “F-scale”) was administered to the subjects after the experiment was completed. This procedure was dictated by practical considerations, but it raises the possibility that authoritarianism scores were inadvertently influenced by the experimental manipulation of the partner’s status. The authors divided the authoritarianism scores into three categories—low, medium, and high.¹⁴ A chi-square test of independence for the condition-by-authoritarianism frequency table (shown in Table 8.2) produces a *p*-value of .08, indicating that there is some ground for believing that the status manipulation affected the authoritarianism scores of the subjects.

Because of the conceptual-rigidity component of authoritarianism, Moore and Krupat expected that low-authoritarian subjects would be *more* responsive than high-authoritarian subjects to the social status of their partner. In other words, authoritarianism and partner’s status are expected to interact—in a particular manner—in determining conformity. The cell means, graphed along with the data in Figure 8.5, appear to confirm the experimenters’ expectations.

¹⁴Moore and Krupat categorized authoritarianism *separately* within each condition. This approach is not strictly justified, but it serves to produce nearly equal cell frequencies—required by the method of computation employed by the authors—for the six combinations of partner’s status and authoritarianism and yields results similar to those reported here. It may have occurred to you that the dummy-regression procedures of the previous chapter are applicable here and do not require the arbitrary categorization of authoritarianism. This analysis appears in Section 8.4. Moore and Krupat do report the difference between slopes for the within-condition regressions of conformity on authoritarianism.

The standard deviation of conformity in one cell (high-authoritarian, low-status partner) is appreciably larger than in the others. Upon inspection of the data, it is clear that the relatively large dispersion in this cell is due to two subjects, Numbers 16 and 19, who have atypically high conformity scores of 24 and 23.¹⁵

8.2.2 Two-Way ANOVA by Dummy Regression

When there are two factors, we can model their main effects and interactions by coding dummy regressors for each factor and forming all pairwise products between them. We require $r - 1$ dummy regressors to represent the r levels of the row factor R , $c - 1$ dummy regressors to represent the c levels of the column factor C , and consequently $(r - 1)(c - 1)$ interaction regressors. Including the intercept, there are $1 + (r - 1) + (c - 1) + (r - 1)(c - 1) = r \times c$ regressors and corresponding parameters, and we can therefore capture any pattern of the $r \times c$ cell means.

For example, when $c = 3$ and $r = 2$, applying our usual practice of treating the last level of each factor as the baseline level, we have the following coding of regressors, corresponding to the six cells formed from the levels of R and C :

R	C	R_1	C_1	C_2	$R_1 \times C_1$	$R_1 \times C_2$
1	1	1	1	0	1	0
1	2	1	0	1	0	1
1	3	1	0	0	0	0
2	1	0	1	0	0	0
2	2	0	0	1	0	0
2	3	0	0	0	0	0

The dummy-coded two-way ANOVA model is then

$$Y_i = \alpha + \beta_1 R_{i1} + \gamma_1 C_{i1} + \gamma_2 C_{i2} + \delta_{11} R_{i1} C_{i1} + \delta_{12} R_{i1} C_{i2} + \varepsilon_i \quad (8.5)$$

Each of the six cell means μ_{jk} can be written in terms of the parameters of the model; for example, for $j = 1$ and $k = 1$,

$$\begin{aligned} \mu_{11} &= E(Y_i | R = 1, C = 1) \\ &= \alpha + \beta_1 \times 1 + \gamma_1 \times 1 + \gamma_2 \times 0 + \delta_{11} \times 1 \times 1 + \delta_{12} \times 1 \times 0 \\ &= \alpha + \beta_1 + \gamma_1 + \delta_{11} \end{aligned}$$

and for $j = 2$ and $k = 3$,

$$\begin{aligned} \mu_{23} &= E(Y_i | R = 2, C = 3) \\ &= \alpha + \beta_1 \times 0 + \gamma_1 \times 0 + \gamma_2 \times 0 + \delta_{11} \times 0 \times 0 + \delta_{12} \times 0 \times 0 \\ &= \alpha \end{aligned}$$

We can therefore solve for the six parameters in terms of the cell means:

¹⁵See Exercise 8.12.

$$\begin{aligned}
 \alpha &= \mu_{23} \\
 \beta_1 &= \mu_{13} - \mu_{23} \\
 \gamma_1 &= \mu_{21} - \mu_{23} \\
 \gamma_2 &= \mu_{22} - \mu_{23} \\
 \delta_{11} &= \mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} \\
 \delta_{12} &= \mu_{12} - \mu_{13} - \mu_{22} + \mu_{23}
 \end{aligned} \tag{8.6}$$

Although the parameters are linear functions of the cell means, they do not have entirely straightforward interpretations, particularly for the main effects of R and C .¹⁶ Nevertheless, if we construct incremental F -tests in conformity with the principle of marginality (i.e., “type II tests”), we can test for interaction between the row and column factors and, if interactions are absent, for row and column main effects.

Applying this approach to Moore and Krupat’s data, where the response variable is conformity, and where R is partner’s status, with two levels, and C is authoritarianism, with three levels, I fit the following models to the data:

Model	Terms	Regression Sum of Squares
1	$R, C, R \times C$	391.436
2	R, C	215.947
3	R	204.332
4	C	3.733

producing the two-way ANOVA table,

Source	Models Contrasted	SS	df	MS	F	p
Partner’s status	2 – 4	212.214	1	212.214	10.12	.003
Authoritarianism	2 – 3	11.615	2	5.807	0.28	.76
Status × Authoritarianism	1 – 2	175.489	2	87.745	4.18	.02
Residuals	from Model 1	817.764	39	20.968		
Total		1209.200	44			

Thus, the interaction between partner’s status and authoritarianism is statistically significant, and we would not interpret the tests of the main effect, which are marginal to the interaction and assume that interaction is absent. As usual, I estimated the error variance, $\hat{\sigma}_e^2 = 20.968$, from the largest model fit to the data, Model 1.

This approach—coding dummy regressors for the main effects of factors and taking products to form interaction regressors—can be extended to three- and higher-way ANOVA models. In three-way ANOVA, for example, with factors A , B , and C , we would form dummy regressors for each factor and then compute all two-way and three-way products of regressors from different sets—that is, for $A \times B$, $A \times C$, $B \times C$, and $A \times B \times C$. As long as we compute incremental F -tests that conform to the principle of marginality (e.g., A after B , C , and BC ; AB after A , B , C ,

¹⁶But see Exercise 8.13.

AC , and BC ; and ABC after A , B , C , AB , AC , and BC), we will test sensible hypotheses about the main effects, the two-way interactions, and the three-way interaction among the three factors.

As long as we construct tests that conform to the principle of marginality, we can code main effects in two- and higher-way ANOVA using dummy regressors, forming interaction regressors as all products of main-effect regressors for the main effects marginal to each interaction.

8.2.3 The Two-Way ANOVA Model

Because interpretation of results in two-way ANOVA depends crucially on the presence or absence of interaction, our first concern is to test the null hypothesis of no interaction. Based on the discussion in Section 8.2.1, this hypothesis can be expressed in terms of the cell means:

$$H_0: \mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'} \quad \text{for all } j, j' \text{ and } k, k' \quad (8.7)$$

In words: The row effects are the same within all levels of the column factor. By rearranging the terms in Equation 8.7, we can write the null hypothesis in the following alternative but equivalent manner:

$$H_0: \mu_{jk} - \mu_{jk'} = \mu_{j'k} - \mu_{j'k'} \quad \text{for all } j, j' \text{ and } k, k' \quad (8.8)$$

That is, the column effects are invariant across rows. Once more, we see the symmetry of the concept of interaction.

It is convenient, following the presentation in the previous section, to express hypotheses concerning main effects in terms of the marginal means. Thus, for the row classification, we have the null hypothesis

$$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{r\cdot} \quad (8.9)$$

and for the column classification

$$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot c} \quad (8.10)$$

Formulated in this manner, the main-effect null hypotheses (Equations 8.9 and 8.10) are testable whether interactions are present or absent, but these hypotheses are generally of interest only when the interactions are nil.

The two-way ANOVA model, suitably defined, provides a convenient means for testing the hypotheses concerning interactions and main effects (in Equations 8.7, 8.9, and 8.10). The model is

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk} \quad (8.11)$$

where Y_{ijk} is the i th observation in row j , column k of the RC table; μ is the general mean of Y ; α_j and β_k are main-effect parameters, for row effects and column effects, respectively; γ_{jk} are interaction parameters; and ε_{ijk} are errors satisfying the usual linear-model assumptions. Taking expectations, Equation 8.11 becomes

$$\mu_{jk} \equiv E(Y_{ijk}) = \mu + \alpha_j + \beta_k + \gamma_{jk} \quad (8.12)$$

Because there are $r \times c$ population cell means and $1 + r + c + (r \times c)$ parameters in Equation 8.12, the parameters of the model are not uniquely determined by the cell means. By reasoning that is familiar from Section 8.1.2 on the one-way ANOVA model, the indeterminacy of Equation 8.12 can be overcome by imposing $1 + r + c$ independent “identifying” restrictions on its parameters. Although—from one point of view—any restrictions will do, it is convenient to select restrictions that make it simple to test the hypotheses of interest.

With this purpose in mind, we specify the following sigma constraints on the model parameters:

$$\begin{aligned} \sum_{j=1}^r \alpha_j &= 0 \\ \sum_{k=1}^c \beta_k &= 0 \\ \sum_{j=1}^r \gamma_{jk} &= 0 \quad \text{for all } k = 1, \dots, c \\ \sum_{k=1}^c \gamma_{jk} &= 0 \quad \text{for all } j = 1, \dots, r \end{aligned} \quad (8.13)$$

At first glance, it seems as if we have specified too many constraints, for Equations 8.13 define $1 + 1 + c + r$ restrictions. One of the restrictions on the interactions is redundant, however.¹⁷ In shorthand form, the sigma constraints specify that each set of parameters sums to 0 over each of its coordinates.

The constraints produce the following solution for model parameters in terms of population cell and marginal means:

$$\begin{aligned} \mu &= \mu.. \\ \alpha_j &= \mu_{j.} - \mu.. \\ \beta_k &= \mu_{.k} - \mu.. \\ \gamma_{jk} &= \mu_{jk} - \mu - \alpha_j - \beta_k \\ &= \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu.. \end{aligned} \quad (8.14)$$

The hypothesis of no row main effects (Equation 8.9) is, therefore, equivalent to H_0 : all $\alpha_j = 0$, for under this hypothesis

$$\mu_{1.} = \mu_{2.} = \dots = \mu_{r.} = \mu..$$

Likewise, the hypothesis of no column main effects (Equation 8.10) is equivalent to H_0 : all $\beta_k = 0$, because then

$$\mu_{.1} = \mu_{.2} = \dots = \mu_{.c} = \mu..$$

Finally, it is not difficult to show that the hypothesis of no interactions (given in Equation 8.7 or 8.8) is equivalent to H_0 : all $\gamma_{jk} = 0$.¹⁸

¹⁷See Exercise 8.2.

¹⁸See Exercise 8.3.

8.2.4 Fitting the Two-Way ANOVA Model to Data

Because the least-squares estimator of μ_{jk} is the sample cell mean

$$\bar{Y}_{jk} = \frac{\sum_{i=1}^{n_{jk}} Y_{ijk}}{n_{jk}}$$

least-squares estimators of the constrained model parameters follow immediately from Equations 8.14:

$$\begin{aligned} M &\equiv \hat{\mu} = \bar{Y}_{..} = \frac{\sum \sum \bar{Y}_{jk}}{r \times c} \\ A_j &\equiv \hat{\alpha}_j = \bar{Y}_j - \bar{Y}_{..} = \frac{\sum_k \bar{Y}_{jk}}{c} - \bar{Y}_{..} \\ B_k &\equiv \hat{\beta}_k = \bar{Y}_{.k} - \bar{Y}_{..} = \frac{\sum_j \bar{Y}_{jk}}{r} - \bar{Y}_{..} \\ C_{jk} &\equiv \hat{\gamma}_{jk} = \bar{Y}_{jk} - \bar{Y}_j - \bar{Y}_{.k} + \bar{Y}_{..} \end{aligned}$$

The residuals are just the deviations of the observations from their cell means because the fitted values are the cell means:

$$\begin{aligned} E_{ijk} &= Y_{ijk} - (M + A_j + B_k + C_{jk}) \\ &= Y_{ijk} - \bar{Y}_{jk} \end{aligned}$$

In testing hypotheses about sets of model parameters, however, we require incremental sums of squares for each set, and there is no general way of calculating these sums of squares directly.¹⁹ As in one-way ANOVA, the restrictions on the two-way ANOVA model can be used to produce deviation-coded regressors. Incremental sums of squares can then be calculated in the usual manner. To illustrate this procedure, we will first examine a two-row \times three-column classification. The extension to the general $r \times c$ classification is straightforward and is described subsequently.

In light of the restriction $\alpha_1 + \alpha_2 = 0$ on the row effects of the 2×3 classification, α_2 can be deleted from the model, substituting $-\alpha_1$. Similarly, because $\beta_1 + \beta_2 + \beta_3 = 0$, the column main effect β_3 can be replaced by $-\beta_1 - \beta_2$. More generally, $-\sum_{j=1}^{r-1} \alpha_j$ replaces α_r , and $-\sum_{k=1}^{c-1} \beta_k$ replaces β_c . Because there are, then, $r - 1$ independent α_j parameters and $c - 1$ independent β_k parameters, the degrees of freedom for row and column main effects are, respectively, $r - 1$ and $c - 1$.

The interactions in the 2×3 classification satisfy the following constraints:²⁰

$$\begin{aligned} \gamma_{11} + \gamma_{12} + \gamma_{13} &= 0 \\ \gamma_{21} + \gamma_{22} + \gamma_{23} &= 0 \\ \gamma_{11} + \gamma_{21} &= 0 \\ \gamma_{12} + \gamma_{22} &= 0 \\ \gamma_{13} + \gamma_{23} &= 0 \end{aligned}$$

¹⁹An exception occurs when all the cell frequencies are equal—see Section 8.2.6.

²⁰Recall that although there are five such constraints, the fifth follows from the first four—you may want to show this—and there are therefore only four *independent* constraints on the interaction parameters.

We can, as a consequence, delete all the interaction parameters except γ_{11} and γ_{12} , substituting for the remaining four parameters in the following manner:

$$\begin{aligned}\gamma_{13} &= -\gamma_{11} - \gamma_{12} \\ \gamma_{21} &= -\gamma_{11} \\ \gamma_{22} &= -\gamma_{12} \\ \gamma_{23} &= -\gamma_{13} = \gamma_{11} + \gamma_{12}\end{aligned}$$

More generally, we can write all $r \times c$ interaction parameters in terms of $(r-1)(c-1)$ of the γ_{jk} s, and there are, therefore, $(r-1)(c-1)$ degrees of freedom for interaction.

These observations lead to the following coding of regressors for the 2×3 classification:

Cell		(α_1)	(β_1)	(β_2)	(γ_{11})	(γ_{12})
Row	Column	R_1	C_1	C_2	$R_1 C_1$	$R_1 C_2$
1	1	1	1	0	1	0
1	2	1	0	1	0	1
1	3	1	-1	-1	-1	-1
2	1	-1	1	0	-1	0
2	2	-1	0	1	0	-1
2	3	-1	-1	-1	1	1

That is, for example, according to the third row of this table,

$$\begin{aligned}\mu_{13} &= \mu + \alpha_1 - \beta_1 - \beta_2 - \gamma_{11} - \gamma_{12} \\ &= \mu + \alpha_1 + \beta_3 + \gamma_{13}\end{aligned}$$

as required.

I have constructed these regressors to reflect the constraints on the model, but they can also be coded mechanically by applying these rules:

1. There are $r-1$ regressors for the row main effects; the j th such regressor, R_j , is coded according to the deviation-coding scheme:

$$R_{ij} = \begin{cases} 1 & \text{if observation } i \text{ is in row } j, \\ -1 & \text{if observation } i \text{ is in row } r \text{ (the last row),} \\ 0 & \text{if observation } i \text{ is in any other row.} \end{cases}$$

2. There are $c-1$ regressors for the column main effects; the k th such regressor, C_k , is coded according to the deviation-coding scheme:

$$C_{ik} = \begin{cases} 1 & \text{if observation } i \text{ is in column } k, \\ -1 & \text{if observation } i \text{ is in column } c \text{ (the last column),} \\ 0 & \text{if observation } i \text{ is in any other column.} \end{cases}$$

3. There are $(r-1)(c-1)$ regressors for the RC interactions. These interaction regressors consist of all pairwise products of the $r-1$ main-effect regressors for rows and $c-1$ main-effect regressors for columns.

The two-way ANOVA model $Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$ incorporates the main effects and interactions of two factors. This model is overparametrized, but it may be fit to data by placing suitable restrictions on its parameters. A convenient set of restrictions is provided by sigma constraints, specifying that each set of parameters (α_j , β_k , and γ_{jk}) sums to 0 over each of its coordinates. As in one-way ANOVA, sigma constraints lead to deviation-coded regressors.

8.2.5 Testing Hypotheses in Two-Way ANOVA

I have specified constraints on the two-way ANOVA model so that testing hypotheses about the parameters of the constrained model is equivalent to testing hypotheses about the interactions and main effects of the two factors. Tests for interactions and main effects can be constructed by the incremental sum-of-squares approach.

For ease of reference, I will write $SS(\alpha, \beta, \gamma)$ to denote the regression sum of squares for the full model, which includes both sets of main effects and the interactions. The regression sums of squares for other models are similarly represented. For example, for the no-interaction model, we have $SS(\alpha, \beta)$, and for the model that omits the column main-effect regressors, we have $SS(\alpha, \gamma)$. This last model violates the principle of marginality because it includes the interaction regressors but omits the column main effects. Nevertheless, as I will explain presently, the model plays a role in constructing the incremental sum of squares for testing the column main effects.

As usual, incremental sums of squares are given by *differences* between the regression sums of squares for alternative models, one of which is “nested” within (i.e., is a special case of) the other. I will use the following notation for incremental sums of squares in ANOVA:²¹

$$\begin{aligned} SS(\gamma|\alpha, \beta) &= SS(\alpha, \beta, \gamma) - SS(\alpha, \beta) \\ SS(\alpha|\beta, \gamma) &= SS(\alpha, \beta, \gamma) - SS(\beta, \gamma) \\ SS(\beta|\alpha, \gamma) &= SS(\alpha, \beta, \gamma) - SS(\alpha, \gamma) \\ SS(\alpha|\beta) &= SS(\alpha, \beta) - SS(\beta) \\ SS(\beta|\alpha) &= SS(\alpha, \beta) - SS(\alpha) \end{aligned}$$

We read $SS(\gamma|\alpha, \beta)$, for example, as “the sum of squares for interaction *after* the main effects” and $SS(\alpha|\beta)$ as “the sum of squares for the row main effects *after* the column main effects and *ignoring* the interactions.” The residual sum of squares is

²¹You may encounter variations of the SS notation. One common approach (used, e.g., in Searle, 1971) is to include the grand mean μ in the arguments to the sum-of-squares function and to let $R(\cdot)$ denote the “raw” (rather than mean deviation) sum of squares. Thus, in this scheme, $R(\mu, \alpha, \beta) = \sum \sum \sum \hat{Y}_{ijk}^2$ is the raw sum of squares for the no-interaction model, while

$$\begin{aligned} R(\alpha, \beta|\mu) &= R(\mu, \alpha, \beta) - R(\mu) \\ &= \sum \sum \sum (\hat{Y}_{ijk} - \bar{Y})^2 \\ &= SS(\alpha, \beta) \end{aligned}$$

is the mean deviation explained sum of squares for the same model. (The $\hat{Y}_{ijk} = \hat{Y}_{jk}$ are the least-squares fitted values from the no-interaction model.)

Table 8.3 Two-Way Analysis of Variance, Showing Alternative Tests for Row and Column Main Effects

Source	df	Sum of Squares	H_0
R	$r - 1$	$SS(\alpha \beta, \gamma)$ $SS(\alpha \beta)$	all $\alpha_j = 0$ ($\mu_{j\cdot} = \mu_{j'\cdot}$) all $\alpha_j = 0 \text{all } \gamma_{jk} = 0$ ($\mu_{j\cdot} = \mu_{j'\cdot} \text{no interaction}$)
C	$c - 1$	$SS(\beta \alpha, \gamma)$ $SS(\beta \alpha)$	all $\beta_k = 0$ ($\mu_{\cdot k} = \mu_{\cdot k'} \text{no interaction}$) all $\beta_k = 0 \text{all } \gamma_{jk} = 0$ ($\mu_{\cdot k} = \mu_{\cdot k'} \text{no interaction}$)
RC	$(r - 1)(c - 1)$	$SS(\gamma \alpha, \beta)$	all $\gamma_{jk} = 0$ ($\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'} = 0$)
Residuals	$n - rc$	TSS - SS(α, β, γ)	
Total	$n - 1$	TSS	

NOTE: Each incremental F -test is formulated by dividing an effect mean square by the residual mean square (where each mean square is the corresponding sum of squares divided by its degrees of freedom). The hypothesis tested by each such F -test is expressed both in terms of constrained model parameters and (in parentheses) in terms of cell or marginal means.

$$\begin{aligned} RSS &= \sum \sum \sum E_i^2 \\ &= \sum \sum \sum (Y_{ijk} - \bar{Y}_{jk})^2 \\ &= TSS - SS(\alpha, \beta, \gamma) \end{aligned}$$

The incremental sum of squares for interaction, $SS(\gamma|\alpha, \beta)$, is appropriate for testing the null hypothesis of no interaction, $H_0: \text{all } \gamma_{jk} = 0$. In the presence of interactions, we can use $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$ to test hypotheses concerning main effects (i.e., differences among row and column marginal means), but—as I have explained—these hypotheses are usually not of interest when the interactions are important.

In the *absence* of interactions, $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$ can be used to test for main effects, but the use of $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$ is also appropriate. If, however, interactions are *present*, then F -tests based on $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$ do not test the main-effect null hypotheses $H_0: \text{all } \alpha_j = 0$ and $H_0: \text{all } \beta_k = 0$; instead, the interaction parameters become implicated in these tests. These remarks are summarized in Table 8.3.

Certain authors (e.g., Nelder, 1976, 1977) prefer main-effects tests based on $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$ because, if interactions are absent, tests based on these sums of squares follow from the principle of marginality and are more powerful than those based on $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$ —indeed, these tests are *maximally* powerful for the main effects if the interactions are absent. Other authors (e.g., Hocking & Speed, 1975) prefer $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$ because, in the presence of interactions, tests based on these sums of squares have a straightforward (if usually uninteresting) interpretation. I believe that either approach is reasonable but

have a preference for tests that conform to the principle of marginality—here, those based on $\text{SS}(\alpha|\beta)$ and $\text{SS}(\beta|\alpha)$.²²

It is important to understand, however, that while $\text{SS}(\alpha)$ and $\text{SS}(\beta)$ are useful as building blocks of $\text{SS}(\alpha|\beta)$ and $\text{SS}(\beta|\alpha)$, it is, in general, *inappropriate* to use $\text{SS}(\alpha)$ and $\text{SS}(\beta)$ to test hypotheses about the R and C main effects: Each of these sums of squares depends on the other set of main effects (and the interactions, if they are present). A main effect is a *partial* effect, so we need to control for rows in assessing the column main effects and vice versa.

Testing hypotheses about the sigma-constrained parameters is equivalent to testing interaction-effect and main-effect hypotheses about cell and marginal means. There are two reasonable procedures for testing main-effect hypotheses in two-way ANOVA: Tests based on $\text{SS}(\alpha|\beta, \gamma)$ and $\text{SS}(\beta|\alpha, \gamma)$ (“type III” tests) employ models that violate the principle of marginality, but the tests are valid whether or not interactions are present. Tests based on $\text{SS}(\alpha|\beta)$ and $\text{SS}(\beta|\alpha)$ (“type II” tests) conform to the principle of marginality but are valid only if interactions are absent, in which case they are maximally powerful.

For the Moore and Krupat conformity data, factor R is partner’s status and factor C is authoritarianism. Sums of squares for various models fit to the data are as follows:

$$\begin{aligned}\text{SS}(\alpha, \beta, \gamma) &= 391.44 \\ \text{SS}(\alpha, \beta) &= 215.95 \\ \text{SS}(\alpha, \gamma) &= 355.42 \\ \text{SS}(\beta, \gamma) &= 151.87 \\ \text{SS}(\alpha) &= 204.33 \\ \text{SS}(\beta) &= 3.73 \\ \text{TSS} &= 1209.20\end{aligned}$$

The ANOVA for the experiment is shown in Table 8.4. The predicted status \times authoritarianism interaction proves to be statistically significant. A researcher would not normally report both sets of main-effect sums of squares; in this instance, where the interactions probably are not negligible, $\text{SS}(\alpha|\beta)$ and $\text{SS}(\beta|\alpha)$ do not test hypotheses about main effects, as I have explained.

8.2.6 Equal Cell Frequencies

Equal cell frequencies simplify—but do not change fundamentally—the procedures of the preceding section. When all the cell frequencies are the same, the deviation regressors for

²²In the SAS statistical computer package, $\text{SS}(\alpha|\beta)$ and $\text{SS}(\beta|\alpha)$ are called “Type II” sums of squares, while $\text{SS}(\alpha|\beta, \gamma)$ and $\text{SS}(\beta|\alpha, \gamma)$ are called “Type III” sums of squares. This terminology has become widespread.

The *sequential* sums of squares $\text{SS}(\alpha)$, $\text{SS}(\beta|\alpha)$, and $\text{SS}(\gamma|\alpha, \beta)$ are similarly termed “Type I” sums of squares. Some researchers are attracted to the sequential sums of squares because they add to the regression sum of squares for the full model, $\text{SS}(\alpha, \beta, \gamma)$. This attraction is misguided, however, because $\text{SS}(\alpha)$ does not test for row main effects. We should focus on the hypotheses to be tested, not on a superficial property of the sums of squares, such as the fact that they add up in a simple manner.

Table 8.4 Analysis-of-Variance Table for Moore and Krupat's Conformity Experiment

Source	SS	df	MS	F	p
Partner's status		1			
$\alpha \beta, \gamma$	239.57		239.57	11.43	.002
$\alpha \beta$	212.22		212.22	10.12	.003
Authoritarianism		2			
$\beta \alpha, \gamma$	36.02		18.01	0.86	.43
$\beta \alpha$	11.62		5.81	0.28	.76
Partner's status \times Authoritarianism	175.49	2	87.74	4.18	.02
Residuals	817.76	39	20.97		
Total	1209.2	44			

NOTE: Alternative tests are shown for the partner's status and authoritarianism main effects.

different sets of effects are uncorrelated. Equal-cell-frequencies data are often termed *balanced* or *orthogonal*.²³

Uncorrelated main-effect and interaction regressors permit a unique decomposition of the regression sum of squares for the model, $SS(\alpha, \beta, \gamma)$, into components due to the three sets of effects. Indeed, for balanced data,

$$\begin{aligned} SS(\alpha|\beta, \gamma) &= SS(\alpha|\beta) = SS(\alpha) \\ SS(\beta|\alpha, \gamma) &= SS(\beta|\alpha) = SS(\beta) \\ SS(\gamma|\alpha, \beta) &= SS(\gamma) \end{aligned}$$

and hence

$$SS(\alpha, \beta, \gamma) = SS(\alpha) + SS(\beta) + SS(\gamma)$$

These results lead to simple direct formulas for the several sums of squares:

$$\begin{aligned} SS(\alpha) &= n'c \sum_{j=1}^r (\bar{Y}_{j\cdot} - \bar{Y}_{..})^2 \\ SS(\beta) &= n'r \sum_{k=1}^c (\bar{Y}_{\cdot k} - \bar{Y}_{..})^2 \\ SS(\gamma) &= n' \sum_{j=1}^r \sum_{k=1}^c (\bar{Y}_{jk} - \bar{Y}_{j\cdot} - \bar{Y}_{\cdot k} + \bar{Y}_{..})^2 \end{aligned}$$

where $n' = n/rc$ is the number of observations in each cell of the *RC* table.

8.2.7 Some Cautionary Remarks

R. A. Fisher (1925) originally formulated ANOVA for balanced data. Yet, as early as 1934, Fisher's colleague at the Rothamsted Experimental Station in England, Frank Yates, extended

²³See Chapter 10, on the geometry of linear models, for an explanation of the term *orthogonal*.

ANOVA to unbalanced data. Apart from approximate methods motivated by the desire to reduce the effort of calculation, Yates (1934) suggested two approaches to the two-way classification, naming both for the computational techniques that he developed. The first approach, which he called “the method of weighted squares of means,” calculates (using my notation) the main-effect sums of squares $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$, and the interaction sum of squares $SS(\gamma|\alpha, \beta)$. Yates’s second approach, which he called “the method of fitting constants,” assumes that interactions are absent and calculates $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$.

Considering the apparent simplicity of the two-way classification and the lucidity of Yates’s treatment of it, it is ironic that the analysis of unbalanced data has become the subject of controversy and confusion. While it is not my purpose to present a complete account of the “debate” concerning the proper handling of unbalanced data—and while it is tempting to ignore this debate altogether—there are two reasons for addressing the topic briefly here: (1) You may encounter confused applications of ANOVA or may have occasion to consult other accounts of the method, and (2) computer programs for ANOVA are occasionally misleading or vague in their documentation and output or even incorrect in their calculations (see Francis, 1973).²⁴

Much of the confusion about the analysis of unbalanced data has its source in the restrictions—or other techniques—that are used to solve the “overparametrized” (i.e., unrestricted) two-way ANOVA model. Imagine, for example, that we use dummy (0, 1) coding rather than deviation (-1, 0, 1) coding to fit the model to the data, as in Section 8.2.3.

Let $SS^*(\cdot)$ denote the regression sum of squares for a dummy-coded model. For the full model and the main-effects model, we obtain the same sums of squares as before; that is,

$$\begin{aligned} SS(\alpha, \beta, \gamma) &= SS^*(\alpha, \beta, \gamma) \\ SS(\alpha, \beta) &= SS^*(\alpha, \beta) \end{aligned}$$

Likewise (because they are just the two one-way ANOVAs)

$$\begin{aligned} SS(\alpha) &= SS^*(\alpha) \\ SS(\beta) &= SS^*(\beta) \end{aligned}$$

And because these regression sums of squares are the same, so are the incremental sums of squares that depend on them:

$$\begin{aligned} SS(\gamma|\alpha, \beta) &= SS^*(\gamma|\alpha, \beta) \\ SS(\alpha|\beta) &= SS^*(\alpha|\beta) \\ SS(\beta|\alpha) &= SS^*(\beta|\alpha) \end{aligned}$$

In general, however,

$$\begin{aligned} SS(\alpha, \gamma) &\neq SS^*(\alpha, \gamma) \\ SS(\beta, \gamma) &\neq SS^*(\beta, \gamma) \end{aligned}$$

and, consequently (also in general),

²⁴With respect to the second point, it is good practice to test a computer program with known data before trusting it to analyze new data. This advice applies not just to ANOVA calculations but generally.

$$\text{SS}(\alpha|\beta, \gamma) \neq \text{SS}^*(\alpha|\beta, \gamma)$$

$$\text{SS}(\beta|\alpha, \gamma) \neq \text{SS}^*(\beta|\alpha, \gamma)$$

The general lesson to be drawn from these results is that tests that conform to the principle of marginality [here, those based on $\text{SS}(\gamma|\alpha, \beta)$, $\text{SS}(\alpha|\beta)$, and $\text{SS}(\beta|\alpha)$] *do not* depend on the specific restrictions that were employed to identify the model (i.e., remove the indeterminacy in the overparametrized model), while tests that “violate” the principle of marginality [those based on $\text{SS}(\alpha|\beta, \gamma)$ and $\text{SS}(\beta|\alpha, \gamma)$] *do* depend on the specific restrictions.

I showed that $\text{SS}(\alpha|\beta, \gamma)$ and $\text{SS}(\beta|\alpha, \gamma)$, based on the sigma constraints, are appropriate for testing hypotheses about main effects in the potential presence of interactions. It follows that $\text{SS}^*(\alpha|\beta, \gamma)$ and $\text{SS}^*(\beta|\alpha, \gamma)$ *do not* properly test these hypotheses. It is important, in this context, to select constraints that test reasonable hypotheses about cell and marginal means. The SS notation is frequently used carelessly, without attention to the constraints that are employed and to the hypotheses that follow from them.²⁵

8.3 Higher-Way Analysis of Variance

The methods of the previous section can be extended to any number of factors. I will consider the three-way classification in some detail before commenting briefly on the general case.

8.3.1 The Three-Way Classification

It is convenient to label the factors in the three-way classification as A , B , and C , with a , b , and c levels, consecutively. A response-variable observation is represented by Y_{ijkm} , where the first subscript gives the index of the observation within its cell. The number of observations sampled in cell $\{j, k, m\}$ is n_{jkm} , and μ_{jkm} is the population mean in this cell. Quantities such as $\mu_{...}$, $\mu_{j..}$, and $\mu_{jk..}$ denote marginal means formed by averaging over the dotted subscripts.

The three-way ANOVA model is

$$\begin{aligned} Y_{ijkm} &= \mu_{jkm} + \varepsilon_{ijkm} \\ &= \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)} + \alpha_{AB(jk)} \\ &\quad + \alpha_{AC(jm)} + \alpha_{BC(km)} + \alpha_{ABC(jkm)} + \varepsilon_{ijkm} \end{aligned} \tag{8.15}$$

To avoid the proliferation of symbols, I have introduced a new and easily extended notation for model parameters: The first set of subscripts (e.g., AB) indicates the factors to which a parameter pertains, while the parenthetical subscripts [e.g., (j, k)] index factor categories.

We make the usual linear-model assumptions about the errors ε_{ijkm} and constrain all sets of parameters to sum to 0 over every coordinate; for example,

²⁵Further discussions on the points raised in this section may be found in a variety of sources, including Hocking and Speed (1975); Speed and Hocking (1976); Speed, Hocking, and Hackney (1978); Speed and Monlezun (1979); Searle, Speed, and Henderson (1981); and Steinhorst (1982). Also see Section 9.1.1 and Exercise 9.15.

$$\begin{aligned}\sum_{j=1}^a \alpha_{A(j)} &= 0 \\ \sum_{j=1}^a \alpha_{AB(jk)} &= \sum_{k=1}^b \alpha_{AB(jk)} = 0 \quad \text{for all } j, k \\ \sum_{j=1}^a \alpha_{ABC(jkm)} &= \sum_{k=1}^b \alpha_{ABC(jkm)} = \sum_{m=1}^c \alpha_{ABC(jkm)} = 0 \quad \text{for all } j, k, m\end{aligned}$$

The sigma constraints for $\alpha_{B(k)}$, $\alpha_{C(m)}$, $\alpha_{AC(jm)}$, and $\alpha_{BC(km)}$ follow similar patterns.

The three-way ANOVA model includes parameters for main effects ($\alpha_{A(j)}$, $\alpha_{B(k)}$, and $\alpha_{C(m)}$), for *two-way interactions* between each pair of factors ($\alpha_{AB(jk)}$, $\alpha_{AC(jm)}$, and $\alpha_{BC(km)}$), and for *three-way interactions* among all three factors ($\alpha_{ABC(jkm)}$). The two-way interactions have the same interpretation as in two-way ANOVA: If, for instance, A and B interact, then the effect of either factor on the response variable varies across the levels of the other factor. Similarly, if the ABC interaction is nonzero, then the joint effect of any pair of factors (say, A and B) varies across the categories of the remaining factor (C).

In formulating models and interpreting effects in three-way ANOVA, we may again appeal to the principle of marginality. Thus, main effects (e.g., of A) are generally not interpreted if they are marginal to nonnull interactions (AB , AC , or ABC). Likewise, a *lower-order* interaction (such as AB) is usually not interpreted if it has a nonnull *higher-order relative* (ABC): If the joint effects of A and B are different in different categories of C , then it is not generally sensible to speak of the *unconditional* AB effects, without reference to a specific category of C .

Deviation regressors for main effects in the three-way classification can be coded as before; regressors for interactions are formed by taking all possible products of the main effects that “compose” the interaction. Here, for example, is the coding for $a = 2$, $b = 2$, and $c = 3$:

<i>Cell</i> <i>jk</i> <i>m</i>	<i>A</i>	<i>B</i>	C_1	C_2	AB	AC_1	AC_2	BC_1	BC_2	ABC_1	ABC_2
111	1	1	1	0	1	1	0	1	0	1	0
112	1	1	0	1	1	0	1	0	1	0	1
113	1	1	-1	-1	1	-1	-1	-1	-1	-1	-1
121	1	-1	1	0	-1	1	0	-1	0	-1	0
122	1	-1	0	1	-1	0	1	0	-1	0	-1
123	1	-1	-1	-1	-1	-1	-1	1	1	1	1
211	-1	1	1	0	-1	-1	0	1	0	-1	0
212	-1	1	0	1	-1	0	-1	0	1	0	-1
213	-1	1	-1	-1	-1	1	1	-1	-1	1	1
221	-1	-1	1	0	1	-1	0	-1	0	1	0
222	-1	-1	0	1	1	0	-1	0	-1	0	1
223	-1	-1	-1	-1	1	1	1	1	-1	-1	-1

The following points are noteworthy:

- The 12 cell means are expressed in terms of an equal number of independent parameters (including the general mean, μ), underscoring the point that three-way interactions may

be required to account for the pattern of cell means. More generally in the three-way classification, there are abc cells and the same number of independent parameters:

$$\begin{aligned} 1 + (a - 1) + (b - 1) + (c - 1) + (a - 1)(b - 1) + (a - 1)(c - 1) + (b - 1)(c - 1) \\ + (a - 1)(b - 1)(c - 1) \\ = abc. \end{aligned}$$

- The degrees of freedom for a set of effects correspond, as usual, to the number of independent parameters in the set. There are, for example, $a - 1$ degrees of freedom for the A main effects, $(a - 1)(b - 1)$ degrees of freedom for the AB interactions, and $(a - 1)(b - 1)(c - 1)$ degrees of freedom for the ABC interactions.

Solving for the constrained parameters in terms of populations means produces the following results:

$$\begin{aligned} \mu &= \mu \dots \\ \alpha_{A(j)} &= \mu_{j..} - \mu \dots \\ \alpha_{AB(jk)} &= \mu_{jk.} - \mu - \alpha_{A(j)} - \alpha_{B(k)} \\ &= \mu_{jk.} - \mu_{j..} - \mu_{.k.} + \mu \dots \\ \alpha_{ABC(jkm)} &= \mu_{jkm} - \mu - \alpha_{A(j)} - \alpha_{B(k)} - \alpha_{C(m)} - \alpha_{AB(jk)} - \alpha_{AC(jm)} - \alpha_{BC(km)} \\ &= \mu_{jkm} - \mu_{jk.} - \mu_{j..m} - \mu_{.km} + \mu_{j..} + \mu_{.k.} + \mu_{..m} - \mu \dots \end{aligned}$$

(The patterns for $\alpha_{B(k)}$, $\alpha_{C(m)}$, $\alpha_{AC(jm)}$, and $\alpha_{BC(km)}$ are similar and are omitted for brevity.) As in two-way ANOVA, therefore, the null hypothesis

$$H_0: \text{all } \alpha_{A(j)} = 0$$

is equivalent to

$$H_0: \mu_{1..} = \mu_{2..} = \dots = \mu_{a..}$$

and the hypothesis

$$H_0: \text{all } \alpha_{AB(jk)} = 0$$

is equivalent to

$$H_0: \mu_{jk.} - \mu_{j'k.} = \mu_{jk'} - \mu_{j'k'} \text{ for all } j, j' \text{ and } k, k'$$

Likewise, some algebraic manipulation²⁶ shows that the null hypothesis

$$H_0: \text{all } \alpha_{ABC(jkm)} = 0$$

is equivalent to

$$\begin{aligned} H_0: (\mu_{jkm} - \mu_{j'km}) - (\mu_{jk'm} - \mu_{j'k'm}) \\ = (\mu_{jkm'} - \mu_{j'km'}) - (\mu_{jk'm'} - \mu_{j'k'm'}) \quad (8.16) \\ \text{for all } j, j'; k, k'; \text{ and } m, m' \end{aligned}$$

The second-order differences in Equation (8.16) are equal when the pattern of AB interactions is invariant across categories of factor C —an intuitively reasonable extension of the

²⁶See Exercise 8.4.

Table 8.5 General Three-Way ANOVA Table, Showing Incremental Sums of Squares for Terms Involving Factor A

Source	df	<i>Sum of Squares</i>	H_0
A	$a-1$	$SS(A B, C, AB, AC, BC, ABC)$ $SS(A B, C, BC)$	$\alpha_A = 0$ $\alpha_A = 0 \alpha_{AB} = \alpha_{AC} = \alpha_{ABC} = 0$
AB	$(a-1)(b-1)$	$SS(AB A, B, C, AC, BC, ABC)$ $SS(AB A, B, C, AC, BC)$	$\alpha_{AB} = 0$ $\alpha_{AB} = 0 \alpha_{ABC} = 0$
ABC	$(a-1)(b-1)(c-1)$	$SS(ABC A, B, C, AB, AC, BC)$	$\alpha_{ABC} = 0$
Residuals	$n-abc$	TSS - SS(A, B, C, AB, AC, BC, ABC)	
Total	$n-1$	TSS	

NOTE: Alternative tests are shown for the A main effects and AB interactions.

notion of no interaction to three factors. Rearranging the terms in Equation 8.16 produces similar results for AC and BC , demonstrating that three-way interaction—like two-way interaction—is symmetric in the factors. As in two-way ANOVA, this simple relationship between model parameters and population means depends on the sigma constraints, which were imposed on the overparametrized model in Equation 8.15.

Incremental F -tests can be constructed in the usual manner for the parameters of the three-way ANOVA model. A general ANOVA table, adapting the SS notation of Section 8.2.5 and showing alternative tests for main effects and lower-order interactions, is sketched in Table 8.5. Once more, for compactness, only tests involving factor A are shown. Note that a main-effect hypothesis such as H_0 : all $\alpha_{A(j)} = 0$ is of interest even when the BC interactions are present because A is not marginal to BC .

8.3.2 Higher-Order Classifications

Extension of ANOVA to more than three factors is algebraically and computationally straightforward. The general p -way classification can be described by a model containing terms for every combination of factors; the highest-order term, therefore, is for the p -way interactions. If the p -way interactions are nonzero, then the joint effects of any $p-1$ factors vary across the levels of the remaining factor. In general, we can be guided by the principle of marginality in interpreting effects.

Three-way interactions, however, are reasonably complex, and the even greater complexity of higher-order interactions can make their interpretation difficult. Yet, at times, we may expect to observe a high-order interaction of a particular sort, as when a specific *combination* of characteristics predisposes individuals to act in a certain manner.²⁷ On the other hand, it is common to find that high-order interactions are not statistically significant or that they are negligibly small relative to other effects.

²⁷An alternative to specifying a high-order interaction would be simply to introduce a dummy regressor, coded 1 for the combination of categories in question and 0 elsewhere.

There is, moreover, no rule of data analysis that requires us to fit and test all possible interactions. In working with higher-way classifications, we may limit our consideration to effects that are of theoretical interest, or at least to effects that are substantively interpretable. It is fairly common, for example, for researchers to fit models containing only main effects:

$$Y_{ijk...r} = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \cdots + \alpha_{P(r)} + \varepsilon_{ijk...r}$$

This approach, sometimes called *multiple-classification analysis* or *MCA* (Andrews, Morgan, & Sonquist, 1973),²⁸ is analogous to an additive multiple regression. In a similar spirit, a researcher might entertain models that include only main effects and two-way interactions.

The ANOVA model and procedures for testing hypotheses about main effects and interactions extend straightforwardly to three-way and higher-way classifications. In each case, the highest-order interaction corresponds to the number of factors in the model. It is not necessary, however, to specify a model that includes all terms through the highest-order interaction.

Cell means for an illustrative four-way classification appear in Figure 8.6, which shows mean vocabulary score in the U.S. General Social Surveys as a function of level of education (less than high school, high school, junior college, bachelor's degree, or graduate degree), age group (five bins, from 18–29 to 60 or more), place of birth (foreign born or native born), and sex.²⁹ The 18,665 observations in the data set are therefore divided across $5 \times 5 \times 2 \times 2 = 100$ cells. Most cells have a substantial number of observations, but some—especially among the foreign born—are sparse, and although there are data in every cell, there is, for example, only one foreign-born male, 50 to 59 years of age, with a junior-college education.

The vertical lines in Figure 8.6 represent ± 2 standard errors around the means; in cells with a very small number of observations, some of these intervals extend beyond the range of the vertical axis (and, indeed, in the cell with only one observation, the interval is infinite). Discounting the means that are highly variable, the pattern of change in mean vocabulary score with education appears quite similar across cells, and education seems to have a much stronger impact on vocabulary score than do the other factors.

Although vocabulary score is discrete and a disguised proportion, its distribution is reasonably well behaved, and I therefore proceed with a four-way ANOVA, shown in Table 8.6. The tests in this ANOVA table conform to the principle of marginality. Thus, for example, the main-effect sum of squares for education is computed after age, place of birth, sex, and all two- and three-way interactions among these factors, but ignoring all the interactions of which education is a lower-order relative.

One of the important uses of a statistical model is to “smooth” the data, eliminating features of the data that are unimportant.³⁰ In a large data set, such as this one, even trivial effects can

²⁸The term *multiple-classification analysis* is unfortunate because it is equally descriptive of any ANOVA model fit to the p -way classification.

²⁹The GSS vocabulary data were introduced in Chapter 3.

³⁰Of course, what counts as “unimportant” varies by context, and in some circumstances, even a relatively minor feature of the data may prove to be of interest.

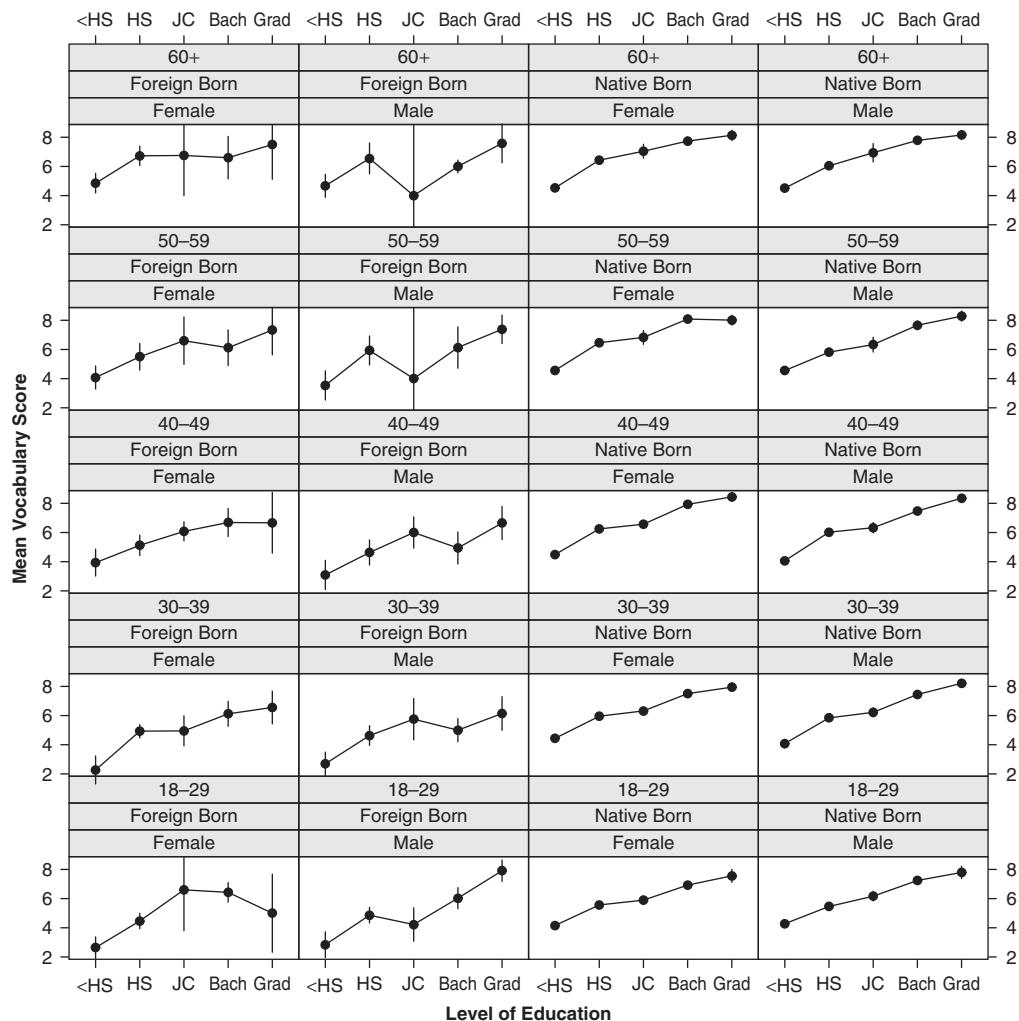


Figure 8.6 Mean vocabulary score by level of education, age group, place of birth, and sex, using data from the U.S. General Social Surveys. The education levels represented are less than high school (< HS), high school (HS), junior college (JC), bachelor's degree (Bach), and graduate degree (Grad). The dots represent the cell means. The vertical line around each dot is ± 2 standard errors around the corresponding cell mean. In some cells, these intervals extend beyond the end points of the vertical axis, while in other cells, the lines are so short that they are not discernible.

prove to be “statistically significant,” and we may wish to ignore such effects in describing the data. Chapter 22 presents methods for selecting a statistical model to summarize data based on considerations other than p -values. Anticipating that discussion, I have settled on a model for the vocabulary data that includes main effects of sex, place of birth, education, and age group, and the two-way interaction between place of birth and age group. This model, with $R^2 = .267$, accounts for almost as much variation in the vocabulary scores as the full model,

Table 8.6 Four-Way ANOVA of Vocabulary Score by Sex, Place of Birth, Education, and Age Groups

Source	Sum of Squares	df	Mean Square	F	p
Sex (S)	127.	1	127.00	37.58	<<.0001
Place of Birth (B)	1,122.	1	1122.00	331.13	<<.0001
Education (E)	20,556.	4	5139.00	1516.38	<<.0001
Age Group (A)	1,211.	4	302.75	89.31	<<.0001
S×B	3.	1	3.00	0.80	.37
S×E	62.	4	15.50	4.56	.001
S×A	104.	4	26.00	7.66	<.0001
B×E	102.	4	25.50	7.50	<.0001
B×A	240.	4	60.00	17.73	<<.0001
E×A	103.	16	6.44	1.90	.02
S×B×E	37.	4	9.25	2.76	.03
S×B×A	18.	4	4.50	1.32	.26
S×E×A	74.	16	4.63	1.36	.15
B×E×A	110.	16	6.87	2.03	.009
S×B×E×A	98.	16	6.13	1.81	.024
Residuals	62,917.	18,565	3.39		
Total	86,833.	18,664			

NOTE: The various sums of squares are computed in conformity with the principle of marginality.

for which $R^2 = .275$, despite the fact that the former has only 15 coefficients and the latter 100 coefficients! Still, given the large sample, an incremental F -test reveals highly statistically significant lack of fit:

$$F_0 = \frac{18,665 - 100}{85} \times \frac{.275 - .267}{1 - .267} = 2.38$$

$$df = 85 \text{ and } 18,565$$

$$p << .0001$$

Figure 8.7 shows “effect displays” for the simplified model.³¹ In computing each effect, other explanatory variables are held to average values—in the case of factors (and all of the explanatory variables here are factors), to their observed distribution in the data. It is apparent from Figure 8.7 that education is by far the largest influence on vocabulary. The sex main effect, in contrast, is quite small—only a fraction of a word on the 10-word test. Age apparently makes more of a difference to the vocabulary scores of the foreign born than of the native born, and the vocabulary advantage of the native born grows smaller with age.

Computing the Effect Display*

To compute the effects in Figure 8.7, each variable in a high-order term is allowed to range over its values, while other explanatory variables are set to “average” values. In the case of a factor, we fix the regressors coding the main effects for the factors to their means, which is

³¹Effect displays were introduced in Section 7.3.4.

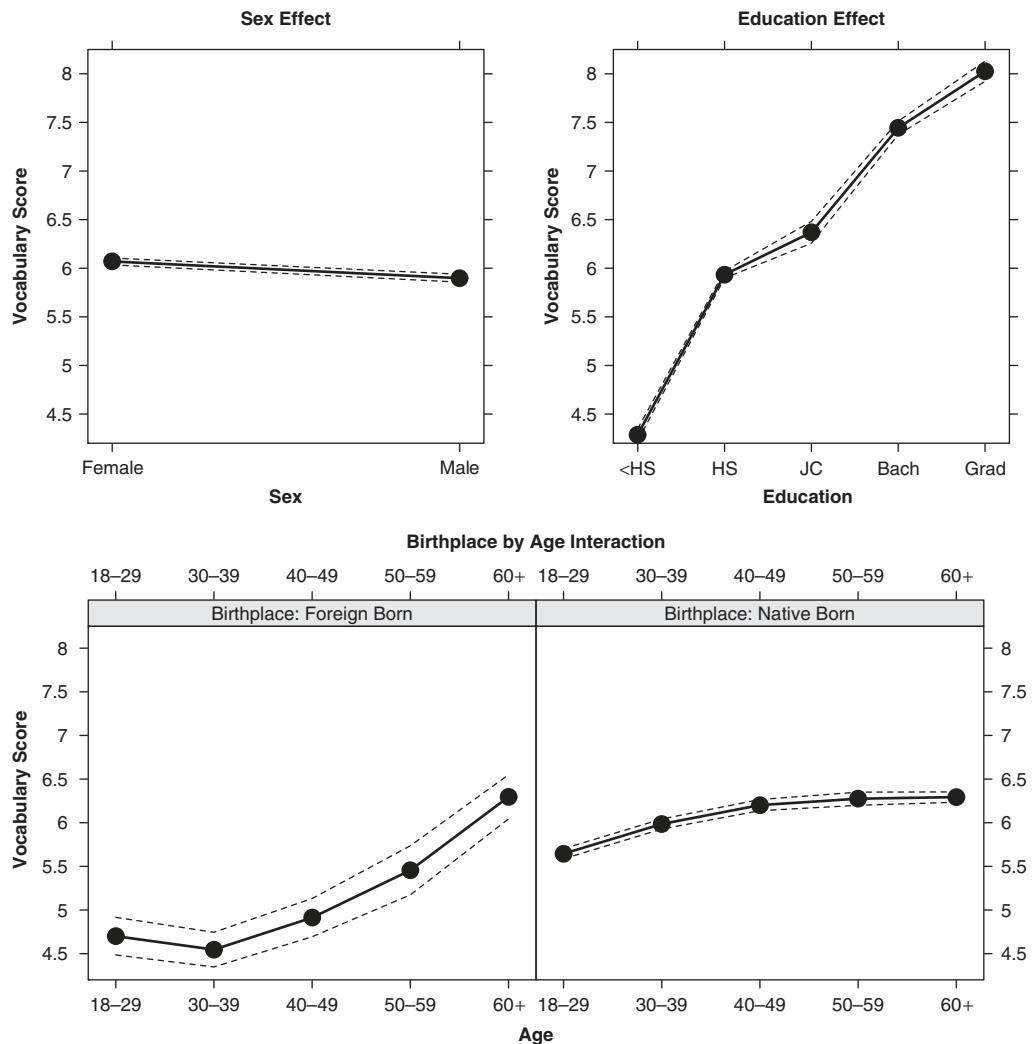


Figure 8.7 Effect displays for the simplified model fit to the GSS vocabulary data, showing the main effects of sex and education, as well as the interaction between place of birth and age—the high-order terms in the model. The broken lines give pointwise 95% confidence intervals around the estimated effects.

equivalent to fixing the distribution of the factor to the sample proportions at its various levels; interaction regressors are fixed to the products of the main-effect regressors marginal to the interaction. Effect displays computed in this manner are invariant with respect to the coding of factors, as long as the model obeys the principle of marginality.

Table 8.7 shows the quantities used to compute two of the fitted values in the effect displays in Figure 8.7: (1) the effect of membership in the “graduate degree” category for the main effect of education and (2) the effect of membership in the combination of categories “foreign born” and “40–49” for the interaction between place of birth and age.

Table 8.7 Computation of Effects for the Model Including Main Effects of Sex and Education and the Interaction Between Place of Birth and Age

Regressor	Coefficient B_j	Regressor Mean, \bar{X}_j	Graduate x_{j1}	Foreign, 40–49 x_{j2}
Constant	6.0350	1.0000	1.0000	1.0000
Sex (Female)	0.0861	0.1442	0.1442	0.1442
Education (<HS)	-2.1247	0.1308	-1.0000	0.1308
Education (HS)	-0.4778	0.4669	-1.0000	0.4669
Education (JC)	-0.0431	-0.0102	-1.0000	-0.0102
Education (Bach)	1.0315	0.0817	-1.0000	0.0817
Birthplace (Foreign)	-0.4490	-0.8649	-0.8649	1.0000
Age (18–29)	-0.4585	0.0018	0.0018	0.0000
Age (30–39)	-0.3654	0.0114	0.0114	0.0000
Age (40–49)	-0.0738	-0.0343	-0.0343	1.0000
Age (50–59)	0.2337	-0.0893	-0.0893	0.0000
Foreign \times 18–29	-0.0231	—	-0.8649×0.0018	0.0000
Foreign \times 30–39	-0.2704	—	-0.8649×0.0114	0.0000
Foreign \times 40–49	-0.1956	—	-0.8649×-0.0343	1.0000
Foreign \times 50–59	0.0389	—	-0.8649×-0.0893	0.0000

NOTE: The column labeled x_{j1} contains the values of the regressors used to compute the effects of membership in the “graduate degree” category of the education factor; the column labeled x_{j2} contains the values of the regressors used to compute the effects of membership in the combination of categories “foreign born” and “40–49” for the interaction of place of birth with age.

- To calculate the first effect from Table 8.7, we have

$$\begin{aligned}\hat{y}_1 &= \sum_{j=1}^{15} B_j x_{j1} \\ &= 6.0350 \times 1 + 0.0861 \times 0.1442 - 2.1247 \times -1.0000 \\ &\quad + \cdots + 0.0389 \times -0.8649 \times -0.0893 \\ &= 8.0264\end{aligned}$$

Note that in this computation, the regression constant is multiplied by 1 and that because “graduate degree” is the *last* category of education, all the regressors for education take on the value -1 by virtue of the sigma constraint on the coefficients of the education main effect. Other main-effect regressors are set to their mean values and interaction regressors to the products of the mean values of the main effects composing the interactions.

- Similarly, to compute the second effect,

$$\begin{aligned}\hat{y}_2 &= \sum_{j=1}^{15} B_j x_{j2} \\ &= 6.0350 \times 1 + 0.0861 \times 0.1442 + \cdots - 0.0738 \times 1 \\ &\quad + \cdots - 0.1956 \times 1 + 0.0389 \times 0 \\ &= 4.9129\end{aligned}$$

Here, the deviation regressor for place of birth takes on the value 1 (i.e., foreign), as do the regressor for the 40–49 age category and the product regressor for these two categories.

Because the effects are just weighted sums of the regression coefficients, their standard errors can be computed from the coefficient sampling variances and covariances.³²

8.3.3 Empty Cells in ANOVA

As the number of factors increases, the number of cells grows at a much faster rate: For p dichotomous factors, for example, the number of cells is 2^p . One consequence of this proliferation is that some combinations of factor categories may not be observed; that is, certain cells in the p -way classification may be empty.

Nevertheless, we can use our deviation-coding approach to estimation and testing in the presence of empty cells as long as the *marginal* frequency tables corresponding to the effects that we entertain contain no empty cells. For example, in a two-way classification with an empty cell, we can safely fit the main-effects model (see below), because the one-way frequency counts for each factor separately contain no 0s. The full model with interactions is not covered by the rule, however, because the two-way table of counts contains a 0 frequency. By extension, the rule *never* covers the p -way interaction when there is a 0 cell in a p -way classification.³³

To illustrate the difficulties produced by empty cells, I will develop a very simple example for a 2×2 classification with cell frequencies:

	C_1	C_2	Row marginal
R_1	n_{11}	n_{12}	$n_{11} + n_{12}$
R_2	n_{21}	0	n_{21}
Column marginal	$n_{11} + n_{21}$	n_{12}	n

That is, the cell frequency n_{22} is 0. Because there are no observations in this cell, we cannot estimate the cell mean μ_{22} . Writing out the other cell means in terms of the sigma-restricted model parameters produces three equations:

$$\begin{aligned}\mu_{11} &= \mu + \alpha_1 + \beta_1 + \gamma_{11} \\ \mu_{12} &= \mu + \alpha_1 + \beta_2 + \gamma_{12} = \mu + \alpha_1 - \beta_1 - \gamma_{11} \\ \mu_{21} &= \mu + \alpha_2 + \beta_1 + \gamma_{21} = \mu - \alpha_1 + \beta_1 - \gamma_{11}\end{aligned}$$

There are, then, four independent parameters (μ , α_1 , β_1 , and γ_{11}) but only three population means in observed cells, so the parameters are not uniquely determined by the means.

³²See Exercise 9.14.

³³It may be possible, however, to estimate and test effects not covered by this simple rule, but determining whether tests are possible and specifying sensible hypotheses to be tested are considerably more complex in this instance. For details see, for example, Searle (1971, pp. 318–324), Hocking and Speed (1975, pp. 711–712), and Speed et al. (1978, pp. 110–111). The advice given in Section 8.2.7 regarding care in the use of computer programs for ANOVA of unbalanced data applies even more urgently when there are empty cells.

Now imagine that we can reasonably specify the *absence* of two-way interactions for these data. Then, according to our general rule, we should be able to estimate and test the R and C main effects because there are observations at each level of R and at each level of C . The equations relating cell means to independent parameters become

$$\begin{aligned}\mu_{11} &= \mu + \alpha_1 + \beta_1 \\ \mu_{12} &= \mu + \alpha_1 + \beta_2 = \mu + \alpha_1 - \beta_1 \\ \mu_{21} &= \mu + \alpha_2 + \beta_1 = \mu - \alpha_1 + \beta_1\end{aligned}$$

Solving for the parameters in terms of the cell means produces³⁴

$$\begin{aligned}\mu &= \frac{\mu_{12} + \mu_{21}}{2} \\ \alpha_1 &= \frac{\mu_{11} - \mu_{21}}{2} \\ \beta_1 &= \frac{\mu_{11} - \mu_{12}}{2}\end{aligned}$$

These results make sense, for, in the *absence* of interaction:

- The cell means μ_{12} and μ_{21} are “balanced” with respect to both sets of main effects, and therefore their average serves as a suitable definition of the grand mean.
- The difference $\mu_{11} - \mu_{21}$ gives the effect of changing R while C is held constant (at Level 1), which is a suitable definition of the main effect of R .
- The difference $\mu_{11} - \mu_{12}$ gives the effect of changing C while R is held constant (at Level 1), which is a suitable definition of the main effect of C .

8.4 Analysis of Covariance

Analysis of covariance (ANCOVA) is a term used to describe linear models that contain both qualitative and quantitative explanatory variables. The method is, therefore, equivalent to dummy-variable regression, discussed in the previous chapter, although the ANCOVA model is parametrized differently from the dummy-regression model.³⁵ Traditional applications of ANCOVA use an additive model (i.e., without interactions). The traditional additive ANCOVA model is a special case of the more general model that I present here.

In ANCOVA, an ANOVA formulation is used for the main effects and interactions of the qualitative explanatory variables (i.e., the factors), and the quantitative explanatory variables (or *covariates*) are expressed as deviations from their means. Neither of these variations represents an essential change, however, for the ANCOVA model provides the same fit to the data as the dummy-regression model. Moreover, if tests are formulated following the principle of marginality, then precisely the same sums of squares are obtained for the two parameterizations. Nevertheless, the ANCOVA parameterization makes it simple to formulate sensible (if

³⁴The 2×2 classification with one empty cell is especially simple because the number of parameters in the main-effects model is equal to (i.e., no fewer than) the number of observed cell means. This is not generally the case, making a general analysis considerably more complex.

³⁵Usage here is not wholly standardized, and the terms *dummy regression* and *analysis of covariance* are often taken as synonymous.

ordinarily uninteresting) tests for lower-order terms in the presence of their higher-order relatives.

I will use Moore and Krupat's study of conformity and authoritarianism to illustrate ANCOVA. When we last encountered these data, both explanatory variables—partner's status and authoritarianism—were treated as factors.³⁶ Partner's status is dichotomous, but authoritarianism is a quantitative score (the "F-scale"), which was arbitrarily categorized for the two-way ANOVA. Here, I will treat authoritarianism more naturally as a covariate.

A dummy-regression formulation, representing authoritarianism by X , and coding $D = 1$ in the low partner's status group and $D = 0$ in the high partner's status group produces the following fit to the data (with estimated standard errors in parentheses below the coefficients):

$$\begin{aligned}\hat{Y} &= 20.79 - 0.1511X - 15.53D + 0.2611(X \times D) \\ (3.26) \quad (0.0717) \quad (4.40) \quad (0.0970) \\ R^2 &= .2942\end{aligned}\tag{8.17}$$

It makes sense, in this model, to test whether the interaction coefficient is statistically significant (clearly it is), but—as explained in the previous chapter—it is not sensible to construe the coefficients of X and D as “main effects” of authoritarianism and partner's status: The coefficient of X is the authoritarianism slope in the high-status group, while the coefficient of D is the difference in the regression lines for the two groups at an authoritarianism score of $X = 0$.

An ANCOVA model for the Moore and Krupat experiment is

$$Y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \bar{X}) + \gamma_j(X_{ij} - \bar{X}) + \varepsilon_{ij}\tag{8.18}$$

where

- Y_{ij} is the conformity score for subject i in category j of partner's status;
- μ is the general level of conformity;
- α_j is the main effect of membership in group j of partner's status;
- β is the main-effect slope of authoritarianism, X ;
- γ_j is the interaction between partner's status and authoritarianism for group j ;
- ε_{ij} is the error; and
- the mean authoritarianism score \bar{X} is computed over all the data.

To achieve a concrete understanding of the model in Equation 8.18, let us—as is our usual practice—write out the model separately for each group:

$$\begin{aligned}\text{Low status: } Y_{i1} &= \mu + \alpha_1 + \beta(X_{i1} - \bar{X}) + \gamma_1(X_{i1} - \bar{X}) + \varepsilon_{i1} \\ &= \mu + \alpha_1 + (\beta + \gamma_1)(X_{i1} - \bar{X}) + \varepsilon_{i1} \\ \text{High status: } Y_{i2} &= \mu + \alpha_2 + \beta(X_{i2} - \bar{X}) + \gamma_2(X_{i2} - \bar{X}) + \varepsilon_{i2} \\ &= \mu + \alpha_2 + (\beta + \gamma_2)(X_{i2} - \bar{X}) + \varepsilon_{i2}\end{aligned}$$

It is immediately apparent that there are too many parameters: We are fitting one line in each of two groups, which requires four parameters, but there are six parameters in the model— μ , α_1 , α_2 , β , γ_1 , and γ_2 .

³⁶See Section 8.2.

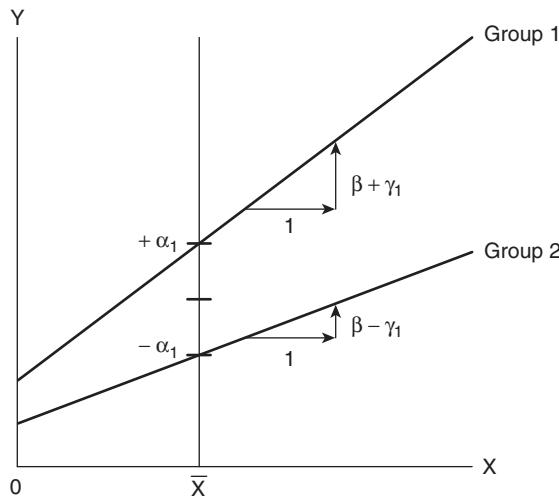


Figure 8.8 The analysis-of-covariance model for two groups, permitting different within-group slopes.

We require two restrictions, and to provide them we will place sigma constraints on the α s and γ s:

$$\begin{aligned}\alpha_1 + \alpha_2 &= 0 \Rightarrow \alpha_2 = -\alpha_1 \\ \gamma_1 + \gamma_2 &= 0 \Rightarrow \gamma_2 = -\gamma_1\end{aligned}$$

Under these constraints, the two regression equations become

$$\begin{aligned}\text{Low status: } Y_{i1} &= \mu + \alpha_1 + (\beta + \gamma_1)(X_{i1} - \bar{X}) + \varepsilon_{i1} \\ \text{High status: } Y_{i2} &= \mu - \alpha_1 + (\beta - \gamma_1)(X_{i2} - \bar{X}) + \varepsilon_{i2}\end{aligned}$$

The parameters of the constrained model therefore have the following straightforward interpretations (see Figure 8.8):

- μ is midway between the two regression lines above the mean of the covariate, \bar{X} .
- α_1 is half the difference between the two regression lines, again above \bar{X} .
- β is the average of the slopes $\beta + \gamma_1$ and $\beta - \gamma_1$ for the two within-group regression lines.
- γ_1 is half the difference between the slopes of the two regression lines.

Note, in particular, that the constrained parameters α_1 and β are reasonably interpreted as “main effects”—that is, the partial effect of one explanatory variable averaged over the other explanatory variable—even when interactions are present in the model. As usual, however, main effects are likely not of interest when interactions are present.

To fit the model to the data, we need to code a deviation regressor S for partner’s status:

<i>Partner's Status</i>	<i>S</i>
Low	1
High	-1

Then we can regress conformity on S , the mean deviation scores for X , and the product of S and $X - \bar{X}$,

$$Y_{ij} = \mu + \alpha_1 S_{ij} + \beta(X_{ij} - \bar{X}) + \gamma_1[S_{ij}(X_{ij} - \bar{X})] + \varepsilon_{ij}$$

producing the following fit to the Moore and Krupat data:

$$\begin{aligned}\hat{Y} &= 12.14 - 2.139S - 0.02055(X - \bar{X}) + 0.1306[S(X - \bar{X})] \\ (0.68) &\quad (0.681) \quad (0.04850) \quad (0.0485) \\ R^2 &= .2942\end{aligned}\tag{8.19}$$

Because each set of effects has one degree of freedom, incremental F -tests for the main effects and interactions are equivalent to the t -tests produced by dividing each coefficient by its standard error. It is apparent, then, that the partner's status \times authoritarianism interaction is statistically significant, as is the status main effect, but the authoritarianism main effect is not. You can verify that the two regression lines derived from the fitted ANCOVA model (8.19) are the same as those derived from the dummy-regression model (8.17).³⁷

ANCOVA is an alternative parameterization of the dummy-regression model, employing deviation-coded regressors for factors and expressing covariates as deviations from their means. The ANCOVA model can incorporate interactions among factors and between factors and covariates.

8.5 Linear Contrasts of Means

I have explained how the overparametrized ANOVA model can be fit to data by placing a sufficient number of linear restrictions on the parameters of the model. Different restrictions produce different regressors and hence different parameter estimates but identical sums of squares—at least for models that conform to the principle of marginality. We have examined in some detail two schemes for coding regressors for a factor: dummy (0, 1) coding and deviation (1, 0, -1) coding. The coefficients for a set of dummy-coded regressors compare each level of a factor with the baseline level, while the coefficients for a set of deviation-coded regressors compare each level (but the last) with the average of the levels.

We do not generally test hypotheses about individual coefficients for dummy-coded or deviation-coded regressors, but we can do so, if we wish. For dummy-coded regressors in

³⁷See Exercise 8.8.

Table 8.8 Data From Friendly and Franklin's (1980) Experiment on the Effects of Presentation on Recall

Condition		
SFR	B	M
39	40	40
25	38	39
37	39	34
25	37	37
29	39	40
39	24	36
21	30	36
39	39	38
24	40	36
25	40	30

NOTE: The data in the table are the number of words correctly recalled by each subject on the final trial of the experiment.

one-way ANOVA, a t -test or F -test of $H_0: \alpha_1 = 0$, for example, is equivalent to testing for the difference in means between the first group and the baseline group, $H_0: \mu_1 = \mu_m$. For deviation-coded regressors, testing $H_0: \alpha_1 = 0$ is equivalent to testing for the difference between the mean of the first group and the average of all the group means, $H_0: \mu_1 = \mu..$

In this section, I will explain a simple procedure for coding regressors that permits us to test specific hypotheses about *linear contrasts* (also called *linear comparisons*) among group means.³⁸ Although I will develop this technique for one-way ANOVA, contrast-coded regressors can also be employed for any factor in a two-way or higher-way ANOVA or in an ANCOVA.³⁹

For concreteness, let us examine the data in Table 8.8, which are drawn from an experimental study by Friendly and Franklin (1980) of the effects of presentation format on learning and memory.⁴⁰ Subjects participating in the experiment read a list of 40 words. Then, after performing a brief distracting task, the subjects were asked to recall as many of the words as possible. This procedure was repeated for five trials. Thirty subjects were randomly assigned to three conditions: In the control or “standard free recall” (*SFR*) condition, the order of presentation of the words on the list was randomized for each of the five trials of the experiment. In the two experimental conditions, recalled words were presented in the order in which they were listed by the subject on the previous trial. In one of these conditions (labeled *B*), the recalled words were presented as a group *before* the forgotten ones, while in the other condition (labeled *M* for *meshed*), the recalled and forgotten words were interspersed. Friendly and Franklin expected that making the order of presentation contingent on the subject’s previous performance would enhance recall. The data recorded in the table are the number of words correctly recalled by each subject for the final trial of the experiment.

³⁸A more general treatment of this topic may be found in Section 9.1.2.

³⁹See Exercise 8.11.

⁴⁰I am grateful to Michael Friendly of York University for providing these data.

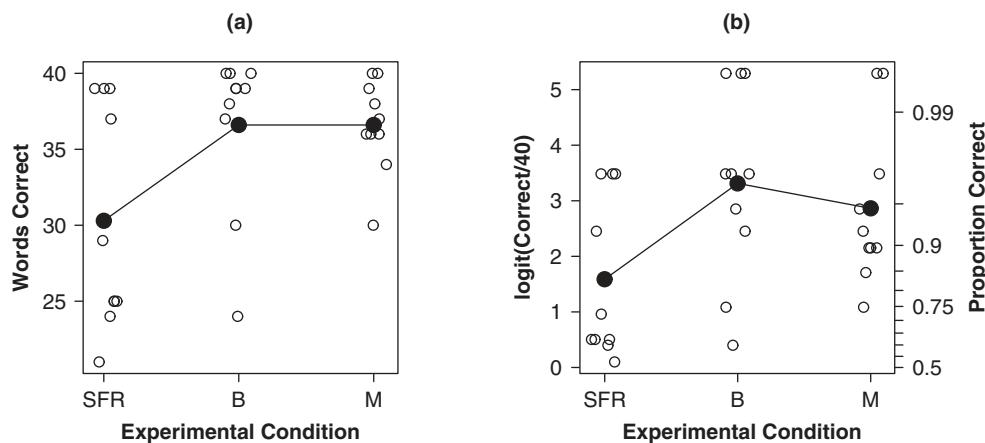


Figure 8.9 Horizontally jittered scatterplots for the Friendly and Franklin memory data: (a) for number of words correct; (b) for the logit of the proportion of words correct.

Means and standard deviations for Friendly and Franklin's memory data are as follows:

	<i>Experimental Condition</i>		
	<i>SFR</i>	<i>B</i>	<i>M</i>
Mean	30.30	36.60	36.60
Standard deviation	7.33	5.34	3.03

The mean number of words recalled is higher in the experimental conditions than in the control; the control group also has the largest standard deviation. A jittered scatterplot of the number of words recalled by condition, shown in Figure 8.9(a), reveals a problem with the data: The data are disguised proportions (number correctly recalled of 40 words), and many subjects—particularly in the *B* and *M* conditions—are at or near the maximum. This “ceiling effect” produces negatively skewed distributions in the two experimental conditions. Logit-transforming the data helps, as shown in Figure 8.9(b).⁴¹ The means and standard deviations for the transformed data are as follows:

	<i>Experimental Condition</i>		
	<i>SFR</i>	<i>B</i>	<i>M</i>
Mean	1.59	3.31	2.86
Standard deviation	1.46	1.71	1.43

⁴¹Because some subjects recalled all the words correctly, I mapped the proportions to the interval [.005, .995] prior to computing the logits, as explained in Section 4.5.

The logit transformation, therefore, has also made the group standard deviations more similar.

A linear contrast in the group means tests the hypothesis that a particular linear combination of population group means is 0. For the Friendly and Franklin memory experiment, we might wish to test the null hypothesis that the mean for the control group is no different from the average of the means for the experimental groups:⁴²

$$H_0: \mu_1 = \frac{\mu_2 + \mu_3}{2}$$

and the hypothesis that the means for the two experimental groups are the same:

$$H_0: \mu_2 = \mu_3$$

The first hypothesis can be rewritten as

$$H_0: 1\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 = 0$$

and the second hypothesis as

$$H_0: 0\mu_1 + 1\mu_2 - 1\mu_3 = 0$$

Then, the weights for the group means in these null hypotheses can be used to formulate two linear-contrast regressors, C_1 and C_2 :

Group	C_1	C_2
(1) SFR	1	0
(2) B	$-\frac{1}{2}$	1
(3) M	$-\frac{1}{2}$	-1

This simple approach to coding linear contrasts will work as long as the following conditions are satisfied:⁴³

1. We need one linear contrast for each degree of freedom. An m -category factor therefore requires $m - 1$ contrasts.
2. Each column of the contrast-coding table must sum to 0.
3. The products of corresponding codes for *different* contrasts must also sum to 0. For the illustration,

$$(1 \times 0) + (-\frac{1}{2} \times 1) + (-\frac{1}{2} \times -1) = 0$$

When there are equal numbers of observations in the groups, these rules ensure that the contrast regressors are uncorrelated. As a consequence, the regression sum of squares for the ANOVA can be decomposed into components due to the contrasts. When the group frequencies are unequal, the regression sum of squares does not decompose in this simple manner, but properly formulated contrasts are still useful for testing hypotheses about the population group means. Because each contrast has one degree of freedom, we can test it by a t -test (dividing

⁴²It would also be reasonable to compare each experimental group with the control group. The comparison could be easily accomplished by using dummy coding, treating the control group as the baseline category.

⁴³See Section 9.1.2 for an explanation of these rules and for a more flexible and general approach to constructing contrasts.

the estimated coefficient for the contrast by its standard error) or—equivalently—by an incremental F -test.

Linear contrasts permit the researcher to test specific hypotheses about means within the framework of ANOVA. A factor with m categories gives rise to $m - 1$ contrasts, one for each degree of freedom. A simple procedure for constructing contrasts requires that the codes for each contrast sum to 0 and that the products of codes for each pair of contrasts also sum to 0.

For Friendly and Franklin's experiment, the fitted model (working with the logits) is

$$\begin{aligned}\hat{Y} &= 2.5880 - 0.9998C_1 + 0.2248C_2 \\ &\quad (0.2804) \quad (0.3966) \quad (0.3434) \\ R^2 &= .2008\end{aligned}$$

and the ANOVA table is

Source	SS	df	MS	F	p
Groups	16.005	2	8.002	3.39	.049
C_1	14.994	1	14.994	6.35	.018
C_2	1.011	1	1.011	0.43	.52
Residuals	63.696	27	2.359		
Total	79.701	29			

We therefore have evidence that the two experimental conditions promoted higher levels of recall than in the control condition (contrast C_1), but no evidence for the superiority of one experimental treatment relative to the other (contrast C_2). Because there are equal numbers of observations in the three groups, the sums of squares for the contrasts add to the sum of squares for groups (i.e., to the regression sum of squares).

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 8.1. *The usual t -statistic for testing a difference between the means of two independently sampled groups, under the assumptions of normality and equal group variances, is

$$t_0 = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{SE}(\bar{Y}_1 - \bar{Y}_2)}$$

where

$$\begin{aligned}\text{SE}(\bar{Y}_1 - \bar{Y}_2) &= S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ S^2 &= \frac{\sum_{j=1}^{n_1} (Y_{i1} - \bar{Y}_1)^2 + \sum_{j=1}^{n_2} (Y_{i2} - \bar{Y}_2)^2}{n_1 + n_2 - 2}\end{aligned}$$

Here, \bar{Y}_1 and \bar{Y}_2 are the means of the two groups, n_1 and n_2 are the numbers of observations in the groups, and Y_{i1} and Y_{i2} are the observations themselves. Let F_0 be the one-way ANOVA F -statistic for testing the null hypothesis $H_0: \mu_1 = \mu_2$. Prove that $t_0^2 = F_0$ and that, consequently, the two tests are equivalent.

Exercise 8.2. *Show that one of the restrictions on the interaction parameters of the two-way ANOVA model,

$$\begin{aligned}\sum_{j=1}^r \gamma_{jk} &= 0 \quad \text{for } k = 1, \dots, c \\ \sum_{k=1}^c \gamma_{jk} &= 0 \quad \text{for } j = 1, \dots, r\end{aligned}$$

is redundant. [Hint: Construct a table of the interaction parameters, labeling the rows $1, 2, \dots, r - 1, r$ and the columns $1, 2, \dots, c$. Insert a column for row sums after the last column and a row for column sums after the last row. At the bottom-right corner of the table is the overall sum of the interaction parameters, $\sum_{j=1}^r \sum_{k=1}^c \gamma_{jk}$. (This table looks much like the table of cell means on page 187, with γ s replacing the μ s.) Then place a 0 in each entry of the column of row sums, corresponding to the r restrictions $\sum_{k=1}^c \gamma_{jk} = 0$. From these restrictions, show that $\sum \sum \gamma_{jk} = 0$, and place this 0 in the lower-right corner. Next, specify 0s for all but the last column sum, $\sum_{j=1}^r \gamma_{jk} = 0$, for $k = 1, \dots, c - 1$. Finally, show that the last column sum, $\sum_{j=1}^r \gamma_{jc}$, is necessarily 0.]

Exercise 8.3. *Demonstrate that the hypothesis

$$H_0: \text{all } \gamma_{jk} = 0$$

for the sigma-constrained two-way ANOVA model is equivalent to the null hypothesis of no interaction stated in terms of the cell means:

$$\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'} \quad \text{for all } j, j' \text{ and } k, k'$$

[Hint: Write out each of the interaction parameters γ_{jk} , $\gamma_{j'k}$, $\gamma_{jk'}$, and $\gamma_{j'k'}$ in terms of the cell and marginal means (e.g., $\gamma_{jk} = \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..}$). Then show that when the γ s are all 0, $\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'}$ (i.e., $0 - 0 = 0 - 0$) implies that $\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'}$.]

Exercise 8.4. *Show that in the sigma-constrained three-way ANOVA model, the null hypothesis

$$H_0: \text{all } \alpha_{ABC(jkm)} = 0$$

is equivalent to the hypothesis given in Equation 8.16 on page 179. (Hint: See Exercise 8.3.)

Exercise 8.5. The geometry of effects in three-way ANOVA: Contrived parameter values for a three-way ANOVA model (each set satisfying the sigma constraints) are given in the following tables:

$\alpha_{A(j)}$		
A_1		A_2
2		-2
$\alpha_{B(k)}$		
B_1		B_2
-3		3
$\alpha_{C(m)}$		
C_1	C_2	C_3
1	-3	2
$\alpha_{AB(jk)}$		
	B_1	B_2
A_1	-2	2
A_2	2	-2
$\alpha_{AC(jm)}$		
	C_1	C_2
A_1	1	1
A_2	-1	-1
	C_3	
		-2
		2
$\alpha_{BC(km)}$		
	C_1	C_2
B_1	0	3
B_2	0	-3
	C_3	
		-3
		3
$\alpha_{ABC(jkm)}$		
	C_1	C_2
A_1B_1	1	-2
A_1B_2	-1	2
A_2B_1	-1	2
A_2B_2	1	-2
	C_3	
		1
		-1
		-1
		1

Use these parameter values to construct population cell means for each of the following models (simply sum the parameters that pertain to each of the 12 cells of the design):

- (a) Main effects only,

$$\mu_{jkm} = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)}$$

- (b) One two-way interaction,

$$\mu_{jkm} = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)} + \alpha_{AC(jm)}$$

- (c) All three two-way interactions,

$$\mu_{jkm} = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)} + \alpha_{AB(jk)} + \alpha_{AC(jm)} + \alpha_{BC(km)}$$

- (d) The full model,

$$\mu_{jkm} = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)} + \alpha_{AB(jk)} + \alpha_{AC(jm)} + \alpha_{BC(km)} + \alpha_{ABC(jkm)}$$

For each of these models, do the following:

- (i) Draw a graph of the cell means, placing factor C on the horizontal axis. Use different lines (solid and broken) or line colors for the levels of factor A and different symbols (e.g., circle and square) for the levels of factor B . Note that there will be four connected profiles of means on each of these plots, one profile for each combination of categories of A and B across the three levels of C . Attempt to interpret the graphs in terms of the effects that are included in each model.
- (ii) Using the table of means generated from each of models (c) and (d), plot (for each model) the six *differences* across the levels of factor B , $\mu_{j1m} - \mu_{j2m}$, by the categories of factors A and C . Can you account for the different patterns of these two graphs in terms of the presence of three-way interactions in the second graph but not in the first?

Exercise 8.6. Adjusted means (continued): The notion of an “adjusted” mean was introduced in Exercises 7.2 and 7.3. Consider the main-effects model for the p -way classification:

$$\mu_{jk...r} \equiv E(Y_{ijk...r}) = \mu + \alpha_{A(j)} + \alpha_{B(k)} + \cdots + \alpha_{P(r)}$$

- (a) Show that if we constrain each set of effects to sum to 0, then the population marginal mean for category j of factor A is $\mu_{j...} = \mu + \alpha_{A(j)}$.
- (b) Let us define the analogous sample quantity, $\tilde{Y}_{j...} \equiv M + A_{A(j)}$, to be the *adjusted mean* in category j of factor A . How is this quantity to be interpreted?
- (c) Does the definition of the adjusted mean in part (b) depend fundamentally on the constraint that each set of effects sums to 0?
- (d) Can the idea of an adjusted mean be extended to ANOVA models that include interactions? (Cf. the discussion of effect displays in this and the preceding chapter.)

Exercise 8.7. ANOVA with equal cell frequencies: In higher-way ANOVA, as in two-way ANOVA, when cell frequencies are equal, the sum of squares for each set of effects can be calculated directly from the parameter estimates for the full model or, equivalently, in terms of cell and marginal means. To get the sum of squares for a particular set of effects, we simply need to square the parameter estimate associated with each cell, sum over all cells, and multiply by the common cell frequency, n' . For example, for a balanced three-way ANOVA,

$$\begin{aligned}
 \text{SS}(\alpha_{AB}) &= n' \sum_{j=1}^a \sum_{k=1}^b \sum_{m=1}^c A_{AB(jk)}^2 \\
 &= n' c \sum_{j=1}^a \sum_{k=1}^b A_{AB(jk)}^2 \\
 &= n' c \sum_{j=1}^a \sum_{k=1}^b (\bar{Y}_{jk\cdot} - \bar{Y}_{j\cdot\cdot} - \bar{Y}_{\cdot k\cdot} + \bar{Y}_{\dots})^2
 \end{aligned}$$

Write out similar expressions for $\text{SS}(\alpha_A)$ and $\text{SS}(\alpha_{ABC})$ in three-way ANOVA. Show that

$$\text{RSS} = (n' - 1) \sum_{j=1}^a \sum_{k=1}^b \sum_{m=1}^c S_{jkm}^2$$

where

$$S_{jkm}^2 = \frac{\sum_{i=1}^{n'} (Y_{ijkm} - \bar{Y}_{jkm})^2}{n' - 1}$$

is the variance in cell j, k, m of the design.

Exercise 8.8. Calculate the fitted regression equation for each group (low and high partner's status) in Moore and Krupat's conformity data using the dummy regression in Equation 8.17 (page 188). Calculate the fitted regression equation for each group using the analysis of covariance in Equation 8.19. Why must the two sets of equations be the same (within rounding error)?

Exercise 8.9. Adjusted means (concluded): The notion of an *adjusted mean* was discussed in Exercises 7.2, 7.3, and 8.6. Now consider the ANCOVA model for two factors, R and C , and two covariates, X_1 and X_2 :

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \delta_1(X_{ijk1} - \bar{X}_1) + \delta_2(X_{ijk2} - \bar{X}_2) + \varepsilon_{ijk}$$

Note that this formulation of the model permits interactions between the factors but not between the factors and the covariates.

- (a) How can the ANCOVA model be used to compute adjusted cell means for the $r \times c$ combinations of levels of the factors R and C ?
- (b) In computing adjusted means, is anything gained by expressing the covariates as deviations from their respective means rather than as raw scores?
- (c) If the interactions between the factors γ_{jk} are deleted from the model, how can we calculate adjusted means for the r categories of R and the c categories of C ?

The calculation of adjusted means in additive ANCOVA models is a traditional use of the ANCOVA. Further information on adjusted means can be found in Searle, Speed, and Milliken (1980). Adjusted means are special cases of effect displays, as developed in this and the preceding chapter.

Exercise 8.10. Testing contrasts using group means: Suppose that we wish to test a hypothesis concerning a contrast of group means in a one-way ANOVA:

$$H_0: c_1\mu_1 + c_2\mu_2 + \cdots + c_m\mu_m = 0$$

where $c_1 + c_2 + \cdots + c_k = 0$. Define the *sample value of the contrast* as

$$C \equiv c_1\bar{Y}_1 + c_2\bar{Y}_2 + \cdots + c_m\bar{Y}_m$$

and let

$$C'^2 \equiv \frac{C^2}{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_m^2}{n_m}}$$

C'^2 is the sum of squares for the contrast.

- (a) *Show that under the null hypothesis

- (i) $E(C) = 0$.
- (ii) $V(C) = \sigma_e^2 \left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_m^2}{n_m} \right)$.
- (iii) $t_0 = C'/S_E$ follows a *t*-distribution with $n - m$ degrees of freedom. [Hint: The \bar{Y}_j are independent, and each is distributed as $N(\mu_j, \sigma_e^2/n_j)$.]

- (d) Using Friendly and Franklin's memory data, verify that the test statistics obtained by the method of this exercise [i.e., in (a) (iii)] are the same as those produced by the incremental sum-of-squares approach, used in the text.

Exercise 8.11. *Contrasts in two-way ANOVA: A simple approach to formulating contrasts in two-way (and higher-way) ANOVA is first to specify contrasts separately for each set of main effects, obtaining interaction contrasts by forming all pairwise products of the main-effect contrasts. Then, as long as the main-effect contrasts satisfy the rules (on page 193), the interaction contrasts will as well. Imagine, for example, a 3×2 classification arising from an experiment in which the first factor consists of a control group (R_1) and two experimental groups (R_2 and R_3). The second factor is, say, gender, with categories male (C_1) and female (C_2). A possible set of main-effect contrasts for this experiment is

Group	Row contrast	
	A_1	A_2
R_1	2	0
R_2	-1	1
R_3	-1	-1

Column contrast	
Gender	B
C_1	1
C_2	-1

The following table shows the full set of main-effect and interaction contrasts for all six cells of the design (with the parameter for each contrast in parentheses):

Cell		(δ_1)	(δ_1)	(β)	(ζ_1)	(ζ_2)
Row	Column	A_1	A_2	B	A_1B	A_2B
1	1	2	0	1	2	0
1	2	2	0	-1	-2	0
2	1	-1	1	1	-1	1
2	2	-1	1	-1	1	-1
3	1	-1	-1	1	-1	-1
3	2	-1	-1	-1	1	1

Note that we use 2 degrees of freedom for condition main effects, 1 degree of freedom for the gender main effect, and 2 degrees of freedom for interaction. Explain the meaning of the following null hypotheses:

- (a) $H_0: \zeta_1 = 0$.
- (b) $H_0: \zeta_2 = 0$.
- (c) $H_0: \delta_1 = 0$.
- (d) $H_0: \delta_2 = 0$.
- (e) $H_0: \beta = 0$.

Exercise 8.12. Reanalyze Moore and Krupat's conformity data eliminating the two outlying observations, Subjects 16 and 19. Perform *both* a two-way ANOVA, treating authoritarianism as a factor, and an ANCOVA, treating authoritarianism as a covariate.

Exercise 8.13. *Equations 8.6 (page 167) show how the parameters in a dummy-coded two-way ANOVA model can be expressed in terms of the cell means μ_{jk} when the number of levels of the row factor R is $r = 2$ and the number of levels of the column factor C is $c = 3$.

- (a) Explain why the hypotheses $H_0: \beta_1 = 0$ and $H_0: \gamma_1 = \gamma_2 = 0$ cannot reasonably be construed as tests of the R and C main effects when the interaction parameters δ_{11} and δ_{12} are nonzero.
- (b) Now suppose that δ_{11} and δ_{12} are zero but that we fit the full model in Equation 8.5 to the data. Explain why $H_0: \beta_1 = 0$ and $H_0: \gamma_1 = \gamma_2 = 0$ now test hypotheses about the R and C main effects but do so in a nonoptimal manner (i.e., with low power).

Summary

- One-way ANOVA examines the relationship between a quantitative response variable and a factor. The omnibus F -statistic for the regression of the response variable on 0/1 dummy regressors constructed from the factor tests for differences in the response means across levels of the factor.

- The one-way ANOVA model,

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

is underdetermined because it uses $m + 1$ parameters to model m group means. This indeterminacy can be removed, however, by placing a restriction on its parameters. Setting one of the α_j s to 0 leads to $(0, 1)$ dummy-regressor coding. Constraining the α_j s to sum to 0 leads to $(1, 0, -1)$ deviation-regressor coding. The two coding schemes are equivalent in that they provide the same fit to the data, producing the same regression and residual sums of squares.

- As long as we construct tests that conform to the principle of marginality, we can code main effects in two- and higher-way ANOVA using dummy regressors, forming interaction regressors as all products of main-effect regressors for the main effects marginal to each interaction.
- The two-way ANOVA model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

incorporates the main effects and interactions of two factors. The factors interact when the profiles of population cell means are not parallel. The two-way ANOVA model is overparametrized, but it can be fit to data by placing suitable restrictions on its parameters. A convenient set of restrictions is provided by sigma constraints, specifying that each set of parameters (α_j , β_k , and γ_{jk}) sums to 0 over each of its coordinates. Testing hypotheses about the sigma-constrained parameters is equivalent to testing interaction-effect and main-effect hypotheses about cell and marginal means. There are two reasonable procedures for testing main-effect hypotheses in two-way ANOVA: Tests based on $SS(\alpha|\beta, \gamma)$ and $SS(\beta|\alpha, \gamma)$ (“type III” tests) employ models that violate the principle of marginality but are valid whether or not interactions are present. Tests based on $SS(\alpha|\beta)$ and $SS(\beta|\alpha)$ (“type II” tests) conform to the principle of marginality but are valid only if interactions are absent, in which case they are maximally powerful.

- The ANOVA model and procedures for testing hypotheses about main effects and interactions extend straightforwardly to three-way and higher-way classifications. In each case, the highest-order interaction corresponds to the number of factors in the model. It is not necessary, however, to specify a model that includes all terms through the highest-order interaction. Effect displays for the high-order terms in a model can clarify the interpretation of the model.
- It is possible to fit an ANOVA model to a classification containing empty cells when the marginal frequency tables corresponding to the terms in the model have no empty cells.
- ANCOVA is an alternative parameterization of the dummy-regression model, employing deviation-coded regressors for factors and expressing covariates as deviations from their means. The ANCOVA model can incorporate interactions among factors and between factors and covariates.
- Linear contrasts permit the researcher to test specific hypotheses about means within the framework of ANOVA. A factor with m categories gives rise to $m - 1$ contrasts, one for each degree of freedom. A simple procedure for constructing contrasts requires that the codes for each contrast sum to 0 and that the products of codes for each pair of contrasts also sum to 0.

9

Statistical Theory for Linear Models*

The purpose of this chapter is twofold: to deepen your knowledge of linear models and linear least-squares estimation and to provide a basis for more advanced work in social statistics—in the remainder of this book and more generally. Relying on the mathematical tools of linear algebra and elementary calculus,¹ we will revisit with greater rigor many of the topics treated informally in Chapters 5 through 8, developing the statistical theory on which the methods described in those chapters depend. The chapter concludes with an introduction to instrumental-variables estimation and two-stage least squares. The next chapter, on the vector geometry of linear models, provides intuitive insight into the statistical theory of linear models and facilitates the formal development of some topics, such as degrees of freedom.

9.1 Linear Models in Matrix Form

The general linear model is given by the equation

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

I have substituted the notationally more convenient β_0 for the regression constant α ; I will, for the time being, suppose that the X -values are fixed, hence the lowercase x_{ij} .²

Collecting the regressors into a row vector, appending a 1 for the constant, and placing the corresponding parameters in a column vector permits us to rewrite the linear model as

$$\begin{aligned} Y_i &= [1, x_{i1}, x_{i2}, \dots, x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i \\ &= \underset{(1 \times k+1)}{\mathbf{x}'_i} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} + \varepsilon_i \end{aligned}$$

For a sample of n observations, we have n such equations, which can be combined into a single matrix equation:

¹See online Appendices B and C for introductions to linear algebra and calculus.

²See Section 9.6 for a discussion of random regressors.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (9.1)$$

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

As we will see, with suitable specification of the contents of \mathbf{X} , called the *model matrix*, Equation 9.1 serves not only for multiple regression but for linear models generally.³

Because $\boldsymbol{\varepsilon}$ is a vector random variable, the assumptions of the linear model can be compactly restated in matrix form. The errors are assumed to be independent and normally distributed with zero expectation and common variance. Thus, $\boldsymbol{\varepsilon}$ follows a multivariate-normal distribution with expectation $E(\boldsymbol{\varepsilon}) = \mathbf{0}_{(n \times 1)}$ and covariance matrix $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2 \mathbf{I}_n$; in symbols, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$. The distribution of \mathbf{y} follows immediately:

$$\begin{aligned} \boldsymbol{\mu} &\equiv E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} \\ V(\mathbf{y}) &= E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = E[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'] \\ &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2 \mathbf{I}_n \end{aligned} \quad (9.2)$$

Furthermore, because it is simply a translation of $\boldsymbol{\varepsilon}$ to a different expectation, \mathbf{y} is also normally distributed: $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}_n)$.

The general linear model can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{y} is an $n \times 1$ vector of response-variable observations; \mathbf{X} is an $n \times k + 1$ matrix of regressors (called the model matrix), including an initial column of 1s for the constant regressor; $\boldsymbol{\beta}$ is a $k + 1 \times 1$ vector of parameters to be estimated; and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. The assumptions of the linear model can be compactly written as $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$.

9.1.1 Dummy Regression and Analysis of Variance

The model matrices for dummy-regression and analysis-of-variance (ANOVA) models—especially the latter—are strongly patterned. Consider the dummy-regression model

$$Y_i = \alpha + \beta x_i + \gamma d_i + \delta(x_i d_i) + \varepsilon_i$$

where (for concreteness) Y is income, x is years of education, and the dummy regressor d is coded 1 for men and 0 for women.⁴ In matrix form, this model becomes

³The model matrix is often called the *design matrix*, a term that is especially appropriate in experimental applications where the explanatory variables, and hence the regressors that compose the \mathbf{X} matrix, derive from the design of the experiment.

⁴This example was discussed in Chapter 7. Here, x and d are treated as fixed; random regressors are considered in Section 9.6.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ \hline 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \hline \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

To emphasize the pattern of the model matrix, the n_1 observations for women (for whom d and hence xd are 0) precede the $n - n_1$ observations for men.

Now consider the overparametrized one-way ANOVA model⁵

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad \text{for groups } j = 1, \dots, m.$$

The matrix form of the model is

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{12} \\ \vdots \\ Y_{n_2,2} \\ \vdots \\ Y_{1,m-1} \\ \vdots \\ Y_{n_{m-1},m-1} \\ \hline Y_{1m} \\ \vdots \\ Y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \\ \alpha_m \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1,1} \\ \hline \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2,2} \\ \vdots \\ \varepsilon_{1,m-1} \\ \vdots \\ \varepsilon_{n_{m-1},m-1} \\ \hline \varepsilon_{1m} \\ \vdots \\ \varepsilon_{n_m,m} \end{bmatrix} \quad (9.3)$$

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

It is apparent that the model matrix is of rank m , one less than the number of columns, because the first column of \mathbf{X} is the sum of the others. One solution is to delete a column, implicitly setting the corresponding parameter to 0. Deleting the last column of the model matrix, for example, sets $\alpha_m = 0$, establishing the last category as the baseline for a dummy-coding scheme.

Alternatively, imposing the sigma constraint $\sum_{j=1}^m \alpha_j = 0$ on the parameters leads to the following *full-rank* model matrix \mathbf{X}_F , composed of deviation-coded regressors; labeling each column of the matrix with the parameter to which it pertains,

⁵See Section 8.1.

$$\mathbf{X}_F = \begin{array}{c} \left[\begin{array}{ccccc} (\mu) & (\alpha_1) & (\alpha_2) & \cdots & (\alpha_{m-1}) \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \end{array} \right] \\ \hline \left[\begin{array}{ccccc} 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{array} \right] \\ \hline \left[\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{array} \right] \\ \hline \left[\begin{array}{ccccc} 1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{array} \right] \end{array} \quad (9.4)$$

There is, then, the following relationship between the group means $\boldsymbol{\mu} = \{\mu_j\}$ and the parameters of the constrained model:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{m-1} \\ \mu_m \end{bmatrix}_{(m \times 1)} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & \cdots & -1 \end{bmatrix}_{(m \times m)} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}_{(m \times 1)} \quad (9.5)$$

$$\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}_F$$

In this *parametric equation*, \mathbf{X}_B is the *row basis* of the full-rank model matrix, consisting of the m unique rows of \mathbf{X}_F , one for each group, and $\boldsymbol{\beta}_F$ is the parameter vector associated with the full-rank model matrix.

By construction, the $m \times m$ matrix \mathbf{X}_B is of full column rank and hence nonsingular, allowing us to invert \mathbf{X}_B and solve uniquely for the constrained parameters in terms of the cell means: $\boldsymbol{\beta}_F = \mathbf{X}_B^{-1} \boldsymbol{\mu}$. The solution follows a familiar pattern:⁶

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} = \begin{bmatrix} \mu. \\ \mu_1 - \mu. \\ \mu_2 - \mu. \\ \vdots \\ \mu_{m-1} - \mu. \end{bmatrix}$$

Let us next examine a two-way ANOVA model. To make the example manageable, suppose that there are two rows and three columns in the design. Imposing sigma constraints on the main effects and interactions produces the parametric equation:

⁶See Exercise 9.1(a) and Section 8.1.

$$\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} \quad (9.6)$$

$$\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}_F$$

(6×1) (6×6) (6×1)

As in one-way ANOVA, the row basis of the full-rank model matrix is nonsingular by construction, yielding the following solution for the parameters in terms of the cell means:⁷

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} = \begin{bmatrix} \mu.. \\ \mu_{1.} - \mu.. \\ \mu_{.1} - \mu.. \\ \mu_{.2} - \mu.. \\ \mu_{11} - \mu_{1.} - \mu_{.1} + \mu.. \\ \mu_{12} - \mu_{1.} - \mu_{.2} + \mu.. \end{bmatrix}$$

The model matrices for dummy-regression and ANOVA models are strongly patterned. In ANOVA, the relationship between group or cell means and the parameters of the linear model is expressed by the parametric equation $\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}_F$, where $\boldsymbol{\mu}$ is the vector of means, \mathbf{X}_B is the row basis of the full-rank model matrix, and $\boldsymbol{\beta}_F$ is the parameter vector associated with the full-rank model matrix. Solving the parametric equation for the parameters yields $\boldsymbol{\beta}_F = \mathbf{X}_B^{-1} \boldsymbol{\mu}$.

9.1.2 Linear Contrasts

The relationship between group means and the parameters of the ANOVA model is given by the parametric equation $\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}_F$; thus, as I have explained, the parameters are linear functions of the group means, $\boldsymbol{\beta}_F = \mathbf{X}_B^{-1} \boldsymbol{\mu}$. The full-rank parameterizations of the one-way ANOVA model that we have considered—dummy coding and deviation coding—permit us to test the null hypothesis of no differences among group means, but the individual parameters are not usually of interest. In certain circumstances, however, we can formulate \mathbf{X}_B so that the individual parameters of $\boldsymbol{\beta}_F$ incorporate interesting contrasts among group means.⁸

In Friendly and Franklin's (1980) memory experiment,⁹ for example, subjects attempted to recall words under three experimental conditions:

⁷See Exercise 9.1(b) and Section 8.2.3.

⁸Linear contrasts were introduced in Section 8.5.

⁹See Section 8.5.

1. the “standard free recall” (*SFR*) condition, in which words were presented in random order;
2. the “before” (*B*) condition, in which remembered words were presented before those forgotten on the previous trial; and
3. the “meshed” (*M*) condition, in which remembered words were interspersed with forgotten words but were presented in the order in which they were recalled.

I defined linear contrasts to test two null hypotheses:

1. $H_0: \mu_1 = (\mu_2 + \mu_3)/2$, that the mean of the *SFR* condition does not differ from the average of the means of the other two conditions and
2. $H_0: \mu_2 = \mu_3$, that the means of the *B* and *M* conditions do not differ.

These hypotheses can be written as linear functions of the group means: (1) $H_0: 1\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 = 0$, and (2) $H_0: 0\mu_1 + 1\mu_2 - 1\mu_3 = 0$. Then each hypothesis can be coded in a parameter of the model, employing the following relationship between parameters and group means:¹⁰

$$\begin{bmatrix} \mu \\ \zeta_1 \\ \zeta_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad (9.7)$$

$$\beta_F = \mathbf{X}_B^{-1} \boldsymbol{\mu}$$

One parameter, μ , is used to code the average of the group means, leaving two parameters to represent differences among the three group means. The hypothesis $H_0: \zeta_1 = 0$ is equivalent to the first null hypothesis; $H_0: \zeta_2 = 0$ is equivalent to the second null hypothesis.

Because the *rows* of \mathbf{X}_B^{-1} in Equation 9.7 are orthogonal, the *columns* of \mathbf{X}_B are orthogonal as well: Each column of \mathbf{X}_B is equal to the corresponding row of \mathbf{X}_B^{-1} divided by the sum of squared entries in that row;¹¹ thus,

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & \frac{2}{3} & 0 \\ 1 & -\frac{1}{3} & \frac{1}{2} \\ 1 & -\frac{1}{3} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \mu \\ \zeta_1 \\ \zeta_2 \end{bmatrix}$$

$$\boldsymbol{\mu} = \mathbf{X}_B \beta_F$$

The one-to-one correspondence between rows of \mathbf{X}_B^{-1} and columns of \mathbf{X}_B makes it simple to specify the latter matrix directly. Moreover, we can rescale the columns of the row basis for more convenient coding, as shown in Equation 9.8, without altering the hypotheses incorporated in the contrast coefficients: If, for example, $\zeta_1 = 0$, then any multiple of ζ_1 is 0 as well.

$$\mathbf{X}_B = \begin{bmatrix} 1 & 2 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \quad (9.8)$$

Although it is convenient to define contrasts that are orthogonal in the row basis, it is not necessary to do so. It is always possible to work backward from \mathbf{X}_B^{-1} (which expresses the

¹⁰Because it is natural to form hypotheses as linear combinations of means, I start by specifying \mathbf{X}_B^{-1} directly, rather than \mathbf{X}_B .

¹¹See Exercise 9.2.

parameters of the model as linear functions of the population group means) to \mathbf{X}_B , as long as the comparisons specified by \mathbf{X}_B^{-1} are linearly independent. Linear independence is required to ensure that \mathbf{X}_B^{-1} is nonsingular.¹²

If there are equal numbers (say n') of observations in the several groups, then an orthogonal model-matrix basis \mathbf{X}_B implies an orthogonal full-rank model matrix \mathbf{X}_F —because \mathbf{X}_F is produced by repeating each of the rows of \mathbf{X}_B an equal number (n') of times. The columns of an orthogonal model matrix represent independent sources of variation in the response variable, and therefore a set of orthogonal contrasts partitions the regression sum of squares into one-degree-of-freedom components, each testing a hypothesis of interest. When it is applicable, this is an elegant approach to linear-model analysis. Linear comparisons may well be of interest, however, even if group frequencies are unequal, causing contrasts that are orthogonal in \mathbf{X}_B to be correlated in \mathbf{X}_F .¹³

9.2 Least-Squares Fit

To find the least-squares coefficients, we write the fitted linear model as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where $\mathbf{b} = [B_0, B_1, \dots, B_k]'$ is the vector of fitted coefficients, and $\mathbf{e} = [E_1, E_2, \dots, E_n]'$ is the vector of residuals. We seek the coefficient vector \mathbf{b} that minimizes the residual sum of squares, expressed as a function of \mathbf{b} :

$$\begin{aligned} S(\mathbf{b}) &= \sum E_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \end{aligned} \quad (9.9)$$

Although matrix multiplication is not generally commutative, each product in Equation 9.9 is (1×1) ; thus, $\mathbf{y}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{y}$, justifying the transition to the last line of the equation.¹⁴

From the point of view of the coefficient vector \mathbf{b} , Equation 9.9 consists of a constant, a linear form in \mathbf{b} , and a quadratic form in \mathbf{b} . To minimize $S(\mathbf{b})$, we find its vector partial derivative with respect to \mathbf{b} :

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$$

Setting this derivative to $\mathbf{0}$ produces the matrix form of the normal equations for the linear model:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \quad (9.10)$$

There are $k + 1$ normal equations in the same number of unknown coefficients. If $\mathbf{X}'\mathbf{X}$ is nonsingular—that is, of rank $k + 1$ —then we can uniquely solve for the least-squares coefficients:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

¹²See Exercise 9.3.

¹³These conclusions are supported by the vector geometry of linear models, described in Chapter 10.

¹⁴See Exercise 9.4.

The rank of $\mathbf{X}'\mathbf{X}$ is equal to the rank of \mathbf{X} :

- Because the rank of \mathbf{X} can be no greater than the smaller of n and $k + 1$, for the least-squares coefficients to be unique, we require at least as many observations (n) as there are coefficients in the model ($k + 1$). This requirement is intuitively sensible: We cannot, for example, fit a unique line to a single data point, nor can we fit a unique plane to two data points. In most applications, n greatly exceeds $k + 1$.
- The $k + 1$ columns of \mathbf{X} must be linearly independent. This requirement implies that no regressor can be a perfect linear function of others and that only the constant regressor can be invariant.¹⁵

In applications, these requirements are usually met: $\mathbf{X}'\mathbf{X}$, therefore, is generally nonsingular, and the least-squares coefficients are uniquely defined.¹⁶

The second partial derivative of the sum of squared residuals is

$$\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}'} = 2\mathbf{X}'\mathbf{X}$$

Because $\mathbf{X}'\mathbf{X}$ is positive-definite when \mathbf{X} is of full rank, the solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ represents a *minimum* of $S(\mathbf{b})$.

If the model matrix \mathbf{X} is of full-column rank, then the least-squares coefficients are given by $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The matrix $\mathbf{X}'\mathbf{X}$ contains sums of squares and products among the regressors (including the constant regressor, $X_0 = 1$); the vector $\mathbf{X}'\mathbf{y}$ contains sums of cross products between the regressors and the response variable. Forming these matrix products and expressing the normal equations (Equation 9.10) in scalar format yields a familiar pattern:¹⁷

$$\begin{aligned} B_0 n &+ B_1 \sum x_{i1} &+ \cdots + B_k \sum x_{ik} &= \sum Y_i \\ B_0 \sum x_{i1} &+ B_1 \sum x_{i1}^2 &+ \cdots + B_k \sum x_{i1}x_{ik} &= \sum x_{i1}Y_i \\ \vdots && \vdots & \\ B_0 \sum x_{ik} &+ B_1 \sum x_{ik}x_{i1} &+ \cdots + B_k \sum x_{ik}^2 &= \sum x_{ik}Y_i \end{aligned}$$

To write an explicit solution to the normal equations in scalar form would be impractical, even for small values of k .

For Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations, the sums of squares and products are as follows:¹⁸

¹⁵If another regressor is invariant, then it is a multiple of the constant regressor, $X_0 = 1$.

¹⁶We will see in Section 13.1, however, that even when \mathbf{X} is of rank $k + 1$, *near-collinearity* of its columns can cause statistical difficulties.

¹⁷See Section 5.2.2.

¹⁸Cf. the scalar calculations for Duncan's regression, which appear in Section 5.2.1.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 45 & 1884 & 2365 \\ 1884 & 105,148 & 122,197 \\ 2365 & 122,197 & 163,265 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2146 \\ 118,229 \\ 147,936 \end{bmatrix}$$

The inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1021058996 & -0.0008495732 & -0.0008432006 \\ -0.0008495732 & 0.0000801220 & -0.0000476613 \\ -0.0008432006 & -0.0000476613 & 0.0000540118 \end{bmatrix}$$

and thus the least-squares regression coefficients are

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} -6.06466 \\ 0.59873 \\ 0.54583 \end{bmatrix}$$

9.2.1 Deficient-Rank Parameterization of Linear Models

As I explained, the unconstrained, overparametrized one-way ANOVA model shown in Equation 9.3 (page 204) has a rank-deficient model matrix \mathbf{X} : There are $m + 1$ columns in the model matrix but it is of rank m . In this text, I deal with overparametrized models by placing sufficient constraints on the parameters of the model to identify the parameters uniquely, incorporating these constraints in the coding of the model matrix. For example, placing a sigma constraint on the parameters of the one-way ANOVA model reduces the model matrix to full rank, as in Equation 9.4. Although the least-squares parameter estimates and the interpretation of the parameters depend on the constraints employed, many fundamental quantities, such as fitted values $\hat{\mathbf{y}}$, residuals \mathbf{e} , and consequently the regression and residual sums of squares for the fitted model, do not depend on which constraints are selected. These conclusions extend to other linear models, such as two-way ANOVA and analysis of covariance.¹⁹

An alternative to placing explicit constraints on the parameters is to form the normal equations using the rank-deficient model matrix \mathbf{X} ,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

The rank of $\mathbf{X}'\mathbf{X}$ is the same as the rank of \mathbf{X} ; the sum-of-squares and products matrix $\mathbf{X}'\mathbf{X}$ is therefore singular, and the normal equations do not have a unique solution. We can, however, select an arbitrary solution to the normal equations by employing a generalized inverse $(\mathbf{X}'\mathbf{X})^-$ of $\mathbf{X}'\mathbf{X}$.²⁰

$$\mathbf{b}^- = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$$

Because the generalized inverse of a singular matrix is not unique, the obtained value of \mathbf{b}^- depends on the specific generalized inverse employed. This approach is, therefore, equivalent to reducing \mathbf{X} to full rank by placing constraints on the parameters, with the constraints implicit in the selection of $(\mathbf{X}'\mathbf{X})^-$.

¹⁹See Section 10.4 on the vector geometry of ANOVA models.

²⁰Generalized inverses are described in online Appendix B on matrices, linear algebra, and vector geometry.

Direct application of overparametrized linear models is worth mentioning, if only briefly, because some statistical software (such as SAS) and some treatments of linear models (such as Searle, 1971) employ this approach.

9.3 Properties of the Least-Squares Estimator

In this section, I derive a number of fundamental results concerning the least-squares estimator \mathbf{b} of the linear-model parameter vector $\boldsymbol{\beta}$. These results serve several related purposes:

- They establish certain desirable properties of the least-squares estimator that hold under the assumptions of the linear model.
- They furnish a basis for using the least-squares coefficients to make statistical inferences about $\boldsymbol{\beta}$.²¹
- They provide a foundation for generalizing the linear model in several directions.²²

9.3.1 The Distribution of the Least-Squares Estimator

With the model matrix \mathbf{X} fixed, the least-squares coefficients \mathbf{b} result from a linear transformation of the response variable; that is, \mathbf{b} is a *linear estimator*:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}$$

defining $\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The expected value of \mathbf{b} is easily established from the expectation of \mathbf{y} (given previously in Equation 9.2 on page 203):

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

The least-squares estimator \mathbf{b} is therefore an unbiased estimator of $\boldsymbol{\beta}$.

The covariance matrix of the least-squares estimator is similarly derived:

$$V(\mathbf{b}) = \mathbf{M}V(\mathbf{y})\mathbf{M}' = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma_{\varepsilon}^2\mathbf{I}_n[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]'$$

Moving the scalar error variance σ_{ε}^2 to the front of this expression, and noting that $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of a symmetric matrix and is thus itself symmetric, we get

$$V(\mathbf{b}) = \sigma_{\varepsilon}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_{\varepsilon}^2(\mathbf{X}'\mathbf{X})^{-1}$$

The sampling variances and covariances of the regression coefficients, therefore, depend only on the model matrix and the variance of the errors.

To derive $E(\mathbf{b})$ and $V(\mathbf{b})$, we do not require the assumption of normality—only the assumptions of linearity [i.e., $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$], constant variance, and independence [$V(\mathbf{y}) = \sigma_{\varepsilon}^2\mathbf{I}_n$.]²³ If \mathbf{y} is normally distributed, however, then so is \mathbf{b} , for—as I have explained— \mathbf{b} results from a linear transformation of \mathbf{y} :

²¹See Section 9.4.

²²See, for example, Section 12.2.2 and Chapters 14, 15, and 16.

²³In general, independence of the observations *implies* that the elements of \mathbf{y} are uncorrelated [i.e., that $V(\mathbf{y})$ is diagonal], but the reverse is not the case: The elements of \mathbf{y} *could be* uncorrelated even if the observations are not independent. That is, independence is a stronger condition than uncorrelation. For normally distributed \mathbf{y} , however, uncorrelation and independence coincide.

Table 9.1 Comparison Between Simple Regression Using Scalars and Multiple Regression Using Matrices

	<i>Simple Regression</i>	<i>Multiple Regression</i>
Model	$Y = \alpha + \beta x + \varepsilon$	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$
Least-squares estimator	$B = \frac{\sum x^* Y^*}{\sum x^{*2}}$ $= (\sum x^{*2})^{-1} \sum x^* Y^*$	$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Sampling variance	$V(B) = \frac{\sigma_\varepsilon^2}{\sum x^{*2}}$ $= \sigma_\varepsilon^2 (\sum x^{*2})^{-1}$	$\mathbf{V}(\mathbf{b}) = \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$
Distribution	$B \sim N[\beta, \sigma_\varepsilon^2 (\sum x^{*2})^{-1}]$	$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}]$

NOTE: Subscripts are suppressed in this table; in particular, $x^* \equiv x_i - \bar{x}$ and $Y^* \equiv Y_i - \bar{Y}$.

SOURCE: Adapted from Wonnacott and Wonnacott (1979, Table 12-1), *Econometrics, Second Edition*. Copyright © John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

$$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}]$$

There is a striking parallel, noted by Wonnacott and Wonnacott (1979) and detailed in Table 9.1, between the scalar formulas for least-squares simple regression and the matrix formulas for the general linear model (“multiple regression”). This sort of structural parallel is common in statistical applications when matrix methods are used to generalize a scalar result: Matrix notation is productive precisely because of the generality and simplicity that it achieves.

Under the full set of assumptions for the linear model, the distribution of the least-squares regression coefficients is

$$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}]$$

9.3.2 The Gauss-Markov Theorem

One of the primary theoretical justifications for least-squares estimation is the Gauss-Markov theorem, which states that if the errors are independently distributed with zero expectation and constant variance, then the least-squares estimator \mathbf{b} is the most efficient linear unbiased estimator of $\boldsymbol{\beta}$. That is, of all unbiased estimators that are linear functions of the observations, the least-squares estimator has the smallest sampling variance and, hence, the smallest mean-squared error. For this reason, the least-squares estimator is sometimes termed *BLUE*, an acronym for *best linear unbiased estimator*.²⁴

²⁴As I explained in Section 6.1.2, the comfort provided by the Gauss-Markov theorem is often an illusion, because the restriction to linear estimators is artificial. Under the additional assumption of normality, however, it is possible to show that the least-squares estimator is maximally efficient among *all* unbiased estimators (see, e.g., Rao, 1973, p. 319). The strategy of proof of the Gauss-Markov theorem employed in this section is borrowed from Wonnacott and Wonnacott (1979, pp. 428–430), where it is used in a slightly different context.

Let $\tilde{\mathbf{b}}$ represent the best linear unbiased estimator of β . As we know, the least-squares estimator \mathbf{b} is also a linear estimator, $\mathbf{b} = \mathbf{My}$. It is convenient to write $\tilde{\mathbf{b}} = (\mathbf{M} + \mathbf{A})\mathbf{y}$, where \mathbf{A} gives the *difference* between the (as yet undetermined) transformation matrix for the BLUE and that for the least-squares estimator. To show that the BLUE and the least-squares estimator coincide—that is, to establish the Gauss-Markov theorem—we need to demonstrate that $\mathbf{A} = \mathbf{0}$.

Because $\tilde{\mathbf{b}}$ is *unbiased*,

$$\begin{aligned}\beta &= E(\tilde{\mathbf{b}}) = E[(\mathbf{M} + \mathbf{A})\mathbf{y}] = E(\mathbf{My}) + E(\mathbf{Ay}) \\ &= E(\mathbf{b}) + \mathbf{AE}(\mathbf{y}) = \beta + \mathbf{AX}\beta\end{aligned}$$

The matrix product $\mathbf{AX}\beta$, then, is $\mathbf{0}$, regardless of the value of β , and therefore \mathbf{AX} must be $\mathbf{0}$.²⁵

I have, to this point, made use of the linearity and unbias of $\tilde{\mathbf{b}}$. Because $\tilde{\mathbf{b}}$ is the *minimum-variance* linear unbiased estimator, the sampling variances of its elements—that is, the diagonal entries of $V(\tilde{\mathbf{b}})$ —are as small as possible.²⁶ The covariance matrix of $\tilde{\mathbf{b}}$ is given by

$$\begin{aligned}V(\tilde{\mathbf{b}}) &= (\mathbf{M} + \mathbf{A})V(\mathbf{y})(\mathbf{M} + \mathbf{A})' \\ &= (\mathbf{M} + \mathbf{A})\sigma_e^2 \mathbf{I}_n (\mathbf{M} + \mathbf{A})' \\ &= \sigma_e^2 (\mathbf{MM}' + \mathbf{MA}' + \mathbf{AM}' + \mathbf{AA}')\end{aligned}\tag{9.11}$$

I have shown that $\mathbf{AX} = \mathbf{0}$; consequently, \mathbf{AM}' and its transpose \mathbf{MA}' are $\mathbf{0}$, for

$$\mathbf{AM}' = \mathbf{AX}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$$

Equation 9.11 becomes

$$V(\tilde{\mathbf{b}}) = \sigma_e^2 (\mathbf{MM}' + \mathbf{AA}')$$

The sampling variance of the coefficient \tilde{B}_j is the j th diagonal entry of $V(\tilde{\mathbf{b}})$.²⁷

$$V(\tilde{B}_j) = \sigma_e^2 \left(\sum_{i=1}^n m_{ji}^2 + \sum_{i=1}^n a_{ji}^2 \right)$$

Both sums in this equation are sums of squares and hence cannot be negative; because $V(\tilde{B}_j)$ is as small as possible, all the a_{ji} must be 0. This argument applies to each coefficient in $\tilde{\mathbf{b}}$, and so every row of \mathbf{A} must be $\mathbf{0}$, implying that $\mathbf{A} = \mathbf{0}$. Finally,

$$\tilde{\mathbf{b}} = (\mathbf{M} + \mathbf{0})\mathbf{y} = \mathbf{My} = \mathbf{b}$$

demonstrating that the BLUE is the least-squares estimator.

²⁵See Exercise 9.7.

²⁶It is possible to prove a more general result: The best linear unbiased estimator of $\mathbf{a}'\beta$ (an arbitrary linear combination of regression coefficients) is $\mathbf{a}'\mathbf{b}$, where \mathbf{b} is the least-squares estimator (see, e.g., Seber, 1977, p. 49).

²⁷Actually, the variance of the constant \tilde{B}_0 is the *first* diagonal entry of $V(\tilde{\mathbf{b}})$; the variance of \tilde{B}_j is therefore the $(j+1)$ st entry. To avoid this awkwardness, I will index the covariance matrix of $\tilde{\mathbf{b}}$ (and later, that of \mathbf{b}) from 0 rather than from 1.

9.3.3 Maximum-Likelihood Estimation

Under the assumptions of the linear model, the least-squares estimator \mathbf{b} is also the maximum-likelihood estimator of β .²⁸ This result establishes an additional justification for least squares when the assumptions of the model are reasonable, but even more important, it provides a basis for generalizing the linear model.²⁹

As I have explained, under the assumptions of the linear model, $\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma_\varepsilon^2 \mathbf{I}_n)$. Thus, for the i th observation, $Y_i \sim N(\mathbf{x}'_i \beta, \sigma_\varepsilon^2)$, where \mathbf{x}'_i is the i th row of the model matrix \mathbf{X} . In equation form, the probability density for observation i is

$$p(y_i) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp \left[-\frac{(y_i - \mathbf{x}'_i \beta)^2}{2\sigma_\varepsilon^2} \right]$$

Because the n observations are independent, their joint probability density is the product of their marginal densities:

$$\begin{aligned} p(\mathbf{y}) &= \frac{1}{(\sigma_\varepsilon \sqrt{2\pi})^n} \exp \left[-\frac{\sum (y_i - \mathbf{x}'_i \beta)^2}{2\sigma_\varepsilon^2} \right] \\ &= \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma_\varepsilon^2} \right] \end{aligned} \quad (9.12)$$

Although this equation also follows directly from the multivariate-normal distribution of \mathbf{y} , the development from $p(y_i)$ to $p(\mathbf{y})$ will prove helpful when we consider random regressors.³⁰

From Equation 9.12, the log-likelihood is

$$\log_e L(\beta, \sigma_\varepsilon^2) = -\frac{n}{2} \log_e 2\pi - \frac{n}{2} \log_e \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \quad (9.13)$$

To maximize the likelihood, we require the partial derivatives of Equation 9.13 with respect to the parameters β and σ_ε^2 . Differentiation is simplified when we notice that $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ is the sum of squared errors:

$$\begin{aligned} \frac{\partial \log_e L(\beta, \sigma_\varepsilon^2)}{\partial \beta} &= -\frac{1}{2\sigma_\varepsilon^2} (2\mathbf{X}'\mathbf{X}\beta - 2\mathbf{X}'\mathbf{y}) \\ \frac{\partial \log_e L(\beta, \sigma_\varepsilon^2)}{\partial \sigma_\varepsilon^2} &= -\frac{n}{2} \left(\frac{1}{\sigma_\varepsilon^2} \right) + \frac{1}{2\sigma_\varepsilon^4} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

Setting these partial derivatives to 0 and solving for the maximum-likelihood estimators $\hat{\beta}$ and $\hat{\sigma}_\varepsilon^2$ produces

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \\ \hat{\sigma}_\varepsilon^2 &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n} \end{aligned}$$

The maximum-likelihood estimator $\hat{\beta}$ is therefore the same as the least-squares estimator \mathbf{b} . In fact, this identity is clear directly from Equation 9.12, without formal maximization of the

²⁸The method of maximum likelihood is introduced in online Appendix D on probability and estimation.

²⁹See, for example, the discussions of transformations in Section 12.5 and of nonlinear least squares in Chapter 17.

³⁰See Section 9.6.

likelihood: The likelihood is large when the negative exponent is small, and the numerator of the exponent contains the sum of squared errors; minimizing the sum of squared residuals, therefore, maximizes the likelihood.

The maximum-likelihood estimator $\hat{\sigma}_\varepsilon^2$ of the error variance is biased; consequently, we prefer the similar, unbiased estimator $S_E^2 = \mathbf{e}'\mathbf{e}/(n - k - 1)$ to $\hat{\sigma}_\varepsilon^2$.³¹ As n increases, however, the bias of $\hat{\sigma}_\varepsilon^2$ shrinks toward 0: As a maximum-likelihood estimator, $\hat{\sigma}_\varepsilon^2$ is consistent.

9.4 Statistical Inference for Linear Models

The results of the previous section, along with some to be established in Chapter 10, provide a basis for statistical inference in linear models.³² I have already shown that the least-squares coefficients \mathbf{b} have certain desirable properties as point estimators of the parameters β . In this section, I will describe tests for individual coefficients, for several coefficients, and for general linear hypotheses.

9.4.1 Inference for Individual Coefficients

We saw that the least-squares estimator \mathbf{b} follows a normal distribution with expectation β and covariance matrix $\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$.³³ Consequently, an individual coefficient B_j is normally distributed with expectation β_j and sampling variance $\sigma_\varepsilon^2 v_{jj}$, where v_{jj} is the j th diagonal entry of $(\mathbf{X}'\mathbf{X})^{-1}$.³⁴ The ratio $(B_j - \beta_j)/\sigma_\varepsilon\sqrt{v_{jj}}$, therefore, follows the unit-normal distribution $N(0, 1)$, and to test the hypothesis $H_0: \beta_j = \beta_j^{(0)}$, we can calculate the test statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\sigma_\varepsilon\sqrt{v_{jj}}}$$

comparing the obtained value of the statistic to quantiles of the unit-normal distribution. This result is not of direct practical use, however, because in applications of linear models, we do not know σ_ε^2 .

Although the error variance is unknown, we have available the unbiased estimator $S_E^2 = \mathbf{e}'\mathbf{e}/(n - k - 1)$. Employing this estimator, we can estimate the covariance matrix of the least-squares coefficients:

$$\widehat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n - k - 1}(\mathbf{X}'\mathbf{X})^{-1}$$

The standard error of the coefficient B_j is, therefore, given by $SE(B_j) = S_E\sqrt{v_{jj}}$, the square root of the j th diagonal entry of $\widehat{V}(\mathbf{b})$.

It can be shown that $(n - k - 1)S_E^2/\sigma_\varepsilon^2 = \mathbf{e}'\mathbf{e}/\sigma_\varepsilon^2$ follows a chi-square distribution with $n - k - 1$ degrees of freedom.³⁵ We recently discovered that $(B_j - \beta_j)/\sigma_\varepsilon\sqrt{v_{jj}}$ is distributed as

³¹See Section 10.3 for a derivation of the expectation of S_E^2 .

³²The results of this section justify and extend the procedures for inference described in Chapter 6.

³³See Section 9.3.1.

³⁴Recall that we index the rows and columns of $(\mathbf{X}'\mathbf{X})^{-1}$ from 0 through k .

³⁵See Section 10.3.

$N(0, 1)$. It can be further established that the estimators B_j and S_E^2 are independent,³⁶ and so the ratio

$$t = \frac{(B_j - \beta_j)/\sigma_\varepsilon \sqrt{v_{jj}}}{\sqrt{\frac{\mathbf{e}'\mathbf{e}/\sigma_\varepsilon^2}{n-k-1}}} = \frac{B_j - \beta_j}{S_E \sqrt{v_{jj}}}$$

follows a t -distribution with $n - k - 1$ degrees of freedom. Heuristically, in estimating σ_ε with S_E , we must replace the normal distribution with the more spread-out t -distribution to reflect the additional source of variability.

To test the hypothesis $H_0: \beta_j = \beta_j^{(0)}$, therefore, we calculate the test statistic

$$t_0 = \frac{B_j - \beta_j^{(0)}}{\text{SE}(B_j)}$$

comparing the obtained value of t_0 with the quantiles of t_{n-k-1} . Likewise, a $100(1 - a)\%$ confidence interval for β_j is given by

$$\beta_j = B_j \pm t_{a/2, n-k-1} \text{SE}(B_j)$$

where $t_{a/2, n-k-1}$ is the critical value of t_{n-k-1} with a probability of $a/2$ to the right.

For Duncan's occupational prestige regression, for example, the estimated error variance is $S_E^2 = 178.73$, and so the estimated covariance matrix of the regression coefficients is

$$\begin{aligned}\widehat{V}(\mathbf{b}) &= 178.73(\mathbf{X}'\mathbf{X})^{-1} \\ &= \begin{bmatrix} 18.249387 & -0.151844 & -0.150705 \\ -0.151844 & 0.014320 & -0.008519 \\ -0.150705 & -0.008519 & 0.009653 \end{bmatrix}\end{aligned}$$

The standard errors of the regression coefficients are³⁷

$$\text{SE}(B_0) = \sqrt{18.249387} = 4.272$$

$$\text{SE}(B_1) = \sqrt{0.014320} = 0.1197$$

$$\text{SE}(B_2) = \sqrt{0.009653} = 0.09825$$

The estimated covariance matrix of the least-squares coefficients is $\widehat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1}$. The standard errors of the regression coefficients are the square-root diagonal entries of this matrix. Under the assumptions of the model, $(B_j - \beta_j)/\text{SE}(B_j) \sim t_{n-k-1}$, providing a basis for hypothesis tests and confidence intervals for individual coefficients.

9.4.2 Inference for Several Coefficients

Although we often test regression coefficients individually, these tests may not be sufficient, for, in general, the least-squares estimators of different parameters are correlated: The

³⁶See Exercise 9.8.

³⁷Compare with the results given in Section 6.1.3.

off-diagonal entries of $V(\mathbf{b}) = \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}$, giving the sampling covariances of the least-squares coefficients, are 0 only when the regressors themselves are uncorrelated.³⁸ Furthermore, in certain applications of linear models—such as dummy regression, analysis of variance, and polynomial regression—we are more interested in related sets of coefficients than in the individual members of these sets.

Simultaneous tests for sets of coefficients, taking their intercorrelations into account, can be constructed by the likelihood-ratio principle. Suppose that we fit the model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \quad (9.14)$$

obtaining the least-squares estimate $\mathbf{b} = [B_0, B_1, \dots, B_k]',$ along with the maximum-likelihood estimate of the error variance, $\hat{\sigma}_e^2 = \mathbf{e}'\mathbf{e}/n.$ We wish to test the null hypothesis that a subset of regression parameters is 0; for convenience, let these coefficients be the first $q \leq k,$ so that the null hypothesis is $H_0: \beta_1 = \cdots = \beta_q = 0.$ The null hypothesis corresponds to the model

$$\begin{aligned} Y &= \beta_0 + 0x_1 + \cdots + 0x_q + \beta_{q+1}x_{q+1} + \cdots + \beta_k x_k + \varepsilon \\ &= \beta_0 + \beta_{q+1}x_{q+1} + \cdots + \beta_k x_k + \varepsilon \end{aligned} \quad (9.15)$$

which is a specialization (or restriction) of the more general model (9.14). Fitting the restricted model (9.15) by least-squares regression of Y on x_{q+1} through $x_k,$ we obtain $\mathbf{b}_0 = [B'_0, 0, \dots, 0, B'_{q+1}, \dots, B'_k]',$ and $\hat{\sigma}_{\varepsilon_0}^2 = \mathbf{e}'_0\mathbf{e}_0/n.$ The coefficients in \mathbf{b}_0 generally differ from those in \mathbf{b} (hence the primes), and $\hat{\sigma}_e^2 \leq \hat{\sigma}_{\varepsilon_0}^2,$ because both models are fit by least squares.

The likelihood for the full model (9.14), evaluated at the maximum-likelihood estimates, can be obtained from Equation 9.12 (on page 214):³⁹

$$L = \left(2\pi e \frac{\mathbf{e}'\mathbf{e}}{n} \right)^{-n/2}$$

Likewise, for the restricted model (9.15), the maximized likelihood is

$$L_0 = \left(2\pi e \frac{\mathbf{e}'_0\mathbf{e}_0}{n} \right)^{-n/2}$$

The likelihood ratio for testing H_0 is, therefore,

$$\frac{L_0}{L_1} = \left(\frac{\mathbf{e}'_0\mathbf{e}_0}{\mathbf{e}'\mathbf{e}} \right)^{-n/2} = \left(\frac{\mathbf{e}'\mathbf{e}}{\mathbf{e}'_0\mathbf{e}_0} \right)^{2/n}$$

Because $\mathbf{e}'_0\mathbf{e}_0 \geq \mathbf{e}'\mathbf{e},$ the likelihood ratio is small when the residual sum of squares for the restricted model is appreciably larger than for the general model—circumstances under which we should doubt the truth of the null hypothesis. A test of H_0 is provided by the generalized

³⁸This point pertains to sampling correlations among the k slope coefficients. The regression constant is correlated with the slope coefficients unless all the regressors—save the constant regressor—have means of 0 (i.e., are in mean deviation form). Expressing the regressors in mean deviation form, called *centering*, has certain computational advantages (it tends to reduce rounding errors in least-squares calculations), but it does not affect the slope coefficients or the sampling covariances among them.

³⁹See Exercise 9.9. The notation here is potentially confusing: $e \approx 2.718$ is the mathematical constant; \mathbf{e} is the vector of residuals.

likelihood-ratio test statistic, $G_0^2 = -2 \log_e(L_0/L_1)$, which is asymptotically distributed as χ_q^2 under the null hypothesis.

It is unnecessary to use this asymptotic result, however, for an exact test can be obtained:⁴⁰ As mentioned in the previous section, $\text{RSS}/\sigma_\varepsilon^2 = \mathbf{e}'\mathbf{e}/\sigma_\varepsilon^2$ is distributed as χ^2 with $n - k - 1$ degrees of freedom. By a direct extension of this result, if the null hypothesis is true, then $\text{RSS}_0/\sigma_\varepsilon^2 = \mathbf{e}'_0\mathbf{e}_0/\sigma_\varepsilon^2$ is distributed as χ^2 with $n - (k - q) - 1 = n - k + q - 1$ degrees of freedom. Consequently, the difference $(\text{RSS}_0 - \text{RSS})/\sigma_\varepsilon^2$ has a χ^2 distribution with $(n - k + q - 1) - (n - k - 1) = q$ degrees of freedom, equal to the number of parameters set to 0 in the restricted model. It can be shown that $(\text{RSS}_0 - \text{RSS})/\sigma_\varepsilon^2$ and $\text{RSS}/\sigma_\varepsilon^2$ are independent, and so the ratio

$$F_0 = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n - k - 1)}$$

is distributed as F with q and $n - k - 1$ degrees of freedom. This is, of course, the incremental F -statistic.⁴¹

Although it is sometimes convenient to find an incremental sum of squares by fitting alternative linear models to the data, it is also possible to calculate this quantity directly from the least-squares coefficient vector \mathbf{b} and the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ for the full model: Let $\mathbf{b}_1 = [B_1, \dots, B_q]'$ represent the coefficients of interest selected from among the entries of \mathbf{b} ; and let \mathbf{V}_{11} represent the square submatrix consisting of the entries in the q rows and q columns of $(\mathbf{X}'\mathbf{X})^{-1}$ that pertain to the coefficients in \mathbf{b}_1 .⁴² Then it can be shown that the incremental sum of squares $\text{RSS}_0 - \text{RSS}$ is equal to $\mathbf{b}'_1\mathbf{V}_{11}^{-1}\mathbf{b}_1$, and thus the incremental F -statistic can be written $F_0 = \mathbf{b}'_1\mathbf{V}_{11}^{-1}\mathbf{b}_1/qS_E^2$. To test the more general hypothesis $H_0: \beta_1 = \beta_1^{(0)}$ (where $\beta_1^{(0)}$ is not necessarily $\mathbf{0}$), we can compute

$$F_0 = \frac{(\mathbf{b}_1 - \beta_1^{(0)})'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \beta_1^{(0)})}{qS_E^2} \quad (9.16)$$

which is distributed as $F_{q,n-k-1}$ under H_0 .

Recall that the omnibus F -statistic for the hypothesis $H_0: \beta_1 = \dots = \beta_k = 0$ is

$$F_0 = \frac{\text{RegSS}/k}{\text{RSS}/(n - k - 1)}$$

The denominator of this F -statistic estimates the error variance σ_ε^2 , whether or not the null hypothesis is true.⁴³ The expectation of the regression sum of squares, it may be shown,⁴⁴ is

$$E(\text{RegSS}) = \beta_1'(\mathbf{X}^{*\prime}\mathbf{X}^*)\beta_1 + k\sigma_\varepsilon^2$$

where $\beta_1 \equiv [\beta_1, \dots, \beta_k]'$ is the vector of regression coefficients, excluding the constant, and $\mathbf{X}_{(n \times k)}^* \equiv \{x_{ij} - \bar{x}_j\}$ is the matrix of mean deviation regressors, omitting the constant regressor.

⁴⁰The F -test that follows is exact when the assumptions of the model hold—including the assumption of normality. Of course, the asymptotically valid likelihood-ratio chi-square test also depends on these assumptions.

⁴¹See Section 6.1.3.

⁴²Note the difference between the vector \mathbf{b}_1 (used here) and the vector \mathbf{b}_0 (used previously): \mathbf{b}_1 consists of coefficients extracted from \mathbf{b} , which, in turn, results from fitting the *full* model; in contrast, \mathbf{b}_0 consists of the coefficients—including those set to 0 in the hypothesis—that result from fitting the *restricted* model.

⁴³See Section 10.3.

⁴⁴See Seber (1977, Chapter 4).

When H_0 is true (and $\beta_1 = \mathbf{0}$), the numerator of the F -statistic (as well as its denominator) estimates σ_ε^2 , but when H_0 is false, $E(\text{RegSS}/k) > \sigma_\varepsilon^2$, because $\mathbf{X}^{*\prime}\mathbf{X}^*$ is positive definite, and thus $\beta_1'(\mathbf{X}^{*\prime}\mathbf{X}^*)\beta_1 > 0$ for $\beta_1 \neq \mathbf{0}$. Under these circumstances, we tend to observe numerators that are larger than denominators and F -statistics that are greater than 1.⁴⁵

An incremental F -test for the hypothesis $H_0: \beta_1 = \dots = \beta_q = 0$, where $1 \leq q \leq k$, is given by $F_0 = (n - k - 1)(\text{RSS}_0 - \text{RSS})/q \text{ RSS}$, where RSS is the residual sum of squares for the full model, and RSS_0 is the residual sum of squares for the model that deletes the q regressors corresponding to the parameters in H_0 . Under the null hypothesis, $F_0 \sim F_{q,n-k-1}$. The incremental F -statistic can also be computed directly as $F_0 = \mathbf{b}'_1 \mathbf{V}_{11}^{-1} \mathbf{b}_1 / q S_E^2$, where $\mathbf{b}_1 = [B_1, \dots, B_q]'$ contains the coefficients of interest extracted from among the entries of \mathbf{b} , and \mathbf{V}_{11} is the square submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$ consisting of the q rows and columns pertaining to the coefficients in \mathbf{b}_1 .

9.4.3 General Linear Hypotheses

Even more generally, we can test the linear hypothesis

$$H_0: \underset{(q \times k+1)}{\mathbf{L}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} = \underset{(q \times 1)}{\mathbf{c}}$$

where \mathbf{L} and \mathbf{c} contain prespecified constants, and the *hypothesis matrix* \mathbf{L} is of full row rank $q \leq k + 1$. The resulting F -statistic,

$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})}{q S_E^2} \quad (9.17)$$

follows an F -distribution with q and $n - k - 1$ degrees of freedom if H_0 is true.

To understand the structure of Equation 9.17, recall that $\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}]$. As a consequence,

$$\mathbf{L}\mathbf{b} \sim N_q[\mathbf{L}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{L}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{L}']$$

Under H_0 , $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, and thus

$$(\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c}) / \sigma_\varepsilon^2 \sim \chi_q^2$$

Equation 9.17 is general enough to encompass all the hypothesis tests that we have considered thus far, along with others. In Duncan's occupational prestige regression, for example, to test the omnibus null hypothesis $H_0: \beta_1 = \beta_2 = 0$, we can take

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

⁴⁵The expectation of F_0 is not precisely 1 when H_0 is true because the expectation of a ratio of random variables is not necessarily the ratio of their expectations. See online Appendix D on probability and estimation.

and $\mathbf{c} = [0, 0]'$. To test the hypothesis that the education and income coefficients are equal, $H_0: \beta_1 = \beta_2$, which is equivalent to $H_0: \beta_1 - \beta_2 = 0$, we can take $\mathbf{L} = [0, 1, -1]$ and $\mathbf{c} = [0]$.⁴⁶

The F -statistic $F_0 = (\mathbf{L}\mathbf{b} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\mathbf{b} - \mathbf{c})/qS_E^2$ is used to test the general linear hypothesis $H_0: \underset{(q \times k+1)}{\mathbf{L}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} = \underset{(q \times 1)}{\mathbf{c}}$, where the rank- q hypothesis matrix \mathbf{L} and right-hand-side vector \mathbf{c} contain prespecified constants. Under the hypothesis, $F_0 \sim F_{q,n-k-1}$.

9.4.4 Joint Confidence Regions

The F -test of Equation 9.16 (on page 218) can be inverted to construct a *joint confidence region* for β_1 . If $H_0: \beta_1 = \beta_1^{(0)}$ is correct, then

$$\Pr \left[\frac{(\mathbf{b}_1 - \beta_1^{(0)})' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1^{(0)})}{qS_E^2} \leq F_{a,q,n-k-1} \right] = 1 - a$$

where $F_{a,q,n-k-1}$ is the critical value of F with q and $n - k - 1$ degrees of freedom, corresponding to a right-tail probability of a . The joint confidence region for β_1 is thus

$$\text{all } \beta_1 \text{ for which } (\mathbf{b}_1 - \beta_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1) \leq qS_E^2 F_{a,q,n-k-1} \quad (9.18)$$

That is, any parameter vector β_1 that satisfies this inequality is *within* the confidence region and is acceptable as a hypothesis; any parameter vector that does not satisfy the inequality is unacceptable. The boundary of the joint confidence region (obtained when the left-hand side of the inequality in Equation 9.18 equals the right-hand side) is an ellipsoid centered at the estimates \mathbf{b}_1 in the q -dimensional space of the parameters β_1 .

Like a confidence interval, a joint confidence region is a portion of the parameter space constructed so that, with repeated sampling, a preselected percentage of regions will contain the true parameter values. Unlike a confidence interval, however, which pertains to a *single* coefficient β_j , a joint confidence region encompasses all *combinations* of values for the parameters β_1, \dots, β_q that are *simultaneously* acceptable at the specified level of confidence. The familiar confidence interval is just a one-dimensional confidence region, and there is a simple

⁴⁶Examples of these calculations appear in Exercise 9.10. The hypothesis that two regression coefficients are equal is sensible only if the corresponding explanatory variables are measured on the same scale. This is arguably the case for income and education in Duncan's regression, because both explanatory variables are percentages. Closer scrutiny suggests, however, that these explanatory variables are *not* commensurable: There is no reason to suppose that the percentage of occupational incumbents with at least a high school education is on the same scale as the percentage earning in excess of \$3500.

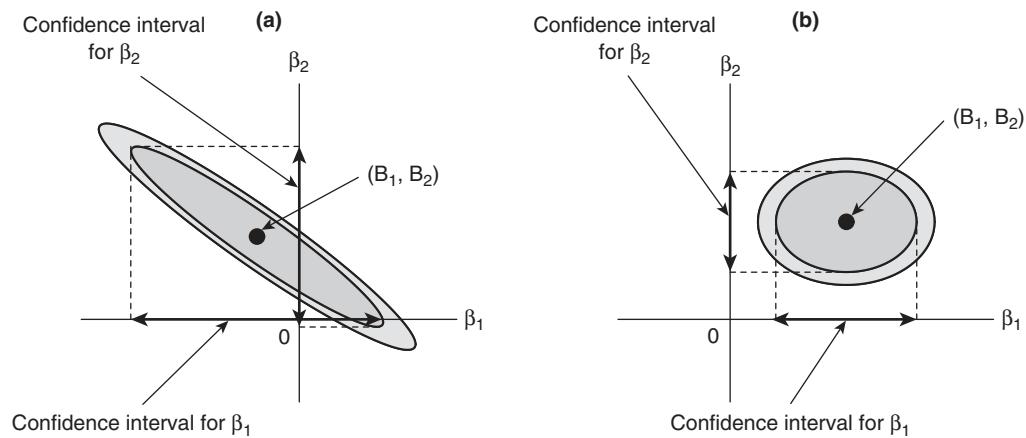


Figure 9.1 Illustrative joint confidence ellipses for the slope coefficients β_1 and β_2 in multiple-regression analysis. The outer ellipse is drawn at a level of confidence of 95%; the inner ellipse (the confidence interval–generating ellipse) is drawn so that its perpendicular shadows on the axes are 95% confidence intervals for the individual β s. In (a), the Xs are positively correlated, producing a joint confidence ellipse that is negatively tilted. In (b), the Xs are uncorrelated, producing a joint confidence ellipse with axes parallel to the axes of the parameter space.

relationship between the confidence interval for a single coefficient and the confidence region for several coefficients (as I will explain shortly).

The essential nature of joint confidence regions is clarified by considering the two-dimensional case, which can be directly visualized. To keep the mathematics as simple as possible, let us work with the slope coefficients β_1 and β_2 from the two-explanatory-variable model, $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$. In this instance, the joint confidence region of Equation 9.18 becomes all (β_1, β_2) for which

$$[B_1 - \beta_1, B_2 - \beta_2] \begin{bmatrix} \sum x_{i1}^{*2} & \sum x_{i1}^* x_{i2}^* \\ \sum x_{i1}^* x_{i2}^* & \sum x_{i2}^{*2} \end{bmatrix} \begin{bmatrix} B_1 - \beta_1 \\ B_2 - \beta_2 \end{bmatrix} \leq 2S_E^2 F_{a, 2, n-3} \quad (9.19)$$

where the $x_{ij}^* \equiv x_{ij} - \bar{x}_j$ are deviations from the means of X_1 and X_2 , and the matrix \mathbf{V}_{11}^{-1} contains mean deviation sums of squares and products for the explanatory variables.⁴⁷ The boundary of the confidence region, obtained when the equality holds, is an ellipse centered at (B_1, B_2) in the $\{\beta_1, \beta_2\}$ plane.

Illustrative joint confidence ellipses are shown in Figure 9.1. When the explanatory variables are *uncorrelated*, the sum of cross-products $\sum x_{i1}^* x_{i2}^*$ vanishes, and the axes of the confidence ellipse are parallel to the axes of the parameter space, as in Figure 9.1(b). When the explanatory variables are *correlated*, in contrast, the ellipse is “tilted,” as in Figure 9.1(a).

Specializing (9.18) to a single coefficient produces the confidence interval for β_1 :

⁴⁷See Exercise 9.11.

$$\text{all } \beta_1 \text{ for which } (B_1 - \beta_1)^2 \frac{\sum x_{i2}^{*2}}{\sum x_{i1}^{*2} \sum x_{i2}^{*2} - (\sum x_{i1}^* x_{i2}^*)^2} \leq S_E^2 F_{a,1,n-3} \quad (9.20)$$

which is written more conventionally as⁴⁸

$$B_1 - t_{a,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}} \leq \beta_1 \leq B_1 + t_{a,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}}$$

The individual confidence intervals for the regression coefficients are very nearly the perpendicular “shadows” (i.e., projections) of the joint confidence ellipse onto the β_1 and β_2 axes. The only slippage here is due to the right-hand-side constant: $2S_E^2 F_{a,2,n-3}$ for the joint confidence region and $S_E^2 F_{a,1,n-3}$ for the confidence interval.

Consider a 95% region and interval, for example. If the residual degrees of freedom $n - 3$ are large, then $2F_{.05,2,n-3} \approx \chi_{.05,2}^2 = 5.99$, while $F_{.05,1,n-3} \approx \chi_{.05,1}^2 = 3.84$. Put another way, using $5.99S_E^2$ in place of $3.84S_E^2$ produces individual intervals at approximately the $1 - Pr(\chi_1^2 > 5.99) = .986$ (rather than .95) level of confidence (but a *joint* 95% confidence region). Likewise, if we construct the joint confidence region using the multiplier 3.84, the resulting smaller ellipse produces shadows that give approximate 95% confidence intervals for *individual* coefficients [and a smaller *joint* level of confidence of $1 - Pr(\chi_2^2 > 3.84) = .853$]. This *confidence interval–generating ellipse* is shown along with the joint confidence ellipse in Figure 9.1.⁴⁹

Figure 9.1(a) illustrates how correlated regressors can lead to ambiguous inferences: Because the individual confidence intervals include 0, we cannot reject the separate hypotheses that *either* β_1 or β_2 is 0. Because the point (0, 0) is *outside* of the joint confidence region, however, we can reject the hypothesis that *both* β_1 and β_2 are 0. In contrast, in Figure 9.1(b), where the explanatory variables are uncorrelated, there is a close correspondence between inferences based on the separate confidence intervals and those based on the joint confidence region.

Still more generally, the confidence interval–generating ellipse can be projected onto *any* line through the origin of the $\{\beta_1, \beta_2\}$ plane. Each such line represents a specific linear combination of β_1 and β_2 , and the shadow of the ellipse on the line gives the corresponding confidence interval for that linear combination of the parameters.⁵⁰ This property is illustrated in Figure 9.2 for the linear combination $\beta_1 + \beta_2$; the line representing $\beta_1 + \beta_2$ is drawn through the origin and the point (1, 1), the coefficients of the parameters in the linear combination. Directions in which the ellipse is narrow, therefore, correspond to linear combinations of the parameters that are relatively precisely estimated.

It is illuminating to examine more closely the relationship between the joint confidence region for the regression coefficients and the joint distribution of the X -values. I have already remarked that the orientation of the confidence region reflects the correlation of the X s, but it is possible to be much more precise. Consider the quadratic form $(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}_{XX}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$, where \mathbf{x}

⁴⁸See Exercise 9.12.

⁴⁹The individual intervals constructed from the larger joint confidence ellipse—called *Scheffé intervals*, after the statistician Henry Scheffé (1959)—can be thought of as incorporating a penalty for examining several coefficients simultaneously. The difference between the Scheffé interval—the shadow of the joint confidence region (for which the multiplier is $kS_E^2 F_{a,k,n-3}$)—and the individual confidence interval (for which the multiplier is $S_E^2 F_{a,1,n-3}$) grows larger as the number of coefficients k increases.

⁵⁰See Monette (1990).

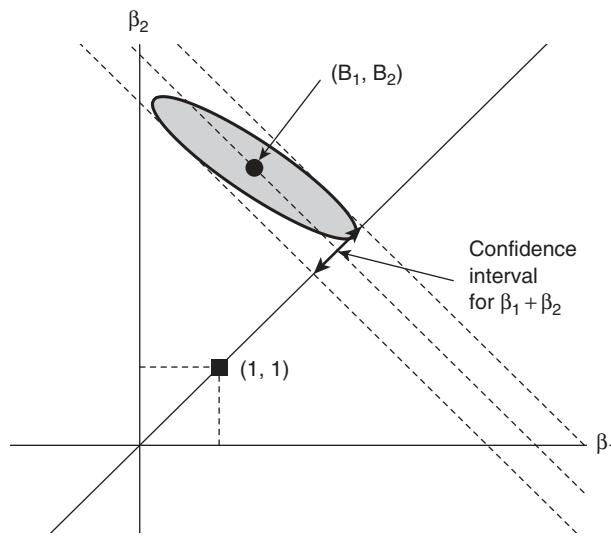


Figure 9.2 To locate the 95% confidence interval for the linear combination of coefficients $\beta_1 + \beta_2$, find the perpendicular shadow of the confidence interval–generating ellipse on the line through the origin and the point $(1, 1)$. The regression coefficients (B_1, B_2) and confidence interval–generating ellipse are not the same as in the previous figure.

SOURCE: Adapted from Monette (1990, Figure 5.7A), in *Modern Methods of Data Analysis*, copyright ©1990 by Sage Publications, Inc. Reprinted with permission of Sage Publications, Inc.

is a $k \times 1$ vector of explanatory-variable values, \bar{x} is the vector of means of the X s, and S_{XX} is the sample covariance matrix of the X s. Setting the quadratic form to 1 produces the equation of an ellipsoid—called the *standard data ellipsoid*—centered at the means of the explanatory variables.

For two explanatory variables, the standard data *ellipse* has the equation

$$\frac{n-1}{\sum x_{i1}^2 \sum x_{i2}^2 - (\sum x_{i1}^* x_{i2}^*)^2} [x_1 - \bar{x}_1, x_2 - \bar{x}_2] \times \begin{bmatrix} \sum x_{i2}^{*2} & -\sum x_{i1}^* x_{i2}^* \\ -\sum x_{i1}^* x_{i2}^* & \sum x_{i1}^{*2} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{bmatrix} = 1 \quad (9.21)$$

representing an ellipse whose horizontal shadow is twice the standard deviation of X_1 and whose vertical shadow is twice the standard deviation of X_2 . These properties are illustrated in Figure 9.3, which shows scatterplots for highly correlated and uncorrelated X s. The major axis of the data ellipse has a positive tilt when the X s are positively correlated, as in Figure 9.3(a).

This representation of the data is most compelling when the explanatory variables are normally distributed. In this case, the means and covariance matrix of the X s are sufficient statistics for their joint distribution, and the standard data ellipsoid estimates a constant-density contour of the joint distribution. Even when—as is typical—the explanatory variables are *not* multivariate normal, however, the standard ellipsoid is informative because of the role of the means, variances, and covariances of the X s in the least-squares fit.

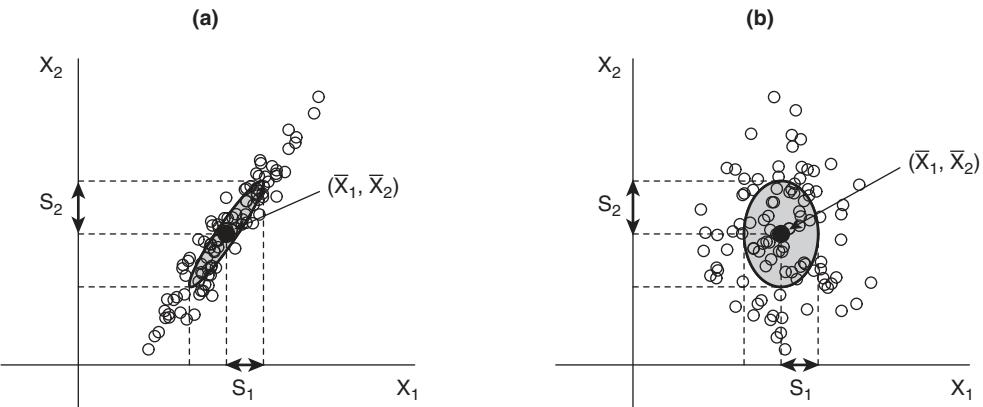


Figure 9.3 Scatterplot and standard data ellipse for (a) two highly correlated regressors and (b) two uncorrelated regressors, X_1 and X_2 . In each panel, the standard ellipse is centered at the point of means (\bar{X}_1, \bar{X}_2) ; its shadows on the axes give the standard deviations of the two variables. (The standard deviations are the half-widths of the shadows.) The data in these figures (along with data on Y) gave rise to the joint confidence ellipses in Figure 9.1. In each case, the confidence ellipse is the rescaled and translated 90° rotation of the data ellipse. Positively correlated X s, as in panel (a), produce negatively correlated coefficients, as in Figure 9.1(a).

SOURCE: Adapted from Monette (1990, Figure 5.7A), in *Modern Methods of Data Analysis*, copyright © 1990 by Sage Publications, Inc. Reprinted with permission of Sage Publications, Inc.

The joint confidence ellipse (Equation 9.19 on page 221) for the slope coefficients and the standard data ellipse (Equation 9.21) of the X s are, except for a constant scale factor and their respective centers, inverses of each other—that is, the confidence ellipse is (apart from its size and location) the 90° rotation of the data ellipse. In particular, if the data ellipse is positively tilted, reflecting a *positive* correlation between the X s, then the confidence ellipse is negatively tilted, reflecting *negatively* correlated coefficient estimates. Likewise, directions in which the data ellipse is relatively *thick*, reflecting a substantial amount of data, are directions in which the confidence ellipse is relatively *thin*, reflecting substantial information about the corresponding linear combination of regression coefficients. Thus, when the X s are strongly positively correlated (and assuming, for simplicity, that the standard deviations of X_1 and X_2 are similar), there is a great deal of information about $\beta_1 + \beta_2$ but little about $\beta_1 - \beta_2$ (as in Figure 9.2).⁵¹

The joint confidence region for the q parameters β_1 , given by $(\mathbf{b}_1 - \beta_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1) \leq q S_E^2 F_{a, q, n-k-1}$, represents the combinations of values of these parameters that are jointly acceptable at the $1 - a$ level of confidence. The boundary of the joint confidence region is an ellipsoid in the q -dimensional parameter space, reflecting the correlational structure and dispersion of the X s.

⁵¹See Exercise 9.3.

9.5 Multivariate Linear Models

The *multivariate linear model* accommodates two or more *response* variables. The theory of multivariate linear models is developed very briefly in this section. Much more extensive treatments may be found in the recommended reading for this chapter.⁵²

Specification, estimation, and testing of multivariate linear models largely parallel univariate linear models. The multivariate general linear model is

$$\mathbf{Y}_{(n \times m)} = \mathbf{X}_{(n \times k+1)} \mathbf{B}_{(k+1 \times m)} + \mathbf{E}_{(n \times m)}$$

where \mathbf{Y} is a matrix of n observations on m response variables; \mathbf{X} is a model matrix with columns for $k + 1$ regressors, including an initial column for the regression constant; \mathbf{B} is a matrix of regression coefficients, one column for each response variable; and \mathbf{E} is a matrix of errors.⁵³ The contents of the model matrix are exactly as in the univariate linear model and may contain, therefore, dummy regressors representing factors, interaction regressors, and so on.

The assumptions of the multivariate linear model concern the behavior of the errors: Let $\boldsymbol{\varepsilon}'_i$ represent the i th row of \mathbf{E} . Then $\boldsymbol{\varepsilon}'_i \sim \mathbf{N}_m(\mathbf{0}, \Sigma)$, where Σ is a nonsingular error-covariance matrix, constant across observations; $\boldsymbol{\varepsilon}'_i$ and $\boldsymbol{\varepsilon}'_{i'}$ are independent for $i \neq i'$; and \mathbf{X} is fixed or independent of \mathbf{E} .⁵⁴

The maximum-likelihood estimator of \mathbf{B} in the multivariate linear model is equivalent to equation-by-equation least squares for the individual responses:⁵⁵

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Procedures for statistical inference in the multivariate linear model, however, take account of the fact that there are several, generally correlated, responses.

Paralleling the decomposition of the total sum of squares into regression and residual sums of squares in the univariate linear model, there is in the multivariate linear model a decomposition of the total *sum-of-squares-and-cross-products* (SSP) matrix into regression and residual SSP matrices. We have

$$\begin{aligned} \mathbf{SSP}_T &= \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}'\bar{\mathbf{y}} \\ &= \hat{\mathbf{E}}'\hat{\mathbf{E}} + (\hat{\mathbf{Y}}'\hat{\mathbf{Y}} - n\bar{\mathbf{y}}'\bar{\mathbf{y}}') \\ &= \mathbf{SSP}_R + \mathbf{SSP}_{\text{Reg}} \end{aligned}$$

where $\bar{\mathbf{y}}$ is the $(m \times 1)$ vector of means for the response variables, $\hat{\mathbf{Y}} \equiv \mathbf{X}\hat{\mathbf{B}}$ is the matrix of fitted values, and $\hat{\mathbf{E}} \equiv \mathbf{Y} - \hat{\mathbf{Y}}$ is the matrix of residuals.

Many hypothesis tests of interest can be formulated by taking differences in $\mathbf{SSP}_{\text{Reg}}$ (or, equivalently, \mathbf{SSP}_R) for nested models. Let \mathbf{SSP}_H represent the incremental SSP matrix for a hypothesis. Multivariate tests for the hypothesis are based on the m eigenvalues L_j of

⁵²Some applications of multivariate linear models are given in the data analysis exercises for the chapter.

⁵³A typographical note: \mathbf{B} and \mathbf{E} are, respectively, the uppercase Greek letters Beta and Epsilon. Because these are indistinguishable from the corresponding Roman letters B and E, I will denote the estimated regression coefficients as $\hat{\mathbf{B}}$ and the residuals as $\hat{\mathbf{E}}$.

⁵⁴We can write more compactly that $\text{vec}(\mathbf{E}) \sim \mathbf{N}_{nm}(\mathbf{0}, \mathbf{I}_n \otimes \Sigma)$. Here, $\text{vec}(\mathbf{E})$ ravelles the error matrix row-wise into a vector, and \otimes is the Kronecker-product operator (see online Appendix B on matrices, linear algebra, and vector geometry).

⁵⁵See Exercise 9.16.

$\mathbf{SSP}_H \mathbf{SSP}_R^{-1}$ (the hypothesis SSP matrix “divided by” the residual SSP matrix), that is, the values of L for which⁵⁶

$$\det(\mathbf{SSP}_H \mathbf{SSP}_R^{-1} - L\mathbf{I}_m) = 0.$$

The several commonly employed multivariate test statistics are functions of these eigenvalues:

$$\begin{aligned} \text{Pillai-Bartlett Trace, } T_{PB} &= \sum_{j=1}^m \frac{L_j}{1-L_j} \\ \text{Hotelling-Lawley Trace, } T_{HL} &= \sum_{j=1}^m L_j \\ \text{Wilks's Lambda, } \Lambda &= \prod_{j=1}^m \frac{1}{1+L_j} \\ \text{Roy's Maximum Root, } L_1 & \end{aligned} \tag{9.22}$$

There are F approximations to the null distributions of these test statistics. For example, for Wilks's Lambda, let s represent the degrees of freedom for the term that we are testing (i.e., the number of columns of the model matrix \mathbf{X} pertaining to the term). Define

$$\begin{aligned} r &\equiv n - k - 1 - \frac{m - s + 1}{2} \\ u &\equiv \frac{ms - 2}{4} \\ t &\equiv \begin{cases} \frac{\sqrt{m^2s^2 - 4}}{m^2 + s^2 - 5} & \text{for } m^2 + s^2 - 5 > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9.23}$$

Rao (1973, p. 556) shows that under the null hypothesis,

$$F_0 \equiv \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \times \frac{rt - 2u}{ms} \tag{9.24}$$

follows an approximate F -distribution with ms and $rt - 2u$ degrees of freedom and that this result is exact if $\min(m, s) \leq 2$ (a circumstance under which all four test statistics are equivalent).

Even more generally, suppose that we want to test the linear hypothesis

$$H_0: \mathbf{L}_{(q \times k+1)} \mathbf{B}_{(k+1 \times m)} = \mathbf{C}_{(q \times m)} \tag{9.25}$$

where \mathbf{L} is a hypothesis matrix of full-row rank $q \leq k + 1$, and the right-hand-side matrix \mathbf{C} consists of constants (usually zeroes). Then the SSP matrix for the hypothesis is

$$\mathbf{SSP}_H = (\widehat{\mathbf{B}'\mathbf{L}' - \mathbf{C}'}) [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1} (\mathbf{L}\widehat{\mathbf{B}} - \mathbf{C})$$

and the various test statistics are based on the $p \equiv \min(q, m)$ nonzero eigenvalues of $\mathbf{SSP}_H \mathbf{SSP}_R^{-1}$ (and the formulas in Equations 9.22, 9.23, and 9.24 are adjusted by substituting p for m).

⁵⁶Eigenvalues and determinants are described in online Appendix B on matrices, linear algebra, and vector geometry.

When a multivariate response arises because a variable is measured on different occasions or under different circumstances (but for the same individuals), it is also of interest to formulate hypotheses concerning comparisons among the responses. This situation, called a *repeated-measures design*, can be handled by linearly transforming the responses using a suitable model matrix, for example, extending the linear hypothesis in Equation 9.25 to

$$H_0: \begin{matrix} \mathbf{L} \\ (q \times k+1) \end{matrix} \quad \begin{matrix} \mathbf{B} \\ (k+1 \times m) \end{matrix} \quad \begin{matrix} \mathbf{P} \\ (m \times v) \end{matrix} = \begin{matrix} \mathbf{C} \\ (q \times v) \end{matrix}$$

Here, the matrix \mathbf{P} provides contrasts in the responses (see, e.g., Hand & Taylor, 1987, or O'Brien & Kaiser, 1985). The SSP matrix for the hypothesis is

$$\mathbf{SSP}_H = (\mathbf{P}' \widehat{\mathbf{B}}' \mathbf{L}' - \mathbf{C}') \left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{L}' \right]^{-1} (\mathbf{L} \widehat{\mathbf{B}} \mathbf{P} - \mathbf{C})$$

and test statistics are based on the $p \equiv \min(q, v)$ nonzero eigenvalues of $\mathbf{SSP}_H (\mathbf{P}' \mathbf{SSP}_R \mathbf{P})^{-1}$.

The multivariate linear model accommodates several response variables:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Under the assumption that the rows ε'_i of the error matrix \mathbf{E} are independent and multivariately normally distributed with mean $\mathbf{0}$ and common nonsingular covariance matrix Σ , the maximum-likelihood estimators of the regression coefficients are given by

$$\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

Hypothesis tests for the multivariate linear model closely parallel those for the univariate linear model, with sum-of-squares-and-products (SSP) matrices in the multivariate case generalizing the role of sums of squares in the univariate case. Several commonly employed test statistics are based on the eigenvalues of $\mathbf{SSP}_H \mathbf{SSP}_R^{-1}$, where \mathbf{SSP}_H is the SSP matrix for a hypothesis, and \mathbf{SSP}_R is the residual SSP matrix.

9.6 Random Regressors

The theory of linear models developed in this chapter has proceeded from the premise that the model matrix \mathbf{X} is *fixed*. If we repeat a study, we expect the response-variable observations \mathbf{y} to change, but if \mathbf{X} is fixed, then the explanatory-variable values are constant across replications of the study. This situation is realistically descriptive of an experiment, where the explanatory variables are manipulated by the researcher. Most research in the social sciences, however, is observational rather than experimental, and in an observational study (e.g., survey research⁵⁷), we would typically obtain different explanatory-variable values on replication of the study. In observational research, therefore, \mathbf{X} is *random* rather than fixed.

⁵⁷Randomized comparative experiments can be carried out in the context of a sample survey by varying aspects of questions asked of respondents. See, e.g., Auspurg and Hinz (in press).

It is remarkable that the statistical theory of linear models applies even when \mathbf{X} is random, as long as certain assumptions are met. For fixed explanatory variables, the assumptions underlying the model take the form $\varepsilon \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$. That is, the distribution of the error is the same for all observed combinations of explanatory-variable values represented by the distinct rows of the model matrix. When \mathbf{X} is random, we need to assume that this property holds for *all possible* combinations of explanatory-variable values in the population that is sampled: That is, \mathbf{X} and ε are assumed to be independent, and thus the *conditional* distribution of the error for a sample of explanatory variable values $\varepsilon|\mathbf{X}_0$ is $N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$, regardless of the *particular* sample $\mathbf{X}_0 = \{x_{ij}\}$ that is chosen.

Because \mathbf{X} is random, it has some (multivariate) probability distribution. It is not necessary to make assumptions about this distribution, however, beyond (1) requiring that \mathbf{X} is measured without error and that \mathbf{X} and ε are independent (as just explained), (2) assuming that the distribution of \mathbf{X} does not depend on the parameters β and σ_ε^2 of the linear model, and (3) stipulating that the covariance matrix of the X s is nonsingular (i.e., that no X is invariant or a perfect linear function of the others in the population). In particular, we need *not* assume that the *regressors* (as opposed to the *errors*) are normally distributed. This is fortunate, for many regressors are highly non-normal—dummy regressors and polynomial regressors come immediately to mind, not to mention many quantitative explanatory variables.

It would be unnecessarily tedious to recapitulate the entire argument of this chapter, but I will show that some key results hold, under the new assumptions, when the explanatory variables are random. The other results of the chapter can be established for random regressors in a similar manner.

For a particular sample of X -values, \mathbf{X}_0 , the conditional distribution of \mathbf{y} is

$$\begin{aligned} E(\mathbf{y}|\mathbf{X}_0) &= E[(\mathbf{X}\beta + \varepsilon)|\mathbf{X}_0] = \mathbf{X}_0\beta + E(\varepsilon|\mathbf{X}_0) \\ &= \mathbf{X}_0\beta \end{aligned}$$

Consequently, the conditional expectation of the least-squares estimator is

$$\begin{aligned} E(\mathbf{b}|\mathbf{X}_0) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}_0\right] = (\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_0E(\mathbf{y}|\mathbf{X}_0) \\ &= (\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{X}'_0\mathbf{X}_0\beta = \beta \end{aligned}$$

Because we can repeat this argument for *any* value of \mathbf{X} , the least-squares estimator \mathbf{b} is conditionally unbiased for any and every such value; it is therefore *unconditionally* unbiased as well, $E(\mathbf{b}) = \beta$.

Suppose now that we use the procedures of the previous section to perform statistical inference for β . For concreteness, imagine that we calculate a p -value for the omnibus null hypothesis $H_0: \beta_1 = \dots = \beta_k = 0$. Because $\varepsilon|\mathbf{X}_0 \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$, as was required when we treated \mathbf{X} as fixed, the p -value obtained is correct for $\mathbf{X} = \mathbf{X}_0$ (i.e., for the sample at hand). There is, however, nothing special about a particular \mathbf{X}_0 : The error vector ε is independent of \mathbf{X} , and so the distribution of ε is $N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$ for any and every value of \mathbf{X} . The p -value, therefore, is *unconditionally* valid.

Finally, I will show that the maximum-likelihood estimators of β and σ_ε^2 are unchanged when \mathbf{X} is random, as long as the new assumptions hold: When \mathbf{X} is random, sampled observations consist not just of response-variable values (Y_1, \dots, Y_n) but also of explanatory-variable values $(\mathbf{x}'_1, \dots, \mathbf{x}'_n)$. The observations themselves are denoted $[Y_1, \mathbf{x}'_1], \dots, [Y_n, \mathbf{x}'_n]$. Because

these observations are sampled independently, their joint probability density is the product of their marginal densities:

$$p(\mathbf{y}, \mathbf{X}) = p([y_1, \mathbf{x}'_1], \dots, [y_n, \mathbf{x}'_n]) = p(y_1, \mathbf{x}'_1) \times \dots \times p(y_n, \mathbf{x}'_n)$$

Now, the probability density $p(y_i, \mathbf{x}'_i)$ for observation i can be written as $p(y_i | \mathbf{x}'_i)p(\mathbf{x}'_i)$. According to the linear model, the conditional distribution of y_i given \mathbf{x}'_i is normal:

$$p(y_i | \mathbf{x}'_i) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp \left[-\frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma_\varepsilon^2} \right]$$

Thus, the joint probability density for all observations becomes

$$\begin{aligned} p(\mathbf{y}, \mathbf{X}) &= \prod_{i=1}^n p(\mathbf{x}'_i) \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp \left[-\frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma_\varepsilon^2} \right] \\ &= \left[\prod_{i=1}^n p(\mathbf{x}'_i) \right] \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\varepsilon^2} \right] \\ &= p(\mathbf{X})p(\mathbf{y}|\mathbf{X}) \end{aligned}$$

As long as $p(\mathbf{X})$ does not depend on the parameters $\boldsymbol{\beta}$ and σ_ε^2 , we can ignore the joint density of the X s in maximizing $p(\mathbf{y}, \mathbf{X})$ with respect to the parameters. Consequently, the maximum-likelihood estimator of $\boldsymbol{\beta}$ is the least-squares estimator, as was the case for fixed \mathbf{X} .⁵⁸

The statistical theory of linear models, formulated under the supposition that the model matrix \mathbf{X} is fixed with respect to repeated sampling, is also valid when \mathbf{X} is random, as long as three additional requirements are satisfied: (1) the model matrix \mathbf{X} and the errors ε are independent; (2) the distribution of \mathbf{X} , which is otherwise unconstrained, does not depend on the parameters $\boldsymbol{\beta}$ and σ_ε^2 of the linear model; and (3) the covariance matrix of the X s is nonsingular.

9.7 Specification Error

To generalize our treatment of misspecified structural relationships,⁵⁹ it is convenient to work with probability limits.⁶⁰ Suppose that the response variable Y is determined by the model

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \varepsilon = \mathbf{X}_1^* \boldsymbol{\beta}_1 + \mathbf{X}_2^* \boldsymbol{\beta}_2 + \varepsilon$$

where the error ε behaves according to the usual assumptions. I have, for convenience, expressed each variable as deviations from its expectation [e.g., $\mathbf{y}^* \equiv \{Y_i - E(Y)\}$] and have

⁵⁸Cf. Section 9.3.3.

⁵⁹See Section 6.3.

⁶⁰Probability limits are introduced in online Appendix D.

partitioned the model matrix into two sets of regressors; the parameter vector is partitioned in the same manner.⁶¹

Imagine that we ignore \mathbf{X}_2^* , so that $\mathbf{y}^* = \mathbf{X}_1^*\boldsymbol{\beta}_1 + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\boldsymbol{\varepsilon}} \equiv \mathbf{X}_2^*\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. The least-squares estimator for $\boldsymbol{\beta}_1$ in the model that omits \mathbf{X}_2^* is

$$\begin{aligned}\mathbf{b}_1 &= (\mathbf{X}_1^{*\prime}\mathbf{X}_1^*)^{-1}\mathbf{X}_1^{*\prime}\mathbf{y}^* \\ &= \left(\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{y}^* \\ &= \left(\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\frac{1}{n}\mathbf{X}_1^{*\prime}(\mathbf{X}_1^*\boldsymbol{\beta}_1 + \mathbf{X}_2^*\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}_1 + \left(\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{X}_2^*\boldsymbol{\beta}_2 + \left(\frac{1}{n}\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\frac{1}{n}\mathbf{X}_1^{*\prime}\boldsymbol{\varepsilon}\end{aligned}$$

Taking probability limits produces

$$\begin{aligned}\text{plim } \mathbf{b}_1 &= \boldsymbol{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\beta}_2 + \Sigma_{11}^{-1}\sigma_{1\varepsilon} \\ &= \boldsymbol{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\beta}_2\end{aligned}$$

where

- $\boldsymbol{\Sigma}_{11} \equiv \text{plim}(1/n)\mathbf{X}_1^{*\prime}\mathbf{X}_1^*$ is the population covariance matrix for \mathbf{X}_1 ;
- $\boldsymbol{\Sigma}_{12} \equiv \text{plim}(1/n)\mathbf{X}_1^{*\prime}\mathbf{X}_2^*$ is the matrix of population covariances between \mathbf{X}_1 and \mathbf{X}_2 ; and
- $\boldsymbol{\sigma}_{1\varepsilon} \equiv \text{plim}(1/n)\mathbf{X}_1^{*\prime}\boldsymbol{\varepsilon}$ is the vector of population covariances between \mathbf{X}_1 and $\boldsymbol{\varepsilon}$, which is $\mathbf{0}$ by the assumed independence of the error and the explanatory variables.

The asymptotic (or population) covariance of \mathbf{X}_1 and $\tilde{\boldsymbol{\varepsilon}}$ is not generally $\mathbf{0}$, however, as is readily established:

$$\begin{aligned}\text{plim } \frac{1}{n}\mathbf{X}_1^{*\prime}\tilde{\boldsymbol{\varepsilon}} &= \text{plim } \frac{1}{n}\mathbf{X}_1^{*\prime}(\mathbf{X}_2^*\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_2 + \boldsymbol{\sigma}_{1\varepsilon} = \boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_2\end{aligned}$$

The estimator \mathbf{b}_1 , therefore, is consistent if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ —that is, if the excluded regressors in \mathbf{X}_2 are uncorrelated with the included regressors in \mathbf{X}_1 . In this case, incorporating \mathbf{X}_2^* in the error does not induce a correlation between \mathbf{X}_1^* and the compound error $\tilde{\boldsymbol{\varepsilon}}$. The estimated coefficients \mathbf{b}_1 are also consistent if $\boldsymbol{\beta}_2 = \mathbf{0}$: Excluding *irrelevant* regressors is unproblematic.⁶²

The omission of regressors from a linear model causes the coefficients of the included regressors to be inconsistent, unless (1) the omitted regressors are uncorrelated with the included regressors or (2) the omitted regressors have coefficients of 0 and hence are irrelevant.

⁶¹Expressing the variables as deviations from their expectations eliminates the constant β_0 .

⁶²Including irrelevant regressors also does not cause the least-squares estimator to become inconsistent; after all, if the assumptions of the model hold, then \mathbf{b} is a consistent estimator of $\boldsymbol{\beta}$ even if some of the elements of $\boldsymbol{\beta}$ are 0. (Recall, however, Exercise 6.9.)

9.8 Instrumental Variables and Two-Stage Least Squares

Under certain circumstances, *instrumental-variables estimation* allows us to obtain consistent estimates of regression coefficients when (some) explanatory variables are correlated with the regression error. I develop the topic briefly, along with an extension of instrumental-variables estimation called *two-stage least squares*, deferring applications to the data analysis exercises.

9.8.1 Instrumental-Variable Estimation in Simple Regression

Let us begin with the simple-regression model, $Y = \alpha + \beta X + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. As in the preceding section, we can eliminate the intercept α from the model by expressing each of X and Y as deviations from their expectations, $X^* \equiv X - E(X)$ and $Y^* \equiv Y - E(Y)$:

$$Y^* = \beta X^* + \varepsilon \quad (9.26)$$

Under the assumption that X and the error ε are independent, multiplying Equation 9.26 through by X^* and taking expectations produces the following result:

$$\begin{aligned} E(X^* Y^*) &= \beta E(X^{*2}) + E(X^* \varepsilon) \\ \sigma_{XY} &= \beta \sigma_X^2 + 0 \end{aligned}$$

where σ_{XY} is the population covariance of X and Y , and σ_X^2 is the variance of X . Because X and the error are independent, $E(X^* \varepsilon)$, the covariance of X and ε , is 0. Solving for the population regression coefficient, $\beta = \sigma_{XY}/\sigma_X^2$. Finally, to obtain a consistent estimator of β , we can substitute the sample covariance S_{XY} and variance S_X^2 for their population analogs, of which they are consistent estimators. We obtain $B_{OLS} = S_{XY}/S_X^2$, which we recognize as the ordinary least-squares (OLS) slope coefficient in simple regression.

Now suppose that it is unreasonable to assume that X and the error are independent but that there is a third observed variable, Z , that is independent of the error ε but correlated with X , so that $\sigma_{Z\varepsilon} = 0$ and $\sigma_{ZX} \neq 0$. Then, in the same manner as before, but multiplying Equation 9.26 through by $Z^* \equiv Z - E(Z)$,

$$\begin{aligned} E(Z^* Y^*) &= \beta E(Z^* X^*) + E(Z^* \varepsilon) \\ \sigma_{ZY} &= \beta \sigma_{ZX} + 0 \end{aligned}$$

and $\beta = \sigma_{ZY}/\sigma_{ZX}$. The variable Z is called an *instrumental variable* (or *instrument*).⁶³ Substituting sample for population covariances produces the consistent *instrumental-variable (IV) estimator* of β , $B_{IV} = S_{ZY}/S_{ZX}$. The requirement that Z be correlated with X (in addition to being independent of the error) is analogous to the stipulation in OLS estimation that the explanatory variable X is not invariant. Notice that if the explanatory variable and the instrumental variable are the same, $Z = X$, then $B_{IV} = B_{OLS}$.⁶⁴

How might we justify the assumption that Z and ε are not related? The justification requires substantive reasoning about the research problem at hand—no different, in principle, from

⁶³Instrumental-variable estimation was introduced in the context of the simple-regression model in Exercise 6.14 (page 126).

⁶⁴Moreover, under these circumstances, the Gauss-Markov theorem (Section 9.3.2) ensures that X is the *best* instrument, producing the smallest coefficient standard error.

asserting that X and ε are independent to justify causal inference in OLS regression. For example, suppose that an experiment is conducted in which half the first-grade students in a school district, selected at random, are given vouchers to defray the cost of attending a private school. At the end of the year, students are administered a standardized exam covering the academic content of the first grade.

Imagine that the researchers conducting the experiment are interested in the effect of private school attendance, in comparison to public school attendance, on average student achievement. If every student who received a voucher attended a private school and every student who did not receive a voucher attended a public school, analysis of the results of the experiment would be reasonably straightforward. Let us suppose, however, that this is not the case and that some students receiving vouchers attended public schools and some not receiving vouchers attended private schools.

Treating the test scores as a numeric response, Y , and the kind of school *actually attended* as a dummy regressor, X , coded 1 for private school attendance and 0 for public school attendance, the coefficient β in Equation 9.26 represents the difference in mean achievement between comparable students attending private and public schools. The OLS regression of Y on X is equivalent to a difference-of-means t -test between the group of students attending private schools and the group attending public schools.

Under these circumstances, however, it is unreasonable to assume that X and the error—partly comprising the omitted causes of student achievement, beyond kind of school—are independent, because the kind of school that the students attend is at least partly determined by their families' characteristics. For example, students *not receiving* a voucher might nevertheless be more likely to attend private schools if their families are relatively wealthy, and children *receiving* a voucher might be more likely nevertheless to attend public schools if their families are relatively poor.

On the other hand, the random assignment itself, also treated as a dummy variable, Z , coded 1 if a student receives a voucher and 0 otherwise, is reasonably assumed to be independent of the error. Because the kind of school that a student attends is also likely to be related to receipt of a private school voucher, Z can serve as an instrumental variable, to obtain an unbiased estimate of β , the effect on achievement of private versus public school attendance.⁶⁵

9.8.2 Instrumental-Variables Estimation in Multiple Regression

We can generalize the instrumental-variable estimator to the multiple-regression model, not bothering this time to center the variables at their expectations and working (as in Section 9.7) with probability limits. The familiar linear model is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \varepsilon \\ \varepsilon &\sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n) \end{aligned} \tag{9.27}$$

where \mathbf{y} is the $n \times 1$ response vector for a sample of n observations; \mathbf{X} is the $n \times k + 1$ model matrix, with an initial columns of 1s; $\boldsymbol{\beta}$ is the $k + 1 \times 1$ vector of regression coefficients to be estimated; and ε is the $n \times 1$ error vector.

⁶⁵For further discussion of this imaginary study, see Exercise 9.17. Using random assignment as an instrumental variable for an experimental “treatment” was proposed by Angrist, Imbens, and Rubin (1996), who discuss subtleties that I have glossed over here.

If the variables in \mathbf{X} were asymptotically uncorrelated with ε , we would be able to estimate β consistently by OLS regression, but let us imagine that this is not the case. Imagine further, however, that there exists an $n \times k + 1$ instrumental-variable matrix \mathbf{Z} , also including an initial column of 1s (and possibly including *some* of the other columns of \mathbf{X}), such that

$$\begin{aligned}\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) &= \boldsymbol{\sigma}_{ZY}^{(R)} \\ \text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) &= \boldsymbol{\Sigma}_{ZX}^{(R)}, \text{ nonsingular} \\ \text{plim}\left(\frac{1}{n}\mathbf{Z}'\varepsilon\right) &= \boldsymbol{\sigma}_{Z\varepsilon}^{(R)} = \mathbf{0} \\ \text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right) &= \boldsymbol{\sigma}_{ZZ}^{(R)}\end{aligned}$$

where the superscript (R) indicates that these are population *raw-moment* vectors and matrices of mean sums of squares and cross-products. The requirement that $\boldsymbol{\sigma}_{Z\varepsilon}^{(R)} = \mathbf{0}$ stipulates that the instrumental variables are asymptotically uncorrelated with the errors; the requirement that $\boldsymbol{\Sigma}_{ZX}^{(R)}$ is nonsingular is the IV analog of ruling out perfect collinearity and implies that each of the X 's must be correlated with the Z 's.

Multiplying Equation 9.27 through by $(1/n)\mathbf{Z}'$ and taking probability limits produces

$$\begin{aligned}\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) &= \text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\beta + \text{plim}\left(\frac{1}{n}\mathbf{Z}'\varepsilon\right) \\ \boldsymbol{\sigma}_{Zy}^{(R)} &= \boldsymbol{\Sigma}_{ZX}^{(R)}\beta + \mathbf{0}\end{aligned}$$

Then, solving for the population regression coefficients,

$$\beta = \boldsymbol{\Sigma}_{ZX}^{(R)-1} \boldsymbol{\sigma}_{ZY}^{(R)}$$

Consequently, the IV estimator

$$\begin{aligned}\mathbf{b}_{IV} &\equiv \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1} \left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) \\ &= (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}\end{aligned}\tag{9.28}$$

is a consistent estimator of β .

The asymptotic covariance matrix of \mathbf{b}_{IV} is given by⁶⁶

$$\mathcal{V}(\mathbf{b}_{IV}) = \frac{\sigma_\varepsilon^2}{n} \boldsymbol{\Sigma}_{ZX}^{(R)-1} \boldsymbol{\Sigma}_{ZZ}^{(R)} \boldsymbol{\Sigma}_{XZ}^{(R)-1}\tag{9.29}$$

This result cannot, of course, be applied directly because we do not know either the error variance, σ_ε^2 , or the population moments. As in least-squares estimation, we can estimate the error variance from the residuals:

⁶⁶See Exercise 9.18.

$$\mathbf{e}_{\text{IV}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{IV}}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{\mathbf{e}'_{\text{IV}} \mathbf{e}_{\text{IV}}}{n - k - 1}$$

Substituting sample for population moment matrices,

$$\begin{aligned}\hat{\mathcal{V}}(\mathbf{b}_{\text{IV}}) &= \frac{\hat{\sigma}_\varepsilon^2}{n} \left(\frac{1}{n} \mathbf{Z}' \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{Z}' \mathbf{Z} \left(\frac{1}{n} \mathbf{X}' \mathbf{Z} \right)^{-1} \\ &= \hat{\sigma}_\varepsilon^2 (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{X}' \mathbf{Z})^{-1}\end{aligned}\quad (9.30)$$

Tests and confidence regions can then be based on \mathbf{b}_{IV} and $\hat{\mathcal{V}}(\mathbf{b}_{\text{IV}})$.⁶⁷ Notice that if all the explanatory variables and the instrumental variables are identical, $\mathbf{X} = \mathbf{Z}$, then the IV and OLS estimators and their respective covariance matrices coincide.⁶⁸

If (some of) the $k + 1$ columns of the model matrix \mathbf{X} are correlated with the error ε in the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, then the OLS estimator of β is inconsistent. Suppose, however, that there exists a matrix \mathbf{Z} of instrumental variables with $k + 1$ columns (some of which may be the same as columns of \mathbf{X}) that are uncorrelated with the error but correlated with \mathbf{X} . Then, $\mathbf{b}_{\text{IV}} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y}$ is a consistent estimator of β , with estimated asymptotic covariance matrix $\hat{\mathcal{V}}(\mathbf{b}_{\text{IV}}) = \hat{\sigma}_\varepsilon^2 (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{Z} (\mathbf{X}' \mathbf{Z})^{-1}$, where $\hat{\sigma}_\varepsilon^2 = \mathbf{e}'_{\text{IV}} \mathbf{e}_{\text{IV}} / (n - k - 1)$ and $\mathbf{e}_{\text{IV}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{\text{IV}}$, as in OLS regression.

9.8.3 Two-Stage Least Squares

The instrumental-variables estimator of the preceding section requires that we have exactly as many instrumental variables in \mathbf{Z} as explanatory variables in \mathbf{X} —that is, $k + 1$ (including the regression constant). If there are fewer IVs than explanatory variables, then there will be fewer than $k + 1$ IV estimating equations, comprising the rows of

$$\mathbf{Z}' \mathbf{X} \mathbf{b}_{\text{IV}} = \mathbf{Z}' \mathbf{y} \quad (9.31)$$

Because there are $k + 1$ parameters in β to estimate, \mathbf{b}_{IV} will be underdetermined.

If, alternatively, there are *more than* $k + 1$ IVs in \mathbf{Z} , then Equations (9.31) will be overdetermined. It is important to understand, however, that this situation is an embarrassment of riches: We could obtain consistent IV estimates of β by discarding IVs until we have just the right number, $k + 1$. To do so, however, would be arbitrary and would waste information that we could deploy to increase the precision of estimation. *Two-stage least-squares (2SLS)* estimation, originally developed in the 1950s by the econometricians Theil (cited in Theil, 1971, p. 452) and Basmann (1957), is a method for reducing the IVs to just the right number, not by discarding surplus IVs but by combining the available IVs in an advantageous manner.

⁶⁷As in Sections 9.4.3 and 9.4.4.

⁶⁸See Exercise 9.19.

An instrumental variable must be correlated with at least some of the X s while remaining uncorrelated with the error, and good instrumental variables—those that produce precise estimates of the regression coefficients—must be as correlated as possible with the X s. In the first stage of 2SLS, we regress the X s on the Z s, computing fitted values, \hat{X} s, which are the linear combinations of the Z s most highly correlated with the original X s. Because the \hat{X} s are linear combinations of the Z s, they too are uncorrelated with the errors.⁶⁹ The fitted values are obtained from the multivariate least-squares regression of \mathbf{X} on \mathbf{Z} :⁷⁰

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

In a typical application, some of the columns of \mathbf{X} also are among the IVs in \mathbf{Z} (this is always true of the regression constant, for example), and for these explanatory variables, the observed and fitted values are identical, $\hat{X} = X$.⁷¹

In the second stage of 2SLS, we either apply the \hat{X} s as instruments or, equivalently, perform a least-squares regression of Y on the \hat{X} s (justifying the name “two-stage least-squares”). The first approach leads to

$$\mathbf{b}_{2SLS} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \quad (9.32)$$

while the second approach leads to

$$\mathbf{b}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \quad (9.33)$$

Showing that $\hat{\mathbf{X}}'\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$ demonstrates the equivalence of the two approaches.⁷² When the number of IVs in \mathbf{Z} is $k + 1$, $\mathbf{b}_{2SLS} = \mathbf{b}_{2SLS}$ —that is, the two-stage least-squares estimator (Equation 9.31) and the IV estimator (Equation 9.28) coincide.⁷³

Finally, the covariance matrix of the 2SLS estimator follows from Equation 9.30, with $\hat{\mathbf{X}}$ in the role of \mathbf{Z} :

$$\hat{\mathcal{V}}(\mathbf{b}_{2SLS}) = \hat{\sigma}_\varepsilon^2 \left(\hat{\mathbf{X}}'\mathbf{X} \right)^{-1} \hat{\mathbf{X}}'\hat{\mathbf{X}} \left(\mathbf{X}'\hat{\mathbf{X}} \right)^{-1} \quad (9.34)$$

The estimated error variance $\hat{\sigma}_\varepsilon^2$ is computed from the residuals as for any IV estimator.

If there are fewer instrumental variables in the IV matrix \mathbf{Z} than regression coefficients to estimate (corresponding to the $k + 1$ columns of model matrix \mathbf{X}), then the IV estimating equations are underdetermined, preventing us from solving uniquely for \mathbf{b}_{IV} . If, however, there are *more* instrumental variables than regression coefficients, then the IV estimating equations employed directly will be overdetermined. The surplus information available in the IVs is used efficiently by the two-stage least-squares estimator $\mathbf{b}_{2SLS} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$ where $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. The estimated asymptotic covariance matrix of the 2SLS estimator is $\hat{\mathcal{V}}(\mathbf{b}_{2SLS}) = \hat{\sigma}_\varepsilon^2 \left(\hat{\mathbf{X}}'\mathbf{X} \right)^{-1} \hat{\mathbf{X}}'\hat{\mathbf{X}} \left(\mathbf{X}'\hat{\mathbf{X}} \right)^{-1}$.

⁶⁹Because the regression coefficients used to compute the \hat{X} s are themselves subject to sampling variation, rather than fixed values, this argument applies asymptotically as the sample size n grows.

⁷⁰See Section 9.5.

⁷¹See Exercise 9.20(a).

⁷²See Exercise 9.20.

⁷³See Exercise 9.20(c).

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 9.1. *Solving the parametric equations in one-way and two-way ANOVA:

- (a) Show that the parametric equation (Equation 9.5, page 205) in one-way ANOVA has the general solution

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} = \begin{bmatrix} \mu.. \\ \mu_1 - \mu.. \\ \mu_2 - \mu.. \\ \vdots \\ \mu_{m-1} - \mu.. \end{bmatrix}$$

- (b) Show that the parametric equation (Equation 9.6, page 205) in two-way ANOVA, with two rows and three columns, has the solution

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} = \begin{bmatrix} \mu.. \\ \mu_{1..} - \mu.. \\ \mu_{.1} - \mu.. \\ \mu_{.2} - \mu.. \\ \mu_{11} - \mu_{1..} - \mu_{.1} + \mu.. \\ \mu_{12} - \mu_{1..} - \mu_{.2} + \mu.. \end{bmatrix}$$

Exercise 9.2. *Orthogonal contrasts (see Section 9.1.2): Consider the equation $\beta_F = \mathbf{X}_B^{-1}\mu$ relating the parameters β_F of the full-rank ANOVA model to the cell means μ . Suppose that \mathbf{X}_B^{-1} is constructed so that its rows are orthogonal. Show that the columns of the row basis \mathbf{X}_B of the model matrix are also orthogonal and further that each column of \mathbf{X}_B is equal to the corresponding row of \mathbf{X}_B^{-1} divided by the sum of squared entries in that row. (*Hint:* Multiply \mathbf{X}_B^{-1} by its transpose.)

Exercise 9.3. Nonorthogonal contrasts: Imagine that we want to compare each of three groups in a one-way ANOVA with a fourth (control) group. We know that coding three dummy regressors, treating Group 4 as the baseline category, will accomplish this purpose. Starting with the equation

$$\begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

show that the row basis $\mathbf{X}_B = (\mathbf{X}_B^{-1})^{-1}$ of the model matrix is equivalent to dummy coding.

Exercise 9.4. Verify that each of the terms in the sum-of-squares function (see Equation 9.9 on page 208)

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

is (1×1) , justifying writing

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - (\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$$

Exercise 9.5. Standardized regression coefficients:⁷⁴

- (a) *Show that the standardized coefficients can be computed as $\mathbf{b}^* = \mathbf{R}_{XX}^{-1} \mathbf{r}_{Xy}$, where \mathbf{R}_{XX} is the correlation matrix of the explanatory variables, and \mathbf{r}_{Xy} is the vector of correlations between the explanatory variables and the response variable. [Hints: Let $\mathbf{Z}_X \equiv \{(X_{ij} - \bar{X}_j)/S_j\}_{(n \times k)}$ contain the standardized explanatory variables, and let $\mathbf{z}_y \equiv \{(Y_i - \bar{Y})/S_Y\}_{(n \times 1)}$ contain the standardized response variable. The regression equation for the standardized variables in matrix form is $\mathbf{z}_y = \mathbf{Z}_X \mathbf{b}^* + \mathbf{e}^*$. Multiply both sides of this equation by $\mathbf{Z}'_X/(n - 1)$.]
- (b) The correlation matrix in Table 9.2 is taken from Blau and Duncan's (1967) work on social stratification. Using these correlations, along with the results in part (a), find the standardized coefficients for the regression of current occupational status on father's education, father's occupational status, respondent's education, and the status of the respondent's first job. Why is the slope for father's education so small? Is it reasonable to conclude that father's education is unimportant as a cause of the respondent's occupational status (recall Section 6.3)?
- (c) *Prove that the squared multiple correlation for the regression of Y on X_1, \dots, X_k can be written as

$$R^2 = B_1^* r_{r1} + \dots + B_k^* r_{rk} = \mathbf{r}'_{yX} \mathbf{b}^*$$

[Hint: Multiply $\mathbf{z}_y = \mathbf{Z}_X \mathbf{b}^* + \mathbf{e}^*$ through by $\mathbf{z}'_y/(n - 1)$.] Use this result to calculate the multiple correlation for Blau and Duncan's regression.

Exercise 9.6. Using the general result $V(\mathbf{b}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1}$, show that the sampling variances of A and B in simple-regression analysis are

$$V(A) = \frac{\sigma_e^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2}$$

$$V(B) = \frac{\sigma_e^2}{\sum (X_i - \bar{X})^2}$$

Table 9.2 Correlations for Blau and Duncan's Stratification Data, $n \approx 20,700$: X_1 =Father's Education; X_2 =Father's Occupational Status; X_3 =Respondent's Education; X_4 =Status of Respondent's First Job; Y =Respondent's Current Occupational Status

	X_1	X_2	X_3	X_4	Y
X_1	1.000				
X_2	.516	1.000			
X_3	.453	.438	1.000		
X_4	.332	.417	.538	1.000	
Y	.322	.405	.596	.541	1.000

SOURCE: Blau and Duncan (1967, p. 169).

⁷⁴Standardized regression coefficients were introduced in Section 5.2.4.

Exercise 9.7. *A crucial step in the proof of the Gauss-Markov theorem (Section 9.3.2) uses the fact that the matrix product \mathbf{AX} must be $\mathbf{0}$ because $\mathbf{AX}\beta = \mathbf{0}$. Why is this the case? [Hint: The key here is that $\mathbf{AX}\beta = \mathbf{0}$ regardless of the value of β . Consider, for example, $\beta = [1, 0, \dots, 0]'$ (i.e., one possible value of β). Show that this implies that the first row of \mathbf{AX} is $\mathbf{0}$. Then consider $\beta = [0, 1, \dots, 0]'$, and so on.]

Exercise 9.8. *For the statistic

$$t = \frac{B_j - \beta_j}{S_E \sqrt{v_{jj}}}$$

to have a t -distribution, the estimators B_j and S_E must be independent. [Here, v_{jj} is the j th diagonal entry of $(\mathbf{X}'\mathbf{X})^{-1}$.] The coefficient B_j is the j th element of \mathbf{b} , and $S_E = \sqrt{\mathbf{e}'\mathbf{e}/(n - k - 1)}$ is a function of the residuals \mathbf{e} . Because both \mathbf{b} and \mathbf{e} are normally distributed, it suffices to prove that their covariance is $\mathbf{0}$. Demonstrate that this is the case. [Hint: Use $C(\mathbf{e}, \mathbf{b}) = E[\mathbf{e}(\mathbf{b} - \beta)']$, and begin by showing that $\mathbf{b} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$.]

Exercise 9.9. *Using Equation 9.12 (page 214), show that the maximized likelihood for the linear model can be written as

$$L = \left(2\pi e \frac{\mathbf{e}'\mathbf{e}}{n} \right)^{-n/2}$$

Exercise 9.10. Using Duncan's regression of occupational prestige on income and education, and performing the necessary calculations, verify that the omnibus null hypothesis $H_0: \beta_1 = \beta_2 = 0$ can be tested as a general linear hypothesis, using the hypothesis matrix

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and right-hand-side vector $\mathbf{c} = [0, 0]'$. Then verify that the $H_0: \beta_1 = \beta_2$ can be tested using $\mathbf{L} = [0, 1, -1]$ and $\mathbf{c} = [0]$. (Cf. Exercise 6.7.)

Exercise 9.11. *Consider the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$. Show that the matrix \mathbf{V}_{11}^{-1} (see Equation 9.16 on page 218) for the slope coefficients β_1 and β_2 contains mean deviation sums of squares and products for the explanatory variables; that is,

$$\mathbf{V}_{11}^{-1} = \begin{bmatrix} \sum x_{i1}^{*2} & \sum x_{i1}^* x_{i2}^* \\ \sum x_{i1}^* x_{i2}^* & \sum x_{i2}^{*2} \end{bmatrix}$$

Now show, more generally, for the model $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i$, that the matrix \mathbf{V}_{11}^{-1} for the slope coefficients β_1, \dots, β_k contains mean deviation sums of squares and products for the explanatory variables.

Exercise 9.12. *Show that Equation 9.20 (page 222) for the confidence interval for β_1 can be written in the more conventional form

$$B_1 - t_{a,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}} \leq \beta_1 \leq B_1 + t_{a,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}}$$

Exercise 9.13. Using Figure 9.2 (on page 223), show how the confidence interval-generating ellipse can be used to derive a confidence interval for the difference of the parameters $\beta_1 - \beta_2$.

Compare the confidence interval for this linear combination with that for $\beta_1 + \beta_2$. Which combination of parameters is estimated more precisely? Why? What would happen if the regressors X_1 and X_2 were *negatively* correlated?

Exercise 9.14. Prediction: One use of a fitted regression equation is to *predict* response-variable values for particular “future” combinations of explanatory-variable scores. Suppose, therefore, that we fit the model $y = \mathbf{X}\beta + \varepsilon$, obtaining the least-squares estimate \mathbf{b} of β . Let $\mathbf{x}'_0 = [1, x_{01}, \dots, x_{0k}]$ represent a set of explanatory-variable scores for which a prediction is desired, and let Y_0 be the (generally unknown, or not yet known) corresponding value of Y . The explanatory-variable vector \mathbf{x}'_0 does not necessarily correspond to an observation in the sample for which the model was fit.

- (a) *If we use $\hat{Y}_0 = \mathbf{x}'_0\mathbf{b}$ to estimate $E(Y_0)$, then the error in estimation is $\delta \equiv \hat{Y}_0 - E(Y_0)$. Show that if the model is correct, then $E(\delta) = 0$ [i.e., \hat{Y}_0 is an unbiased estimator of $E(Y_0)$] and that $V(\delta) = \sigma^2_\varepsilon \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0$.
- (b) *We may be interested not in estimating the *expected* value of Y_0 but in predicting or forecasting the *actual* value $Y_0 = \mathbf{x}'_0\beta + \varepsilon_0$ that will be observed. The error in the forecast is then

$$D \equiv \hat{Y}_0 - Y_0 = \mathbf{x}'_0\mathbf{b} - (\mathbf{x}'_0\beta + \varepsilon_0) = \mathbf{x}'_0(\mathbf{b} - \beta) - \varepsilon_0$$

Show that $E(D) = 0$ and that $V(D) = \sigma^2_\varepsilon [1 + \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0]$. Why is the variance of the forecast error D greater than the variance of δ found in part (a)?

- (c) Use the results in parts (a) and (b), along with the Canadian occupational prestige regression (see Section 5.2.2), to predict the prestige score for an occupation with an average income of \$12,000, an average education of 13 years, and 50% women. Place a 90% confidence interval around the prediction, assuming (i) that you wish to estimate $E(Y_0)$ and (ii) that you wish to forecast an actual Y_0 score. (Because σ^2_ε is not known, you will need to use S_E^2 and the t -distribution.)
- (d) Suppose that the methods of this problem are used to forecast a value of Y for a combination of X 's very different from the X values in the data to which the model was fit. For example, calculate the estimated variance of the forecast error for an occupation with an average income of \$50,000, an average education of 0 years, and 100% women. Is the estimated variance of the forecast error large or small? Does the variance of the forecast error adequately capture the uncertainty in using the regression equation to predict Y in this circumstance?

Exercise 9.15. Suppose that the model matrix for the two-way ANOVA model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

is reduced to full rank by imposing the following constraints (for $r = 2$ rows and $c = 3$ columns):

$$\alpha_2 = 0$$

$$\beta_3 = 0$$

$$\gamma_{21} = \gamma_{22} = \gamma_{13} = \gamma_{23} = 0$$

These constraints imply dummy-variable (0/1) coding of the full-rank model matrix.⁷⁵

⁷⁵Cf. the discussion of dummy-coding in two-way ANOVA in Section 8.2.2.

- (a) Write out the row basis of the full-rank model matrix under these constraints.
- (b) Solve for the parameters of the constrained model in terms of the cell means. What is the nature of the hypotheses H_0 : all $\alpha_j = 0$ and H_0 : all $\beta_k = 0$ for this parametrization of the model? Are these hypotheses generally sensible?
- (c) Let $\text{SS}^*(\alpha, \beta, \gamma)$ represent the regression sum of squares for the full model, calculated under the constraints defined above; let $\text{SS}^*(\alpha, \beta)$ represent the regression sum of squares for the model that deletes the interaction regressors; and so on. Using the Moore and Krupat data (discussed in Section 8.2), confirm that

$$\text{SS}^*(\alpha|\beta) = \text{SS}(\alpha|\beta)$$

$$\text{SS}^*(\beta|\alpha) = \text{SS}(\beta|\alpha)$$

$$\text{SS}^*(\gamma|\alpha, \beta) = \text{SS}(\gamma|\alpha, \beta)$$

but that

$$\text{SS}^*(\alpha|\beta, \gamma) \neq \text{SS}(\alpha|\beta, \gamma)$$

$$\text{SS}^*(\beta|\alpha, \gamma) \neq \text{SS}(\beta|\alpha, \gamma)$$

where $\text{SS}(\cdot)$ and $\text{SS}(\cdot|\cdot)$ give regression and incremental sums of squares under the usual sigma constraints and deviation-coded $(1, 0, -1)$ regressors.

- (d) Analyze the Moore and Krupat data using one or more computer programs available to you. How do the programs calculate sums of squares in two-way ANOVA? Does the documentation accompanying the programs clearly explain how the sums of squares are computed?

Exercise 9.16. *Show that the equation-by-equation least-squares estimator $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the maximum-likelihood estimator of the regression coefficients \mathbf{B} in the multivariate general linear model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where the model matrix \mathbf{X} is fixed, and the distribution of the errors is $\varepsilon_i \sim \mathbf{N}_m(\mathbf{0}, \Sigma)$, with ε_i and $\varepsilon_{i'}$ independent for $i \neq i'$. Show that the MLE of the error-covariance matrix is $\frac{1}{n}\hat{\mathbf{E}}'\hat{\mathbf{E}}$, where $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{XB}$.

Exercise 9.17. Intention to treat: Recall the imaginary example in Section 9.8.1 in which students were randomly provided vouchers to defray the cost of attending a private school. In the text, we imagined that the researchers want to determine the effect of private versus public school attendance on academic achievement, and straightforward estimation of this effect is compromised by the fact that some students who received a voucher did not attend a private school, and some who did not receive a voucher nevertheless did attend a private school. We dealt with this problem by using provision of a voucher as an instrumental variable. How, if at all, would the situation change if the goal of the research were to determine the effect on achievement of *providing a voucher* rather than the effect of *actually attending* a private school? From a social-policy perspective, why might provision of a voucher be the explanatory variable of more direct interest? This kind of analysis is termed *intention to treat*.

Exercise 9.18. *The asymptotic covariance matrix of the IV estimator is⁷⁶

$$\mathcal{V} = \frac{1}{n} \text{plim} [n(\mathbf{b}_{IV} - \boldsymbol{\beta})(\mathbf{b}_{IV} - \boldsymbol{\beta})']$$

⁷⁶See online Appendix D.4 on asymptotic distribution theory.

The IV estimator itself (Equation 9.28) can be written as

$$\mathbf{b}_{IV} = \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}$$

(Reader: Why?) Then,

$$\mathcal{V} = \frac{1}{n} \text{plim} \left[n(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1} \right]$$

Starting with this formula, show that (repeating Equation 9.29)

$$\mathcal{V}(\mathbf{b}_{IV}) = \frac{\sigma_{\varepsilon}^2}{n} \boldsymbol{\Sigma}_{ZX}^{(R)-1} \boldsymbol{\Sigma}_{ZZ}^{(R)} \boldsymbol{\Sigma}_{XZ}^{(R)-1}$$

Caveat: This relatively simple derivation of $\mathcal{V}(\mathbf{b}_{IV})$ appears, for example, in Johnston (1972, Section 9-3), but it (although not the result itself) is technically flawed (see McCallum, 1973).

Exercise 9.19. Show that when the model matrix \mathbf{X} is used as the IV matrix \mathbf{Z} in instrumental-variables estimation, the IV and OLS estimators and their covariance matrices coincide. See Equations 9.28 and 9.29 (on page 233).

Exercise 9.20. Two-stage least-squares estimation:

- (a) Suppose that the column \mathbf{x}_1 in the model matrix \mathbf{X} also appears in the matrix \mathbf{Z} of instrumental variables in 2SLS estimation. Explain why $\hat{\mathbf{x}}_1$ in the first-stage regression simply reproduces \mathbf{x}_1 ; that is, $\hat{\mathbf{x}}_1 = \mathbf{x}_1$.
- (b) *Show that the two formulas for the 2SLS estimator (Equations 9.32 and 9.33 on page 235) are equivalent by demonstrating that $\hat{\mathbf{X}}'\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$.
- (c) *Show that when the number of instrumental variables in \mathbf{Z} is the same as the number of columns in the model matrix \mathbf{X} (i.e., $k+1$), the 2SLS estimator (Equation 9.32 on page 235) is equivalent to the direct IV estimator (Equation 9.28 on page 233). (*Hint:* It is probably simplest to demonstrate this result using the tools of the next chapter, on the vector geometry of linear models.)

Summary

- The general linear model can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{y} is an $n \times 1$ vector of response-variable observations; \mathbf{X} is an $n \times k+1$ matrix of regressors (called the model matrix), including an initial column of 1s for the constant regressor; $\boldsymbol{\beta}$ is a $k+1 \times 1$ vector of parameters to be estimated; and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. The assumptions of the linear model can be compactly written as $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I}_n)$.
- The model matrices for dummy-regression and ANOVA models are strongly patterned. In ANOVA, the relationship between group or cell means and the parameters of the linear model is expressed by the parametric equation $\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}_F$, where $\boldsymbol{\mu}$ is the vector of means, \mathbf{X}_B is the row basis of the full-rank model matrix, and $\boldsymbol{\beta}_F$ is the parameter vector associated with the full-rank model matrix. Solving the parametric equation for the parameters yields $\boldsymbol{\beta}_F = \mathbf{X}_B^{-1} \boldsymbol{\mu}$. Linear contrasts are regressors that are coded to incorporate specific hypotheses about the group means in the parameters of the model.
- If the model matrix \mathbf{X} is of full-column rank, then the least-squares coefficients are given by $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Under the full set of assumptions for the linear model,

$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1}]$. The least-squares estimator is also the most efficient unbiased estimator of $\boldsymbol{\beta}$ and the maximum-likelihood estimator of $\boldsymbol{\beta}$.

- The estimated covariance matrix of the least-squares coefficients is $\hat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1}$. The standard errors of the regression coefficients are the square-root diagonal entries of this matrix. Under the assumptions of the model, $(B_j - \beta_j)/SE(B_j) \sim t_{n-k-1}$, providing a basis for hypothesis tests and confidence intervals for individual coefficients.
- An incremental F -test for the hypothesis $H_0: \beta_1 = \dots = \beta_q = 0$, where $1 \leq q \leq k$, is given by

$$F_0 = \frac{(\text{RSS}_0 - \text{RSS}/q)}{\text{RSS}/(n - k - 1)}$$

where RSS is the residual sum of squares for the full model, and RSS_0 is the residual sum of squares for the model that deletes the q regressors corresponding to the parameters in H_0 . Under the null hypothesis, $F_0 \sim F_{q, n-k-1}$. The incremental F -statistic can also be computed directly as $F_0 = \mathbf{b}'_1 \mathbf{V}_{11}^{-1} \mathbf{b}_1 / q S_E^2$, where $\mathbf{b}_1 = [B_1, \dots, B_q]'$ contains the coefficients of interest extracted from among the entries of \mathbf{b} , and \mathbf{V}_{11} is the square submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$ consisting of the q rows and columns pertaining to the coefficients in \mathbf{b}_1 .

- The F -statistic

$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})}{q S_E^2}$$

is used to test the general linear hypothesis $H_0: \underset{(q \times k+1)}{\mathbf{L}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} = \underset{(q \times 1)}{\mathbf{c}}$, where the rank- q hypothesis matrix \mathbf{L} and right-hand-side vector \mathbf{c} contain prespecified constants. Under the hypothesis, $F_0 \sim F_{q, n-k-1}$.

- The joint confidence region for the q parameters $\boldsymbol{\beta}_1$, given by

$$\text{all } \boldsymbol{\beta}_1 \text{ for which } (\mathbf{b}_1 - \boldsymbol{\beta}_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \boldsymbol{\beta}_1) \leq q S_E^2 F_{a, q, n-k-1}$$

represents the combinations of values of these parameters that are jointly acceptable at the $1 - a$ level of confidence. The boundary of the joint confidence region is an ellipsoid in the q -dimensional parameter space, reflecting the correlational structure and dispersion of the X s.

- The multivariate linear model accommodates several response variables:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

Under the assumption that the rows $\boldsymbol{\varepsilon}_i'$ of the error matrix \mathbf{E} are independent and multivariately normally distributed with mean $\mathbf{0}$ and common nonsingular covariance matrix $\boldsymbol{\Sigma}$, the maximum-likelihood estimators of the regression coefficients are given by

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$$

Hypothesis tests for the multivariate linear model closely parallel those for the univariate linear model, with sum-of-squares-and-products (SSP) matrices in the multivariate case generalizing the role of sums of squares in the univariate case. Several commonly

employed test statistics are based on the eigenvalues of $\mathbf{SSP}_H \mathbf{SSP}_R^{-1}$, where \mathbf{SSP}_H is the SSP matrix for a hypothesis, and \mathbf{SSP}_R is the residual SSP matrix.

- The statistical theory of linear models, formulated under the supposition that the model matrix \mathbf{X} is fixed with respect to repeated sampling, is also valid when \mathbf{X} is random, as long as three additional requirements are satisfied:
 1. the model matrix \mathbf{X} is measured without error and is independent of the errors ε ;
 2. the distribution of \mathbf{X} , which is otherwise unconstrained, does not depend on the parameters β and σ_ε^2 of the linear model; and
 3. the covariance matrix of the X s is nonsingular.
- The omission of regressors from a linear model causes the coefficients of the included regressors to be inconsistent, unless
 1. the omitted regressors are uncorrelated with the included regressors or
 2. the omitted regressors have coefficients of 0 and hence are irrelevant.
- If (some of) the $k + 1$ columns of the model matrix \mathbf{X} are correlated with the error ε in the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, then the OLS estimator of β is inconsistent. Suppose, however, that there exists a matrix \mathbf{Z} of instrumental variables with $k + 1$ columns (some of which may be the same as columns of \mathbf{X}) that are uncorrelated with the error but correlated with \mathbf{X} . Then, $\mathbf{b}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ is a consistent estimator of β , with estimated asymptotic covariance matrix $\hat{\mathcal{V}}(\mathbf{b}_{IV}) = \hat{\sigma}_\varepsilon^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}$, where $\hat{\sigma}_\varepsilon^2 = \mathbf{e}'_{IV}\mathbf{e}_{IV}/(n - k - 1)$ and $\mathbf{e}_{IV} = \mathbf{y} - \mathbf{X}\mathbf{b}_{IV}$, as in OLS regression.
- If there are fewer instrumental variables in the IV matrix \mathbf{Z} than regression coefficients to estimate (corresponding to the $k + 1$ columns of model matrix \mathbf{X}), then the IV estimating equations are underdetermined, preventing us from solving uniquely for \mathbf{b}_{IV} . If, however, there are *more* instrumental variables than regression coefficients, then the IV estimating equations employed directly will be overdetermined. The surplus information available in the IVs is used efficiently by the two-stage least-squares estimator $\mathbf{b}_{2SLS} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$ where $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. The estimated asymptotic covariance matrix of the 2SLS estimator is $\hat{\mathcal{V}}(\mathbf{b}_{2SLS}) = \hat{\sigma}_\varepsilon^2(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{X}}(\mathbf{X}'\hat{\mathbf{X}})^{-1}$.

Recommended Reading

There are many texts that treat the theory of linear models more abstractly, more formally, and with greater generality than I have in this chapter.

- Seber (1977) is a reasonably accessible text that develops in a statistically more sophisticated manner most of the topics discussed in the last five chapters. Seber also pays more attention to issues of computation and develops some topics that I do not.
- Searle (1971) presents a very general treatment of linear models, including a much broader selection of ANOVA models, stressing the analysis of unbalanced data. Searle directly analyzes model matrices of less than full rank (elaborating the material in

Section 9.2.1), an approach that—in my opinion—makes the subject more complex than it needs to be. Despite its relative difficulty, however, the presentation is of exceptionally high quality.

- Hocking (1985) and Searle (1987) cover much the same ground as Searle (1971) but stress the use of “cell-means” models, avoiding some of the complications of overparametrized models for ANOVA. These books also contain a very general presentation of the theory of linear statistical models.
- A fine and accessible paper by Monette (1990) develops in more detail the geometric representation of regression analysis using ellipses (a topic that is usually treated only in difficult sources). Friendly, Monette, and Fox (2013) present a comprehensive overview of the role of elliptical geometry in statistics.
- There are many general texts on multivariate statistical methods. Krzanowski (1988) and Morrison (2005) provide wide-ranging and accessible introductions to the subject, including the multivariate linear model. The statistical theory of multivariate linear models is developed in detail by Anderson (2003) and Rao (1973).
- Instrumental-variables estimation is a standard topic in econometrics texts; see, for example, Greene (2003, Section 5.4).

10

The Vector Geometry of Linear Models*

As is clear from the previous chapter, linear algebra is the algebra of linear models. Vector geometry provides a spatial representation of linear algebra and therefore furnishes a powerful tool for understanding linear models. The geometric understanding of linear models is venerable: R. A. Fisher's (1915) development of the central notion of degrees of freedom in statistics was closely tied to vector geometry, for example.

Few points in this book are developed exclusively in geometric terms. The reader who takes the time to master the geometry of linear models, however, will find the effort worthwhile: Certain ideas—including degrees of freedom—are most simply developed or understood from the geometric perspective.¹

The chapter begins by describing the geometric vector representation of simple and multiple regression. Geometric vectors are then employed (in the spirit of Fisher's seminal paper) to explain the connection between degrees of freedom and unbiased estimation of the error variance in linear models. Finally, vector geometry is used to illuminate the essential nature of overparametrized analysis-of-variance (ANOVA) models.

10.1 Simple Regression

We can write the simple-regression model in vector form in the following manner:

$$\mathbf{y} = \alpha \mathbf{1}_n + \beta \mathbf{x} + \boldsymbol{\varepsilon} \quad (10.1)$$

where $\mathbf{y} \equiv [Y_1, Y_2, \dots, Y_n]'$, $\mathbf{x} \equiv [x_1, x_2, \dots, x_n]'$, $\boldsymbol{\varepsilon} \equiv [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]'$, and $\mathbf{1}_n \equiv [1, 1, \dots, 1]'$; α and β are the population regression coefficients.² As before, we will assume that $\boldsymbol{\varepsilon} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. The fitted regression equation is, similarly,

$$\mathbf{y} = A \mathbf{1}_n + B \mathbf{x} + \mathbf{e} \quad (10.2)$$

where $\mathbf{e} \equiv [E_1, E_2, \dots, E_n]'$ is the vector of residuals, and A and B are the least-squares regression coefficients. From Equation 10.1, we have

$$E(\mathbf{y}) = \alpha \mathbf{1}_n + \beta \mathbf{x}$$

Analogously, from Equation 10.2,

$$\hat{\mathbf{y}} = A \mathbf{1}_n + B \mathbf{x}$$

¹The basic vector geometry on which this chapter depends is developed in online Appendix B.

²Note that the X -values are treated as fixed. As in the previous chapter, the development of the vector geometry of linear models is simpler for fixed X , but the results apply as well when X is random.

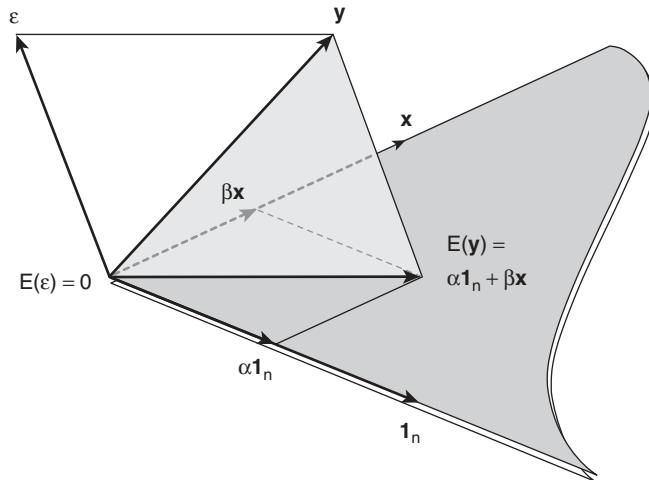


Figure 10.1 The vector geometry of the simple-regression model, showing the three-dimensional subspace spanned by the vectors \mathbf{x} , \mathbf{y} , and $\mathbf{1}_n$. Because the expected error is $\mathbf{0}$, the expected- \mathbf{Y} vector, $E(\mathbf{y})$, lies in the plane spanned by $\mathbf{1}_n$ and \mathbf{x} .

We are familiar with a seemingly natural geometric representation of $\{X, Y\}$ data—the scatterplot—in which the axes of a two-dimensional coordinate space are defined by the variables X and Y and where the observations are represented as points in the space according to their $\{x_i, Y_i\}$ coordinates. The scatterplot is a valuable data-analytic tool as well as a device for thinking about regression analysis.

I will now exchange the familiar roles of variables and observations, defining an n -dimensional coordinate space for which the *axes* are given by the *observations* and in which the *variables* are plotted as *vectors*. Of course, because there are generally many more than three observations, it is not possible to visualize the full vector space of the observations.³ Our interest, however, often inheres in two- and three-dimensional subspaces of this larger n -dimensional vector space. In these instances, as we will see presently, visual representation is both possible and illuminating. Moreover, the geometry of higher-dimensional subspaces can be grasped by analogy to the two- and three-dimensional cases.

The two-dimensional *variable space* (i.e., in which the variables define the axes) and the n -dimensional *observation space* (in which the observations define the axes) each contains a complete representation of the $(n \times 2)$ data matrix $[\mathbf{x}, \mathbf{y}]$. The formal duality of these spaces means that properties of the data, or of models meant to describe them, have equivalent representations in both spaces. Sometimes, however, the geometric representation of a property will be easier to understand in one space than in the other.

The simple-regression model of Equation 10.1 is shown geometrically in Figure 10.1. The subspace depicted in this figure is of dimension 3 and is spanned by the vectors \mathbf{x} , \mathbf{y} , and $\mathbf{1}_n$. Because \mathbf{y} is a vector random variable that varies from sample to sample, the vector diagram necessarily represents a *particular* sample. The other vectors shown in the diagram clearly lie in the subspace spanned by \mathbf{x} , \mathbf{y} , and $\mathbf{1}_n$: $E(\mathbf{y})$ is a linear combination of \mathbf{x} and $\mathbf{1}_n$ (and thus lies

³See Exercise 10.1 for a scaled-down, two-dimensional example, however.

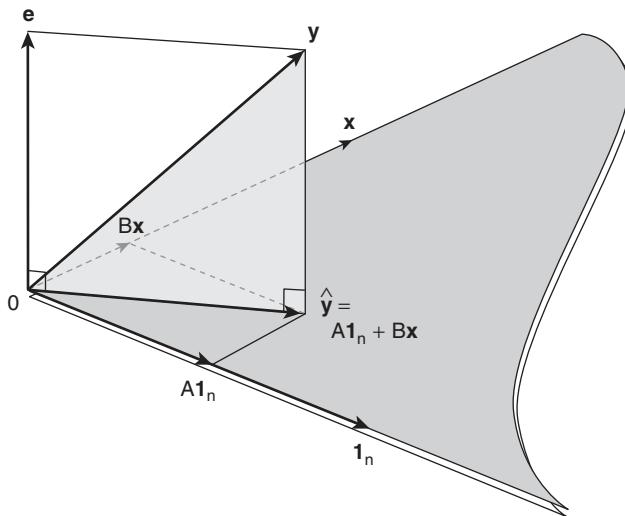


Figure 10.2 The vector geometry of least-squares fit in simple regression. Minimizing the residual sum of squares is equivalent to making the \mathbf{e} vector as short as possible. The $\hat{\mathbf{y}}$ vector is, therefore, the orthogonal projection of \mathbf{y} onto the $\{\mathbf{1}_n, \mathbf{x}\}$ plane.

in the $\{\mathbf{1}_n, \mathbf{x}\}$ plane), and the error vector ε is $\mathbf{y} - \alpha\mathbf{1}_n - \beta\mathbf{x}$. Although ε is nonzero in this sample, on average, over many samples, $E(\varepsilon) = \mathbf{0}$.

Figure 10.2 represents the least-squares simple regression of Y on X , for the same data as shown in Figure 10.1. The peculiar geometry of Figure 10.2 requires some explanation: We know that the fitted values $\hat{\mathbf{y}}$ are a linear combination of $\mathbf{1}_n$ and \mathbf{x} and hence lie in the $\{\mathbf{1}_n, \mathbf{x}\}$ plane. The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ has length $\|\mathbf{e}\| = \sqrt{\sum E_i^2}$ —that is, the square root of the residual sum of squares. The least-squares criterion interpreted geometrically, therefore, specifies that \mathbf{e} must be as short as possible. Because the length of \mathbf{e} is the distance between \mathbf{y} and $\hat{\mathbf{y}}$, this length is minimized by taking $\hat{\mathbf{y}}$ as the orthogonal projection of \mathbf{y} onto the $\{\mathbf{1}_n, \mathbf{x}\}$ plane, as shown in the diagram.

Variables, such as X and Y in simple regression, can be treated as vectors— \mathbf{x} and \mathbf{y} —in the n -dimensional space whose axes are given by the observations. Written in vector form, the simple-regression model is $\mathbf{y} = \alpha\mathbf{1}_n + \beta\mathbf{x} + \varepsilon$. The least-squares regression, $\mathbf{y} = \mathbf{A}\mathbf{1}_n + \mathbf{B}\mathbf{x} + \mathbf{e}$, is found by projecting \mathbf{y} orthogonally onto the plane spanned by $\mathbf{1}_n$ and \mathbf{x} , thus minimizing the sum of squared residuals, $\|\mathbf{e}\|^2$.

10.1.1 Variables in Mean Deviation Form

We can simplify the vector representation for simple regression by eliminating the constant regressor $\mathbf{1}_n$ and, with it, the intercept coefficient A . This simplification is worthwhile for two reasons:

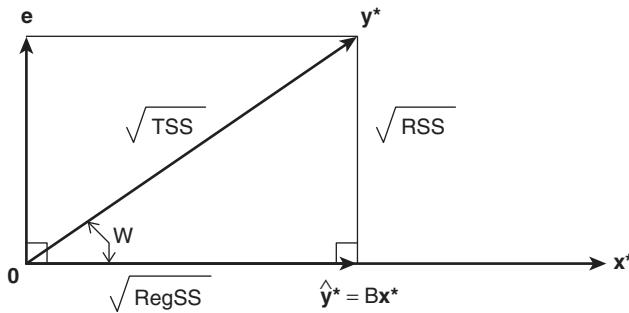


Figure 10.3 The vector geometry of least-squares fit in simple regression for variables in mean deviation form. The analysis of variance for the regression follows from the Pythagorean theorem. The correlation between X and Y is the cosine of the angle W separating the \mathbf{x}^* and \mathbf{y}^* vectors.

1. Our diagram is reduced from three to two dimensions. When we turn to multiple regression—introducing a second explanatory variable—eliminating the constant allows us to work with a three-dimensional rather than a four-dimensional subspace.
2. The ANOVA for the regression appears in the vector diagram when the constant is eliminated, as I will shortly explain.

To get rid of A , recall that $\bar{Y} = A + B\bar{x}$; subtracting this equation from the fitted model $\hat{Y}_i = A + Bx_i + E_i$ produces

$$\hat{Y}_i - \bar{Y} = B(x_i - \bar{x}) + E_i$$

Expressing the variables in mean deviation form eliminates the regression constant. Defining $\mathbf{y}^* \equiv \{Y_i - \bar{Y}\}$ and $\mathbf{x}^* \equiv \{x_i - \bar{x}\}$, the vector form of the fitted regression model becomes

$$\mathbf{y}^* = B\mathbf{x}^* + \mathbf{e} \quad (10.3)$$

The vector diagram corresponding to Equation 10.3 is shown in Figure 10.3. By the same argument as before,⁴ $\hat{Y}_i \equiv \{\hat{Y}_i - \bar{Y}\}$ is a multiple of \mathbf{x}^* , and the length of \mathbf{e} is minimized by taking \hat{Y}_i as the orthogonal projection of \mathbf{y}^* onto \mathbf{x}^* . Thus,

$$B = \frac{\mathbf{x}^* \cdot \mathbf{y}^*}{\|\mathbf{x}^*\|^2} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

which is the familiar formula for the least-squares slope in simple regression.⁵

Sums of squares appear on the vector diagram as the squared lengths of vectors. I have already remarked that

$$\text{RSS} = \sum E_i^2 = \|\mathbf{e}\|^2$$

⁴The mean deviations for the fitted values are $\{\hat{Y}_i - \bar{Y}\}$ because the mean of the fitted values is the same as the mean of Y . See Exercise 10.2.

⁵See Section 5.1.

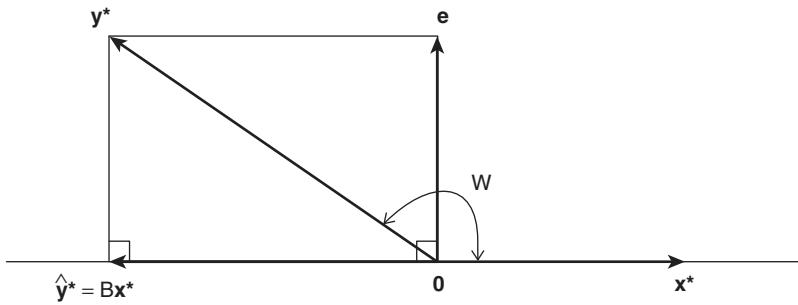


Figure 10.4 The vector geometry of least-squares fit for a negative relationship between X and Y . Here B is negative, so $\hat{\mathbf{y}} = B\mathbf{x}$ points in the direction opposite to \mathbf{x} .

Similarly,

$$\text{TSS} = \sum (Y_i - \bar{Y})^2 = \|\mathbf{y}^*\|^2$$

and

$$\text{RegSS} = \sum (\hat{Y}_i - \bar{Y})^2 = \|\hat{\mathbf{y}}^*\|^2$$

The ANOVA for the regression, $\text{TSS} = \text{RegSS} + \text{RSS}$, follows from the Pythagorean theorem.

The correlation coefficient is

$$r = \sqrt{\frac{\text{RegSS}}{\text{TSS}}} = \frac{\|\hat{\mathbf{y}}^*\|}{\|\mathbf{y}^*\|}$$

The vectors $\hat{\mathbf{y}}^*$ and \mathbf{y}^* are, respectively, the adjacent side and hypotenuse for the angle W in the right triangle whose vertices are given by the tips of $\mathbf{0}$, \mathbf{y}^* , and $\hat{\mathbf{y}}^*$. Thus, $r = \cos W$: The correlation between two variables (here, X and Y) is the cosine of the angle separating their mean deviation vectors. When this angle is 0, one variable is a perfect linear function of the other, and $r = \cos 0 = 1$. When the vectors are orthogonal, $r = \cos 90^\circ = 0$. We will see shortly that when two variables are *negatively* correlated, $90^\circ < W \leq 180^\circ$.⁶ The correlation $r = \cos W$ can be written directly as⁷

$$r = \frac{\mathbf{x}^* \cdot \mathbf{y}^*}{\|\mathbf{x}^*\| \|\mathbf{y}^*\|} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (Y_i - \bar{Y})^2}} \quad (10.4)$$

Figure 10.4 illustrates an inverse relationship between X and Y . All the conclusions that we based on Figure 10.3 still hold. Because B is now negative, $\hat{\mathbf{y}}^* = B\mathbf{x}^*$ is a negative multiple of the \mathbf{x}^* vector, pointing in the *opposite* direction from \mathbf{x}^* . The correlation is still the cosine of

⁶We need only consider angles between 0° and 180° for we can always examine the *smaller* of the two angles separating \mathbf{x}^* and \mathbf{y}^* . Because $\cos W = \cos(360^\circ - W)$, this convention is of no consequence.

⁷This is the alternative formula for the correlation coefficient presented in Section 5.1 (Equation 5.4 on page 90). The vector representation of simple regression, therefore, demonstrates the equivalence of the two formulas for r —the direct formula and the definition in terms of sums of squares.

W , but now we need to take the *negative* root of $\sqrt{\|\hat{\mathbf{y}}^*\|^2/\|\mathbf{y}^*\|^2}$ if we wish to define r in terms of vector lengths; Equation 10.4 produces the proper sign because $\mathbf{x}^* \cdot \mathbf{y}^*$ is negative.

Writing X and Y in mean deviation form, as the vectors \mathbf{x}^* and \mathbf{y}^* , eliminates the constant term and thus permits representation of the fitted regression in two (rather than three) dimensions: $\mathbf{y}^* = B\mathbf{x}^* + \mathbf{e}$. The ANOVA for the regression, $TSS = \text{RegSS} + \text{RSS}$, is represented geometrically as $\|\mathbf{y}^*\|^2 = \|\hat{\mathbf{y}}^*\|^2 + \|\mathbf{e}\|^2$. The correlation between X and Y is the cosine of the angle separating the vectors \mathbf{x}^* and \mathbf{y}^* .

10.1.2 Degrees of Freedom

The vector representation of simple regression helps clarify the concept of degrees of freedom. In general, sums of squares for linear models are the squared lengths of variable vectors. The degrees of freedom associated with a sum of squares represent the dimension of the subspace to which the corresponding vector is confined.

- Consider, first, the vector \mathbf{y} in Figure 10.2 (page 247): This vector can be located anywhere in the n -dimensional observation space. The *uncorrected* sum of squares $\sum Y_i^2 = \|\mathbf{y}\|^2$, therefore, has n degrees of freedom.
- When we convert Y to mean deviation form (as in Figure 10.3 on page 248), we confine the \mathbf{y}^* vector to an $(n - 1)$ -dimensional subspace, “losing” 1 degree of freedom in the process. This is easily seen for vectors in two-dimensional space: Let $\mathbf{y} = [Y_1, Y_2]'$, and $\mathbf{y}^* = [Y_1 - \bar{Y}, Y_2 - \bar{Y}]'$. Then, because $\bar{Y} = (Y_1 + Y_2)/2$, we can write

$$\mathbf{y}^* = \left[Y_1 - \frac{Y_1 + Y_2}{2}, Y_2 - \frac{Y_1 + Y_2}{2} \right]' = \left[\frac{Y_1 - Y_2}{2}, \frac{Y_2 - Y_1}{2} \right]' = [Y_1^*, -Y_1^*]'$$

Thus, all vectors \mathbf{y}^* lie on a line through the origin, as shown in Figure 10.5: The subspace of all vectors \mathbf{y}^* is one dimensional. Algebraically, by subtracting the mean from each of its coordinates, we have imposed a linear restriction on \mathbf{y}^* , ensuring that its entries sum to zero, $\sum(Y_i - \bar{Y}) = 0$; among the n values of $Y_i - \bar{Y}$, only $n - 1$ are linearly independent. The total sum of squares $TSS = \|\mathbf{y}^*\|^2 = \sum (Y_i - \bar{Y})^2$, therefore, has $n - 1$ degrees of freedom.

We can extend this reasoning to the residual and regression sums of squares:

- In Figure 10.3, $\hat{\mathbf{y}}^*$ is a multiple of \mathbf{x}^* . The vector \mathbf{x}^* , in turn, is fixed and spans a one-dimensional subspace. Because $\hat{\mathbf{y}}^*$ necessarily lies somewhere in this one-dimensional subspace, $\text{RegSS} = \|\hat{\mathbf{y}}^*\|^2$ has 1 degree of freedom.
- The degrees of freedom for the residual sum of squares can be determined from either Figure 10.2 or Figure 10.3. In Figure 10.2, \mathbf{y} lies somewhere in the n -dimensional observation space. The vectors \mathbf{x} and $\mathbf{1}_n$ are fixed and together span a subspace of dimension 2 within the larger observation space. The location of the residual vector \mathbf{e} depends

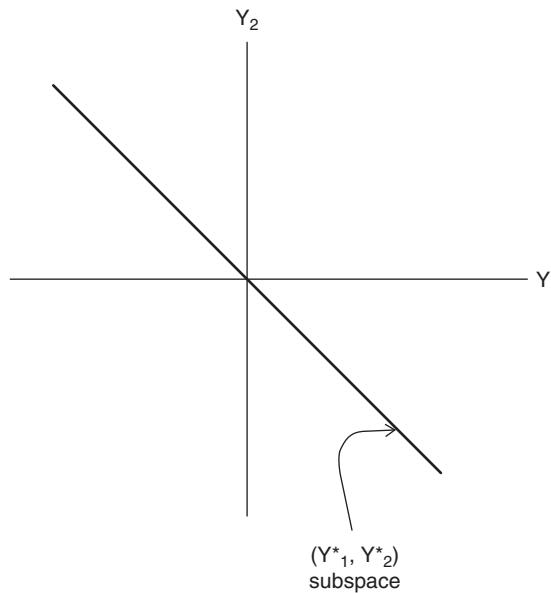


Figure 10.5 When $n = 2$, the mean deviation vector $\mathbf{y}^* = [Y_1 - \bar{Y}, Y_2 - \bar{Y}]'$ is confined to a one-dimensional subspace (i.e., a line) of the two-dimensional observation space.

on \mathbf{y} , but in any event, \mathbf{e} is orthogonal to the plane spanned by \mathbf{x} and $\mathbf{1}_n$. Consequently, \mathbf{e} lies in a subspace of dimension $n - 2$ (the orthogonal complement of the subspace spanned by \mathbf{x} and $\mathbf{1}_n$), and $\text{RSS} = \|\mathbf{e}\|^2$ has $n - 2$ degrees of freedom. Algebraically, the least-squares residuals \mathbf{e} satisfy two independent linear restrictions— $\sum E_i = 0$ (i.e., $\mathbf{e} \cdot \mathbf{1}_n = 0$) and $\sum E_i x_i = 0$ (i.e., $\mathbf{e} \cdot \mathbf{x} = 0$)—accounting for the “loss” of 2 degrees of freedom.⁸

- Alternatively, referring to Figure 10.3, \mathbf{y}^* lies in the $(n - 1)$ -dimensional subspace of mean deviations; the residual vector \mathbf{e} is orthogonal to \mathbf{x}^* , both of which also lie in the $(n - 1)$ -dimensional mean deviation subspace; hence, RSS has $(n - 1) - 1 = n - 2$ degrees of freedom.

Degrees of freedom in simple regression correspond to the dimensions of subspaces to which variable vectors associated with sums of squares are confined: (1) The \mathbf{y}^* vector lies in the $(n - 1)$ -dimensional subspace of mean deviations but is otherwise unconstrained; TSS, therefore, has $n - 1$ degrees of freedom. (2) The $\hat{\mathbf{y}}^*$ vector lies somewhere along the one-dimensional subspace spanned by \mathbf{x}^* ; RegSS, therefore, has 1 degree of freedom. (3) The \mathbf{e} vector lies in the $(n - 1)$ -dimensional subspace of mean deviations and is constrained to be orthogonal to \mathbf{x}^* ; RSS, therefore, has $(n - 1) - 1 = n - 2$ degrees of freedom.

⁸It is also the case that $\sum E_i \hat{Y}_i = \mathbf{e} \cdot \hat{\mathbf{y}} = 0$, but this constraint follows from the other two. See Exercise 10.3.

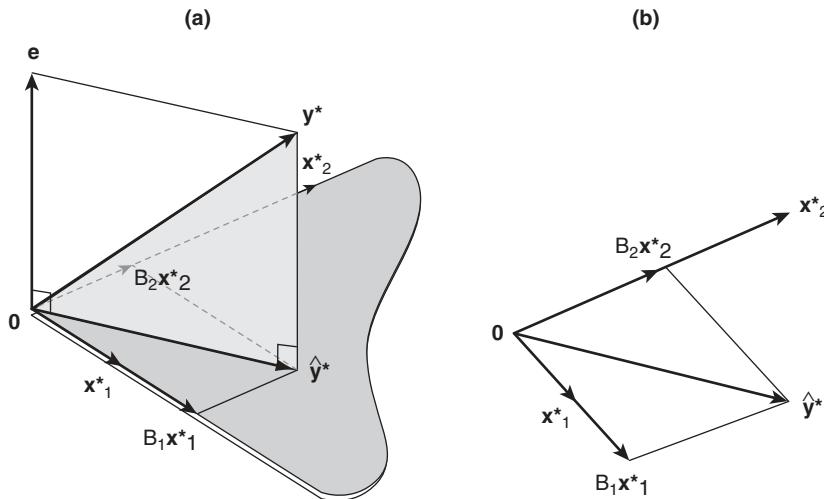


Figure 10.6 The vector geometry of least-squares fit in multiple regression, with the variables in mean deviation form. The vectors y^* , x_1^* , and x_2^* span a three-dimensional subspace, shown in (a). The fitted- Y vector, \hat{y}^* , is the orthogonal projection of y^* onto the plane spanned by x_1^* and x_2^* . The $\{x_1^*, x_2^*\}$ plane is shown in (b). In this illustration, both B_1 and B_2 are positive, with $B_1 > 1$ and $B_2 < 1$.

10.2 Multiple Regression

To develop the vector geometry of multiple regression, I will work primarily with the two-explanatory-variable model: Virtually all important points can be developed for this case, and by expressing the variables in mean deviation form (eliminating the constant regressor), the subspace of interest is confined to three dimensions and consequently can be visualized.

Consider, then, the fitted model

$$\mathbf{y} = A\mathbf{1}_n + B_1 \mathbf{x}_1 + B_2 \mathbf{x}_2 + \mathbf{e} \quad (10.5)$$

where \mathbf{y} is, as before, the vector of response-variable observations; \mathbf{x}_1 and \mathbf{x}_2 are explanatory-variable vectors; \mathbf{e} is the vector of residuals; and $\mathbf{1}_n$ is a vector of 1s. The least-squares regression coefficients are A , B_1 , and B_2 . From each observation of Equation 10.5, let us subtract $\bar{Y} = A + B_1 \bar{x}_1 + B_2 \bar{x}_2$, obtaining

$$\mathbf{y}^* = B_1 \mathbf{x}_1^* + B_2 \mathbf{x}_2^* + \mathbf{e} \quad (10.6)$$

In Equation 10.6, \mathbf{y}^* , \mathbf{x}_1^* , and \mathbf{x}_2^* are vectors of mean deviations.

Figure 10.6(a) shows the three-dimensional vector diagram for the fitted model of Equation 10.6, while Figure 10.6(b) depicts the explanatory-variable plane. The fitted values $\hat{y}^* = B_1 \mathbf{x}_1^* + B_2 \mathbf{x}_2^*$ are a linear combination of the regressors, and the vector \hat{y}^* , therefore, lies in the $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane. By familiar reasoning, the least-squares criterion implies that the residual vector \mathbf{e} is orthogonal to the explanatory-variable plane and, consequently, that \hat{y}^* is the orthogonal projection of \mathbf{y}^* onto this plane.

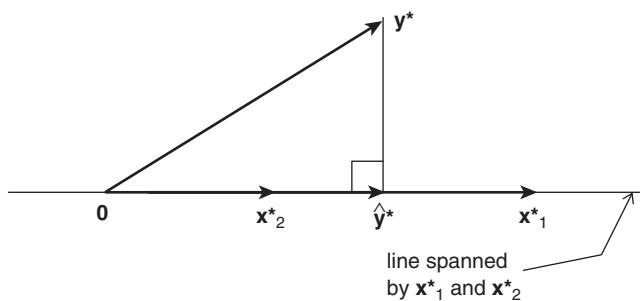


Figure 10.7 When the explanatory variables are perfectly collinear, x_1^* and x_2^* span a line rather than a plane. The \hat{y}^* vector can still be found by projecting y^* orthogonally onto this line, but the regression coefficients B_1 and B_2 , expressing \hat{y}^* as a linear combination of x_1^* and x_2^* , are not unique.

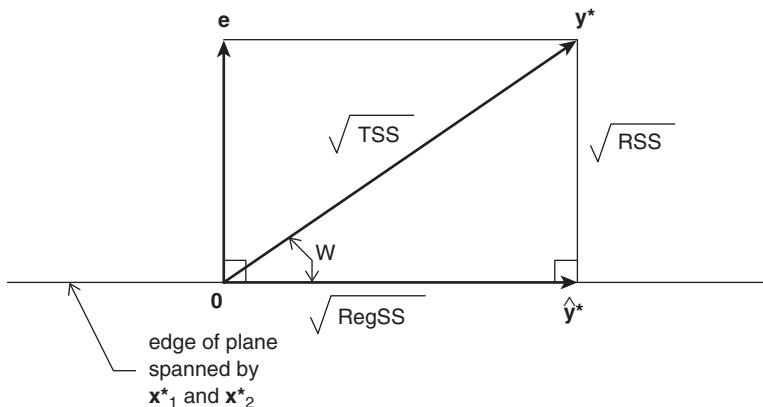


Figure 10.8 The analysis of variance for multiple regression appears in the plane spanned by y^* and \hat{y}^* . The multiple correlation R is the cosine of the angle W separating y^* and \hat{y}^* .

The regression coefficients B_1 and B_2 are uniquely defined as long as x_1^* and x_2^* are not collinear. This is the geometric version of the requirement that the explanatory variables may not be perfectly correlated. If the regressors are collinear, then they span a *line* rather than a plane; although we can still find the fitted values by orthogonally projecting y^* onto this line, as shown in Figure 10.7, we cannot express \hat{y}^* *uniquely* as a linear combination of x_1^* and x_2^* .

The ANOVA for the multiple-regression model appears in the plane spanned by y^* and \hat{y}^* , as illustrated in Figure 10.8. The residual vector also lies in this plane (because $e = y^* - \hat{y}$), while the regressor plane $\{x_1^*, x_2^*\}$ is perpendicular to it. As in simple-regression analysis, $TSS = \|y^*\|^2$, $RegSS = \|\hat{y}^*\|^2$, and $RSS = \|e\|^2$. The equation $TSS = RegSS + RSS$ follows from the Pythagorean theorem.

It is also clear from Figure 10.8 that $R = \sqrt{RegSS/TSS} = \cos W$. Thus, the *multiple* correlation is the *simple* correlation between the observed and fitted response-variable values, Y and \hat{Y} . If there is a *perfect* linear relationship between Y and the explanatory variables, then y^* lies in

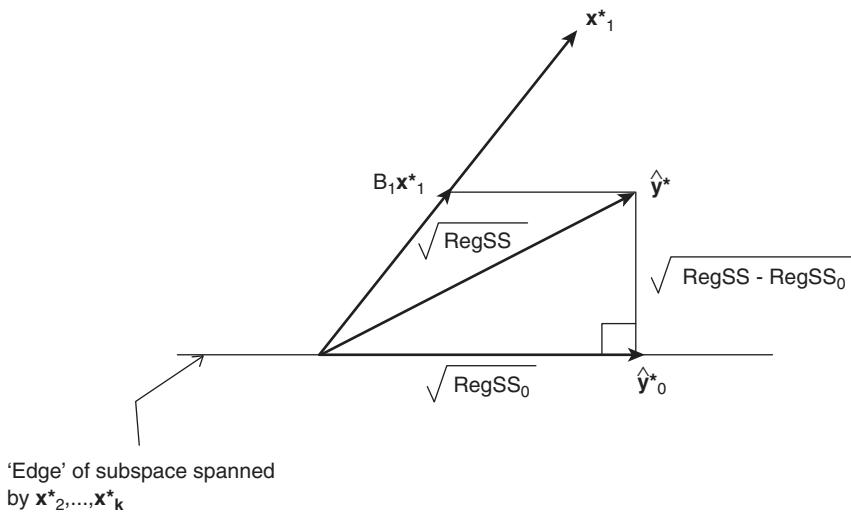


Figure 10.9 The incremental sum of squares for the hypothesis $H_0: \beta_1 = 0$. The vector \hat{y}_0^* is for the regression of Y on X_2, \dots, X_k (i.e., excluding X_1), while the vector \hat{y}^* is for the regression of Y on all the X s, including X_1 .

the regressor plane, $y^* = \hat{y}^*$, $e = \mathbf{0}$, $W = 0^\circ$, and $R = 1$; if, at the other extreme, there is *no* linear relationship between Y and the explanatory variables, then $y^* = e$, $\hat{y}^* = \mathbf{0}$, $W = 90^\circ$, and $R = 0$.

The fitted multiple-regression model for two explanatory variables is written in vector form as $\mathbf{y} = A\mathbf{1}_n + B_1\mathbf{x}_1 + B_2\mathbf{x}_2 + \mathbf{e}$. Putting Y and the X s in mean deviation form eliminates the constant, $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \mathbf{e}$, and permits a representation in three (rather than four) dimensions. The fitted values, $\hat{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^*$, are found by projecting \mathbf{y}^* orthogonally onto the plane spanned by \mathbf{x}_1^* and \mathbf{x}_2^* . The ANOVA for the regression, which is essentially the same as in simple regression, appears in the plane spanned by \mathbf{y}^* and \hat{y}^* . The multiple correlation R is the cosine of the angle separating \mathbf{y}^* and \hat{y}^* and, consequently, is the simple correlation between the observed and fitted Y values.

Figure 10.9 shows the vector geometry of the incremental F -test for the hypothesis $H_0: \beta_1 = 0$ in a model with k explanatory variables. RegSS —the regression sum of squares from the full model, where Y is regressed on all the X s—is decomposed into two orthogonal components: RegSS_0 (for the regression of Y on X_2, \dots, X_k) and the incremental sum of squares $\text{RegSS} - \text{RegSS}_0$.

The vector representation of regression analysis also helps clarify the relationship between simple and multiple regression. Figure 10.10(a) is drawn for two positively correlated regressors. The fitted response-variable vector is, from our previous work, the orthogonal projection of \mathbf{y}^* onto the $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane. To find the multiple-regression coefficient B_1 , we project \hat{y}^* parallel to \mathbf{x}_2^* , locating $B_1\mathbf{x}_1^*$, as shown in Figure 10.10(b), which depicts the regressor plane. The coefficient B_2 is located similarly.

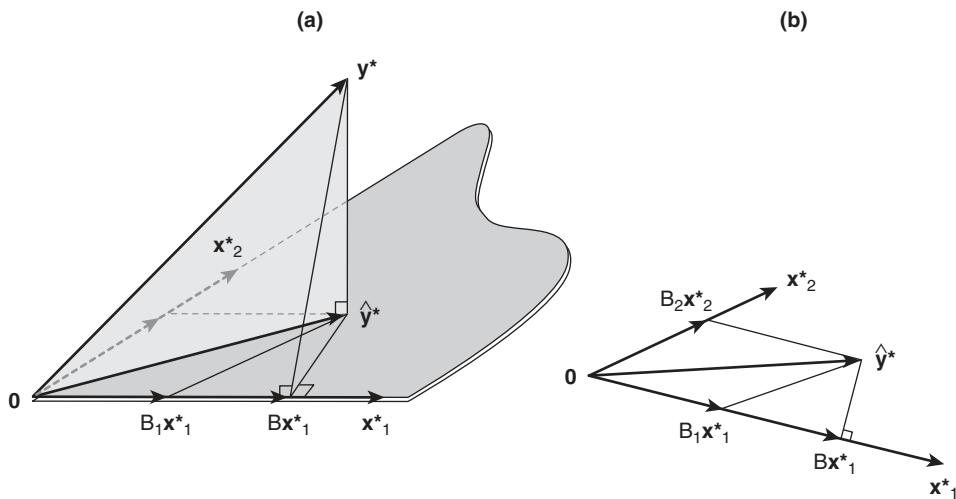


Figure 10.10 When the Xs are correlated (here positively), the slope B for the simple regression of Y on X_1 alone generally differs from the slope B_1 in the multiple regression of Y on both X_1 and X_2 . The least-squares fit is in (a), the regressor plane in (b).

To find the slope coefficient B for the *simple* regression of Y on X_1 , we need to project y^* onto x_1^* *alone*, obtaining Bx_1^* ; this result also appears in Figure 10.10(a). Because $x_1^* \cdot y^* = x_1^* \cdot \hat{y}^*$,⁹ the vector $B_1x_1^*$ is also the orthogonal projection of \hat{y}^* onto x_1^* , as shown in Figure 10.10(a) and (b). In this instance, projecting \hat{y}^* perpendicular to x_1^* (simple regression) rather than parallel to x_2^* (multiple regression) causes the simple-regression slope B to exceed the multiple-regression slope B_1 .

The situation changes fundamentally if the explanatory variables X_1 and X_2 are *uncorrelated*, as illustrated in Figure 10.11(a) and (b). Here, $B = B_1$. Another advantage of orthogonal regressors is revealed in Figure 10.11(b): There is a unique partition of the regression sum of squares into components due to each of the two regressors. We have¹⁰

$$\text{RegSS} = \hat{y}^* \cdot \hat{y}^* = B_1^2 x_1^* \cdot x_1^* + B_2^2 x_2^* \cdot x_2^*$$

In contrast, when the regressors are *correlated*, as in Figure 10.10(b), no such partition is possible, for then

$$\text{RegSS} = \hat{y}^* \cdot \hat{y}^* = B_1^2 x_1^* \cdot x_1^* + B_2^2 x_2^* \cdot x_2^* + 2B_1 B_2 x_1^* \cdot x_2^* \quad (10.7)$$

The last term in Equation 10.7 can be positive or negative, depending on the signs of the regression coefficients and of the correlation between X_1 and X_2 .¹¹

⁹See Exercise 10.5.

¹⁰See Exercise 10.6.

¹¹Occasionally, $2B_1 B_2 x_1^* \cdot x_2^*$ is interpreted as the variation in Y due to the “overlap” between the correlated explanatory variables X_1 and X_2 , and one may even see the terms in Equation 10.7 represented as areas in a Venn diagram. That this interpretation (and associated Venn diagram representation) is nonsense follows from the observation that the “overlap” can be negative.

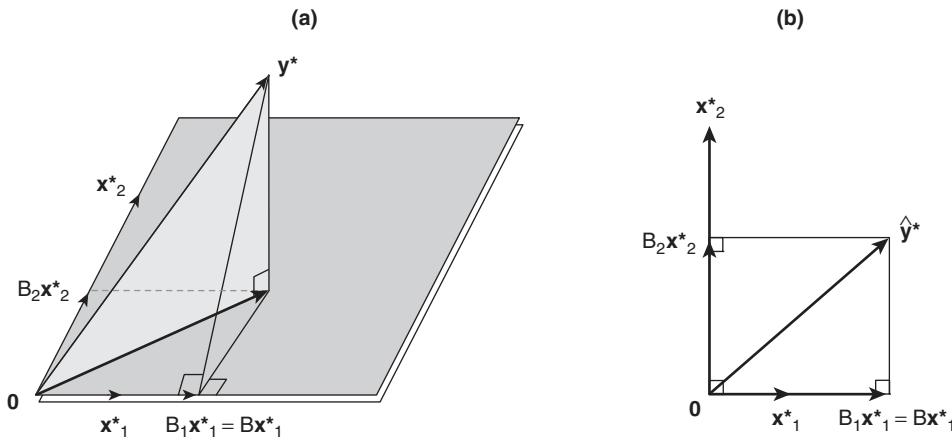


Figure 10.11 When the X s are uncorrelated, the simple-regression slope B and the multiple-regression slope B_1 are the same. The least-squares fit is in (a), the regressor plane in (b).

When the explanatory variables in multiple regression are orthogonal (uncorrelated), the regression sum of squares can be partitioned into components due to each explanatory variable: $\|\hat{y}^*\|^2 = B_1^2\|\mathbf{x}_1^*\|^2 + B_2^2\|\mathbf{x}_2^*\|^2$. When the explanatory variables are correlated, however, no such partition is possible.

As in simple regression, degrees of freedom in multiple regression correspond to the dimension of subspaces of the observation space. Because the \mathbf{y}^* vector, as a vector of mean deviations, is confined to a subspace of dimension $n - 1$, there are $n - 1$ degrees of freedom for TSS $= \|\mathbf{y}^*\|^2$. The fitted-value vector $\hat{\mathbf{y}}^*$ necessarily lies in the fixed $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane, which is a subspace of dimension 2; thus, RegSS $= \|\hat{\mathbf{y}}^*\|^2$ has 2 degrees of freedom. Finally, the residual vector \mathbf{e} is orthogonal to the explanatory-variable plane, and, therefore, RSS $= \|\mathbf{e}\|^2$ has $(n - 1) - 2 = n - 3$ degrees of freedom.

More generally, k noncollinear regressors in mean deviation form generate a subspace of dimension k . The fitted response-variable vector $\hat{\mathbf{y}}^*$ is the orthogonal projection of \mathbf{y}^* onto this subspace, and, therefore, RegSS has k degrees of freedom. Likewise, because \mathbf{e} is orthogonal to the k -dimensional regressor subspace, RSS has $(n - 1) - k = n - k - 1$ degrees of freedom.

As in simple regression, degrees of freedom in multiple regression follow from the dimensionality of the subspaces to which the \mathbf{y}^* , $\hat{\mathbf{y}}^*$, and \mathbf{e} vectors are confined.

10.3 Estimating the Error Variance

The connection between degrees of freedom and unbiased variance estimation is subtle but yields *relatively* simply to the geometric point of view. This section uses the vector geometry

of regression to show that $S_E^2 = \sum E_i^2 / (n - k - 1)$ is an unbiased estimator of the error variance, σ_ε^2 .

Even when the errors in a linear model are independent and normally distributed with zero means and constant variance, $\varepsilon \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$, the least-squares residuals are correlated and generally have different variances, $\mathbf{e} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{Q})$. The matrix $\mathbf{Q} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is non-diagonal, singular, and of rank $n - k - 1$.¹²

Following Putter (1967), we can transform the least-squares residuals into an independent and identically distributed set by selecting an orthonormal basis for the error subspace, defining transformed residuals in the following manner:

$$\underset{(n-k-1 \times 1)}{\mathbf{z}} \equiv \underset{(n-k-1 \times n)}{\mathbf{G}} \underset{(n \times 1)}{\mathbf{e}}$$

The transformation matrix \mathbf{G} is selected so that it is orthonormal and orthogonal to \mathbf{X} :

$$\begin{aligned} \mathbf{G}\mathbf{G}' &= \mathbf{I}_{n-k-1} \\ \mathbf{G}\mathbf{X} &= \underset{(n-k-1 \times k+1)}{\mathbf{0}} \end{aligned}$$

The transformed residuals then have the following properties:¹³

$$\begin{aligned} \mathbf{z} &= \mathbf{Gy} \\ E(\mathbf{z}) &= \mathbf{0} \\ V(\mathbf{z}) &= \sigma_\varepsilon^2 \mathbf{I}_{n-k-1} \end{aligned}$$

If the elements of ε are independent and normally distributed with constant variance, then so are the elements of \mathbf{z} . There are, however, n of the former and $n - k - 1$ of the latter. Furthermore, the transformation matrix \mathbf{G} (and hence \mathbf{z}) is not unique—there are infinitely many ways of selecting an orthonormal basis for the error subspace.¹⁴

Transforming \mathbf{e} to \mathbf{z} suggests a simple method for deriving an estimator of the error variance σ_ε^2 . The entries of \mathbf{z} have zero expectations and common variance σ_ε^2 , so

$$E(\mathbf{z}'\mathbf{z}) = \sum_{i=1}^{n-k-1} E(Z_i^2) = (n - k - 1)\sigma_\varepsilon^2$$

Thus, an unbiased estimator of the error variance is given by

$$S_E^2 \equiv \frac{\mathbf{z}'\mathbf{z}}{n - k - 1}$$

Moreover, because the Z_i are independent and normally distributed,

$$\frac{\mathbf{z}'\mathbf{z}}{\sigma_\varepsilon^2} = \frac{(n - k - 1)S_E^2}{\sigma_\varepsilon^2} \sim \chi_{n-k-1}^2$$

The estimator S_E^2 can be computed *without* finding transformed residuals, for the length of the least-squares residual vector \mathbf{e} is the same as the length of the vector of transformed residuals \mathbf{z} ; that is, $\sqrt{\mathbf{e}'\mathbf{e}} = \sqrt{\mathbf{z}'\mathbf{z}}$. This result follows from the observation that \mathbf{e} and \mathbf{z} are the *same* vector represented according to alternative bases: (1) \mathbf{e} gives the coordinates of the residuals relative to

¹²See Exercise 10.10.

¹³See Exercise 10.11.

¹⁴The transformed residuals are useful not only for exploring properties of least-squares estimation but also in diagnosing certain linear-model problems (see, e.g., Putter, 1967; Theil, 1971, chap. 5).

the natural basis of the n -dimensional observation space; (2) \mathbf{z} gives the coordinates of the residuals relative to an arbitrary orthonormal basis for the $(n - k - 1)$ -dimensional error subspace. A vector does not change its length when the basis changes, and, therefore,

$$S_E^2 = \frac{\mathbf{z}'\mathbf{z}}{n - k - 1} = \frac{\mathbf{e}'\mathbf{e}}{n - k - 1}$$

which is our usual estimator of the error variance.

Heuristically, although \mathbf{e} contains n elements, there are, as I have explained, $k + 1$ linear dependencies among them. In calculating an unbiased estimator of the error variance, we need to divide by the residual degrees of freedom rather than by the number of observations.

An unbiased estimator of the error variance σ_ε^2 can be derived by transforming the n correlated residuals \mathbf{e} to $n - k - 1$ independently and identically distributed residuals \mathbf{z} , employing an orthonormal basis \mathbf{G} for the $(n - k - 1)$ -dimensional error subspace: $\mathbf{z} = \mathbf{Ge}$. If the errors are independent and normally distributed, with zero means and common variance σ_ε^2 , then so are the elements of \mathbf{z} . Thus, $\mathbf{z}'\mathbf{z}/(n - k - 1)$ is an unbiased estimator of the error variance, and because \mathbf{z} and \mathbf{e} are the same vector represented according to alternative bases, $\mathbf{z}'\mathbf{z}/(n - k - 1) = \mathbf{e}'\mathbf{e}/(n - k - 1)$, which is our usual estimator of error variance, S_E^2 .

10.4 Analysis-of-Variance Models

Recall the overparametrized one-way ANOVA model¹⁵

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad \text{for } i = 1, \dots, n_j; j = 1, \dots, m$$

The \mathbf{X} matrix for this model (with parameters labeling the columns) is

$$\mathbf{X}_{(n \times m+1)} = \begin{bmatrix} (\mu) & (\alpha_1) & (\alpha_2) & \cdots & (\alpha_{m-1}) & (\alpha_m) \\ \hline 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 1 & 0 & \cdots & 0 & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

¹⁵See Section 8.1.

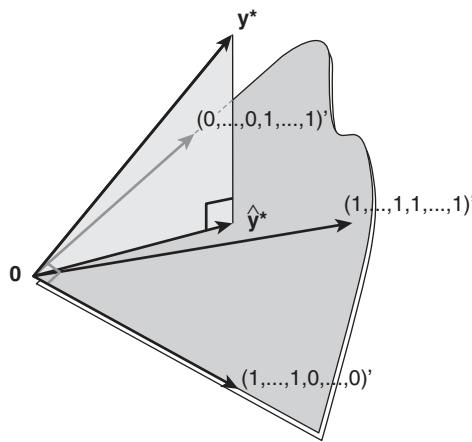


Figure 10.12 The vector geometry of least-squares fit for the overparametrized one-way ANOVA model when there are two groups. The $m+1 = 3$ columns of the model matrix are collinear and span a subspace of dimension $m = 2$.

The $m + 1$ columns of the model matrix span a subspace of dimension m . We can project the response-variable vector \mathbf{y} onto this subspace, locating the fitted-value vector $\hat{\mathbf{y}}$. Because they are collinear, the columns of \mathbf{X} do not provide a basis for the subspace that they span, and, consequently, the individual parameter estimates are not uniquely determined. This situation is illustrated in Figure 10.12 for $m = 2$. Even in the absence of uniquely determined parameters, however, we have no trouble calculating the regression sum of squares for the model because we can find $\hat{\mathbf{y}}$ by picking an arbitrary basis for the column space of \mathbf{X} . The dummy-coding and deviation-coding schemes of Chapter 8 select alternative bases for the column space of the model matrix: Dummy coding simply deletes the last column to provide a basis for the column space of \mathbf{X} ; deviation coding constructs a new basis for the column space of \mathbf{X} .

In the overparametrized one-way ANOVA model, $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, the $m + 1$ columns of the model matrix \mathbf{X} are collinear and span a subspace of dimension m . We can, however, still find $\hat{\mathbf{y}}$ for the model by projecting \mathbf{y} orthogonally onto this subspace, most simply by selecting an arbitrary basis for the column space of the model matrix. Conceived in this light, dummy coding and deviation coding are two techniques for constructing a basis for the column space of \mathbf{X} .

Let us turn next to the overparametrized two-way ANOVA model:¹⁶

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk} \quad \text{for } i = 1, \dots, n_{jk}; j = 1, \dots, r; k = 1, \dots, c$$

We will consider the simplest case, where $j = k = 2$. It suffices to examine the parametric equation for the model, relating the four cell means μ_{jk} to the nine parameters of the model:

¹⁶See Section 8.2.

$$\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} = \left[\begin{array}{c|cc|cc|cccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{array} \right] \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix}$$

$\boldsymbol{\mu} = \mathbf{X}_B \boldsymbol{\beta}$

Note that the four columns in \mathbf{X}_B representing the interactions are linearly independent, and hence the corresponding columns of \mathbf{X} span the *full* column space of the model matrix. The subspaces spanned by the main effects, each consisting of two linearly independent columns, lie in the space spanned by the interaction regressors—the main-effect subspaces are literally *marginal to* (i.e., contained in) the interaction subspace. Finally, the constant regressor is marginal both to the interaction subspace and to each of the main-effect subspaces: The constant regressor is simply the sum of the interaction regressors or of either set of main-effect regressors. Understood in this light, the deviation-coding method of Chapter 8 selects a convenient full-rank basis for the model matrix.

In the overparametrized two-way ANOVA model, $Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$, the interaction regressors provide a basis for the full column space of the model matrix. The model-matrix columns for the two sets of main effects are therefore marginal to (i.e., subspaces of) the interaction space. The column for the constant regressor is marginal to the main-effect subspaces as well as to the interactions.

Exercises

Exercise 10.1. Here is a very small (contrived) data set with two variables and two observations:

Observations	Variables	
	X	Y
1	1	2
2	3	5

Construct a scatterplot for the two observations in the $\{X, Y\}$ variable space, and then construct a vector diagram showing \mathbf{x} and \mathbf{y} in the observation space.

Exercise 10.2. Show that the average fitted value, $\bar{\hat{Y}}$, is the same as the average response-variable value, \bar{Y} . [Hint: Form the sum, $\sum Y_i = \sum(\hat{Y}_i + E_i)$.]

Exercise 10.3. *Show that the constraints $\mathbf{e} \cdot \mathbf{x} = 0$ and $\mathbf{e} \cdot \mathbf{1}_n = 0$ imply that $\mathbf{e} \cdot \hat{\mathbf{y}} = 0$. (Hint: $\hat{\mathbf{y}}$ lies in the plane spanned by \mathbf{x} and $\mathbf{1}_n$.)

Exercise 10.4. Using Duncan's occupational prestige data (discussed, e.g., in Chapter 5), construct the geometric vector representation for the regression of prestige on education, showing the \mathbf{x}^* , \mathbf{y}^* , $\hat{\mathbf{y}}^*$, and \mathbf{e} vectors drawn to scale. Find the angle between \mathbf{x}^* and \mathbf{y}^* .

Exercise 10.5. Prove that $\mathbf{x}_1^* \cdot \mathbf{y}^* = \mathbf{x}_1^* \cdot \hat{\mathbf{y}}^*$. (Hint: $\mathbf{y}^* = \hat{\mathbf{y}}^* + \mathbf{e}$, and \mathbf{e} is orthogonal to \mathbf{x}_1^* .)

Exercise 10.6. Show that when X_1 and X_2 are uncorrelated, the regression sum of squares can be written as

$$\text{RegSS} = \hat{\mathbf{y}}^* \cdot \hat{\mathbf{y}}^* = B_1^2 \mathbf{x}_1^* \cdot \mathbf{x}_1^* + B_2^2 \mathbf{x}_2^* \cdot \mathbf{x}_2^*$$

(Hint: Use $\hat{y}^* = B_1 \mathbf{x}_1^* + B_2 \mathbf{x}_2^*$.)

Exercise 10.7. Exercise 10.4 (continued): Using Duncan's occupational prestige data, construct the geometric representation for the regression of prestige Y on income X_1 and education X_2 . Draw separate graphs for (a) the $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane, showing the $\hat{\mathbf{y}}^*$ vector, B_1 , and B_2 , and (b) the $\{\mathbf{y}^*, \hat{\mathbf{y}}^*\}$ plane, showing \mathbf{e} . Draw all vectors to scale. (Hint: Calculate the correlation between X_1 and X_2 to find the angle between \mathbf{x}_1^* and \mathbf{x}_2^* .)

Exercise 10.8. Nearly collinear regressors: Construct the geometric vector representation of a regression with two explanatory variables in mean deviation form, $\hat{\mathbf{y}}^* = B_1 \mathbf{x}_1^* + B_2 \mathbf{x}_2^*$, distinguishing between two cases: (a) X_1 and X_2 are highly correlated, so that the angle separating the \mathbf{x}_1^* and \mathbf{x}_2^* vectors is small, and (b) X_1 and X_2 are uncorrelated, so that the \mathbf{x}_1^* and \mathbf{x}_2^* vectors are orthogonal. By examining the regressor plane, show that slight changes in the position of the $\hat{\mathbf{y}}^*$ vector (due, e.g., to sampling fluctuations) can cause dramatic changes in the regression coefficients B_1 and B_2 in case (a) but not in case (b). The problem of collinearity is discussed further in Chapter 13.

Exercise 10.9. Partial correlation (see Exercise 5.8):

- (a) Illustrate how the partial correlation $r_{Y1|2}$ can be represented using geometric vectors. Draw the vectors \mathbf{y}^* , \mathbf{x}_1^* , and \mathbf{x}_2^* , and define $\mathbf{e}_1 \equiv \{E_{i1|2}\}$ and $\mathbf{e}_Y \equiv \{E_{iY|2}\}$ (where i is the subscript for observations).
- (b) *Use the vector diagram in part (a) to show that the incremental F -test for the hypothesis $H_0: \beta_1 = 0$ can be written as

$$F_0 = \frac{(n - k - 1)r_{Y1|2}^2}{1 - r_{Y1|2}^2}$$

Recalling part (b) of Exercise 5.8, why is this result intuitively plausible?

Exercise 10.10. *Show that the matrix $\mathbf{Q} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (see Section 10.3) is nondiagonal, singular, and of rank $n - k - 1$. (Hints: Verify that the rows of \mathbf{Q} satisfy the $k + 1$ constraints implied by $\mathbf{Q}\mathbf{X} = \mathbf{0}$. If \mathbf{Q} is singular and diagonal, then some of its diagonal entries must be 0; show that this is not generally the case.)

Exercise 10.11. *Prove that when the least-squares residuals are transformed according to the equation $\mathbf{z} = \mathbf{Ge}$, where the $n - k - 1 \times n$ transformation matrix \mathbf{G} is orthonormal and orthogonal to \mathbf{X} , the transformed residuals \mathbf{z} have the following properties: $\mathbf{z} = \mathbf{Gy}$, $E(\mathbf{z}) = \mathbf{0}$, and $V(\mathbf{z}) = \sigma_e^2 \mathbf{I}_{n-k-1}$. (See Section 10.3.)

Exercise 10.12. Exercise 9.15 (continued): Let $\text{SS}(\cdot)$ give sums of squares for the two-way ANOVA model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

using deviation-coded regressors (i.e., employing sigma constraints to reduce the model matrix to full rank), and let $\text{SS}^*(\cdot)$ give sums of squares for the same model using dummy-coded regressors. Working with the model for $r = 2$ rows and $c = 3$ columns, use the geometric vector representation of the model to explain why

$$\begin{aligned}\text{SS}^*(\alpha|\beta) &= \text{SS}(\alpha|\beta) \\ \text{SS}^*(\beta|\alpha) &= \text{SS}(\beta|\alpha) \\ \text{SS}^*(\gamma|\alpha, \beta) &= \text{SS}(\gamma|\alpha, \beta)\end{aligned}$$

but that, in general,

$$\begin{aligned}\text{SS}^*(\alpha|\beta, \gamma) &\neq \text{SS}(\alpha|\beta, \gamma) \\ \text{SS}^*(\beta|\alpha, \gamma) &\neq \text{SS}(\beta|\alpha, \gamma)\end{aligned}$$

[*Hints:* Show that (i) the subspaces spanned by the deviation and dummy regressors for each of the two sets of main effects are the same, (ii) the subspaces spanned by the deviation and dummy regressors for the full set of effects (main effects and interactions) are the same, but (iii) the subspaces spanned by the deviation and dummy interaction regressors are *different*.]

Summary

- Variables, such as X and Y in simple regression, can be treated as vectors— \mathbf{x} and \mathbf{y} —in the n -dimensional space whose axes are given by the observations. Written in vector form, the simple-regression model is $\mathbf{y} = \alpha \mathbf{1}_n + \beta \mathbf{x} + \mathbf{e}$. The least-squares regression, $\hat{\mathbf{y}} = A \mathbf{1}_n + B \mathbf{x} + \mathbf{e}$, is found by projecting \mathbf{y} orthogonally onto the plane spanned by $\mathbf{1}_n$ and \mathbf{x} , thus minimizing the sum of squared residuals $\|\mathbf{e}\|^2$.
- Writing X and Y in mean deviation form, as the vectors \mathbf{x}^* and \mathbf{y}^* , eliminates the constant term and thus permits representation of the fitted regression in two (rather than three) dimensions: $\hat{\mathbf{y}}^* = B \mathbf{x}^* + \mathbf{e}$. The ANOVA for the regression, $\text{TSS} = \text{RegSS} + \text{RSS}$, is represented geometrically as $\|\mathbf{y}^*\|^2 = \|\hat{\mathbf{y}}^*\|^2 + \|\mathbf{e}\|^2$. The correlation between X and Y is the cosine of the angle separating the vectors \mathbf{x}^* and \mathbf{y}^* .
- Degrees of freedom in simple regression correspond to the dimensions of subspaces to which variable vectors associated with sums of squares are confined:
 - The \mathbf{y}^* vector lies in the $(n - 1)$ -dimensional subspace of mean deviations but is otherwise unconstrained; $\text{TSS} = \|\mathbf{y}^*\|^2$, therefore, has $n - 1$ degrees of freedom.
 - The $\hat{\mathbf{y}}^*$ vector lies somewhere along the one-dimensional subspace spanned by \mathbf{x}^* ; $\text{RegSS} = \|\hat{\mathbf{y}}^*\|^2$, therefore, has 1 degree of freedom.

- The \mathbf{e} vector lies in the $(n - 1)$ -dimensional subspace of mean deviations and is constrained to be orthogonal to \mathbf{x}^* ; RSS = $\|\mathbf{e}\|^2$, therefore, has $(n - 1) - 1 = n - 2$ degrees of freedom.
- The fitted multiple-regression model for two explanatory variables is written in vector form as $\mathbf{y} = A\mathbf{1}_n + B_1\mathbf{x}_1 + B_2\mathbf{x}_2 + \mathbf{e}$. Putting Y and the X 's in mean deviation form eliminates the constant, $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \mathbf{e}$, and permits a representation in three (rather than four) dimensions. The fitted values, $\hat{\mathbf{y}}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^*$, are found by projecting \mathbf{y}^* orthogonally onto the plane spanned by \mathbf{x}_1^* and \mathbf{x}_2^* . The ANOVA for the regression, which is essentially the same as in simple regression, appears in the plane spanned by \mathbf{y}^* and $\hat{\mathbf{y}}^*$. The multiple correlation R is the cosine of the angle separating \mathbf{y}^* and $\hat{\mathbf{y}}^*$, and, consequently, is the simple correlation between the observed and fitted Y -values.
- When the explanatory variables in multiple regression are orthogonal (uncorrelated), the regression sum of squares can be partitioned into components due to each explanatory variable: $\|\hat{\mathbf{y}}^*\|^2 = B_1^2\|\mathbf{x}_1^*\|^2 + B_2^2\|\mathbf{x}_2^*\|^2$. When the explanatory variables are correlated, however, no such partition is possible.
- As in simple regression, degrees of freedom in multiple regression follow from the dimensionality of the subspaces to which the various vectors are confined.
 - The \mathbf{y}^* vector lies in the $(n - 1)$ -dimensional subspace of mean deviations; TSS, therefore, has $n - 1$ degrees of freedom.
 - The $\hat{\mathbf{y}}^*$ vector lies somewhere in the plane spanned by \mathbf{x}_1^* and \mathbf{x}_2^* ; RegSS, therefore, has 2 degrees of freedom. More generally, k explanatory variables $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_k^*$ span a subspace of dimension k , and $\hat{\mathbf{y}}^*$ is the orthogonal projection of \mathbf{y}^* onto this subspace; thus, RegSS has k degrees of freedom.
 - The \mathbf{e} vector is constrained to be orthogonal to the two-dimensional subspace spanned by \mathbf{x}_1^* and \mathbf{x}_2^* ; RSS, therefore, has $(n - 1) - 2 = n - 3$ degrees of freedom. More generally, \mathbf{e} is orthogonal to the k -dimensional subspace spanned by $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_k^*$, and so RSS has $(n - 1) - k = n - k - 1$ degrees of freedom.
- An unbiased estimator of the error variance σ_e^2 can be derived by transforming the n correlated residuals \mathbf{e} to $n - k - 1$ independently and identically distributed residuals \mathbf{z} , employing an orthonormal basis \mathbf{G} for the $(n - k - 1)$ -dimensional error subspace: $\mathbf{z} = \mathbf{Ge}$. If the errors are independent and normally distributed, with zero means and common variance σ_e^2 , then so are the elements of \mathbf{z} . Thus, $\mathbf{z}'\mathbf{z}/(n - k - 1)$ is an unbiased estimator of the error variance, and because \mathbf{z} and \mathbf{e} are the same vector represented according to alternative bases, $\mathbf{z}'\mathbf{z}/(n - k - 1) = \mathbf{e}'\mathbf{e}/(n - k - 1)$, which is our usual estimator of error variance, S_E^2 .
- In the overparametrized one-way ANOVA model, $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, the $m + 1$ columns of the model matrix \mathbf{X} are collinear and span a subspace of dimension m . We can, however, still find $\hat{\mathbf{y}}$ for the model by projecting \mathbf{y} orthogonally onto this subspace, most simply by selecting an arbitrary basis for the column space of the model matrix. Conceived in this light, dummy coding and deviation coding are two techniques for constructing a basis for the column space of \mathbf{X} .

- In the overparametrized two-way ANOVA model, $Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$, the interaction regressors provide a basis for the full column space of the model matrix. The model-matrix columns for the two sets of main effects are therefore marginal to (i.e., subspaces of) the interaction space. The column for the constant regressor is marginal to the main-effect subspaces as well as to the interactions.

Recommended Reading

- There are several advanced texts that treat linear models from a strongly geometric perspective, including Dempster (1969) and Stone (1987). Both these books describe multivariate (i.e., multiple response-variable) generalizations of linear models, and both demand substantial mathematical sophistication. Also see Christensen (2011).
- In a text on matrix algebra, vector geometry, and associated mathematical topics, Green and Carroll (1976) focus on the geometric properties of linear models and related multivariate methods. The pace of the presentation is relatively leisurely, and the strongly geometric orientation provides insight into both the mathematics and the statistics.
- Wonnacott and Wonnacott (1979) invoke vector geometry to explain a variety of statistical topics, including some not covered in the present text—such as instrumental-variables estimation and structural-equation models.

PART III

Linear-Model Diagnostics

11

Unusual and Influential Data

As we have seen, linear statistical models—particularly linear regression analysis—make strong assumptions about the structure of data, assumptions that often do not hold in applications. The method of least squares, which is typically used to fit linear models to data, can be very sensitive to the structure of the data and may be markedly influenced by one or a few unusual observations.

We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation.¹ A less drastic response is also possible, however: We can adapt and extend the methods for examining and transforming data described in Chapters 3 and 4 to diagnose problems with a linear model that has been fit to data and—often—to suggest solutions.

I will pursue this strategy in this and the next two chapters:

- The current chapter deals with unusual and influential data.
- Chapter 12 takes up a variety of problems, including nonlinearity, nonconstant error variance, and non-normality.
- Collinearity is the subject of Chapter 13.

Taken together, the diagnostic and corrective methods described in these chapters greatly extend the practical application of linear models. These methods are often the difference between a crude, mechanical data analysis and a careful, nuanced analysis that accurately describes the data and therefore supports meaningful interpretation of them.

Another point worth making at the outset is that many problems can be anticipated and dealt with through careful examination of the data *prior* to building a regression model. Consequently, if you use the methods for examining and transforming data discussed in Chapters 3 and 4, you will be much less likely to encounter the difficulties detailed in the current part of the text on “postfit” linear-model diagnostics.

11.1 Outliers, Leverage, and Influence

In simple regression analysis, an *outlier* is an observation whose response-variable value is *conditionally* unusual *given* the value of the explanatory variable: See Figure 11.1. In contrast, a *univariate outlier* is a value of Y or X that is *unconditionally* unusual; such a value may or may not be a regression outlier.

¹Methods for nonparametric regression were introduced informally in Chapter 2 and will be described in more detail in Chapter 18. Robust regression is the subject of Chapter 19.

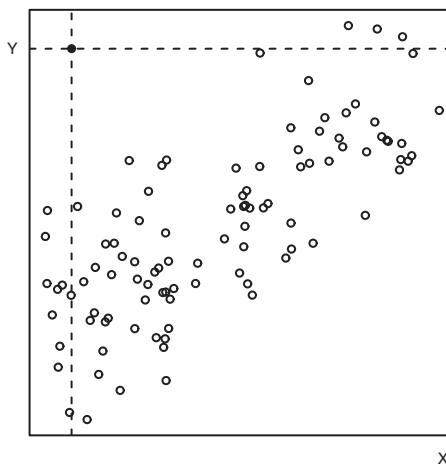


Figure 11.1 The black point is a regression outlier because it combines a relatively large value of Y with a relatively small value of X , even though neither its X -value nor its Y -value is unusual individually. Because of the positive relationship between Y and X , points with small X -values also tend to have small Y -values, and thus the black point is far from other points with similar X -values.

Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis and because their presence may be a signal that the model fails to capture important characteristics of the data. Some central distinctions are illustrated in Figure 11.2 for the simple-regression model $Y = \alpha + \beta X + \varepsilon$.

Regression outliers appear in Figure 11.2(a) and (b). In Figure 11.2(a), the outlying observation has an X -value that is at the center of the X -distribution; as a consequence, deleting the outlier has relatively little impact on the least-squares fit, leaving the slope B unchanged and affecting the intercept A only slightly. In Figure 11.2(b), however, the outlier has an unusually large X -value, and thus its deletion markedly affects both the slope and the intercept.² Because of its unusual X -value, the outlying right-most observation in Figure 11.2(b) exerts strong *leverage* on the regression coefficients, while the outlying middle observation in Figure 11.2(a) is at a low-leverage point. The combination of high leverage with a regression outlier therefore produces substantial *influence* on the regression coefficients. In Figure 11.2(c), the right-most observation has no influence on the regression coefficients even though it is a high-leverage point, because this observation is in line with the rest of the data—it is not a regression outlier.

The following heuristic formula helps to distinguish among the three concepts of influence, leverage, and discrepancy (“outlyingness”):

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}$$

A simple and transparent example, with real data from Davis (1990), appears in Figure 11.3. These data record the measured and reported weight of 183 male and female subjects

²When, as here, an observation is far away from and out of line with the rest of data, it is difficult to know what to make of it: Perhaps the relationship between Y and X in Figure 11.2(b) is nonlinear.

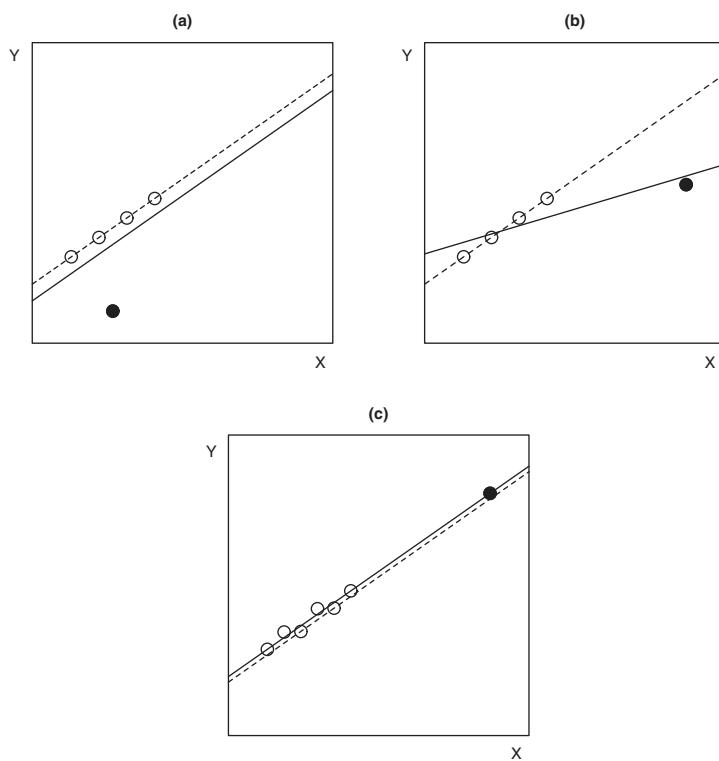


Figure 11.2 Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all the data, while the broken line gives the least-squares regression with the unusual data point (the black circle) omitted. (a) An outlier near the mean of X has low leverage and little influence on the regression coefficients. (b) An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients. (c) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect but are, in fact, coincident.

who engage in programs of regular physical exercise.³ Davis's data can be treated in two ways:

- We could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F , coded 1 for women and 0 for men), and an interaction regressor (formed as the product $MW \times F$). This specification follows from the reasonable assumption that measured weight, and possibly sex, can affect reported weight. The results are as follows (with coefficient standard errors in parentheses):

$$\begin{aligned}\widehat{RW} &= 1.36 + 0.990MW + 40.0F - 0.725(MW \times F) \\ &\quad (3.28) \quad (0.043) \quad (3.9) \quad (0.056) \\ R^2 &= 0.89 \quad S_E = 4.66\end{aligned}$$

³Davis's data were introduced in Chapter 2.

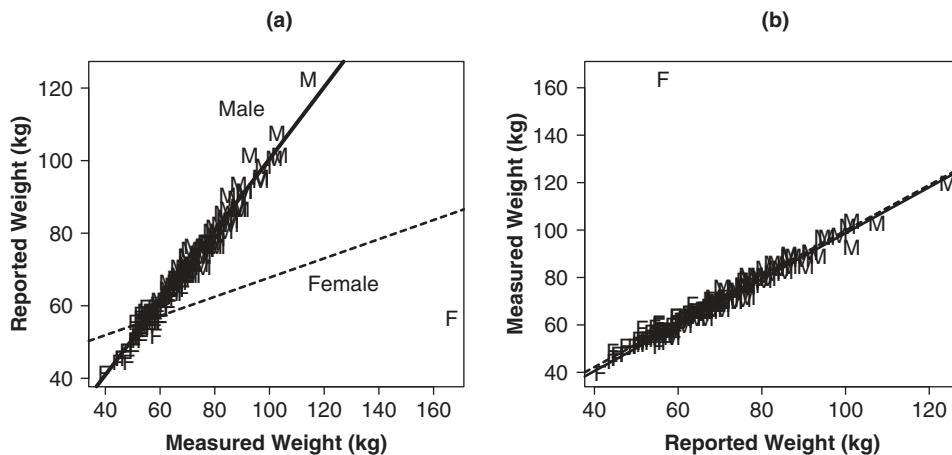


Figure 11.3 Regressions for Davis's data on reported and measured weight for women (F) and men (M). Panel (a) shows the least-squares linear regression line for each group (the solid line for men, the broken line for women) for the regression of reported on measured weight. The outlying observation has a large impact on the fitted line for women. Panel (b) shows the fitted regression lines for the regression of measured on reported weight; here, the outlying observation makes little difference to the fit, and the least-squares lines for men and women are nearly the same.

Were these results taken seriously, we would conclude that men are unbiased reporters of their weights (because $A = 1.36 \approx 0$ and $B_1 = 0.990 \approx 1$), while women tend to over-report their weights if they are relatively light and under-report if they are relatively heavy (the intercept for women is $1.36 + 40.0 = 41.4$ and the slope is $0.990 - 0.725 = 0.265$). Figure 11.3(a), however, makes it clear that the differential results for women and men are due to one female subject whose reported weight is about average (for women) but whose measured weight is extremely large. Recall that this subject's measured weight in kilograms and height in centimeters were erroneously switched. Correcting the data produces the regression

$$\widehat{RW} = 1.36 + 0.990MW + 1.98F - 0.0567(MW \times F)$$

$$(1.58) \quad (0.021) \quad (2.45) \quad (0.0385)$$

$$R^2 = 0.97 \quad S_E = 2.24$$

which suggests that both women and men are approximately unbiased reporters of their weight.

- We could (as in our previous analysis of Davis's data) treat measured weight as the response variable, regressing it on reported weight, sex, and their interaction—reflecting a desire to use reported weight as a predictor of measured weight. For the *uncorrected* data,

$$\begin{aligned}\widehat{MW} &= 1.79 + 0.969RW + 2.07F - 0.00953(RW \times F) \\ &\quad (5.92) (0.076) \quad (9.30) \quad (0.147) \\ R^2 &= 0.70 \quad S_E = 8.45\end{aligned}$$

The outlier does not have much impact on the coefficients for this regression (both the dummy-variable coefficient and the interaction coefficient are small) precisely because the value of RW for the outlying observation is near \overline{RW} for women [see Figure 11.3(b)]. There is, however, a marked effect on the multiple correlation and regression standard error: For the corrected data, $R^2 = 0.97$ and $S_E = 2.25$.

Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis and because they may indicate that the model fails to capture important features of the data. It is useful to distinguish among high-leverage observations, regression outliers, and influential observations. Influence on the regression coefficients is the product of leverage and outlyingness.

11.2 Assessing Leverage: Hat-Values

The so-called *hat-value* h_i is a common measure of leverage in regression. These values are so named because it is possible to express the fitted values \widehat{Y}_j (“Y-hat”) in terms of the observed values Y_i :

$$\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{ij}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^n h_{ij}Y_i$$

Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value \widehat{Y}_j : If h_{ij} is large, then the i th observation can have a considerable impact on the j th fitted value. It can be shown that $h_{ii} = \sum_{j=1}^n h_{ij}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on *all* the fitted values. The hat-values are bounded between $1/n$ and 1 (i.e., $1/n \leq h_i \leq 1$), and the average hat-value is $\bar{h} = (k+1)/n$ (where k is the number of regressors in the model, excluding the constant).⁴

In simple-regression analysis, the hat-values measure distance from the mean of X :⁵

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^n (X_j - \overline{X})^2}$$

In multiple regression, h_i measures distance from the centroid (point of means) of the X s, taking into account the correlational and variational structure of the X s, as illustrated for $k = 2$ explanatory variables in Figure 11.4. *Multivariate* outliers in the X -space are thus

⁴For derivations of this and other properties of leverage, outlier, and influence diagnostics, see Section 11.8.

⁵See Exercise 11.1. Note that the sum in the denominator is over the subscript j because the subscript i is already in use.

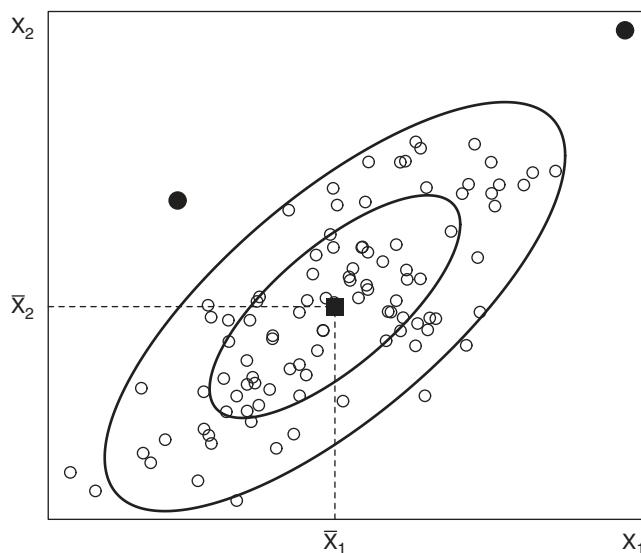


Figure 11.4 Elliptical contours of constant leverage (constant hat-values h_i) for $k = 2$ explanatory variables. Two high-leverage points appear, both represented by black circles. One point has unusually large values for each of X_1 and X_2 , but the other is unusual only in combining a moderately large value of X_2 with a moderately small value of X_1 . The centroid (point of means) is marked by the black square. (The contours of constant leverage are proportional to the standard data ellipse, introduced in Chapter 9.)

high-leverage observations. The response-variable values are not at all involved in determining leverage.

For Davis's regression of reported weight on measured weight, the largest hat-value by far belongs to the 12th subject, whose measured weight was wrongly recorded as 166 kg: $h_{12} = 0.714$. This quantity is many times the average hat-value, $\bar{h} = (3 + 1)/183 = 0.0219$.

Figure 11.5(a) shows an *index plot* of hat-values from Duncan's regression of the prestige of 45 occupations on their income and education levels (i.e., a scatterplot of hat-values vs. the observation indices).⁶ The horizontal lines in this graph are drawn at twice and three times the average hat-values, $\bar{h} = (2 + 1)/45 = 0.06667$.⁷ Figure 11.5(b) shows a scatterplot for the explanatory variables education and income: *Railroad engineers* and *conductors* have high leverage by virtue of their relatively high income for their moderately low level of education, while *ministers* have high leverage because their level of income is relatively low given their moderately high level of education.

Observations with unusual combinations of explanatory-variable values have high *leverage* in a least-squares regression. The hat-values h_i provide a measure of leverage. The average hat-value is $\bar{h} = (k + 1)/n$.

⁶Duncan's regression was introduced in Chapter 5.

⁷See Section 11.5 on numerical cutoffs for diagnostic statistics.

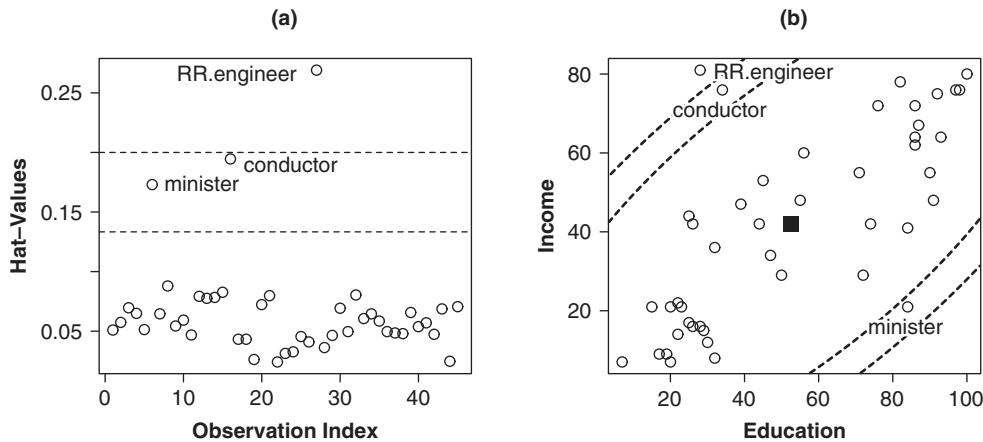


Figure 11.5 (a) An index plot of hat-values for Duncan's occupational prestige regression, with horizontal lines at $2 \times \bar{h}$ and $3 \times \bar{h}$. (b) A scatterplot of education by income, with contours of constant leverage at $2 \times \bar{h}$ and $3 \times \bar{h}$ given by the broken lines. (Note that the ellipses extend beyond the boundaries of the graph.) The centroid is marked by the black square.

11.3 Detecting Outliers: Studentized Residuals

To identify an outlying observation, we need an index of the unusualness of Y given the X s. Discrepant observations usually have large residuals, but it turns out that even if the errors ε_i have equal variances (as assumed in the general linear model), the residuals E_i do not:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

High-leverage observations, therefore, tend to have small residuals—an intuitively sensible result because these observations can pull the regression surface toward them.

Although we can form a *standardized residual* by calculating

$$E'_i \equiv \frac{E_i}{S_E \sqrt{1 - h_i}}$$

this measure is slightly inconvenient because its numerator and denominator are not independent, preventing E'_i from following a t -distribution: When $|E_i|$ is large, the standard error of the regression, $S_E = \sqrt{\sum E_i^2 / (n - k - 1)}$, which contains E_i^2 , tends to be large as well.

Suppose, however, that we refit the model deleting the i th observation, obtaining an estimate $S_{E(-i)}$ of σ_ε that is based on the *remaining* $n - 1$ observations. Then the *studentized residual*

$$E_i^* \equiv \frac{E_i}{S_{E(-i)} \sqrt{1 - h_i}} \tag{11.1}$$

has an independent numerator and denominator and follows a t -distribution with $n - k - 2$ degrees of freedom.

An alternative, but equivalent, procedure for defining the studentized residuals employs a “mean-shift” outlier model:

$$Y_j = \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} + \gamma D_j + \varepsilon_j \quad (11.2)$$

where D is a dummy regressor set to 1 for observation i and 0 for all other observations:

$$D_j = \begin{cases} 1 & \text{for } j = i \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\begin{aligned} E(Y_i) &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma \\ E(Y_j) &= \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} \quad \text{for } j \neq i \end{aligned}$$

It would be natural to specify the model in Equation 11.2 if, *before examining the data*, we suspected that observation i differed from the others. Then, to test $H_0: \gamma = 0$ (i.e., the null hypothesis that the i th observation is *not* an outlier), we can calculate $t_0 = \hat{\gamma}/\text{SE}(\hat{\gamma})$. This test statistic is distributed as t_{n-k-2} under H_0 and (it turns out) is the studentized residual E_i^* of Equation 11.1.

Hoaglin and Welsch (1978) arrive at the studentized residuals by successively omitting each observation, calculating its residual based on the regression coefficients obtained for the remaining sample, and dividing the resulting residual by its standard error. Finally, Beckman and Trussell (1974) demonstrate the following simple relationship between studentized and standardized residuals:

$$E_i^* = E'_i \sqrt{\frac{n - k - 2}{n - k - 1 - E'^2_i}} \quad (11.3)$$

If n is large, then the factor under the square root in Equation 11.3 is close to 1, and the distinction between standardized and studentized residuals essentially disappears.⁸ Moreover, for large n , the hat-values are generally small, and so it is usually the case that

$$E_i^* \approx E'_i \approx \frac{E_i}{S_E}$$

Equation 11.3 also implies that E_i^* is a monotone function of E'_i , and thus the rank order of the studentized and standardized residuals is the same.

11.3.1 Testing for Outliers in Linear Models

Because in most applications we do not suspect a *particular* observation in advance, but rather want to look for *any* outliers that may occur in the data, we can, in effect, refit the mean-shift model n times,⁹ once for each observation, producing studentized residuals E_i^* ,

⁸Here, as elsewhere in statistics, terminology is not wholly standard: E_i^* is sometimes called a *deleted studentized residual*, an *externally studentized residual*, or even a *standardized residual*; likewise, E'_i is sometimes called an *internally studentized residual*, or simply a *studentized residual*. It is therefore helpful, especially in small samples, to determine exactly what is being calculated by a computer program.

⁹It is not necessary *literally* to perform n auxiliary regressions. Equation 11.3, for example, permits the computation of studentized residuals with little effort.

E_2^*, \dots, E_n^* . Usually, our interest then focuses on the largest absolute E_i^* , denoted E_{\max}^* . Because we have in effect picked the biggest of n test statistics, however, it is not legitimate simply to use t_{n-k-2} to find a p -value for E_{\max}^* : For example, even if our model is wholly adequate, and disregarding for the moment the dependence among the E_i^* 's, we would expect to obtain about 5% of E_i^* 's beyond $t_{.025} \approx \pm 2$, about 1% beyond $t_{.005} \approx \pm 2.6$, and so forth.

One solution to this problem of simultaneous inference is to perform a *Bonferroni adjustment* to the p -value for the *largest absolute* E_i^* .¹⁰ The Bonferroni test requires either a special t -table or, even more conveniently, a computer program that returns accurate p -values for values of t far into the tail of the t -distribution. In the latter event, suppose that $p' = \Pr(t_{n-k-2} > E_{\max}^*)$. Then the Bonferroni p -value for testing the statistical significance of E_{\max}^* is $p = 2np'$. The factor 2 reflects the two-tail character of the test: We want to detect large negative as well as large positive outliers.

Beckman and Cook (1983) show that the Bonferroni adjustment is usually exact in testing the largest studentized residual. A much larger E_{\max}^* is required for a statistically significant result than would be the case for an ordinary individual t -test.

In Davis's regression of reported weight on measured weight, the largest studentized residual by far belongs to the incorrectly recorded 12th observation, with $E_{12}^* = -24.3$. Here, $n - k - 2 = 183 - 3 - 2 = 178$, and $\Pr(t_{178} > 24.3) \approx 10^{-58}$. The Bonferroni p -value for the outlier test is thus $p \approx 2 \times 183 \times 10^{-58} \approx 4 \times 10^{-56}$, an unambiguous result.

Put alternatively, the 5% critical value for E_{\max}^* in this regression is the value of t_{178} with probability $.025/183 = 0.0001366$ to the right. That is, $E_{\max}^* = t_{178, .0001366} = 3.714$; this critical value contrasts with $t_{178, .025} = 1.973$, which would be appropriate for testing an *individual* studentized residual identified in advance of inspecting the data.

For Duncan's occupational prestige regression, the largest studentized residual belongs to *ministers*, with $E_{\text{minister}}^* = 3.135$. The associated Bonferroni p -value is $2 \times 45 \times \Pr(t_{45-2-2} > 3.135) = .143$, showing that it is not terribly unusual to observe a studentized residual this big in a sample of 45 observations.

11.3.2 Anscombe's Insurance Analogy

Thus far, I have treated the identification (and, implicitly, the potential correction, removal, or accommodation) of outliers as a hypothesis-testing problem. Although this is by far the most common procedure in practice, a more reasonable (if subtle) general approach is to assess the potential costs and benefits for estimation of discarding an unusual observation.

Imagine, for the moment, that the observation with the largest E_i^* is simply an unusual data point but one generated by the assumed statistical model:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

with independent errors ε_i that are each distributed as $N(0, \sigma_\varepsilon^2)$. To discard an observation under these circumstances would *decrease* the efficiency of estimation, because when the

¹⁰See online Appendix D on probability and estimation for a discussion of Bonferroni inequalities and their role in simultaneous inference. A graphical alternative to testing for outliers is to construct a quantile-comparison plot for the studentized residuals, comparing the sample distribution of these quantities with the t -distribution for $n - k - 2$ degrees of freedom. See the discussion of non-normality in the next chapter.

model—including the assumption of normality—is correct, the least-squares estimators are maximally efficient among all unbiased estimators of the regression coefficients.¹¹

If, however, the observation in question does not belong with the rest (e.g., because the mean-shift model applies), then to eliminate it may make estimation more efficient. Anscombe (1960) developed this insight by drawing an analogy to insurance: To obtain *protection* against “bad” data, one purchases a *policy* of outlier rejection, a policy paid for by a small *premium* in efficiency when the policy inadvertently rejects “good” data.¹²

Let q denote the desired premium, say 0.05—that is, a 5% increase in estimator mean-squared error if the model holds for all of the data. Let z represent the unit-normal deviate corresponding to a tail probability of $q(n - k - 1)/n$. Following the procedure derived by Anscombe and Tukey (1963), compute $m = 1.4 + 0.85z$ and then find

$$E'_q = m \left(1 - \frac{m^2 - 2}{4(n - k - 1)} \right) \sqrt{\frac{n - k - 1}{n}} \quad (11.4)$$

The largest absolute *standardized* residual can be compared with E'_q to determine whether the corresponding observation should be rejected as an outlier. This cutoff can be translated to the studentized-residual scale using Equation 11.3:

$$E_q^* = E'_q \sqrt{\frac{n - k - 2}{n - k - 1 - E'^2_q}} \quad (11.5)$$

In a real application, of course, we should inquire about discrepant observations rather than simply throwing them away.¹³

For example, for Davis’s regression of reported on measured weight, $n = 183$ and $k = 3$; so, for the premium $q = 0.05$, we have

$$\frac{q(n - k - 1)}{n} = \frac{0.05(183 - 3 - 1)}{183} = 0.0489$$

From the quantile function of the standard-normal distribution, $z = 1.66$, from which $m = 1.4 + 0.85 \times 1.66 = 2.81$. Then, using Equation 11.4, $E'_q = 2.76$, and using Equation 11.5, $E_q^* = 2.81$. Because $E_{\max}^* = |E_{12}^*| = 24.3$ is much larger than E_q^* , the 12th observation is identified as an outlier.

In Duncan’s occupational prestige regression, $n = 45$ and $k = 2$. Thus, with premium $q = 0.05$,

$$\frac{q(n - k - 1)}{n} = \frac{0.05(45 - 2 - 1)}{45} = 0.0467$$

The corresponding unit-normal deviate is $z = 1.68$, yielding $m = 1.4 + 0.85 \times 1.68 = 2.83$, $E'_q = 2.63$, and $E_q^* = 2.85 < |E_{\text{minister}}^*| = 3.135$, suggesting that *ministers* be rejected as an outlier, even though the Bonferroni test did not declare this observation to be a “statistically significant” outlier.

¹¹See Chapter 9.

¹²An alternative is to employ a robust estimator, which is somewhat less efficient than least squares when the model is correct but much more efficient when outliers are present. See Chapter 19.

¹³See the discussion in Section 11.7.

A regression *outlier* is an observation with an unusual response-variable value given its combination of explanatory-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination, a Bonferroni test for the largest absolute E_i^* , or Anscombe's insurance analogy. If the model is correct (and there are no true outliers), then each studentized residual follows a t -distribution with $n - k - 2$ degrees of freedom.

11.4 Measuring Influence

As noted previously, influence on the regression coefficients combines leverage and discrepancy. The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

$$D_{ij} = B_j - B_{j(-i)} \quad \text{for } i = 1, \dots, n \text{ and } j = 0, 1, \dots, k$$

where the B_j are the least-squares coefficients calculated for all the data, and the $B_{j(-i)}$ are the least-squares coefficients calculated with the i th observation omitted. (So as not to complicate the notation here, I denote the least-squares intercept A as B_0 .) To assist in interpretation, it is useful to scale the D_{ij} by (deleted) coefficient standard errors:

$$D_{ij}^* = \frac{D_{ij}}{\text{SE}_{(-i)}(B_j)}$$

Following Belsley, Kuh, and Welsch (1980), the D_{ij} are often termed DFBETA_{ij} , and the D_{ij}^* are called DFBETAS_{ij} .

One problem associated with using the D_{ij} or the D_{ij}^* is their large number— $n(k + 1)$ of each. Of course, these values can be more quickly and effectively examined graphically than in numerical tables. We can, for example, construct an index plot of the D_{ij}^* 's for each coefficient, $j = 0, 1, \dots, k$ (see below for an example). A more informative, if more complex, alternative is to construct a scatterplot matrix of the D_{ij}^* with index plots (or some other univariate display) on the diagonal.¹⁴ Nevertheless, it is useful to have a single summary index of the influence of each observation on the least-squares fit.

Cook (1977) has proposed measuring the “distance” between the B_j and the corresponding $B_{j(-i)}$ by calculating the F -statistic for the “hypothesis” that $\beta_j = B_{j(-i)}$, for $j = 0, 1, \dots, k$. This statistic is recalculated for each observation $i = 1, \dots, n$. The resulting values should not literally be interpreted as F -tests—Cook’s approach merely exploits an *analogy* to testing to produce a measure of distance that is independent of the scales of the X -variables. Cook’s distance can be written (and simply calculated) as

$$D_i = \frac{E_i'^2}{k + 1} \times \frac{h_i}{1 - h_i}$$

In effect, the first term in the formula for Cook’s D is a measure of discrepancy, and the second is a measure of leverage. We look for values of D_i that stand out from the rest.

¹⁴This interesting display was suggested to me by Michael Friendly of York University.

Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D statistic provides a summary index of influence on the coefficients.

Belsley et al. (1980) have suggested the very similar measure¹⁵

$$\text{DFFITS}_i = E_i^* \sqrt{\frac{h_i}{1 - h_i}}$$

Except for unusual data configurations, Cook's $D_i \approx \text{DFFITS}_i^2 / (k + 1)$.

Because all the deletion statistics depend on the hat-values and residuals, a graphical alternative to either of these general influence measures is to plot the E_i^* against the h_i and to look for observations for which both are big. A slightly more sophisticated (and more informative) version of this plot displays circles of area proportional to Cook's D instead of points (see Figure 11.6). We can follow up by examining the D_{ij} or D_{ij}^* for the observations with the largest few D_i , $|\text{DFFITS}_i|$, or a combination of large h_i and $|E_i^*|$.

For Davis's regression of reported weight on measured weight, all the indices of influence point to the obviously discrepant 12th observation:

$$\text{Cook's } D_{12} = 85.9 \text{ (next largest, } D_{115} = 0.085\text{)}$$

$$\text{DFFITS}_{12} = -38.4 \text{ (next largest, DFFITS}_{115} = 0.603\text{)}$$

$$\text{DFBETAS}_{0,12} = \text{DFBETAS}_{1,12} = 0$$

$$\text{DFBETAS}_{2,12} = 20.0, \text{DFBETAS}_{3,12} = -24.8$$

Note that the outlying Observation 12, which is for a female subject, has no impact on the male intercept B_0 (i.e., A) and slope B_1 but does exert considerable influence on the dummy-variable coefficient B_2 and the interaction coefficient B_3 .

Turning our attention to Duncan's occupational prestige regression, Figure 11.6 shows a “bubble plot” of studentized residuals by hat-values, with the areas of the circles proportional to the Cook's distances of the observations. Several noteworthy observations are identified on the plot: *ministers* and *conductors*, who combine relatively high leverage with relatively large studentized residuals; *railroad engineers*, who have very high leverage but a small studentized residual; and *reporters*, who have a relatively large (negative) residual but lower leverage. Index plots of D_{ij}^* for the income and education coefficients in the regression appear in Figure 11.7: *Ministers* and *conductors* serve to decrease the income coefficient and increase the education coefficient—in the case of *ministers* by more than one standard error.

11.4.1 Influence on Standard Errors

In developing the concept of influence in regression, I have focused on changes in regression coefficients. Other regression outputs are also subject to influence, however. One important

¹⁵Other global measures of influence are available; see Chatterjee and Hadi (1988, chap. 4) for a comparative treatment.

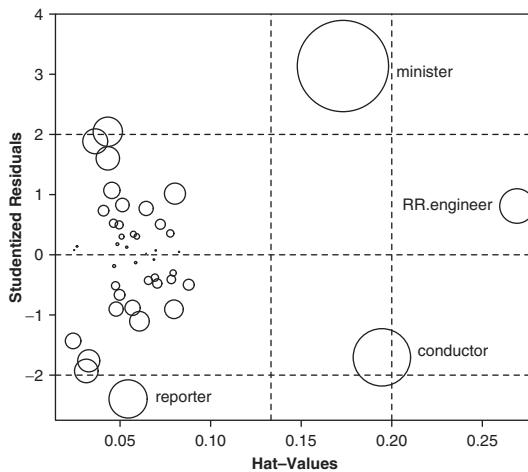


Figure 11.6 “Bubble plot” of Cook’s D_s , studentized residuals, and hat-values, for Duncan’s regression of occupational prestige on income and education. Each point is plotted as a circle with area proportional to D . Horizontal reference lines are drawn at studentized residuals of 0 and ± 2 ; vertical reference lines are drawn at hat-values of $2\bar{h}$ and $3\bar{h}$ (see Section 11.5 on numerical cutoffs for diagnostic statistics). Several observations are identified on the plot: *Ministers* and *conductors* have large hat-values and relatively large residuals; *reporters* have a relatively large negative residual but a small hat-value; *railroad engineers* have a large hat-value but a small residual.

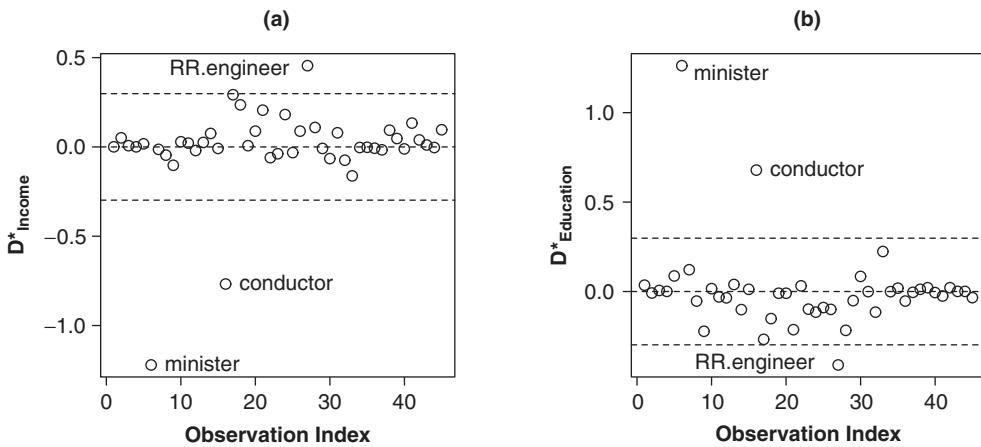


Figure 11.7 Index plots of D_{ij}^* for the (a) income and (b) education coefficients in Duncan’s occupational prestige regression. The horizontal lines in the graphs are drawn at $D^* = 0$ and the rule-of-thumb cutoffs $\pm 2/\sqrt{n}$ (see Section 11.5). The observations *ministers* and *conductors* stand out. *Railroad engineers* are beyond the cutoffs for both coefficients but do not stand out from the other observations to the same degree.

regression output is the set of coefficient sampling variances and covariances, which capture the precision of estimation in regression.

Reexamine, for example, Figure 11.2(c) on page 268, in which a high-leverage observation exerts no influence on the regression coefficients because it is in line with the rest of the data. Recall, as well, that the standard error of the least-squares slope in simple regression is¹⁶

$$\text{SE}(B) = \frac{S_E}{\sqrt{\sum (X_i - \bar{X})^2}}$$

By increasing the variance of X , therefore, a high-leverage but in-line observation serves to decrease $\text{SE}(B)$ even though it does not influence the regression coefficients A and B . Depending on the context, such an observation may be considered beneficial—because it increases the precision of estimation—or it may cause us to exaggerate our confidence in the estimate B .

In multiple regression, we can examine the impact of deleting each observation in turn on the size of the joint confidence region for the regression coefficients.¹⁷ The size of the joint confidence region is analogous to the length of a confidence interval for an individual regression coefficient, which, in turn, is proportional to the standard error of the coefficient. The squared length of a confidence interval is, therefore, proportional to the sampling variance of the coefficient, and, analogously, the squared size of a joint confidence region is proportional to the “generalized variance” of a set of coefficients.

An influence measure proposed by Belsley et al. (1980) closely approximates the squared ratio of volumes of the deleted and full-data confidence regions for the regression coefficients:¹⁸

$$\text{COVRATIO}_i = \frac{1}{(1 - h_i) \left(\frac{n-k-2+E_i^*}{n-k-1} \right)^{k+1}}$$

Observations that *increase* the precision of estimation have values of COVRATIO that are larger than 1; those that *decrease* the precision of estimation have values smaller than 1. Look for values of COVRATIO, therefore, that differ considerably from 1.

As was true of measures of influence on the regression coefficients, both the hat-value and the (studentized) residual figure in COVRATIO. A large hat-value produces a large COVRATIO, however, even when—indeed, especially when—the studentized residual is small because a high-leverage in-line observation improves the precision of estimation. In contrast, a discrepant, low-leverage observation might not change the coefficients much, but it decreases the precision of estimation by increasing the estimated error variance; such an observation, with small h_i and large E_i^* , produces a COVRATIO_i well below 1.

For Davis’s regression of reported weight on measured weight, sex, and their interaction, by far the most extreme value is $\text{COVRATIO}_{12} = 0.0103$. The 12th observation, therefore, *decreases* the precision of estimation by a factor of $1/0.0103 \approx 100$. In this instance, a very large leverage, $h_{12} = 0.714$, is more than offset by a massive residual, $E_{12}^* = -24.3$.

¹⁶See Chapter 6.

¹⁷See Section 9.4.4 for a discussion of joint confidence regions.

¹⁸Alternative, similar measures have been suggested by several authors. Chatterjee and Hadi (1988, Chapter 4) provide a comparative discussion.

In Duncan's occupational prestige regression, the smallest and largest values are $\text{COVRATIO}_{\text{minister}} = 0.682$ and $\text{COVRATIO}_{\text{railroad engineer}} = 1.402$. Thus, the discrepant, relatively high-leverage observation *minister* decreases the precision of estimation, while the in-line, high-leverage observation *railroad engineer* increases it.

11.4.2 Influence on Collinearity

Other characteristics of a regression analysis can also be influenced by individual observations, including the degree of collinearity among the explanatory variables.¹⁹ I will not address this issue in any detail, but the following points may prove helpful:²⁰

- Influence on collinearity is one of the factors reflected in influence on coefficient standard errors. Measures such as COVRATIO, however, also reflect influence on the error variance and on the variation of the X s. Moreover, COVRATIO and similar measures examine the sampling variances and covariances of *all* the regression coefficients, including the regression constant, while a consideration of collinearity generally excludes the constant. Nevertheless, our concern for collinearity reflects its impact on the precision of estimation, which is precisely what is addressed by COVRATIO.
- Collinearity-influential points are those that either induce or weaken correlations among the X s. Such points usually—but not always—have large hat-values. Conversely, points with large hat-values often influence collinearity.
- Individual points that induce collinearity are obviously problematic. More subtly, points that weaken collinearity also merit examination because they may cause us to be overly confident in our results—analogous to the increased (or apparently increased) precision of estimation induced by an unusually large X -value for an in-line observation in simple regression.
- It is frequently possible to detect collinearity-influential points by plotting explanatory variables against each other, as in a scatterplot matrix or a three-dimensional rotating plot. This approach may fail, however, if the collinear relations in question involve more than two or three explanatory variables at a time.

11.5 Numerical Cutoffs for Diagnostic Statistics

I have deliberately refrained from suggesting specific numerical criteria for identifying noteworthy observations on the basis of measures of leverage and influence: I believe that it is generally more effective to examine the distributions of these quantities directly to locate unusual values. For studentized residuals, the hypothesis-testing and insurance approaches provide numerical cutoffs, but even these criteria are no substitute for graphical examination of the residuals.

Still, numerical cutoffs can be of some use, as long as they are not given too much weight and especially when they are employed to enhance graphical displays: A line can be drawn on

¹⁹See Chapter 13 for a general treatment of collinearity.

²⁰See Chatterjee and Hadi (1988, Chapters 4 and 5) for more information about influence on collinearity.

a graph at the value of a numerical cutoff, and observations that exceed the cutoff can be identified individually.²¹

Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic. Cutoffs may be absolute, or they may be adjusted for sample size.²² For some diagnostic statistics, such as measures of influence, absolute cutoffs are unlikely to identify noteworthy observations in large samples. This characteristic reflects the ability of large samples to absorb discrepant data without markedly changing the results, but it is still often of interest to identify *relatively* influential points, even if no observation has strong *absolute* influence, because unusual data may prove to be substantively interesting. An outlier, for example, may teach us something unexpected about the process under investigation.²³

The cutoffs presented below are, as explained briefly here, derived from statistical theory. An alternative and universally applicable data-based criterion is simply to examine the most extreme (e.g., 5% of) values of a diagnostic statistic.

11.5.1 Hat-Values

Belsley et al. (1980) suggest that hat-values exceeding about twice the average $\bar{h} = (k + 1)/n$ are noteworthy. This size-adjusted cutoff was derived as an approximation identifying the most extreme 5% of cases when the X s are multivariate normal, and the number of regressors k and degrees of freedom for error $n - k - 1$ are relatively large. The cutoff is nevertheless recommended by these authors as a rough general guide even when the regressors are not normally distributed. In small samples, using $2 \times \bar{h}$ tends to nominate too many points for examination, and $3 \times \bar{h}$ can be used instead.²⁴

11.5.2 Studentized Residuals

Beyond the issues of “statistical significance” and estimator robustness and efficiency discussed above, it sometimes helps to call attention to residuals that are relatively large. Recall that, under ideal conditions, about 5% of studentized residuals are outside the range $|E_i^*| \leq 2$. It is, therefore, reasonable, for example, to draw lines at ± 2 on a display of studentized residuals to draw attention to observations outside this range.

11.5.3 Measures of Influence

Many cutoffs have been suggested for various measures of influence. A few are presented here:

²¹See, for example, Figures 11.5 (page 272) and 11.6 (page 278).

²²See Belsley et al. (1980, Chapter 2) for further discussion of these distinctions.

²³See the discussion in Section 11.7.

²⁴See Chatterjee and Hadi (1988, Chapter 4) for a discussion of alternative cutoffs for hat-values.

- *Standardized change in regression coefficients.* The D_{ij}^* are scaled by standard errors, and, consequently, $|D_{ij}^*| > 1$ or 2 suggests itself as an absolute cutoff. As explained above, however, this criterion is unlikely to nominate observations in large samples. Belsley et al. (1980) propose the size-adjusted cutoff $2/\sqrt{n}$ for identifying noteworthy D_{ij}^* s.
- *Cook's D and DFFITS.* Several numerical cutoffs have been recommended for Cook's D and for DFFITS—exploiting the analogy between D and an F -statistic, for example. Chatterjee and Hadi (1988) suggest the size-adjusted cutoff²⁵

$$|\text{DFFITS}_i| > 2\sqrt{\frac{k+1}{n-k-1}}$$

Because of the approximate relationship between DFFITS and Cook's D , it is simple to translate this criterion into

$$D_i > \frac{4}{n-k-1}$$

Absolute cutoffs for D , such as $D_i > 1$, risk missing relatively influential data.

- *COVRATIO.* Belsley et al. (1980) suggest the size-adjusted cutoff

$$|\text{COVRATIO}_i - 1| > \frac{3(k+1)}{n}$$

11.6 Joint Influence

As illustrated in Figure 11.8, subsets of observations can be *jointly influential* or can offset each other's influence. *Influential subsets* or *multiple outliers* can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially. It can be important, however, to refit the model after deleting each point because the presence of a single influential value can dramatically affect the fit at other points. Still, the sequential approach is not always successful.

11.6.1 Added-Variable Plots

Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach impractical.²⁶ An attractive alternative is to employ graphical methods, and an especially useful influence graph is the *added-variable* (or *AV*) plot (also called a *partial-regression plot* or a *partial-regression leverage plot*).

²⁵Also see Cook (1977), Belsley et al. (1980), and Velleman and Welsch (1981).

²⁶Cook and Weisberg (1980), for example, extend the D statistic to a subset of p observations indexed by the vector subscript $\mathbf{i} = (i_1, i_2, \dots, i_p)'$:

$$D_{\mathbf{i}} = \frac{\mathbf{d}_{\mathbf{i}}'(\mathbf{X}'\mathbf{X})\mathbf{d}_{\mathbf{i}}}{(k+1)S_E^2}$$

where $\mathbf{d}_{\mathbf{i}} = \mathbf{b} - \mathbf{b}_{(-\mathbf{i})}$ gives the impact on the regression coefficients of deleting the subset \mathbf{i} . See Belsley et al. (1980, Chapter 2) and Chatterjee and Hadi (1988) for further discussions of deletion diagnostics based on subsets of observations. There are, however, $n!/[p!(n-p)!]$ subsets of size p —typically a prohibitively large number, even for modest values of p .

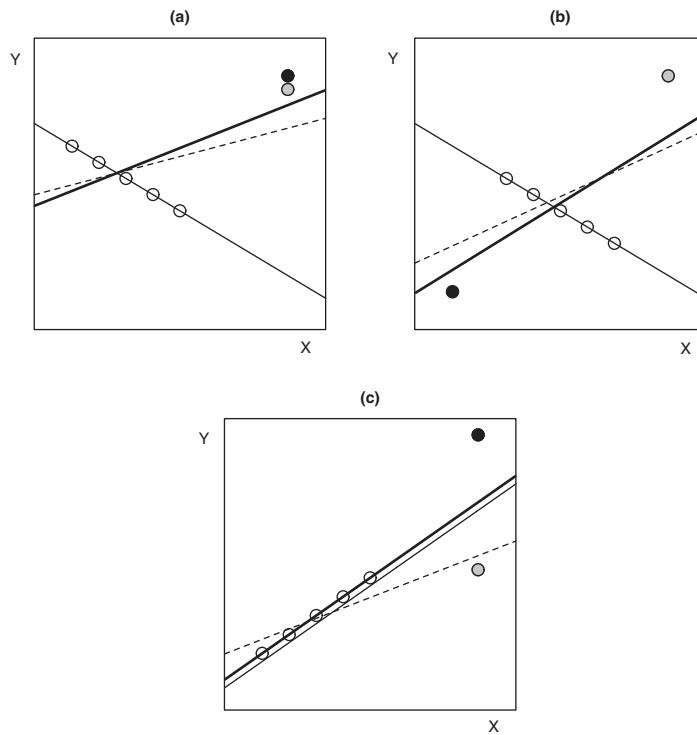


Figure 11.8 Jointly influential data in simple regression. In each graph, the heavier solid line gives the least-squares regression for all of the data, the broken line gives the regression with the black circle deleted, and the lighter solid line gives the regression with both the black circle and the gray circle deleted. (a) Jointly influential observations located close to one another: Deletion of both observations has a much greater impact than deletion of only one. (b) Jointly influential observations located on opposite sides of the data. (c) Observations that offset one another: The regression with both observations deleted is the same as for the whole data set (the two lines are separated slightly for visual effect).

Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all the X s with the exception of X_1 —that is, the residuals from the fitted regression equation.

$$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$

The parenthetical superscript (1) indicates the omission of X_1 from the right-hand side of the regression equation. Likewise, $X_i^{(1)}$ is the residual from the least-squares regression of X_1 on all the other X s:

$$X_{i1} = C^{(1)} + D_2^{(1)}X_{i2} + \cdots + D_k^{(1)}X_{ik} + X_i^{(1)}$$

This notation emphasizes the interpretation of the residuals $Y^{(1)}$ and $X^{(1)}$ as the parts of Y and X_1 that remain when the contributions of X_2, \dots, X_k are “removed.”

The residuals $Y^{(1)}$ and $X^{(1)}$ have the following interesting properties:

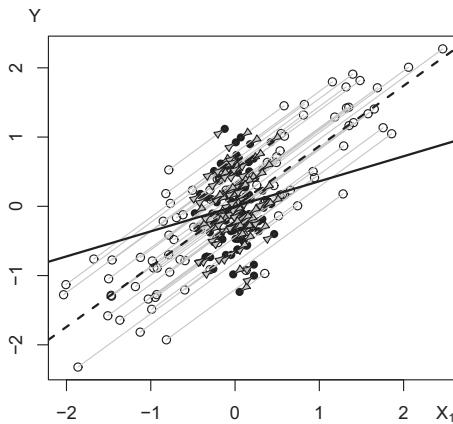


Figure 11.9 The marginal scatterplot (open circles) for Y and X_1 superimposed on the added-variable plot (filled circles) for X_1 in the regression of Y on X_1 and X_2 . The variables Y and X_1 are centered at their means to facilitate the comparison of the two sets of points. The arrows show how the points in the marginal scatterplot map into those in the AV plot. In this contrived data set, X_1 and X_2 are highly correlated ($r_{12} = .98$), and so the conditional variation in X_1 (represented by the horizontal spread of the filled points) is much less than its marginal variation (represented by the horizontal spread of the open points). The broken line gives the slope of the marginal regression of Y on X_1 alone, while the solid line gives the slope B_1 of X_1 in the multiple regression of Y on both X s.

1. The slope from the least-squares regression of $Y^{(1)}$ on $X^{(1)}$ is simply the least-squares slope B_1 from the *full* multiple regression.
2. The residuals from the simple regression of $Y^{(1)}$ on $X^{(1)}$ are the same as those from the full regression; that is,

$$Y_i^{(1)} = B_1 X_i^{(1)} + E_i \quad (11.6)$$

No constant is required here because both $Y^{(1)}$ and $X^{(1)}$ are least-squares residuals and therefore have means of 0, forcing the regression through the origin.

3. The variation of $X^{(1)}$ is the *conditional variation* of X_1 holding the other X s constant and, as a consequence, the standard error of B_1 in the auxiliary simple regression (Equation 11.6),

$$\text{SE}(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)2}}}$$

is the same as the *multiple-regression* standard error of B_1 .²⁷ Unless X_1 is uncorrelated with the other X s, its conditional variation is smaller than its *marginal variation* $\sum (X_{i1} - \bar{X}_1)^2$ —much smaller, if X_1 is strongly collinear with the other X s (see Figure 11.9).

²⁷There is slight slippage here with respect to the degrees of freedom for error: S_E is from the multiple regression, with $n - k - 1$ degrees of freedom for error. We need not subtract the mean of $X_i^{(1)}$ to calculate the standard error of the slope because the mean of these residuals is already 0.

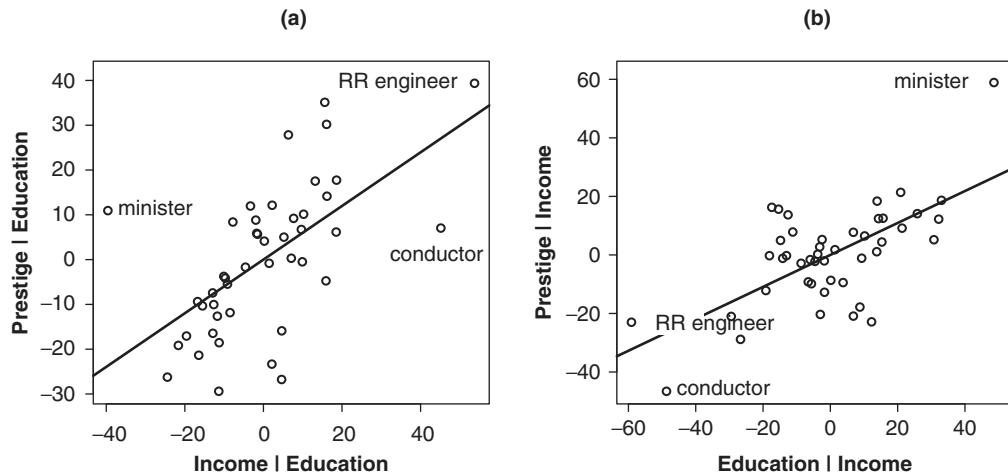


Figure 11.10 Added-variable plots for Duncan's regression of occupational prestige on the (a) income and (b) education levels of 45 U.S. occupations in 1950. Three unusual observations, *ministers*, *conductors*, and *railroad engineers*, are identified on the plots. The added-variable plot for the intercept A is not shown.

Plotting $Y^{(1)}$ against $X^{(1)}$ permits us to examine the leverage and influence of the observations on B_1 . Because of properties 1 to 3, this plot also provides a visual impression of the precision of the estimate B_1 . Similar added-variable plots can be constructed for the other regressors.²⁸

Plot $Y^{(j)}$ versus $X^{(j)}$ for each $j = 1, \dots, k$

Subsets of observations can be jointly influential. Added-variable plots are useful for detecting joint influence on the regression coefficients. The added-variable plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all the other X s.

Illustrative added-variable plots are shown in Figure 11.10, using data from Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations. Recall (from Chapter 5) that Duncan's regression yields the following least-squares fit:

$$\widehat{\text{Prestige}} = -6.06 + 0.599 \times \text{Income} + 0.546 \times \text{Education}$$

$$(4.27) \quad (0.120) \quad (0.098)$$

$$R^2 = 0.83 \quad S_E = 13.4$$

²⁸We can also construct an added-variable plot for the intercept A , by regressing the “constant regressor” $X_0 = 1$ and Y on X_1 through X_k , with no constant in these regression equations.

The added-variable plot for income in Figure 11.10(a) reveals three observations that exert substantial leverage on the income coefficient. Two of these observations serve to decrease the income slope: *ministers*, whose income is unusually low given the educational level of the occupation, and *railroad conductors*, whose income is unusually high given education. The third occupation, *railroad engineers*, is above the fitted regression but is not as discrepant; it, too, has relatively high income given education. Remember that the horizontal variable in this added-variable plot is the residual from the regression of income on education, and thus values far from 0 in this direction are for occupations with incomes that are unusually high or low given their levels of education.

The added-variable plot for education in Figure 11.10(b) shows that the same three observations have relatively high leverage on the education coefficient: *Ministers* and *railroad conductors* tend to increase the education slope, while *railroad engineers* appear to be closer in line with the rest of the data. Recall that our attention was also called to these occupations when we examined the individual-observation diagnostic statistics: hat-values, studentized residuals, Cook's distances, and so on.

Deleting *ministers* and *conductors* produces the fitted regression

$$\widehat{\text{Prestige}} = -6.41 + 0.867 \times \text{Income} + 0.332 \times \text{Education}$$

$$(3.65) \quad (0.122) \quad (0.099)$$

$$R^2 = 0.88 \quad S_E = 11.4$$

which, as expected from the added-variable plots, has a larger income slope and smaller education slope than the original regression. The coefficient standard errors are likely optimistic, however, because relative outliers have been trimmed away. Deleting *railroad engineers*, along with *ministers* and *conductors*, further increases the income slope and decreases the education slope, but the change is not dramatic: $B_{\text{Income}} = 0.931$, $B_{\text{Education}} = 0.285$.

Added-variable plots can be straightforwardly extended to pairs of regressors in a model with more than two X s. We can, for example, regress each of X_1 , X_2 , and Y on the remaining regressors, X_3, \dots, X_k , obtaining residuals $X_{i1}^{(12)}$, $X_{i2}^{(12)}$, and $Y_i^{(12)}$. We then plot $Y^{(12)}$ against $X_1^{(12)}$ and $X_2^{(12)}$ to produce a dynamic three-dimensional scatterplot on which the partial-regression plane can be displayed.²⁹

11.6.2 Forward Search

Atkinson and Riani (2000) suggest a fundamentally different approach, termed a *forward search*, for locating multiple unusual observations: They begin by fitting a regression model to a small subset of the data that is almost surely free of outliers and then proceed to add observations one at a time to this subset, refitting the model at each step and monitoring regression outputs such as coefficients, t -statistics, residuals, hat-values, and Cook's distances.

To implement the forward search, Atkinson and Riani (2000) begin with a robust-regression fit to the data, employing a method that is highly resistant to outliers.³⁰ Residuals from this

²⁹See Cook and Weisberg (1989) for a discussion of three-dimensional added-variable plots. An alternative, two-dimensional extension of added-variable plots to subsets of coefficients is described in Section 11.8.4.

³⁰The method that they employ, *least median of squares* (or *LMS*) regression, is similar in its properties to *least-trimmed-squares* (*LTS*) regression, which is described in Chapter 19 on robust regression.

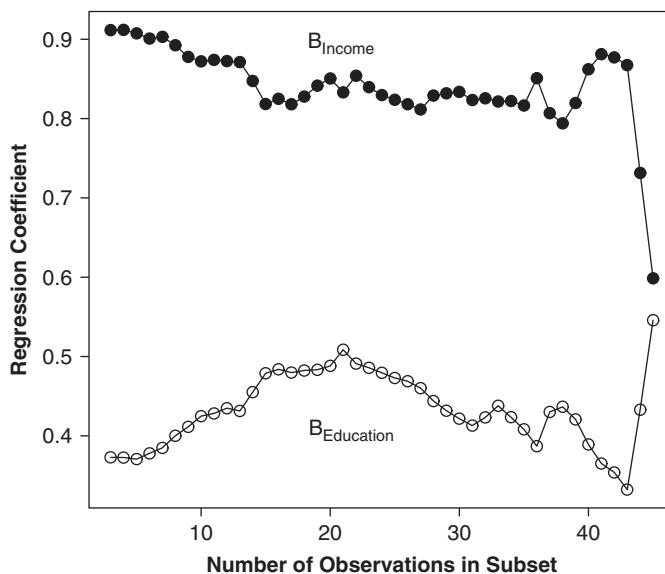


Figure 11.11 Forward-search trajectories of the coefficients for income and education in Duncan's occupational prestige regression. The points added at the last two steps are for *conductors* and *ministers*.

resistant fit are computed, and (in a model with $k + 1$ regression coefficients) the $k + 1$ observations with the smallest residuals are selected. The least-squares regression coefficients are then computed for the initial subset of observations, and residuals from this least-squares fit are computed for all n observations. Because there are equal numbers of observations and parameters at the first step, the residuals for the $k + 1$ observations employed to obtain the initial fit are necessarily 0.³¹ The additional observation with the next smallest residual is added to the subset, the least-squares regression coefficients are recomputed for the resulting $k + 2$ observations, and new residuals are found from the updated fit. Suppose, at any step, that there are m observations in the subset used to compute the current fit: The $m + 1$ observations used in the subsequent step are those with the smallest residuals from the current fit; usually, but not necessarily, these will include the m observations used to determine the current least-squares fit.

Figure 11.11 applies the forward search to Duncan's occupational prestige regression, monitoring the trajectory of the two slope coefficients as observations are added to an initial subset of $k + 1 = 3$ occupations. It is clear from this graph that although the income and education coefficients are nearly identical to one another in the least-squares fit to all 45 observations (at the far right),³² this result depends on the presence of just two observations, which enter in the last two steps of the forward search. It should come as no surprise that these observations are *conductors* and *ministers*. In this instance, therefore, the jointly influential observations were also revealed by more conventional, and less computationally intensive, methods.

³¹Care must be taken that the initial subset of $k + 1$ observations is not perfectly collinear.

³²Because both income and education are percentages (of, recall, relatively high-income earners and high school graduates), it makes at least superficial sense to compare their coefficients in this manner.

Atkinson and Riani's forward search adds observations successively to an initial small subset that is almost surely uncontaminated by unusual data. By monitoring outputs such as regression coefficients, this strategy can reveal unusual groups of observations that are missed by more conventional methods.

11.7 Should Unusual Data Be Discarded?

The discussion thus far in this chapter has implicitly assumed that outlying and influential data are simply discarded. In practice, although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

- It is important to investigate *why* an observation is unusual. Truly “bad” data (e.g., an error in data entry as in Davis’s data on measured and reported weight) can often be corrected or, if correction is not possible, thrown away. When a discrepant data point is correct, we may be able to understand why the observation is unusual. For Duncan’s regression, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation and for a reason not shared by other occupations. In a case like this, where an outlying observation has characteristics that render it unique, we may choose to set it aside from the rest of the data.
- Alternatively, outliers, high-leverage points, or influential data may motivate model respecification, and the pattern of unusual data may suggest the introduction of additional explanatory variables. We noticed, for example, that both conductors and railroad engineers had high leverage in Duncan’s regression because these occupations combined relatively high income with relatively low education. Perhaps this combination of characteristics is due to a high level of unionization of these occupations in 1950, when the data were collected. If so, and if we can ascertain the levels of unionization of all of the occupations, we could enter this as an explanatory variable, perhaps shedding further light on the process determining occupational prestige.³³ Furthermore, in some instances, transformation of the response variable or of an explanatory variable may draw apparent outliers toward the rest of the data, by rendering the error distribution more symmetric or by eliminating nonlinearity. We must, however, be careful to avoid “overfitting” the data—permitting a small portion of the data to determine the form of the model.³⁴
- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data. Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously down-weights outlying data rather than simply discarding them. Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not

³³This example is entirely speculative, but I mean simply to illustrate how unusual data can suggest respecification of a regression model.

³⁴See the discussion of nonlinearity in Chapter 12.

very different from careful application of least squares, and, indeed, robust-regression weights can be used to identify outliers.³⁵

- Finally, in large samples, unusual data substantially alter the results only in extreme instances. Identifying unusual observations in a large sample, therefore, should be regarded more as an opportunity to learn something about the data not captured by the model that we have fit, rather than as an occasion to reestimate the model with the unusual observations removed.

Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. “Bad” data can often be corrected. “Good” observations that are unusual may provide insight into the structure of the data and may motivate respecification of the statistical model used to summarize the data.

11.8 Some Statistical Details*

11.8.1 Hat-Values and the Hat-Matrix

Recall, from Chapter 9, the matrix form of the general linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The fitted model is given by $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} + \mathbf{e}$, in which the vector of least-squares estimates is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The least-squares fitted values are therefore a linear function of the observed response-variable values:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Here, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the *hat-matrix*, so named because it transforms \mathbf{y} into $\hat{\mathbf{y}}$ (“y-hat”). The hat-matrix is symmetric ($\mathbf{H} = \mathbf{H}'$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$), as can easily be verified.³⁶ Consequently, the diagonal entries of the hat-matrix $h_i \equiv h_{ii}$, which we called the *hat-values*, are

$$h_i \equiv \mathbf{h}_i'\mathbf{h}_i = \sum_{j=1}^n h_{ij}^2 = h_i^2 + \sum_{j \neq i} h_{ij}^2 \quad (11.7)$$

where (because of symmetry) the elements of \mathbf{h}_i comprise both the i th row and the i th column of \mathbf{H} .

Equation 11.7 implies that $0 \leq h_i \leq 1$. If the model matrix \mathbf{X} includes the constant regressor $\mathbf{1}_n$, then $1/n \leq h_i$. Because \mathbf{H} is a projection matrix,³⁷ projecting \mathbf{y} orthogonally onto the $(k+1)$ -dimensional subspace spanned by the columns of \mathbf{X} , it follows that $\sum h_i = k+1$, and thus $\bar{h} = (k+1)/n$ (as stated in Section 11.2).

³⁵See Chapter 19 on robust regression.

³⁶See Exercise 11.2.

³⁷See Chapter 10 for the vector geometry of linear models.

I mentioned as well that when there are several explanatory variables in the model, the leverage h_i of the i th observation is directly related to the distance of this observation from the center of the explanatory-variable point cloud. To demonstrate this property of the hat-values, it is convenient to rewrite the fitted model with all variables in mean deviation form: $\mathbf{y}^* = \mathbf{X}^* \mathbf{b}_1 + \mathbf{e}$, where $\mathbf{y}^* \equiv \{Y_i - \bar{Y}\}$ is the “centered” response-variable vector; $\mathbf{X}^* \equiv \{X_{ij} - \bar{X}_j\}$ contains the centered explanatory variables, but no constant regressor, which is no longer required; and \mathbf{b}_1 is the vector of least-squares slopes (suppressing the regression intercept). Then the hat-value for the i th observation is

$$h_i^* = \mathbf{h}_i^{*'} \mathbf{h}_i^* = \mathbf{x}_i^{*'} (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{x}_i^* = h_i - \frac{1}{n}$$

where $\mathbf{x}_i^{*'} = [X_{i1} - \bar{X}_1, \dots, X_{ik} - \bar{X}_k]$ is the i th row of \mathbf{X}^* (and \mathbf{x}_i^* is the i th row of \mathbf{X}^* written as a column vector).

As Weisberg (1985, p. 112) has pointed out, $(n - 1)h_i^*$ is the *generalized* or *Mahalanobis distance* between \mathbf{x}_i' and $\bar{\mathbf{x}}'$, where $\bar{\mathbf{x}}' = [\bar{X}_1, \dots, \bar{X}_k]$ is the mean vector or *centroid* of the explanatory variables. The Mahalanobis distances, and hence the hat-values, do not change if the explanatory variables are rescaled. Indeed, the Mahalanobis distances and hat-values are invariant with respect to any nonsingular linear transformation of \mathbf{X} .

11.8.2 The Distribution of the Least-Squares Residuals

The least-squares residuals are given by

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \end{aligned}$$

Thus,

$$E(\mathbf{e}) = (\mathbf{I} - \mathbf{H})E(\boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\mathbf{0} = \mathbf{0}$$

and

$$V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})V(\boldsymbol{\varepsilon})(\mathbf{I} - \mathbf{H})' = \sigma_\varepsilon^2(\mathbf{I} - \mathbf{H})$$

because $\mathbf{I} - \mathbf{H}$, like \mathbf{H} itself, is symmetric and idempotent. The matrix $\mathbf{I} - \mathbf{H}$ is not diagonal, and therefore the residuals are generally correlated, even when the errors are (as assumed here) independent. The diagonal entries of $\mathbf{I} - \mathbf{H}$ generally differ from one another, and so the residuals generally have different variances (as stated in Section 11.3):³⁸ $V(e_i) = \sigma_\varepsilon^2(1 - h_i)$.

11.8.3 Deletion Diagnostics

Let $\mathbf{b}_{(-i)}$ denote the vector of least-squares regression coefficients calculated with the i th observation omitted. Then, $\mathbf{d}_i \equiv \mathbf{b} - \mathbf{b}_{(-i)}$ represents the influence of observation i on the regression coefficients. The influence vector \mathbf{d}_i can be calculated efficiently as³⁹

³⁸Balanced ANOVA models are an exception: Here, all the hat-values are equal. See Exercise 11.3.

³⁹See Exercise 11.4.

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \frac{E_i}{1 - h_i} \quad (11.8)$$

where \mathbf{x}'_i is the i th row of the model matrix \mathbf{X} (and \mathbf{x}_i is the i th row written as a column vector).

Cook's D_i is the F -statistic for testing the "hypothesis" that $\boldsymbol{\beta} = \mathbf{b}_{(-i)}$:

$$\begin{aligned} D_i &= \frac{(\mathbf{b} - \mathbf{b}_{(-i)})' \mathbf{X}' \mathbf{X} (\mathbf{b} - \mathbf{b}_{(-i)})}{(k+1) S_E^2} \\ &= \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})' (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{(k+1) S_E^2} \end{aligned}$$

An alternative interpretation of D_i , therefore, is that it measures the aggregate influence of observation i on the fitted values $\hat{\mathbf{y}}$. This is why Belsley et al. (1980) call their similar statistic "DFFITS." Using Equation 11.8,

$$\begin{aligned} D_i &= \frac{E_i^2}{S_E^2(k+1)} \times \frac{h_i}{(1-h_i)^2} \\ &= \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i} \end{aligned}$$

which is the formula for Cook's D given in Section 11.4.

11.8.4 Added-Variable Plots and Leverage Plots

In vector form, the fitted multiple-regression model is

$$\begin{aligned} \mathbf{y} &= A\mathbf{1}_n + B_1\mathbf{x}_1 + B_2\mathbf{x}_2 + \cdots + B_k\mathbf{x}_k + \mathbf{e} \\ &= \hat{\mathbf{y}} + \mathbf{e} \end{aligned} \quad (11.9)$$

where the fitted-value vector $\hat{\mathbf{y}}$ is the orthogonal projection of \mathbf{y} onto the subspace spanned by the regressors $\mathbf{1}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$.⁴⁰ Let $\mathbf{y}^{(1)}$ and $\mathbf{x}^{(1)}$ be the projections of \mathbf{y} and \mathbf{x}_1 , respectively, onto the orthogonal complement of the subspace spanned by $\mathbf{1}_n$ and $\mathbf{x}_2, \dots, \mathbf{x}_k$ (i.e., the residual vectors from the least-squares regressions of Y and X_1 on the other X s). Then, by the geometry of projections, the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is $B_1\mathbf{x}^{(1)}$, and $\mathbf{y}^{(1)} - B_1\mathbf{x}^{(1)} = \mathbf{e}$, the residual vector from the overall least-squares regression, given in Equation 11.9.⁴¹

Sall (1990) suggests the following generalization of added-variable plots, which he terms *leverage plots*: Consider the general linear hypothesis⁴²

$$H_0: \underset{(q \times k+1)}{\mathbf{L}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} = \underset{(q \times 1)}{\mathbf{0}} \quad (11.10)$$

For example, in the regression of occupational prestige (Y) on education (X_1), income (X_2), and type of occupation (represented by the dummy regressors D_1 and D_2),

⁴⁰See Chapter 10.

⁴¹See Exercises 11.5 and 11.6.

⁴²See Section 9.4.3.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$$

the hypothesis matrix

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

is used to test the hypothesis $H_0: \gamma_1 = \gamma_2 = 0$ that there is no partial effect of type of occupation.

The residuals for the full model, unconstrained by the hypothesis in Equation 11.10, are the usual least-squares residuals, $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$. The estimated regression coefficients under the hypothesis are⁴³

$$\mathbf{b}_0 = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

and the residuals constrained by the hypothesis are given by

$$\mathbf{e}_0 = \mathbf{e} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

where

$$\mathbf{u} = \mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{b}$$

Thus, the incremental sum of squares for H_0 is⁴⁴

$$\|\mathbf{e}_0 - \mathbf{e}\|^2 = \mathbf{b}'\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{L}\mathbf{b}$$

The leverage plot is a scatterplot with

$$\mathbf{v}_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\mathbf{u}$$

on the horizontal axis, and

$$\mathbf{v}_y = \mathbf{v}_x + \mathbf{e}$$

on the vertical axis. The leverage plot, so defined, has the following properties:

- The residuals around the horizontal line at $V_y = 0$ are the constrained least-squares residuals \mathbf{e}_0 under the hypothesis H_0 .
- The least-squares line fit to the leverage plot has an intercept of 0 and a slope of 1; the residuals around this line are the unconstrained least-squares residuals, \mathbf{e} . The incremental sum of squares for H_0 is thus the regression sum of squares for the line.
- When the hypothesis matrix \mathbf{L} is formulated with a single row to test the coefficient of an individual regressor, the leverage plot specializes to the usual added-variable plot, with the horizontal axis rescaled so that the least-squares intercept is 0 and the slope 1.

Leverage plots, however, have the following disquieting property, which limits their usefulness: Even when an observation strongly influences the regression coefficients in a hypothesis, it may not influence the sum of squares for the hypothesis. For example, removing a particular

⁴³For this and other results pertaining to leverage plots, see Sall (1990).

⁴⁴See Exercise 11.7.

observation might increase a formerly small regression coefficient and decrease a formerly large one, so that the F -statistic for the hypothesis that both coefficients are zero is unaltered.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 11.1. *Show that, in simple-regression analysis, the hat-value is

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

[Hint: Evaluate $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ for $\mathbf{x}_i' = (1, X_i)$.]

Exercise 11.2. *Show that the hat-matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is symmetric ($\mathbf{H} = \mathbf{H}'$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$).

Exercise 11.3. *Show that in a one-way ANOVA with equal numbers of observations in the several groups, all the hat-values are equal to each other. By extension, this result implies that the hat-values in any balanced ANOVA are equal. Why?

Exercise 11.4. *Using Duncan's regression of occupational prestige on the educational and income levels of occupations, verify that the influence vector for the deletion of *ministers* on the regression coefficients, $\mathbf{d}_i = \mathbf{b} - \mathbf{b}_{(-i)}$, can be written as

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{E_i}{1 - h_i}$$

where \mathbf{x}_i is the i th row of the model matrix \mathbf{X} (i.e., the row for *ministers*) written as a column. [A much more difficult problem is to show that this formula works in general; see, e.g., Belsley, et al. (1980, pp. 69–83) or Velleman and Welsch (1981).]

Exercise 11.5. *Consider the two-explanatory-variable linear-regression model, with variables written as vectors in mean deviation form (as in Section 10.4): $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \mathbf{e}$. Let $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ represent the residual vectors from the regression (i.e., orthogonal projection) of \mathbf{x}_1^* and \mathbf{y}^* , respectively, on \mathbf{x}_2^* . Drawing the three-dimensional diagram of the subspace spanned by \mathbf{x}_1^* , \mathbf{x}_2^* , and \mathbf{y}^* , prove geometrically that the coefficient for the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is B_1 .

Exercise 11.6. *Extending the previous exercise, now consider the more general model $\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \cdots + B_k\mathbf{x}_k^* + \mathbf{e}$. Let $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ represent the residual vectors from the projections of \mathbf{x}_1^* and \mathbf{y}^* , respectively, onto the subspace spanned by $\mathbf{x}_2^*, \dots, \mathbf{x}_k^*$. Prove that the coefficient for the orthogonal projection of $\mathbf{y}^{(1)}$ onto $\mathbf{x}^{(1)}$ is B_1 .

Exercise 11.7. *Show that the incremental sum of squares for the general linear hypothesis $H_0: \mathbf{L}\beta = \mathbf{0}$ can be written as

$$\|\mathbf{e}_0 - \mathbf{e}\|^2 = \mathbf{b}'\mathbf{L}'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}\mathbf{L}\mathbf{b}$$

[Hint: $\|\mathbf{e}_0 - \mathbf{e}\|^2 = (\mathbf{e}_0 - \mathbf{e})'(\mathbf{e}_0 - \mathbf{e})$.]

Summary

- Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis and because they may indicate that the model fails to capture important features of the data.
- Observations with unusual combinations of explanatory-variable values have high *leverage* in a least-squares regression. The hat-values h_i provide a measure of leverage. A rough cutoff for noteworthy hat-values is $h_i > 2\bar{h} = 2(k + 1)/n$.
- A regression *outlier* is an observation with an unusual response-variable value given its combination of explanatory-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination, a Bonferroni test for the largest absolute $|E_i^*|$, or Anscombe's insurance analogy. If the model is correct (and there are no "bad" observations), then each studentized residual follows a t -distribution with $n - k - 2$ degrees of freedom.
- Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D statistic provides a summary index of influence on the coefficients. A rough cutoff for noteworthy values of D is $D_i > 4/(n - k - 1)$.
- It is also possible to investigate the influence of individual observations on other regression "outputs," such as coefficient standard errors and collinearity.
- Subsets of observations can be jointly influential. Added-variable plots are useful for detecting joint influence on the regression coefficients. The added-variable plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all the other X 's.
- Atkinson and Riani's forward search adds observations successively to an initial small subset that is almost surely uncontaminated by unusual data. By monitoring outputs such as regression coefficients, this strategy can reveal unusual groups of observations that are missed by more conventional methods.
- Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. "Bad" data can often be corrected. "Good" observations that are unusual may provide insight into the structure of the data and may motivate respecification of the statistical model used to summarize the data.

Recommended Reading

There is a large journal literature on methods for identifying unusual and influential data. Fortunately, there are several texts that present this literature in a more digestible form:⁴⁵

- Although it is now more than three decades old, Cook and Weisberg (1982) is, in my opinion, still the best book-length presentation of methods for assessing leverage, outliers, and influence. There are also good discussions of other problems, such as nonlinearity and transformations of the response and explanatory variables.

⁴⁵Also see the recommended readings given at the end of the following chapter.

- Chatterjee and Hadi (1988) is a thorough text dealing primarily with influential data and collinearity; other problems—such as nonlinearity and nonconstant error variance—are treated briefly.
- Belsley, Kuh, and Welsch (1980) is a seminal text that discusses influential data and the detection of collinearity.⁴⁶
- Barnett and Lewis (1994) present an encyclopedic survey of methods for outlier detection, including methods for detecting outliers in linear models.
- Atkinson and Riani (2000) describe in detail methods for detecting influential data based on a “forward search”; these methods were presented briefly in Section 11.6.2.

⁴⁶I believe that Belsley et al.’s (1980) approach to diagnosing collinearity is fundamentally flawed—see the discussion of collinearity in Chapter 13.

12

Diagnosing Non-Normality, Nonconstant Error Variance, and Nonlinearity

Chapters 11, 12, and 13 show how to detect and correct problems with linear models that have been fit to data. The previous chapter focused on problems with specific observations. The current chapter and the next deal with more general problems with the specification of the model.

The first three sections of this chapter take up the problems of non-normally distributed errors, nonconstant error variance, and nonlinearity. The treatment here stresses simple graphical methods for detecting these problems, along with transformations of the data to correct problems that are detected.

Subsequent sections describe tests of nonconstant error variance and nonlinearity for discrete explanatory variables, diagnostic methods based on embedding the usual linear model in a more general nonlinear model that incorporates transformations as parameters, and diagnostics that seek to detect the underlying dimensionality of the regression.

To illustrate the methods described in this chapter, I will primarily use data drawn from the 1994 wave of Statistics Canada's Survey of Labour and Income Dynamics (SLID). The SLID data set includes 3,997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.¹ Regressing the composite hourly wage rate (i.e., the wage rate computed from all sources of employment, in dollars per hour) on a dummy variable for sex (code 1 for males), education (in years), and age (also in years) produces the following results:

$$\begin{aligned}\widehat{\text{Wages}} = & -8.124 + 3.474 \times \text{Male} & + 0.2613 \times \text{Age} \\ & (0.599) \quad (0.2070) & (0.0087) \\ & + 0.9296 \times \text{Education} & \\ & (0.0343) & \end{aligned} \tag{12.1}$$

$$R^2 = .3074$$

The coefficient standard errors, in parentheses below the coefficients, reveal that all the regression coefficients are precisely estimated (and highly statistically significant), as is to be

¹I assumed that individuals for whom the composite hourly wage rate is missing are not employed. There are, in addition, 150 people who are missing data on education and who are excluded from the analysis reported here.

expected in a sample of this size. The regression also accounts for more than 30% of the variation in hourly wages.

Although we will get quite a bit of mileage from this example, it is somewhat artificial: (1) A careful data analyst (using the methods for examining and transforming data introduced in Chapters 3 and 4) would not specify the model in this form. Indeed, on substantive grounds, we should not expect linear relationships between wages and age and, possibly, between wages and education.² (2) We should entertain the obvious possibility that the effects of age and education on income may be different for men and women (i.e., that sex may interact with age and education), a possibility that we will pursue later in the chapter. (3) A moderately large sample such as this presents the opportunity to introduce additional explanatory variables into the analysis.

12.1 Non-Normally Distributed Errors

The assumption of normally distributed errors is almost always arbitrary. Nevertheless, the central limit theorem ensures that, under very broad conditions, inference based on the least-squares estimator is approximately valid in all but small samples. Why, then, should we be concerned about non-normal errors?

- Although the *validity* of least-squares estimation is robust—the levels of tests and the coverage of confidence intervals are approximately correct in large samples even when the assumption of normality is violated—the *efficiency* of least squares is not robust: Statistical theory assures us that the least-squares estimator is the most efficient unbiased estimator only when the errors are normal. For some types of error distributions, however, particularly those with heavy tails, the efficiency of least-squares estimation decreases markedly. In these cases, the least-squares estimator becomes much less efficient than robust estimators (or least-squares augmented by diagnostics).³ To a great extent, heavy-tailed error distributions are problematic because they give rise to outliers, a problem that I addressed in the previous chapter.

A commonly quoted justification of least-squares estimation—the Gauss-Markov theorem—states that the least-squares coefficients are the most efficient unbiased estimators that are *linear* functions of the observations Y_i . This result depends on the assumptions of linearity, constant error variance, and independence but does not require the assumption of normality.⁴ Although the restriction to linear estimators produces simple formulas for coefficient standard errors, it is not compelling in the light of the vulnerability of least squares to heavy-tailed error distributions.

- Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit. This fit is a conditional mean (of Y given the X s), and the mean is not a good measure of the center of a highly skewed distribution. Consequently, we may prefer to transform the response to produce a symmetric error distribution.

²Unfortunately, in my experience, careful data analysis is far from the norm, and it is not hard to find examples of egregiously misspecified regressions with large R^2 's that satisfied the people who performed them.

³Robust estimation is discussed in Chapter 19.

⁴A proof of the Gauss-Markov theorem appears in Section 9.3.2.

- A multimodal error distribution suggests the omission of one or more discrete explanatory variables that divide the data naturally into groups. An examination of the distribution of the residuals may, therefore, motivate respecification of the model.

Although there are tests for non-normal errors, I will instead describe graphical methods for examining the distribution of the residuals, employing univariate displays introduced in Chapter 3.⁵ These methods are more useful than tests for pinpointing the nature of the problem and for suggesting solutions.

One such graphical display is the quantile-comparison plot. Recall from the preceding chapter that we compare the sample distribution of the studentized residuals, E_i^* , with the quantiles of the unit-normal distribution, $N(0, 1)$, or with those of the t -distribution for $n - k - 2$ degrees of freedom. Unless n is small, of course, the normal and t -distributions are nearly identical. We choose to plot *studentized* residuals because they have equal variances and are t -distributed, but, in larger samples, standardized or raw residuals will convey much the same impression.

Even if the model is correct, however, the studentized residuals are not an *independent* random sample from t_{n-k-2} : Different residuals are correlated with one another.⁶ These correlations depend on the configuration of the X -values, but they are generally negligible unless the sample size is small. Furthermore, at the cost of some computation, it is possible to adjust for the dependencies among the residuals in interpreting a quantile-comparison plot.⁷

The quantile-comparison plot is especially effective in displaying the tail behavior of the residuals: Outliers, skewness, heavy tails, or light tails all show up clearly. Other univariate graphical displays effectively supplement the quantile-comparison plot. In large samples, a histogram with many bars conveys a good impression of the shape of the residual distribution and generally reveals multiple modes more clearly than does the quantile-comparison plot. In smaller samples, a more stable impression is formed by smoothing the histogram of the residuals with a nonparametric density estimator (which is also a reasonable display in large samples).

Figure 12.1 shows the distribution of the studentized residuals from the SLID regression of Equation 12.1. The broken lines in the quantile-comparison plot [Figure 12.1(a)] represent a pointwise 95% confidence envelope computed under the assumption that the errors are normally distributed (according to the method described in Section 12.1.1). The window width for the kernel density estimate [Figure 12.1(b)] is 3/4 of the “optimal” value for normally distributed data and was selected by visual trial and error. It is clear from both graphs that the residual distribution is positively skewed. The density estimate suggests, in addition, that there may be more than one mode to the distribution.

A positive skew in the residuals can usually be corrected by moving the *response variable* down the ladder of powers and roots. Trial and error suggests that the log transformation of wages renders the distribution of the residuals much more symmetric, as shown in Figure 12.2.

⁵See the discussion of Box-Cox transformations in Section 12.5.1, however.

⁶*Different residuals are correlated because the off-diagonal entries of the hat-matrix (i.e., h_{ij} for $i \neq j$) are generally nonzero; see Section 11.8.

⁷See Section 12.1.1.

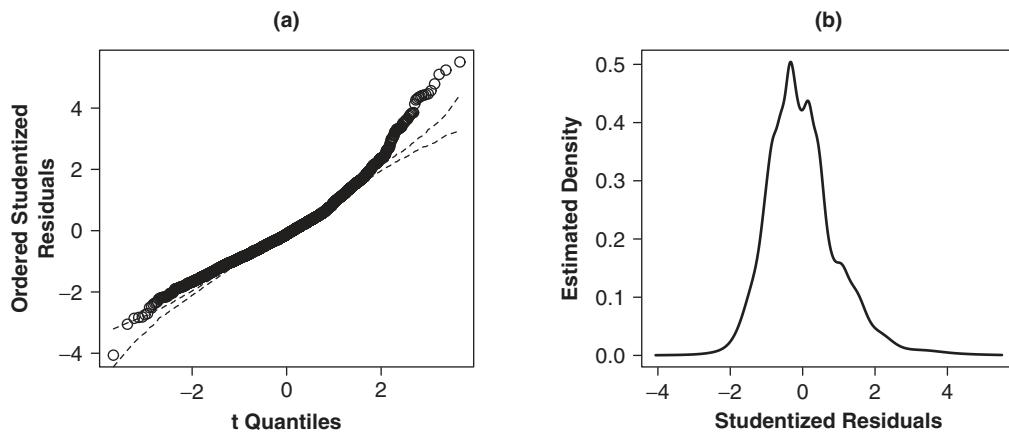


Figure 12.1 (a) Quantile-comparison plot and (b) kernel-density estimate for the studentized residuals from the SLID regression. The broken lines in the quantile-comparison plot represent a pointwise 95% simulated confidence envelope (described in Section 12.1.1).

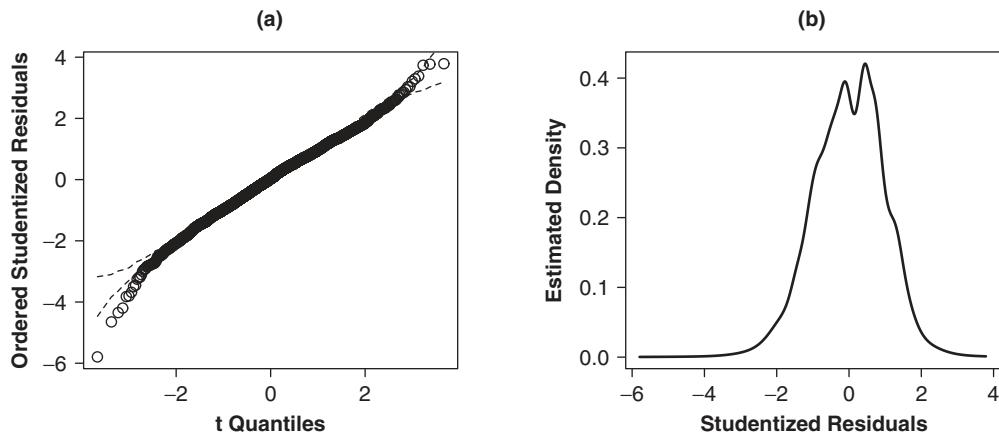


Figure 12.2 (a) Quantile-comparison plot and (b) kernel-density estimate for the studentized residuals from the SLID regression with wages log-transformed.

The residuals from the transformed regression appear to be heavy-tailed, a characteristic that in a small sample would lead us to worry about the efficiency of the least-squares estimator.⁸

A cube-root transformation (i.e., the 1/3 power, not shown) does an even better job (reducing the long left tail of the residual distribution in Figure 12.2), but because it produces very

⁸Heavy-tailed error distributions suggest robust estimation, as described in Chapter 19, but in a sample this large, robust regression produces essentially the same results as least squares.

similar results in the regression, I prefer the more easily interpreted log transformation. The fitted regression equation using the log (base 2) of wages is as follows:

$$\begin{aligned}\widehat{\log_2 \text{Wages}} &= 1.585 + 0.3239 \times \text{Male} + 0.02619 \times \text{Age} \\ &\quad (0.055) \quad (0.0189) \quad (0.00079) \\ &\quad + 0.08061 \times \text{Education} \\ &\quad (0.00313)\end{aligned}\tag{12.2}$$

$R^2 = 0.3213$

We will return to the interpretation of the regression with the log-transformed response after we fix some other problems with the SLID model.⁹

Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multi-modal errors compromise the interpretation of the least-squares fit. Non-normality can often be detected by examining the distribution of the least-squares residuals and frequently can be corrected by transforming the data.

12.1.1 Confidence Envelopes by Simulated Sampling*

Atkinson (1985) has suggested the following procedure for constructing an approximate confidence “envelope” in a residual quantile-comparison plot, taking into account the correlational structure of the explanatory variables. Atkinson’s procedure employs *simulated sampling* and uses the assumption of normally distributed errors.¹⁰

1. Fit the regression model as usual, obtaining fitted values \widehat{Y}_i and the estimated regression standard error S_E .
2. Construct m samples, each consisting of n simulated Y values; for the j th such sample, the simulated value for observation i is

$$Y_{ij}^s = \widehat{Y}_i + S_E Z_{ij}$$

where Z_{ij} is a random draw from the unit-normal distribution. In other words, we sample from a “population” in which the expectation of Y_i is \widehat{Y}_i ; the true standard deviation of the errors is S_E ; and the errors are normally distributed.

3. Regress the n simulated observations for sample j on the X s in the original sample, obtaining simulated studentized residuals, $E_{1j}^*, E_{2j}^*, \dots, E_{nj}^*$. Because this regression employs the original X -values, the simulated studentized residuals reflect the correlational structure of the X s.

⁹For the present, recall that increasing \log_2 wages by 1 implies doubling wages; more generally, adding x to \log_2 wages multiplies wages by 2^x .

¹⁰The notion of simulated sampling from a population constructed from the observed data is the basis of “bootstrapping,” discussed in Chapter 21. Atkinson’s procedure described here is an application of the *parametric bootstrap*.

4. Order the studentized residuals for sample j from smallest to largest, as required by a quantile-comparison plot: $E_{(1)j}^*, E_{(2)j}^*, \dots, E_{(n)j}^*$.
5. To construct an estimated $(100 - a)\%$ confidence interval for $E_{(i)}^*$ (the i th ordered studentized residual), find the $a/2$ and $1 - a/2$ empirical quantiles of the m simulated values $E_{(i)1}^*, E_{(i)2}^*, \dots, E_{(i)m}^*$. For example, if $m = 20$ and $a = .05$, then the smallest and largest of the $E_{(1)j}^*$ provide a 95% confidence interval for $E_{(1)}^*$: $[E_{(1)(1)}^*, E_{(1)(20)}^*]$.¹¹ The confidence limits for the n ordered studentized residuals are graphed as a confidence envelope on the quantile-comparison plot, along with the studentized residuals themselves.

A weakness of Atkinson's procedure is that the probability of *some* studentized residual straying outside the confidence limits by chance is greater than a , which is the probability that an *individual* studentized residual falls outside its confidence interval. Because the joint distribution of the studentized residuals is complicated, however, to construct a correct joint-confidence envelope would require even more calculation. As well, in small samples, where there are few residual degrees of freedom, even radical departures from normally distributed errors can give rise to apparently normally distributed residuals; Andrews (1979) presents an example of this phenomenon, which he terms "supernormality."

12.2 Nonconstant Error Variance

As we know, one of the assumptions of the regression model is that the variation of the response variable around the regression surface—the error variance—is everywhere the same:

$$V(\varepsilon) = V(Y|x_1, \dots, x_k) = \sigma_\varepsilon^2$$

Constant error variance is often termed *homoscedasticity*; similarly, *nonconstant* error variance is termed *heteroscedasticity*. Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, the efficiency of the least-squares estimator is impaired, and the usual formulas for coefficient standard errors are inaccurate—the degree of the problem depending on the degree to which error variances differ, the sample size, and the configuration of the X -values in the regression. In this section, I will describe graphical methods for detecting nonconstant error variances and methods for dealing with the problem when it is detected.¹²

12.2.1 Residual Plots

Because the regression surface is k -dimensional and embedded in a space of $k + 1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct

¹¹Selecting the smallest and largest of the 20 simulated values corresponds to our simple convention that the proportion of the data below the j th of m order statistics is $(j - 1/2)/m$. Here, $(1 - 1/2)/20 = .025$ and $(20 - 1/2)/20 = .975$, defining 95% confidence limits. Atkinson uses a slightly different convention. To estimate the confidence limits more accurately, it helps to make m larger and perhaps to use a more sophisticated version of the bootstrap (see Chapter 21). The envelopes in Figures 12.1(a) and 12.2(a) are based on $m = 100$ replications.

¹²Tests for heteroscedasticity are discussed in Section 12.4 on discrete data and in Section 12.5 on maximum-likelihood methods.

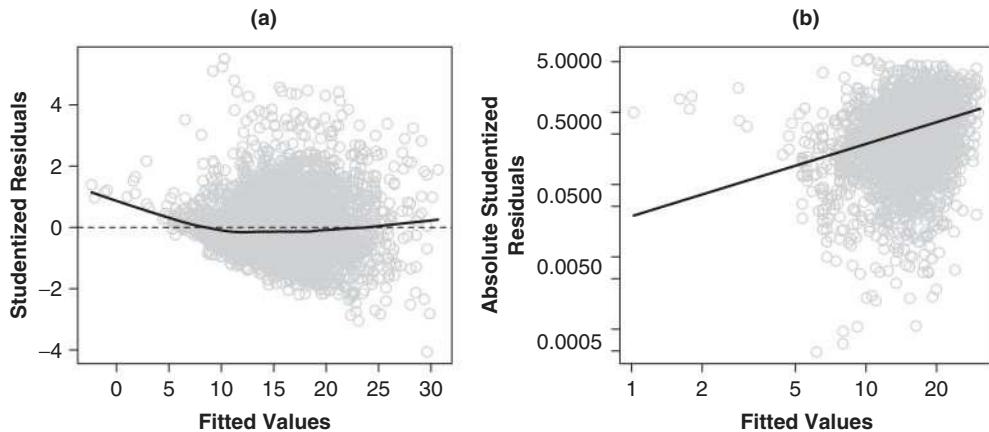


Figure 12.3 (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals. The solid line in panel (a) is fit by lowess, with a span of 0.4. The line in panel (b) is produced by robust linear regression. The points are plotted in gray to avoid obscuring the lines.

graphical examination of the data when k is larger than 1 or 2. Nevertheless, it is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X . The former situation can often be detected by plotting residuals against fitted values and the latter by plotting residuals against each X .¹³

Plotting residuals against Y (as opposed to \hat{Y}) is generally unsatisfactory because the plot is “tilted”: $Y = \hat{Y} + E$, and consequently the linear correlation between the *observed response* Y and the residuals E is $\sqrt{1 - R^2}$.¹⁴ In contrast, the least-squares fit ensures that the correlation between the *fitted values* \hat{Y} and E is precisely 0, producing a plot that is much easier to examine for evidence of nonconstant spread.

Because the least-squares *residuals* have unequal variances even when the assumption of constant *error* variance is correct, it is preferable to plot studentized residuals against fitted values. A pattern of changing spread is often more easily discerned in a plot of absolute studentized residuals, $|E_i^*|$, or squared studentized residuals, E_i^{*2} , against \hat{Y} . Finally, if the values of \hat{Y} are all positive, then we can plot $\log|E_i^*|$ (log spread) against $\log \hat{Y}$ (log level). A line, with slope b fit to this plot, suggests the variance-stabilizing transformation $Y^{(p)}$, with $p = 1 - b$.¹⁵

Figure 12.3 shows a plot of studentized residuals against fitted values and a spread-level plot of studentized residuals for the SLID regression of Equation 12.1 (page 296); several points with negative fitted values were omitted. It is apparent from both graphs that the residual spread tends to increase with the level of the response, suggesting transforming the response *down* the ladder of powers and roots. The slope of the line fit to the spread-level plot in

¹³These displays are not infallible, however. See Cook (1994) and the discussion in Section 12.6.

¹⁴See Exercise 12.1.

¹⁵This is an application of Tukey’s rule for selecting a transformation, introduced in Section 4.4. Other analytic methods for choosing a variance-stabilizing transformation are discussed in Section 12.5.

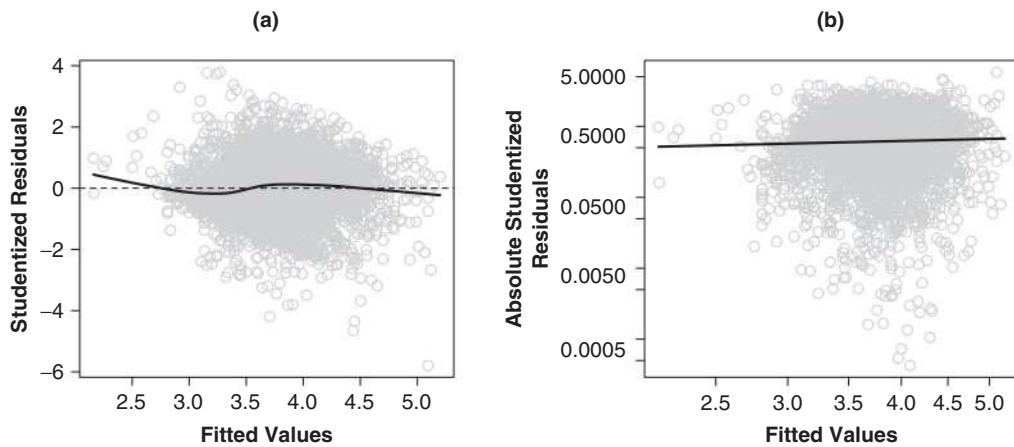


Figure 12.4 (a) Plot of studentized residuals versus fitted values and (b) spread-level plot for studentized residuals following log transformation of wages in the SLID regression.

Figure 12.3(b) is $b = 0.9994$, corresponding to the power transformation $1 - 0.9994 = 0.0006 \approx 0$ (i.e., the log transformation).¹⁶

In the previous section, the log transformation of wages made the distribution of the studentized residuals more symmetric. The same transformation approximately stabilizes the residual variance, as illustrated in diagnostic plots shown in Figure 12.4. This outcome is not surprising because the heavy right tail of the residual distribution and nonconstant spread are both common consequences of the lower bound of 0 for the response variable.

Transforming Y changes the shape of the error distribution, but it also alters the shape of the regression of Y on the X s. At times, eliminating nonconstant spread also makes the relationship of Y to the X s more nearly linear, but this is not a necessary consequence of stabilizing the error variance, and it is important to check for nonlinearity following transformation of the response variable. Of course, because there is generally no reason to suppose that the regression is linear *prior* to transforming Y , we should check for nonlinearity in any event.¹⁷

Nonconstant residual spread sometimes is symptomatic of the omission of important effects from the model. Suppose, for example, that there is an omitted categorical explanatory variable, such as urban versus rural residence, that interacts with education in affecting wages; in particular, suppose that the education slope, although positive in both urban and rural areas, is steeper in urban areas. Then the omission of urban/rural residence and its interaction with education could produce a fan-shaped residual plot even if the errors from the correct model have constant variance.¹⁸ The detection of this type of specification error requires insight into the process generating the data and cannot rely on diagnostics alone.

¹⁶The line in Figure 12.3(b) was fit by M estimation using the Huber weight function—a method of robust regression described in Chapter 19. In this example, however, nearly the same results are provided by least-squares, for which $b = 0.9579$. The plots in Figures 12.3(a) and 12.4(a) suggest that there is some unmodeled nonlinearity. Although plotting residuals versus fitted values may, as here, reveal a problem with the specified functional form of a regression model, the indication is insufficiently specific to know where precisely the problem or problems lie and how to deal with them, an issue to which we will turn in Section 12.3.

¹⁷See Section 12.3.

¹⁸See Exercise 12.2 for an illustration of this phenomenon.

12.2.2 Weighted-Least-Squares Estimation*

Weighted-least-squares (WLS) regression provides an alternative approach to estimation in the presence of nonconstant error variance. Suppose that the errors from the linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ are independent and normally distributed, with zero means but *different* variances: $\varepsilon_i \sim N(0, \sigma_i^2)$. Suppose further that the variances of the errors are known up to a constant of proportionality σ_ε^2 , so that $\sigma_i^2 = \sigma_\varepsilon^2/w_i^2$. Then, the likelihood for the model is¹⁹

$$L(\beta, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \right]$$

where Σ is the covariance matrix of the errors,

$$\Sigma = \sigma_\varepsilon^2 \times \text{diag}\{1/w_1^2, \dots, 1/w_n^2\} \equiv \sigma_\varepsilon^2 \times \mathbf{W}^{-1}$$

The maximum-likelihood estimators of β and σ_ε^2 are then

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y} \\ \hat{\sigma}_\varepsilon^2 &= \frac{\sum (w_i E_i)^2}{n}\end{aligned}$$

where the residuals are defined in the usual manner as the difference between the observed response and the fitted values, $\mathbf{e} = \{E_i\} = \mathbf{y} - \mathbf{X}\hat{\beta}$. This procedure is equivalent to minimizing the *weighted sum of squares* $\sum w_i^2 E_i^2$, according greater weight to observations with smaller variance—hence the term *weighted least squares*. The estimated asymptotic covariance matrix of $\hat{\beta}$ is given by

$$\hat{\mathcal{V}}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$$

In practice, we would need to estimate the weights W_i or know that the error variance is systematically related to some observable variable. In the first instance, for example, we could use the residuals from a preliminary *ordinary-least-squares* (OLS) regression to obtain estimates of the error variance within different subsets of observations, formed by partitioning the data according to one or more categorical variables. Basing the weights on a preliminary estimate of error variances can, however, seriously bias the estimated covariance matrix $\hat{\mathcal{V}}(\hat{\beta})$, because the sampling error in the estimates should reflect the additional source of uncertainty.²⁰

In the second instance, suppose that inspection of a residual plot for the preliminary OLS fit suggests that the magnitude of the errors is proportional to the first explanatory variable, X_1 . We can then use $1/X_{i1}$ as the weights W_i . Dividing both sides of the regression equation by X_{i1} produces

$$\frac{Y_i}{X_{i1}} = \alpha \frac{1}{X_{i1}} + \beta_1 + \beta_2 \frac{X_{i2}}{X_{i1}} + \cdots + \beta_k \frac{X_{ik}}{X_{i1}} + \frac{\varepsilon_i}{X_{i1}} \quad (12.3)$$

Because the standard deviations of the errors are proportional to X_1 , the “new” errors $\varepsilon'_i \equiv \varepsilon_i/X_{i1}$ have constant variance, and Equation 12.3 can be estimated by OLS regression of

¹⁹See Exercise 12.3 for this and other results pertaining to weighted-least-squares estimation; also see the discussion of generalized least squares in Section 16.1.

²⁰In this case, it is probably better to obtain an honest estimate of the coefficient covariance matrix from the bootstrap, described in Chapter 21, or to estimate the within-group variances simultaneously with the regression parameters.

Y/X_1 on $1/X_1, X_2/X_1, \dots, X_k/X_1$. Note that the constant from this regression estimates β_1 , while the coefficient of $1/X_1$ estimates α ; the remaining coefficients are straightforward.

12.2.3 Correcting OLS Standard Errors for Nonconstant Variance*

The covariance matrix of the OLS estimator is

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (12.4)$$

Under the standard assumptions, including the assumption of constant error variance, $V(\mathbf{y}) = \sigma_e^2 \mathbf{I}_n$, Equation 12.4 simplifies to the usual formula, $V(\mathbf{b}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1}$.²¹ If, however, the errors are heteroscedastic but independent, then $\Sigma \equiv V(\mathbf{y}) = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$, and

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Because $E(\varepsilon_i) = 0$, the variance of the i th error is $\sigma_i^2 = E(\varepsilon_i^2)$, which suggests the possibility of estimating $V(\mathbf{b})$ by

$$\tilde{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (12.5)$$

with $\hat{\Sigma} = \text{diag}\{E_1^2, \dots, E_n^2\}$, where E_i is the OLS residual for observation i . White (1980) shows that Equation 12.5 provides a consistent estimator of $V(\mathbf{b})$.²²

Subsequent work has suggested small modifications to White's coefficient-variance estimator, and in particular simulation studies by Long and Ervin (2000) support the use of

$$\tilde{V}^*(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^*\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (12.6)$$

where $\hat{\Sigma}^* = \text{diag}\{E_i^2/(1 - h_i)^2\}$ and h_i is the hat-value associated with observation i .²³ In large samples, in which individual hat-values are almost surely very small, the distinction between the coefficient-variance estimators in Equations 12.5 and 12.6 essentially disappears.

For the original SLID regression model in Equation 12.1 (on page 296), coefficient standard errors computed by the usual formula, by White's approach (in Equation 12.5), and by the modification to White's approach (in Equation 12.6) are as follows:

²¹See Section 9.3.

²²White's coefficient-variance estimator is sometimes called a *sandwich estimator* because the matrix $\mathbf{X}'\hat{\Sigma}\mathbf{X}$ is “sandwiched between” the two occurrences of $(\mathbf{X}'\mathbf{X})^{-1}$ in Equation 12.5. Coefficient standard errors computed by this approach are also often termed *Huber-White standard errors*, because of their introduction in an earlier paper by Huber (1967).

²³See Sections 11.2 and 11.8 for a discussion of hat-values. Long and Ervin call the coefficient-variance estimator in Equation 12.6 “HC3” for “heteroscedasticity-consistent” estimator number 3—one of several such estimators considered in their paper. Cribari-Neto (2004) suggests using

$$\hat{\Sigma}^* = \text{diag}\left\{E_i^2/(1 - h_i)^{d_i}\right\}$$

where $d_i = \min(4, h_i/\bar{h})$, producing a coefficient-covariance matrix estimator that he terms “HC4”; the HC4 estimator can perform better in small samples with influential observations.

Coefficient	Standard Error of Coefficient		
	Traditional OLS	White-Adjusted	Modified White-Adjusted
Constant	0.5990	0.6358	0.6370
Male	0.2070	0.2071	0.2074
Age	0.008664	0.008808	0.008821
Education	0.03426	0.03847	0.03854

In this instance, therefore, the adjusted standard errors are very close to the usual OLS standard errors—despite the strong evidence that we uncovered of nonconstant error variance.

An advantage of White's approach for coping with heteroscedasticity is that knowledge of the *pattern* of nonconstant error variance (e.g., increased variance with the level of Y or with an X) is not required. If, however, the heteroscedasticity problem is severe, and the corrected coefficient standard errors therefore are considerably larger than those produced by the usual formula, then discovering the pattern of nonconstant variance and taking account of it—by a transformation or WLS estimation—offers the possibility of more efficient estimation. In any event, as the next section shows, unequal error variance typically distorts statistical inference only when the problem is severe.

12.2.4 How Nonconstant Error Variance Affects the OLS Estimator*

The impact of nonconstant error variance on the efficiency of the OLS estimator and on the validity of least-squares inference depends on several factors, including the sample size, the degree of variation in the σ_i^2 , the configuration of the X -values, and the relationship between the error variance and the X 's. It is therefore not possible to develop wholly general conclusions concerning the harm produced by heteroscedasticity, but the following simple case is nevertheless instructive.

Suppose that $Y_i = \alpha + \beta X_i + \varepsilon_i$, where the errors are independent and normally distributed, with zero means but with different standard deviations proportional to X , so that $\sigma_i = \sigma_\varepsilon X_i$ (where all of the $X_i > 0$). Then the OLS estimator B is less efficient than the WLS estimator $\hat{\beta}$, which, under these circumstances, is the most efficient unbiased estimator of β .²⁴

Formulas for the sampling variances of B and $\hat{\beta}$ are easily derived.²⁵ The efficiency of the OLS estimator relative to the optimal WLS estimator is given by $V(\hat{\beta})/V(B)$, and the relative precision of the OLS estimator is the square root of this ratio, that is, $SD(\hat{\beta})/SD(B)$ (interpretable, e.g., as the relative width of confidence intervals for the two estimators).

Now suppose that X is uniformly distributed over the interval $[x_0, ax_0]$, where both x_0 and a are positive numbers, so that a is the ratio of the largest to the smallest value of X (and,

²⁴This property of the WLS estimator requires the assumption of normality. Without normal errors, the WLS estimator is still the most efficient *linear* unbiased estimator—an extension of the Gauss-Markov theorem. See Exercise 12.4.

²⁵ $V(B)$ here is the *correct* sampling variance of the OLS estimator, taking heteroscedasticity into account, *not* the variance of B computed by the usual formula, which assumes constant error variance. See Exercise 12.5 for this and other results described in this section.

consequently, of the largest to the smallest σ_i). The relative precision of the OLS estimator stabilizes quickly as the sample size grows and exceeds 90% when $a = 2$ and 85% when $a = 3$, even when n is as small as 20. For $a = 10$, the penalty for using OLS is greater, but even here the relative precision of OLS exceeds 65% for $n \geq 20$.

The validity of statistical inferences based on OLS estimation (as opposed to the efficiency of the OLS estimator) is even less sensitive to common patterns of nonconstant error variance. Here, we need to compare the expectation of the usual estimator $\widehat{V}(B)$ of $V(B)$, which is typically biased when the error variance is not constant, with the true sampling variance of B . The square root of $E[\widehat{V}(B)]/V(B)$ expresses the result in relative standard deviation terms. For the illustration, where the standard deviation of the errors is proportional to X , and where X is uniformly distributed, this ratio is 98% when $a = 2$, 97% when $a = 3$, and 93% when $a = 10$, all for $n \geq 20$.

The results in this section suggest that nonconstant error variance is a serious problem only when the magnitude (i.e., the standard deviation) of the errors varies by more than a factor of about 3—that is, when the largest error variance is more than about 10 times the smallest. Because there are other distributions of the X 's for which the deleterious effects of heteroscedasticity can be more severe, a safer rule of thumb is to worry about nonconstant error variance when the magnitude of the errors varies by more than a factor of about 2—or, equivalently, when the ratio of largest to smallest error variance exceeds 4. One cautious approach is always to compare (modified) White-adjusted coefficient standard errors with the usual standard errors, preferring the former (or, if the pattern is known, correcting for nonconstant error variance) when the two disagree.

It is common for the variance of the errors to increase with the level of the response variable. This pattern of nonconstant error variance (“heteroscedasticity”) can often be detected in a plot of residuals against fitted values. Strategies for dealing with nonconstant error variance include transformation of the response variable to stabilize the variance, the substitution of weighted-least-squares estimation for ordinary least squares, and the correction of coefficient standard errors for heteroscedasticity. A rough rule is that nonconstant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more (or, more conservatively, about 4 or more).

12.3 Nonlinearity

The assumption that the average error, $E(\varepsilon)$, is everywhere 0 implies that the specified regression surface accurately reflects the dependency of the conditional average value of Y on the X 's. Conversely, violating the assumption of linearity implies that the model fails to capture the systematic pattern of relationship between the response and explanatory variables. The term *nonlinearity*, therefore, is not used in the narrow sense here, although it includes the possibility that a partial relationship assumed to be linear is, in fact, nonlinear: If, for example, two explanatory variables specified to have additive effects instead interact, then the average error is not 0 for all combinations of X -values, constituting nonlinearity in the broader sense.

If nonlinearity, in the broad sense, is slight, then the fitted model can be a useful approximation even though the regression surface $E(Y|X_1, \dots, X_k)$ is not captured precisely. In other instances, however, the model can be seriously misleading.

The regression surface is generally high dimensional, even after accounting for regressors (such as dummy variables, interactions, polynomial terms, and regression-spline terms) that are functions of a smaller number of fundamental explanatory variables.²⁶ As in the case of non-constant error variance, therefore, it is necessary to focus on particular patterns of departure from linearity. The graphical diagnostics discussed in this section are two-dimensional (and three-dimensional) projections of the $(k+1)$ -dimensional point cloud of observations $\{Y_i, X_{i1}, \dots, X_{ik}\}$.

12.3.1 Component-Plus-Residual Plots

Although it is useful in multiple regression to plot Y against each X (e.g., in one row of a scatterplot matrix), these plots often do not tell the whole story—and can be misleading—because our interest centers on the *partial* relationship between Y and each X (“controlling” for the other X s), not on the *marginal* relationship between Y and an individual X (“ignoring” the other X s). Residual-based plots are consequently more promising for detecting nonlinearity in multiple regression.

Plotting residuals or studentized residuals against each X , perhaps augmented by a nonparametric-regression smoother, is frequently helpful for detecting departures from linearity. As Figure 12.5 illustrates, however, simple residual plots cannot distinguish between monotone and nonmonotone nonlinearity. This distinction is lost in the residual plots because the least-squares fit ensures that the residuals are linearly uncorrelated with each X . The distinction is important because monotone nonlinearity frequently can be “corrected” by simple transformations.²⁷ In Figure 12.5, for example, case (a) might be modeled by $Y = \alpha + \beta\sqrt{X} + \varepsilon$, while case (b) cannot be linearized by a power transformation of X and might instead be dealt with by the quadratic regression, $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$.²⁸

In contrast to simple residual plots, added-variable plots, introduced in the previous chapter for detecting influential data, can reveal nonlinearity and suggest whether a relationship is monotone. These plots are not always useful for locating a transformation, however: The added-variable plot adjusts X_j for the other X s, but it is the *unadjusted* X_j that is transformed in respecifying the model. Moreover, as Cook (1998, Section 14.5) shows, added-variable plots are biased toward linearity when the correlations among the explanatory variables are large. *Component-plus-residual (CR) plots*, also called *partial-residual plots*, are often an effective

²⁶Polynomial regression—for example, the model $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$ —is discussed in Section 17.1. In this simple quadratic model, there are two regressors (X and X^2) but only one explanatory variable (X). Regression splines, discussed in Section 17.2, similarly construct several regressors from an X .

²⁷Recall the material in Section 4.3 on linearizing transformations.

²⁸Case (b) could, however, be accommodated by a more complex transformation of X , of the form $Y = \alpha + \beta(X - \gamma)^\lambda + \varepsilon$. In the illustration, γ could be taken as \bar{X} and λ as 2. More generally, γ and λ could be estimated from the data, for example, by nonlinear least squares (as described in Section 17.1). I will not pursue this approach here because there are other obstacles to estimating this more general transformation (see, e.g., Exercise 17.2).

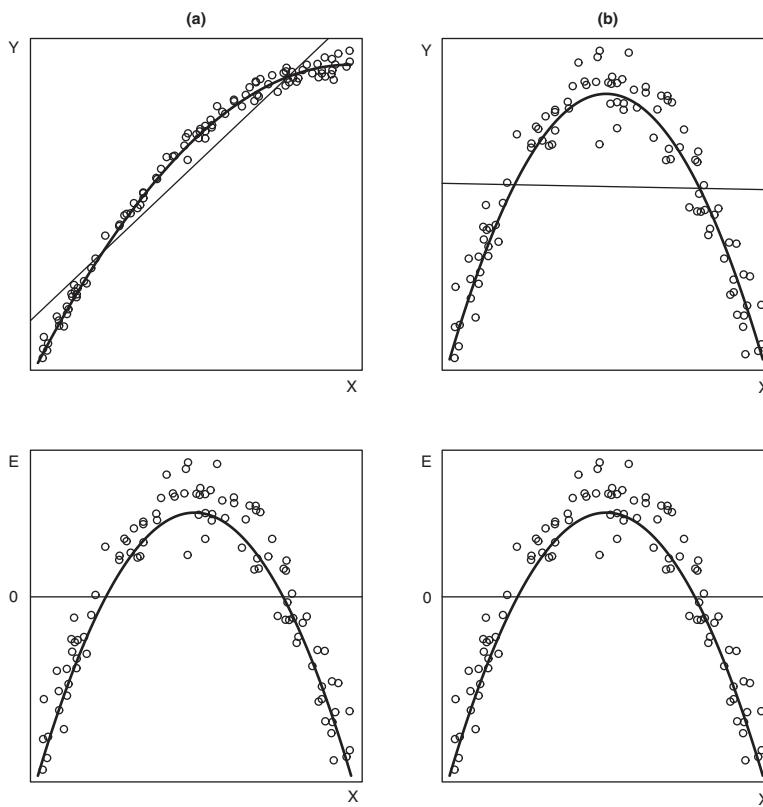


Figure 12.5 The residual plots of E versus X (in the lower panels) are identical, even though the regression of Y on X in (a) is monotone while that in (b) is nonmonotone.

alternative. Component-plus-residual plots are not as suitable as added-variable plots for revealing leverage and influence, however.

Define the *partial residual* for the j th explanatory variable as

$$E_i^{(j)} = E_i + B_j X_{ij}$$

In words, add back the linear component of the partial relationship between Y and X_j to the least-squares residuals, which may include an unmodeled nonlinear component. Then plot $E^{(j)}$ versus X_j . By construction, the multiple-regression coefficient B_j is the slope of the simple linear regression of $E^{(j)}$ on X_j , but nonlinearity may be apparent in the plot as well. Again, a nonparametric-regression smoother may help in interpreting the plot.

Figure 12.6 shows component-plus-residual plots for age and education in the SLID regression of log wages on these variables and sex (Equation 12.2 on page 300). Both plots look nonlinear: It is not entirely clear whether the partial relationship of log wages to age is monotone, simply tending to level off at the higher ages, or whether it is nonmonotone, turning back down at the far right. In the former event, we should be able to linearize the relationship by moving age *down* the ladder of powers because the bulge points to the left. In the latter event, a quadratic partial regression might work. In contrast, the partial relationship of log wages to

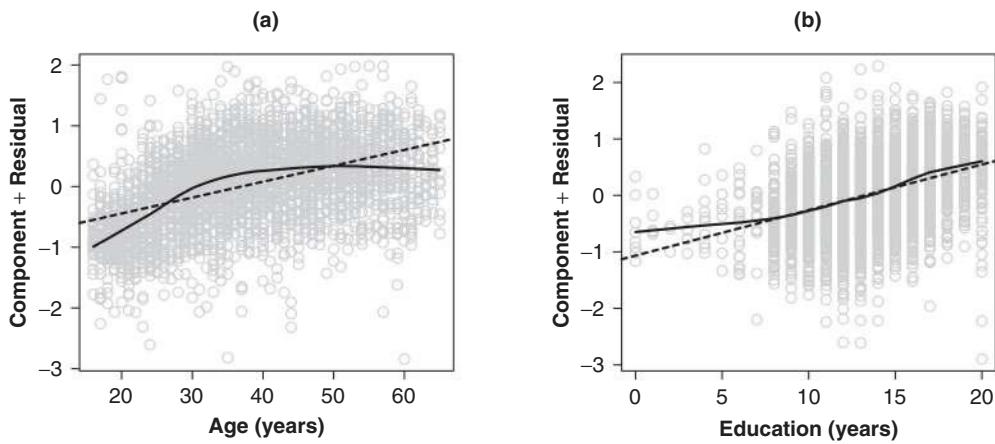


Figure 12.6 Component-plus-residual plots for age and education in the SLID regression of log wages on these variables and sex. The solid lines are for lowess smooths with spans of 0.4, and the broken lines are for linear least-squares fits.

education is clearly monotone, and the departure from linearity is not great—except at the lowest levels of education, where data are sparse; we should be able to linearize this partial relationship by moving education *up* the ladder of powers, because the bulge points to the right.

Trial-and-error experimentation suggests that the quadratic specification for age works better than a transformation; a quadratic in age, along with squaring education, produces the following fit to the data:²⁹

$$\begin{aligned} \widehat{\log_2 \text{Wages}} = & 0.5725 + 0.3195 \times \text{Male} + 0.1198 \times \text{Age} \\ & (0.0834) \quad (0.0180) \quad (0.0046) \\ & - 0.001230 \times \text{Age}^2 + 0.002605 \times \text{Education}^2 \quad (12.7) \\ & \quad (0.000059) \quad (0.000113) \end{aligned}$$

$$R^2 = .3892$$

I will consider the interpretation of this model shortly, but first let us examine component-plus-residual plots for the new fit. Because the model is now nonlinear in both age and education, there are two ways to proceed:

1. We can plot partial residuals for each of age and education against the corresponding explanatory variable. In the case of age, the partial residuals are computed as

$$E_i^{(\text{Age})} = 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2 + E_i \quad (12.8)$$

and for education,

$$E_i^{(\text{Education})} = 0.002605 \times \text{Education}_i^2 + E_i \quad (12.9)$$

²⁹See Exercise 12.7 for the alternative of transforming age.

The corresponding component-plus-residual plots are shown in the upper panels of Figure 12.7. The solid lines in these graphs are the *partial fits* (i.e., the components) for the two explanatory variables,

$$\begin{aligned}\hat{Y}_i^{(\text{Age})} &= 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2 \\ \hat{Y}_i^{(\text{Education})} &= 0.002605 \times \text{Education}_i^2\end{aligned}\quad (12.10)$$

The broken lines are lowess smooths, computed with $\text{span} = 0.4$. We look for the components to be close to the lowess smooths.

2. We can plot the partial residuals (as defined in Equations 12.8 and 12.9) against the partial fits (Equation 12.10) for the two variables. These plots are in the two lower panels of Figure 12.7. Here, the solid lines are least-squares lines, and, as before, the broken lines are lowess smooths. We look for the lowess smooths to be close to the least-squares lines.

It is apparent from the component-plus-residual plots in Figure 12.7 that the respecified model has done a good job of capturing the nonlinearity in the partial relationships of log wages with age and education—except possibly at the very highest ages, where the quadratic fit for age may exaggerate the downturn in wages.

To this point, then, we have log-transformed wages to make the distribution of the residuals more symmetric and to stabilize the error variance, and we have fit a quadratic regression in age and power-transformed education to linearize the relationship of log wages to these variables. The result is the fitted model in Equation 12.7. Two of its characteristics make this model difficult to interpret:

1. The transformations of wages and education move these variables from their familiar scales (i.e., dollars per hour and years, respectively).
2. Because the linear term in age is marginal to the squared term, the two terms are not separately interpretable. More precisely, the coefficient of the linear term, 0.1198, is the slope of the regression surface in the direction of age at age 0—clearly not a meaningful quantity—and twice the coefficient of the squared term in age, $2 \times (-0.001230) = -0.002460$, is the *change* in the age slope per year of age; the slope consequently declines with age and eventually becomes negative.³⁰

Interpretation is therefore greatly facilitated by graphing the partial regressions, using the effect-display framework developed in Chapters 7 and 8. Effect displays for age and education appear in Figure 12.8. In the effect display for age, for example, education and the dummy regressor for sex are set to their average values (which, in the latter case, represents the proportion of men in the SLID data set). The effects are graphed on the log-wages scale employed in the model, but I show wages in dollars on the axis at the right of each panel. An alternative

³⁰*The slope of the partial fit for age is the derivative $d(0.1198 \times \text{Age} - 0.001230 \times \text{Age}^2)/d\text{Age} = 0.1198 - 0.002460\text{Age}$. These points are developed in more detail in the discussion of polynomial regression in Section 17.1.

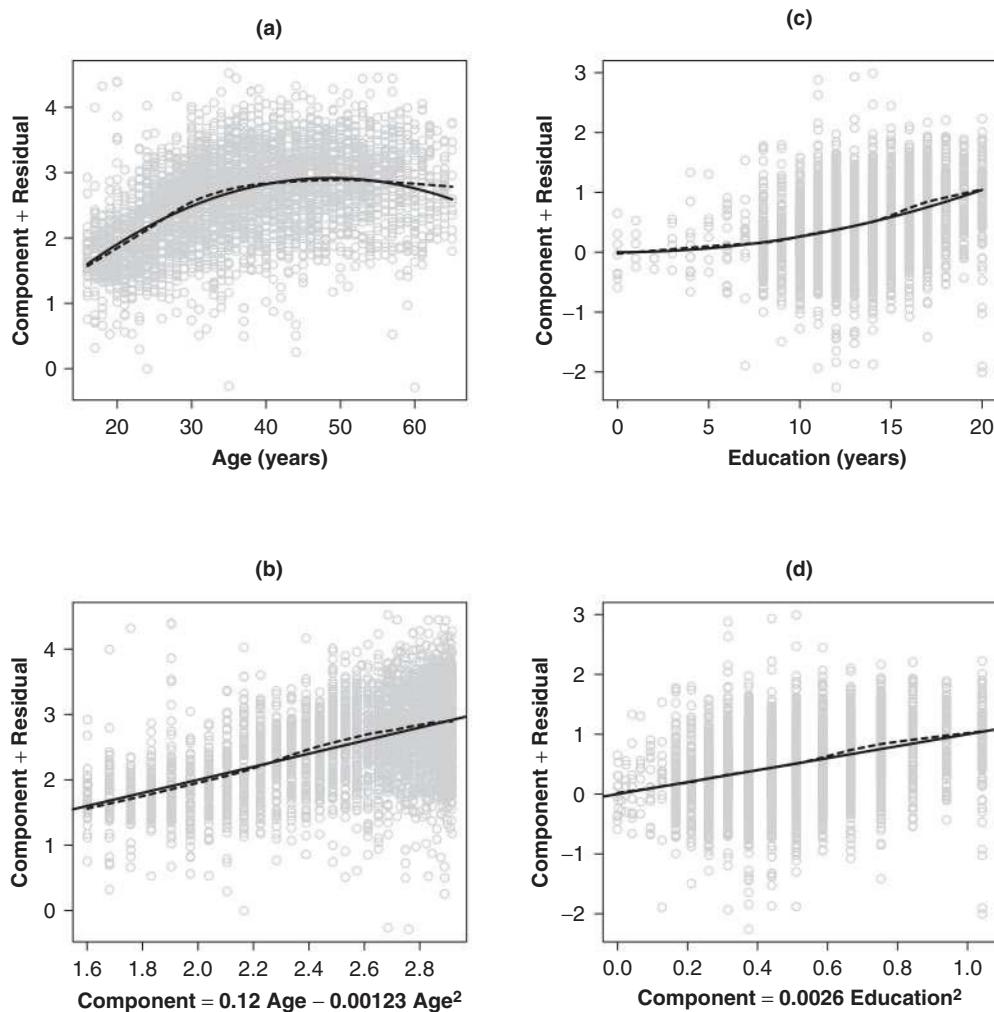


Figure 12.7 Component-plus-residual plots for age [panels (a) and (b)] and education [panels (c) and (d)] in the respecified model fit to the SLID data. In panels (a) and (c), partial residuals for age and education are plotted against the corresponding explanatory variable, with the component for the explanatory variable graphed as a solid line. In panels (b) and (d), the partial residuals are plotted against each component, and the solid line is a least-squares line. In all four panels, the broken line represents a lowess smooth with a span of 0.4.

would be to graph the effects directly on the dollar scale. The 95% pointwise confidence envelopes around the effects show that they are precisely estimated.³¹

³¹I could have shown an effect display for sex as well, but this effect is readily ascertained directly from the coefficient: Holding age and education constant, men earn on average $2^{0.3195} = 1.248$ times as much as (i.e., 25% more than) women.

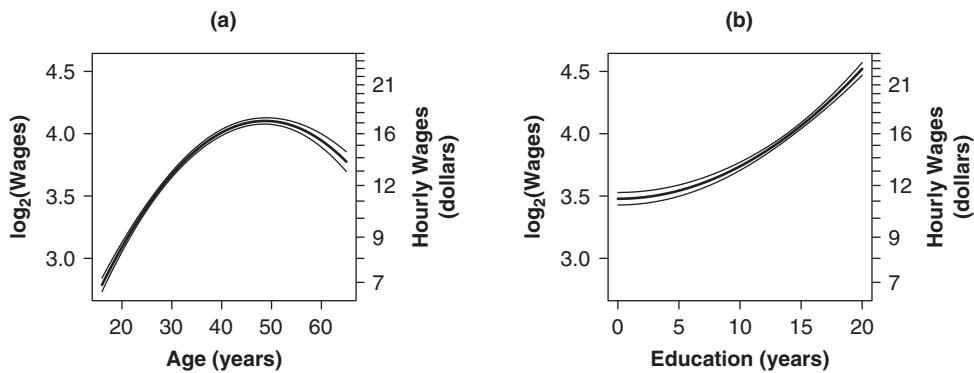


Figure 12.8 Effect displays for age and education in the model in Equation 12.7 (page 310). The lighter lines give 95% pointwise confidence envelopes around the fits.

Table 12.1 Coefficients for the Regression of Log Wages on Sex, Age, Education and the Interactions Between Sex and Age and Between Sex and Education

Coefficient	Estimate	Standard Error
Constant	0.8607	0.1155
Male	-0.3133	0.1641
Age	0.1024	0.0064
Age ²	-0.001072	0.000083
Education ²	-0.003230	0.000166
Male × Age	-0.03694	0.00910
Male × Age ²	-0.0003392	0.0001171
Male × Education ²	-0.001198	0.000225

12.3.2 Component-Plus-Residual Plots for Models With Interactions

Traditionally, component-plus-residual plots are applied to additive terms in a linear model, but these displays can be adapted to models with interactions. Suppose, for example, that we augment the SLID regression model with interactions between sex and age and between sex and education, retaining (at least tentatively) the quadratic specification of the age effect and the square of education. Estimated coefficients and standard errors for the new model are in Table 12.1. The R^2 for this model is .4029. The two interactions are highly statistically significant by incremental F -tests: For the interaction of sex with age, we have $F_0 = 31.33$, with 2 and 3989 degrees of freedom, for which $p \ll .0001$, and for the interaction of sex with education, we have $F_0 = 28.26$ with 1 and 3989 degrees of freedom, for which $p \ll .0001$ as well.³²

³²The latter test could be computed from the t -value obtained by dividing the Male × Education² coefficient by its standard error.

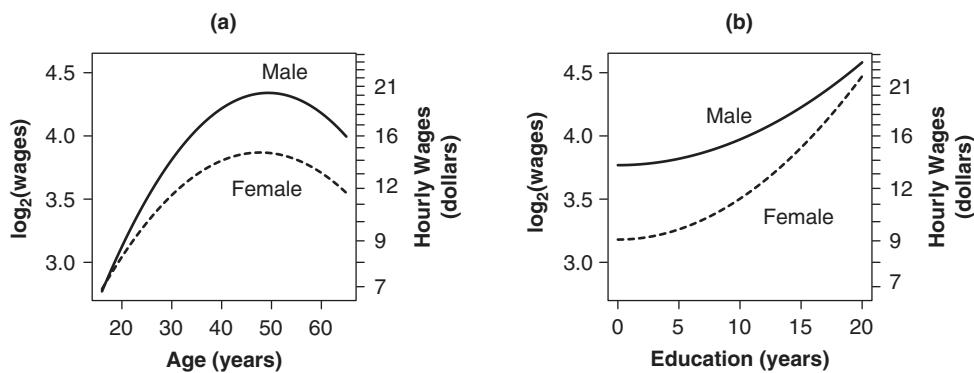


Figure 12.9 Effect displays for (a) the interaction between sex and age and (b) the interaction between sex and education, in the model summarized in Table 12.1.

The complex structure of the model makes it difficult to interpret directly from the coefficients: For example, the coefficient for the sex dummy variable is the log-income advantage of men at age 0 and education 0. Effect displays for the high-order terms in the model, shown in Figure 12.9, are straightforward, however: At average education, men's and women's average income is similar at the lowest ages, but men's income initially rises more rapidly and then falls off more rapidly at the highest ages. Likewise, at average age, men's income advantage is greatest at the lowest levels of education and the advantage declines, while average income itself rises as education goes up.

To construct component-plus-residual plots for this model, we can divide the data by sex and, for each sex, plot partial residuals against the partial fit. The results, shown in Figures 12.10 and 12.11, suggest that the model fits the data adequately (although the quadratic specification for age may exaggerate the decline in income at the highest ages).

It is straightforward to extend component-plus-residual plots to three dimensions by forming partial residuals for two quantitative explanatory variables simultaneously and plotting the partial residuals against these X s. A 3D component-plus-residual plot can reveal not only nonlinearity in the partial relationship of Y to each of two X s but also unmodeled interaction between the X s.³³ An alternative is to “slice” the data by one X and then to represent two X s simultaneously in a sequence of two-dimensional scatterplots, allocating partial residuals to the slice to which they are closest and plotting against the unsliced X . This approach can be generalized to more than two X s simultaneously.³⁴

12.3.3 When Do Component-Plus-Residual Plots Work?

Circumstances under which regression plots, including component-plus-residual plots, are informative about the structure of data have been extensively studied.³⁵ It is unreasonable to

³³See Cook (1998) and Cook and Weisberg (1994, 1999).

³⁴This approach is implemented by Sanford Weisberg and me in the **effects** package for the **R** statistical computing environment.

³⁵Much of this work is due to Cook and his colleagues; see, in particular, Cook (1993), on which the current section is based, and Cook (1994). Cook and Weisberg (1994, 1999) provide accessible summaries.

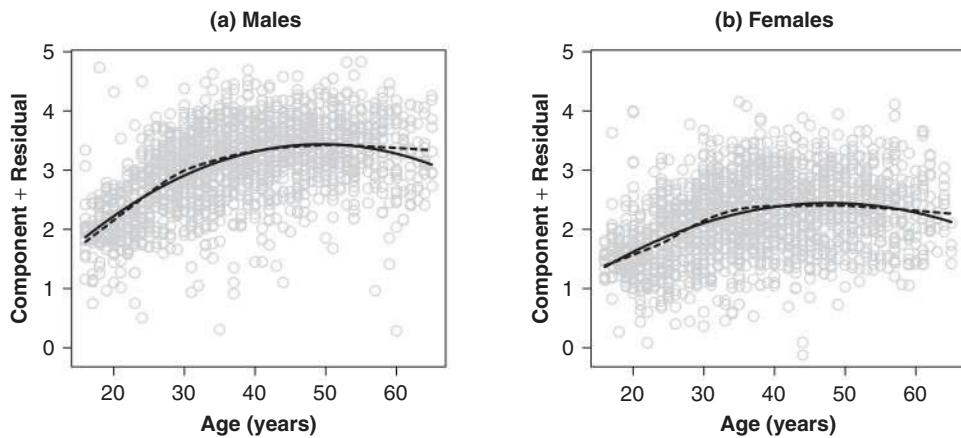


Figure 12.10 Component-plus-residual plots for the sex-by-age interaction. The solid lines give the partial fit (i.e., the component), while the broken lines are for a lowess smooth with a span of 0.4.

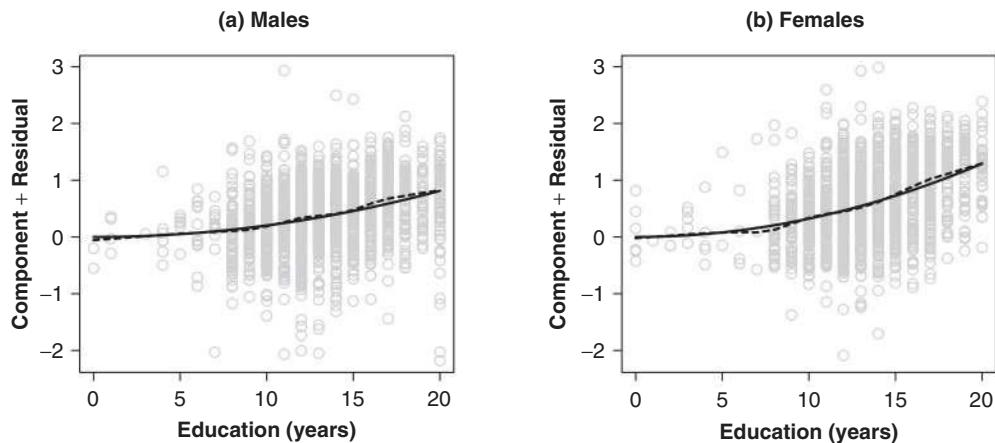


Figure 12.11 Component-plus-residual plots for the sex-by-education interaction.

expect that lower-dimensional displays can always uncover structure in a higher-dimensional problem. We may, for example, discern an interaction between two explanatory variables in a three-dimensional scatterplot, but it is not possible to do so in two separate two-dimensional plots, one for each explanatory variable.

It is important, therefore, to understand when graphical displays work and why they sometimes fail: First, understanding the circumstances under which a plot is effective may help us to produce those circumstances. Second, understanding why plots succeed and why they fail may help us to construct more effective displays. Both of these aspects will be developed below.

To provide a point of departure for this discussion, imagine that the following model accurately describes the data:

$$Y_i = \alpha + f(X_{i1}) + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad (12.11)$$

That is, the partial relationship between Y and X_1 is (potentially) nonlinear, characterized by the function $f(X_1)$, while the other explanatory variables, X_2, \dots, X_k , enter the model linearly.

We do not know in advance the shape of the function $f(X_1)$ and indeed do not know that the partial relationship between Y and X_1 is nonlinear. Instead of fitting the true model (Equation 12.11) to the data, therefore, we fit the “working model”:

$$Y_i = \alpha' + \beta'_1 X_{i1} + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

The primes indicate that the estimated coefficients for this model do not, in general, estimate the corresponding parameters of the true model (Equation 12.11), nor is the “error” of the working model the same as the error of the true model.

Suppose, now, that we construct a component-plus-residual plot for X_1 in the working model. The partial residuals estimate

$$\varepsilon_i^{(1)} = \beta'_1 X_{i1} + \varepsilon'_i \quad (12.12)$$

What we would really like to estimate, however, is $f(X_{i1}) + \varepsilon_i$, which, apart from random error, will tell us the partial relationship between Y and X_1 . Cook (1993) shows that $\varepsilon_i^{(1)} = f(X_{i1}) + \varepsilon_i$, as desired, under either of two circumstances:

1. The function $f(X_1)$ is linear after all, in which case the population analogs of the partial residuals in Equation 12.12 are appropriately linearly related to X_1 .
2. The *other* explanatory variables X_2, \dots, X_k are each linearly related to X_1 . That is,

$$E(X_{ij}) = \alpha_{j1} + \beta_{j1} X_{i1} \quad \text{for } j = 2, \dots, k \quad (12.13)$$

If, in contrast, there are *nonlinear* relationships between the other X s and X_1 , then the component-plus-residual plot for X_1 may not reflect the true partial regression $f(X_1)$.³⁶

The second result suggests a practical procedure for improving the chances that component-plus-residual plots will provide accurate evidence of nonlinearity: If possible, transform the explanatory variables to linearize the relationships among them (using, e.g., the unconditional Box-Cox procedure described in Section 4.6). Evidence suggests that weak nonlinearity is not especially problematic, but strong nonlinear relationships among the explanatory variables can invalidate the component-plus-residual plot as a useful diagnostic display.³⁷

Mallows (1986) has suggested a variation on the component-plus-residual plot that sometimes reveals nonlinearity more clearly. I will focus on X_1 , but the spirit of Mallows’s suggestion is to draw a plot for each X in turn. First, construct a working model with a quadratic term in X_1 along with the usual linear term:

³⁶Note that each of the other X s is regressed on X_1 , *not* vice versa.

³⁷See Exercise 12.6.

$$Y_i = \alpha' + \beta'_1 X_{i1} + \gamma_1 X_{i1}^2 + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

Then, after fitting the working model, form the “augmented” partial residual

$$E_i^{(1)} = E'_i + B'_1 X_{i1} + C_1 X_{i1}^2$$

Note that B'_1 generally differs from the regression coefficient for X_1 in the original model, which does not include the squared term. Finally, plot $E^{(1)}$ versus X_1 .

The circumstances under which the augmented partial residuals accurately capture the true partial-regression function $f(X_1)$ are closely analogous to the linear case (see Cook, 1993); either

1. the function $f(X_1)$ is a quadratic in X_1 ³⁸ or
2. the regressions of the other explanatory variables on X_1 are quadratic:

$$E(X_{ij}) = \alpha_{j1} + \beta_{j1} X_{i1} + \gamma_{j1} X_{i1}^2 \quad \text{for } j = 2, \dots, k \quad (12.14)$$

This is a potentially useful result if we cannot transform away nonlinearity among the explanatory variables—as is the case, for example, when the relationships among the explanatory variables are not monotone.

Mallows’s approach can be generalized to higher-order polynomials.

The premise of this discussion, expressed in Equation 12.11, is that Y is a nonlinear function of X_1 but linearly related to the other X s. In real applications of component-plus-residual plots, however, it is quite possible that there is more than one nonlinear partial relationship, and we typically wish to examine each explanatory variable in turn. Suppose, for example, that the relationship between Y and X_1 is linear, that the relationship between Y and X_2 is nonlinear, and that X_1 and X_2 are correlated. The component-plus-residual plot for X_1 can, in this situation, show apparent nonlinearity—sometimes termed a “leakage” effect. If more than one component-plus-residual plot shows evidence of nonlinearity, it may, therefore, be advisable to refit the model and reconstruct the component-plus-residual plots after correcting the most dramatic instance of nonlinearity.³⁹

Applied to the SLID regression of log wages on sex, age, and education, Mallows’s augmented component-plus-residual plots look very much like traditional component-plus-residual plots.⁴⁰

Simple forms of nonlinearity can often be detected in component-plus-residual plots. Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an explanatory variable, for example). Component-plus-residual plots reliably reflect nonlinearity when there are not strong nonlinear relationships among the explanatory variables in a regression.

³⁸This condition covers a linear partial relationship as well—that is where $\gamma_1 = 0$.

³⁹An iterative formalization of this procedure provides a basis for fitting nonparametric additive regression models, discussed in Section 18.2.2.

⁴⁰See Exercise 12.8.

CERES Plots*

Cook (1993) provides a still more general procedure, which he calls CERES (for “Combining conditional Expectations and RESiduals”): Let

$$\hat{X}_{ij} = \hat{g}_{j1}(X_{i1})$$

represent the estimated regression of X_j on X_1 , for $j = 2, \dots, k$. These regressions may be linear (as in Equation 12.13), quadratic (as in Equation 12.4), or nonparametric. Of course, the functions $\hat{g}_{j1}(X_1)$ will generally be different for different X_j s. Once the regression functions for the other explanatory variables are found, form the working model

$$Y_i = \alpha'' + \beta_2'' X_{i2} + \cdots + \beta_k'' X_{ik} + \gamma_{12} \hat{X}_{i2} + \cdots + \gamma_{1k} \hat{X}_{ik} + \varepsilon_i''$$

The residuals from this model are then combined with the estimates of the γ s,

$$E_i^{(1)} = E_i'' + C_{12} \hat{X}_{i2} + \cdots + C_{1k} \hat{X}_{ik}$$

and plotted against X_1 .

CERES plots for the SLID regression of log wages on sex, age, and education are very similar to traditional component-plus-residual plots.⁴¹

12.4 Discrete Data

As explained in Chapter 3, discrete explanatory and response variables often lead to plots that are difficult to interpret, a problem that can be partially rectified by “jittering” the plotted points.⁴² A discrete *response* variable also violates the assumption that the errors in a linear model are normally distributed. This problem, like that of a limited response variable (i.e., one that is bounded below or above), is only serious in extreme cases—for example, when there are very few distinct response values or where a large proportion of the data assumes a small number of unique values, conditional on the values of the explanatory variables. In these cases, it is best to use statistical models for categorical response variables.⁴³

Discrete *explanatory* variables, in contrast, are perfectly consistent with the general linear model, which makes no distributional assumptions about the X s, other than independence between the X s and the errors. Indeed, because it partitions the data into groups, a discrete X (or combination of X s) facilitates straightforward tests of nonlinearity and nonconstant error variance.

12.4.1 Testing for Nonlinearity (“Lack of Fit”)

Recall the data on vocabulary and education from the U.S. General Social Survey, introduced in Chapter 3. Years of education in this data set range between 0 and 20. Suppose that we model the relationship between vocabulary score and education in two ways:

⁴¹See Exercise 12.8.

⁴²See Section 3.2.

⁴³See Chapter 14.

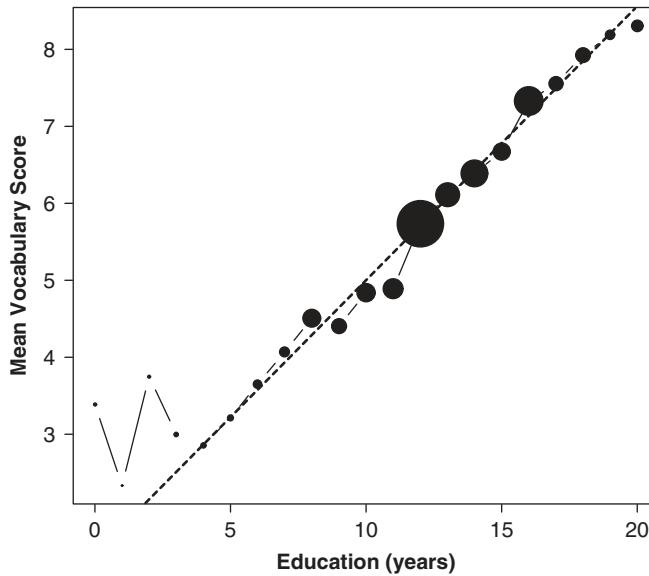


Figure 12.12 Mean vocabulary score by years of education. The size of the points is proportional to the number of observations at each educational level. The broken line is for the least-squares regression of vocabulary on education.

1. Fit a linear regression of vocabulary on education:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (12.15)$$

2. Model education in a one-way ANOVA with a set of dummy regressors. There are 21 distinct values of education, yielding 20 dummy regressors (treating 0 years of education as the baseline category):

$$Y_i = \alpha' + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{20} D_{i,20} + \varepsilon'_i \quad (12.16)$$

Figure 12.12 contrasts these two models visually, showing the mean vocabulary score at each level of education (corresponding to Equation 12.16) and the least-squares regression line (corresponding to Equation 12.15). The area of the points representing the means is proportional to the number of observations at each educational level.

Contrasting the two models produces a test for nonlinearity because the model in Equation 12.5, specifying a linear relationship between vocabulary and education, is a special case of the model given in Equation 12.16, which can capture *any* pattern of relationship between $E(Y)$ and X . The resulting incremental F -test for nonlinearity appears in Table 12.2. There is, therefore, very strong evidence of a departure from linearity. Nevertheless, the linear regression of vocabulary on education accounts for almost all the variation among the means: The R^2 for the linear regression is $25,340/101,436 = 0.2498$, and for the more general one-way ANOVA (i.e., dummy regression), $R^2 = 26,099/101,436 = 0.2573$. In such a large sample, with more

Table 12.2 Analysis of Variance for Vocabulary Test Scores, Showing the Incremental F -Test for Nonlinearity of the Relationship Between Vocabulary and Education

Source	SS	df	F	p
Education				
(Model 12.6)	26,099.	20	374.44	<.0001
Linear				
(Model 12.15)	25,340.	1	7270.99	<.0001
Nonlinear				
(“lack of fit”)	759.	19	11.46	<.0001
Error				
(“pure error”)	75,337.	21,617		
Total	101,436.	21,637		

than 20,000 observations, even this relatively small difference is statistically significant. Small though it may be, the departure from linearity in Figure 12.12 nevertheless makes some substantive sense: Discounting the means at very low levels of education where there are little data, there are small jumps in average vocabulary scores at 12 and 16 years of education—corresponding to graduation from high school and university.

The incremental F -test for nonlinearity can easily be extended to a discrete explanatory variable—say X_1 —in a multiple-regression model. Here, we need to contrast the general model

$$Y_i = \alpha' + \gamma_1 D_{i1} + \cdots + \gamma_{m-1} D_{im-1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon'_i$$

with the model specifying a linear effect of X_1

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where D_1, \dots, D_{m-1} are dummy regressors constructed to represent the m distinct values of X_1 .

Consider, by way of illustration, the additive model that we fit in the previous section to the SLID data (Equation 12.7 on page 310), regressing log wages on sex, a second-degree polynomial in age, and the square of education. Does this specification adequately capture the shape of the partial regressions of log wages on age and education? Because both age and education are discrete, we can fit a model that treats age and education as factors, with 50 and 21 levels, respectively. We contrast this larger model with two null models: one that treats age as a second-order polynomial and the other that includes the square of education, producing the following regression sums of squares and degrees of freedom:

Model	RegSS	df
Sex, Age (as factor), Education (as factor)	855.64	70
Sex, Age (as quadratic), Education (as factor)	827.02	23
Sex, Age (as factor), Education ²	847.53	51

The residual sum of squares for the full model is RSS = 1249.10 on 3,926 degrees of freedom, and consequently incremental F -tests for lack of fit are

$$\text{Age: } F_0 = \frac{\frac{855.64 - 827.02}{70 - 23}}{\frac{1249.10}{3926}} = 1.91, df = 47, 3926, p = .0002$$

$$\text{Education: } F_0 = \frac{\frac{855.64 - 847.53}{70 - 51}}{\frac{1249.10}{3926}} = 1.34, df = 19, 3926, p = .15$$

The lack of fit from the specified partial relationship of log wages to age is statistically significant, while the lack of fit for education is not. Even in the case of age, however, lack of fit is substantively small: For the full model $R^2 = .4065$ and for the model treating age as a second-degree polynomial $R^2 = .3929$. That this difference of about 1% in explained variation is statistically significant is testimony to the power of the test in a moderately large sample.⁴⁴

A slightly more elaborate example uses the model for the SLID data in Table 12.1 (in the previous section, on page 313). This model specifies a regression of log wages on sex, a quadratic for age, and the square of education but also permits interactions between sex and age and between sex and education. To test for lack of fit, we contrast the following models:

Model	RegSS	df
Sex, Age (as factor), Education (as factor), with interactions	903.10	138
Sex, Age (as quadratic), Education (as factor), with interactions	858.72	44
Sex, Age (as factor), Education ² , with interactions	893.14	101

These models, especially the full model in which both age and education are treated as factors, have many parameters to estimate, but with nearly 4000 observations, we can spend many degrees of freedom on the model and still have plenty left to estimate the error variance: The residual sum of squares for the full model is RSS = 1201.64 on 3,858 degrees of freedom. Incremental F -tests for lack of fit are as follows:

Age (and its interaction with Sex):

$$F_0 = \frac{\frac{903.10 - 858.72}{138 - 44}}{\frac{1201.64}{3858}} = 1.51, df = 94, 3858, p = .0011$$

Education (and its interaction with Sex):

$$F_0 = \frac{\frac{903.10 - 893.14}{138 - 101}}{\frac{1201.64}{3858}} = 0.86, df = 37, 3858, p = .70$$

Thus, as in the preceding example, there is a small but statistically significant lack of fit entailed by using a quadratic for age (the R^2 for the full model is .4291 versus .4080 for the

⁴⁴All alternative to testing for nonlinearity is to use a model-selection criterion that takes the parsimony of the models into account. Model-selection criteria are discussed in Section 22.1, and I invite the reader to apply, for example, the AIC and BIC to examples in the current section.

much more parsimonious model with a quadratic in age), and there is no evidence of lack of fit using the square of education in the regression.⁴⁵

Another approach to testing for nonlinearity exploits the fact that a polynomial of degree $m - 1$ can perfectly capture the relationship between Y and a discrete X with m categories, regardless of the specific form of this relationship. We remove one term at a time from the model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_{m-1} X_i^{m-1} + \varepsilon_i$$

beginning with X^{m-1} . If the decrement in the regression sum of squares is nonsignificant (by an incremental F -test on 1 degree of freedom), then we proceed to remove X^{m-2} , and so on.⁴⁶ This “step-down” approach has the potential advantage of parsimony because we may well require more than one term (i.e., a linear relationship) but fewer than $m - 1$ terms (i.e., a relationship of arbitrary form). High-degree polynomials, however, are usually difficult to interpret.⁴⁷

12.4.2 Testing for Nonconstant Error Variance

A discrete X (or combination of X s) partitions the data into m groups (as in analysis of variance). Let Y_{ij} denote the i th of n_j response-variable scores in group j . If the error variance is constant across groups, then the within-group sample variances

$$S_j^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{n_j - 1}$$

should be similar. Tests that examine the S_j^2 directly, such as Bartlett’s (1937) classic (and commonly employed) test, do not maintain their validity well when the distribution of the errors is non-normal.

Many alternative tests have been proposed. In a large-scale simulation study, Conover, Johnson, and Johnson (1981) found that the following simple F -test (called “Levene’s test”) is both robust and powerful:⁴⁸ Calculate the values

$$Z_{ij} \equiv |Y_{ij} - \tilde{Y}_j|$$

where \tilde{Y}_j is the median response-variable value in group j . Then perform a one-way analysis of variance of the Z_{ij} over the m groups. If the error variance is not constant across the groups, then the group means \bar{Z}_j will tend to differ, producing a large value of the F -test statistic.⁴⁹

⁴⁵We could, in principle, go further in testing for lack of fit, specifying a model that divides the data by combinations of levels of sex, age, and education and comparing this model with our current model for the SLID data. We run into the “curse of dimensionality,” however (see Section 2.2 and Chapter 18): There are $2 \times 50 \times 21 = 2100$ combinations of values of the three explanatory variables and “only” about 4,000 observations in the data set.

⁴⁶As usual, the estimate of error variance in the denominator of these F -tests is taken from the full model with all $m - 1$ terms.

⁴⁷*There is a further, technical difficulty with this procedure: The several powers of X are usually highly correlated, sometimes to the point that least-squares calculations break down. A solution is to orthogonalize the power regressors prior to fitting the model. See the discussion of polynomial regression in Section 17.1.

⁴⁸An alternative, less robust, version of Levene’s test uses deviations from the group means rather than from the group medians.

⁴⁹This test ironically exploits the robustness of the validity of the F -test in one-way ANOVA. (The irony lies in the common use of tests of constant variance as a preliminary to tests of differences in means.)

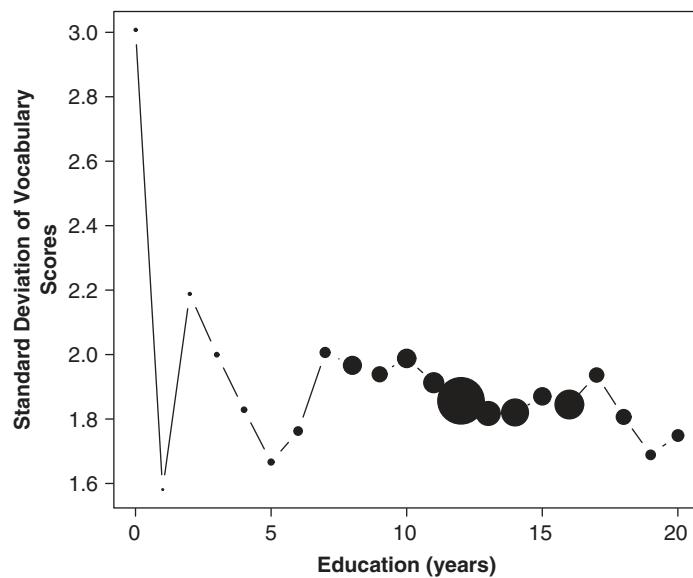


Figure 12.13 Standard deviation of vocabulary scores by education. The relative size of the points is proportional to the number of respondents at each level of education.

For the vocabulary data, for example, where education partitions the 21,638 observations into $m = 21$ groups, this test gives $F_0 = 4.26$, with 20 and 21,617 degrees of freedom, for which $p \ll .0001$. There is, therefore, strong evidence of nonconstant spread in vocabulary across the categories of education, though, as revealed in Figure 12.13, the within-group standard deviations are not very different (discounting the small numbers of individuals with very low levels of education).⁵⁰

Discrete explanatory variables divide the data into groups. A simple incremental F -test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means. Likewise, tests of nonconstant variance can be based on comparisons of spread in the different groups.

12.5 Maximum-Likelihood Methods*

A statistically sophisticated approach to selecting a transformation of Y or an X is to embed the usual linear model in a more general nonlinear model that contains a parameter for the

⁵⁰The tendency of the standard deviations to decline slightly with increasing education is likely due to a “ceiling” effect—at higher levels of education, the vocabulary scores push toward the upper boundary of 10.

transformation. If several variables are potentially to be transformed, or if the transformation is complex, then there may be several such parameters.⁵¹

Suppose that the transformation is indexed by a single parameter λ (e.g., the power transformation $Y \rightarrow Y^\lambda$) and that we can write down the likelihood for the model as a function of the transformation parameter and the usual regression parameters: $L(\lambda, \alpha, \beta_1, \dots, \beta_k, \sigma_\varepsilon^2)$.⁵² Maximizing the likelihood yields the maximum-likelihood estimate of λ along with the MLEs of the other parameters. Now suppose that $\lambda = \lambda_0$ represents *no* transformation (e.g., $\lambda_0 = 1$ for the power transformation Y^λ). A likelihood-ratio test, Wald test, or score test of $H_0: \lambda = \lambda_0$ assesses the evidence that a transformation is required.

A disadvantage of the likelihood-ratio and Wald tests in this context is that they require finding the MLE, which usually necessitates iteration (i.e., a repetitive process of successively closer approximations). In contrast, the slope of the log-likelihood at λ_0 —on which the score test depends—generally can be assessed or approximated without iteration and therefore is faster to compute.

Often, the score test can be formulated as the t -statistic for a new regressor, called a *constructed variable*, to be added to the linear model. An added-variable plot for the constructed variable then can reveal whether one or a small group of observations is unduly influential in determining the transformation or, alternatively, whether evidence for the transformation is spread throughout the data.

12.5.1 Box-Cox Transformation of Y

Box and Cox (1964) suggested a power transformation of Y with the object of normalizing the error distribution, stabilizing the error variance, and straightening the relationship of Y to the X s.⁵³ The general Box-Cox model is

$$Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

where the errors ε_i are independently $N(0, \sigma_\varepsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

For the Box-Cox transformation to make sense, all the Y_i must be positive.⁵⁴

For a particular choice of λ , the conditional maximized log-likelihood is⁵⁵

⁵¹ Models of this type are fundamentally nonlinear and can be treated by the general methods of Chapter 17 as well as by the methods described in the present section.

⁵² See online Appendix D for a general introduction to maximum-likelihood estimation.

⁵³ Subsequent work (Hernandez & Johnson, 1980) showed that Box and Cox's method principally serves to normalize the error distribution.

⁵⁴ Strictly speaking, the requirement that the Y_i are positive precludes the possibility that they are normally distributed (because the normal distribution is unbounded), but this is not a serious practical difficulty unless many Y values stack up near 0. If there are 0 or negative values of Y , we can use a start to make all the Y -values positive, or we can use the Yeo-Johnson family of modified power transformations described in Exercise 4.4.

⁵⁵ See Exercise 12.9. Equation 12.17 is not technically a log-likelihood because if the distribution of Y^λ is normal for some particular value (say, λ') of λ , it is not normal for other values of $\lambda \neq \lambda'$. Box and Cox's method is, therefore, strictly speaking *analogous* to maximum likelihood. I am grateful to Sanford Weisberg for pointing this out to me.

$$\begin{aligned}\log_e L(\alpha, \beta_1, \dots, \beta_k, \sigma_\varepsilon^2 | \lambda) &= -\frac{n}{2}(1 + \log_e 2\pi) \\ &\quad - \frac{n}{2} \log_e \hat{\sigma}_\varepsilon^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log_e Y_i\end{aligned}\tag{12.17}$$

where $\hat{\sigma}_\varepsilon^2(\lambda) = \sum E_i^2(\lambda)/n$ and where the $E_i(\lambda)$ are the residuals from the least-squares regression of $Y^{(\lambda)}$ on the X s. The least-squares coefficients from this regression are the maximum-likelihood estimates of α and the β s, conditional on the value of λ .

A simple procedure for finding the maximum-likelihood estimator $\hat{\lambda}$, then, is to evaluate the maximized $\log_e L$ (called the *profile log-likelihood*) for a range of values of λ , say between -2 and $+2$. If this range turns out not to contain the maximum of the log-likelihood, then the range can be expanded. To test $H_0: \lambda = 1$, calculate the likelihood-ratio statistic

$$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \hat{\lambda})]$$

which is asymptotically distributed as χ^2 with one degree of freedom under H_0 . Alternatively (but equivalently), a 95% confidence interval for λ includes those values for which

$$\log_e L(\lambda) > \log_e L(\lambda = \hat{\lambda}) - 1.92$$

The number 1.92 comes from $\frac{1}{2} \times \chi^2_{1,0.05} = \frac{1}{2} \times 1.96^2$.

Figure 12.14 shows a plot of the profile log-likelihood against λ for the original SLID regression of composite hourly wages on sex, age, and education (Equation 12.1 on page 296). In constructing this graph, I have “zeroed in” on the maximum-likelihood estimate of λ : I originally plotted the profile log-likelihood over the wider range $\lambda = -2$ to $\lambda = 2$. The maximum-likelihood estimate of λ is $\hat{\lambda} = 0.09$, and a 95% confidence interval, marked out by the intersection of the line near the top of the graph with the profile log-likelihood, runs from 0.04 to 0.13. Recall that we previously employed a log transformation for these data (i.e., $\lambda = 0$) to make the residual distribution more nearly normal and to stabilize the error variance. Although $\lambda = 0$ is outside the confidence interval, it represents essentially the same transformation of wages as $\lambda = 0.09$ (indeed, the correlation between log wages and wages^{0.09} is 0.9996). I prefer the log transformation for interpretability.

Atkinson (1985) proposed an approximate score test for the Box-Cox model, based on the constructed variable

$$G_i = Y_i \left[\log_e \left(\frac{Y_i}{\tilde{Y}} \right) - 1 \right]$$

where \tilde{Y} is the *geometric mean* of Y :⁵⁶

$$\tilde{Y} \equiv (Y_1 \times Y_2 \times \dots \times Y_n)^{1/n}$$

This constructed variable is obtained by a linear approximation to the Box-Cox transformation $Y^{(\lambda)}$ evaluated at $\lambda = 1$. The augmented regression, including the constructed variable, is then

$$Y_i = \alpha' + \beta'_1 X_{i1} + \dots + \beta'_k X_{ik} + \phi G_i + \varepsilon'_i$$

⁵⁶It is more practical to compute the geometric mean as $\tilde{Y} = \exp[(\sum \log_e Y_i)/n]$.

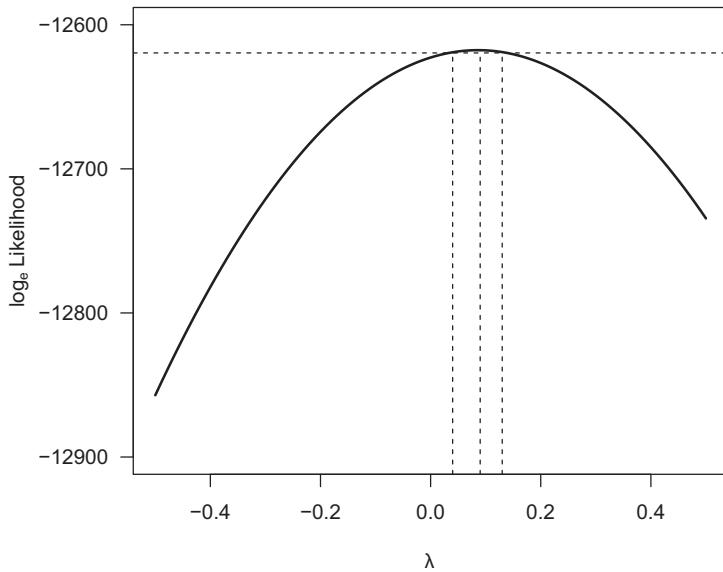


Figure 12.14 Box-Cox transformations for the SLID regression of wages on sex, age, and education. The maximized (profile) log-likelihood is plotted against the transformation parameter λ . The intersection of the line near the top of the graph with the profile log-likelihood curve marks off a 95% confidence interval for λ . The maximum of the log-likelihood corresponds to the MLE of λ .

The t -test of $H_0: \phi = 0$, that is, $t_0 = \hat{\phi}/\text{SE}(\hat{\phi})$, assesses the need for a transformation. The quantities $\hat{\phi}$ and $\text{SE}(\hat{\phi})$ are obtained from the least-squares regression of Y on X_1, \dots, X_k and G . An estimate of λ (though not the MLE) is given by $\tilde{\lambda} = 1 - \hat{\phi}$; and the added-variable plot for the constructed variable G shows influence and leverage on $\hat{\phi}$ and hence on the choice of λ .

Atkinson's constructed-variable plot for the SLID regression is shown in Figure 12.15. Although the trend in the plot is not altogether linear, it appears that evidence for the transformation of Y is spread generally through the data and does not depend unduly on a small number of observations. The coefficient of the constructed variable in the regression is $\hat{\phi} = 1.454$, with $\text{SE}(\hat{\phi}) = 0.026$, providing overwhelmingly strong evidence of the need to transform Y . The suggested transformation, $\tilde{\lambda} = 1 - 1.454 = -0.454$, is far from the MLE.

12.5.2 Box-Tidwell Transformation of the Xs

Now, consider the model

$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$

where the errors are independently distributed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and all the X_{ij} are positive. The parameters of this model— α , β_1, \dots, β_k , $\gamma_1, \dots, \gamma_k$, and σ_ε^2 —could be estimated by general

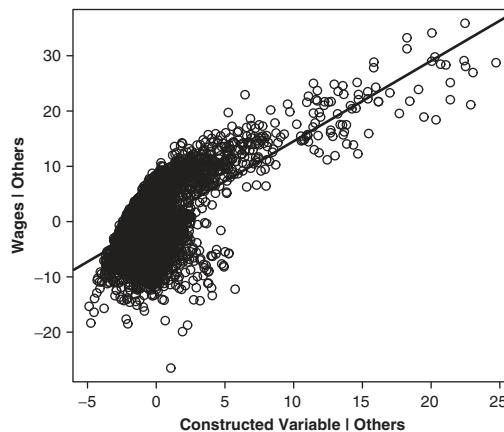


Figure 12.15 Constructed-variable plot for the Box-Cox transformation of wages in the SLID regression. The least-squares line is shown on the plot.

nonlinear least squares, but Box and Tidwell (1962) suggest instead a computationally more efficient procedure that also yields a constructed-variable diagnostic:⁵⁷

1. Regress Y on X_1, \dots, X_k , obtaining A, B_1, \dots, B_k .
2. Regress Y on X_1, \dots, X_k and the constructed variables $X_1 \log_e X_1, \dots, X_k \log_e X_k$, obtaining A', B'_1, \dots, B'_k and D_1, \dots, D_k . Because of the presence of the constructed variables in this second regression, in general $A \neq A'$ and $B_j \neq B'_j$. As in the Box-Cox model, the constructed variables result from a linear approximation to $X_j^{\gamma_j}$ evaluated at $\gamma_j = 1$.⁵⁸
3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis $H_0: \delta_j = 0$, where δ_j is the population coefficient of $X_j \log_e X_j$ in Step 2. Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the X s.
4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is given by

$$\tilde{\gamma}_j = 1 + \frac{D_j}{B_j}$$

Recall that B_j is from the *initial* (i.e., Step 1) regression (not from Step 2).

This procedure can be iterated through Steps 1, 2, and 4 until the estimates of the transformation parameters stabilize, yielding the MLEs $\hat{\gamma}_j$.

By way of example, I will work with the SLID regression of log wages on sex, education, and age. The dummy regressor for sex is not a candidate for transformation, of course, but I will consider power transformations of age and education. Recall that we were initially

⁵⁷Nonlinear least-squares regression is described in Section 17.4.

⁵⁸See Exercise 12.10.

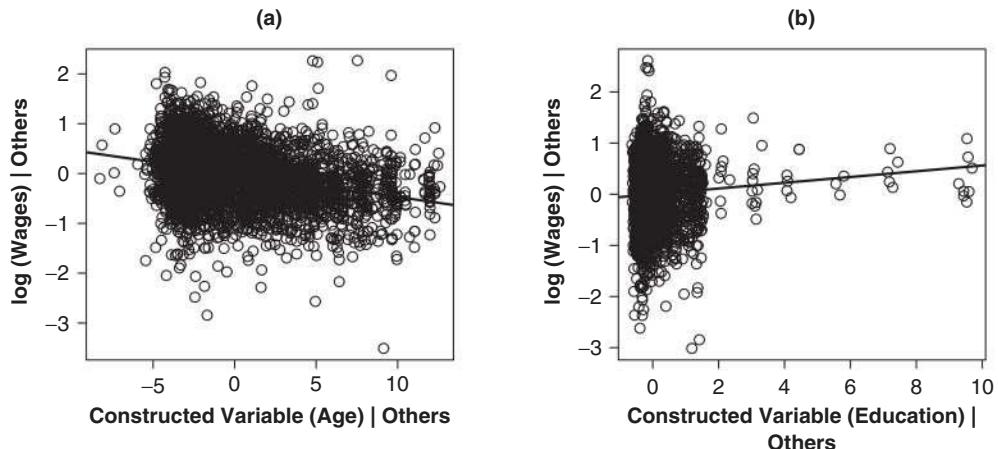


Figure 12.16 Constructed-variable plots for the Box-Tidwell transformation of (a) age and (b) education in the SLID regression of log wages on sex, age, and education.

undecided about whether to model the age effect as a quadratic or as a transformation down the ladder of powers and roots. To make power transformations of age more effective, I use a negative start of 15 (where age ranges from 16 to 65). As well, there are a few 0 values of education, and so I will use a start of 1 for education. Adding constants to the values of age and education changes the intercept but not the age and education coefficients in the initial regression.

The coefficients of $(\text{Age} - 15) \times \log_e(\text{Age} - 15)$ and $(\text{Education} + 1) \times \log_e(\text{Education} + 1)$ in the step-2 augmented model are, respectively, $D_{\text{Age}} = -0.04699$ with $\text{SE}(D_{\text{Age}}) = 0.00231$, and $D_{\text{Education}} = 0.05612$ with $\text{SE}(D_{\text{Education}}) = 0.01254$. Although both score tests are, consequently, statistically significant, there is much stronger evidence of the need to transform age than education.

The first-step estimates of the transformation parameters are

$$\begin{aligned}\tilde{\gamma}_{\text{Age}} &= 1 + \frac{D_{\text{Age}}}{B_{\text{Age}}} = 1 + \frac{-0.04699}{0.02619} = -0.79 \\ \tilde{\gamma}_{\text{Education}} &= 1 + \frac{D_{\text{Education}}}{B_{\text{Education}}} = 1 + \frac{0.05612}{0.08061} = 1.69\end{aligned}$$

The fully iterated MLEs of the transformation parameters are $\hat{\gamma}_{\text{Age}} = 0.051$ and $\hat{\gamma}_{\text{Education}} = 1.89$ —very close to the log transformation of started-age and the square of education.

Constructed-variable plots for the transformation of age and education, shown in Figure 12.16, suggest that evidence for the transformation of age is spread throughout the data but that there are some high-leverage, and hence potentially influential, observations determining the transformation of education. Accordingly, I proceeded to remove observations for which the education constructed variable exceeds 1 and found that when I did this, I obtained a similar estimate of the transformation parameter for education ($\hat{\gamma}_{\text{Education}} = 2.40$).

A statistically sophisticated general approach to selecting a transformation of Y or an X is to embed the linear-regression model in a more general model that contains a parameter for the transformation. The Box-Cox procedure selects a power transformation of Y to normalize the errors. The Box-Tidwell procedure selects power transformations of the X s to linearize the regression of Y on the X s. In both cases, “constructed-variable” plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.

12.5.3 Nonconstant Error Variance Revisited

Breusch and Pagan (1979) developed a score test for heteroscedasticity based on the specification

$$\sigma_i^2 \equiv V(\varepsilon_i) = g(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_p Z_{ip})$$

where Z_1, \dots, Z_p are known variables and where the function $g(\cdot)$ is quite general (and need not be explicitly specified). The same test was independently derived by Cook and Weisberg (1983). The score statistic for the hypothesis that the σ_i^2 are all the same, which is equivalent to $H_0: \gamma_1 = \cdots = \gamma_p = 0$, can be formulated as an auxiliary-regression problem.

Let $U_i \equiv E_i^2 / \hat{\sigma}_\varepsilon^2$, where $\hat{\sigma}_\varepsilon^2 = \sum E_i^2 / n$ is the MLE of the error variance.⁵⁹

The U_i are a type of standardized squared residuals. Regress U on the Z s:

$$U_i = \eta_0 + \eta_1 Z_{i1} + \cdots + \eta_p Z_{ip} + \omega_i \quad (12.18)$$

Breusch and Pagan (1979) show that the score statistic

$$S_0^2 = \frac{\sum (\widehat{U}_i - \bar{U})^2}{2}$$

is asymptotically distributed as χ^2 with p degrees of freedom under the null hypothesis of constant error variance. Here, the \widehat{U}_i are fitted values from the regression of U on the Z s, and thus S_0^2 is half the regression sum of squares from fitting Equation 12.18.

To apply this result, it is, of course, necessary to select Z s, the choice of which depends on the suspected pattern of nonconstant error variance. If several patterns are suspected, then several score tests can be performed. Employing X_1, \dots, X_k in the auxiliary regression (Equation 12.8), for example, permits detection of a tendency of the error variance to increase with the values of one or more of the explanatory variables in the main regression.

Likewise, Cook and Weisberg (1983) suggest regressing U on the fitted values from the main regression (i.e., fitting the auxiliary regression $U_i = \eta_0 + \eta_1 \widehat{Y}_i + \omega_i$), producing a one-degree-of-freedom score test to detect the common tendency of the error variance to increase with the level of the response variable. When the error variance follows this pattern, the auxiliary regression of U on \widehat{Y} provides a more powerful test than the more general regression of U

⁵⁹Note the division by n rather than by $n - 1$ in $\hat{\sigma}_\varepsilon^2$. See Section 9.3.3 on maximum-likelihood estimation of the linear model.

on the X s. A similar, but more complex, procedure was described by Anscombe (1961), who suggests correcting detected heteroscedasticity by transforming Y to the Box-Cox power $Y^{(\lambda)}$ with $\tilde{\lambda} = 1 - \frac{1}{2}\hat{\eta}_1\bar{Y}$.

Finally, White (1980) proposed a score test based on a comparison of his heteroscedasticity-corrected estimator of coefficient sampling variance with the usual estimator of coefficient variance.⁶⁰ If the two estimators are sufficiently different, then doubt is cast on the assumption of constant error variance. White's test can be implemented as an auxiliary regression of the squared residuals from the main regression, E_i^2 , on all the X s together with all the squares and pairwise products of the X s. Thus, for $k = 2$ explanatory variables in the main regression, we would fit the model

$$E_i^2 = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \delta_{11} X_{i1}^2 + \delta_{22} X_{i2}^2 + \delta_{12} X_{i1} X_{i2} + v_i$$

In general, there will be $p = k(k + 3)/2$ terms in the auxiliary regression, plus the constant. The score statistic for testing the null hypothesis of constant error variance is $S_0^2 = nR^2$, where R^2 is the squared multiple correlation from the auxiliary regression. Under the null hypothesis, S_0^2 follows an asymptotic χ^2 distribution with p degrees of freedom.

Because all these score tests are potentially sensitive to violations of model assumptions other than constant error variance, it is important, in practice, to supplement the tests with graphical diagnostics, as suggested by Cook and Weisberg (1983). When there are several Z s, a simple diagnostic is to plot U_i against \hat{U}_i , the fitted values from the auxiliary regression. We can also construct added-variable plots for the Z s in the auxiliary regression. When U_i is regressed on \hat{Y}_i , these plots convey essentially the same information as the plot of studentized residuals against fitted values proposed in Section 12.2.

Simple score tests are available to determine the need for a transformation and to test for nonconstant error variance.

Applied to the initial SLID regression of wages on sex, age, and education, an auxiliary regression of U on \hat{Y} yields $\hat{U} = -0.3449 + 0.08652\hat{Y}$, and $S_0^2 = 567.66/2 = 283.83$ on 1 degree of freedom. There is, consequently, very strong evidence that the error variance increases with the level of the response variable. The suggested variance-stabilizing transformation using Anscombe's rule is $\tilde{\lambda} = 1 - \frac{1}{2}(0.08652)(15.545) = 0.33$. Compare this value with those produced by the Box-Cox model ($\hat{\lambda} = 0.09$, in Section 12.5.1) and by trial and error or a spread-level plot ($\lambda = 0$, i.e., the log transformation, in Section 12.2).

An auxiliary regression of U on the explanatory variables in the main regression yields $S_0^2 = 579.08/2 = 289.54$ on $k = 3$ degrees of freedom and thus also provides strong evidence against constant error variance. The score statistic for the more general test is not much larger than that for the regression of U on \hat{Y} , implying that the pattern of nonconstant error variance is indeed for the spread of the errors to increase with the level of Y .

To perform White's test, I regressed the squared residuals from the initial SLID model on the dummy regressor for sex, age, education, the squares of age and education, and the pairwise products of the variables. It does not, of course, make sense to square the dummy

⁶⁰White's coefficient-variance estimator is described in Section 12.2.3.

regressor for sex. The resulting regression produced an R^2 of .03989 and thus a score statistic of $S_0^2 = 3997 \times 0.03989 = 159.5$ on $p = 8$ degrees of freedom, which also provides very strong evidence of nonconstant error variance.

12.6 Structural Dimension

In discussing the use and potential failure of component-plus-residual plots as a diagnostic for nonlinearity, I explained that it is unreasonable to expect that a collection of two- or three-dimensional graphs can, in every instance, adequately capture the dependence of Y on the X s: The surface representing this dependence lies, after all, in a space of $k + 1$ dimensions. Relying primarily on Cook (1994), I will now briefly consider the geometric notion of dimension in regression analysis, along with the implications of this notion for diagnosing problems with regression models that have been fit to data.⁶¹ The *structural dimension* of a regression problem corresponds to the dimensionality of the smallest subspace of the X s required to represent the dependency of Y on the X s.

Let us initially suppose that the distribution of Y is completely *independent* of the explanatory variables X_1, \dots, X_k . Then, in Cook and Weisberg's (1994) terminology, an "ideal summary" of the data is simply the univariate, unconditional distribution of Y —represented, say, by the density function $p(y)$. In a sample, we could compute a density estimate, a histogram, or some other univariate display. In this case, the structural dimension of the data is 0.

Now suppose that Y depends on the X s only through the linear regression

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

where $E(\varepsilon) = 0$ and the distribution of the error is independent of the X s. Then the expectation of Y conditional on the X s is a linear function of the X s:

$$E(Y|x_1, \dots, x_k) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

A plot of Y against $\alpha + \beta_1 X_1 + \dots + \beta_k X_k$, therefore, constitutes an ideal summary of the data. This two-dimensional plot shows the systematic component of Y in an edge-on view of the regression hyperplane and also shows the conditional variation of Y around the hyperplane (i.e., the variation of the errors). Because the subspace spanned by the linear combination $\alpha + \beta_1 X_1 + \dots + \beta_k X_k$ is one-dimensional, the structural dimension of the data is 1. In a sample, the ideal summary is a two-dimensional scatterplot of Y_i against $\hat{Y}_i = A + B_1 X_{i1} + \dots + B_k X_{ik}$; the regression line in this plot is an edge-on view of the fitted least-squares surface.

The structural dimension of the data can be 1 even if the regression is *nonlinear* or if the errors are not identically distributed, as long as the expectation of Y and the distribution of the errors depend only on a single linear combination of the X s—that is, a subspace of dimension 1. The structural dimension is 1, for example, if

$$E(Y|x_1, \dots, x_k) = f(\alpha + \beta_1 x_1 + \dots + \beta_k x_k) \tag{12.19}$$

and

⁶¹An extended discussion of structural dimension, at a more elementary level than Cook (1994), may be found in Cook and Weisberg (1994, 1999).

$$V(Y|x_1, \dots, x_k) = g(\alpha + \beta_1 x_1 + \dots + \beta_k x_k) \quad (12.20)$$

where the mean function $f(\cdot)$ and the variance function $g(\cdot)$, although generally different functions, depend on the *same* linear function of the X s. In this case, a plot of Y against $\alpha + \beta_1 X_1 + \dots + \beta_k X_k$ is still an ideal summary of the data, showing the nonlinear dependency of the expectation of Y on the X s, along with the pattern of nonconstant error variance.

Similarly, we hope to see these features of the data in a sample plot of Y against \hat{Y} from the *linear* regression of Y on the X s (even though the linear regression does not itself capture the dependency of Y on the X s). It turns out, however, that the plot of Y against \hat{Y} can fail to reflect the mean and variance functions accurately if the X s themselves are not linearly related—even when the true structural dimension is 1 (i.e., when Equations 12.19 and 12.20 hold).⁶² This, then, is another context in which linearly related explanatory variables are desirable.⁶³ Linearly related explanatory variables are not required here if the true regression is linear—something that, however, we are typically not in a position to know prior to examining the data.

The structural dimension of a regression is the dimensionality of the smallest subspace of the explanatory variables required, along with the response variable, to represent the dependence of Y on the X s. When Y is completely independent of the X s, the structural dimension is 0, and an ideal summary of the data is simply the unconditional distribution of Y . When the linear-regression model holds—or when the conditional expectation and variance of Y are functions of a single linear combination of the X s—the structural dimension is 1.

The structural dimension of the data exceeds 1 if Equations 12.19 and 12.20 do not both hold. If, for example, the mean function depends on one linear combination of the X s,

$$E(Y|x_1, \dots, x_k) = f(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)$$

and the variance function on a *different* linear combination

$$V(Y|x_1, \dots, x_k) = g(\gamma + \delta_1 x_1 + \dots + \delta_k x_k)$$

then the structural dimension is 2.

Correspondingly, if the mean function depends on *two different* linear combinations of the X s, implying interaction among the X s,

$$E(Y|x_1, \dots, x_k) = f(\alpha + \beta_1 x_1 + \dots + \beta_k x_k, \gamma + \delta_1 x_1 + \dots + \delta_k x_k)$$

while the errors are independent of the X s, then the structural dimension is also 2. When the structural dimension is 2, a plot of Y against \hat{Y} (from the linear regression of Y on the X s) is necessarily incomplete.

⁶²See Exercise 12.11.

⁶³The requirement of linearity here is, in fact, stronger than pairwise linear relationships among the X s: The regression of any linear function of the X s on any set of linear functions of the X s must be linear. If the X s are multivariate normal, then this condition is necessarily satisfied (although it may be satisfied even if the X s are not normal). It is not possible to check for linearity in this strict sense when there are more than two or three X s, but there is some evidence that checking pairs—and perhaps triples—of X s is usually sufficient. See Cook and Weisberg (1994). Cf. Section 12.3.3 for the conditions under which component-plus-residual plots are informative.

These observations are interesting, but their practical import—beyond the advantage of linearly related regressors—is unclear: Short of modeling the regression of Y on the X s nonparametrically, we can never be sure that we have captured all the structure of the data in a lower-dimensional subspace of the explanatory variables.

There is, however, a further result that *does* have direct practical application: Suppose that the explanatory variables are linearly related and that there is one-dimensional structure. Then the *inverse regressions* of each of the explanatory variables on the response variable have the following properties:

$$\begin{aligned} E(X_j|y) &= \mu_j + \eta_j m(y) \\ V(X_j|y) &\approx \sigma_j^2 + \eta_j^2 v(y) \end{aligned} \tag{12.21}$$

Equation 12.21 has two special features that are useful in checking whether a one-dimensional structure is reasonable for a set of data.⁶⁴

1. Most important, the functions $m(\cdot)$ and $v(\cdot)$, through which the means and variances of the X s depend on Y , are the same for all the X s. Consequently, if the scatterplot of X_1 against Y shows a linear relationship, for example, then the scatterplots of each of X_2, \dots, X_k against Y must also show linear relationships. If one of these relationships is quadratic, in contrast, then the others must be quadratic. Likewise, if the variance of X_1 increases linearly with the level of Y , then the variances of the other X s must also be linearly related to Y . There is only one exception: The constant η_j can be 0, in which case the mean and variance of the corresponding X_j are *unrelated* to Y .
2. The constant η_j appears in the formula for the conditional mean of X_j and η_j^2 in the formula for its conditional variance, placing constraints on the patterns of these relationships. If, for example, the mean of X_1 is unrelated to Y , then the variance of X_1 should also be unrelated to Y .

The sample inverse regressions of the X s on Y can be conveniently examined in the first column of the scatterplot matrix for $\{Y, X_1, \dots, X_k\}$. An illustrative application is shown in Figure 12.17, for the regression of prestige on education, income, and percent women, for the Canadian occupational prestige data.⁶⁵ Here, I have log-transformed income and taken the logit of percent women to make the relationships among the explanatory variables more nearly linear. The inverse-response plots in the first column of the scatterplot matrix show roughly similar patterns, as required for a one-dimensional structure.

If the structural dimension is 1, and if the explanatory variables are linearly related to one another, then the inverse regressions of the explanatory variables on the response variable all have the same general form.

⁶⁴Equation 12.21 is the basis for formal dimension-testing methods, such as *sliced inverse regression* (Duan & Li, 1991) and related techniques. See Cook and Weisberg (1994, 1999) for an introductory treatment of dimension testing and for additional references.

⁶⁵This data set was introduced in Chapter 2 and used for an example of multiple regression in Chapter 4.

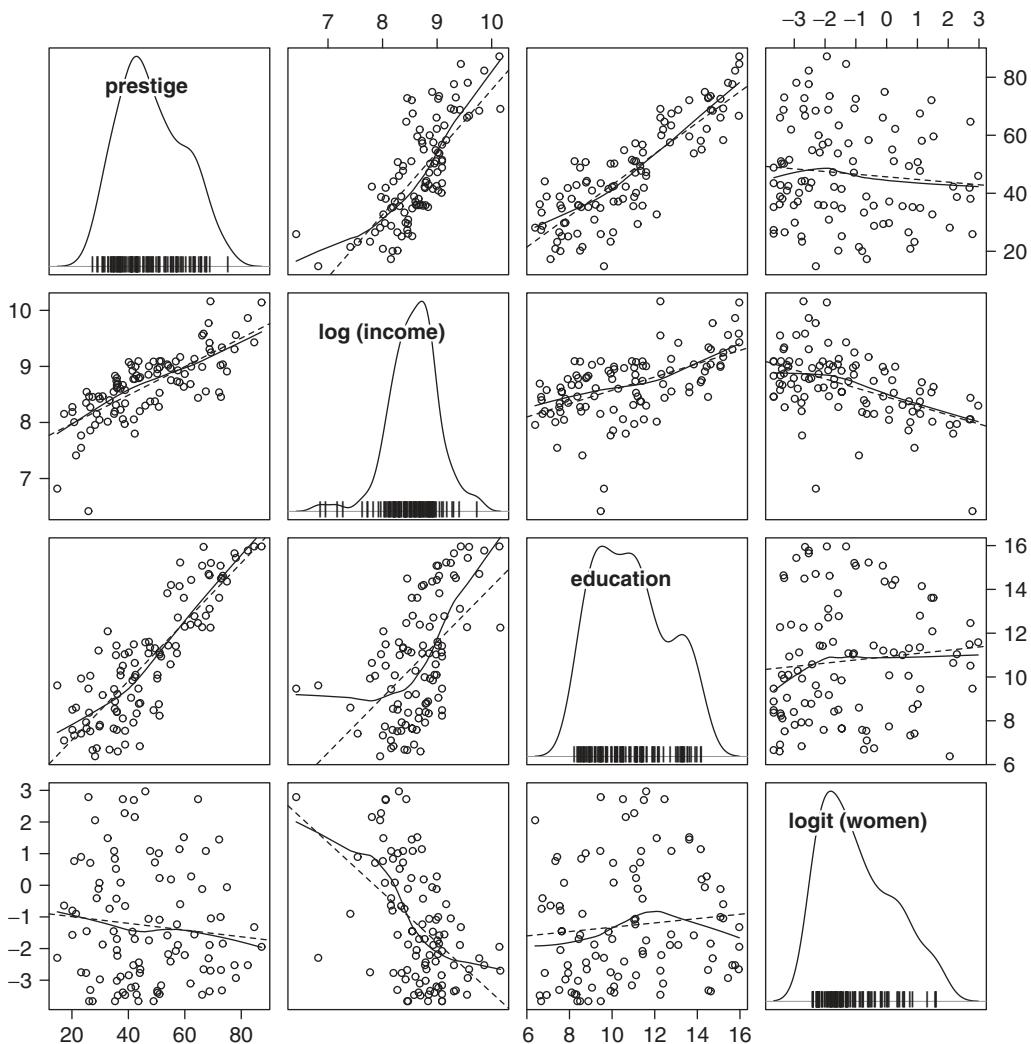


Figure 12.17 Scatterplot matrix for prestige, log of income, education, and logit of percent women in the Canadian occupational prestige data. The inverse response plots are in the first column of the scatterplot matrix. In each panel, the solid line is for a lowess smooth with span 3/4, while the broken line is the least-squares line. Kernel-density estimates are given on the diagonal, with the rug-plot at the bottom of each diagonal panel showing the location of the observations.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 12.1. *Show that the correlation between the least-squares residuals E_i and the response-variable values Y_i is $\sqrt{1 - R^2}$. [Hint: Use the geometric vector representation of multiple regression (developed in Chapter 10), examining the plane in which the \mathbf{e} , \mathbf{y}^* , and $\hat{\mathbf{y}}^*$ vectors lie.]

Exercise 12.2. Nonconstant variance and specification error: Generate 100 observations according to the following model:

$$Y = 10 + (1 \times X) + (1 \times D) + (2 \times X \times D) + \varepsilon$$

where $\varepsilon \sim N(0, 10^2)$; the values of X are $1, 2, \dots, 50, 1, 2, \dots, 50$; the first 50 values of D are 0; and the last 50 values of D are 1. Then regress Y on X alone (i.e., omitting D and XD), $Y = A + BX + E$. Plot the residuals E from this regression against the fitted values \hat{Y} . Is the variance of the residuals constant? How do you account for the pattern in the plot?

Exercise 12.3. *Weighted-least-squares estimation: Suppose that the errors from the linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ are independent and normally distributed, but with different variances, $\varepsilon_i \sim N(0, \sigma_i^2)$, and that $\sigma_i^2 = \sigma_\varepsilon^2/w_i^2$. Show that:

- (a) The likelihood for the model is

$$L(\beta, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \Sigma (\mathbf{y} - \mathbf{X}\beta) \right]$$

where

$$\Sigma = \sigma_\varepsilon^2 \times \text{diag}\{1/w_1^2, \dots, 1/w_n^2\} \equiv \sigma_\varepsilon^2 \mathbf{W}^{-1}$$

- (b) The maximum-likelihood estimators of β and σ_ε^2 are

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y} \\ \hat{\sigma}_\varepsilon^2 &= \frac{\sum (E_i/w_i)^2}{n} \end{aligned}$$

where $\mathbf{e} = \{E_i\} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

- (c) The MLE is equivalent to minimizing the weighted sum of squares $\sum w_i^2 E_i^2$.
(d) The estimated asymptotic covariance matrix of $\hat{\beta}$ is given by

$$\hat{\mathcal{V}}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$$

Exercise 12.4. *Show that when the covariance matrix of the errors is

$$\Sigma = \sigma_\varepsilon^2 \times \text{diag}\{1/W_1^2, \dots, 1/W_n^2\} \equiv \sigma_\varepsilon^2 \mathbf{W}^{-1}$$

the weighted-least-squares estimator

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y} \\ &= \mathbf{M} \mathbf{y} \end{aligned}$$

is the minimum-variance linear unbiased estimator of β (*Hint:* Adapt the proof of the Gauss-Markov theorem for OLS estimation given in Section 9.3.2.)

Exercise 12.5. *The impact of nonconstant error variance on OLS estimation: Suppose that $Y_i = \alpha + \beta x_i + \varepsilon_i$, with independent errors, $\varepsilon_i \sim N(0, \sigma_i^2)$, and $\sigma_i = \sigma_\varepsilon x_i$. Let B represent the OLS estimator and $\hat{\beta}$ the WLS estimator of β .

- (a) Show that the sampling variance of the OLS estimator is

$$V(B) = \frac{\sum (X_i - \bar{X})^2 \sigma_i^2}{\left[\sum (X_i - \bar{X})^2 \right]^2}$$

and that the sampling variance of the WLS estimator is

$$V(\hat{\beta}) = \frac{\sigma_e^2}{\sum w_i^2 (X_i - \tilde{X})^2}$$

where $\tilde{X} \equiv (\sum w_i^2 X_i) / (\sum w_i^2)$. (*Hint:* Write each slope estimator as a linear function of the Y_i .)

- (b) Now suppose that x is uniformly distributed over the interval $[x_0, ax_0]$, where $x_0 > 0$ and $a > 0$, so that a is the ratio of the largest to the smallest σ_i . The efficiency of the OLS estimator relative to the optimal WLS estimator is $V(\hat{\beta})/V(B)$, and the relative precision of the OLS estimator is the square root of this ratio, that is, $SD(\hat{\beta})/SD(B)$. Calculate the relative precision of the OLS estimator for all combinations of $a = 2, 3, 5, 10$ and $n = 5, 10, 20, 50, 100$. For example, when $a = 3$ and $n = 10$, you can take the x -values as 1, 1.222, 1.444, ..., 2.778, 3. Under what circumstances is the OLS estimator much less precise than the WLS estimator?

- (c) The usual variance estimate for the OLS slope (assuming constant error variance) is

$$\hat{V}(B) = \frac{S_E^2}{\sum (X_i - \bar{X})^2}$$

where $S_E^2 = \sum E_i^2 / (n - 2)$. Kmenta (1986, Section 8.2) shows that the expectation of this variance estimator (under nonconstant error variance σ_i^2) is

$$E[\hat{V}(B)] = \frac{\bar{\sigma}^2}{\sum (X_i - \bar{X})^2} - \frac{\sum (X_i - \bar{X})^2 (\sigma_i^2 - \bar{\sigma}^2)}{(n - 2)[\sum (X_i - \bar{X})^2]^2}$$

where $\bar{\sigma}^2 \equiv \sum \sigma_i^2 / n$. (*Prove this result.) Kmenta also shows that the true variance of the OLS slope estimator, $V(B)$ [derived in part (a)], is generally different from $E[\hat{V}(B)]$. If $\sqrt{E[\hat{V}(B)]/V(B)}$ is substantially below 1, then the usual formula for the standard deviation of B will lead us to believe that the OLS estimator is more precise than it really is. Calculate $\sqrt{E[\hat{V}(B)]/V(B)}$ under the conditions of part (b), for $a = 5, 10, 20, 50$ and $n = 5, 10, 20, 50, 100$. What do you conclude about the robustness of validity of OLS inference with respect to nonconstant error variance?

Exercise 12.6. Experimenting with component-plus-residual plots: Generate random samples of 100 observations according to each of the following schemes. In each case, construct the component-plus-residual plots for X_1 and X_2 . Do these plots accurately capture the partial relationships between Y and each of X_1 and X_2 ? Whenever they appear, ε and δ are $N(0, 1)$ and independent of each other and of the other variables.

- (a) Independent X s and a linear regression: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $Y = X_1 + X_2 + 0.1\epsilon$.
- (b) Linearly related X s and a linear regression: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1 + 0.1\delta$; $Y = X_1 + X_2 + 0.1\epsilon$.
- (c) Independent X s and a nonlinear regression on one X : X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.1\epsilon$.
- (d) Linearly related X s and a nonlinear regression on one X : X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1 + 0.1\delta$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.1\epsilon$. (Note the “leakage” here from X_1 to X_2 .)
- (e) Nonlinearly related X s and a linear regression: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $Y = X_1 + X_2 + 0.02\epsilon$.
- (f) Nonlinearly related X s and a linear regression on one X : X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $Y = 2(X_1 - 0.5)^2 + X_2 + 0.02\epsilon$. (Note how strong a nonlinear relationship between the X s and how small an error variance in the regression are required for the effects in this example to be noticeable.)

Exercise 12.7. Consider an alternative analysis of the SLID data in which log wages is regressed on sex, transformed education, and transformed age—that is, try to straighten the relationship between log wages and age by a transformation rather than by a quadratic regression. How successful is this approach? (*Hint:* Use a negative start, say age -15 , prior to transforming age.)

Exercise 12.8. Apply Mallows’s procedure to construct augmented component-plus-residual plots for the SLID regression of log wages on sex, age, and education. *Then apply Cook’s CERES procedure to this regression. Compare the results of these two procedures with each other and with the ordinary component-plus-residual plots in Figure 12.6. Do the more complex procedures give clearer indications of nonlinearity in this case?

Exercise 12.9. *Box-Cox transformations of Y : In matrix form, the Box-Cox regression model given in Section 12.5.1 can be written as

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- (a) Show that the probability density for the observations is given by

$$p(y) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left[-\frac{\sum_{i=1}^n (Y_i^{(\lambda)} - \mathbf{x}'_i \boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right] \prod_{i=1}^n Y_i^{\lambda-1}$$

- where \mathbf{x}'_i is the i th row of \mathbf{X} . (*Hint:* $Y_i^{\lambda-1}$ is the Jacobian of the transformation from Y_i to ε_i .)
- (b) For a given value of λ , the *conditional* maximum-likelihood estimator of $\boldsymbol{\beta}$ is the least-squares estimator

$$\mathbf{b}_\lambda = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(\lambda)}$$

(Why?) Show that the maximized log-likelihood can be written as

$$\begin{aligned} \log_e L(\alpha, \beta_1, \dots, \beta_k, \sigma_\varepsilon^2 | \lambda) \\ = -\frac{n}{2}(1 + \log_e 2\pi) - \frac{n}{2} \log_e \hat{\sigma}_\varepsilon^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log_e Y_i \end{aligned}$$

as stated in the text.

Recall from footnote 55 (page 324) that the distribution of Y^λ cannot really be normal for more than one value of λ .

Exercise 12.10. *Box-Tidwell transformations of the X s: Recall the Box-Tidwell model

$$Y_i = \alpha + \beta_1 X_i^{\gamma_1} + \dots + \beta_k X_i^{\gamma_k} + \varepsilon_i$$

and focus on the first regressor, X_1 . Show that the first-order Taylor-series approximation for $X_1^{\gamma_1}$ at $\gamma_1 = 1$ is

$$X_1^{\gamma_1} \approx X_1 + (\gamma_1 - 1)X_1 \log_e X_1$$

providing the basis for the constructed variable $X_1 \log_e X_1$.

Exercise 12.11. *Experimenting with structural dimension: Generate random samples of 100 observations according to each of the following schemes. In each case, fit the linear regression of Y on X_1 and X_2 , and plot the values of Y against the resulting fitted values \hat{Y} . Do these plots accurately capture the dependence of Y on X_1 and X_2 ? To decide this question in each case, it may help (1) to draw graphs of $E(Y|x_1, x_2) = f(\alpha + \beta_1 x_1 + \beta_2 x_2)$ and $V(Y|x_1, x_2) = g(\alpha + \beta_1 x_1 + \beta_2 x_2)$ over the observed range of values for $\alpha + \beta_1 X_1 + \beta_2 X_2$ and (2) to plot a nonparametric-regression smooth in the plot of Y against \hat{Y} . Whenever they appear, ε and δ are $N(0, 1)$ and independent of each other and of the other variables.

- (a) Independent X s, a linear regression, and constant error variance: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $E(Y|x_1, x_2) = x_1 + x_2$; $V(Y|x_1, x_2) = 0.1\varepsilon$.
- (b) Independent X s, mean and variance of Y dependent on the same linear function of the X s: X_1 and X_2 independent and uniformly distributed on the interval $[0, 1]$; $E(Y|x_1, x_2) = (x_1 + x_2 - 1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1 + x_2 - 1| \times \varepsilon$.
- (c) Linearly related X s, mean and variance of Y dependent on the same linear function of the X s: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = X_1 + 0.1\delta$; $E(Y|x_1, x_2) = (x_1 + x_2 - 1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1 + x_2 - 1| \times \varepsilon$.
- (d) Nonlinearly related X s, mean and variance of Y dependent on the same linear function of the X s: X_1 uniformly distributed on the interval $[0, 1]$; $X_2 = |X_1 - 0.5|$; $E(Y|x_1, x_2) = (x_1 + x_2 - 1)^2$; $V(Y|x_1, x_2) = 0.1 \times |x_1 + x_2 - 1| \times \varepsilon$.

Summary

- Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit. Non-normality can often be detected by examining the distribution of the least-squares residuals and frequently can be corrected by transforming the data.

- It is common for the variance of the errors to increase with the level of the response variable. This pattern of nonconstant error variance (“heteroscedasticity”) can often be detected in a plot of residuals against fitted values. Strategies for dealing with nonconstant error variance include transformation of the response variable to stabilize the variance, the substitution of weighted-least-squares estimation for ordinary least squares, and the correction of coefficient standard errors for heteroscedasticity. A rough rule is that nonconstant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more (or, more conservatively, about 4 or more).
- Simple forms of nonlinearity can often be detected in component-plus-residual plots. Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an explanatory variable, for example). Component-plus-residual plots reliably reflect nonlinearity when there are not strong nonlinear relationships among the explanatory variables in a regression. More complex versions of these displays, such as augmented component-plus-residual plots and CERES plots, are more robust.
- Discrete explanatory variables divide the data into groups. A simple incremental F -test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means. Likewise, tests of nonconstant variance can be based on comparisons of spread in the different groups.
- A statistically sophisticated general approach to selecting a transformation of Y or an X is to embed the linear-regression model in a more general model that contains a parameter for the transformation. The Box-Cox procedure selects a power transformation of Y to normalize the errors. The Box-Tidwell procedure selects power transformations of the X s to linearize the regression of Y on the X s. In both cases, “constructed-variable” plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.
- Simple score tests are available to determine the need for a transformation and to test for nonconstant error variance.
- The structural dimension of a regression is the dimensionality of the smallest subspace of the explanatory variables required, along with the response variable, to represent the dependence of Y on the X s. When Y is completely independent of the X s, the structural dimension is 0, and an ideal summary of the data is simply the unconditional distribution of Y . When the linear-regression model holds—or when the conditional expectation and variance of Y are functions of a single linear combination of the X s—the structural dimension is 1. If the structural dimension is 1, and if the explanatory variables are linearly related to one another, then the inverse regressions of the explanatory variables on the response variable all have the same general form.

Recommended Reading

Methods for diagnosing problems in regression analysis and for visualizing regression data have been the subject of a great deal of research in statistics. The following texts summarize the state of the art and include extensive references to the journal literature.

- Cook and Weisberg (1994, 1999) present a lucid and accessible treatment of many of the topics discussed in this chapter. They also describe a freely available computer program written in Lisp-Stat, called Arc, that implements the graphical methods presented in their books (and much more). See Cook (1998) for a more advanced treatment of much the same material. Also see Weisberg (2014) for an accessible account of these methods.
- Cleveland (1993) describes novel graphical methods for regression data, including two-dimensional, three-dimensional, and higher-dimensional displays.
- Atkinson (1985) has written an interesting, if somewhat idiosyncratic, book that stresses the author's important contributions to regression diagnostics. There is, therefore, an emphasis on diagnostics that yield constructed-variable plots. This text includes a strong treatment of transformations and a discussion of the extension of least-squares diagnostics to generalized linear models (e.g., logistic regression, as described in Chapters 14 and 15⁶⁶).

⁶⁶See, in particular, Section 15.4.

13

Collinearity and Its Purported Remedies

As I have explained, when there is a perfect linear relationship among the regressors in a linear model, the least-squares coefficients are not uniquely defined.¹ A strong, but less-than-perfect, linear relationship among the X s causes the least-squares coefficients to be unstable: Coefficient standard errors are large, reflecting the imprecision of estimation of the β s; consequently, confidence intervals for the β s are broad, and hypothesis tests have low power. Small changes in the data—even, in extreme cases, due to rounding errors—can greatly alter the least-squares coefficients, and relatively large changes in the coefficients from the least-squares values hardly increase the sum of squared residuals from its minimum (i.e., the least-squares coefficients are not sharply defined).

This chapter describes methods for detecting collinearity and techniques that are often employed for dealing with collinearity when it is present. I would like to make three important points at the outset, however:

1. Except in certain specific contexts—such as time-series regression² or regression with aggregated data—collinearity is a comparatively rare problem in social science applications of linear models. Insufficient variation in explanatory variables, small samples, and large error variance (i.e., weak relationships) are much more frequently the source of imprecision in estimation.
2. Methods that are commonly employed as cures for collinearity—in particular, biased estimation and variable selection—can easily be worse than the disease. A principal goal of this chapter is to explain the substantial limitations of this statistical snake oil.
3. It is not at all obvious that the detection of collinearity in data has practical implications. There are, as mentioned in Point 1, several sources of imprecision in estimation, which can augment or partially offset each other. The standard errors of the regression coefficients are the “bottom line”: If the coefficient estimates are sufficiently precise, then the degree of collinearity is irrelevant; if the estimated coefficients are *insufficiently* precise, then knowing that the culprit is collinearity is of use only if the study can be redesigned to decrease the correlations among the X s. In observational studies, where the X s are sampled along with Y , it is usually impossible to influence their correlational

¹See Sections 5.2 and 9.2.

²See the example developed below. Chapter 16 describes methods for time-series regression that take account of dependence among the errors.

structure, but it may very well be possible to improve the precision of estimation by increasing the sample size or by decreasing the error variance.³

13.1 Detecting Collinearity

We have encountered the notion of collinearity at several points, and it is therefore useful to summarize what we know:

- When there is a perfect linear relationship among the X s,

$$c_1X_{i1} + c_2X_{i2} + \cdots + c_kX_{ik} = c_0$$

where the constants c_1, c_2, \dots, c_k are not all 0,

1. the least-squares normal equations do not have a unique solution, and
2. the sampling variances of the regression coefficients are infinite.

Perfect collinearity is usually the product of some error in formulating the linear model, such as failing to employ a baseline category in dummy regression.

*Points 1 and 2 follow from the observation that the matrix $\mathbf{X}'\mathbf{X}$ of sums of squares and products is singular. Moreover, because the columns of \mathbf{X} are perfectly collinear, the regressor subspace is of deficient dimension.

- When collinearity is less than perfect:

1. The sampling variance of the least-squares slope coefficient B_j is

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_e^2}{(n - 1)S_j^2}$$

where R_j^2 is the squared multiple correlation for the regression of X_j on the other X s, and $S_j^2 = \sum(X_{ij} - \bar{X}_j)^2/(n - 1)$ is the variance of X_j . The term $1/(1 - R_j^2)$, called the *variance-inflation factor* (VIF), directly and straightforwardly indicates the impact of collinearity on the precision of B_j . Because the precision of estimation of β_j is most naturally expressed as the width of the confidence interval for this parameter, and because the width of the confidence interval is proportional to the standard deviation of B_j (not its variance), I recommend examining the square root of the VIF in preference to the VIF itself. Figure 13.1 reveals that the linear relationship among the X s must be very strong before collinearity seriously impairs the precision of estimation: It is not until R_j approaches .9 that the precision of estimation is halved.

Because of its simplicity and direct interpretation, the VIF (or its square root) is the basic diagnostic for collinearity. It is not, however, applicable to sets of related

³The error variance can sometimes be decreased by improving the procedures of the study or by introducing additional explanatory variables. The latter remedy may, however, increase collinearity and may change the nature of the research. It may be possible, in some contexts, to increase precision by increasing the variation of the X s, but only if their values are under the control of the researcher, in which case collinearity could also be reduced. Sometimes, however, researchers may be able to exert indirect control over the variational and correlational structure of the X s by selecting a research setting judiciously or by designing an advantageous sampling procedure.

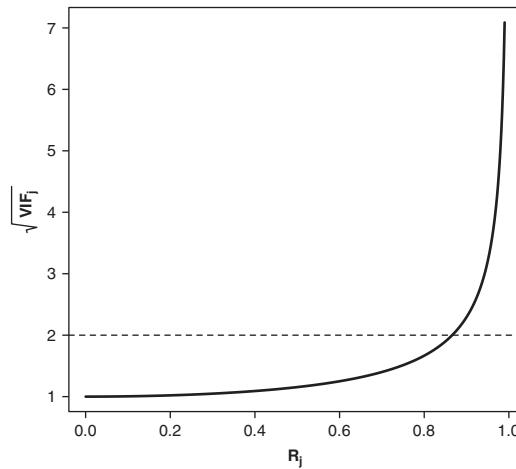


Figure 13.1 Precision of estimation (square root of the variance-inflation factor) of β_j as a function of the multiple correlation R_j between X_j and the other explanatory variables. It is not until the multiple correlation gets very large that the precision of estimation is seriously degraded.

regressors, such as sets of dummy-variable coefficients, or coefficients for polynomial regressors.⁴

2. When X_1 is strongly collinear with the other regressors, the residuals $X^{(1)}$ from the regression of X_1 on X_2, \dots, X_k show little variation—most of the variation in X_1 is accounted for by its regression on the other X s. The added-variable plot graphs the residuals from the regression of Y on X_2, \dots, X_k against $X^{(1)}$, converting the multiple regression into a simple regression.⁵ Because the explanatory variable in this plot, $X^{(1)}$, is nearly invariant, the slope B_1 is subject to substantial sampling variation.⁶
3. *Confidence intervals for individual regression coefficients are projections of the confidence interval–generating ellipse. Because this ellipse is the inverse—that is, the rescaled, 90° rotation—of the data ellipse for the explanatory variables, the individual confidence intervals for the coefficients are wide. If the correlations among the X s are positive, however, then there is considerable information in the data about the *sum* of the regression coefficients, if not about individual coefficients.⁷

⁴Section 13.1.2 describes a generalization of variance inflation to sets of related regressors.

⁵More precisely, the multiple regression is converted into a sequence of simple regressions, for each X in turn. Added-variable plots are discussed in Section 11.6.1, particularly Figure 11.9 (page 284).

⁶See Stine (1995) for a nice graphical interpretation of this point.

⁷See the discussion of joint confidence regions for regression coefficients in Section 9.4.4 and in particular Figure 9.2 (page 223).

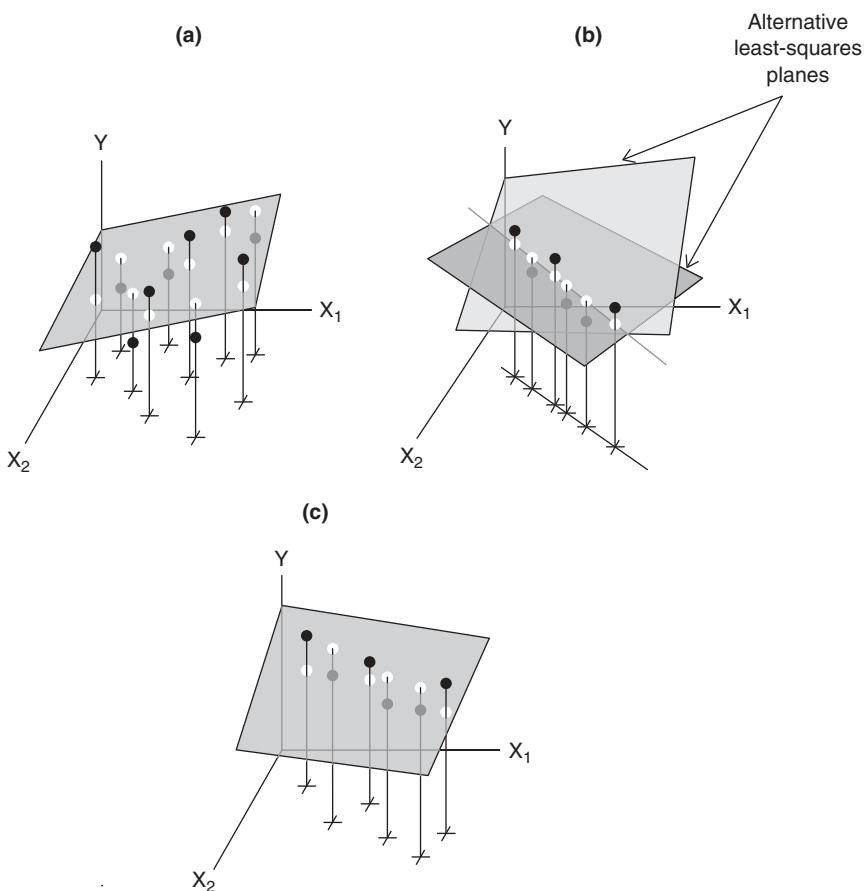


Figure 13.2 The impact of collinearity on the stability of the least-squares regression plane. In (a), the correlation between X_1 and X_2 is small, and the regression plane therefore has a broad base of support. In (b), X_1 and X_2 are perfectly correlated; the least-squares plane is not uniquely defined. In (c), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2 ; the least-squares plane is uniquely defined, but it is not well supported by the data.

When the regressors in a linear model are perfectly collinear, the least-squares coefficients are not unique. Strong, but less-than-perfect, collinearity substantially increases the sampling variances of the least-squares coefficients and can render them useless as estimators. The variance-inflation factor $VIF_j = 1/(1 - R_j^2)$ indicates the deleterious impact of collinearity on the precision of the estimate B_j .

Collinearity is sometimes termed *multicollinearity*, which has the virtue of emphasizing that collinear relationships are not limited to strong correlations between pairs of explanatory variables.

Figures 13.2 and 13.3 provide further insight into collinearity, illustrating its effect on estimation when there are two explanatory variables in a regression. The black and gray dots in

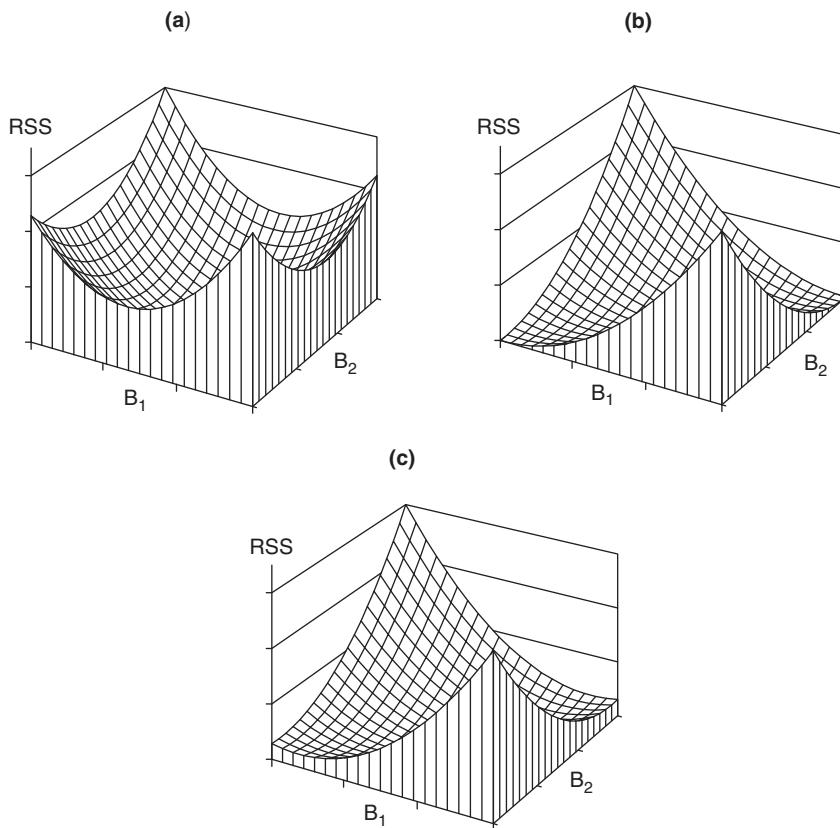


Figure 13.3 The residual sum of squares as a function of the slope coefficients B_1 and B_2 . In each graph, the vertical axis is scaled so that the least-squares value of RSS is at the bottom of the axis. When, as in (a), the correlation between the explanatory variables X_1 and X_2 is small, the residual sum of squares has a well-defined minimum, much like a deep bowl. When there is a perfect linear relationship between X_1 and X_2 , as in (b), the residual sum of squares is flat at its minimum, above a line in the $\{B_1, B_2\}$ plane: The least-squares values of B_1 and B_2 are not unique. When, as in (c), there is a strong, but less-than-perfect, linear relationship between X_1 and X_2 , the residual sum of squares is nearly flat at its minimum, so values of B_1 and B_2 quite different from the least-squares values are associated with residual sums of squares near the minimum.

Figure 13.2 represent the data points (the gray dots are below the regression plane), while the white dots represent fitted values lying in the regression plane; the +s show the projection of the data points onto the $\{X_1, X_2\}$ plane. Figure 13.3 shows the sum of squared residuals as a function of the slope coefficients B_1 and B_2 . The residual sum of squares is at a minimum, of course, when the B s are equal to the least-squares estimates; the vertical axis is scaled so that the minimum is at the “floor” of the graphs.⁸

⁸For each pair of slopes B_1 and B_2 , the intercept A is chosen to make the residual sum of squares as small as possible.

In Figure 13.2(a), the correlation between the explanatory variables X_1 and X_2 is slight, as indicated by the broad scatter of points in the $\{X_1, X_2\}$ plane. The least-squares regression plane, also shown in this figure, therefore has a firm base of support. Correspondingly, Figure 13.3(a) shows that small changes in the regression coefficients are associated with relatively large increases in the residual sum of squares—the sum-of-squares function is like a deep bowl, with steep sides and a well-defined minimum.

In Figure 13.2(b), X_1 and X_2 are perfectly collinear. Because the explanatory-variable observations form a line in the $\{X_1, X_2\}$ plane, the least-squares regression plane, in effect, also reduces to a line. The plane can tip about this line without changing the residual sum of squares, as Figure 13.3(b) reveals: The sum-of-squares function is flat at its minimum along a line defining pairs of values for B_1 and B_2 —rather like a sheet of paper with two corners raised—and thus there are an infinite number of pairs of coefficients (B_1, B_2) that yield the minimum RSS.

Finally, in Figure 13.2(c), the linear relationship between X_1 and X_2 is strong, although not perfect. The support afforded to the least-squares plane is tenuous, so that the plane can be tipped without causing large increases in the residual sum of squares, as is apparent in Figure 13.3(c)—the sum-of-squares function is like a shallow bowl with a nearly flat bottom and hence a poorly defined minimum.

Illustrative data on Canadian women's labor force participation in the postwar period, drawn from B. Fox (1980), are shown in Figure 13.4. These are time-series data, with yearly observations from 1946 through 1975. Fox was interested in determining how women's labor force participation (measured here as the percentage of adult women in the workforce) is related to several factors indicative of the supply of and demand for women's labor. The explanatory variables in the analysis include the total fertility rate (the expected number of births to a cohort of 1000 women who proceed through their childbearing years at current age-specific fertility rates), men's and women's average weekly wages (expressed in constant 1935 dollars and adjusted for current tax rates), per-capita consumer debt (also in constant dollars), and the prevalence of part-time work (measured as the percentage of the active workforce working 34 hours a week or less). Women's wages, consumer debt, and the prevalence of part-time work were expected to affect women's labor force participation positively, while fertility and men's wages were expected to have negative effects.

The time-series plots in Figure 13.4 do not bode well for the regression of women's labor force participation on the other variables: Several of the explanatory variables evidence strong linear trends over time and consequently are highly correlated with one another. Moreover, to control for factors that change regularly with time but are not included explicitly in the regression model, the author also included time (years, from 1 to 30) as an explanatory variable.⁹ Correlations among the variables in the data set, including time, are given in Table 13.1; some of these correlations are very large. As mentioned previously, time-series regression is a research context in which collinearity problems are common.

⁹The specification of time as an explanatory variable in a time-series regression is a common (if crude) strategy. As well, the use of time-series data in regression casts doubt on the assumption that errors from different observations are independent because the observation for one period is likely to share unmeasured characteristics with observations from other periods close to it in time. In the present case, however, examination of the least-squares residuals supports the reasonableness of the assumption of independent errors. Time-series regression is taken up at greater depth in Chapter 16.

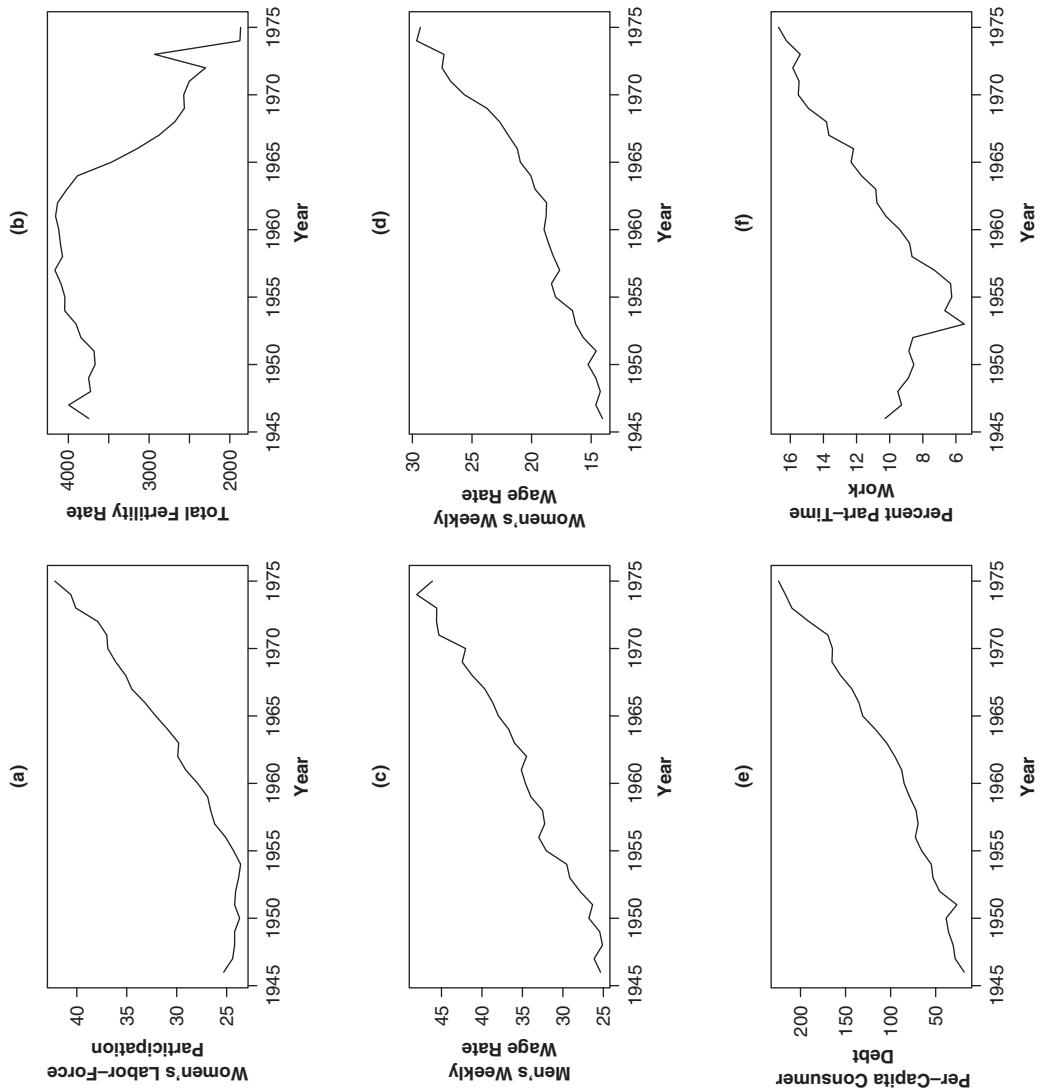


Figure 13.4 Time-series data on Canadian women's labor-force participation and other variables.

SOURCE: B. Fox (1980).

Table 13.1 Correlations Among the Variables in B. Fox's Canadian Women's Labor Force Data

	<i>L</i>	<i>F</i>	<i>M</i>	<i>W</i>	<i>D</i>	<i>P</i>	<i>T</i>
Labor-Force Participation	1.0000						
Fertility	-.9011	1.0000					
Men's Wages	.9595	-.8118	1.0000				
Women's Wages	.9674	-.8721	.9830	1.0000			
Consumer Debt	.9819	-.8696	.9861	.9868	1.0000		
Part-Time Work	.9504	-.8961	.8533	.8715	.8875	1.0000	
Time	.9531	-.7786	.9891	.9637	.9805	.8459	1.0000

The plot of the total fertility rate in Figure 13.4(b) also suggests a possible error in the data: There is an unusual jump in the total fertility rate for 1973. As it turns out, the TFR for this year was misrecorded as 2931; the correct value is 1931. This correction is reflected in the analysis reported below.¹⁰

The results of a least-squares regression of women's labor force participation on the several explanatory variables prove disappointing despite a very large R^2 of .9935. Table 13.2 shows estimated regression coefficients, standard errors, and *p*-values for the slope coefficients (for the hypothesis that each coefficient is 0). All the coefficients have the anticipated signs, but some are small (taking into account their units of measurement, of course), and most have very large standard errors despite the large multiple correlation.

Square-root variance-inflation factors for the slope coefficients in the model are as follows:

<i>Fertility</i>	<i>Men's Wages</i>	<i>Women's Wages</i>
3.89	10.67	8.21
<i>Consumer Debt</i>	<i>Part-Time Work</i>	<i>Time</i>
11.47	2.75	9.75

All are large, and most are very large, contributing to the big standard errors that I noted.

13.1.1. Principal Components*

The method of principal components, introduced by Karl Pearson (1901) and Harold Hotelling (1933), provides a useful representation of the correlational structure of a set of variables. I will describe the method briefly here, with particular reference to its application to

¹⁰This example illustrates once again the importance of examining data prior to their analysis. Apparently, I had not learned that lesson sufficiently when I used these data in the 1980s.

Table 13.2 Regression of Women's Labor Force Participation on Several Explanatory Variables

Coefficient	Estimate	Standard Error	p
Constant	16.80	3.72	
Fertility	-0.000001949	0.0005011	.99
Men's Wages	-0.02919	0.1502	.85
Women's Wages	0.01984	0.1744	.91
Consumer Debt	0.06397	0.01850	.0021
Part-Time Work	0.6566	0.0821	<.0001
Time	0.004452	0.1107	.97

collinearity in regression; more complete accounts can be obtained from texts on multivariate statistics (e.g., Morrison, 2005, chap. 8). Because the material in this section is relatively complex, the section includes a summary; you may, on first reading, wish to pass lightly over most of the section and refer primarily to the summary and to the two-variable case, which is treated immediately prior to the summary.¹¹

We begin with the vectors of standardized regressors, $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k$. Because vectors have length equal to the square root of their sum of squared elements, each \mathbf{z}_j has length $\sqrt{n - 1}$. As we will see, the *principal components* $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_p$ provide an orthogonal basis for the regressor subspace.¹² The first principal component, \mathbf{w}_1 , is oriented so as to account for maximum collective variation in the \mathbf{z}_j s; the second principal component, \mathbf{w}_2 , is orthogonal to \mathbf{w}_1 and—under this restriction of orthogonality—is oriented to account for maximum remaining variation in the \mathbf{z}_j s; the third component, \mathbf{w}_3 , is orthogonal to \mathbf{w}_1 and \mathbf{w}_2 ; and so on. Each principal component is scaled so that its variance is equal to the combined regressor variance for which it accounts.

There are as many principal components as there are linearly independent regressors: $p = \text{rank}(\mathbf{Z}_X)$, where $\mathbf{Z}_X \equiv [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k]$. Although the method of principal components is more general, I will assume throughout most of this discussion that the regressors are not *perfectly* collinear and, consequently, that $p = k$.

Because the principal components lie in the regressor subspace, each is a linear combination of the regressors. Thus, the first principal component can be written as

$$\begin{aligned}\mathbf{w}_1 &= A_{11}\mathbf{z}_1 + A_{21}\mathbf{z}_2 + \cdots + A_{k1}\mathbf{z}_k \\ &= \mathbf{Z}_X \mathbf{a}_1\end{aligned}$$

The variance of the first component is

$$S_{W_1}^2 = \frac{1}{n - 1} \mathbf{w}'_1 \mathbf{w}_1 = \frac{1}{n - 1} \mathbf{a}'_1 \mathbf{Z}'_X \mathbf{Z}_X \mathbf{a}_1 = \mathbf{a}'_1 \mathbf{R}_{XX} \mathbf{a}_1$$

where $\mathbf{R}_{XX} \equiv [1/(n - 1)]\mathbf{Z}'_X \mathbf{Z}_X$ is the correlation matrix of the regressors.

¹¹Online Appendix B on matrices, linear algebra, and vector geometry provides background for this section.

¹²It is also possible to find principal components of the *unstandardized* regressors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, but these are not generally interpretable unless all the X s are measured on the same scale.

We want to maximize $S_{W_1}^2$, but, to make maximization meaningful, it is necessary to constrain the coefficients \mathbf{a}_1 . In the absence of a constraint, $S_{W_1}^2$ can be made arbitrarily large simply by picking large coefficients. The normalizing constraint

$$\mathbf{a}'_1 \mathbf{a}_1 = 1 \quad (13.1)$$

proves convenient, but any constraint of this general form would do.¹³

We can maximize $S_{W_1}^2$ subject to the restriction of Equation 13.1 by employing a Lagrange multiplier L_1 , defining¹⁴

$$F_1 \equiv \mathbf{a}'_1 \mathbf{R}_{XX} \mathbf{a}_1 - L_1 (\mathbf{a}'_1 \mathbf{a}_1 - 1)$$

Then, differentiating this equation with respect to \mathbf{a}_1 and L_1 ,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX} \mathbf{a}_1 - 2L_1 \mathbf{a}_1 \\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}'_1 \mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces the equations

$$\begin{aligned}(\mathbf{R}_{XX} - L_1 \mathbf{I}_k) \mathbf{a}_1 &= \mathbf{0} \\ \mathbf{a}'_1 \mathbf{a}_1 &= 1\end{aligned} \quad (13.2)$$

The first line of Equations 13.2 has nontrivial solutions for \mathbf{a}_1 only when $(\mathbf{R}_{XX} - L_1 \mathbf{I}_k)$ is singular—that is, when $|\mathbf{R}_{XX} - L_1 \mathbf{I}_k| = 0$. The multiplier L_1 , therefore, is an eigenvalue of \mathbf{R}_{XX} , and \mathbf{a}_1 is the corresponding eigenvector, scaled so that $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

There are, however, k solutions to Equations 13.2, corresponding to the k eigenvalue-eigenvector pairs of \mathbf{R}_{XX} , so we must decide which solution to choose. From the first line of Equations 13.2, we have $\mathbf{R}_{XX} \mathbf{a}_1 = L_1 \mathbf{a}_1$. Consequently,

$$S_{W_1}^2 = \mathbf{a}'_1 \mathbf{R}_{XX} \mathbf{a}_1 = L_1 \mathbf{a}'_1 \mathbf{a}_1 = L_1$$

Because our purpose is to *maximize* $S_{W_1}^2$ (subject to the constraint on \mathbf{a}_1), we must select the *largest* eigenvalue of \mathbf{R}_{XX} to define the first principal component.

The second principal component is derived similarly, under the further restriction that it is orthogonal to the first; the third that it is orthogonal to the first two; and so on.¹⁵ It turns out that the second principal component corresponds to the second-largest eigenvalue of \mathbf{R}_{XX} , the third to the third-largest eigenvalue, and so on. We order the eigenvalues of \mathbf{R}_{XX} so that¹⁶

$$L_1 \geq L_2 \geq \cdots \geq L_k > 0$$

The matrix of principal-component coefficients

¹³Normalizing the coefficients so that $\mathbf{a}'_1 \mathbf{a}_1 = 1$ causes the variance of the first principal component to be equal to the combined variance of the standardized regressors accounted for by this component, as will become clear presently.

¹⁴See online Appendix C on calculus for an explanation of the method of Lagrange multipliers for constrained optimization.

¹⁵See Exercise 13.1.

¹⁶Recall that we are assuming that \mathbf{R}_{XX} is of full rank, and hence none of its eigenvalues is 0. It is possible, but unlikely, that two or more eigenvalues of \mathbf{R}_{XX} are equal. In this event, the orientation of the principal components corresponding to the equal eigenvalues is not unique, although the subspace spanned by these components—and for which they constitute a basis—is unique.

$$\underset{(k \times k)}{\mathbf{A}} \equiv [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k]$$

contains normalized eigenvectors of \mathbf{R}_{XX} . This matrix is, therefore, orthonormal: $\mathbf{A}'\mathbf{A} = \mathbf{AA}' = \mathbf{I}_k$.

The principal components

$$\underset{(n \times k)}{\mathbf{W}} = \underset{(n \times k)}{\mathbf{Z}_X} \underset{(k \times k)}{\mathbf{A}} \quad (13.3)$$

have covariance matrix

$$\begin{aligned} \frac{1}{n-1} \mathbf{W}'\mathbf{W} &= \frac{1}{n-1} \mathbf{A}'\mathbf{Z}_X'\mathbf{Z}_X\mathbf{A} \\ &= \mathbf{A}'\mathbf{R}_{XX}\mathbf{A} = \mathbf{A}'\mathbf{AL} = \mathbf{L} \end{aligned}$$

where $\mathbf{L} \equiv \text{diag}[L_1, L_2, \dots, L_k]$ is the diagonal matrix of eigenvalues of \mathbf{R}_{XX} ; the covariance matrix of the principal components is, therefore, orthogonal, as required. Furthermore,

$$\text{trace}(\mathbf{L}) = \sum_{j=1}^k L_j = k = \text{trace}(\mathbf{R}_{XX})$$

and thus the principal components partition the combined variance of the standardized variables Z_1, Z_2, \dots, Z_k .

Solving Equation 13.3 for \mathbf{Z}_X produces

$$\mathbf{Z}_X = \mathbf{WA}^{-1} = \mathbf{WA}'$$

and, consequently,

$$\mathbf{R}_{XX} = \frac{1}{n-1} \mathbf{Z}_X'\mathbf{Z}_X = \frac{1}{n-1} \mathbf{A}\mathbf{W}'\mathbf{W}\mathbf{A}' = \mathbf{ALA}'$$

Finally,

$$\mathbf{R}_{XX}^{-1} = (\mathbf{A}')^{-1} \mathbf{L}^{-1} \mathbf{A}^{-1} = \mathbf{AL}^{-1}\mathbf{A}' \quad (13.4)$$

We will use this result presently in our investigation of collinearity.

Two Variables

The vector geometry of principal components is illustrated for two variables in Figure 13.5. The symmetry of this figure is peculiar to the two-dimensional case. The length of each principal-component vector is the square root of the sum of squared orthogonal projections of \mathbf{z}_1 and \mathbf{z}_2 on the component. The direction of \mathbf{w}_1 is chosen to maximize the combined length of these projections and hence to maximize the length of \mathbf{w}_1 . Because the subspace spanned by \mathbf{z}_1 and \mathbf{z}_2 is two-dimensional, \mathbf{w}_2 is simply chosen to be orthogonal to \mathbf{w}_1 . Note that $\|\mathbf{w}_j\|^2 = L_j(n-1)$.¹⁷

¹⁷There is a small subtlety here: The subspace spanned by each component is one-dimensional, and the length of each component is fixed by the corresponding eigenvalue, but these factors determine the orientation of the component only up to a rotation of 180° —that is, a change in sign.

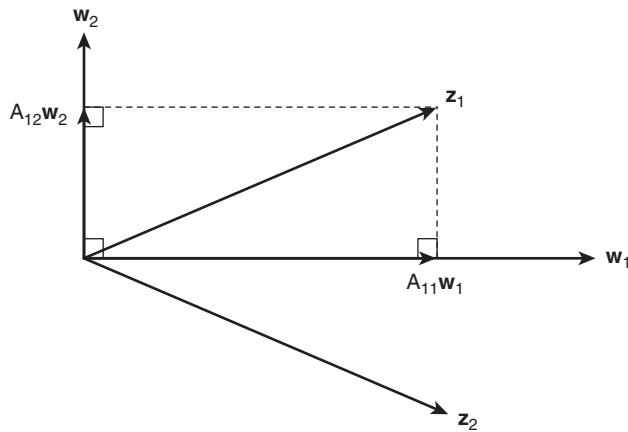


Figure 13.5 Vector geometry of principal components for two positively correlated standardized variables \mathbf{z}_1 and \mathbf{z}_2 .

It is clear from the figure that as the correlation between Z_1 and Z_2 increases, the first principal component grows at the expense of the second; thus, L_1 gets larger and L_2 smaller. If, alternatively, \mathbf{z}_1 and \mathbf{z}_2 are orthogonal, then $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = \sqrt{n-1}$ and $L_1 = L_2 = 1$.

The algebra of the two-variable case is also quite simple. The eigenvalues of \mathbf{R}_{XX} are the solutions of the characteristic equation

$$\begin{vmatrix} 1-L & r_{12} \\ r_{12} & 1-L \end{vmatrix} = 0$$

that is,

$$(1-L)^2 - r_{12}^2 = L^2 - 2L + 1 - r_{12}^2 = 0$$

Using the quadratic formula to find the roots of the characteristic equation yields

$$\begin{aligned} L_1 &= 1 + \sqrt{r_{12}^2} \\ L_2 &= 1 - \sqrt{r_{12}^2} \end{aligned} \tag{13.5}$$

And so, consistent with the geometry of Figure 13.5, as the magnitude of the correlation between the two variables increases, the variation attributed to the first principal component also grows. If r_{12} is positive, then solving for \mathbf{A} from the relation $\mathbf{R}_{XX}\mathbf{A} = \mathbf{LA}$ under the restriction $\mathbf{A}'\mathbf{A} = \mathbf{I}_2$ gives¹⁸

$$\mathbf{A} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \tag{13.6}$$

¹⁸Exercise 13.2 derives the solution for $r_{12} < 0$.

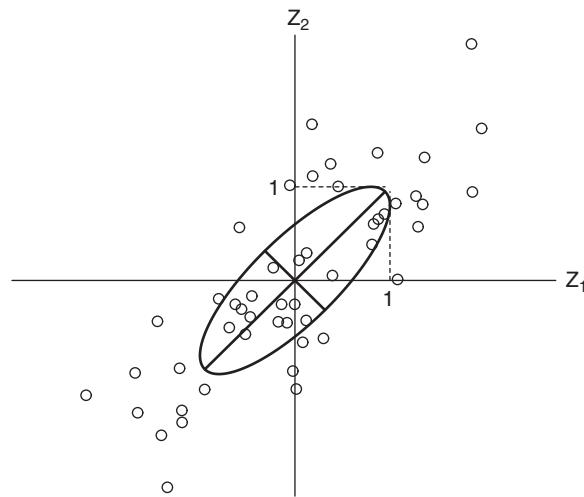


Figure 13.6 The principal components for two standardized variables Z_1 and Z_2 are the principal axes of the standard data ellipse $\mathbf{z}'\mathbf{R}_{XX}^{-1}\mathbf{z} = 1$. The first eigenvalue L_1 of \mathbf{R}_{XX} gives the half-length of the major axis of the ellipse; the second eigenvalue L_2 gives the half-length of the minor axis. In this illustration, the two variables are correlated $r_{12} = .8$, so L_1 is large and L_2 is small.

The generalization to k standardized regressors is straightforward: If the variables are orthogonal, then all $L_j = 1$ and all $\|\mathbf{w}_j\| = \sqrt{n - 1}$. As collinearities among the variables increase, some eigenvalues become large while others grow small. Small eigenvalues and the corresponding short principal components represent dimensions along which the regressor subspace has (nearly) collapsed. Perfect collinearities are associated with eigenvalues of 0.

The Data Ellipsoid

The principal components have an interesting interpretation in terms of the standard data ellipsoid for the Z s.¹⁹ The data ellipsoid is given by the equation

$$\mathbf{z}'\mathbf{R}_{XX}^{-1}\mathbf{z} = 1$$

where $\mathbf{z} \equiv (Z_1, \dots, Z_k)'$ is a vector of values for the k standardized regressors. Because the variables are standardized, the data ellipsoid is centered at the origin, and the shadow of the ellipsoid on each axis is of length 2 (i.e., 2 standard deviations). It can be shown that the principal components correspond to the principal axes of the data ellipsoid, and, furthermore, that the half-length of each axis is equal to the square root of the corresponding eigenvalue L_j of \mathbf{R}_{XX} .²⁰ These properties are depicted in Figure 13.6 for $k = 2$. When the variables are uncorrelated, the data ellipse becomes circular, and each axis has a half-length of 1.

¹⁹The standard data ellipsoid was introduced in Section 9.4.4.

²⁰See Exercise 13.3. These relations also hold for *unstandardized* variables. That is, the principal components calculated from the covariance matrix \mathbf{S}_{XX} give the principal axes of the standard data ellipsoid $(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{S}_{XX}^{-1})(\mathbf{x} - \bar{\mathbf{x}})$, and the half-length of the j th principal axis of this ellipsoid is equal to the square root of the j th eigenvalue of \mathbf{S}_{XX} .

Summary

- The principal components of the k standardized regressors \mathbf{Z}_X are a new set of k variables derived from \mathbf{Z}_X by a linear transformation: $\mathbf{W} = \mathbf{Z}_X \mathbf{A}$, where \mathbf{A} is the $(k \times k)$ transformation matrix.
- The transformation \mathbf{A} is selected so that the columns of \mathbf{W} are orthogonal—that is, the principal components are uncorrelated. In addition, \mathbf{A} is constructed so that the first component accounts for maximum variance in the Z s, the second for maximum variance under the constraint that it is orthogonal to the first, and so on. Each principal component is scaled so that its variance is equal to the variance in the Z s for which it accounts. The principal components therefore partition the variance of the Z s.
- The transformation matrix \mathbf{A} contains (by columns) normalized eigenvectors of \mathbf{R}_{XX} , the correlation matrix of the regressors. The columns of \mathbf{A} are ordered by their corresponding eigenvalues: The first column corresponds to the largest eigenvalue and the last column to the smallest. The eigenvalue L_j associated with the j th component represents the collective variation in the Z s attributable to that component.
- If there are perfect collinearities in \mathbf{Z}_X , then some eigenvalues of \mathbf{R}_{XX} will be 0, and there will be fewer than k principal components, the number of components corresponding to $\text{rank}(\mathbf{Z}_X) = \text{rank}(\mathbf{R}_{XX})$. Near collinearities are associated with small eigenvalues and correspondingly short principal components.

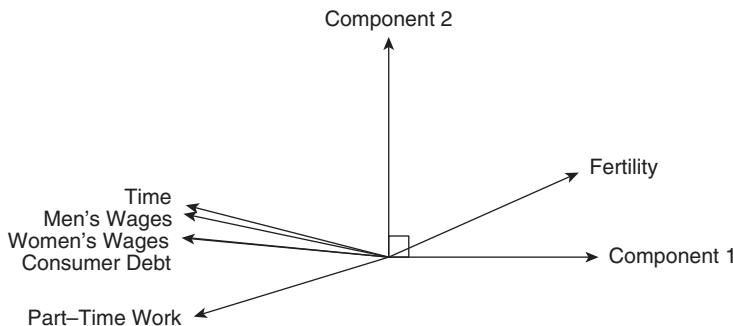
Principal components can be used to explicate the correlational structure of the explanatory variables in regression. The principal components are a derived set of variables that form an orthogonal basis for the subspace of the standardized X s. The first principal component spans the one-dimensional subspace that accounts for maximum variation in the standardized X s. The second principal component accounts for maximum variation in the standardized X s, under the constraint that it is orthogonal to the first. The other principal components are similarly defined; unless the X s are perfectly collinear, there are as many principal components as there are X s. Each principal component is scaled to have variance equal to the collective variance in the standardized X s for which it accounts. Collinear relations among the explanatory variables, therefore, correspond to very short principal components, which represent dimensions along which the regressor subspace has nearly collapsed.

A principal-components analysis for the explanatory variables in B. Fox's Canadian women's labor force regression is summarized in Table 13.3, which shows the coefficients of the principal components (i.e., the elements of \mathbf{A}), along with the eigenvalues of the correlation matrix of the explanatory variables and the cumulative percentage of variation in the X s accounted for by the principal components. The first two principal components account for almost 98% of the variation in the six explanatory variables.

The principal-components analysis is graphed in Figure 13.7. Here, the variables—including the two principal components—are standardized to common length, and the variables are

Table 13.3 Principal-Components Coefficients for the Explanatory Variables in B. Fox's Regression

Variable	Principal Component					
	W_1	W_2	W_3	W_4	W_5	W_6
Fertility	0.3849	0.6676	0.5424	0.2518	-0.1966	-0.0993
Men's Wages	-0.4159	0.3421	-0.0223	0.1571	0.7055	-0.4326
Women's Wages	-0.4196	0.1523	-0.2658	0.7292	-0.2791	0.3472
Consumer Debt	-0.4220	0.1591	-0.0975	-0.2757	-0.6188	-0.5728
Part-Time Work	-0.3946	-0.4693	0.7746	0.1520	-0.0252	-0.0175
Time	-0.4112	0.4106	0.1583	-0.5301	0.0465	0.5951
Eigenvalue	5.5310	0.3288	0.1101	0.0185	0.0071	0.0045
Cumulative percentage	92.18	97.66	99.50	99.81	99.93	100.00

**Figure 13.7** Orthogonal projections of the six explanatory variables onto the subspace spanned by the first two principal components. All the variables, including the components, are standardized to common length. The projections of the vectors for women's wages and consumer debt are nearly coincident (the two vectors are essentially on top of one another).

projected orthogonally onto the subspace spanned by the first two principal components. Because the first two components account for almost all the variation in the variables, the vectors for the variables lie very close to this subspace. Consequently, the projections of the vectors are almost as long as the vectors themselves, and the cosines of the angles between the projected vectors closely approximate the correlations between the variables. The projected vectors for women's wages and consumer debt are nearly coincident, reflecting the near-perfect correlation between the two variables.

It is clear that the explanatory variables divide into two subsets: time, men's wages, women's wages, and consumer debt (which are all highly positively correlated) in one subset, as well as fertility and part-time work (which are strongly negatively correlated) in the other. Correlations *between* the two sets of variables are also quite high. In effect, the subspace of the explanatory variables has collapsed into two dimensions.

Diagnosing Collinearity

I explained earlier that the sampling variance of the regression coefficient B_j is

$$V(B_j) = \frac{\sigma_e^2}{(n-1)S_j^2} \times \frac{1}{1-R_j^2}$$

It can be shown that $VIF_j = 1/(1-R_j^2)$ is the j th diagonal entry of \mathbf{R}_{XX}^{-1} (see Theil, 1971, p. 166). Using Equation 13.4, the variance inflation factors can be expressed as functions of the eigenvalues of \mathbf{R}_{XX} and the principal components; specifically,

$$VIF_j = \sum_{l=1}^k \frac{A_{jl}^2}{L_l}$$

Thus, it is the small eigenvalues that contribute to large sampling variance, but only for those regressors that have large coefficients associated with the corresponding short principal components. This result is sensible, for small eigenvalues, and their short components correspond to collinear relations among the regressors; regressors with large coefficients for these components are the regressors implicated in the collinearities (see below).

The relative size of the eigenvalues serves as an indicator of the degree of collinearity present in the data. The square root of the ratio of the largest to smallest eigenvalue, $K \equiv \sqrt{L_1/L_k}$, called the *condition number*, is a commonly employed standardized index of the global instability of the least-squares regression coefficients: A large condition number (say, 10 or more) indicates that relatively small changes in the data tend to produce large changes in the least-squares solution. In this event, \mathbf{R}_{XX} is said to be *ill conditioned*.

It is instructive to examine the condition number in the simplified context of the two-regressor model. From Equations 13.5 (page 352),

$$K = \sqrt{\frac{L_1}{L_2}} = \sqrt{\frac{1 + \sqrt{r_{12}^2}}{1 - \sqrt{r_{12}^2}}}$$

and thus $K = 10$ corresponds to $r_{12}^2 = .9608$, for which $VIF = 26$ (and $\sqrt{VIF} \approx 5$).

Belsley, Kuh, and Welsh (1980, chap. 3) define a *condition index* $K_j \equiv \sqrt{L_1/L_j}$ for each principal component of \mathbf{R}_{XX} .²¹ Then, the number of large condition indices points to the number of different collinear relations among the regressors.

²¹Primarily for computational accuracy, Belsley et al. (1980, chap. 3) develop diagnostic methods for collinearity in terms of the *singular-value decomposition* of the regressor matrix, scaled so that each variable has a sum of squares of 1. I employ an equivalent eigenvalue-eigenvector approach because of its conceptual simplicity and broader familiarity. The eigenvectors of \mathbf{R}_{XX} , it turns out, are the squares of the singular values of $(1/\sqrt{n-1})\mathbf{Z}_X$. Indeed, the condition number K defined here is actually the condition number of $(1/\sqrt{n-1})\mathbf{Z}_X$ (and hence of \mathbf{Z}_X). Information on the singular-value decomposition and its role in linear-model analysis can be found in Belsley et al. (1980, chap. 3) and in Mandel (1982).

A more substantial difference between my approach and that of Belsley et al. is that they base their analysis not on the correlation matrix of the X s but rather on $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}$ is the regressor matrix, including the constant regressor, with columns normed to unit length. Consider an explanatory variable that is uncorrelated with the others but that has scores that are far from 0. Belsley et al. would say that this explanatory variable is “collinear with the constant regressor.” This seems to me a corruption of the notion of collinearity, which deals fundamentally with the inability to separate the effects of highly correlated explanatory variables and should not change with linear transformations of individual explanatory variables. See Belsley (1984) and the associated commentary for various points of view on this issue.

The condition indices for B. Fox's Canadian women's labor force regression, reported in Table 13.2 (on page 349), are as follows:

K_1	K_2	K_3	K_4	K_5	K_6
1.00	4.10	7.09	17.27	27.99	35.11

The last three condition indices are therefore very large, suggesting an unstable regression.

Chatterjee and Price (1991, chap. 7) employ the principal-component coefficients to estimate near collinearities: A component \mathbf{w}_l associated with a very small eigenvalue $L_l \approx 0$ is itself approximately equal to the zero vector; consequently,

$$A_{1l}\mathbf{z}_1 + A_{2l}\mathbf{z}_2 + \cdots + A_{kl}\mathbf{z}_k \approx \mathbf{0}$$

and we can use the large A_{jl} s to specify a linear combination of the Z s that is approximately equal to 0.²²

13.1.2 Generalized Variance Inflation*

The methods for detecting collinearity described thus far are not fully applicable to models that include related sets of regressors, such as dummy regressors constructed from a polytomous categorical variable or polynomial regressors. The reasoning underlying this qualification is subtle but can be illuminated by appealing to the vector representation of linear models.²³

The correlations among a set of dummy regressors are affected by the choice of baseline category. Similarly, the correlations among a set of polynomial regressors in an explanatory variable X are affected by adding a constant to the X -values. Neither of these changes alters the fit of the model to the data, however, so neither is fundamental. It is, indeed, always possible to select an orthogonal basis for the dummy-regressor or polynomial-regressor subspace (although such a basis does not employ dummy variables or simple powers of X). What is fundamental is the subspace itself and not the arbitrarily chosen basis for it.²⁴

We are not concerned, therefore, with the “artificial” collinearity among dummy regressors or polynomial regressors in the same set. We are instead interested in the relationships between the subspaces generated to represent the effects of *different* explanatory variables. As a consequence, we can legitimately employ variance-inflation factors to examine the impact of collinearity on the coefficients of numerical regressors, or on any single-degree-of-freedom effects, even when sets of dummy regressors or polynomial regressors are present in the model.

Fox and Monette (1992) generalize the notion of variance inflation to sets of related regressors. Rewrite the linear model as

²²See Exercise 13.4 for an application to B. Fox's regression.

²³The vector geometry of linear models is developed in Chapter 10.

²⁴A specific basis may be a poor computational choice, however, if it produces numerically unstable results. Consequently, researchers are sometimes advised to pick a category with many cases to serve as the baseline for a set of dummy regressors or to subtract the mean from X prior to constructing polynomial regressors; the latter procedure is called *centering*. Neither of these practices fundamentally alters the model but may lead to more accurate calculations.

$$\mathbf{y}_{(n \times 1)} = \alpha \mathbf{1}_{(n \times 1)} + \mathbf{X}_1_{(n \times p)} \boldsymbol{\beta}_1_{(p \times 1)} + \mathbf{X}_2_{(n \times k-p)} \boldsymbol{\beta}_2_{(k-p \times 1)} + \boldsymbol{\varepsilon}_{(n \times 1)}$$

where the p regressors of interest (e.g., a set of dummy regressors) are in \mathbf{X}_1 , while the remaining $k - p$ regressors (with the exception of the constant) are in \mathbf{X}_2 . Fox and Monette (1992) show that the squared ratio of the size (i.e., length when $p = 1$, area when $p = 2$, volume when $p = 3$, or hyper-volume when $p > 3$) of the joint confidence region for $\boldsymbol{\beta}_1$ to the size of the same region for orthogonal but otherwise similar data is

$$\text{GVIF}_1 = \frac{\det \mathbf{R}_{11} \det \mathbf{R}_{22}}{\det \mathbf{R}} \quad (13.7)$$

Here, \mathbf{R}_{11} is the correlation matrix for \mathbf{X}_1 , \mathbf{R}_{22} is the correlation matrix for \mathbf{X}_2 , and \mathbf{R} is the matrix of correlations among all the variables.²⁵ The *generalized variance-inflation factor* (GVIF) is independent of the bases selected for the subspaces spanned by the columns of each of \mathbf{X}_1 and \mathbf{X}_2 . If \mathbf{X}_1 contains only one column, then the GVIF reduces to the familiar variance-inflation factor. To make generalized variance-inflation factors comparable across dimensions, Fox and Monette suggest reporting $\text{GVIF}^{p/2}$ —analogous to reporting $\sqrt{\text{VIF}}$ for a single coefficient.

The notion of variance inflation can be extended to sets of related regressors, such as dummy regressors and polynomial regressors, by considering the size of the joint confidence region for the related coefficients.

13.2 Coping With Collinearity: No Quick Fix

Consider the regression of a response variable Y on two explanatory variables X_1 and X_2 : When X_1 and X_2 are strongly collinear, the data contain little information about the impact of X_1 on Y holding X_2 constant statistically because there is little variation in X_1 when X_2 is fixed.²⁶ (Of course, the same is true for X_2 fixing X_1 .) Because B_1 estimates the partial effect of X_1 controlling for X_2 , this estimate is imprecise.

Although there are several strategies for dealing with collinear data, none magically extracts nonexistent information from the data. Rather, the research problem is redefined, often subtly and implicitly. Sometimes the redefinition is reasonable; usually it is not. The ideal solution to the problem of collinearity is to collect new data in such a manner that the problem is avoided—for example, by experimental manipulation of the X s, or through a research setting (or sampling procedure) in which the explanatory variables of interest are not strongly related. Unfortunately, these solutions are rarely practical. Several less adequate strategies for coping with collinear data are briefly described in this section.

²⁵An interesting observation is that Equation 13.7 can alternatively be applied to the correlation matrix of the estimated regression coefficients, rather than to the correlation matrix of the X s, yielding exactly the same result: See Exercise 13.5. This observation provides a basis for generalizing variance inflation beyond linear models: See, e.g., Section 15.4.3. I am grateful to Henric Nilsson for pointing this out to me.

²⁶This observation again invokes the added-variable plot; see, e.g., Figure 11.9 (page 284).

13.2.1 Model Respecification

Although collinearity is a data problem, not (necessarily) a deficiency of the model, one approach to the problem is to respecify the model. Perhaps, after further thought, several regressors in the model can be conceptualized as alternative indicators of the same underlying construct. Then these measures can be combined in some manner, or one can be chosen to represent the others. In this context, high correlations among the X s in question indicate high reliability—a fact to be celebrated, not lamented. Imagine, for example, an international analysis of factors influencing infant mortality, in which gross national product per capita, energy use per capita, and hours of Internet use per capita are among the explanatory variables and are highly correlated. A researcher may choose to treat these variables as indicators of the general level of economic development.

Alternatively, we can reconsider whether we really need to control for X_2 (for example) in examining the relationship of Y to X_1 . Generally, though, respecification of this variety is possible only where the original model was poorly thought out or where the researcher is willing to abandon some of the goals of the research. For example, suppose that in a time-series regression examining determinants of married women's labor force participation, collinearity makes it impossible to separate the effects of men's and women's wage levels. There may be good theoretical reason to want to know the effect of women's wage level on their labor force participation, holding men's wage level constant, but the data are simply uninformative about this question. It may still be of interest, however, to determine the partial relationship between *general* wage level and women's labor force participation, controlling for other explanatory variables in the analysis.²⁷

13.2.2 Variable Selection

A common, but usually misguided, approach to collinearity is variable selection, where some automatic procedure is employed to reduce the regressors in the model to a less highly correlated set.²⁸ *Forward-selection* methods add explanatory variables to the model one at a time. At each step, the variable that yields the largest increment in R^2 is selected. The procedure stops, for example, when the increment is smaller than a preset criterion.²⁹ *Backward-elimination* methods are similar, except that the procedure starts with the full model and deletes variables one at a time. *Forward/backward*—or *stepwise*—methods combine the two approaches, allowing variables to enter or leave at each step. Often the term *stepwise regression* is used for all these variations.

These methods frequently are abused by naive researchers who seek to interpret the order of entry of variables into the regression equation as an index of their “importance.” This practice is potentially misleading: For example, suppose that there are two highly correlated explanatory variables that have nearly identical large correlations with Y ; only one of these explanatory

²⁷In the example developed in this chapter, however, men's and women's wages are not only highly correlated with each other but with other variables (such as time) as well.

²⁸Variable selection methods are discussed in a more general context and in greater detail in Chapter 22.

²⁹Often, the stopping criterion is calibrated by the incremental F for adding a variable to the model or by using an index of model quality, such as those discussed in Chapter 22.

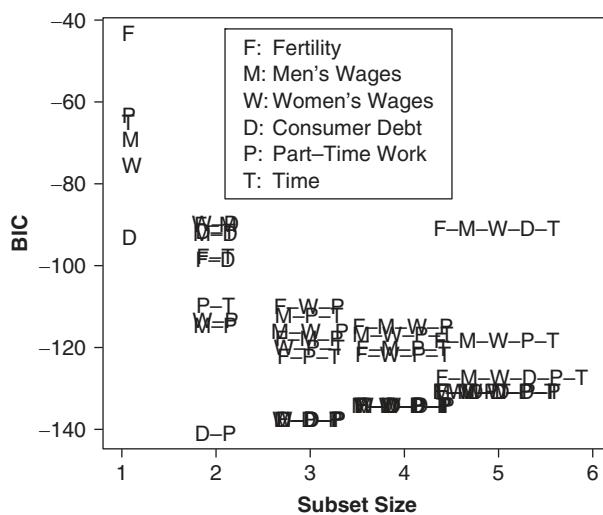


Figure 13.8 Variable selection for B. Fox's Canadian women's labor force regression. Up to the 10 "best" subsets of each size are shown, using the BIC model selection criterion.

variables will enter the regression equation because the other can contribute little additional information. A small modification to the data, or a new sample, could easily reverse the result.

A technical objection to stepwise methods is that they can fail to turn up the optimal subset of regressors of a given size (i.e., the subset that maximizes R^2). Advances in computer power and in computing procedures make it feasible to examine all subsets of regressors even when k is quite large.³⁰ Aside from optimizing the selection criterion, subset techniques also have the advantage of revealing alternative, nearly equivalent models and thus avoid the misleading appearance of producing a uniquely "correct" result.³¹

Figure 13.8 shows the result of applying an all-subset method of variable selection to the Canadian women's labor force regression. For each subset size of $p = 1$ to 6 explanatory variables, up to 10 "best" models are displayed.³² The criterion of model quality employed in this graph is the *Bayesian information criterion* (or *BIC*): Smaller values indicate a better-fitting model. Unlike the R^2 , which never declines when an additional variable is added to the model, the BIC "penalizes" the fit for the number of parameters in the model and therefore can prefer a smaller model to a larger one.³³ According to the BIC, the best model includes the two explanatory variables consumer debt and part-time work.³⁴ There are several models with three

³⁰For k explanatory variables, the number of subsets, excluding the null subset with no predictors, is $2^k - 1$. See Exercise 13.6.

³¹There are algorithms available to find the optimal subset of a given size without examining all possible subsets (see, e.g., Furnival & Wilson, 1974). When the data are highly collinear, however, the optimal subset of a given size may be only trivially "better" than many of its competitors.

³²The numbers of distinct subsets of one to six explanatory variables are 6, 15, 20, 15, 6, and 1, consecutively.

³³See Chapter 22 for a discussion of the BIC and other model selection criteria.

³⁴*Not surprisingly, this best subset of size 2 includes one variable from each of the two highly correlated subsets that were identified in the principal-components analysis of the preceding section; these are also the two explanatory variables whose coefficients are statistically significant in the initial regression.

explanatory variables that are slightly worse than the best model of size 2 but that are essentially indistinguishable from each other; the same is true for models with four and five explanatory variables.³⁵

In applying variable selection, it is essential to keep the following caveats in mind:

- Most important, variable selection results in a respecified model that usually does not address the research questions that were originally posed. In particular, if the original model is correctly specified, and if the included and omitted variables are correlated, then coefficient estimates following variable selection are biased.³⁶ Consequently, these methods are most useful for pure prediction problems, in which the values of the regressors for the data to be predicted will be within the configuration of X -values for which selection was employed. In this case, it is possible to get good estimates of $E(Y)$ even though the regression coefficients themselves are biased. If, however, the X -values for a new observation differ greatly from those used to obtain the estimates, then the predicted Y can be badly biased.
- When regressors occur in sets (e.g., of dummy variables), then these sets should generally be kept together during selection. Likewise, when there are hierarchical relations among regressors, these relations should be respected: For example, an interaction regressor should not appear in a model that does not contain the main effects marginal to the interaction.
- Because variable selection optimizes the fit of the model to the sample data, coefficient standard errors calculated following explanatory-variable selection—and hence confidence intervals and hypothesis tests—almost surely overstate the precision of results. There is, therefore, a very considerable risk of capitalizing on chance characteristics of the sample.³⁷
- As discussed in Chapter 22, variable selection has applications to statistical modeling even when collinearity is not an issue. For example, it is generally not problematic to eliminate regressors that have small, precisely estimated coefficients, thus producing a more parsimonious model. Indeed, in a very large sample, we may feel justified in deleting regressors with trivially small but “statistically significant” coefficients.

13.2.3 Biased Estimation

Still another general approach to collinear data is biased estimation. The essential idea here is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance. The hoped-for result is a smaller mean-squared error of estimation of the β s than is provided by the least-squares estimates. By far the most common biased estimation method is *ridge regression* (due to Hoerl & Kennard, 1970a, 1970b).

³⁵Raftery (1995) suggests that a difference in BIC less than 2 provides “weak” evidence for the superiority of one model relative to another; similarly, a difference between 2 and 6 provides “positive” evidence, between 6 and 10 “strong” evidence, and greater than 10 “very strong” evidence for the relative superiority of a model. In the current application, the difference in BIC between the best-fitting models of sizes 2 and 3 is 3.3. Again, see Chapter 22 for a more extensive discussion of the BIC.

³⁶See Sections 6.3, 9.7, and 13.2.5.

³⁷This issue is pursued in Chapter 22 on model selection.

Like variable selection, biased estimation is not a magical panacea for collinearity. Ridge regression involves the arbitrary selection of a “ridge constant”, which controls the extent to which ridge estimates differ from the least-squares estimates: The larger the ridge constant, the greater the bias and the smaller the variance of the ridge estimator. Unfortunately, but as one might expect, to pick an optimal ridge constant—or even a good one—generally requires knowledge about the unknown β s that we are trying to estimate. My principal reason for mentioning biased estimation here is to caution against its routine use.

Ridge Regression*

The ridge-regression estimator for the *standardized* regression coefficients is given by

$$\mathbf{b}_d^* \equiv (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1} \mathbf{r}_{Xy} \quad (13.8)$$

where \mathbf{R}_{XX} is the correlation matrix for the explanatory variables, \mathbf{r}_{Xy} is the vector of correlations between the explanatory variables and the response, and $d \geq 0$ is a scalar constant. When $d = 0$, the ridge and least-squares estimators coincide: $\mathbf{b}_0^* = \mathbf{b}^* = \mathbf{R}_{XX}^{-1} \mathbf{r}_{Xy}$. When the data are collinear, some off-diagonal entries of \mathbf{R}_{XX} are generally large, making this matrix ill conditioned. Heuristically, the ridge-regression method improves the conditioning of \mathbf{R}_{XX} by inflating its diagonal entries.

Although the least-squares estimator \mathbf{b}^* is unbiased, its entries tend to be too large in absolute value, a tendency that is magnified as collinearity increases. In practice, researchers working with collinear data often compute wildly large regression coefficients. The ridge estimator may be thought of as a “shrunken” version of the least-squares estimator, correcting the tendency of the latter to produce coefficients that are too far from 0.

The ridge estimator of Equation 13.8 can be rewritten as³⁸

$$\mathbf{b}_d^* = \mathbf{U} \mathbf{b}^* \quad (13.9)$$

where $\mathbf{U} \equiv (\mathbf{I}_k + d\mathbf{R}_{XX}^{-1})^{-1}$. As d increases, the entries of \mathbf{U} tend to grow smaller, and, therefore, \mathbf{b}_d^* is driven toward $\mathbf{0}$. Hoerl and Kennard (1970a) show that for any value of $d > 0$, the squared length of the ridge estimator is less than that of the least-squares estimator: $\mathbf{b}_d^{*'} \mathbf{b}_d^* < \mathbf{b}^{*'} \mathbf{b}^*$.

The expected value of the ridge estimator can be determined from its relation to the least-squares estimator, given in Equation 13.9; treating the X -values, and hence \mathbf{R}_{XX} and \mathbf{U} , as fixed,

$$E(\mathbf{b}_d^*) = \mathbf{U} E(\mathbf{b}^*) = \mathbf{U} \boldsymbol{\beta}^*$$

The bias of \mathbf{b}_d^* is, therefore,

$$\text{bias}(\mathbf{b}_d^*) \equiv E(\mathbf{b}_d^*) - \boldsymbol{\beta}^* = (\mathbf{U} - \mathbf{I}_k) \boldsymbol{\beta}^*$$

and because the departure of \mathbf{U} from \mathbf{I}_k increases with d , the bias of the ridge estimator is an increasing function of d .

The variance of the ridge estimator is also simply derived:³⁹

³⁸See Exercise 13.8.

³⁹See Exercise 13.9.

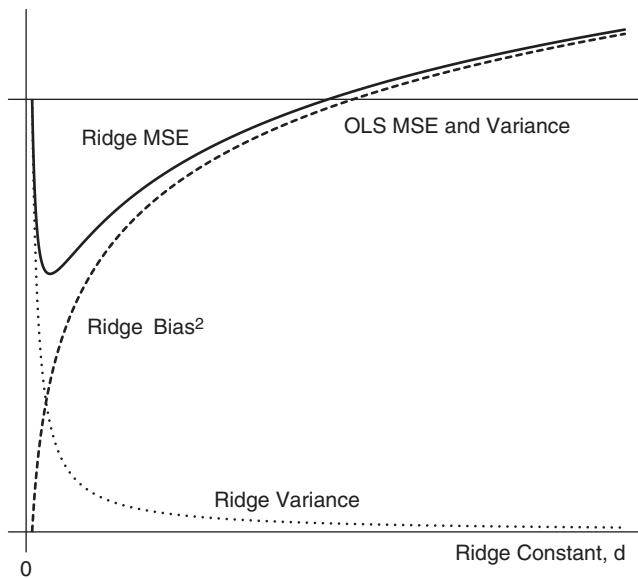


Figure 13.9 Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d . The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d , the MSE error of the ridge estimator is below the variance of the OLS estimator.

$$V(\mathbf{b}_d^*) = \frac{\sigma_\varepsilon^{*2}}{n-1} (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1} \mathbf{R}_{XX} (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1} \quad (13.10)$$

where σ_ε^{*2} is the error variance for the standardized regression. As d increases, the inverted term $(\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}$ is increasingly dominated by $d\mathbf{I}_k$. The sampling variance of the ridge estimator, therefore, is a decreasing function of d . This result is intuitively reasonable because the estimator itself is driven toward $\mathbf{0}$.

The mean-squared error of the ridge estimator is the sum of its squared bias and sampling variance. Hoerl and Kennard (1970a) prove that it is always possible to choose a positive value of the ridge constant d so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. These ideas are illustrated heuristically in Figure 13.9. As mentioned, however, the optimal value of d depends on the unknown population regression coefficients.

The central problem in applying ridge regression is to find a value of d for which the trade-off of bias against variance is favorable. In deriving the properties of the ridge estimator, I treated d as fixed. If d is determined from the data, however, it becomes a random variable, casting doubt on the conceptual basis for the ridge estimator. A number of methods have been proposed for selecting d . Some of these are rough and qualitative, while others incorporate

specific formulas or procedures for estimating the optimal value of d . All these methods, however, have only ad hoc justifications.⁴⁰

There have been many random-sampling simulation experiments exploring the properties of ridge estimation along with other methods meant to cope with collinear data. While these studies are by no means unanimous in their conclusions, the ridge estimator often performs well in comparison with least-squares estimation and in comparison with other biased-estimation methods. On the basis of evidence from simulation experiments, it would, however, be misleading to recommend a particular procedure for selecting the ridge constant d , and, indeed, the dependence of the optimal value of d on the unknown regression parameters makes it unlikely that there is a generally best way of finding d . Several authors critical of ridge regression (e.g., Draper & Smith, 1998, p. 395) have noted that simulations supporting the method generally incorporate restrictions on parameter values especially suited to ridge regression.⁴¹

Because the ridge estimator is biased, standard errors based on Equation 13.10 cannot be used in the normal manner for statistical inferences concerning the population regression coefficients. Indeed, as Obenchain (1977) has pointed out, under the assumptions of the linear model, confidence intervals centered at the least-squares estimates paradoxically retain their optimal properties regardless of the degree of collinearity: In particular, they are the *shortest* possible intervals at the stated level of confidence (Scheffé, 1959, chap. 2). An interval centered at the ridge estimate of a regression coefficient is, therefore, *wider* than the corresponding least-squares interval, even if the ridge estimator has smaller mean-squared error than the least-squares estimator.

13.2.4 Prior Information About the Regression Coefficients

A final approach to estimation with collinear data is to introduce additional prior information (i.e., relevant information external to the data at hand) that reduces the ambiguity produced by collinearity. There are several different ways in which prior information can be brought to bear on a regression, including Bayesian analysis,⁴² but I will present a very simple case to illustrate the general point. More complex methods are beyond the scope of this discussion and are, in any event, difficult to apply in practice.⁴³

Suppose that we wish to estimate the model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where Y is savings, X_1 is income from wages and salaries, X_2 is dividend income from stocks, and X_3 is interest income. Imagine that we have trouble estimating β_2 and β_3 because X_2 and X_3 are highly correlated in our data. Suppose further that we have reason to believe that $\beta_2 = \beta_3$, and denote the common quantity β_* . If X_2 and X_3 were *not* so highly correlated, then

⁴⁰Exercise 13.10 describes a qualitative method proposed by Hoerl and Kennard in their 1970 papers.

⁴¹See Section 13.2.5. Simulation studies of ridge regression and other biased estimation methods are too numerous to cite individually here. References to and comments on this literature can be found in many sources, including Draper and Van Nostrand (1979), Vinod (1978), and Hocking (1976). Vinod and Ullah (1981) present an extensive treatment of ridge regression and related methods.

⁴²Bayesian inference is introduced in online Appendix D on probability and estimation.

⁴³See, for example, Belsley et al. (1980, pp. 193–204) and Theil (1971, pp. 346–352).

we could reasonably test this belief as a hypothesis. In the current situation, we can fit the model

$$Y = \alpha + \beta_1 X_1 + \beta_* (X_2 + X_3) + \varepsilon$$

incorporating our belief in the equality of β_2 and β_3 in the specification of the model and thus eliminating the collinearity problem (along with the possibility of testing the belief that the two coefficients are equal).⁴⁴

13.2.5 Some Comparisons

Although I have presented them separately, the several approaches to collinear data have much in common:

- Model respecification can involve variable selection, and variable selection, in effect, respecifies the model.
- Variable selection implicitly constrains the coefficients of deleted regressors to 0.
- Variable selection produces biased coefficient estimates if the deleted variables have nonzero β s and are correlated with the included variables (as they will be for collinear data).⁴⁵ As in ridge regression and similar biased estimation methods, we might hope that the trade-off of bias against variance is favorable, and that, therefore, the mean-squared error of the regression estimates is smaller following variable selection than before. Because the bias and hence the mean-squared error depend on the unknown regression coefficients, however, we have no assurance that this will be the case. Even if the coefficients obtained following selection have smaller mean-squared error, their superiority can easily be due to the very large variance of the least-squares estimates when collinearity is high than to acceptably small bias. We should be especially careful about removing a variable that is causally prior to an explanatory variable of central interest.
- Certain types of prior information (as in the hypothetical example presented in the previous section) result in a respecified model.
- It can be demonstrated that biased-estimation methods like ridge regression place prior constraints on the values of the β s. Ridge regression imposes the restriction $\sum_{j=1}^k B_j^{*2} \leq c$, where c is a decreasing function of the ridge constant d ; the ridge estimator finds least-squares coefficients subject to this constraint (Draper & Smith, 1998, pp. 392–393). In effect, large absolute standardized coefficients are ruled out a priori, but the specific constraint is imposed indirectly through the choice of d .

The primary lesson to be drawn from these remarks is that mechanical model selection and modification procedures disguise the substantive implications of modeling decisions.

⁴⁴To test $H_0: \beta_2 = \beta_3$ simply entails contrasting the two models (see Exercise 6.7). In the present context, however, where X_2 and X_3 are very highly correlated, this test has virtually no power: If the second model is wrong, then we cannot, as a practical matter, detect it. We need either to accept the second model on theoretical grounds or to admit that we cannot estimate β_2 and β_3 .

⁴⁵Bias due to the omission of explanatory variables is discussed in a general context in Sections 6.3 and 9.7.

Consequently, these methods generally cannot compensate for weaknesses in the data and are no substitute for judgment and thought.

Several methods have been proposed for dealing with collinear data. Although these methods are sometimes useful, none can be recommended generally: When the X s are highly collinear, the data contain little information about the partial relationship between each X and Y , controlling for the other X s. To resolve the intrinsic ambiguity of collinear data, it is necessary either to introduce information external to the data or to redefine the research question asked of the data. Neither of these general approaches should be undertaken mechanically. Methods that are commonly (and, more often than not, unjustifiably) employed with collinear data include model respecification, variable selection (stepwise and subset methods), biased estimation (e.g., ridge regression), and the introduction of additional prior information. Comparison of the several methods shows that they have more in common than it appears at first sight.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 13.1. *The second principal component is

$$\begin{aligned}\mathbf{w}_2 &= A_{12}\mathbf{z}_1 + A_{22}\mathbf{z}_2 + \cdots + A_{k2}\mathbf{z}_k \\ &= \mathbf{Z}_X \mathbf{a}_2\end{aligned}$$

$(n \times 1)$ $(n \times k)$ $(k \times 1)$

with variance

$$S_{W_2}^2 = \mathbf{a}_2' \mathbf{R}_{XX} \mathbf{a}_2$$

We need to maximize this variance subject to the *normalizing constraint* $\mathbf{a}_2' \mathbf{a}_2 = 1$ and the *orthogonality constraint* $\mathbf{w}_1' \mathbf{w}_2 = 0$. Show that the orthogonality constraint is equivalent to $\mathbf{a}_1' \mathbf{a}_2 = 0$. Then, using two Lagrange multipliers, one for the normalizing constraint and the other for the orthogonality constraint, show that \mathbf{a}_2 is an eigenvector corresponding to the second-largest eigenvalue of \mathbf{R}_{XX} . Explain how this procedure can be extended to derive the remaining $k - 2$ principal components.

Exercise 13.2. *Find the matrix \mathbf{A} of principal-component coefficients when $k = 2$ and r_{12} is negative. (Cf. Equation 13.6 on page 352.)

Exercise 13.3. *Show that when $k = 2$, the principal components of \mathbf{R}_{XX} correspond to the principal axes of the data ellipse for the standardized regressors Z_1 and Z_2 ; show that the half-length of each axis is equal to the square root of the corresponding eigenvalue of \mathbf{R}_{XX} . Now extend this reasoning to the principal axes of the data ellipsoid for the standardized regressors when $k > 2$.

Exercise 13.4. *Use the principal-components analysis of the explanatory variables in B. Fox's time-series regression, given in Table 13.3, to estimate the nearly collinear relationships among the variables corresponding to small principal components. Which variables appear to be involved in each nearly collinear relationship?

Exercise 13.5. *Show that Equation 13.7 (page 358) applied to the correlation matrix of the least-squares regression coefficients, computed from the coefficient covariance matrix $S_E^2(\mathbf{X}'\mathbf{X})^{-1}$, produces the same generalized variance-inflation factor as when it is applied to the correlation matrix of the X s.

Exercise 13.6. Why are there $2^k - 1$ distinct subsets of k explanatory variables? Evaluate this quantity for $k = 2, 3, \dots, 15$.

Exercise 13.7. Apply the backward, forward, and forward/backward stepwise regression methods to B. Fox's Canadian women's labor force participation data. Compare the results of these procedures with those shown in Figure 13.8, based on the application of the BIC to all subsets of predictors.

Exercise 13.8. *Show that the ridge-regression estimator of the standardized regression coefficients,

$$\mathbf{b}_d^* = (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}\mathbf{r}_{Xy}$$

can be written as a linear transformation $\mathbf{b}_d^* = \mathbf{Ub}^*$ of the usual least-squares estimator $\mathbf{b}^* = \mathbf{R}_{XX}^{-1}\mathbf{r}_{Xy}$, where the transformation matrix is $\mathbf{U} \equiv (\mathbf{I}_k + d\mathbf{R}_{XX}^{-1})^{-1}$.

Exercise 13.9. *Show that the variance of the ridge estimator is

$$V(\mathbf{b}_d^*) = \frac{\sigma_\varepsilon^{*2}}{n-1} (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1} \mathbf{R}_{XX} (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}$$

[Hint: Express the ridge estimator as a linear transformation of the standardized response variable, $\mathbf{b}_d^* = (\mathbf{R}_{XX} + d\mathbf{I}_k)^{-1}[1/(n-1)]\mathbf{Z}'_X \mathbf{z}_y$.]

Exercise 13.10. *Finding the ridge constant d : Hoerl and Kennard suggest plotting the entries in \mathbf{b}_d^* against values of d ranging between 0 and 1. The resulting graph, called a *ridge trace*, both furnishes a visual representation of the instability due to collinearity and (ostensibly) provides a basis for selecting a value of d . When the data are collinear, we generally observe dramatic changes in regression coefficients as d is gradually increased from 0. As d is increased further, the coefficients eventually stabilize and then are driven slowly toward $\mathbf{0}$. The estimated error variance, S_E^{*2} , which is minimized at the least-squares solution ($d = 0$), rises slowly with increasing d . Hoerl and Kennard recommend choosing d so that the regression coefficients are stabilized and the error variance is not unreasonably inflated from its minimum value. (A number of other methods have been suggested for selecting d , but none avoids the fundamental difficulty of ridge regression—that good values of d depend on the unknown β s.) Construct a ridge trace, including the regression standard error S_E^* , for B. Fox's Canadian women's labor force participation data. Use this information to select a value of the ridge constant d , and compare the resulting ridge estimates of the regression parameters with the least-squares estimates. Make this comparison for both standardized and unstandardized coefficients. In applying ridge regression to these data, B. Fox selected $d = 0.05$.

Summary

- When the regressors in a linear model are perfectly collinear, the least-squares coefficients are not unique. Strong, but less-than-perfect, collinearity substantially increases the sampling variances of the least-squares coefficients and can render them useless as estimators.
- The sampling variance of the least-squares slope coefficient B_j is

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{(n - 1)S_j^2}$$

where R_j^2 is the squared multiple correlation for the regression of X_j on the other X s, and $S_j^2 = \sum (X_{ij} - \bar{X}_j)^2 / (n - 1)$ is the variance of X_j . The variance-inflation factor $VIF_j = 1/(1 - R_j^2)$ indicates the deleterious impact of collinearity on the precision of the estimate B_j . The notion of variance inflation can be extended to sets of related regressors, such as dummy regressors and polynomial regressors, by considering the size of the joint confidence region for the related coefficients.

- Principal components can be used to explicate the correlational structure of the explanatory variables in regression. The principal components are a derived set of variables that form an orthogonal basis for the subspace of the standardized X s. The first principal component spans the one-dimensional subspace that accounts for maximum variation in the standardized X s. The second principal component accounts for maximum variation in the standardized X s, under the constraint that it is orthogonal to the first. The other principal components are similarly defined; unless the X s are perfectly collinear, there are as many principal components as there are X s. Each principal component is scaled to have variance equal to the collective variance in the standardized X s for which it accounts. Collinear relations among the explanatory variables, therefore, correspond to very short principal components, which represent dimensions along which the regressor subspace has nearly collapsed.
- Several methods have been proposed for dealing with collinear data. Although these methods are sometimes useful, none can be recommended generally: When the X s are highly collinear, the data contain little information about the partial relationship between each X and Y , controlling for the other X s. To resolve the intrinsic ambiguity of collinear data, it is necessary either to introduce information external to the data or to redefine the research question asked of the data (or, as is usually impractical, to collect more informative data). Neither of these general approaches should be undertaken mechanically. Methods that are commonly (and, more often than not, unjustifiably) employed with collinear data include model respecification, variable selection (stepwise and subset methods), biased estimation (e.g., ridge regression), and the introduction of additional prior information. Comparison of the several methods shows that they have more in common than it appears at first sight.

PART IV

Generalized Linear Models

14

Logit and Probit Models for Categorical Response Variables

This chapter and the next deal with generalized linear models—the extension of linear models to variables that have specific non-normal conditional distributions:

- Rather than dive directly into generalized linear models in their full generality, the current chapter takes up linear logit and probit models for categorical response variables. Beginning with this most important special case allows for a gentler introduction to the topic, I believe. As well, I develop some models for categorical data that are not subsumed by the generalized linear model described in the next chapter.
- Chapter 15 is devoted to the generalized linear model, which has as special cases the linear models of Part II of the text and the dichotomous logit and probit models of the current chapter. Chapter 15 focuses on generalized linear models for count data and develops diagnostic methods for generalized linear models that parallel many of the diagnostics for linear models fit by least-squares, introduced in Part III.

All the statistical models described in previous chapters are for quantitative response variables. It is unnecessary to document the prevalence of qualitative/categorical data in the social sciences. In developing the general linear model, I introduced qualitative *explanatory* variables through the device of coding dummy-variable regressors.¹ There is no reason that qualitative variables should not also appear as response variables, affected by other variables, both qualitative and quantitative.

This chapter deals primarily with logit models for qualitative and ordered-categorical response variables, although related probit models are also briefly considered. The first section of the chapter describes logit and probit models for dichotomous response variables. The second section develops similar statistical models for polytomous response variables, including ordered categories. The third and final section discusses the application of logit models to contingency tables, where the explanatory variables, as well as the response, are categorical.

14.1 Models for Dichotomous Data

Logit and probit models express a qualitative response variable as a function of several explanatory variables, much in the manner of the general linear model. To understand why these

¹See Chapter 7.

models are required, let us begin by examining a representative problem, attempting to apply linear least-squares regression to it. The difficulties that are encountered point the way to more satisfactory statistical models for qualitative data.

In September 1988, 15 years after the coup of 1973, the people of Chile voted in a plebiscite to decide the future of the military government headed by General Augusto Pinochet. A *yes* vote would yield 8 more years of military rule; a *no* vote would set in motion a process to return the country to civilian government. As you are likely aware, the *no* side won the plebiscite, by a clear if not overwhelming margin.

Six months before the plebiscite, the independent research center FLACSO/Chile conducted a national survey of 2700 randomly selected Chilean voters.² Of these individuals, 868 said that they were planning to vote *yes*, and 889 said that they were planning to vote *no*. Of the remainder, 558 said that they were undecided, 187 said that they planned to abstain, and 168 did not answer the question. I will look here only at those who expressed a preference.³

Figure 14.1 plots voting intention against a measure of support for the status quo. As seems natural, voting intention appears as a dummy variable, coded 1 for *yes*, 0 for *no*. As we will see presently, this coding makes sense in the context of a dichotomous response variable. Because many points would otherwise be overplotted, voting intention is jittered in the graph (although not in the calculations that follow). Support for the status quo is a scale formed from a number of questions about political, social, and economic policies: High scores represent general support for the policies of the military regime. (For the moment, disregard the lines plotted in this figure.)

We are used to thinking of a regression as a conditional average. Does this interpretation make sense when the response variable is dichotomous? After all, an average between 0 and 1 represents a “score” for the dummy response variable that cannot be realized by any individual. In the population, the conditional average $E(Y|x_i)$ is simply the proportion of 1s among those individuals who share the value x_i for the explanatory variable—the conditional probability π_i of sampling a *yes* in this group;⁴ that is,

$$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$

and, thus,

$$E(Y|x_i) = \pi_i(1) + (1 - \pi_i)(0) = \pi_i \quad (14.1)$$

If X is discrete, then in a sample we can calculate the conditional proportion for Y at each value of X . The collection of these conditional proportions represents the sample nonparametric regression of the dichotomous Y on X . In the present example, X is continuous, but we

²FLACSO is an acronym for Facultad Latinoamericana de Ciencias Sociales, a respected institution that conducts social research and trains graduate students in several Latin American countries. During the Chilean military dictatorship, FLACSO/Chile was associated with the opposition to the military government. I worked on the analysis of the survey described here as part of a joint project between FLACSO in Santiago, Chile, and the Centre for Research on Latin America and the Caribbean at York University, Toronto.

³It is, of course, difficult to know how to interpret ambiguous responses such as “undecided.” It is tempting to infer that respondents were afraid to state their opinions, but there is other evidence from the survey that this is not the case. Few respondents, for example, uniformly refused to answer sensitive political questions, and the survey interviewers reported little resistance to the survey.

⁴Notice that π_i is a *probability*, not the mathematical constant $\pi \approx 3.14159$. A Greek letter is used because π_i can be estimated but not observed directly.

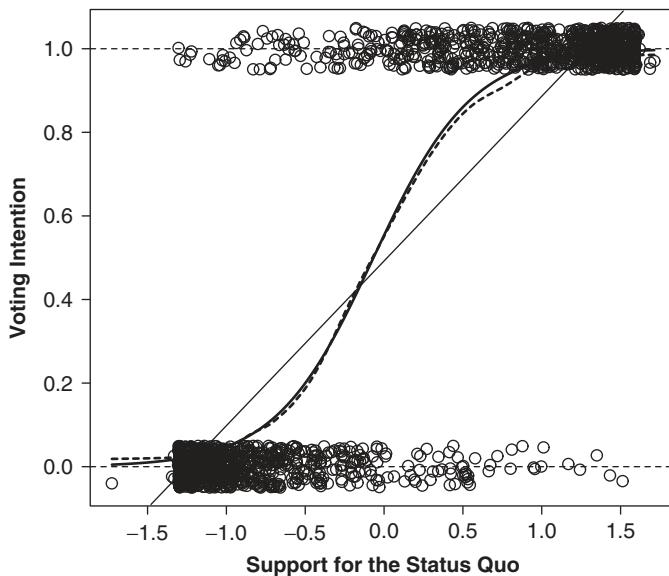


Figure 14.1 Scatterplot of voting intention (1 represents yes, 0 represents no) by a scale of support for the status quo, for a sample of Chilean voters surveyed prior to the 1988 plebiscite. The points are jittered vertically to minimize overplotting. The solid straight line shows the linear least-squares fit; the solid curved line shows the fit of the logistic-regression model (described in the next section); the broken line represents a nonparametric kernel regression with a span of 0.4.

can nevertheless resort to strategies such as local averaging, as illustrated in Figure 14.1.⁵ At low levels of support for the status quo, the conditional proportion of yes responses is close to 0; at high levels, it is close to 1; and in between, the nonparametric regression curve smoothly approaches 0 and 1 in a gentle, elongated S-shaped pattern.

14.1.1 The Linear-Probability Model

Although nonparametric regression works here, it would be useful to capture the dependency of Y on X as a simple function. To do so will be especially helpful when we introduce additional explanatory variables. As a first effort, let us try linear regression with the usual assumptions:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (14.2)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and ε_i and $\varepsilon_{i'}$ are independent for $i \neq i'$. If X is random, then we assume that it is independent of ε .

⁵The nonparametric-regression line in Figure 14.1 was fit by kernel regression—a method based on locally weighted averaging, which is similar to locally weighted regression (lowess, which was introduced in Chapter 2 for smoothing scatterplots). Unlike lowess, however, the kernel estimator of a proportion cannot be outside the interval from 0 to 1. Both the kernel-regression estimator and other nonparametric-regression methods that are more appropriate for a dichotomous response are described in Chapter 18. The span for the kernel regression (i.e., the fraction of the data included in each local average) is 0.4.

Under Equation 14.2, $E(Y_i) = \alpha + \beta X_i$, and so, from Equation 14.1,

$$\pi_i = \alpha + \beta X_i$$

For this reason, the linear-regression model applied to a dummy response variable is called the *linear-probability model*. This model is untenable, but its failure will point the way toward more adequate specifications:

- Because Y_i can take on only the values 0 and 1, the conditional distribution of the error ε_i is dichotomous as well—and, hence, is not normally distributed, as assumed: If $Y_i = 1$, which occurs with probability π_i , then

$$\varepsilon_i = 1 - E(Y_i) = 1 - (\alpha + \beta X_i) = 1 - \pi_i$$

Alternatively, if $Y_i = 0$, which occurs with probability $1 - \pi_i$, then

$$\varepsilon_i = 0 - E(Y_i) = 0 - (\alpha + \beta X_i) = 0 - \pi_i = -\pi_i$$

Because of the central limit theorem, however, the assumption of normality is not critical to least-squares estimation of the normal-probability model, as long as the sample size is sufficiently large.

- The variance of ε cannot be constant, as we can readily demonstrate: If the assumption of linearity holds over the range of the data, then $E(\varepsilon_i) = 0$. Using the relations just noted,

$$V(\varepsilon_i) = \pi_i(1 - \pi_i)^2 + (1 - \pi_i)(-\pi_i)^2 = \pi_i(1 - \pi_i)$$

The heteroscedasticity of the errors bodes ill for ordinary least-squares estimation of the linear probability model, but only if the probabilities π_i get close to 0 or 1.⁶ Goldberger (1964, pp. 248–250) has proposed a correction for heteroscedasticity employing weighted least squares.⁷ Because the variances $V(\varepsilon_i)$ depend on the π_i , however, which, in turn, are functions of the unknown parameters α and β , we require preliminary estimates of the parameters to define weights. Goldberger obtains ad hoc estimates from a preliminary OLS regression; that is, he takes $\hat{V}(\varepsilon_i) = \hat{Y}_i(1 - \hat{Y}_i)$. The fitted values from an OLS regression are not constrained to the interval $[0, 1]$, and so some of these “variances” may be negative.

- This last remark suggests the most serious problem with the linear-probability model: The assumption that $E(\varepsilon_i) = 0$ —that is, the assumption of linearity—is only tenable over a limited range of X -values. If the range of the X s is sufficiently broad, then the linear specification cannot confine π to the unit interval $[0, 1]$. It makes no sense, of course, to interpret a number outside the unit interval as a probability. This difficulty is illustrated in Figure 14.1, in which the least-squares line fit to the Chilean plebiscite data produces fitted probabilities below 0 at low levels and above 1 at high levels of support for the status quo.

⁶See Exercise 14.1. Remember, however, that it is the *conditional probability*, not the *marginal probability*, of Y that is at issue: The overall proportion of 1s can be near .5 (as in the Chilean plebiscite data), and yet the conditional proportion can still get very close to 0 or 1, as is apparent in Figure 14.1.

⁷See Section 12.2.2 for a discussion of weighted-least-squares estimation.

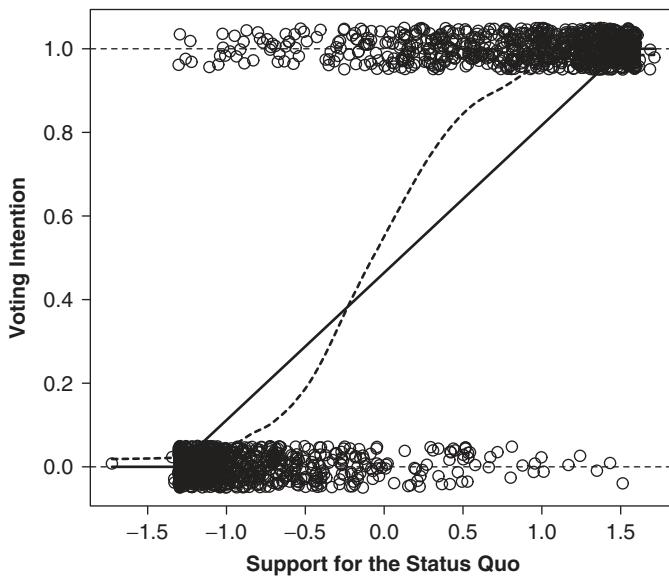


Figure 14.2 The solid line shows the constrained linear-probability model, fit by maximum likelihood to the Chilean plebiscite data. The broken line is for a nonparametric kernel regression with a span of 0.4.

Dummy-regressor variables do not cause comparable difficulties because the general linear model makes no distributional assumptions about the regressors (other than independence from the errors). Nevertheless, for values of π not too close to 0 or 1, the linear-probability model estimated by least squares frequently provides results similar to those produced by the more generally adequate methods described in the remainder of this chapter.

It is problematic to apply least-squares linear regression to a dichotomous response variable: The errors cannot be normally distributed and cannot have constant variance. Even more fundamentally, the linear specification does not confine the probability for the response to the unit interval.

One solution to the problems of the linear-probability model—though not a good general solution—is simply to constrain π to the unit interval while retaining the linear relationship between π and X within this interval:

$$\pi = \begin{cases} 0 & \text{for } 0 > \alpha + \beta X \\ \alpha + \beta X & \text{for } 0 \leq \alpha + \beta X \leq 1 \\ 1 & \text{for } \alpha + \beta X > 1 \end{cases} \quad (14.3)$$

Figure 14.2 shows the fit of this model to the Chilean plebiscite data, with the parameters α and β estimated by maximum likelihood. Although this *constrained linear-probability model* cannot be dismissed on logical grounds, the model has certain unattractive features: Most

important, the abrupt changes in slope at $\pi = 0$ and $\pi = 1$ are usually unreasonable. A smoother relationship between π and X (as characterizes the nonparametric regression in Figure 14.1) is more generally sensible. Moreover, numerical instability can make the constrained linear-probability model difficult to fit to data, and the statistical properties of estimators of the model are hard to derive because of the discontinuities in the slope.⁸

14.1.2 Transformations of π : Logit and Probit Models

A central difficulty of the unconstrained linear-probability model is its inability to ensure that π stays between 0 and 1. What we require to correct this problem is a positive monotone (i.e., nondecreasing) function that maps the *linear predictor* $\eta = \alpha + \beta X$ into the unit interval. A transformation of this type will allow us to retain the fundamentally linear structure of the model while avoiding the contradiction of probabilities below 0 or above 1. Any cumulative probability distribution function (CDF) meets this requirement.⁹ That is, we can respecify the model as

$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i) \quad (14.4)$$

where the CDF $P(\cdot)$ is selected in advance, and α and β are then parameters to be estimated.

If we choose $P(\cdot)$ as the cumulative rectangular distribution, for example, then we obtain the constrained linear-probability model (Equation 14.3).¹⁰ An a priori reasonable $P(\cdot)$ should be both smooth and symmetric and should approach $\pi = 0$ and $\pi = 1$ gradually.¹¹ Moreover, it is advantageous if $P(\cdot)$ is strictly increasing (which requires that $\pi = 0$ and $\pi = 1$ be approached as asymptotes), for then the transformation in Equation 14.4 is one to one, permitting us to rewrite the model as

$$P^{-1}(\pi_i) = \eta_i = \alpha + \beta X_i \quad (14.5)$$

where $P^{-1}(\cdot)$ is the inverse of the CDF $P(\cdot)$ (i.e., the quantile function for the distribution).¹² Thus, we have a linear model (Equation 14.5) for a transformation of π , or—equivalently—a nonlinear model (Equation 14.4) for π itself.

The transformation $P(\cdot)$ is often chosen as the CDF of the unit-normal distribution, $N(0, 1)$,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp(-\frac{1}{2}Z^2) dZ \quad (14.6)$$

or, even more commonly, of the *logistic distribution*

^{8*}Consider the strong constraints that the data place on the maximum-likelihood estimators of α and β : If, as in the illustration, $\hat{\beta} > 0$, then the rightmost observation for which $Y = 0$ can have an X -value no larger than $(1 - \hat{\alpha})/\hat{\beta}$, which is the point at which the estimated regression line hits $\hat{\pi} = 1$, because any 0 to the right of this point would produce a 0 likelihood. Similarly, the leftmost observation for which $Y = 1$ can have an X -value no smaller than $-\hat{\alpha}/\hat{\beta}$, the point at which the regression line hits $\hat{\pi} = 0$. As the sample size grows, these extreme values will tend to move, respectively, to the right and left, making $\hat{\beta}$ smaller.

⁹See online Appendix D on probability and estimation.

¹⁰See Exercise 14.2.

¹¹This is not to say, however, that $P(\cdot)$ needs to be symmetric in every case, just that symmetric $P(\cdot)$ s are more appropriate *in general*. For an example of an asymmetric choice of $P(\cdot)$, see the discussion of the complementary log-log transformation in Chapter 15.

¹²If, alternatively, the CDF levels off (as is the case, e.g., for the rectangular distribution), then the inverse of the CDF does not exist.

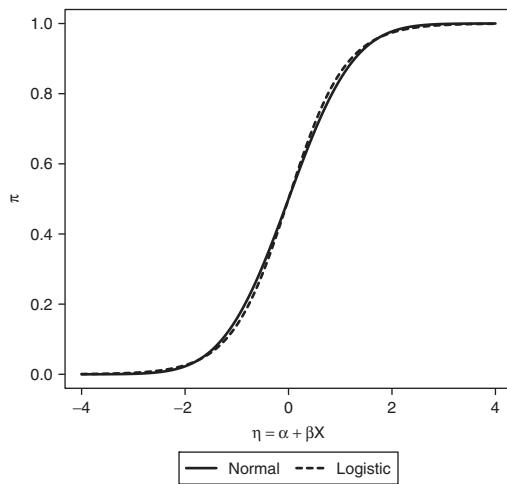


Figure 14.3 Once their variances are equated, the cumulative logistic and cumulative normal distributions—used here to transform $\eta = \alpha + \beta X$ to the unit interval—are virtually indistinguishable.

$$\Lambda(z) = \frac{1}{1 + e^{-z}} \quad (14.7)$$

In these equations, $\pi \approx 3.141$ and $e \approx 2.718$ are the familiar mathematical constants.¹³

- Using the normal distribution $\Phi(\cdot)$ yields the *linear probit model*:

$$\begin{aligned} \pi_i &= \Phi(\alpha + \beta X_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha + \beta X_i} \exp\left(-\frac{1}{2}Z^2\right) dZ \end{aligned}$$

- Using the logistic distribution $\Lambda(\cdot)$ produces the *linear logistic-regression* or *linear logit model*:

$$\begin{aligned} \pi_i &= \Lambda(\alpha + \beta X_i) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta X_i)]} \end{aligned} \quad (14.8)$$

Once their variances are equated—the logistic distribution has variance $\pi^2/3$, not 1—the logit and probit transformations are so similar that it is not possible, in practice, to distinguish between them without a great deal of data, as is apparent in Figure 14.3. It is also clear from this graph that both functions are nearly linear over much of their range, say between about

¹³A note to the reader for whom calculus is unfamiliar: An integral, represented by the symbol \int in Equation 14.6, represents the area under a curve, here the area between $Z = -\infty$ and $Z = z$ under the curve given by the function $\exp(-\frac{1}{2}Z^2)$. The constant $1/\sqrt{2\pi}$ ensures that the total area under the normal density function “integrates” (i.e., adds up) to 1.

$\pi = .2$ and $\pi = .8$. This is why the linear-probability model produces results similar to the logit and probit models, except for extreme values of π_i .

Despite their essential similarity, there are two practical advantages of the logit model compared to the probit model:

1. The equation of the logistic CDF (Equation 14.7) is very simple, while the normal CDF (Equation 14.6) involves an unevaluated integral. This difference is trivial for dichotomous data because very good closed-form approximations to the normal CDF are available, but for polytomous data, where we will require the *multivariate* logistic or normal distribution, the disadvantage of the probit model is somewhat more acute.¹⁴
2. More important, the inverse linearizing transformation for the logit model, $\Lambda^{-1}(\pi)$, is directly interpretable as a *log-odds*, while the inverse transformation for the probit model, the quantile function of the standard-normal distribution, $\Phi^{-1}(\pi)$, does not have a direct interpretation. Rearranging Equation 14.8, we get

$$\frac{\pi_i}{1 - \pi_i} = \exp(\alpha + \beta X_i) \quad (14.9)$$

The ratio $\pi_i/(1 - \pi_i)$ is the *odds* that $Y_i = 1$ (e.g., the odds of voting *yes*), an expression of relative chances familiar from gambling (at least to those who engage in this vice). Unlike the probability scale, odds are unbounded above (though bounded below by 0). Taking the log of both sides of Equation 14.9 produces

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta X_i$$

The inverse transformation $\Lambda^{-1}(\pi) = \log_e[\pi/(1 - \pi)]$, called the *logit* of π , is therefore the log of the odds that Y is 1 rather than 0. As the following table shows, if the odds are “even”—that is, equal to 1, corresponding to $\pi = .5$ —then the logit is 0. The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response variable in a linear-like model. In contrast, probabilities are bounded both below and above (at 0 and 1); odds are unbounded above but bounded below by 0.

<i>Probability</i>	<i>Odds</i> $\frac{\pi}{1 - \pi}$	<i>Logit</i>
		$\log_e \frac{\pi}{1 - \pi}$
.01	$1/99 = 0.0101$	-4.60
.05	$5/95 = 0.0526$	-2.94
.10	$1/9 = 0.1111$	-2.20
.30	$3/7 = 0.4286$	-0.85
.50	$5/5 = 1$	0.00
.70	$7/3 = 2.3333$	0.85
.90	$9/1 = 9$	2.20
.95	$95/5 = 19$	2.94
.99	$99/1 = 99$	4.60

¹⁴See Section 14.2.1. Computation of multivariate-normal probabilities is, however, feasible, so this is not a serious obstacle to fitting probit models to polytomous data.

The logit model is a linear, additive model for the log odds, but (from Equation 14.9) it is also a multiplicative model for the odds:

$$\begin{aligned}\frac{\pi_i}{1 - \pi_i} &= \exp(\alpha + \beta X_i) = \exp(\alpha)\exp(\beta X_i) \\ &= e^\alpha (e^\beta)^{X_i}\end{aligned}$$

So, increasing X by 1 changes the logit by β and multiplies the odds by e^β . For example, if $\beta = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$.¹⁵

Still another way of understanding the parameter β in the logit model is to consider the slope of the relationship between π and X , given by Equation 14.8. Because this relationship is non-linear, the slope is not constant; the slope is $\beta\pi(1 - \pi)$ and hence is at a maximum when $\pi = \frac{1}{2}$, where the slope is $\beta\frac{1}{2}(1 - \frac{1}{2}) = \beta/4$, as illustrated in the following table:¹⁶

π	$\beta\pi(1 - \pi)$
.01	$\beta \times .0099$
.05	$\beta \times .0475$
.10	$\beta \times .09$
.20	$\beta \times .16$
.50	$\beta \times .25$
.80	$\beta \times .16$
.90	$\beta \times .09$
.95	$\beta \times .0475$
.99	$\beta \times .0099$

Notice that the slope of the relationship between π and X does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

The least-squares line fit to the Chilean plebiscite data in Figure 14.1, for example, has the equation

$$\hat{\pi}_{\text{yes}} = 0.492 + 0.394 \times \text{Status quo} \quad (14.10)$$

As I have pointed out, this line is a poor summary of the data. The logistic-regression model, fit by the method of maximum likelihood (to be developed presently), has the equation

$$\log_e \frac{\hat{\pi}_{\text{yes}}}{\hat{\pi}_{\text{no}}} = 0.215 + 3.21 \times \text{Status quo}$$

As is apparent from Figure 14.1, the logit model produces a much more adequate summary of the data, one that is very close to the nonparametric regression. Increasing support for the status quo by one unit multiplies the odds of voting yes by $e^{3.21} = 24.8$. Put alternatively, the slope of the relationship between the fitted probability of voting yes and support for the status quo at

¹⁵The exponentiated coefficient e^β is sometimes called an “odds ratio” because it represents the ratio of the odds of response at two X -values, with the X -value in the numerator one unit larger than that in the denominator.

¹⁶See Exercise 14.3.

$\hat{\pi}_{\text{yes}} = .5$ is $3.21/4 = 0.80$. Compare this value with the slope ($B = 0.39$) from the linear least-squares regression in Equation 14.10.¹⁷

14.1.3 An Unobserved-Variable Formulation

An alternative derivation of the logit or probit model posits an underlying regression for a continuous but unobservable response variable ξ (representing, e.g., the “propensity” to vote *yes*), scaled so that

$$Y_i = \begin{cases} 0 & \text{when } \xi_i \leq 0 \\ 1 & \text{when } \xi_i > 0 \end{cases} \quad (14.11)$$

That is, when ξ crosses 0, the observed discrete response Y changes from *no* to *yes*. The latent variable ξ is assumed to be a linear function of the explanatory variable X and the (usual) unobservable error variable ε :

$$\xi_i = \alpha + \beta X_i - \varepsilon_i \quad (14.12)$$

(It is notationally convenient here—but otherwise inconsequential—to *subtract* the error ε rather than to add it.) We want to estimate the parameters α and β but cannot proceed by least-squares regression of ξ on X because the latent response variable (unlike Y) is not observed.

Using Equations 14.11 and 14.12,

$$\begin{aligned} \pi_i &\equiv \Pr(Y_i = 1) = \Pr(\xi_i > 0) = \Pr(\alpha + \beta X_i - \varepsilon_i > 0) \\ &= \Pr(\varepsilon_i < \alpha + \beta X_i) \end{aligned}$$

If the errors are independently distributed according to the unit-normal distribution, $\varepsilon_i \sim N(0, 1)$, then

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Phi(\alpha + \beta X_i)$$

which is the probit model.¹⁸ Alternatively, if the ε_i follow the similar logistic distribution, then we get the logit model

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Lambda(\alpha + \beta X_i)$$

We will have occasion to return to the unobserved-variable formulation of logit and probit models when we consider models for ordinal categorical data.¹⁹

¹⁷As I have explained, the slope for the logit model is not constant: It is steepest at $\pi = .5$ and flattens out as π approaches 0 and 1. The linear probability model, therefore, will agree more closely with the logit model when the response probabilities do not (as here) attain extreme values. In addition, in this example, we cannot interpret $3.21/4 = 0.80$ as an approximate effect on the probability scale when $\hat{\pi}$ is close to $.5$ because the coefficient is too large: $0.5 + 0.80 = 1.30 > 1$.

¹⁸The variance of the errors is set conveniently to 1. This choice is legitimate because we have not yet fixed the unit of measurement of the latent variable ξ . The location of the ξ scale was implicitly fixed by setting 0 as the point at which the observable response changes from *no* to *yes*. You may be uncomfortable assuming that the errors for an unobservable response variable are normally distributed, because we cannot check the assumption by examining residuals, for example. In most instances, however, we can ensure that the error distribution has any form we please by transforming ξ to make the assumption true. We cannot, however, simultaneously ensure that the true regression is linear. If the latent-variable regression is not linear, then the probit model will not adequately capture the relationship between the dichotomous Y and X . See Section 15.4.2 for a discussion of nonlinearity diagnostics for logit, probit, and other generalized linear model.

¹⁹See Section 14.2.3.

14.1.4 Logit and Probit Models for Multiple Regression

Generalizing the logit and probit models to several explanatory variables is straightforward. All we require is a linear predictor (η_i in Equation 14.13) that is a function of several regressors. For the logit model,

$$\begin{aligned}\pi_i &= \Lambda(\eta_i) = \Lambda(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})]}\end{aligned}\quad (14.13)$$

or, equivalently,

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

For the probit model,

$$\pi_i = \Phi(\eta_i) = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

Moreover, the X s can be as general as in the general linear model, including, for example:

- quantitative explanatory variables,
- transformations of quantitative explanatory variables,
- polynomial (or regression-spline) regressors formed from quantitative explanatory variables,²⁰
- dummy regressors representing qualitative explanatory variables, and
- interaction regressors.

Interpretation of the partial-regression coefficients in the general linear logit model (Equation 14.13) is similar to the interpretation of the slope in the logit simple-regression model, with the additional provision of holding other explanatory variables in the model constant. For example, expressing the model in terms of odds,

$$\begin{aligned}\frac{\pi_i}{1 - \pi_i} &= \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}) \\ &= e^\alpha (e^{\beta_1})^{X_{i1}} \cdots (e^{\beta_k})^{X_{ik}}\end{aligned}$$

Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other X s constant. Similarly, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

More adequate specifications than the linear probability model transform the linear predictor $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ smoothly to the unit interval, using a cumulative probability distribution function $P(\cdot)$. Two such specifications are the probit and the logit models, which use the normal and logistic CDFs, respectively. Although these models are very similar, the logit model is simpler to interpret because it can be written as a linear model for the log odds, $\log_e[\pi_i/(1 - \pi_i)] = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$, or, exponentiating the coefficients, as a multiplicative model for the odds, $\pi_i/(1 - \pi_i) = e^\alpha (e^{\beta_1})^{X_{i1}} \cdots (e^{\beta_k})^{X_{ik}}$.

²⁰See Chapter 17.

The general linear logit and probit models can be fit to data by the method of maximum likelihood. I will concentrate here on outlining maximum-likelihood estimation for the logit model. Details are given in the next section.

Recall that the response variable Y_i takes on the two values 1 and 0 with probabilities π_i and $1 - \pi_i$, respectively. Using a mathematical “trick,” the probability distribution for Y_i can be compactly represented as a single equation:²¹

$$p(y_i) \equiv \Pr(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

where y_i can be 0 or 1.

Now consider a particular sample of n independent observations, y_1, y_2, \dots, y_n (comprising a specific sequence of 0s and 1s). Because the observations are independent, the joint probability for the data is the product of the marginal probabilities:

$$\begin{aligned} p(y_1, y_2, \dots, y_n) &= p(y_1)p(y_2) \cdots p(y_n) \\ &= \prod_{i=1}^n p(y_i) \\ &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^n \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i) \end{aligned} \tag{14.14}$$

From the general logit model (Equation 14.13),

$$\frac{\pi_i}{1 - \pi_i} = \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

and (after some manipulation)²²

$$1 - \pi_i = \frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}$$

Substituting these results into Equation 14.14 expresses the probability of the data in terms of the parameters of the logit model:

$$\begin{aligned} p(y_1, y_2, \dots, y_n) &= \prod_{i=1}^n [\exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})]^{y_i} \\ &\quad \times \left[\frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})} \right]^{1-y_i} \end{aligned}$$

Thinking of this equation as a function of the parameters, and treating the data (y_1, y_2, \dots, y_n) as fixed, produces the likelihood function, $L(\alpha, \beta_1, \dots, \beta_k)$ for the logit model. The values of $\alpha, \beta_1, \dots, \beta_k$ that maximize $L(\alpha, \beta_1, \dots, \beta_k)$ are the maximum-likelihood estimates A, B_1, \dots, B_k .

²¹See Exercise 14.4.

²²See Exercise 14.5.

Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.²³ For an individual coefficient, it is most convenient to test the hypothesis $H_0: \beta_j = \beta_j^{(0)}$ by calculating the Wald statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\text{SE}(B_j)}$$

where $\text{SE}(B_j)$ is the asymptotic (i.e., large-sample) standard error of B_j . To test the most common hypothesis, $H_0: \beta_j = 0$, simply divide the estimated coefficient by its standard error to compute $Z_0 = B_j/\text{SE}(B_j)$; these tests are analogous to *t*-tests for individual coefficients in the general linear model. The test statistic Z_0 follows an asymptotic standard-normal distribution under the null hypothesis, an approximation that is usually reasonably accurate unless the sample size is small.²⁴ Similarly, an asymptotic $100(1 - a)\%$ confidence interval for β_j is given by

$$\beta_j = B_j \pm z_{a/2} \text{SE}(B_j)$$

where $z_{a/2}$ is the value from $Z \sim N(0, 1)$ with probability $a/2$ to the right. Wald tests and joint confidence regions for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.²⁵

It is also possible to formulate a likelihood-ratio test for the hypothesis that several coefficients are simultaneously 0, $H_0: \beta_1 = \cdots = \beta_q = 0$. We proceed, as in least-squares regression, by fitting two models to the data: the full model (Model 1),

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k$$

and the null model (Model 0),

$$\begin{aligned} \text{logit}(\pi) &= \alpha + 0X_1 + \cdots + 0X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k \\ &= \alpha + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k \end{aligned}$$

Fitting each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model. Because the null model is a specialization of the full model, $L_1 \geq L_0$. The likelihood-ratio test statistic for the null hypothesis is

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

Under the null hypothesis, this test statistic has an asymptotic chi-square distribution with q degrees of freedom.

By extension, a test of the omnibus null hypothesis $H_0: \beta_1 = \cdots = \beta_k = 0$ is obtained by specifying a null model that includes only the regression constant, $\text{logit}(\pi) = \alpha$. At the other extreme, the likelihood-ratio test can, of course, be applied to a *single* coefficient, $H_0: \beta_j = 0$, and this test can be inverted to provide a confidence interval for β_j : For example, the 95% confidence interval for β_j includes all values β'_j for which the hypothesis $H_0: \beta_j = \beta'_j$ is acceptable at the .05 level—that is, all values of β'_j for which $2(\log_e L_1 - \log_e L_0) \leq \chi^2_{0.05,1} = 3.84$, where

²³These general procedures are discussed in online Appendix D on probability and estimation.

²⁴Under certain circumstances, however, tests and confidence intervals based on the Wald statistic can break down in logistic regression (see Hauck & Donner, 1977). Tests and confidence intervals based on the likelihood-ratio statistic, described immediately below, are more reliable but more time-consuming to compute.

²⁵See Section 14.1.5.

$\log_e L_1$ is (as before) the maximized log-likelihood for the full model, and $\log_e L_0$ is the maximized log-likelihood for a model in which β_j is constrained to the value β'_j .

An analog to the multiple-correlation coefficient can also be obtained from the log-likelihood. The maximized log-likelihood for the fitted model can be written as²⁶

$$\log_e L = \sum_{i=1}^n [y_i \log_e P_i + (1 - y_i) \log_e(1 - P_i)]$$

where P_i is the fitted probability that $Y_i = 1$,²⁷ that is,

$$P_i = \frac{1}{1 + \exp[-(A + B_1 X_{i1} + \dots + B_k X_{ik})]}$$

Thus, if the fitted model can perfectly predict the Y values ($P_i = 1$ whenever $y_i = 1$, and $P_i = 0$ whenever $y_i = 0$), then $\log_e L = 0$ (i.e., the maximized likelihood is $L = 1$).²⁸ To the extent that predictions are less than perfect, $\log_e L < 0$ (and $0 < L < 1$).

By comparing $\log_e L_0$ for the model containing only the constant to $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y . The quantity $G^2 \equiv -2 \log_e L$, called the *residual deviance* under the model, is a generalization of the residual sum of squares for a linear model.²⁹ Thus,

$$\begin{aligned} R^2 &\equiv 1 - \frac{G_1^2}{G_0^2} \\ &= 1 - \frac{\log_e L_1}{\log_e L_0} \end{aligned}$$

is analogous to R^2 for a linear model.³⁰

The dichotomous logit model can be fit to data by the method of maximum likelihood. Wald tests and likelihood-ratio tests for the coefficients of the model parallel t -tests and incremental F -tests for the general linear model. The deviance for the model, defined as $G^2 = -2 \times$ the maximized log-likelihood, is analogous to the residual sum of squares for a linear model.

To illustrate logistic regression, I turn once again to the 1994 wave of the Statistics Canada Survey of Labour and Income Dynamics (the “SLID”).³¹ Confining our attention to married women between the ages of 20 and 35, I examine how the labor force participation of these

²⁶See Exercise 14.6.

²⁷In a particular sample, y_i is either 0 or 1, so we can interpret this fitted probability as the estimated population proportion of individuals sharing the i th person’s characteristics for whom Y is 1. Other interpretations are also possible, but this is the most straightforward.

²⁸Because, for the logit model, π never quite reaches 0 or 1, the predictions cannot be perfect, but they can approach perfection in the limit.

²⁹See Exercise 14.7 and Chapter 15 on generalized linear models.

³⁰For alternative R^2 measures for logit and probit models, see, for example, Veall and Zimmermann (1996).

³¹The SLID was introduced in Chapter 2.

Table 14.1 Distributions of Variables in the SLID Data Set

Variable	Summary
Labor-Force Participation	Yes, 79%
Region (R)	Atlantic, 23%; Quebec, 13; Ontario, 30; Prairies, 26; BC, 8
Children 0–4 (K04)	Yes, 53%
Children 5–9 (K59)	Yes, 44%
Children 10–14 (K1014)	Yes, 22%
Family Income (I, \$1000s)	5-number summary: 0, 18.6, 26.7, 35.1, 131.1
Education (E, years)	5-number summary: 0, 12, 13, 15, 20

women (defined as working outside the home at some point during the year of the survey) is related to several explanatory variables:

- the region of the country in which the woman resides;
- the presence of children between 0 and 4 years of age in the household, coded as absent or present;
- the presence of children between 5 and 9 years of age;
- the presence of children between 10 and 14 years of age;
- family after-tax income, excluding the woman's own income (if any);³² and
- education, defined as number of years of schooling.

The SLID data set includes 1936 women with valid data on these variables. Some information about the distribution of the variables appears in Table 14.1. Recall that the five-number summary includes the minimum, first quartile, median, third quartile, and maximum of a variable.

In modeling these data, I want to allow for the possibility of interaction between presence of children and each of family income and education in determining women's labor force participation. Table 14.2 shows the residual deviances and number of parameters for each of a series of models fit to the SLID data. These models are formulated so that likelihood-ratio tests of terms in the full model can be computed by taking differences in the residual deviances for the models, in conformity with the principle of marginality, producing "Type II" tests.³³ The residual deviances are the building blocks of likelihood-ratio tests, much as residual sums of squares are the building blocks of incremental *F*-tests in linear models. The tests themselves, with an indication of the models contrasted for each test, appear in an *analysis-of-deviance* table in Table 14.3, closely analogous to an ANOVA table for a linear model.

It is clear from the likelihood-ratio tests in Table 14.3 that none of the interactions approaches statistical significance. Presence of children 4 years old and younger and education have very highly statistically significant coefficients; the terms for region, children 5 to 9 years old, and family income are also statistically significant, while that for children 10 through 14 is not.

³²I excluded from the analysis two women for whom this variable is negative.

³³See Sections 7.3.5 and 8.2.5.

Table 14.2 Models Fit to the SLID Labor Force Participation Data

<i>Model</i>	<i>Terms in the Model</i>	<i>Number of Parameters</i>	<i>Residual Deviance</i>
0	C	1	1988.084
1	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K1014×I, K04×E, K59×E, K1014×E	16	1807.376
2	Model 1 – K04×I	15	1807.378
3	Model 1 – K59×I	15	1808.600
4	Model 1 – K1014×I	15	1807.834
5	Model 1 – K04×E	15	1807.407
6	Model 1 – K59×E	15	1807.734
7	Model 1 – K1014×E	15	1807.938
8	Model 1 – R	12	1824.681
9	C, R, K04, K59, K1014, I, E, K59×I, K1014×I, K59×E, K1014×E	14	1807.408
10	Model 9 – K04	13	1866.689
11	C, R, K04, K59, K1014, I, E, K04×I, K1014×I, K04×E, K1014×E	14	1809.268
12	Model 11 – K59	13	1819.273
13	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K04×E, K59×E	14	1808.310
14	Model 13 – K1014	13	1808.548
15	C, R, K04, K59, K1014, I, E, K04×E, K59×E, K1014×E	13	1808.854
16	Model 15 – I	12	1817.995
17	C, R, K04, K59, K1014, I, E, K04×I, K59×I, K1014×I	13	1808.428
18	Model 17 – E	12	1889.223

NOTE: "C" represents the regression constant; codes for other variables in the model are given in Table 14.1.

Estimated coefficients and standard errors for a summary model including the statistically significant terms are given in Table 14.4. The residual deviance for this model, 1810.444, is only a little larger than the residual deviance for the original model, 1807.376. The Atlantic provinces are the baseline category for the region effects in this model. The column of the table labeled e^{B_j} represents multiplicative effects on the odds scale. Thus, for example, holding the other explanatory variables constant, having children 0 to 4 years old in the household reduces the *odds* of labor force participation by $100(1 - 0.379) = 62.1\%$, and increasing education by 1 year increases the odds of labor force participation by $100(1.246 - 1) = 24.6\%$. As explained, as long as a coefficient is not too large, we can also express effects on the probability scale near $\pi = .5$ by dividing the coefficient by 4: For example (and again, holding other explanatory variables constant), if the probability of labor force participation is near .5 with children 0 to 4 absent, the presence of children of this age decreases the probability by approximately $0.9702/4 = 0.243$ or 24.3%, while an additional year of education increases the probability by approximately $0.2197/4 = .0549$ or 5.5%.

Table 14.3 Analysis of Deviance Table for the SLID Labor Force Participation Logit Model

<i>Term</i>	<i>Models Contrasted</i>	<i>df</i>	G_0^2	<i>p</i>
Region (R)	8-1	4	17.305	.0017
Children 0–4 (K04)	10-9	1	59.281	<.0001
Children 5–9 (K59)	12-11	1	10.005	.0016
Children 10–14 (K1014)	14-12	1	0.238	.63
Family Income (I)	16-15	1	9.141	.0025
Education (E)	18-17	1	80.795	<.0001
K04×I	2-1	1	0.002	.97
K59×I	3-1	1	1.224	.29
K1014×I	4-1	1	0.458	.50
K04×E	5-1	1	0.031	.86
K59×E	6-1	1	0.358	.55
K1014×E	7-1	1	0.562	.45

Table 14.4 Estimates for a Final Model Fit to the SLID Labor Force Participation Data

<i>Coefficient</i>	<i>Estimate</i> (B_j)	<i>Standard Error</i>	e^{B_j}
Constant	-0.3763	0.3398	
Region: Quebec	-0.5469	0.1899	0.579
Region: Ontario	0.1038	0.1670	1.109
Region: Prairies	0.0742	0.1695	1.077
Region: BC	0.3760	0.2577	1.456
Children 0–4	-0.9702	0.1254	0.379
Children 5–9	-0.3971	0.1187	0.672
Family income (\$1000s)	-0.0127	0.0041	0.987
Education (years)	0.2197	0.0250	1.246
Residual deviance	1810.444		

Still another strategy for interpreting a logit model is to graph the high-order terms in the model, producing effect displays, much as we did for linear models.³⁴ The final model for the SLID labor force participation data in Table 14.4 has a simple structure in that there are no interactions or polynomial terms. Nevertheless, it helps to see how each explanatory variable influences the probability of the response holding other explanatory variables to their average values. In Figure 14.4, I plot the terms in the model on the logit scale (given by the left-hand axis in each graph), preserving the linear structure of the model, but I also show corresponding fitted probabilities of labor force participation (on the right-hand axis)—a more familiar scale on which to interpret the results. The vertical axis is the same in each graph, facilitating comparison of the several partial effects.

³⁴See the discussion of effect displays for linear models in Sections 7.3.4 and 8.3.2. Details of effect displays for logit models are developed in a more general context in the next chapter (Section 15.3.4).

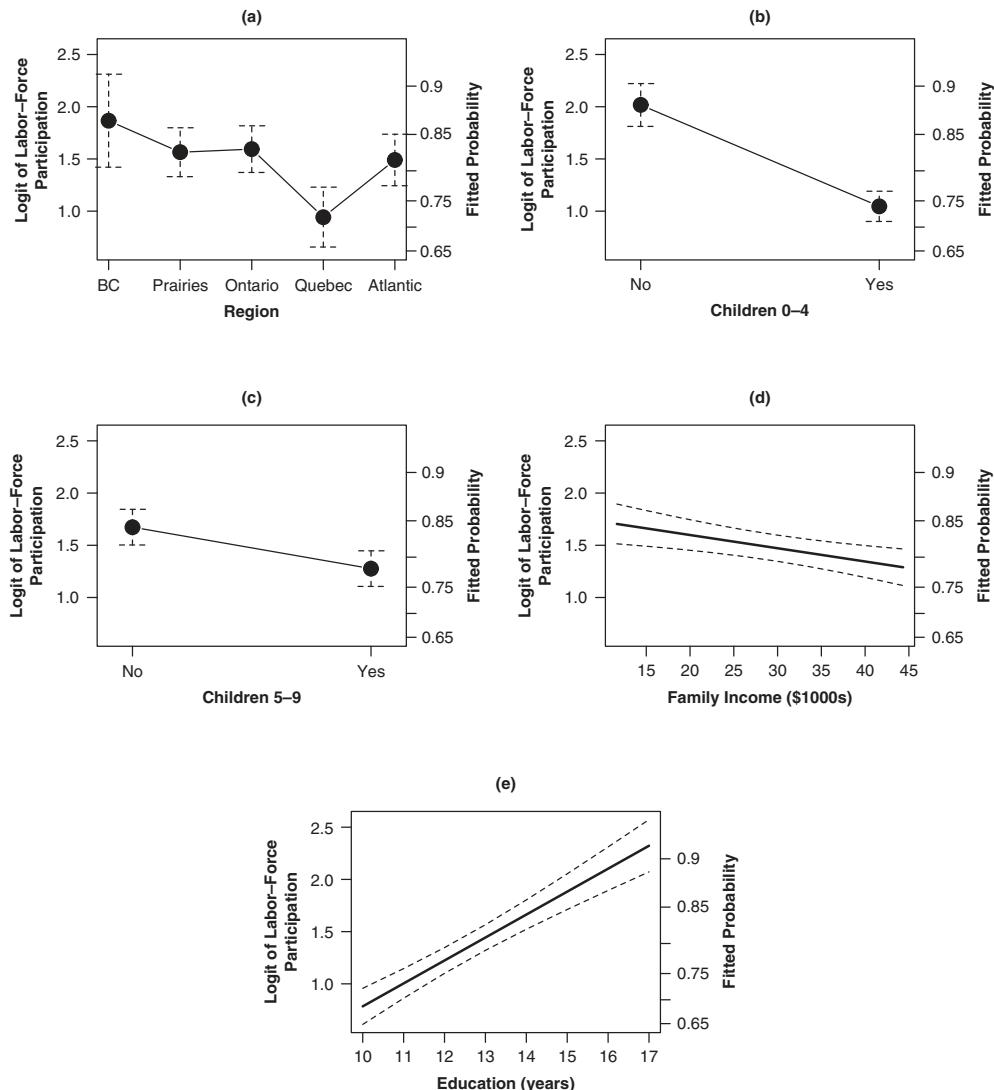


Figure 14.4 Effect displays for the summary logit model fit to the SLID labor force participation data. The error bars and envelopes give pointwise 95% confidence intervals around the estimated effects. The plots for family income and education range between the 10th and 90th percentiles of these variables.

We should not forget that the logit model fit to the SLID data is a parametric model, assuming linear partial relationships (on the logit scale) between labor force participation and the two quantitative explanatory variables, family income and education. There is no more reason to believe that relationships are necessarily linear in logit models than in linear least-squares regression. I will take up diagnostics, including nonlinearity diagnostics, for logit models and other generalized linear models in the next chapter.³⁵

³⁵See Section 15.4.

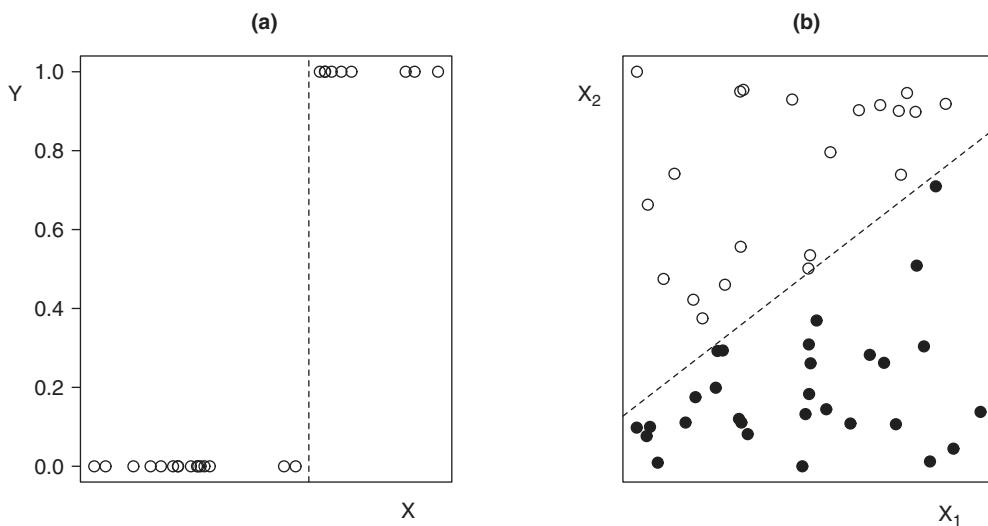


Figure 14.5 Separability in logistic regression: (a) with one explanatory variable, X ; (b) with two explanatory variables, X_1 and X_2 . In panel (b), the solid dots represent observations for which $Y = 1$ and the hollow dots observations for which $Y = 0$.

Woes of Logistic-Regression Coefficients

Just as the least-squares surface flattens out at its minimum when the X s are collinear, the likelihood surface for a logistic regression flattens out at its maximum in the presence of collinearity, so that the maximum-likelihood estimates of the coefficients of the model are not uniquely defined. Likewise, strong, but less-than-perfect, collinearity causes the coefficients to be imprecisely estimated.

Paradoxically, problems for estimation can also occur in logit models when the explanatory variables are very strong predictors of the dichotomous response. One such circumstance, illustrated in Figure 14.5, is *separability*. When there is a single X , the data are separable if the “failures” (0s) and “successes” (1s) fail to overlap [as in Figure 14.5(a)]. In this case, the maximum-likelihood estimate of the slope coefficient β is infinite (either $-\infty$ or $+\infty$, depending on the direction of the relationship between X and Y), and the estimate of the intercept α is not unique. When there are two X s, the data are separable if there is a line in the $\{X_1, X_2\}$ plane that separates successes from failures [as in Figure 14.5(b)]. For three X s, the data are separable if there is a separating plane in the three-dimensional space of the X s, and the generalization to any number of X s is a separating hyperplane—that is, a linear surface of dimension $k - 1$ in the k -dimensional X space.

Still another circumstance that yields infinite coefficients is that of data in which some of the responses become perfectly predictable even in the absence of complete separability. For example, if at one level of a factor all observations are successes, the estimated probability of success for an observation at this level is 1, and the odds of success are $1/0 = \infty$.

Statistical software may or may not detect these problems for estimation. The problems may manifest themselves in failure of the software to converge to a solution, in wildly large estimated coefficients, or in very large coefficient standard errors.

14.1.5 Estimating the Linear Logit Model*

In this section, I will develop the details of maximum-likelihood estimation for the general linear logit model (Equation 14.13 on page 380). It is convenient to rewrite the model in vector form as

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})}$$

where $\mathbf{x}'_i \equiv (1, X_{i1}, \dots, X_{ik})$ is the i th row of the model matrix \mathbf{X} , and $\boldsymbol{\beta} \equiv (\alpha, \beta_1, \dots, \beta_k)'$ is the parameter vector. The probability of n independently sampled observations of Y conditional on \mathbf{X} is, therefore,

$$p(y_1, \dots, y_n | \mathbf{X}) = \prod_{i=1}^n [\exp(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} \right]$$

and the log-likelihood function is

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^n Y_i \mathbf{x}'_i \boldsymbol{\beta} - \sum_{i=1}^n \log_e [1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})]$$

The partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned} \frac{\partial \log_e L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n Y_i \mathbf{x}_i - \sum_{i=1}^n \left[\frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} \right] \mathbf{x}_i \\ &= \sum_{i=1}^n Y_i \mathbf{x}_i - \sum_{i=1}^n \left[\frac{1}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})} \right] \mathbf{x}_i \end{aligned} \quad (14.15)$$

Setting the vector of partial derivatives to $\mathbf{0}$ to maximize the likelihood yields estimating equations

$$\sum_{i=1}^n \left[\frac{1}{1 + \exp(-\mathbf{x}'_i \mathbf{b})} \right] \mathbf{x}_i = \sum_{i=1}^n Y_i \mathbf{x}_i \quad (14.16)$$

where $\mathbf{b} = (A, B_1, \dots, B_k)'$ is the vector of maximum-likelihood estimates.

The estimating equations (14.16) have the following intuitive justification:

$$P_i \equiv \frac{1}{1 + \exp(-\mathbf{x}'_i \mathbf{b})}$$

is the *fitted* probability for observation i (i.e., the estimated value of π_i). The estimating equations, therefore, set the “fitted sum” $\sum P_i \mathbf{x}_i$ equal to the corresponding observed sum $\sum Y_i \mathbf{x}_i$. In matrix form, we can write the estimating equations as $\mathbf{X}' \mathbf{p} = \mathbf{X}' \mathbf{y}$, where $\mathbf{p} = (P_1, \dots, P_n)'$ is the vector of fitted values. Note the essential similarity to the least-squares estimating equations $\mathbf{X}' \mathbf{b} = \mathbf{X}' \hat{\mathbf{y}}$, which can be written $\mathbf{X}' \hat{\mathbf{y}} = \mathbf{X}' \mathbf{y}$.

Because \mathbf{b} is a maximum-likelihood estimator, its estimated asymptotic covariance matrix can be obtained from the inverse of the information matrix³⁶

³⁶See online Appendix D on probability and estimation.

$$\mathcal{I}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]$$

evaluated at $\boldsymbol{\beta} = \mathbf{b}$. Differentiating Equation 14.15 and making the appropriate substitutions,³⁷

$$\begin{aligned}\widehat{\mathcal{V}}(\mathbf{b}) &= \sum_{i=1}^n \left\{ \frac{\exp(-\mathbf{x}_i' \mathbf{b})}{[1 + \exp(-\mathbf{x}_i' \mathbf{b})]^2} \mathbf{x}_i \mathbf{x}_i' \right\}^{-1} \\ &= \left[\sum_{i=1}^n P_i(1 - P_i) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \\ &= (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1}\end{aligned}$$

where $\mathbf{V} \equiv \text{diag}\{P_i(1 - P_i)\}$ contains the estimated variances of the Y_i s. The square roots of the diagonal entries of $\widehat{\mathcal{V}}(\mathbf{b})$ are the asymptotic standard errors, which can be used, as described in the previous section, for Wald-based inferences about individual parameters of the logit model.

As for the linear model estimated by least squares, general linear hypotheses about the parameters of the logit model can be formulated as $H_0: \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, where \mathbf{L} is a $(q \times k+1)$ hypothesis matrix of rank $q \leq k+1$ and \mathbf{c} is a $q \times 1$ vector of fixed elements, typically 0.³⁸ Then the Wald statistic

$$Z_0^2 = (\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})$$

follows an asymptotic chi-square distribution with q degrees of freedom under the hypothesis H_0 . For example, to test the omnibus hypothesis $H_0: \beta_1 = \dots = \beta_k = 0$, we take

$$(\mathbf{L}_{(k \times k+1)}) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = [\mathbf{0}_{(k \times 1)}, \mathbf{I}_k]$$

and $\mathbf{c} = \mathbf{0}_{(k \times 1)}$.

Likewise, the asymptotic $100(1 - a)\%$ joint confidence region for a subset of q parameters β_1 takes the form

$$(\mathbf{b}_1 - \boldsymbol{\beta}_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \boldsymbol{\beta}_1) \leq \chi_{q,a}^2$$

Here, \mathbf{V}_{11} is the $(q \times q)$ submatrix of $\widehat{\mathcal{V}}(\mathbf{b})$ that pertains to the estimates \mathbf{b}_1 of $\boldsymbol{\beta}_1$, and $\chi_{q,a}^2$ is the critical value of the chi-square distribution for q degrees of freedom with probability a to the right.

Unlike the normal equations for a linear model, the logit-model estimating equations (14.16) are nonlinear functions of \mathbf{b} and, therefore, require iterative solution. One common approach to solving the estimating equations is the *Newton-Raphson method*, which can be described as follows:³⁹

³⁷See Exercise 14.8.

³⁸See Section 9.4.3.

³⁹This approach was first applied by R. A. Fisher, in the context of a probit model, and is sometimes termed *Fisher scoring* in his honor.

1. Select initial estimates \mathbf{b}_0 ; a simple choice is $\mathbf{b}_0 = \mathbf{0}$.
2. At each iteration $l + 1$, compute new estimates

$$\mathbf{b}_{l+1} = \mathbf{b}_l + (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p}_l) \quad (14.17)$$

where $\mathbf{p}_l = \{1/[1 + \exp(-\mathbf{x}_i'\mathbf{b}_l)]\}$ is the vector of fitted values from the previous iteration and $\mathbf{V}_l = \text{diag}\{P_{li}(1 - P_{li})\}$.

3. Iterations continue until $\mathbf{b}_{l+1} \approx \mathbf{b}_l$ to the desired degree of accuracy. When convergence takes place,

$$(\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p}_l) \approx \mathbf{0}$$

and thus the estimating equations $\mathbf{X}'\mathbf{p} = \mathbf{X}'\mathbf{y}$ are approximately satisfied. Conversely, if the fitted sums $\mathbf{X}'\mathbf{p}_l$ are very different from the observed sums $\mathbf{X}'\mathbf{y}$, then there will be a large adjustment in \mathbf{b} at the next iteration. The Newton-Raphson procedure conveniently produces the estimated asymptotic covariance matrix of the coefficients $\widehat{\mathcal{V}}(\mathbf{b}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$ as a by-product.

Suppose, now, that we have obtained complete convergence of the Newton-Raphson procedure to the maximum-likelihood estimator \mathbf{b} . From Equation 14.17, we have

$$\mathbf{b} = \mathbf{b} + (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{p})$$

which we can rewrite as

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}^*$$

where⁴⁰

$$\mathbf{y}^* = \mathbf{X}\mathbf{b} + \mathbf{V}^{-1}(\mathbf{y} - \mathbf{p})$$

These formulas suggest an analogy between maximum-likelihood estimation of the linear logit model and weighted-least-squares regression. The analogy is the basis of an alternative method for calculating the maximum-likelihood estimates called *iterative weighted least squares* (IWLS):⁴¹

1. As before, select arbitrary initial values \mathbf{b}_0 .
2. At each iteration l , calculate fitted values $\mathbf{p}_l = \{1/[1 + \exp(-\mathbf{x}_i'\mathbf{b}_l)]\}$, the variance matrix $\mathbf{V}_l = \text{diag}\{P_{li}(1 - P_{li})\}$, and the “pseudoresponse variable” $\mathbf{y}_l^* = \mathbf{X}\mathbf{b}_l + \mathbf{V}_l^{-1}(\mathbf{y} - \mathbf{p}_l)$.
3. Calculate updated estimates by weighted-least-squares regression of the pseudoresponse on the X s, using the current variance matrix for weights:

$$\mathbf{b}_{l+1} = (\mathbf{X}'\mathbf{V}_l\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_l\mathbf{y}_l^*$$

4. Repeat Steps 2 and 3 until the coefficients converge.

⁴⁰See Exercise 14.9.

⁴¹This method is also called *iteratively reweighted least squares* (IRLS). See Section 12.2.2 for an explanation of weighted-least-squares estimation. In fact, the IWLS algorithm is an alternative implementation of the Newton-Raphson method and leads to the same history of iterations.

14.2 Models for Polytomous Data

A limitation of the logit and probit models of the previous section is that they apply only to dichotomous response variables. In the Chilean plebiscite data, for example, many of the voters surveyed indicated that they were undecided, and some said that they planned to abstain or refused to reveal their voting intentions. Polytomous data of this sort are common, and it is desirable to model them in a natural manner—not simply to ignore some of the categories (e.g., restricting attention to those who responded *yes* or *no*) or to combine categories arbitrarily to produce a dichotomy.⁴²

In this section, I will describe three general approaches to modeling polytomous data:⁴³

1. modeling the polytomy directly as a set of unordered categories, using a generalization of the dichotomous logit model;
2. constructing a set of nested dichotomies from the polytomy, fitting an independent logit or probit model to each dichotomy; and
3. extending the unobserved-variable interpretation of the dichotomous logit and probit models to ordered polytomies.

14.2.1 The Polytomous Logit Model

It is possible to generalize the dichotomous logit model to a polytomy by employing the multivariate logistic distribution. This approach has the advantage of treating the categories of the polytomy in a nonarbitrary, symmetric manner (but the disadvantage that the analysis is relatively complex).⁴⁴

Suppose that the response variable Y can take on any of m qualitative values, which, for convenience, we number $1, 2, \dots, m$. To anticipate the example employed in this section, a voter in the 2001 British election voted for (1) the Labour Party, (2) the Conservative Party, or (3) the Liberal Democrats (disregarding smaller and regional parties). Although the categories of Y are numbered, we do not, in general, attribute ordinal properties to these numbers: They are simply category *labels*. Let π_{ij} denote the probability that the i th observation falls in the j th category of the response variable; that is, $\pi_{ij} \equiv \Pr(Y_i = j)$, for $j = 1, \dots, m$.

We have available k regressors, X_1, \dots, X_k , on which the π_{ij} depend. More specifically, this dependence is modeled using the *multivariate logistic distribution*:

$$\pi_{ij} = \frac{\exp(\gamma_{0j} + \gamma_{1j}X_{i1} + \dots + \gamma_{kj}X_{ik})}{1 + \sum_{l=1}^{m-1} \exp(\gamma_{0l} + \gamma_{1l}X_{i1} + \dots + \gamma_{kl}X_{ik})} \quad \text{for } j = 1, \dots, m - 1 \quad (14.18)$$

$$\pi_{im} = 1 - \sum_{j=1}^{m-1} \pi_{ij} \quad (\text{for category } m) \quad (14.19)$$

⁴²Additional statistical models for polytomous data are described, for example, in Agresti (2012).

⁴³A similar probit model based on the multivariate-normal distribution is slightly more difficult to estimate because of the necessity of evaluating a multivariate integral but is sometimes preferred to the polytomous logit model developed in this section (see Exercise 14.12).

This model is sometimes called the *multinomial logit model*.⁴⁴ There is, then, one set of parameters, $\gamma_{0j}, \gamma_{1j}, \dots, \gamma_{kj}$, for each response category but the last. The last category (i.e., category m) functions as a type of baseline. The use of a baseline category is one way of avoiding redundant parameters because of the restriction, reflected in Equation 14.19, that the response category probabilities for each observation must sum to 1:⁴⁵

$$\sum_{j=1}^m \pi_{ij} = 1$$

The denominator of π_{ij} in Equation 14.18 imposes this restriction.

Some algebraic manipulation of Equation 14.18 produces⁴⁶

$$\log_e \frac{\pi_{ij}}{\pi_{im}} = \gamma_{0j} + \gamma_{1j}X_{i1} + \dots + \gamma_{kj}X_{ik} \quad \text{for } j = 1, \dots, m-1$$

The regression coefficients, therefore, represent effects on the log-odds of membership in category j versus the baseline category m . It is also possible to form the log-odds of membership in *any* pair of categories j and j' (other than category m):

$$\begin{aligned} \log_e \frac{\pi_{ij}}{\pi_{ij'}} &= \log_e \left(\frac{\pi_{ij}/\pi_{im}}{\pi_{ij'}/\pi_{im}} \right) \\ &= \log_e \frac{\pi_{ij}}{\pi_{im}} - \log_e \frac{\pi_{ij'}}{\pi_{im}} \\ &= (\gamma_{0j} - \gamma_{0j'}) + (\gamma_{1j} - \gamma_{1j'})X_{i1} + \dots + (\gamma_{kj} - \gamma_{kj'})X_{ik} \end{aligned} \tag{14.20}$$

Thus, the regression coefficients for the logit between any pair of categories are the *differences* between corresponding coefficients for the two categories.

To gain further insight into the polytomous logit model, suppose that the model is specialized to a dichotomous response variable. Then, $m = 2$, and

$$\log_e \frac{\pi_{i1}}{\pi_{i2}} = \log_e \frac{\pi_{i1}}{1 - \pi_{i1}} = \gamma_{01} + \gamma_{11}X_{i1} + \dots + \gamma_{k1}X_{ik}$$

When it is applied to a dichotomy, the polytomous logit model is, therefore, identical to the dichotomous logit model of the previous section.

⁴⁴I prefer to reserve the term *multinomial logit model* for a version of the model that can accommodate counts for the several categories of the response variable in a contingency table formed by discrete explanatory variables. I make a similar distinction between *binary* and *binomial* logit models, with the former term applied to individual observations and the latter to counts of “successes” and “failures” for a dichotomous response. See the discussion of the application of logit models to contingency tables in Section 14.3.

⁴⁵An alternative is to treat the categories symmetrically:

$$\pi_{ij} = \frac{\exp(\gamma_{0j} + \gamma_{1j}X_{i1} + \dots + \gamma_{kj}X_{ik})}{\sum_{l=1}^m \exp(\gamma_{0l} + \gamma_{1l}X_{i1} + \dots + \gamma_{kl}X_{ik})}$$

but to impose a linear restriction—analogous to a sigma constraint in an ANOVA model (see Chapter 8)—on the parameters of the model. This approach produces somewhat more difficult computations, however, and has no real advantages. Although the choice of baseline category is essentially arbitrary and inconsequential, if one of the response categories represents a natural point of comparison, one might as well use it as the baseline.

⁴⁶See Exercise 14.10.

The following example is adapted from work by Andersen, Heath, and Sinnott (2002) on the 2001 British election, using data from the final wave of the 1997–2001 British Election Panel Study (BEPS) (also see Fox & Andersen, 2006). The central issue addressed in the data analysis is the potential interaction between respondents' political knowledge and political attitudes in determining their vote. The response variable, vote, has three categories: Labour, Conservative, and Liberal Democrat; individuals who voted for smaller and regional parties are excluded from the analysis. There are several explanatory variables:

- Attitude toward European integration, an 11-point scale, with high scores representing a negative attitude (so-called Euroskepticism).
- Knowledge of the platforms of the three parties on the issue of European integration, with integer scores ranging from 0 through 3. (Labour and the Liberal Democrats supported European integration, the Conservatives were opposed.)
- Other variables included in the model primarily as “controls”—age, gender, perceptions of national and household economic conditions, and ratings of the three party leaders.

The coefficients of a polytomous logit model fit to the BEPS data are shown, along with their standard errors, in Table 14.5. This model differs from those I have described previously in this text in that it includes the product of two quantitative explanatory variables, representing the *linear-by-linear interaction* between these variables:⁴⁷ Focusing on the Conservative/Liberal Democrat logit, for example, when political knowledge is 0, the slope for attitude toward European integration (“Euroskepticism”) is -0.068 . With each unit increase in political knowledge, the slope for Euroskepticism increases by 0.183 , thus becoming increasingly positive. This result is sensible: Those with more knowledge of the parties’ positions are more likely to vote in conformity with their own position on the issue. By the same token, at low levels of Euroskepticism, the slope for political knowledge is negative, but it increases by 0.183 with each unit increase in Euroskepticism. By a Wald test, this interaction coefficient is highly statistically significant, with $Z = 0.183/0.028 = 6.53$, for which $p \ll .0001$.

A Type II analysis-of-deviance table for the model appears in Table 14.6. Note that each term has two degrees of freedom, representing the two coefficients for the term, one for the Labour/Liberal Democrat logit and the other for the Conservative/Liberal Democrat logit. All the terms in the model are highly statistically significant, with the exception of gender and perception of household economic position.

Although we can therefore try to understand the fitted model by examining its coefficients, there are two obstacles to doing so: (1) As explained, the interaction between political knowledge and attitude toward European integration requires that we perform mental gymnastics to combine the estimated coefficient for the interaction with the coefficients for the “main-effect” regressors that are marginal to the interaction. (2) The structure of the polytomous logit model, which is for log-odds of pairs of categories (each category versus the baseline Liberal Democrat category), makes it difficult to formulate a general understanding of the results.

⁴⁷For more on models of this form, see Section 17.1 on polynomial regression.

Table 14.5 Polytomous Logit Model Fit to the BEPS Data

Coefficient	Labour/Liberal Democrat	
	Estimate	Standard Error
Constant	-0.155	0.612
Age	-0.005	0.005
Gender (male)	0.021	0.144
Perception of Economy	0.377	0.091
Perception of Household Economic Position	0.171	0.082
Evaluation of Blair (Labour leader)	0.546	0.071
Evaluation of Hague (Conservative leader)	-0.088	0.064
Evaluation of Kennedy (Liberal Democrat leader)	-0.416	0.072
Euroscepticism	-0.070	0.040
Political Knowledge	-0.502	0.155
Euroscepticism × Knowledge	0.024	0.021

Coefficient	Conservative/Liberal Democrat	
	Estimate	Standard Error
Constant	0.718	0.734
Age	0.015	0.006
Gender (male)	-0.091	0.178
Perception of Economy	-0.145	0.110
Perception of Household Economic Position	-0.008	0.101
Evaluation of Blair (Labour leader)	-0.278	0.079
Evaluation of Hague (Conservative leader)	0.781	0.079
Evaluation of Kennedy (Liberal Democrat leader)	-0.656	0.086
Euroscepticism	-0.068	0.049
Political Knowledge	-1.160	0.219
Euroscepticism × Knowledge	0.183	0.028

Table 14.6 Analysis of Deviance for the Polytomous Logit Model Fit to the BEPS Data

Source	df	G_0^2	p
Age	2	13.87	.0009
Gender	2	0.45	.78
Perception of Economy	2	30.60	<<.0001
Perception of Household Economic Position	2	5.65	.059
Evaluation of Blair	2	135.37	<<.0001
Evaluation of Hague	2	166.77	<<.0001
Evaluation of Kennedy	2	68.88	<<.0001
Euroscepticism	2	78.03	<<.0001
Political Knowledge	2	55.57	<<.0001
Euroscepticism × Knowledge	2	50.80	<<.0001

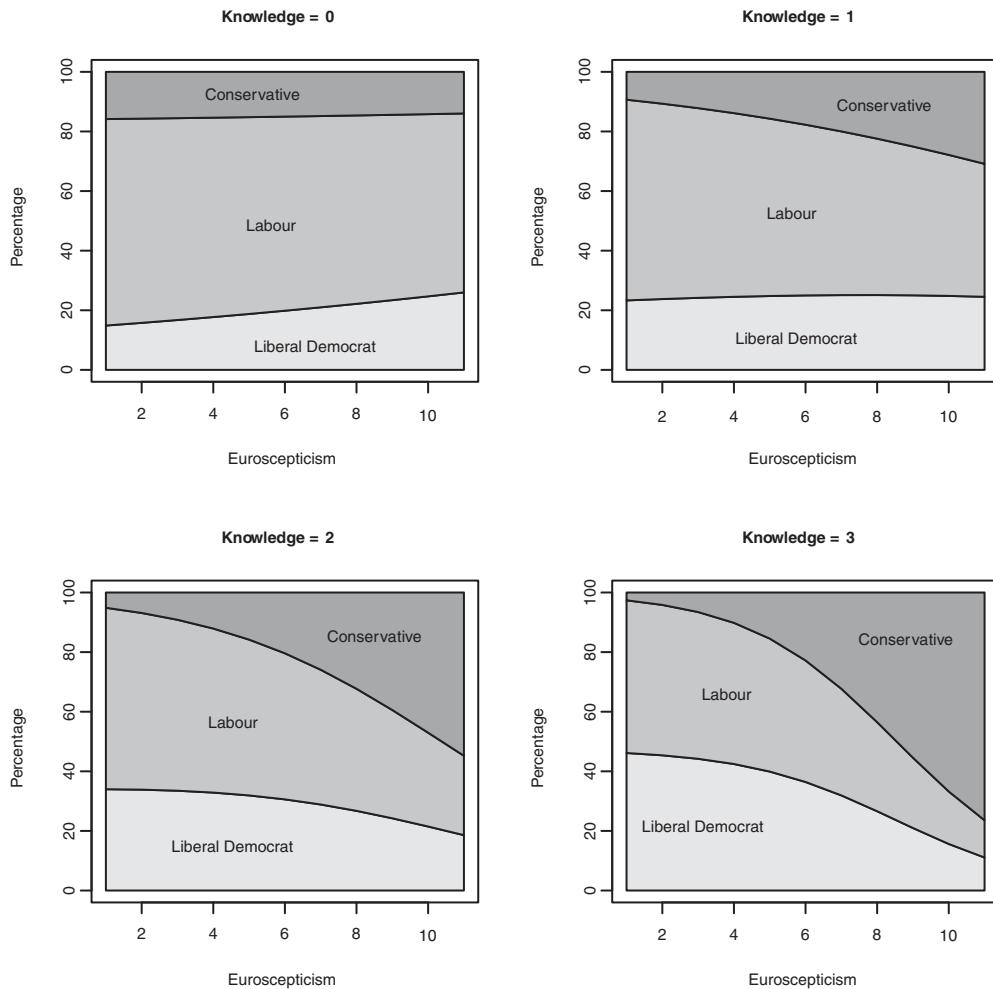


Figure 14.6 Effect display for the interaction between attitude toward European integration and political knowledge in the polytomous logit model for vote in the 2001 British election.

Once more, a graphical representation of the fitted model can greatly aid in its interpretation. An effect plot for the interaction of attitude toward European integration with political knowledge is shown in Figure 14.6. The strategy for constructing this plot is the usual one, adapted to the polytomous logit model: Compute the fitted probability of membership in each of the three categories of the response variable, letting Euroskepticism and knowledge range in combination over their values, while the other explanatory variables are fixed to average values. It is apparent that as political knowledge increases, vote conforms more closely to the respondent's attitude toward European integration.

Details of Estimation*

To fit the polytomous logit model given in Equation 14.18 (on page 392) to data, we may again invoke the method of maximum likelihood. Recall that each Y_i takes on its possible values $1, 2, \dots, m$ with probabilities $\pi_{i1}, \pi_{i2}, \dots, \pi_{im}$. Following Nerlove and Press (1973), let us define indicator variables W_{i1}, \dots, W_{im} so that $W_{ij} = 1$ if $Y_i = j$, and $W_{ij} = 0$ if $Y_i \neq j$; thus,

$$\begin{aligned} p(y_i) &= \pi_{i1}^{w_{i1}} \pi_{i2}^{w_{i2}} \cdots \pi_{im}^{w_{im}} \\ &= \prod_{j=1}^m \pi_{ij}^{w_{ij}} \end{aligned}$$

If the observations are sampled independently, then their joint probability distribution is given by

$$\begin{aligned} p(y_1, \dots, y_n) &= p(y_1) \times \cdots \times p(y_n) \\ &= \prod_{i=1}^n \prod_{j=1}^m \pi_{ij}^{w_{ij}} \end{aligned}$$

For compactness, define the following vectors:

$$\begin{aligned} \mathbf{x}'_i &\equiv (1, X_{i1}, \dots, X_{ik}) \\ \boldsymbol{\gamma}_j &\equiv (\gamma_{0j}, \gamma_{1j}, \dots, \gamma_{kj})' \end{aligned}$$

and the model matrix

$$\mathbf{X}_{(n \times k+1)} \equiv \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}$$

It is convenient to impose the restriction $\sum_{j=1}^m \pi_{ij} = 1$ by setting $\boldsymbol{\gamma}_m = \mathbf{0}$ (making category m the baseline, as explained previously). Then, employing Equations 14.18 and 14.19,

$$p(y_1, \dots, y_n | \mathbf{X}) = \prod_{i=1}^n \prod_{j=1}^m \left[\frac{\exp(\mathbf{x}'_i \boldsymbol{\gamma}_j)}{1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l)} \right]^{w_{ij}} \quad (14.21)$$

and the log-likelihood is

$$\begin{aligned} \log_e L(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_{m-1}) &= \sum_{i=1}^n \sum_{j=1}^m W_{ij} \left\{ \mathbf{x}'_i \boldsymbol{\gamma}_j - \log_e \left[1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l) \right] \right\} \\ &= \sum_{i=1}^n \sum_{j=1}^{m-1} W_{ij} \mathbf{x}'_i \boldsymbol{\gamma}_j - \sum_{i=1}^n \log_e \left[1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l) \right] \end{aligned}$$

because $\sum_{j=1}^m W_{ij} = 1$ and $\boldsymbol{\gamma}_m = \mathbf{0}$; setting $\boldsymbol{\gamma}_m = \mathbf{0}$ accounts for the 1 in the denominator of Equation 14.21, because $\exp(\mathbf{x}'_i \mathbf{0}) = 1$.

Differentiating the log-likelihood with respect to the parameters, and setting the partial derivatives to $\mathbf{0}$, produces the nonlinear estimating equations:⁴⁸

⁴⁸See Exercise 14.11.

$$\begin{aligned}\sum_{i=1}^n W_{ij} \mathbf{x}_i &= \sum_{i=1}^n \mathbf{x}_i \frac{\exp(\mathbf{x}'_i \mathbf{c}_j)}{1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \mathbf{c}_l)} \quad \text{for } j = 1, \dots, m-1 \\ &= \sum_{i=1}^n P_{ij} \mathbf{x}_i\end{aligned}\tag{14.22}$$

where $\mathbf{c}_j \equiv \hat{\gamma}_j$ are the maximum-likelihood estimators of the regression coefficients, and the

$$P_{ij} \equiv \frac{\exp(\mathbf{x}'_i \mathbf{c}_j)}{1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \mathbf{c}_l)}$$

are the fitted probabilities. As in the dichotomous logit model, the maximum-likelihood estimator sets observed sums equal to fitted sums. The estimating equations (14.22) are nonlinear and, therefore, require iterative solution.

Let us stack up all the parameters in a large vector:

$$\boldsymbol{\gamma}_{[(m-1)(k+1) \times 1]} \equiv \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{m-1} \end{bmatrix}$$

The information matrix is⁴⁹

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})_{[(m-1)(k+1) \times (m-1)(k+1)]} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} & \cdots & \mathcal{I}_{1,m-1} \\ \mathcal{I}_{21} & \mathcal{I}_{22} & \cdots & \mathcal{I}_{2,m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{I}_{m-1,1} & \mathcal{I}_{m-1,2} & \cdots & \mathcal{I}_{m-1,m-1} \end{bmatrix}$$

where

$$\begin{aligned}\mathcal{I}_{jj'}_{[(k+1) \times (k+1)]} &= -E \left[\frac{\partial^2 \log_e L(\boldsymbol{\gamma})}{\partial \gamma_j \partial \gamma'_{j'}} \right] \\ &= \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}'_i \exp(\mathbf{x}'_i \boldsymbol{\gamma}_j) [1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l) - \exp(\mathbf{x}'_i \boldsymbol{\gamma}_j)]}{[1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l)]^2}\end{aligned}\tag{14.23}$$

and

$$\begin{aligned}\mathcal{I}_{jj'}_{[(k+1) \times (k+1)]} &= -E \left[\frac{\partial^2 \log_e L(\boldsymbol{\gamma})}{\partial \gamma_j \partial \gamma'_{j'}} \right] \\ &= - \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}'_i \exp[\mathbf{x}'_i (\boldsymbol{\gamma}_{j'} + \boldsymbol{\gamma}_j)]}{[1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_i \boldsymbol{\gamma}_l)]^2}\end{aligned}\tag{14.24}$$

The estimated asymptotic covariance matrix of

$$\mathbf{c} \equiv \begin{bmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{m-1} \end{bmatrix}$$

is obtained from the inverse of the information matrix, replacing $\boldsymbol{\gamma}$ with \mathbf{c} .

⁴⁹See Exercise 14.11.

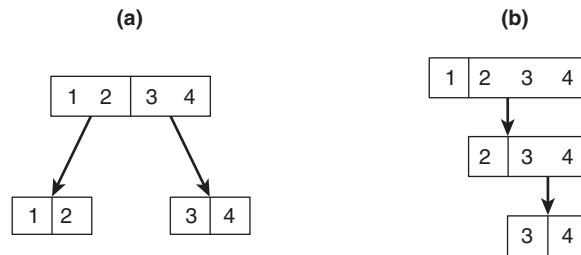


Figure 14.7 Alternative sets of nested dichotomies [(a) and (b)] for a four-category polytomous response variable.

14.2.2 Nested Dichotomies

Perhaps the simplest approach to polytomous data—because it employs the already familiar dichotomous logit or probit model—is to fit separate models to each of a set of dichotomies derived from the polytomy. These dichotomies are constructed so that the likelihood for the polytomous response variable is the product of the likelihoods for the dichotomies—that is, the models are statistically independent even though they are fitted to data from the same sample. The likelihood is separable in this manner if the set of dichotomies is *nested*.⁵⁰ Although the system of nested dichotomies constitutes a model for the polytomy, and although this model often yields fitted probabilities that are very similar to those associated with the polytomous logit model of the previous section, the two models are not equivalent.

A nested set of $m - 1$ dichotomies is produced from an m -category polytomy by successive binary partitions of the categories of the polytomy. Two examples for a four-category variable are shown in Figure 14.7. In part (a) of this figure, the dichotomies are $\{1, 2, 34\}$ (i.e., the combination of Categories 1 and 2 vs. the combination of Categories 3 and 4); $\{1, 2\}$ (Category 1 vs. Category 2); and $\{3, 4\}$ (Category 3 vs. Category 4). In part (b), the nested dichotomies are $\{1, 2, 34\}$, $\{2, 34\}$, and $\{3, 4\}$. This simple—and abstract—example illustrates a key property of nested dichotomies: The set of nested dichotomies selected to represent a polytomy is *not* unique. Because the results of the analysis (e.g., fitted probabilities under the model) and their interpretation depend on the set of nested dichotomies that is selected, this approach to polytomous data is reasonable only when a particular choice of dichotomies is substantively compelling. If the dichotomies are purely arbitrary, or if alternative sets of dichotomies are equally reasonable and interesting, then nested dichotomies should probably not be used to analyze the data.

Nested dichotomies are an especially attractive approach when the categories of the polytomy represent ordered progress through the stages of a process. Imagine, for example, that the categories in Figure 14.7(b) represent adults' attained level of education: (1) less than high school, (2) high school graduate, (3) some postsecondary, or (4) postsecondary degree. Because individuals normally progress through these categories in sequence, the dichotomy $\{1, 2, 34\}$ represents the completion of high school; $\{2, 34\}$ the continuation to postsecondary

⁵⁰A proof of this property of nested dichotomies will be given presently.

education, conditional on high school graduation; and $\{3, 4\}$ the completion of a degree conditional on undertaking a postsecondary education.⁵¹

Why Nested Dichotomies Are Independent*

For simplicity, I will demonstrate the independence of the nested dichotomies $\{1, 2\}$ and $\{1, 2\}$. By repeated application, this result applies generally to any system of nested dichotomies. Let W_{i1} , W_{i2} , and W_{i3} be dummy variables indicating whether the polytomous response variable Y_i is 1, 2, or 3. For example, $W_{i1} = 1$ if $Y_i = 1$, and 0 otherwise. Let Y'_i be a dummy variable representing the first dichotomy, $\{1, 2\}$: That is, $Y'_i = 1$ when $Y_i = 1$ or 2, and $Y'_i = 0$ when $Y_i = 3$. Likewise, let Y''_i be a dummy variable representing the second dichotomy, $\{1, 2\}$: $Y''_i = 1$ when $Y_i = 1$, and $Y''_i = 0$ when $Y_i = 2$; Y''_i is *undefined* when $Y_i = 3$. We need to show that $p(y_i) = p(y'_i)p(y''_i)$. [To form this product, we adopt the convention that $p(y''_i) \equiv 1$ when $Y_i = 3$.]

The probability distribution of Y'_i is given by

$$\begin{aligned} p(y'_i) &= (\pi_{i1} + \pi_{i2})^{y'_i} \pi_{i3}^{1-y'_i} \\ &= (\pi_{i1} + \pi_{i2})^{w_{i1}+w_{i2}} \pi_{i3}^{w_{i3}} \end{aligned} \quad (14.25)$$

where $\pi_{ij} \equiv Pr(Y_i = j)$ for $j = 1, 2, 3$. To derive the probability distribution of Y''_i , note that

$$\begin{aligned} \Pr(Y''_i = 1) &= \Pr(Y_i = 1 | Y_i \neq 3) = \frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}} \\ \Pr(Y''_i = 0) &= \Pr(Y_i = 2 | Y_i \neq 3) = \frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}} \end{aligned}$$

and, thus,

$$\begin{aligned} p(y''_i) &= \left(\frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}} \right)^{y''_i} \left(\frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}} \right)^{1-y''_i} \\ &= \left(\frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}} \right)^{w_{i1}} \left(\frac{\pi_{i2}}{\pi_{i1} + \pi_{i2}} \right)^{w_{i2}} \end{aligned} \quad (14.26)$$

Multiplying Equation 14.25 by Equation 14.26 produces

$$p(y'_i)p(y''_i) = \pi_{i1}^{w_{i1}} \pi_{i2}^{w_{i2}} \pi_{i3}^{w_{i3}} = p(y_i)$$

which is the required result.

Because the dichotomies Y' and Y'' are independent, it is legitimate to combine models for these dichotomies to form a model for the polytomy Y . Likewise, we can sum likelihood-ratio or Wald test statistics for the two dichotomies.

14.2.3 Ordered Logit and Probit Models

Imagine (as in Section 14.1.3) that there is a latent (i.e., unobservable) variable ξ that is a linear function of the X s plus a random error:

⁵¹Fienberg (1980, pp. 110–116) terms ratios of odds formed from these nested dichotomies *continuation ratios*. An example employing nested dichotomies for educational attainment is developed in the data analysis exercises for this chapter.



Figure 14.8 The thresholds $\alpha_1 < \alpha_2 < \dots < \alpha_{m-1}$ divide the latent continuum ξ into m regions, corresponding to the values of the observable variable Y .

$$\xi_i = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

Now, however, suppose that instead of dividing ξ into two regions to produce a dichotomous response, ξ is dissected by $m - 1$ thresholds (i.e., boundaries) into m regions. Denoting the thresholds by $\alpha_1 < \alpha_2 < \dots < \alpha_{m-1}$, and the resulting response by Y , we observe

$$Y_i = \begin{cases} 1 & \text{if } \xi_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < \xi_i \leq \alpha_2 \\ \vdots & \\ m-1 & \text{if } \alpha_{m-2} < \xi_i \leq \alpha_{m-1} \\ m & \text{if } \alpha_{m-1} < \xi_i \end{cases} \quad (14.27)$$

The thresholds, regions, and corresponding values of ξ and Y are represented graphically in Figure 14.8. As in this graph, the thresholds are not in general uniformly spaced.

Using Equation 14.27, we can determine the cumulative probability distribution of Y :

$$\begin{aligned} \Pr(Y_i \leq j) &= \Pr(\xi_i \leq \alpha_j) \\ &= \Pr(\alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i \leq \alpha_j) \\ &= \Pr(\varepsilon_i \leq \alpha_j - \alpha - \beta_1 X_{i1} - \dots - \beta_k X_{ik}) \end{aligned}$$

If the errors ε_i are independently distributed according to the standard normal distribution, then we obtain the ordered probit model.⁵² If the errors follow the similar logistic distribution, then we get the ordered logit model. In the latter event,

$$\begin{aligned} \text{logit}[\Pr(Y_i \leq j)] &= \log_e \frac{\Pr(Y_i \leq j)}{\Pr(Y_i > j)} \\ &= \alpha_j - \alpha - \beta_1 X_{i1} - \dots - \beta_k X_{ik} \end{aligned}$$

Equivalently,

$$\begin{aligned} \text{logit}[\Pr(Y_i > j)] &= \log_e \frac{\Pr(Y_i > j)}{\Pr(Y_i \leq j)} \\ &= (\alpha - \alpha_j) + \beta_1 X_{i1} + \dots + \beta_k X_{ik} \end{aligned} \quad (14.28)$$

for $j = 1, 2, \dots, m - 1$.

The logits in Equation 14.28 are for cumulative categories—at each point contrasting categories above category j with category j and below. The slopes for each of these regression equations are identical; the equations differ only in their intercepts. The logistic-regression

⁵²As in the dichotomous case, we conveniently fix the error variance to 1 to set the scale of the latent variable ξ . The resulting ordered probit model does not have the proportional-odds property described below.

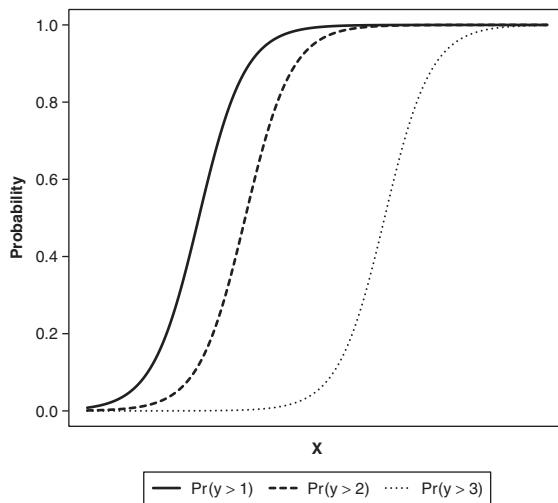


Figure 14.9 The proportional-odds model for four response categories and a single explanatory variable X . The logistic regression curves are horizontally parallel.

SOURCE: Adapted from Agresti (1990, Figure 9.1), *Categorical Data Analysis*. Copyright © 1990 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

surfaces are, therefore, horizontally parallel to each other, as illustrated in Figure 14.9 for $m = 4$ response categories and a single X . (For the more general case, just replace X by the linear predictor $\eta = \beta_1 X_1 + \dots + \beta_k X_k$.)

Put another way, for a fixed set of X s, any two different cumulative log-odds (i.e., logits)—say, at categories j and j' —differ only by the constant $(\alpha_{j'} - \alpha_j)$. The odds, therefore, are *proportional* to one another; that is,

$$\frac{\text{odds}_j}{\text{odds}_{j'}} = \exp(\text{logit}_j - \text{logit}_{j'}) = \exp(\alpha_{j'} - \alpha_j) = \frac{e^{\alpha_{j'}}}{e^{\alpha_j}}$$

where, for example, $\text{odds}_j \equiv \Pr(Y_i > j)/\Pr(Y_i \leq j)$ and $\text{logit}_j \equiv \log_e[\Pr(Y_i > j)/\Pr(Y_i \leq j)]$. For this reason, the model in Equation 14.28 is called the *proportional-odds logit model*.

There are $(k + 1) + (m - 1) = k + m$ parameters to estimate in the proportional-odds model, including the regression coefficients $\alpha, \beta_1, \dots, \beta_k$ and the category thresholds $\alpha_1, \dots, \alpha_{m-1}$. There is, however, an extra parameter in the regression equations (Equation 14.28) because each equation has its own constant, $-\alpha_j$, along with the common constant α . A simple solution is to set $\alpha = 0$ (and to absorb the negative sign into α_j), producing⁵³

$$\text{logit}[\Pr(Y_i > j)] = \alpha_j + \beta_1 X_{i1} + \dots + \beta_k X_{ik} \quad (14.29)$$

In this parameterization, the intercepts α_j are the *negatives* of the category thresholds.

⁵³Setting $\alpha = 0$ implicitly establishes the origin of the latent variable ξ (just as fixing the error variance establishes its unit of measurement). An alternative would be to fix one of the thresholds to 0. These choices are arbitrary and inconsequential.

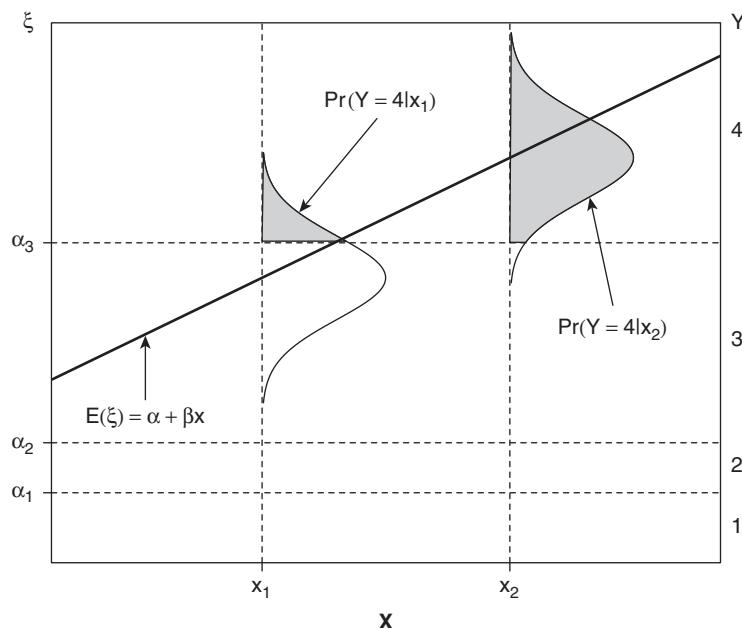


Figure 14.10 The proportional-odds model for four response categories and a single explanatory variable X . The latent response variable ξ has a linear regression on X . The latent continuum ξ and thresholds α_j appear at the left of the graph, the observable response Y at the right. The conditional logistic distribution of the latent variable is shown for two values of the explanatory variable, x_1 and x_2 . The shaded area in each distribution gives the conditional probability that $Y = 4$.

SOURCE: Adapted from Agresti (1990, Figure 9.2), *Categorical Data Analysis*. Copyright © 1990 John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

Figure 14.10 illustrates the proportional-odds model for $m = 4$ response categories and a single X . The conditional distribution of the latent variable ξ is shown for two representative values of the explanatory variable, x_1 [where $\Pr(Y > 3) = \Pr(Y = 4)$ is about .2] and x_2 [where $\Pr(Y = 4)$ is about .98]. McCullagh (1980) explains how the proportional-odds model can be fit by the method of maximum likelihood (and discusses alternatives to this model).

To illustrate the use of the proportional-odds model, I draw on data from the World Values Survey (WVS) of 1995–1997 (European Values Study Group and World Values Survey Association, 2000).⁵⁴ Although the WVS collects data in many countries, to provide a manageable example, I will restrict attention to only four: Australia, Sweden, Norway, and the United States.⁵⁵ The combined sample size for these four countries is 5381. The response variable in the analysis is the answer to the question “Do you think that what the government is doing for people in poverty is about the right amount, too much, or too little?” There are, therefore, three

⁵⁴This illustration is adapted from Fox and Andersen (2006).

⁵⁵Using data from a larger number of countries, we could instead entertain hierarchical mixed-effects models for polytomous data, analogous to the linear and generalized linear mixed-effects models introduced in Chapters 23 and 24.

Table 14.7 Analysis of Deviance Table for the Proportional-Odds Model Fit to the World Values Survey Data

Source	df	G_0^2	p
Country	3	250.881	<.0001
Gender	1	10.749	.0010
Religion	1	4.132	.042
Education	1	4.284	.038
Age	1	49.950	<.0001
Country × Gender	3	3.049	.38
Country × Religion	3	21.143	<.0001
Country × Education	3	12.861	.0049
Country × Age	3	17.529	.0005

ordered categories: *too little*, *about right*, and *too much*. There are several explanatory variables: gender (represented by a dummy variable coded 1 for *men* and 0 for *women*), whether or not the respondent belonged to a religion (coded 1 for *yes*, 0 for *no*), whether or not the respondent had a university degree (coded 1 for *yes* and 0 for *no*), age (in years, ranging from 18–87), and country (entered into the model as a set of three dummy regressors, with *Australia* as the baseline category). Preliminary analysis of the data suggested a roughly linear age effect.

Table 14.7 shows the analysis of deviance for an initial model fit to the data incorporating interactions between country and each of the other explanatory variables. As usual, the likelihood-ratio tests in the table are computed by contrasting the deviances for alternative models, with and without the terms in question. These tests were formulated in conformity with the principle of marginality (i.e., Type II tests). So, for example, the test for the country-by-age interaction was computed by dropping this term from the full model, and the test for the country main effect was computed by dropping the dummy regressors for country from a model that includes only main effects.

With the exception of the interaction between country and gender, all these interactions prove to be statistically significant. Estimated coefficients and their standard errors for a final model, removing the nonsignificant interaction between country and gender, appear in Table 14.8. This table also shows the estimated thresholds between response categories, which are, as explained, the negatives of the intercepts of the proportional-odds model.

Interpretation of the estimated coefficients for the proportional-odds model in Table 14.8 is complicated by the interactions in the model and by the multiple-category response. I will use the interaction between age and country to illustrate: We can see that the age slope is positive in the baseline country of Australia (suggesting that sympathy for the poor declines with age in Australia) and that this slope is nearly zero in Norway (i.e., adding the coefficient for Norway × Age to the baseline Age coefficient), smaller in Sweden than in Australia, and very slightly larger in the United States than in Australia, but a more detailed understanding of the age-by-country interaction is hard to discern from the coefficients alone. Figures 14.11 and 14.12 show alternative effect displays of the age-by-country interaction. The strategy for constructing these displays is the usual one—compute fitted values under the model, letting age and country range over their values while other explanatory variables (i.e., gender, religion, and education) are

Table 14.8 Estimated Proportional-Odds Model Fit to the World Values Survey Data

Coefficient	Estimate	Standard Error
Gender (Men)	0.1744	0.0532
Country (Norway)	0.1516	0.3355
Country (Sweden)	-1.2237	0.5821
Country (United States)	1.2225	0.3068
Religion (Yes)	0.0255	0.1120
Education (Degree)	-0.1282	0.1676
Age	0.0153	0.0026
Country (Norway) \times Religion	-0.2456	0.2153
Country (Sweden) \times Religion	-0.9031	0.5125
Country (United States) \times Religion	0.5706	0.1733
Country (Norway) \times Education	0.0524	0.2080
Country (Sweden) \times Education	0.6359	0.2141
Country (United States) \times Education	0.3103	0.2063
Country (Norway) \times Age	-0.0156	0.0044
Country (Sweden) \times Age	-0.0090	0.0047
Country (United States) \times Age	0.0008	0.0040
<i>Thresholds</i>		
$-\hat{\alpha}_1$ (Too Little About Right)	0.7699	0.1491
$-\hat{\alpha}_2$ (About Right Too Much)	2.5372	0.1537

held to average values. Figure 14.11 plots the fitted probabilities of response (as percentages) by age for each country; Figure 14.12 plots the fitted value of the latent response variable by age for each country and shows the intercategory thresholds.

The proportional-odds model (Equation 14.29 on page 402) constrains corresponding slopes for the $m - 1$ cumulative logits to be equal. By relaxing this strong constraint, and fitting a model to the cumulative logits that permits different slopes along with different intercepts, we can test the proportional-odds assumption:

$$\text{logit}[\Pr(Y_i > j)] = \alpha_j + \beta_{j1}X_{i1} + \dots + \beta_{jk}X_{ik}, \text{ for } j = 1, \dots, m - 1 \quad (14.30)$$

Like the polytomous logit model (14.18 on page 392), this new model has $(m - 1)(k + 1)$ parameters, but the two models are for *different* sets of logits. The deviances and numbers of parameters for the three models fit to the World Values Survey data are as follows:

Model	Residual Deviance	Number of Parameters
Proportional-Odds Model (Equation 14.29)	10,350.12	18
Cumulative Logits, Unconstrained	9,961.63	34
Slopes (Equation 14.30)		
Polytomous Logit Model (Equation 14.18)	9,961.26	34

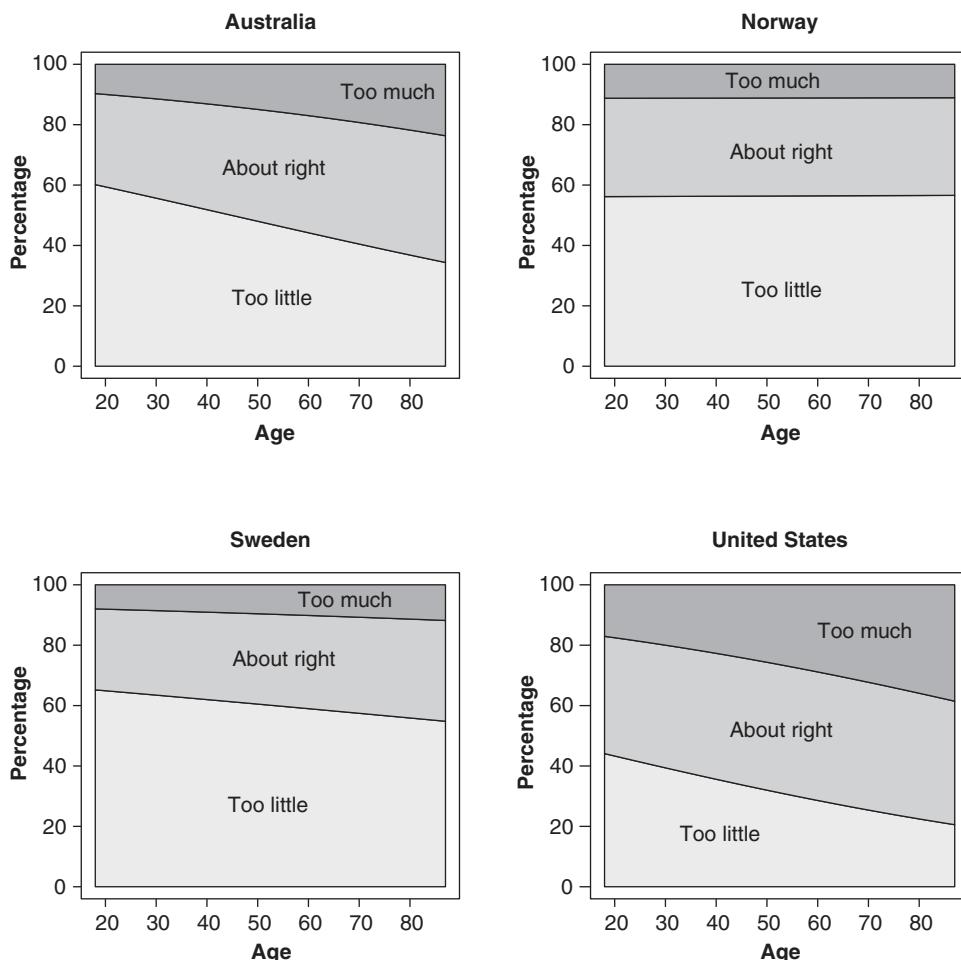


Figure 14.11 Effect display for the interaction of age with country in the proportional-odds model fit to the World Values Survey data. The response variable is assessment of government action for people in poverty.

The likelihood-ratio statistic for testing the assumption of proportional odds is therefore $G_0^2 = 10,350.12 - 9,961.63 = 388.49$, on $34 - 18 = 16$ degrees of freedom. This test statistic is highly statistically significant, leading us to reject the proportional-odds assumption for these data. Note that the deviance for the model that relaxes the proportional-odds assumption is nearly identical to the deviance for the polytomous logit model. This is typically the case, in my experience.⁵⁶

⁵⁶Consequently, if you are working with software that does not compute the unconstrained-slopes model for cumulative logits, it is generally safe to use the polytomous logit model to formulate an approximate likelihood-ratio test for proportional odds. There is also a score test and a Wald test for the proportional-odds assumption (discussed, e.g., in Long, 1997, Section 5.5).

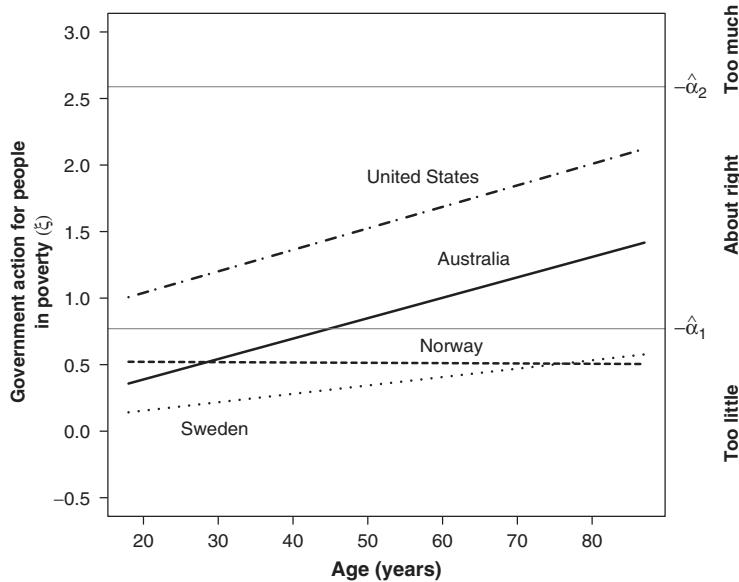


Figure 14.12 Alternative effect display for the proportional-odds model fit to the World Values Survey data, showing fitted values of the latent response. Intercategory thresholds and the corresponding response categories are given at the right of the graph and by the lighter horizontal lines.

14.2.4 Comparison of the Three Approaches

Several approaches can be taken to modeling polytomous data, including (1) modeling the polytomy directly using a logit model based on the multivariate logistic distribution, (2) constructing a set of $m - 1$ nested dichotomies to represent the m categories of the polytomy, and (3) fitting the proportional-odds model to a polytomous response variable with ordered categories.

The three approaches to modeling polytomous data—the polytomous logit model, logit models for nested dichotomies, and the proportional-odds model—address different sets of log-odds, corresponding to different dichotomies constructed from the polytomy. Consider, for example, the ordered polytomy $\{1, 2, 3, 4\}$ —representing, say, four ordered educational categories:

- Treating Category 4 as the baseline, the coefficients of the polytomous logit model apply *directly* to the dichotomies $\{1, 4\}$, $\{2, 4\}$, and $\{3, 4\}$ and *indirectly* to any pair of categories.

Table 14.9 Voter Turnout by Perceived Closeness of the Election and Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

Perceived Closeness	Intensity of Preference	Turnout		Logit $\log \frac{\text{Voted}}{\text{Did Not Vote}}$
		Voted	Did Not Vote	
One-sided	Weak	91	39	0.847
	Medium	121	49	0.904
	Strong	64	24	0.981
Close	Weak	214	87	0.900
	Medium	284	76	1.318
	Strong	201	25	2.084

NOTE: Frequency counts are shown in the body of the table.

- Forming continuation dichotomies (one of several possibilities), the nested-dichotomies approach models $\{1, 234\}$, $\{2, 34\}$, and $\{3, 4\}$.
- The proportional-odds model applies to the cumulative dichotomies $\{1, 234\}$, $\{12, 34\}$, and $\{123, 4\}$, imposing the restriction that only the intercepts of the three regression equations differ.

Which of these models is most appropriate depends partly on the structure of the data and partly on our interest in them. If it fits well, the proportional-odds model would generally be preferred for an ordered response on grounds of parsimony, but this model imposes strong structure on the data and may not fit well. Nested dichotomies should only be used if the particular choice of dichotomies makes compelling substantive sense for the data at hand. The implication, then, is that of these three models, the polytomous logit model has the greatest general range of application.⁵⁷

14.3 Discrete Explanatory Variables and Contingency Tables

When the explanatory variables—as well as the response—are discrete, the joint sample distribution of the variables defines a contingency table of counts: Each cell of the table records the number of observations possessing a particular combination of characteristics. An example, drawn from *The American Voter* (Campbell, Converse, Miller, & Stokes, 1960), a classical study of electoral behavior, appears in Table 14.9. This table, based on data from sample surveys conducted during the 1956 U.S. presidential election campaign and after the election, relates voting turnout in the election to strength of partisan preference (classified as weak, medium, or strong) and perceived closeness of the election (one-sided or close).

⁵⁷But see Exercise 14.12.

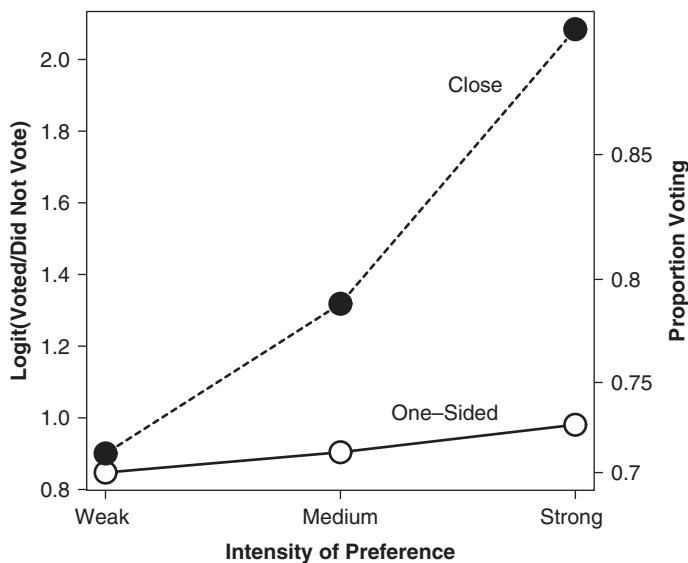


Figure 14.13 Empirical logits for voter turnout by intensity of partisan preference and perceived closeness of the election, for the 1956 U.S. presidential election.

The last column of Table 14.9 gives the *empirical logit* for the response variable,

$$\log_e \frac{\text{Proportion voting}}{\text{Proportion not voting}}$$

for each of the six combinations of categories of the explanatory variables.⁵⁸ For example,

$$\text{logit}(\text{voted}|\text{one-sided, weak preference}) = \log_e \frac{91/130}{39/130} = \log_e \frac{91}{39} = 0.847$$

Because the conditional proportions voting and not voting share the same denominator, the empirical logit can also be written as

$$\log_e \frac{\text{Number voting}}{\text{Number not voting}}$$

The empirical logits from Table 14.9 are graphed in Figure 14.13, much in the manner of profiles of cell means for a two-way ANOVA.⁵⁹ Perceived closeness of the election and intensity of preference appear to interact in affecting turnout: Turnout increases with increasing intensity of preference, but only if the election is perceived to be close. Those with medium or strong preference who perceive the election to be close are more likely to vote than those who

⁵⁸This calculation will fail if there is a 0 frequency in the table because, in this event, the proportion voting or not voting for some combination of explanatory-variable values will be 0. A simple remedy is to add 0.5 to each of the cell frequencies. Adding 0.5 to each count also serves to reduce the bias of the sample logit as an estimator of the corresponding population logit. See Cox and Snell (1989, pp. 31–32).

⁵⁹See Section 8.2.1.

perceive the election to be one-sided; this difference is greater among those with strong partisan preference than those with medium partisan preference.

The methods of this chapter are fully appropriate for tabular data. When, as in Table 14.9, the explanatory variables are qualitative or ordinal, it is natural to use logit or probit models that are analogous to ANOVA models. Treating perceived closeness of the election as the “row” factor and intensity of partisan preference as the “column” factor, for example, yields the model⁶⁰

$$\text{logit } \pi_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk} \quad (14.31)$$

where

- π_{jk} is the conditional probability of voting in combination of categories j of perceived closeness and k of preference (i.e., in cell jk of the explanatory-variable table),
- μ is the general level of turnout in the population,
- α_j is the main effect on turnout of membership in the j th category of perceived closeness,
- β_k is the main effect on turnout of membership in the k th category of preference, and
- γ_{jk} is the interaction effect on turnout of simultaneous membership in categories j of perceived closeness and k of preference.

When all the variables—explanatory as well as response—are discrete, their joint distribution defines a contingency table of frequency counts. It is natural to employ logit models that are analogous to ANOVA models to analyze contingency tables.

Under the usual sigma constraints, Equation 14.31 leads to deviation-coded regressors, as in ANOVA. Adapting the SS(\cdot) notation of Chapter 8,⁶¹ likelihood-ratio tests for main effects and interactions can then be constructed in close analogy to the incremental F -tests for the two-way ANOVA model. Residual deviances under several models for the *American Voter* data are shown in Table 14.10, and the analysis-of-deviance table for these data is given in Table 14.11. The log-likelihood-ratio statistic for testing H_0 : all $\gamma_{jk} = 0$, for example, is

$$\begin{aligned} G_0^2(\gamma|\alpha, \beta) &= G^2(\alpha, \beta) - G^2(\alpha, \beta, \gamma) \\ &= 1363.552 - 1356.434 \\ &= 7.118 \end{aligned}$$

with $6 - 4 = 2$ degrees of freedom, for which $p = .028$. The interaction discerned in Figure 14.13 is, therefore, statistically significant, but not overwhelmingly so.

⁶⁰This formulation assumes the sigma-constrained two-way ANOVA model discussed in Section 8.2.3. Alternatively, we can proceed by coding dummy regressors for the main effects and interaction, as in Section 8.2.2, as long as we restrict ourselves to Type II tests.

⁶¹In Chapter 8, we used SS(\cdot) to denote the *regression* sum of squares for a model including certain terms. Because the deviance is analogous to the *residual* sum of squares, we need to take differences of deviances in the opposite order. Again, as long as we confine ourselves to Type II tests, obeying the principle of marginality (i.e., the main-effect tests for $\alpha|\beta$ and $\beta|\alpha$ in Table 14.11), we can employ dummy regressors instead of deviation-coded regressors.

Table 14.10 Residual Deviances for Models Fit to the *American Voter* Data. Terms: α , Perceived Closeness; β , Intensity of Preference; γ , Closeness \times Preference Interaction

Model	Terms	k+1	Deviance: G^2
1	α, β, γ	6	1356.434
2	α, β	4	1363.552
3	α, γ	4	1368.042
4	β, γ	5	1368.554
5	α	2	1382.658
6	β	3	1371.838

NOTE: The column labeled $k+1$ gives the number of parameters in the model, including the constant μ .

Table 14.11 Analysis-of-Deviance Table for the *American Voter* Data

Source	Models Contrasted	df	G_0^2	p
Perceived closeness		1		
$\alpha \beta$	6–2		8.286	.0040
$\alpha \beta, \gamma$	4–1		12.120	.0005
Intensity of preference		2		
$\beta \alpha$	5–2		19.106	<.0001
$\beta \alpha, \gamma$	3–1		11.608	.0030
Closeness \times Preference		2		
$\gamma \alpha, \beta$	2–1		7.118	.028

NOTE: The table shows alternative likelihood-ratio tests for the main effects of perceived closeness of the election and intensity of partisan preference.

14.3.1 The Binomial Logit Model*

Although the models for dichotomous and polytomous response variables described in this chapter can be directly applied to tabular data, there is some advantage in reformulating these models to take direct account of the replication of combinations of explanatory-variable values. In analyzing dichotomous data, for example, we previously treated each observation individually, so that the dummy response variable Y_i takes on either the value 0 or the value 1.

Suppose, instead, that we group all the n_i observations that share the specific combination of explanatory-variable values $\mathbf{x}'_i = [x_{i1}, x_{i2}, \dots, x_{ik}]$. Let Y_i count the number of these observations that fall in the first of the two categories of the response variable; we arbitrarily term these observations *successes*. The count Y_i can take on any integer value between 0 and n_i . Let m denote the number of *distinct combinations* of the explanatory variables (e.g., $m = 6$ in Table 14.9 on page 408). To take this approach, the explanatory variables need not be qualitative, as long as they are discrete, so that there are replicated combinations of values of the explanatory variables.

As in our previous development of the dichotomous logit model, let π_i represent $\Pr(\text{success}|\mathbf{x}_i)$. Then the success count Y_i follows the binomial distribution:

$$\begin{aligned} p(y_i) &= \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \\ &= \binom{n_i}{y_i} \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{n_i} \end{aligned} \quad (14.32)$$

To distinguish grouped dichotomous data from ungrouped data, I will refer to the former as *binomial data* and the latter as *binary data*.⁶²

Suppose, next, that the dependence of the response probabilities π_i on the explanatory variables is well described by the logit model

$$\log_e \frac{\pi_i}{1 - \pi_i} = \mathbf{x}'_i \boldsymbol{\beta}$$

Substituting this model into Equation 14.32, the likelihood for the parameters is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^m \binom{n_i}{y_i} [\exp(\mathbf{x}'_i \boldsymbol{\beta})]^{y_i} \left(\frac{1}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})} \right)^{n_i}$$

Maximizing the likelihood leads to precisely the same maximum-likelihood estimates, coefficient standard errors, and statistical tests as the binary logit model of Section 14.1.5.⁶³ The binomial logit model nevertheless has the following advantages:

- Because we deal with m binomial observations rather than the larger $n = \sum_{i=1}^m n_i$ binary observations, computations for the binomial logit model are more efficient, especially when the n_i are large.
- The overall residual deviance for the binomial logit model, $-2 \log_e L(\mathbf{b})$, implicitly contrasts the model with a *saturated* model that has one parameter for each of the m combinations of explanatory-variable values (e.g., the full two-way “ANOVA” model with main effects and interactions fit in the previous section to the *American Voter* data). The saturated model necessarily recovers the m empirical logits perfectly and, consequently, has a likelihood of 1 and a log-likelihood of 0. The residual deviance for a less-than-saturated model, therefore, provides a likelihood-ratio test, on $m - k - 1$ degrees of freedom, of the hypothesis that the functional form of the model is correct.⁶⁴ In contrast, the residual deviance for the binary logit model cannot be used for a statistical test because the residual degrees of freedom $n - k - 1$ (unlike $m - k - 1$) grow as the sample size n grows.
- As long as the frequencies n_i are not very small, many diagnostics are much better behaved for the cells of the binomial logit model than for individual binary observations. For example, the individual components of the deviance for the binomial logit model,

$$G_i \equiv \pm \sqrt{-2 \left[Y_i \log_e \frac{n_i P_i}{Y_i} + (n_i - Y_i) \log_e \frac{n_i (1 - P_i)}{n_i - Y_i} \right]}$$

can be compared with the unit-normal distribution to locate outlying cells. Here $P_i = 1/[1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})]$ is the fitted probability of “success” for cell i , and, therefore,

⁶²Binary data can be thought of as a limiting case of binomial data, for which all $n_i = 1$.

⁶³See Exercise 14.13.

⁶⁴This test is analogous to the test for “lack of fit” in a linear model with discrete explanatory variables, described in Section 12.4.1.

$\hat{Y}_i = n_i P_i$ is the expected number of “successes” in this cell. The sign of G_i is selected to agree with that of the simple cell residual, $E_i = Y_i - \hat{Y}_i$.⁶⁵

Although the binary logit model can be applied to tables in which the response variable is dichotomous, it is also possible to use the equivalent binomial logit model; the binomial logit model is based on the frequency counts of “successes” and “failures” for each combination of explanatory-variable values. When it is applicable, the binomial logit model offers several advantages, including efficient computation, a test of the fit of the model based on its residual deviance, and better-behaved diagnostics.

Polytomous data can be handled in a similar manner, employing the multinomial distribution.⁶⁶ Consequently, all the logit and probit models discussed in this chapter have generalizations to data in which there are repeated observations for combinations of values of the explanatory variables. For example, the *multinomial logit model* generalizes the polytomous logit model (Equation 14.8 on page 376); indeed, even when it is fit to individual observations, the polytomous logit model is often called the “multinomial logit model” (as I previously mentioned).⁶⁷

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 14.1. Nonconstant error variance in the linear-probability model: Make a table showing the variance of the error $V(\varepsilon) = \pi(1 - \pi)$ for the following values of π :

.001, .01, .05, .1, .3, .5, .7, .9, .95, .99, .999

When is the heteroscedasticity problem serious?

Exercise 14.2. Show that using the cumulative rectangular distribution as $P(\cdot)$ in the general model

$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i)$$

produces the constrained linear-probability model. (See Section 14.1.2.)

Exercise 14.3. *Show that the slope of the logistic-regression curve, $\pi = 1/[1 + e^{-(\alpha + \beta X)}]$, can be written as $\beta\pi(1 - \pi)$. (Hint: Differentiate π with respect to X , and then substitute for expressions that equal π and $1 - \pi$.)

Exercise 14.4. Substitute first $y_i = 0$ and then $y_i = 1$ into the expression

⁶⁵Diagnostics for logit models and other generalized linear models are discussed in Section 15.4.

⁶⁶See Exercise 14.14.

⁶⁷As in the case of binary data, we can think of individual polytomous observations as multinomial observations in which all the total counts are $n_i = 1$.

$$p(y_i) \equiv \Pr(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

to show that this equation captures $p(0) = 1 - \pi_i$ and $p(1) = \pi_i$.

Exercise 14.5. *Show that, for the logit multiple-regression model,

$$\pi_i = \frac{1}{1 + \exp[-(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})]}$$

the probability that $Y_i = 0$ can be written as

$$1 - \pi_i = \frac{1}{1 + \exp(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}$$

Exercise 14.6. *Show that the maximized likelihood for the fitted logit model can be written as

$$\log_e L = \sum_{i=1}^n [y_i \log_e P_i + (1 - y_i) \log_e(1 - P_i)]$$

where

$$P_i = \frac{1}{1 + \exp[-(A + B_1 X_{i1} + \cdots + B_k X_{ik})]}$$

is the fitted probability that $Y_i = 1$. [Hint: Use $p(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$.]

Exercise 14.7. *Residual deviance in least-squares regression: The log-likelihood for the linear regression model with normal errors can be written as

$$\log_e L(\alpha, \beta_1, \dots, \beta_k, \sigma_\varepsilon^2) = -\frac{n}{2} \log_e(2\pi\sigma_\varepsilon^2) - \frac{\sum_{i=1}^n \varepsilon_i^2}{2\sigma_\varepsilon^2}$$

where $\varepsilon_i = Y_i - (\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})$ (see Section 9.3.3). Let l represent the maximized log-likelihood, treated as a function of the regression coefficients $\alpha, \beta_1, \dots, \beta_k$ but not of the error variance σ_ε^2 , which is regarded as a “nuisance parameter.” Let $l' = -(n/2) \log_e(2\pi\sigma_\varepsilon^2)$ represent the log-likelihood for a model that fits the data perfectly (i.e., for which all $\varepsilon_i = 0$). Then the residual deviance is defined as $-2\sigma_\varepsilon^2(l - l')$. Show that, by this definition, the residual deviance for the normal linear model is just the residual sum of squares. (For the logit model, there is no nuisance parameter, and $l' = 0$; the residual deviance for this model is, therefore, $-2 \log_e L$, as stated in the text. See Chapter 15 for further discussion of the deviance.)

Exercise 14.8. *Evaluate the information matrix for the logit model,

$$\mathcal{I}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log_e L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]$$

and show that the estimated asymptotic covariance matrix of the coefficients is

$$\widehat{\mathcal{V}}(\mathbf{b}) = \left[\sum_{i=1}^n \frac{\exp(-\mathbf{x}'_i \mathbf{b})}{[1 + \exp(-\mathbf{x}'_i \mathbf{b})]^2} \mathbf{x}_i \mathbf{x}'_i \right]^{-1}$$

Exercise 14.9. *Show that the maximum-likelihood estimator for the logit model can be written as

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}^*$$

where

$$\mathbf{y}^* \equiv \mathbf{X}\mathbf{b} + \mathbf{V}^{-1}(\mathbf{y} - \mathbf{p})$$

(Hint: Simply multiply out the equation.)

Exercise 14.10. *Show that the polytomous logit model (Equation 14.18, page 392) can be written in the form

$$\log_e \frac{\pi_{ij}}{\pi_{im}} = \gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik} \quad \text{for } j = 1, \dots, m-1$$

Exercise 14.11. *Derive the estimating equations (Equations 14.22 on page 398) and the information matrix (Equations 14.23 and 14.24) for the polytomous logit model.

Exercise 14.12. Independence from irrelevant alternatives: In the polytomous logit model discussed in Section 14.2.1, the logit for a particular pair of categories depends on the coefficients for those categories but not on those for other categories in the model. Show that this is the case. (Hint: See Equation 14.2.1.) In the context of a discrete-choice model (e.g., Greene, 2003, chap. 21; or Alvarez & Nagler, 1998), this property can be interpreted to mean that the relative odds for a pair of categories is independent of the other categories in the choice set. Why is this often an implausible assumption? (Hint: Consider a multiparty election in a jurisdiction, such as Canada or the United Kingdom, where some parties field candidates in only part of the country, or what happens to the electoral map when a new party is formed.) For this reason, models such as the polytomous probit model that *do not* assume independence from irrelevant alternatives are sometimes preferred.

Exercise 14.13. *Derive the maximum-likelihood estimating equations for the binomial logit model. Show that this model produces the same estimated coefficients as the dichotomous (binary) logit model of Section 14.1. (Hint: Compare the log-likelihood for the binomial model with the log-likelihood for the binary model; by separating individual observations sharing a common set of X -values, show that the former log-likelihood is equal to the latter, except for a constant factor. This constant is irrelevant because it does not influence the maximum-likelihood estimator; moreover, the constant disappears in likelihood-ratio tests.)

Exercise 14.14. *Use the multinomial distribution (see online Appendix D) to specify a polytomous logit model for discrete explanatory variables (analogous to the binomial logit model), where combinations of explanatory-variable values are replicated. Derive the likelihood under the model and the maximum-likelihood estimating equations.

Summary

- It is problematic to apply least-squares linear regression to a dichotomous response variable: The errors cannot be normally distributed and cannot have constant variance. Even more fundamentally, the linear specification does not confine the probability for the response to the unit interval.

- More adequate specifications transform the linear predictor $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ smoothly to the unit interval, using a cumulative probability distribution function $P(\cdot)$. Two such specifications are the probit and the logit models, which use the normal and logistic CDFs, respectively. Although these models are very similar, the logit model is simpler to interpret because it can be written as a linear model for the log odds,

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

or, exponentiating the coefficients, as a multiplicative model for the odds,

$$\frac{\pi_i}{1 - \pi_i} = e^\alpha (e^{\beta_1})^{X_{i1}} \cdots (e^{\beta_k})^{X_{ik}}$$

- The dichotomous logit model can be fit to data by the method of maximum likelihood. Wald tests and likelihood-ratio tests for the coefficients of the model parallel t -tests and incremental F -tests for the general linear model. The residual deviance for the model, defined as $G^2 = -2 \times$ the maximized log-likelihood, is analogous to the residual sum of squares for a linear model.
- Several approaches can be taken to modeling polytomous data, including:
 - modeling the polytomy directly using a logit model based on the multivariate logistic distribution,
 - constructing a set of $m - 1$ nested dichotomies to represent the m categories of the polytomy, and
 - fitting the proportional-odds model to a polytomous response variable with ordered categories.
- When all the variables—explanatory as well as response—are discrete, their joint distribution defines a contingency table of frequency counts. It is natural to employ logit models that are analogous to ANOVA models to analyze contingency tables. Although the binary logit model can be applied to tables in which the response variable is dichotomous, it is also possible to use the equivalent binomial logit model; the binomial logit model is based on the frequency counts of “successes” and “failures” for each combination of explanatory-variable values. When it is applicable, the binomial logit model offers several advantages, including efficient computation, a test of the fit of the model based on its residual deviance, and better-behaved diagnostics. There are analogous logit and probit models, such as the multinomial logit model, for polytomous responses.

Recommended Reading

The topics introduced in this chapter could easily be expanded to fill several books, and there is a large literature—both in journals and texts—dealing with logit and related models for categorical response variables and with the analysis of contingency tables.⁶⁸

⁶⁸Also see the references on generalized linear models given at the end of the next chapter, which briefly describes log-linear models for contingency tables.

- Agresti (2012) presents an excellent and comprehensive overview of statistical methods for qualitative data. The emphasis is on logit and loglinear models for contingency tables, but there is some consideration of logistic regression models and other topics. Also see Agresti (2007) for a briefer and lower-level treatment of much of this material.
- Fienberg's (1980) widely read text on the analysis of contingency tables provides an accessible and lucid introduction to loglinear models and related subjects, such as logit models and models for ordered categories.
- The second edition of Cox and Snell's (1989) classic text concentrates on logit models for dichotomous data but also includes some discussion of polytomous nominal and ordinal data.
- Collett (2003) also focuses on the binary and binomial logit models. The book is noteworthy for its extensive review of diagnostic methods for logit models.
- Greene (2003, chap. 21) includes a broad treatment of models for categorical responses from the point of view of “discrete choice models” in econometrics.
- Long (1997) and Powers and Xie (2008) both present high-quality, accessible expositions for social scientists of statistical models for categorical data.

15

Generalized Linear Models

Due originally to Nelder and Wedderburn (1972), generalized linear models are a remarkable synthesis and extension of familiar regression models such as the linear models described in Part II of this text and the logit and probit models described in the preceding chapter. The current chapter begins with a consideration of the general structure and range of application of generalized linear models; proceeds to examine in greater detail generalized linear models for count data, including contingency tables; briefly sketches the statistical theory underlying generalized linear models; describes the extension of regression diagnostics to generalized linear models; and concludes with a discussion of design-based statistical inference for complex sample surveys.

The unstarred sections of this chapter are perhaps more difficult than the unstarred material in preceding chapters. Generalized linear models have become so central to effective statistical data analysis, however, that it is worth the additional effort required to acquire at least a basic understanding of the subject.

15.1 The Structure of Generalized Linear Models

A *generalized linear model* (or GLM¹) consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i (for the i th of n independently sampled observations), given the values of the explanatory variables in the model. In Nelder and Wedderburn's original formulation, the distribution of Y_i is a member of an *exponential family*, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions. Subsequent work, however, has extended GLMs to multivariate exponential families (such as the multinomial distribution), to certain nonexponential families (such as the two-parameter negative-binomial distribution), and to some situations in which the distribution of Y_i is not specified completely. Most of these ideas are developed later in the chapter.
2. A *linear predictor*—that is, a linear function of regressors

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

As in the linear model, and in the logit and probit models of Chapter 14, the regressors X_{ij} are prespecified functions of the explanatory variables and therefore may include

¹Some authors use the acronym “GLM” to refer to the “*general linear model*”—that is, the linear regression model with normal errors described in Part II of the text—and instead employ “GLIM” to denote *generalized* linear models (which is also the name of a computer program used to fit GLMs).

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	$\eta_i^{1/2}$
Logit	$\log_e \frac{\mu_i}{1 - \mu_i}$	$\frac{1}{1 + e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1 - \mu_i)]$	$1 - \exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response, η_i is the linear predictor, and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

quantitative explanatory variables, transformations of quantitative explanatory variables, polynomial or regression-spline regressors, dummy regressors, interactions, and so on. Indeed, one of the advantages of GLMs is that the structure of the linear predictor is the familiar structure of a linear model.

3. A smooth and invertible linearizing *link function* $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

Because the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or as a nonlinear regression model for the response. The inverse link $g^{-1}(\cdot)$ is also called the *mean function*. Commonly employed link functions and their inverses are shown in Table 15.1. Of these, the *identity link* simply returns its argument unaltered, $\eta_i = g(\mu_i) = \mu_i$, and thus $\mu_i = g^{-1}(\eta_i) = \eta_i$.

The last four link functions in Table 15.1 are for binomial data, where Y_i represents the observed proportion of “successes” in n_i independent binary trials; thus, Y_i can take on any of the $n_i + 1$ values $0, 1/n_i, 2/n_i, \dots, (n_i - 1)/n_i, 1$. Recall from Section 14.3.1 that binomial data also encompass binary data, where all the observations represent $n_i = 1$ trial, and consequently, Y_i is either 0 or 1. The expectation of the response $\mu_i = E(Y_i)$ is then the probability of success, which we symbolized by π_i in the previous chapter. The logit, probit, log-log, and complementary log-log links are graphed in Figure 15.1. In contrast to the logit and probit links (which, as we noted previously, are nearly indistinguishable once the variances of the underlying normal and logistic distributions are equated), the log-log and complementary log-log links approach the asymptotes of 0 and 1 asymmetrically.²

²Because the log-log link can be obtained from the complementary log-log link by exchanging the definitions of “success” and “failure,” it is common for statistical software to provide only one of the two—typically, the complementary log-log link.

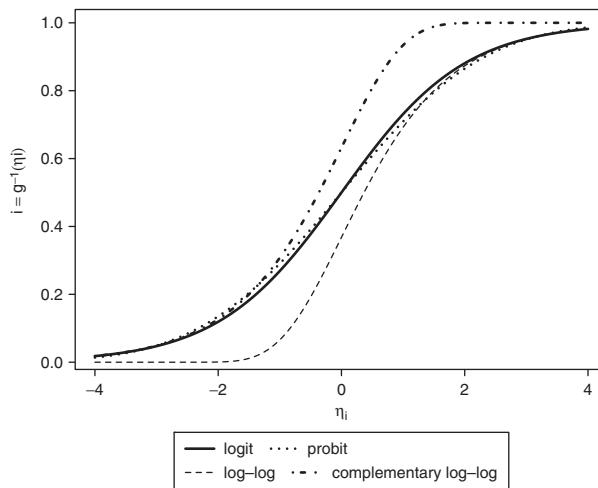


Figure 15.1 Logit, probit, log-log, and complementary log-log links for binomial data. The variances of the normal and logistic distributions have been equated to facilitate the comparison of the logit and probit links [by graphing the cumulative distribution function of $N(0, \pi^2/3)$ for the probit link].

Beyond the general desire to select a link function that renders the regression of Y on the X s linear, a promising link will remove restrictions on the range of the expected response. This is a familiar idea from the logit and probit models discussed in Chapter 14, where the object was to model the probability of “success,” represented by μ_i in our current more general notation. As a probability, μ_i is confined to the unit interval $[0,1]$. The logit, probit, log-log, and complementary log-log links map this interval to the entire real line, from $-\infty$ to $+\infty$. Similarly, if the response Y is a count, taking on only nonnegative integer values, $0, 1, 2, \dots$, and consequently μ_i is an expected count, which (though not necessarily an integer) is also nonnegative, the log link maps μ_i to the whole real line. This is not to say that the choice of link function is entirely determined by the range of the response variable, just that the link should behave reasonably in relation to the range of the response.

A generalized linear model (or GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i (for the i th of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor—that is, a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i [say, $v(\mu_i)$] and, possibly, a *dispersion parameter* ϕ . The variance functions for the commonly used exponential families appear in Table 15.2. The conditional variance of the response in the Gaussian family is a constant, ϕ , which is simply alternative notation for what we previously termed the error variance, σ_ε^2 . In the binomial and Poisson families, the dispersion parameter is set to the fixed value $\phi = 1$.

Table 15.2 also shows the range of variation of the response variable in each family and the so-called *canonical* (or “natural”) *link function* associated with each family. The canonical link simplifies the GLM,³ but other link functions may be used as well. Indeed, one of the strengths of the GLM paradigm—in contrast to transformations of the response variable in linear regression—is that the choice of linearizing transformation is partly separated from the distribution of the response, and the same transformation does not have to both normalize the distribution of Y and make its regression on the X s linear.⁴ The specific links that may be used vary from one family to another and also—to a certain extent—from one software implementation of GLMs to another. For example, it would not be promising to use the identity, log, inverse, inverse-square, or square-root links with binomial data, nor would it be sensible to use the logit, probit, log-log, or complementary log-log link with nonbinomial data.

Table 15.2 Canonical Link, Response Range, and Conditional Variance Function for Exponential Families

Family	Canonical Link	Range of Y_i	$V(Y_i \eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$0, 1, \dots, n_i$	$\frac{\mu_i(1 - \mu_i)}{n_i}$
Poisson	Log	$0, 1, 2, \dots$	μ_i
Gamma	Inverse	$(0, \infty)$	$\phi \mu_i^2$
Inverse-Gaussian	Inverse-square	$(0, \infty)$	$\phi \mu_i^3$

NOTE: ϕ is the dispersion parameter, η_i is the linear predictor, and μ_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

³This point is pursued in Section 15.3.

⁴There is also this more subtle difference: When we transform Y and regress the transformed response on the X s, we are modeling the expectation of the transformed response,

$$E[g(Y_i)] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

In a GLM, in contrast, we model the transformed expectation of the response,

$$g[E(Y_i)] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

While similar in spirit, this is not quite the same thing when (as is true except for the identity link) the link function $g(\cdot)$ is nonlinear. It is also the case that we can address nonlinearity in the relationship of Y to the X s in other ways than through transformation of the response or the expected response—for example, by transforming the X s or by employing polynomial regressors or regression splines.

I assume that the reader is generally familiar with the Gaussian and binomial families and simply give their distributions here for reference. The Poisson, gamma, and inverse-Gaussian distributions are perhaps less familiar, and so I provide some more detail:⁵

- The Gaussian distribution with mean μ and variance σ^2 has density function

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right] \quad (15.1)$$

- The binomial distribution for the proportion Y of successes in n independent binary trials with probability of success μ has probability function

$$p(y) = \binom{n}{ny} \mu^{ny} (1-\mu)^{n(1-y)} \quad (15.2)$$

Here, ny is the observed *number* of successes in the n trials, and $n(1-y)$ is the number of failures, and

$$\binom{n}{ny} = \frac{n!}{(ny)![n(1-y)]!}$$

is the binomial coefficient.

- The Poisson distributions are a discrete family with probability function indexed by the *rate parameter* $\mu > 0$:

$$p(y) = \mu^y \times \frac{e^{-\mu}}{y!} \text{ for } y = 0, 1, 2, \dots$$

The expectation and variance of a Poisson random variable are both equal to μ . Poisson distributions for several values of the parameter μ are graphed in Figure 15.2. As we will see in Section 15.2, the Poisson distribution is useful for modeling count data. As μ increases, the Poisson distribution grows more symmetric and is eventually well approximated by a normal distribution.

- The gamma distributions are a continuous family with probability-density function indexed by the *scale parameter* $\omega > 0$ and *shape parameter* $\psi > 0$:

$$p(y) = \left(\frac{y}{\omega}\right)^{\psi-1} \times \frac{\exp\left(-\frac{y}{\omega}\right)}{\omega\Gamma(\psi)} \text{ for } y > 0 \quad (15.3)$$

where $\Gamma(\cdot)$ is the gamma function.⁶ The expectation and variance of the gamma distribution are, respectively, $E(Y) = \omega\psi$ and $V(Y) = \omega^2\psi$. In the context of a generalized linear model, where, for the gamma family, $V(Y) = \phi\mu^2$ (recall Table 15.2 on page 421), the dispersion parameter is simply the inverse of the shape parameter, $\phi = 1/\psi$. As the

⁵The various distributions used in this chapter are described in a general context in online Appendix D on probability and estimation.

⁶*The gamma function is defined as

$$\Gamma(x) = \int_0^\infty e^{-z} z^{x-1} dz$$

and may be thought of as a continuous generalization of the factorial function in that when x is a nonnegative integer, $x! = \Gamma(x+1)$.

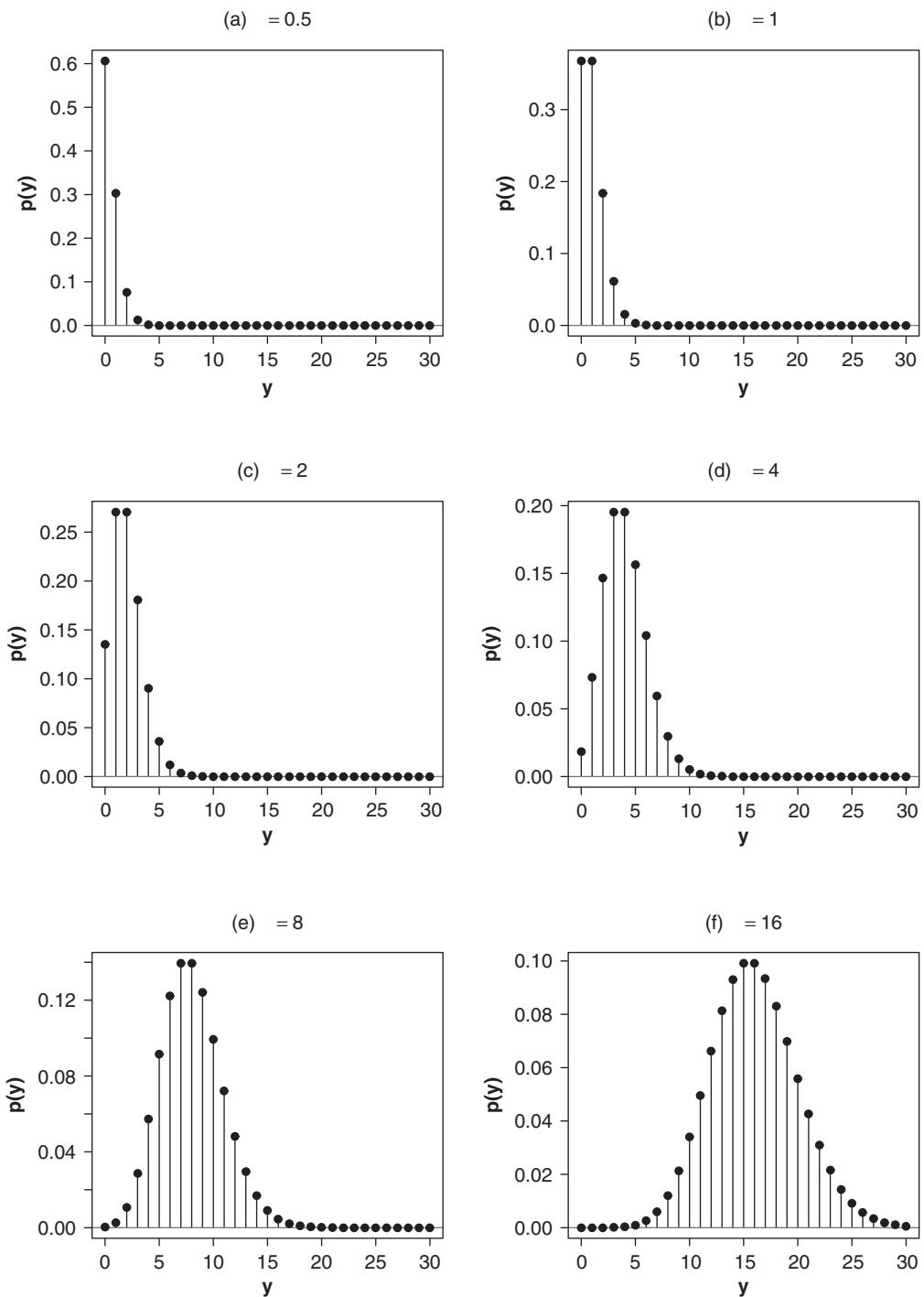


Figure 15.2 Poisson distributions for various values of the rate parameter μ .

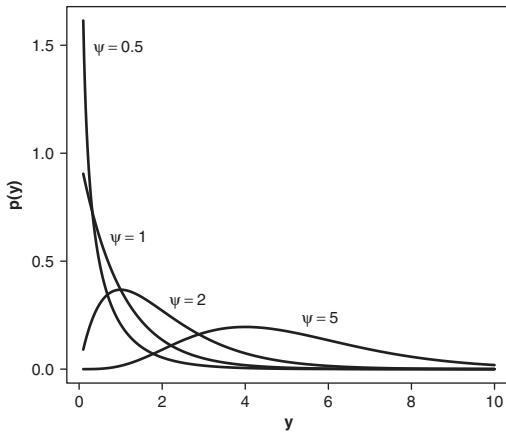


Figure 15.3 Several gamma distributions for scale $\omega = 1$ and various values of the shape parameter ψ .

names of the parameters suggest, the scale parameter in the gamma family influences the spread (and, incidentally, the location) but not the shape of the distribution, while the shape parameter controls the skewness of the distribution. Figure 15.3 shows gamma distributions for scale $\omega = 1$ and several values of the shape parameter ψ . (Altering the scale parameter would change only the labeling of the horizontal axis in the graph.) As the shape parameter gets larger, the distribution grows more symmetric. The gamma distribution is useful for modeling a positive continuous response variable, where the conditional variance of the response grows with its mean but where the *coefficient of variation* of the response, $SD(Y)/\mu$, is constant.

- The inverse-Gaussian distributions are another continuous family indexed by two parameters, μ and λ , with density function

$$p(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2y\mu^2}\right] \text{ for } y > 0$$

The expectation and variance of Y are $E(Y) = \mu$ and $V(Y) = \mu^3/\lambda$. In the context of a GLM, where, for the inverse-Gaussian family, $V(Y) = \phi\mu^3$ (as recorded in Table 15.2 on page 421), λ is the inverse of the dispersion parameter ϕ . Like the gamma distribution, therefore, the variance of the inverse-Gaussian distribution increases with its mean but at a more rapid rate. Skewness also increases with the value of μ and decreases with λ . Figure 15.4 shows several inverse-Gaussian distributions.

A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i and, possibly, a dispersion parameter ϕ . In addition to the familiar Gaussian and binomial families (the latter for proportions), the Poisson family is useful for modeling count data, and the gamma and inverse-Gaussian families for modeling positive continuous data, where the conditional variance of Y increases with its expectation.

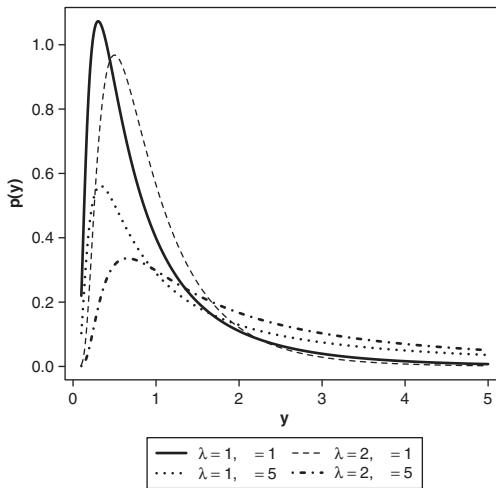


Figure 15.4 Inverse-Gaussian distributions for several combinations of values of the mean μ and inverse-dispersion λ .

15.1.1 Estimating and Testing GLMs

GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic (i.e., large-sample) standard errors of the coefficients.⁷ To test the null hypothesis $H_0: \beta_j = \beta_j^{(0)}$, we can compute the Wald statistic $Z_0 = (B_j - \beta_j^{(0)})/\text{SE}(B_j)$, where $\text{SE}(B_j)$ is the asymptotic standard error of the estimated coefficient B_j . Under the null hypothesis, Z_0 follows a standard normal distribution.⁸

As explained, some of the exponential families on which GLMs are based include an unknown dispersion parameter ϕ . Although this parameter can, in principle, be estimated by maximum likelihood as well, it is more common to use a “method of moments” estimator, which I will denote $\tilde{\phi}$.⁹

As is familiar from the preceding chapter on logit and probit models, the ANOVA for linear models has a close analog in the *analysis of deviance* for GLMs. In the current more general context, the *residual deviance* for a GLM is

$$D_m \equiv 2(\log_e L_s - \log_e L_m)$$

where L_m is the maximized likelihood under the model in question and L_s is the maximized likelihood under a *saturated model*, which dedicates one parameter to each observation and

⁷Details are provided in Section 15.3.2. The method of maximum likelihood is introduced in online Appendix D on probability and estimation.

⁸For models with an estimated dispersion parameter, we can instead compare the Wald statistic to the t -distribution with $n - k - 1$ degrees of freedom. Wald chi-square and F -tests of more general linear hypotheses are described in Section 15.3.2.

⁹Again, see Section 15.3.2.

consequently fits the data as closely as possible. The residual deviance is analogous to (and, indeed, is a generalization of) the residual sum of squares for a linear model.

In GLMs for which the dispersion parameter is fixed to 1 (i.e., binomial and Poisson GLMs), the likelihood-ratio test statistic is simply the difference in the residual deviances for nested models. Suppose that Model 0, with $k_0 + 1$ coefficients, is nested within Model 1, with $k_1 + 1$ coefficients (where, then, $k_0 < k_1$); most commonly, Model 0 would simply omit some of the regressors in Model 1. We test the null hypothesis that the restrictions on Model 1 represented by Model 0 are correct by computing the likelihood-ratio test statistic

$$G_0^2 = D_0 - D_1$$

Under the hypothesis, G_0^2 is asymptotically distributed as chi-square with $k_1 - k_0$ degrees of freedom.

Likelihood-ratio tests can be turned around to provide confidence intervals for coefficients; as mentioned in Section 14.1.4 in connection with logit and probit models, tests and intervals based on the likelihood-ratio statistic tend to be more reliable than those based on the Wald statistic. For example, the 95% confidence interval for β_j includes all values β'_j for which the hypothesis $H_0: \beta_j = \beta'_j$ is acceptable at the .05 level—that is, all values of β'_j for which $2(\log_e L_1 - \log_e L_0) \leq \chi_{.05,1}^2 = 3.84$, where $\log_e L_1$ is the maximized log-likelihood for the full model, and $\log_e L_0$ is the maximized log-likelihood for a model in which β_j is constrained to the value β'_j . This procedure is computationally intensive because it required “profiling” the likelihood—refitting the model for various fixed values β'_j of β_j .

For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an F -test,

$$F_0 = \frac{\frac{D_0 - D_1}{k_1 - k_0}}{\tilde{\phi}}$$

where the estimated dispersion $\tilde{\phi}$, analogous to the estimated error variance for a linear model, is taken from the *largest* model fit to the data (which is not necessarily Model 1). If the largest model has $k + 1$ coefficients, then, under the hypothesis that the restrictions on Model 1 represented by Model 0 are correct, F_0 follows an F -distribution with $k_1 - k_0$ and $n - k - 1$ degrees of freedom. Applied to a Gaussian GLM, this is simply the familiar incremental F -test. The residual deviance divided by the estimated dispersion, $D^* \equiv D/\tilde{\phi}$, is called the *scaled deviance*.¹⁰

As we did for logit and probit models,¹¹ we can base a GLM analog of the squared multiple correlation on the residual deviance: Let D_0 be the residual deviance for the model including only the regression constant α —termed the *null deviance*—and D_1 the residual deviance for the model in question. Then,

$$R^2 \equiv 1 - \frac{D_1}{D_0}$$

represents the proportion of the null deviance accounted for by the model.

¹⁰Usage is not entirely uniform here, and either the residual deviance or the scaled deviance is often simply termed “the deviance.”

¹¹See Section 14.1.4.

GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients.

The ANOVA for linear models has an analog in the analysis of deviance for GLMs. The residual deviance for a GLM is $D_m = 2(\log_e L_s - \log_e L_m)$, where L_m is the maximized likelihood under the model in question and L_s is the maximized likelihood under a saturated model. The residual deviance is analogous to the residual sum of squares for a linear model.

In GLMs for which the dispersion parameter is fixed to 1 (binomial and Poisson GLMs), the likelihood-ratio test statistic is the difference in the residual deviances for nested models and is asymptotically distributed as chi-square under the null hypothesis. For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an incremental F -test.

15.2 Generalized Linear Models for Counts

The basic GLM for count data is the Poisson model with the log link. Consider, by way of example, Michael Ornstein's data on interlocking directorates among 248 dominant Canadian firms, previously discussed in Chapters 3 and 4. The number of interlocks for each firm is the number of ties that a firm maintained by virtue of its board members and top executives also serving as board members or executives of other firms in the data set. Ornstein was interested in the regression of the number of interlocks on other characteristics of the firms—specifically, on their assets (measured in billions of dollars), nation of control (Canada, the United States, the United Kingdom, or another country), and the principal sector of operation of the firm (10 categories, including banking, other financial institutions, heavy manufacturing, etc.).

Examining the distribution of number of interlocks (Figure 15.5) reveals that the variable is highly positively skewed and that there are many 0 counts. Although the conditional distribution of interlocks given the explanatory variables could differ from its marginal distribution, the extent to which the marginal distribution of interlocks departs from symmetry bodes ill for least-squares regression. Moreover, no transformation will spread out the 0s.¹²

The results of the Poisson regression of number of interlocks on assets, nation of control, and sector are summarized in Table 15.3. I set the *United States* as the baseline category for nation of control, and *Construction* as the baseline category for sector—these are the categories with the smallest fitted numbers of interlocks controlling for the other variables in the regression, and the dummy-regressor coefficients are therefore all positive.

¹²Ornstein (1976) in fact performed a linear least-squares regression for these data, although one with a slightly different specification from that given here. He cannot be faulted for having done so, however, inasmuch as Poisson regression models—and, with the exception of loglinear models for contingency tables, other specialized models for counts—were not typically in sociologists' statistical toolkit at the time.

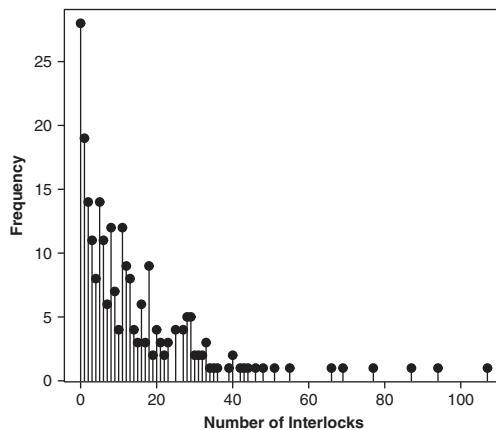


Figure 15.5 The distribution of number of interlocks among 248 dominant Canadian corporations.

Table 15.3 Estimated Coefficients for the Poisson Regression of Number of Interlocks on Assets, Nation of Control, and Sector, for Ornstein's Canadian Interlocking-Directorate Data

Coefficient	Estimate	Standard Error	e^β
Constant	0.8791	0.2101	—
Assets	0.02085	0.00120	1.021
<i>Nation of Control (baseline: United States)</i>			
Canada	0.8259	0.0490	2.284
Other	0.6627	0.0755	1.940
United Kingdom	0.2488	0.0919	1.282
<i>Sector (baseline: construction)</i>			
Wood and paper	1.331	0.213	3.785
Transport	1.297	0.214	3.658
Other financial	1.297	0.211	3.658
Mining, metals	1.241	0.209	3.459
Holding companies	0.8280	0.2329	2.289
Merchandising	0.7973	0.2182	2.220
Heavy manufacturing	0.6722	0.2133	1.959
Agriculture, food, light industry	0.6196	0.2120	1.858
Banking	0.2104	0.2537	1.234

The residual deviance for this model is $D(\text{Assets}, \text{Nation}, \text{Sector}) = 1887.402$ on $n - k - 1 = 248 - 13 - 1 = 234$ degrees of freedom. Deleting each explanatory variable in turn from the model produces the following residual deviances and degrees of freedom:

<i>Explanatory Variables</i>	<i>Residual Deviance</i>	<i>df</i>
Nation, Sector	2278.298	235
Assets, Sector	2216.345	237
Assets, Nation	2248.861	243

Taking differences between these deviances and the residual deviance for the full model yields the following analysis-of-deviance table:

<i>Source</i>	G_0^2	<i>df</i>	<i>p</i>
Assets	390.90	1	<<.0001
Nation	328.94	3	<<.0001
Sector	361.46	9	<<.0001

All the terms in the model are therefore highly statistically significant.

Because the model uses the log link, we can interpret the exponentiated coefficients (i.e., the e^{B_j} , also shown in Table 15.3) as multiplicative effects on the expected number of interlocks. Thus, for example, holding nation of control and sector constant, increasing assets by 1 billion dollars (the unit of the assets variable) multiplies the estimated expected number of interlocks by $e^{0.02085} = 1.021$ —that is, an increase of just over 2%. Similarly, the estimated expected number of interlocks is $e^{0.8259} = 2.284$ times as high in a Canadian-controlled firm as in a comparable U.S.-controlled firm.

As mentioned, the residual deviance for the full model fit to Ornstein’s data is $D_1 = 1887.402$; the deviance for a model fitting only the constant (i.e., the null deviance) is $D_0 = 3737.010$. Consequently, $R^2 = 1 - 1887.402/3737.010 = .495$, revealing that the model accounts for nearly half the deviance in number of interlocks.

The Poisson-regression model is a nonlinear model for the expected response, and I therefore find it generally simpler to interpret the model graphically using effect displays than to examine the estimated coefficients directly. The principles of construction of effect displays for GLMs are essentially the same as for linear models and for logit and probit models:¹³ We usually construct one display for each high-order term in the model, allowing the explanatory variables in that term to range over their values, while holding other explanatory variables in the model to typical values. In a GLM, it is advantageous to plot effects on the scale of the estimated linear predictor, $\hat{\eta}$, a procedure that preserves the linear structure of the model. In a Poisson model with the log link, the linear predictor is on the log-count scale. We can, however, make the display easier to interpret by relabeling the vertical axis in the scale of the expected response, $\hat{\mu}$, most informatively by providing a second vertical axis on the right-hand side of the plot. For a Poisson model, the expected response is a count.

Effect displays for the terms in Ornstein’s Poisson regression are shown in Figure 15.6. This model has an especially simple structure because each high-order term is a main effect—there are no interactions in the model. The effect display for assets shows a one-dimensional

¹³See Section 15.3.4 for details.

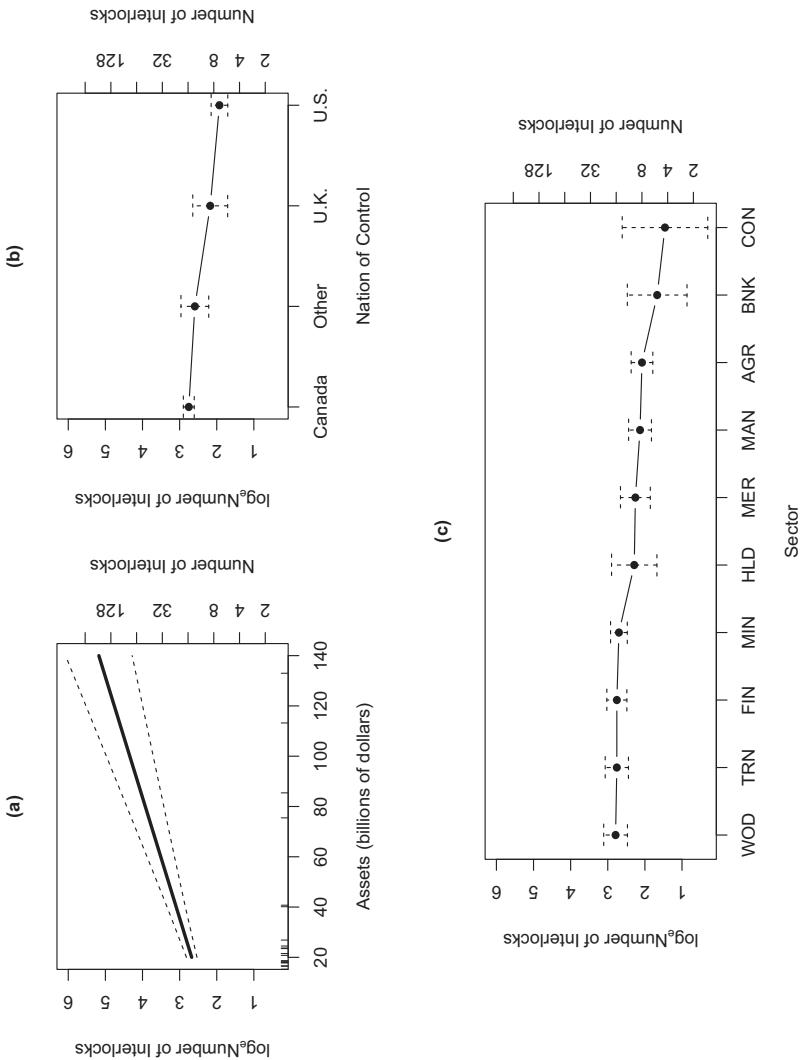


Figure 15.6 Effect displays for (a) assets, (b) nation of control, and (c) sector in the Poisson regression for Ornstein's interlocking-directorate data. The broken lines and error bars give 95% confidence intervals around the fitted effects (computed using the quasi-Poisson model described below). A "rug-plot" at the bottom of panel (a) shows the distribution of assets.

scatterplot (a “rug-plot”) for this variable at the bottom of the graph, revealing that the distribution of assets is highly skewed to the right. Skewness produces some high-leverage observations and suggests the possibility of a nonlinear effect for assets, points that I pursue later in the chapter.¹⁴

15.2.1 Models for Overdispersed Count Data

The residual deviance for the Poisson-regression model fit to the interlocking-directorate data, $D = 1887.4$, is much larger than the 234 residual degrees of freedom for the model. If the Poisson model fits the data reasonably, we would expect the residual deviance to be roughly equal to the residual degrees of freedom.¹⁵ That the residual deviance is so large suggests that the conditional variation of the expected number of interlocks exceeds the variation of a Poisson-distributed variable, for which the variance equals the mean. This common occurrence in the analysis of count data is termed *overdispersion*.¹⁶ Indeed, overdispersion is so common in regression models for count data, and its consequences are potentially so severe, that models such as the quasi-Poisson and negative-binomial GLMs discussed in this section should be employed as a matter of course in preference to the Poisson GLM.

The Quasi-Poisson Model

A simple remedy for overdispersed count data is to introduce a dispersion parameter into the Poisson model, so that the conditional variance of the response is now $V(Y_i|\eta_i) = \phi\mu_i$. If $\phi > 1$, therefore, the conditional variance of Y increases more rapidly than its mean. There is no exponential family corresponding to this specification, and the resulting GLM does not imply a specific probability distribution for the response variable. Rather, the model specifies the conditional mean and variance of Y_i directly. Because the model does not give a probability distribution for Y_i , it cannot be estimated by maximum likelihood. Nevertheless, the usual procedure for maximum-likelihood estimation of a GLM yields the so-called *quasi-likelihood* estimators of the regression coefficients, which share many of the properties of maximum-likelihood estimators.¹⁷

As it turns out, the quasi-likelihood estimates of the regression coefficients are identical to the maximum-likelihood (ML) estimates for the Poisson model. The estimated coefficient standard errors differ, however: If $\tilde{\phi}$ is the estimated dispersion for the model, then the coefficient standard errors for the *quasi-Poisson model* are $\tilde{\phi}^{1/2}$ times those for the Poisson model. In the event of overdispersion, therefore, where $\tilde{\phi} > 1$, the effect of introducing a dispersion parameter and obtaining quasi-likelihood estimates is (realistically) to inflate the coefficient standard

¹⁴See Section 15.4 on diagnostics for GLMs.

¹⁵That is, the ratio of the residual deviance to degrees of freedom can be taken as an estimate of the dispersion parameter ϕ , which, in a Poisson model, is fixed to 1. This deviance-based estimator of the dispersion can perform poorly, however. A generally preferable “method of moments” estimator is given in Section 15.3.

¹⁶Although it is much less common, it is also possible for count data to be *underdispersed*—that is, for the conditional variation of the response to be *less than* the mean. The remedy for underdispersed count data is the same as for overdispersed data; for example, we can fit a quasi-Poisson model with a dispersion parameter, as described immediately below.

¹⁷See Section 15.3.2.

errors. Likewise, F -tests for terms in the model will reflect the estimated dispersion parameter, producing smaller test statistics and larger p -values.

As explained in the following section, we use a method-of-moments estimator for the dispersion parameter. In the quasi-Poisson model, the dispersion estimator takes the form

$$\tilde{\phi} = \frac{1}{n - k - 1} \sum \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ is the fitted expectation of Y_i . Applied to Ornstein's interlocking-directorate regression, for example, we get $\tilde{\phi} = 7.9435$, and, therefore, the standard errors of the regression coefficients for the Poisson model in Table 15.3 are each multiplied by $\sqrt{7.9435} = 2.818$.

I note in passing that there is a similar *quasi-binomial model* for overdispersed proportions, replacing the fixed dispersion parameter of 1 in the binomial distribution with a dispersion parameter ϕ to be estimated from the data. Overdispersed binomial data can arise, for example, when individuals who share the same values of the explanatory variables nevertheless differ in their probability μ of success, a situation that is termed *unmodeled heterogeneity*. Similarly, overdispersion can occur when binomial "trials" are not independent, as required by the binomial distribution—for example, when the trials for each binomial observation are for related individuals, such as members of a family.

The Negative-Binomial Model

There are several routes to models for counts based on the negative-binomial distribution (see, e.g., Long, 1997, Section 8.3; McCullagh & Nelder, 1989, Section 6.2.3). One approach (following McCullagh & Nelder, 1989, p. 233) is to adopt a Poisson model for the count Y_i but to suppose that the expected count μ_i^* is itself an unobservable random variable that is gamma-distributed with mean μ_i and constant scale parameter ω (implying that the gamma shape parameter is $\psi_i = \mu_i/\omega$ ¹⁸). Then the observed count Y_i follows a *negative-binomial distribution*,¹⁹

$$p(y_i) = \frac{\Gamma(y_i + \omega)}{y! \Gamma(\omega)} \times \frac{\mu_i^{y_i} \omega^\omega}{(\mu_i + \omega)^{\mu_i + \omega}} \quad (15.4)$$

with expected value $E(Y_i) = \mu_i$ and variance $V(Y_i) = \mu_i + \mu_i^2/\omega$. Unless the parameter ω is large, therefore, the variance of Y increases more rapidly with the mean than the variance of a Poisson variable. Making the expected value of Y_i a random variable incorporates additional variation among observed counts for observations that share the same values of the explanatory variables and consequently have the same linear predictor η_i .

With the gamma scale parameter ω fixed to a known value, the negative-binomial distribution is an exponential family (in the sense of Equation 15.15 in Section 15.3.1), and a GLM based on this distribution can be fit by iterated weighted least squares (as developed in the next section). If instead—and as is typically the case—the value of ω is unknown and must therefore be estimated from the data, standard methods for GLMs based on exponential families do not apply. We can, however, obtain estimates of both the regression coefficients and ω by the method of maximum likelihood. Applied to Ornstein's interlocking-directorate regression and

¹⁸See Equation 15.3 on page 422.

¹⁹A simpler form of the negative-binomial distribution is given in online Appendix D on probability and estimation.

using the log link, the negative-binomial GLM produces results very similar to those of the quasi-Poisson model (as the reader may wish to verify). The estimated scale parameter for the negative-binomial model is $\hat{\omega} = 1.312$, with standard error $SE(\hat{\omega}) = 0.143$; we have, therefore, strong evidence that the conditional variance of the number of interlocks increases more rapidly than its expected value.²⁰

Zero-Inflated Poisson Regression

A particular kind of overdispersion obtains when there are more 0s in the data than is consistent with a Poisson (or negative-binomial) distribution, a situation that can arise when only certain members of the population are “at risk” of a nonzero count. Imagine, for example, that we are interested in modeling the number of children born to a woman. We might expect that this number is a partial function of such explanatory variables as marital status, age, ethnicity, religion, and contraceptive use. It is also likely, however, that some women (or their partners) are infertile and are distinct from fertile women who, though at risk for bearing children, happen to have none. If we knew which women are infertile, we could simply exclude them from the analysis, but let us suppose that this is not the case. To reiterate, there are two sources of 0s in the data that cannot be perfectly distinguished: women who cannot bear children and those who can but have none.

Several statistical models have been proposed for count data with an excess of 0s, including the *zero-inflated Poisson regression* (or ZIP) model, due to Lambert (1992). The ZIP model consists of two components: (1) A binary logistic-regression model for membership in the *latent class* of individuals for whom the response variable is necessarily 0 (e.g., infertile individuals),²¹ and (2) a Poisson-regression model for the latent class of individuals for whom the response may be 0 or a positive count (e.g., fertile women).²²

Let π_i represent the probability that the response Y_i for the i th individual is necessarily 0. Then

$$\log_e \frac{\pi_i}{1 - \pi_i} = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip} \quad (15.5)$$

where the z_{ij} are regressors for predicting membership in the first latent class, and

$$\begin{aligned} \log_e \mu_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\ p(y_i|x_1, \dots, x_k) &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \text{ for } y_i = 0, 1, 2, \dots \end{aligned} \quad (15.6)$$

where $\mu_i \equiv E(Y_i)$ is the expected count for an individual in the second latent class, and the x_{ij} are regressors for the Poisson submodel. In applications, the two sets of regressors—the X s and the Z s—are often the same, but this is not necessarily the case. Indeed, a particularly simple special case arises when the logistic submodel is $\log_e \pi_i / (1 - \pi_i) = \gamma_0$, a constant, implying that the probability of membership in the first latent class is identical for all observations.

²⁰See Exercise 15.1 for a test of overdispersion based on the negative-binomial GLM.

²¹See Section 14.1 for a discussion of logistic regression.

²²Although this form of the zero-inflated count model is the most common, Lambert (1992) also suggested the use of other binary GLMs for membership in the zero latent class (i.e., probit, log-log, and complementary log-log models) and the alternative use of the negative-binomial distribution for the count submodel (see Exercise 15.2).

The probability of observing a 0 count is

$$p(0) \equiv \Pr(Y_i = 0) = \pi_i + (1 - \pi_i)e^{-\mu_i}$$

and the probability of observing any particular nonzero count y_i is

$$p(y_i) = (1 - \pi_i) \times \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

The conditional expectation and variance of Y_i are

$$\begin{aligned} E(Y_i) &= (1 - \pi_i)\mu_i \\ V(Y_i) &= (1 - \pi_i)\mu_i(1 + \pi_i\mu_i) \end{aligned}$$

with $V(Y_i) > E(Y_i)$ for $\pi_i > 0$ [unlike a pure Poisson distribution, for which $V(Y_i) = E(Y_i) = \mu_i$].²³

*Estimation of the ZIP model would be simple if we knew to which latent class each observation belongs, but, as I have pointed out, that is not true. Instead, we must maximize the somewhat more complex combined log-likelihood for the two components of the ZIP model:²⁴

$$\begin{aligned} \log_e L(\beta, \gamma) = & \sum_{y_i=0} \log_e \left\{ \exp(\mathbf{z}'_i \gamma) + \exp[-\exp(\mathbf{x}'_i \beta)] \right\} + \sum_{y_i>0} [y_i \mathbf{x}'_i \beta - \exp(\mathbf{x}'_i \beta)] \\ & - \sum_{i=1}^n \log_e [1 + \exp(\mathbf{z}'_i \gamma)] - \sum_{y_i>0} \log_e (y_i!) \end{aligned} \quad (15.7)$$

where $\mathbf{z}'_i \equiv [1, z_{i1}, \dots, z_{ip}]$, $\mathbf{x}'_i \equiv [1, x_{i1}, \dots, x_{ik}]$, $\gamma \equiv [\gamma_0, \gamma_1, \dots, \gamma_p]'$, and $\beta \equiv [\alpha, \beta_1, \dots, \beta_k]'$.

The basic GLM for count data is the Poisson model with log link. Frequently, however, when the response variable is a count, its conditional variance increases more rapidly than its mean, producing a condition termed *overdispersion* and invalidating the use of the Poisson distribution. The quasi-Poisson GLM adds a dispersion parameter to handle overdispersed count data; this model can be estimated by the method of quasi-likelihood. A similar model is based on the negative-binomial distribution, which is not an exponential family. Negative-binomial GLMs can nevertheless be estimated by maximum likelihood. The zero-inflated Poisson regression model may be appropriate when there are more zeroes in the data than is consistent with a Poisson distribution.

15.2.2 Loglinear Models for Contingency Tables

The joint distribution of several categorical variables defines a *contingency table*. As discussed in the preceding chapter,²⁵ if one of the variables in a contingency table is treated as the

²³See Exercise 15.2.

²⁴See Exercise 15.2.

²⁵See Section 14.3.

Table 15.4 Voter Turnout by Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

Intensity of Preference	Voter Turnout		
	Voted	Did Not Vote	Total
Weak	305	126	431
Medium	405	125	530
Strong	265	49	314
Total	975	300	1275

response variable, we can fit a logit or probit model (e.g., for a dichotomous response, a binomial GLM) to the table. *Loglinear models*, in contrast, which are models for the associations among the variables in a contingency table, treat the variables symmetrically—they do not distinguish one variable as the response. There is, however, a relationship between loglinear models and logit models that I will develop later in this section. As we will see as well, loglinear models have the formal structure of two-way and higher-way ANOVA models²⁶ and can be fit to data by Poisson regression.

Loglinear models for contingency tables have many specialized applications in the social sciences—for example to “square” tables, such as mobility tables, where the variables in the table have the same categories. The treatment of loglinear models in this section merely scratches the surface.²⁷

Two-Way Tables

I will examine contingency tables for two variables in some detail, for this is the simplest case, and the key results that I establish here extend straightforwardly to tables of higher dimension. Consider the illustrative *two-way table* shown in Table 15.4, constructed from data reported in the *American Voter* (Campbell, Converse, Miller, & Stokes, 1960), introduced in the previous chapter.²⁸ The table relates intensity of partisan preference to voting turnout in the 1956 U.S. presidential election. To anticipate my analysis, the data indicate that voting turnout is positively associated with intensity of partisan preference.

More generally, two categorical variables with r and c categories, respectively, define an $r \times c$ contingency table with r rows and c columns, as shown in Table 15.5, where Y_{ij} is the *observed frequency count* in the i,j th cell of the table. I use a “+” to represent summation over a subscript; thus, $Y_{i+} \equiv \sum_{j=1}^c Y_{ij}$ is the *marginal frequency* in the i th row, $Y_{+j} \equiv \sum_{i=1}^r Y_{ij}$ is the marginal frequency in the j th column, and $n = Y_{++} \equiv \sum_{i=1}^r \sum_{j=1}^c Y_{ij}$ is the number of observations in the sample.

²⁶See Sections 8.2.3 and 8.3.

²⁷More extensive accounts are available in many sources, including Agresti (2012), Fienberg (1980), and Powers and Xie (2008).

²⁸Table 14.9 (page 408) examined the relationship of voter turnout to intensity of partisan preference *and* perceived closeness of the election. The current example collapses the table for these three variables over the categories of perceived closeness to examine the *marginal table* for turnout and preference. I return below to the analysis of the full three-way table.

Table 15.5 General Two-Way Frequency Table

		Variable C				
		1	2	...	c	Total
Variable R		Y ₁₁	Y ₁₂	...	Y _{1c}	Y ₁₊
1		Y ₁₁	Y ₁₂	...	Y _{1c}	Y ₁₊
2		Y ₂₁	Y ₂₂	...	Y _{2c}	Y ₂₊
⋮		⋮	⋮	⋮	⋮	⋮
r		Y _{r1}	Y _{r2}	...	Y _{rc}	Y _{r+}
Total		Y ₊ 1	Y ₊ 2	...	Y ₊ c	n

I assume that the n observations in Table 15.5 are independently sampled from a population with proportion π_{ij} in cell i,j and therefore that the probability of sampling an individual observation in this cell is π_{ij} . Marginal probability distributions π_{i+} and π_{+j} may be defined as above; note that $\pi_{++} = 1$. If the row and column variables are statistically independent in the population, then the joint probability π_{ij} is the product of the marginal probabilities for all i and j : $\pi_{ij} = \pi_{i+}\pi_{+j}$.

Because the observed frequencies Y_{ij} result from drawing a random sample, they are random variables that generally take on different values in different samples. The *expected frequency* in cell i,j is $\mu_{ij} \equiv E(Y_{ij}) = n\pi_{ij}$. If the variables are independent, then we have $\mu_{ij} = n\pi_{i+}\pi_{+j}$. Moreover, because $\mu_{i+} = \sum_{j=1}^c n\pi_{ij} = n\pi_{i+}$ and $\mu_{+j} = \sum_{i=1}^r n\pi_{ij} = n\pi_{+j}$, we may write $\mu_{ij} = \mu_{i+}\mu_{+j}/n$. Taking the log of both sides of this last equation produces

$$\eta_{ij} \equiv \log_e \mu_{ij} = \log_e \mu_{i+} + \log_e \mu_{+j} - \log_e n \quad (15.8)$$

That is, under independence, the log expected frequencies η_{ij} depend additively on the logs of the row marginal expected frequencies, the column marginal expected frequencies, and the sample size. As Fienberg (1980, pp. 13–14) points out, Equation 15.8 is reminiscent of a main-effects two-way ANOVA model, where $-\log_e n$ plays the role of the constant, $\log_e \mu_{i+}$ and $\log_e \mu_{+j}$ are analogous to “main-effect” parameters, and η_{ij} appears in place of the response-variable mean. If we impose ANOVA-like sigma constraints on the model,²⁹ we may reparameterize Equation 15.8 as follows:

$$\eta_{ij} = \mu + \alpha_i + \beta_j \quad (15.9)$$

where $\alpha_+ \equiv \sum \alpha_i = 0$ and $\beta_+ \equiv \sum \beta_j = 0$. Equation (15.9) is the *loglinear model for independence* in the two-way table. Solving for the parameters of the model, we obtain

$$\begin{aligned} \mu &= \frac{\eta_{++}}{rc} \\ \alpha_i &= \frac{\eta_{i+}}{c} - \mu \\ \beta_j &= \frac{\eta_{+j}}{r} - \mu \end{aligned} \quad (15.10)$$

It is important to stress that although the loglinear model is *formally* similar to an ANOVA model, the *meaning* of the two models differs importantly: In analysis of variance, the α_i and

²⁹See Section 8.2.3.

β_j are main-effect parameters, specifying the partial relationship of the (quantitative) response variable to each explanatory variable. The loglinear model in Equation 15.9, in contrast, does not distinguish a response variable and, because it is a model for independence, specifies that the row and column variables in the contingency table are *unrelated*; for this model, the α_i and β_j merely express the relationship of the log expected cell frequencies to the row and column marginals. The model for independence describes rc expected frequencies in terms of

$$1 + (r - 1) + (c - 1) = r + c - 1$$

independent parameters.

By analogy to the two-way ANOVA model, we can add parameters to extend the loglinear model to data for which the row and column classifications are not independent in the population but rather are related in an arbitrary manner:

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (15.11)$$

where $\alpha_+ = \beta_+ = \gamma_{i+} = \gamma_{+j} = 0$ for all i and j . As before, we may write the parameters of the model in terms of the log expected counts η_{ij} . Indeed, the solutions for μ, α_i , and β_j are the same as in Equations 15.10, and

$$\gamma_{ij} = \eta_{ij} - \mu - \alpha_i - \beta_j$$

By analogy to the ANOVA model, the γ_{ij} in the loglinear model are often called “interactions,” but this usage is potentially confusing. I will therefore instead refer to the γ_{ij} as *association parameters* because they represent deviations from independence.

Under the model in Equation 15.11, called the *saturated model* for the two-way table, the number of independent parameters is equal to the number of cells in the table,

$$1 + (r - 1) + (c - 1) + (r - 1)(c - 1) = rc$$

The model is therefore capable of capturing *any* pattern of association in a two-way table.

Remarkably, maximum-likelihood estimates for the parameters of a loglinear model (i.e., in the present case, either the model for independence in Equation 15.9 or the saturated model in Equation 15.11) may be obtained by treating the observed cell counts Y_{ij} as the response variable in a Poisson GLM; the log expected counts η_{ij} are then just the linear predictor for the GLM, as the notation suggests.³⁰

The constraint that all $\gamma_{ij} = 0$ imposed by the model of independence can be tested by a likelihood-ratio test, contrasting the model of independence (Equation 15.9) with the more general

³⁰*The reason that this result is remarkable is that a direct route to a likelihood function for the loglinear model leads to the multinomial distribution (discussed in online Appendix D on probability and estimation), not to the Poisson distribution. That is, selecting n independent observations from a population characterized by cell probabilities π_{ij} results in cell counts following the multinomial distribution,

$$\begin{aligned} p(y_{11}, \dots, y_{rc}) &= \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c y_{ij}!} \prod_{i=1}^r \prod_{j=1}^c \pi_{ij}^{n_{ij}} \\ &= \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c y_{ij}!} \prod_{i=1}^r \prod_{j=1}^c \left(\frac{\mu_{ij}}{n}\right)^{n_{ij}} \end{aligned}$$

Noting that the expected counts μ_{ij} are functions of the parameters of the loglinear model leads to the multinomial likelihood function for the model. It turns out that maximizing this multinomial likelihood is equivalent to maximizing the likelihood for the Poisson GLM described in the text (see, e.g., Fienberg, 1980, app. II).

Table 15.6 Estimated Parameters for the Saturated Loglinear Model Fit in Table 15.4

i	$\hat{\gamma}_{ij}$		$\hat{\alpha}_i$
	$j = 1$	$j = 2$	
1	-0.183	0.183	0.135
2	-0.037	0.037	0.273
3	0.219	-0.219	-0.408
$\hat{\beta}_j$	0.625	-0.625	$\hat{\mu} = 5.143$

model (Equation 15.11). Because the latter is a saturated model, its residual deviance is necessarily 0, and the likelihood-ratio statistic for the hypothesis of independence $H_0: \gamma_{ij} = 0$ is simply the residual deviance for the independence model, which has $(r - 1)(c - 1)$ residual degrees of freedom. Applied to the illustrative two-way table for the *American Voter* data, we get $G_0^2 = 19.428$ with $(3 - 1)(2 - 1) = 2$ degrees of freedom, for which $p < .0001$, suggesting that there is strong evidence that intensity of preference and turnout are related.³¹

Maximum-likelihood estimates of the parameters of the saturated loglinear model are shown in Table 15.6. It is clear from the estimated association parameters $\hat{\gamma}_{ij}$ that turning out to vote, $j = 1$, increases with partisan preference (and, of course, that *not* turning out to vote, $j = 2$, decreases with preference).

Three-Way Tables

The saturated loglinear model for a three-way ($a \times b \times c$) table for variables A , B , and C is defined in analogy to the three-way ANOVA model, although, as in the case of two-way tables, the meaning of the parameters is different:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)} + \alpha_{ABC(ijk)} \quad (15.12)$$

with sigma constraints specifying that each set of parameters sums to 0 over each subscript; for example, $\alpha_{A(+)} = \alpha_{AB(i+)} = \alpha_{ABC(ij+)} = 0$. Given these constraints, we may solve for the parameters in terms of the log expected counts, with the solution following the usual ANOVA pattern; for example,

$$\begin{aligned}\mu &= \frac{\eta_{+++}}{abc} \\ \alpha_{A(i)} &= \frac{\eta_{i++}}{bc} - \mu \\ \alpha_{AB(ij)} &= \frac{\eta_{ij+}}{c} - \mu - \alpha_{A(i)} - \alpha_{B(j)} \\ \alpha_{ABC(ijk)} &= \eta_{ijk} - \mu - \alpha_{A(i)} - \alpha_{B(j)} - \alpha_{C(k)} - \alpha_{AB(ij)} - \alpha_{AC(ik)} - \alpha_{BC(jk)}\end{aligned}$$

³¹This test is very similar to the usual Pearson chi-square test for independence in a two-way table. See Exercise 15.3 for details, and for an alternative formula for calculating the likelihood-ratio test statistic G_0^2 directly from the observed frequencies, Y_{ij} , and estimated expected frequencies under independence, $\hat{\mu}_{ij}$.

The presence of the three-way term α_{ABC} in the model implies that the relationship between any pair of variables (say, A and B) depends on the category of the third variable (say, C).³²

Other loglinear models are defined by suppressing certain terms in the saturated model, that is, by setting parameters to 0. In specifying a restricted loglinear model, we will be guided by the principle of marginality:³³ Whenever a high-order term is included in the model, its lower-order relatives are included as well. Loglinear models of this type are often called *hierarchical*. Nonhierarchical loglinear models may be suitable for special applications, but they are not sensible in general (see Fienberg, 1980). According to the principle of marginality, for example, if α_{AB} appears in the model, so do α_A and α_B .

- If we set all of α_{ABC} , α_{AB} , α_{AC} , and α_{BC} to 0, we produce the model of mutual independence, implying that the variables in the three-way table are completely unrelated:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)}$$

- Setting α_{ABC} , α_{AC} , and α_{BC} to 0 yields the model

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)}$$

which specifies (1) that variables A and B are related, controlling for (i.e., within categories of) variable C ; (2) that this partial relationship is constant across the categories of variable C ; and (3) that variable C is independent of variables A and B taken jointly—that is, if we form the two-way table with rows given by combinations of categories of A and B and columns given by C , the two variables in this table are independent. Note that there are two other models of this sort: one in which α_{AC} is nonzero and another in which α_{BC} is nonzero.

- A third type of model has *two* nonzero two-way terms; for example, setting α_{ABC} and α_{BC} to 0, we obtain

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)}$$

This model implies that (1) variables A and B have a constant partial relationship across the categories of variable C , (2) variables A and C have a constant partial relationship across the categories of variable B , and (3) variables B and C are independent within categories of variable A . Again, there are two other models of this type.

- Finally, consider the model that sets only the three-way term α_{ABC} to 0:

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)}$$

This model specifies that each pair of variables (e.g., A and B) has a constant partial association across the categories of the remaining variable (e.g., C).

These descriptions are relatively complicated because the loglinear models are models of association among variables. As we will see presently, however, if one of the variables in a

³²Here and below I use the shorthand notation α_{ABC} to represent the whole set of $\alpha_{ABC(ijk)}$ parameters and similarly for the other terms in the model.

³³See Section 7.3.2.

Table 15.7 Voter Turnout by Perceived Closeness of the Election and Intensity of Partisan Preference, for the 1956 U.S. Presidential Election

(A) Perceived Closeness	(B) Intensity of Preference	(C) Turnout	
		Voted	Did Not Vote
One-sided	Weak	91	39
	Medium	121	49
	Strong	64	24
	Close	214	87
	Weak	284	76
	Medium	201	25

Table 15.8 Hierarchical Loglinear Models Fit to Table 15.7

High-Order Terms	General	Residual Degrees of Freedom		
		Table 15.7	G_0^2	p
A, B, C	$(a-1)(b-1)+(a-1)(c-1)(b-1)(c-1)$ $+(a-1)(b-1)(c-1)$	7	36.39	<<.0001
AB, C	$(a-1)(c-1)+(b-1)(c-1)+(a-1)(b-1)(c-1)$	5	34.83	<<.0001
AC, B	$(a-1)(b-1)+(b-1)(c-1)+(a-1)(b-1)(c-1)$	5	16.96	.0046
A, BC	$(a-1)(b-1)+(a-1)(c-1)+(a-1)(b-1)(c-1)$	6	27.78	.0001
AB, AC	$(b-1)(c-1)+(a-1)(b-1)(c-1)$	3	15.40	.0015
AB, BC	$(a-1)(c-1)+(a-1)(b-1)(c-1)$	4	26.22	<.0001
AC, BC	$(a-1)(b-1)+(a-1)(b-1)(c-1)$	4	8.35	.079
AB, AC, BC	$(a-1)(b-1)(c-1)$	2	7.12	.028
ABC	0	0	0.0	—

NOTE: The column labeled G_0^2 is the likelihood-ratio statistic for testing each model against the saturated model.

table is taken as the response variable, then the loglinear model is equivalent to a logit model with a simpler interpretation.

Table 15.7 shows a three-way table cross-classifying voter turnout by perceived closeness of the election and intensity of partisan preference, elaborating the two-way table for the *American Voter* data presented earlier in Table 15.4.³⁴ I have fit all hierarchical loglinear models to this three-way table, displaying the results in Table 15.8. Here I employ a compact notation for the high-order terms in each fitted model: For example, AB represents the two-way term α_{AB} and implies that the lower-order relatives of this term— μ , α_A , and α_B —are also in the model. As in the loglinear model for a two-way table, the saturated model has a residual

³⁴This table was also discussed in Chapter 14 (see Table 14.9 on page 408).

deviance of 0, and consequently, the likelihood-ratio statistic to test any model against the saturated model (within which all of the other models are nested and which is the last model shown) is simply the residual deviance for the unsaturated model.

The first model in Table 15.8 is the model of complete independence, and it fits the data very poorly. At the other end, the model with high-order terms AB , AC , and BC , which may be used to test the hypothesis of no three-way association, H_0 : all $\alpha_{ABC(ijk)} = 0$, also has a statistically significant likelihood-ratio test statistic (though not overwhelmingly so), suggesting that the association between any pair of variables in the contingency tables varies over the levels of the remaining variable.

This approach generalizes to contingency tables of any dimension, although the interpretation of high-order association terms can become complicated.

Loglinear Models and Logit Models

As I explained, the loglinear model for a contingency table is a model for association among the variables in the table; the variables are treated symmetrically, and none is distinguished as the response variable. When one of the variables in a contingency table is regarded as the response, however, the loglinear model for the table implies a logit model (identical to the logit model for a contingency table developed in Chapter 14), the parameters of which bear a simple relationship to the parameters of the loglinear model for the table.

For example, it is natural to regard voter turnout in Table 15.7 as a dichotomous response variable, potentially affected by perceived closeness of the election and by intensity of partisan preference. Indeed, this is precisely what we did previously when we analyzed this table using a logit model.³⁵ With this example in mind, let us return to the saturated loglinear model for the three-way table (repeating Equation 15.12):

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{AC(ik)} + \alpha_{BC(jk)} + \alpha_{ABC(ijk)}$$

For convenience, I suppose that the response variable is variable C , as in the illustration. Let Ω_{ij} symbolize the response-variable logit within categories i, j of the two explanatory variables; that is,

$$\begin{aligned}\Omega_{ij} &= \log_e \frac{\pi_{ij1}}{\pi_{ij2}} = \log_e \frac{n\pi_{ij1}}{n\pi_{ij2}} = \log_e \frac{\mu_{ij1}}{\mu_{ij2}} \\ &= \eta_{ij1} - \eta_{ij2}\end{aligned}$$

Then, from the saturated loglinear model for η_{ijk} ,

$$\begin{aligned}\Omega_{ij} &= [\alpha_{C(1)} - \alpha_{C(2)}] + [\alpha_{AC(i1)} - \alpha_{AC(i2)}] \\ &\quad + [\alpha_{BC(j1)} - \alpha_{BC(j2)}] + [\alpha_{ABC(ij1)} - \alpha_{ABC(ij2)}]\end{aligned}\tag{15.13}$$

Noting that the first bracketed term in Equation 15.13 does not depend on the explanatory variables, that the second depends only on variable A , and so forth, let us rewrite this equation in the following manner:

³⁵See Section 14.3.

$$\Omega_{ij} = \omega + \omega_{A(i)} + \omega_{B(j)} + \omega_{AB(ij)} \quad (15.14)$$

where, because of the sigma constraints on the α s,

$$\begin{aligned}\omega &\equiv \alpha_{C(1)} - \alpha_{C(2)} = 2\alpha_{C(1)} \\ \omega_{A(i)} &\equiv \alpha_{AC(i1)} - \alpha_{AC(i2)} = 2\alpha_{AC(i1)} \\ \omega_{B(j)} &\equiv \alpha_{BC(j1)} - \alpha_{BC(j2)} = 2\alpha_{BC(j1)} \\ \omega_{AB(ij)} &\equiv \alpha_{ABC(ij1)} - \alpha_{ABC(ij2)} = 2\alpha_{ABC(ij1)}\end{aligned}$$

Furthermore, because they are defined as twice the α s, the ω s are also constrained to sum to 0 over any subscript:

$$\omega_{A(+)} = \omega_{B(+)} = \omega_{AB(i+)} = \omega_{AB(+j)} = 0, \text{ for all } i \text{ and } j$$

Note that the loglinear model parameters for the association of the *explanatory* variables A and B do not appear in Equation 15.3. This equation (or, equivalently, Equation 15.14), the saturated logit model for the table, therefore shows how the response-variable log odds depend on the explanatory variables and their interactions. In light of the constraints that they satisfy, the ω s are interpretable as ANOVA-like effect parameters, and indeed we have returned to the binomial logit model for a contingency table introduced in the previous chapter: For example, the likelihood-ratio test for the three-way term in the loglinear model for the *American Voter* data (given in the penultimate line of Table 15.8) is identical to the likelihood-ratio test for the interaction between closeness and preference in the logit model for turnout fit to these data (see Table 14.11 on page 411).

A similar argument may also be pursued with respect to *any* unsaturated loglinear model for the three-way table: Each such model implies a model for the response-variable logits. Because, however, our purpose is to examine the effects of the explanatory variables on the response and not to explore the association *between* the explanatory variables, we generally include α_{AB} and its lower-order relatives in *any* model that we fit, thereby treating the association (if any) between variables A and B as given. Furthermore, a similar argument to the one developed here can be applied to a table of any dimension that has a response variable and to a response variable with more than two categories. In the latter event, the loglinear model is equivalent to a *multinomial* logit model for the table, and in any event, we would generally include in the loglinear model a term of dimension one less than the table corresponding to all associations among the explanatory variables.

Loglinear models for contingency tables bear a formal resemblance to analysis-of-variance models and can be fit to data as Poisson generalized linear models with a log link. The loglinear model for a contingency table, however, treats the variables in the table symmetrically—none of the variables is distinguished as a response variable—and consequently the parameters of the model represent the associations among the variables, not the effects of explanatory variables on a response. When one of the variables is construed as the response, the loglinear model reduces to a binomial or multinomial logit model.

15.3 Statistical Theory for Generalized Linear Models*

In this section, I revisit with greater rigor and more detail many of the points raised in the preceding sections.³⁶

15.3.1 Exponential Families

As much else in modern statistics, the insight that many of the most important distributions in statistics could be expressed in the following common “linear-exponential” form was due to R. A. Fisher:

$$p(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \quad (15.15)$$

where

- $p(y; \theta, \phi)$ is the probability-mass function for the discrete random variable Y or the probability-density function for continuous Y .
- $a(\cdot), b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another (see below for examples).
- $\theta = g_c(\mu)$, the *canonical parameter* for the exponential family in question, is a function of the expectation $\mu \equiv E(Y)$ of Y ; moreover, the *canonical link function* $g_c(\cdot)$ does not depend on ϕ .
- $\phi > 0$ is a *dispersion parameter*, which, in some families, takes on a fixed, known value, while in other families it is an unknown parameter to be estimated from the data along with θ .

Consider, for example, the normal or Gaussian distribution with mean μ and variance σ^2 , the density function for which is given in Equation 15.1 (on page 422). To put the normal distribution into the form of Equation 15.15 requires some heroic algebraic manipulation, eventually producing³⁷

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - \theta^2/2}{\phi} - \frac{1}{2} \left[\frac{y^2}{\phi} + \log_e(2\pi\phi) \right] \right\}$$

with $\theta = g_c(\mu) = \mu$; $\phi = \sigma^2$; $a(\phi) = \phi$; $b(\theta) = \theta^2/2$; and $c(y, \phi) = -\frac{1}{2}[y^2/\phi + \log_e(2\pi\phi)]$. Thus, $g_c(\cdot)$ is the identity link.

Now consider the binomial distribution in Equation 15.2 (page 422), where Y is the proportion of “successes” in n independent binary trials, and μ is the probability of success on an individual trial. Written after more algebraic gymnastics as an exponential family,³⁸

³⁶The exposition here owes a debt to Chapter 2 of McCullagh and Nelder (1989), which has become the standard source on GLMs, and to the remarkably lucid and insightful brief treatment of the topic by Firth (1991).

³⁷See Exercise 15.4.

³⁸See Exercise 15.5.

Table 15.9 Functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ for Constructing the Exponential Families

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}[y^2/\phi + \log_e(2\pi\phi)]$
Binomial	$1/n$	$\log_e(1 + e^\theta)$	$\log_e\left(\frac{n}{ny}\right)$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-1}\log_e(y/\phi) - \log_e y - \log_e \Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}[\log_e(\pi\phi y^3) + 1/(\phi y)]$

NOTE: In this table, n is the number of binomial trials, and $\Gamma(\cdot)$ is the gamma function.

$$p(y; \theta, \phi) = \exp\left[\frac{y\theta - \log_e(1 + e^\theta)}{1/n} + \log_e\left(\frac{n}{ny}\right)\right]$$

with $\theta = g_c(\mu) = \log_e[\mu/(1 - \mu)]$; $\phi = 1$; $a(\phi) = 1/n$; $b(\theta) = \log_e(1 + e^\theta)$; and $c(y, \phi) = \log_e\left(\frac{n}{ny}\right)$. The canonical link is therefore the logit link.

Similarly, the Poisson, gamma, and inverse-Gaussian families can all be put into the form of Equation 15.15, using the results given in Table 15.9.³⁹

The advantage of expressing diverse families of distributions in the common exponential form is that general properties of exponential families can then be applied to the individual cases. For example, it is true in general that

$$b'(\theta) \equiv \frac{db(\theta)}{d\theta} = \mu$$

and that

$$V(Y) = a(\phi)b''(\theta) = a(\phi) \frac{d^2 b(\theta)}{d\theta^2} = a(\phi)v(\mu)$$

leading to the results in Table 15.2 (on page 441)⁴⁰. Note that $b'(\cdot)$ is the inverse of the canonical link function. For example, for the normal distribution,

$$\begin{aligned} b'(\theta) &= \frac{d(\theta^2/2)}{d\theta} = \theta = \mu \\ a(\phi)b''(\theta) &= \phi \times 1 = \sigma^2 \\ v(\mu) &= 1 \end{aligned}$$

and for the binomial distribution,

³⁹See Exercise 15.6.

⁴⁰See Exercise 15.7.

$$\begin{aligned}
b'(\theta) &= \frac{d[\log_e(1 + e^\theta)]}{d\theta} = \frac{e^\theta}{1 + e^\theta} = \frac{1}{1 + e^{-\theta}} = \mu \\
a(\phi)b''(\theta) &= \frac{1}{n} \times \left[\frac{e^\theta}{1 + e^\theta} - \left(\frac{e^\theta}{1 + e^\theta} \right)^2 \right] = \frac{\mu(1 - \mu)}{n} \\
v(\mu) &= \mu(1 - \mu)
\end{aligned}$$

The Gaussian, binomial, Poisson, gamma, and inverse-Gaussian distributions can all be written in the common linear-exponential form:

$$p(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another; $\theta = g_c(\mu)$ is the canonical parameter for the exponential family in question; $g_c(\cdot)$ is the canonical link function; and $\phi > 0$ is a dispersion parameter, which takes on a fixed, known value in some families. It is generally the case that $\mu = E(Y) = b'(\theta)$ and that $V(Y) = a(\phi)b''(\theta)$.

15.3.2 Maximum-Likelihood Estimation of Generalized Linear Models

The log-likelihood for an individual observation Y_i follows directly from Equation 15.5 (page 433):

$$\log_e L(\theta_i, \phi; Y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi)$$

For n independent observations, we have

$$\log_e L(\boldsymbol{\theta}, \phi; \mathbf{y}) = \sum_{i=1}^n \frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi) \quad (15.16)$$

where $\boldsymbol{\theta} \equiv \{\theta_i\}$ and $\mathbf{y} \equiv \{Y_i\}$.

Suppose that a GLM uses the link function $g(\cdot)$, so that⁴¹

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

The model therefore expresses the expected values of the n observations in terms of a much smaller number of regression parameters. To get estimating equations for the regression parameters, we have to differentiate the log-likelihood with respect to each coefficient in turn. Let l_i represent the i th component of the log-likelihood. Then, by the chain rule,

⁴¹It is notationally convenient here to write β_0 for the regression constant α .

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{\partial \eta_i}{\partial \beta_j} \text{ for } j = 0, 1, \dots, k \quad (15.17)$$

After some work, we can rewrite Equation 15.17 as⁴²

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij}$$

Summing over observations, and setting the sum to 0, produces the maximum-likelihood estimating equations for the GLM,

$$\sum_{i=1}^n \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \dots, k \quad (15.18)$$

where $a_i \equiv a_i(\phi)/\phi$ does not depend on the dispersion parameter, which is constant across observations. For example, in a Gaussian GLM, $a_i = 1$, while in a binomial GLM, $a_i = 1/n_i$.

Further simplification can be achieved when $g(\cdot)$ is the canonical link. In this case, the maximum-likelihood estimating equations become

$$\sum_{i=1}^n \frac{Y_i x_{ij}}{a_i} = \sum_{i=1}^n \frac{\mu_i x_{ij}}{a_i}$$

setting the “observed sum” on the left of the equation to the “expected sum” on the right. We noted this pattern in the estimating equations for logistic-regression models in the previous chapter.⁴³ Nevertheless, even here the estimating equations are (except in the case of the Gaussian family paired with the identity link) nonlinear functions of the regression parameters and generally require iterative methods for their solution.

Iterative Weighted Least Squares

Let

$$\begin{aligned} Z_i &\equiv \eta_i + (Y_i - \mu_i) \frac{d\eta_i}{d\mu_i} \\ &= \eta_i + (Y_i - \mu_i) g'(\mu_i) \end{aligned}$$

Then

$$E(Z_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

and

$$V(Z_i) = [g'(\mu_i)]^2 a_i v(\mu_i)$$

If, therefore, we could compute the Z_i , we would be able to fit the model by weighted least-squares regression of Z on the X s, using the inverses of the $V(Z_i)$ as weights.⁴⁴ We cannot, however, proceed in this manner, because we do not know the values of the μ_i and η_i , which, indeed, depend on the regression coefficients that we wish to estimate—that is, the argument is

⁴²See Exercise 15.8.

⁴³See Sections 14.1.5 and 14.2.1.

⁴⁴See Section 12.2.2 for a general discussion of weighted least squares.

essentially circular. This observation suggested to Nelder and Wedderburn (1972) the possibility of estimating GLMs by *iterative weighted least squares* (IWLS), cleverly turning the circularity into an iterative procedure:

1. Start with initial estimates of the $\hat{\mu}_i$ and the $\hat{\eta}_i = g(\hat{\mu}_i)$, denoted $\hat{\mu}_i^{(0)}$ and $\hat{\eta}_i^{(0)}$. A simple choice is to set $\hat{\mu}_i^{(0)} = Y_i$.⁴⁵
2. At each iteration l , compute the *working response variable* Z using the values of $\hat{\mu}$ and $\hat{\eta}$ from the preceding iteration,

$$Z_i^{(l-1)} = \eta_i^{(l-1)} + (Y_i - \mu_i^{(l-1)}) g'(\mu_i^{(l-1)})$$

along with weights

$$W_i^{(l-1)} = \frac{1}{[g'(\mu_i^{(l-1)})]^2 a_i v(\mu_i^{(l-1)})}$$

3. Fit a weighted least-squares regression of $Z^{(l-1)}$ on the X s, using the $W^{(l-1)}$ as weights. That is, compute

$$\mathbf{b}^{(l)} = (\mathbf{X}' \mathbf{W}^{(l-1)} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{(l-1)} \mathbf{z}^{(l-1)}$$

where $\mathbf{b}^{(l)}$ is the vector of regression coefficients at the current iteration, \mathbf{X} is (as usual) the model matrix, $\mathbf{W}^{(l-1)} \equiv \text{diag}\{W_i^{(l-1)}\}$ is the diagonal weight matrix, and $\mathbf{z}^{(l-1)} \equiv \{Z_i^{(l-1)}\}$ is the working-response vector.

4. Repeat Steps 2 and 3 until the regression coefficients stabilize, at which point \mathbf{b} converges to the maximum-likelihood estimates of the β s.

Applied to the canonical link, IWLS is equivalent to the Newton-Raphson method (as we discovered for a logit model in the previous chapter); more generally, IWLS implements Fisher's "method of scoring."

Estimating the Dispersion Parameter

We do not require an estimate of the dispersion parameter to estimate the regression coefficients in a GLM. Although it is in principle possible to estimate ϕ by maximum likelihood as well, this is rarely done. Instead, recall that $V(Y_i) = \phi a_i v(\mu_i)$. Solving for the dispersion parameter, we get $\phi = V(Y_i)/a_i v(\mu_i)$, suggesting the *method of moments* estimator

⁴⁵In certain settings, starting with $\hat{\mu}_i^{(0)} = Y_i$ can cause computational difficulties. For example, in a binomial GLM, some of the observed proportions may be 0 or 1—indeed, for binary data, this will be true for *all* the observations—requiring us to divide by 0 or to take the log of 0. The solution is to adjust the starting values, which are in any event not critical, to protect against this possibility. For a binomial GLM, where $Y_i = 0$, we can take $\hat{\mu}_i^{(0)} = 0.5/n_i$, and where $Y_i = 1$, we can take $\hat{\mu}_i^{(0)} = (n_i - 0.5)/n_i$. For binary data, then, all the $\hat{\mu}_i^{(0)}$ are 0.5.

$$\tilde{\phi} = \frac{1}{n - k - 1} \sum \frac{(Y_i - \hat{\mu}_i)^2}{a_i v(\hat{\mu}_i)} \quad (15.19)$$

The estimated asymptotic covariance matrix of the coefficients is then obtained from the last IWLS iteration as

$$\hat{\mathcal{V}}(\mathbf{b}) = \tilde{\phi} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \quad (15.20)$$

Because the maximum-likelihood estimator \mathbf{b} is asymptotically normally distributed, $\hat{\mathcal{V}}(\mathbf{b})$ may be used as the basis for Wald tests of the regression parameters.

The maximum-likelihood estimating equations for generalized linear models take the common form

$$\sum_{i=1}^n \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \dots, k$$

These equations are generally nonlinear and therefore have no general closed-form solution, but they can be solved by IWLS. The estimating equations for the coefficients do not involve the dispersion parameter, which (for models in which the dispersion is not fixed) then can be estimated as

$$\tilde{\phi} = \frac{1}{n - k - 1} \sum \frac{(Y_i - \hat{\mu}_i)^2}{a_i v(\hat{\mu}_i)}$$

The estimated asymptotic covariance matrix of the coefficients is

$$\hat{\mathcal{V}}(\mathbf{b}) = \tilde{\phi} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$$

where \mathbf{b} is the vector of estimated coefficients and \mathbf{W} is a diagonal matrix of weights from the last IWLS iteration.

Quasi-Likelihood Estimation

The argument leading to IWLS estimation rests only on the linearity of the relationship between $\eta = g(\mu)$ and the X 's, as well as on the assumption that $V(Y)$ depends in a particular manner on a dispersion parameter and μ . As long as we can express the transformed mean of Y as a linear function of the X 's and can write down a variance function for Y (expressing the conditional variance of Y as a function of its mean and a dispersion parameter), we can apply the “maximum-likelihood” estimating equations (Equation 15.8 on page 436) and obtain estimates by IWLS—even without committing ourselves to a particular conditional distribution for Y .

This is the method of *quasi-likelihood estimation*, introduced by Wedderburn (1974), and it has been shown to retain many of the properties of maximum-likelihood estimation: Although the quasi-likelihood estimator may not be maximally asymptotically efficient, it is consistent and has the same asymptotic distribution as the maximum-likelihood estimator of a GLM in an exponential family.⁴⁶ We can think of quasi-likelihood estimation of GLMs as analogous to

⁴⁶See, for example, McCullagh and Nelder (1989, chap. 9) and McCullagh (1991).

least-squares estimation of linear regression models with potentially non-normal errors: Recall that as long as the relationship between Y and the X s is linear, the error variance is constant, and the observations are independently sampled, the theory underlying OLS estimation applies—although the OLS estimator may no longer be maximally efficient.⁴⁷

The maximum-likelihood estimating equations, and IWLS estimation, can be applied whenever we can express the transformed mean of Y as a linear function of the X s and can write the conditional variance of Y as a function of its mean and (possibly) a dispersion parameter—even when we do not specify a particular conditional distribution for Y . The resulting quasi-likelihood estimator shares many of the properties of maximum-likelihood estimators.

15.3.3 Hypothesis Tests

Analysis of Deviance

Originally (in Equation 15.6 on page 433), I wrote the log-likelihood for a GLM as a function $\log_e L(\theta, \phi; \mathbf{y})$ of the canonical parameters θ for the observations. Because $\mu_i = g_c^{-1}(\theta_i)$, for the canonical link $g_c(\cdot)$, we can equally well think of the log-likelihood as a function of the expected response and therefore can write the maximized log-likelihood as $\log_e L(\hat{\mu}, \phi; \mathbf{y})$. If we then dedicate a parameter to each observation, so that $\hat{\mu}_i = Y_i$ (e.g., by removing the constant from the regression model and defining a dummy regressor for each observation), the log-likelihood becomes $\log_e L(\mathbf{y}, \phi; \mathbf{y})$. The *residual deviance* under the initial model is twice the difference in these log-likelihoods:

$$\begin{aligned} D(\mathbf{y}; \hat{\mu}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\hat{\mu}, \phi; \mathbf{y})] \\ &= 2 \sum_{i=1}^n [\log_e L(Y_i, \phi; Y_i) - \log_e L(\hat{\mu}_i, \phi; Y_i)] \\ &= 2 \sum_{i=1}^n \frac{Y_i[g(Y_i) - g(\hat{\mu}_i)] - b[g(Y_i)] + b[g(\hat{\mu}_i)]}{a_i} \end{aligned} \tag{15.21}$$

Dividing the residual deviance by the estimated dispersion parameter produces the *scaled deviance*, $D^*(\mathbf{y}; \hat{\mu}) \equiv D(\mathbf{y}; \hat{\mu})/\tilde{\phi}$. As explained in Section 15.1.1, deviances are the building blocks of likelihood-ratio chi-square and F -tests for GLMs.

Applying Equation 15.21 to the Gaussian distribution, where $g_c(\cdot)$ is the identity link, $a_i = 1$, and $b(\theta) = \theta^2/2$, produces (after some simplification)

⁴⁷See Chapter 9.

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum (Y_i - \hat{\mu})^2$$

that is, the residual sum of squares for the model. Similarly, applying Equation 15.21 to the binomial distribution, where $g_c(\cdot)$ is the logit link, $a_i = n_i$, and $b(\theta) = \log_e(1 + e^\theta)$, we get (after quite a bit of simplification)⁴⁸

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum n_i \left[Y_i \log_e \frac{Y_i}{\hat{\mu}_i} + (1 - Y_i) \log_e \frac{1 - Y_i}{1 - \hat{\mu}_i} \right]$$

The residual deviance for a model is twice the difference in the log-likelihoods for the saturated model, which dedicates one parameter to each observation, and the model in question:

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2 \sum_{i=1}^n \frac{Y_i[g(Y_i) - g(\hat{\mu}_i)] - b[g(Y_i)] + b[g(\hat{\mu}_i)]}{a_i} \end{aligned}$$

Dividing the residual deviance by the estimated dispersion parameter produces the scaled deviance, $D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) \equiv D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\tilde{\phi}$.

Testing General Linear Hypotheses

As was the case for linear models,⁴⁹ we can formulate a test for the general linear hypothesis

$$H_0: \mathbf{L}_{(q \times k+1)} \boldsymbol{\beta}_{(k+1 \times 1)} = \mathbf{c}_{(q \times 1)}$$

where the hypothesis matrix \mathbf{L} and right-hand-side vector \mathbf{c} contain prespecified constants; usually, $\mathbf{c} = \mathbf{0}$. For a GLM, the Wald statistic

$$Z_0^2 = (\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})$$

follows an asymptotic chi-square distribution with q degrees of freedom under the hypothesis. The simplest application of this result is to the Wald statistic $Z_0 = B_j/\text{SE}(B_j)$, testing that an individual regression coefficient is 0. Here, Z_0 follows a standard-normal distribution under $H_0: \beta_j = 0$ (or, equivalently, Z_0^2 follows a chi-square distribution with 1 degree of freedom).

Alternatively, when the dispersion parameter is estimated from the data, we can calculate the test statistic

$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})}{q}$$

which is distributed as $F_{q, n-k-1}$ under H_0 . Applied to an individual coefficient, $t_0 = \pm \sqrt{F_0} = B_j/\text{SE}(B_j)$ produces a t -test on $n - k - 1$ degrees of freedom.

⁴⁸See Exercise 15.9, which also develops formulas for the deviance in Poisson, gamma, and inverse-Gaussian models.

⁴⁹See Section 9.4.3.

To test the general linear hypothesis $H_0: \mathbf{L}\beta = \mathbf{c}$, where the hypothesis matrix \mathbf{L} has q rows, we can compute the Wald chi-square test statistic $Z_0^2 = (\mathbf{L}\hat{\beta} - \mathbf{c})'[\mathbf{L}\hat{V}(\mathbf{b})\mathbf{L}]^{-1}(\mathbf{L}\hat{\beta} - \mathbf{c})$, with q degrees of freedom. Alternatively, if the dispersion parameter is estimated from the data, we can compute the F -test statistic $F_0 = (\mathbf{L}\hat{\beta} - \mathbf{c})'[\mathbf{L}\hat{V}(\mathbf{b})\mathbf{L}]^{-1}(\mathbf{L}\hat{\beta} - \mathbf{c})/q$ on q and $n - k - 1$ degrees of freedom.

Testing Nonlinear Hypotheses

It is occasionally of interest to test a hypothesis or construct a confidence interval for a *nonlinear* function of the parameters of a linear or generalized linear model. If the nonlinear function in question is a differentiable function of the regression coefficients, then an approximate asymptotic standard error may be obtained by the *delta method*.⁵⁰

Suppose that we are interested in the function

$$\gamma \equiv f(\beta) = f(\beta_0, \beta_1, \dots, \beta_k)$$

where, for notational convenience, I have used β_0 to denote the regression constant. The function $f(\beta)$ need not use *all* the regression coefficients (see the example below). The maximum-likelihood estimator of γ is simply $\hat{\gamma} = f(\hat{\beta})$ (which, as an MLE, is also asymptotically normal), and the approximate sampling variance of $\hat{\gamma}$ is then

$$\hat{V}(\hat{\gamma}) \approx \sum_{j=0}^k \sum_{j'=0}^k v_{jj'} \times \frac{\partial \hat{\gamma}}{\partial \hat{\beta}_j} \times \frac{\partial \hat{\gamma}}{\partial \hat{\beta}_{j'}}$$

where $v_{jj'}$ is the j, j' 'th element of the estimated asymptotic covariance matrix of the coefficients, $\hat{V}(\hat{\beta})$.

To illustrate the application of this result, imagine that we are interested in determining the maximum or minimum value of a quadratic partial regression.⁵¹ Focusing on the partial relationship between the response variable and a particular X , we have an equation of the form

$$E(Y) = \dots + \beta_1 X + \beta_2 X^2 + \dots$$

Differentiating this equation with respect to X , we get

$$\frac{dE(Y)}{dX} = \beta_1 + 2\beta_2 X$$

Setting the derivative to 0 and solving for X produces the value at which the function reaches a minimum (if β_2 is positive) or a maximum (if β_2 is negative),

⁵⁰The delta method (Rao, 1973) is described in online Appendix D on probability and estimation. The method employs a first-order (i.e., linear) Taylor-series approximation to the nonlinear function. The delta method is appropriate here because the maximum-likelihood (or quasi-likelihood) estimates of the coefficients of a GLM are asymptotically normally distributed. Indeed, the procedure described in this section is applicable whenever the parameters of a regression model are normally distributed and can therefore be applied in a wide variety of contexts—such as to the nonlinear regression models described in Chapter 17. In small samples, however, the delta-method approximation to the standard error may not be adequate, and the bootstrapping procedures described in Chapter 21 will usually provide more reliable results.

⁵¹See Section 17.1 for a discussion of polynomial regression. The application of the delta method to finding the minimum or maximum of a quadratic curve is suggested by Weisberg (2014, Section 7.6).

$$X = -\frac{\beta_1}{2\beta_2}$$

which is a nonlinear function of the regression coefficients β_1 and β_2 .

For example, in Section 12.3.1, using data from the Canadian Survey of Labour and Income Dynamics (the “SLID”), I fit a least-squares regression of log wage rate on a quadratic in age, a dummy regressor for sex, and the square of education, obtaining (repeating, and slightly rearranging, Equation 12.7 on page 310):

$$\begin{aligned}\widehat{\log_2 \text{Wages}} &= 0.5725 + 0.1198 \times \text{Age} - 0.001230 \times \text{Age}^2 \\ &\quad (0.0834) \quad (0.0046) \quad (0.000059) \\ &\quad + 0.3195 \times \text{Male} + 0.002605 \times \text{Education}^2 \\ &\quad (0.0180) \quad (0.000113)\end{aligned}$$

$$R^2 = .3892$$

Imagine that we are interested in the age $\gamma \equiv -\beta_1/(2\beta_2)$ at which wages are at a maximum, holding sex and education constant. The necessary derivatives are

$$\begin{aligned}\frac{\partial \widehat{\gamma}}{\partial \beta_1} &= -\frac{1}{2B_2} = -\frac{1}{2(-0.001230)} = 406.5 \\ \frac{\partial \widehat{\gamma}}{\partial \beta_2} &= \frac{B_1}{2B_2^2} = \frac{0.1198}{2(-0.001230)^2} = 39,593\end{aligned}$$

Our point estimate of γ is

$$\widehat{\gamma} = -\frac{B_1}{2B_2} = -\frac{0.1198}{2 \times 0.001230} = 48.70 \text{ years}$$

The estimated sampling variance is $\widehat{V}(B_1) = 2.115 \times 10^{-5}$ of the age coefficient and $\widehat{V}(B_2) = 3.502 \times 10^{-9}$ of the coefficient of age-squared; the estimated sampling covariance for the two coefficients is $\widehat{C}(B_1, B_2) = -2.685 \times 10^{-7}$. The approximate estimated variance of $\widehat{\gamma}$ is then

$$\begin{aligned}\widehat{V}(\widehat{\gamma}) &\approx (2.115 \times 10^{-5}) \times 406.5^2 - (2.685 \times 10^{-7}) \times 406.5 \times 39,593 \\ &\quad - (2.685 \times 10^{-7}) \times 406.5 \times 39,593 + (3.502 \times 10^{-9}) \times 39,593^2 \\ &= 0.3419\end{aligned}$$

Consequently, the approximate standard error of $\widehat{\gamma}$ is $\text{SE}(\widehat{\gamma}) \approx \sqrt{0.3419} = 0.5847$, and an approximate 95% confidence interval for the age at which income is highest on average is $\gamma = 48.70 \pm 1.96(0.5847) = (47.55, 49.85)$.

The delta method may be used to approximate the standard error of a nonlinear function of regression coefficients in a GLM. If $\gamma \equiv f(\beta_0, \beta_1, \dots, \beta_k)$, then

$$\widehat{V}(\widehat{\gamma}) \approx \sum_{j=0}^k \sum_{j'=0}^k v_{jj'} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_{j'}}$$

15.3.4 Effect Displays

Let us write the GLM in matrix form, with linear predictor

$$\eta = \mathbf{X} \beta$$

and link function $g(\mu) = \eta$, where μ is the expectation of the response vector \mathbf{y} . As described in Section 15.3.2, we compute the maximum-likelihood estimate \mathbf{b} of β , along with the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$ of \mathbf{b} .

Let the rows of \mathbf{X}^* include regressors corresponding to all combinations of values of explanatory variables appearing in a high-order term of the model (or, for continuous explanatory variables, values spanning the ranges of the variables), along with typical values of the remaining regressors. The structure of \mathbf{X}^* with respect to interactions, for example, is the same as that of the model matrix \mathbf{X} . Then the fitted values $\widehat{\eta}^* = \mathbf{X}^* \mathbf{b}$ represent the high-order term in question, and a table or graph of these values—or, alternatively, of the fitted values transformed to the scale of the response variable, $g^{-1}(\widehat{\eta}^*)$ —is an effect display. The standard errors of $\widehat{\eta}^*$, available as the square-root diagonal entries of $\mathbf{X}^* \widehat{\mathcal{V}}(\mathbf{b}) \mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects, the endpoints of which may then also be transformed to the scale of the response. Even more generally, we can compute an effect display for a subset of explanatory variables, whether or not they correspond to a high-order term in the model.

For example, for the Poisson regression model fit to Ornstein's interlocking-directorate data, the effect display for assets in Figure 15.6(a) (page 430) is constructed by letting assets range between its minimum value of 0.062 and maximum of 147.670 billion dollars, fixing the dummy variables for nation of control and sector to their sample means—that is, to the observed proportions of the data in each of the corresponding categories of nation and sector. As noted previously, this is an especially simple example, because the model includes no interactions. The model was fit with the log link, and so the estimated effects, which in general are on the scale of the linear predictor, are on the log-count scale; the right-hand axis of the graph shows the corresponding count scale, which is the scale of the response variable.

Effect displays for GLMs are based on the fitted values $\widehat{\eta}^* = \mathbf{X}^* \mathbf{b}$, representing a high-order term in the model; that is, \mathbf{X}^* has the same general structure as the model matrix \mathbf{X} , with the explanatory variables in the high-term order ranging over their values in the data while other explanatory variables are set to typical values. The standard errors of $\widehat{\eta}^*$, given by the square-root diagonal entries of $\mathbf{X}^* \widehat{\mathcal{V}}(\mathbf{b}) \mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects.

15.4 Diagnostics for Generalized Linear Models

Most of the diagnostics for linear models presented in Chapters 11 and 12 extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least

squares, as described in Section 15.3.2. The final weighted-least-squares fit linearizes the model and provides a quadratic approximation to the log-likelihood. Approximate diagnostics are then either based directly on the WLS solution or are derived from statistics easily calculated from this solution. Seminal work on the extension of linear least-squares diagnostics to GLMs was done by Pregibon (1981); Landwehr, Pregibon, and Shoemaker (1980); Wang (1985, 1987); and Williams (1987). In my experience, and with the possible exception of added-variable plots for non-Gaussian GLMs, these extended diagnostics typically work reasonably well.

15.4.1 Outlier, Leverage, and Influence Diagnostics

Hat-Values

Hat-values, h_i , for a GLM can be taken directly from the final iteration of the IWLS procedure for fitting the model,⁵² and have the usual interpretation—except that, unlike in a linear model, the hat-values in a GLM depend on the response variable Y as well as on the configuration of the X s.

Residuals

Several kinds of residuals can be defined for GLMs:

- Most straightforwardly (but least usefully), *response residuals* are simply the differences between the observed response and its estimated expected value: $Y_i - \hat{\mu}_i$, where

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i) = g^{-1}(A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik})$$

- *Working residuals* are the residuals from the final WLS fit. These may be used to define partial residuals for component-plus-residual plots (see below).
- *Pearson residuals* are casewise components of the *Pearson goodness-of-fit statistic* for the model:⁵³

$$\frac{\hat{\phi}^{1/2}(Y_i - \hat{\mu}_i)}{\sqrt{\hat{V}(Y_i|\eta_i)}}$$

where $\hat{\phi}$ is the estimated dispersion parameter for the model (Equation 15.9 on 436) and $V(y_i|\eta_i)$ is the conditional variance of the response (given in Table 15.2 on page 421).

- *Standardized Pearson residuals* correct for the conditional response variation and for the differential leverage of the observations:

⁵²*The hat-matrix is

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{1/2}$$

where \mathbf{W} is the weight matrix from the final IWLS iteration.

⁵³The Pearson statistic, an alternative to the deviance for measuring the fit of the model to the data, is the sum of squared Pearson residuals.

$$R_{Pi} \equiv \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\nu}(Y_i|\eta_i)(1-h_i)}}$$

- *Deviance residuals*, G_i , are the square roots of the casewise components of the residual deviance (Equation 15.21 on page 449), attaching the sign of the corresponding response residual.
- *Standardized deviance residuals* are

$$R_{Gi} \equiv \frac{G_i}{\sqrt{\tilde{\phi}(1-h_i)}}$$

- Several different approximations to studentized residuals have been proposed. To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn and noting the decline in the deviance, a procedure that is computationally unattractive. Williams suggests the approximation

$$E_i^* \equiv \sqrt{(1-h_i)R_{Gi}^2 + h_i R_{Pi}^2}$$

where, once again, the sign is taken from the response residual. A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.

Influence Measures

An approximation to Cook's distance influence measure, due to Williams (1987), is

$$D_i \equiv \frac{R_{Pi}^2}{k+1} \times \frac{h_i}{1-h_i}$$

Approximate values of influence measures for individual coefficients, DFBETA_{ij} and DFBETAS_{ij}, may be obtained directly from the final iteration of the IWLS procedure.

Wang (1985) suggests an extension of added-variable plots to GLMs that works as follows: Suppose that the focal regressor is X_j . Refit the model with X_j removed, extracting the working residuals from this fit. Then regress X_j on the other X 's by WLS, using the weights from the last IWLS step, obtaining residuals. Finally, plot the working residuals from the first regression against the residuals for X_j from the second regression.

Figure 15.7 shows hat-values, studentized residuals, and Cook's distances for the quasi-Poisson model fit to Ornstein's interlocking directorate data. One observation—number 1, the corporation with the largest assets—stands out by combining a very large hat-value with the biggest absolute studentized residual.⁵⁴ This point is not a statistically significant outlier, however (indeed, the Bonferroni p -value for the largest studentized residual exceeds 1). As shown in the DFBETA plot in Figure 15.8, Observation 1 makes the coefficient of assets substantially smaller than it would otherwise be (recall that the coefficient for assets is 0.02085).⁵⁵ In this case, the approximate DFBETA is quite accurate: If Observation 1 is deleted, the assets

⁵⁴Unfortunately, the data source does not include the names of the firms, but Observation 1 is the largest of the Canadian banks, which, in the 1970s, was (I believe) the Royal Bank of Canada.

⁵⁵I invite the reader to plot the DFBETA values for the other coefficients in the model.

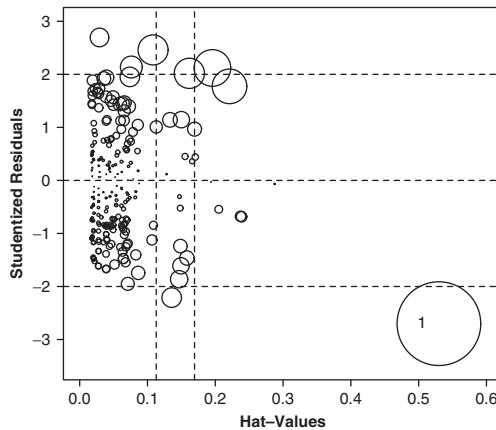


Figure 15.7 Hat-values, studentized residuals, and Cook's distances from the quasi-Poisson regression for Ornstein's interlocking-directorate data. The areas of the circles are proportional to the Cook's distances for the observations. Horizontal lines are drawn at -2 , 0 , and 2 on the studentized-residual scale, vertical lines at twice and three times the average hat-value.

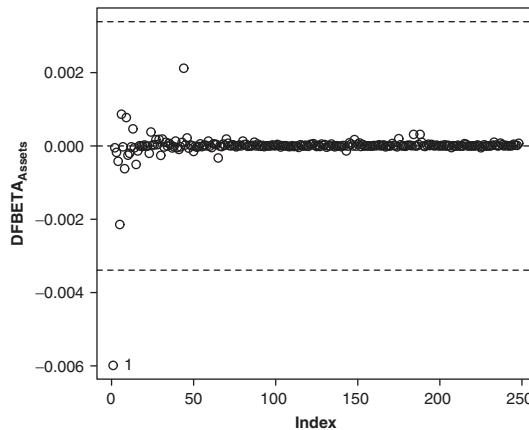


Figure 15.8 Index plot of DFBETA for the assets coefficient. The horizontal lines are drawn at 0 and $\pm \text{SE}(B_{\text{Assets}})$.

coefficient increases to 0.02602 . Before concluding that Observation 1 requires special treatment, however, consider the check for nonlinearity in the next section.

15.4.2 Nonlinearity Diagnostics

Component-plus-residual and CERES plots also extend straightforwardly to GLMs. Nonparametric smoothing of the resulting scatterplots can be important to interpretation,

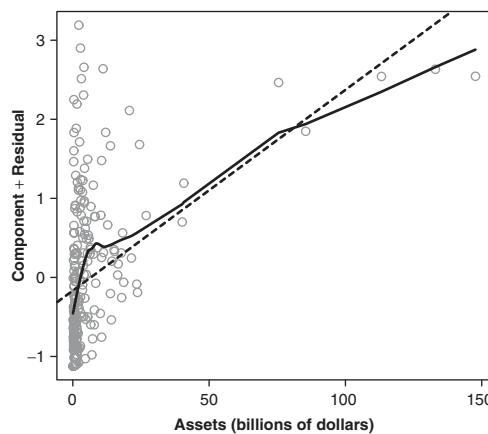


Figure 15.9 Component-plus-residual plot for assets in the interlocking-directorate quasi-Poisson regression. The broken line shows the least-squares fit to the partial residuals; the solid line is for a nonrobust lowess smooth with a span of 0.9.

especially in models for binary response variables, where the discreteness of the response makes the plots difficult to examine. Similar (if typically less extreme) effects can occur for binomial and count data.

Component-plus-residual and CERES plots use the linearized model from the last step of the IWLS fit. For example, the partial residual for X_j adds the working residual to $B_j X_{ij}$; the component-plus-residual plot then graphs the partial residual against X_j . In smoothing a component-plus-residual plot for a non-Gaussian GLM, it is generally preferable to use a nonrobust smoother.

A component-plus-residual plot for assets in the quasi-Poisson regression for the interlocking-directorate data is shown in Figure 15.9. Assets is so highly positively skewed that the plot is difficult to examine, but it is nevertheless apparent that the partial relationship between number of interlocks and assets is nonlinear, with a much steeper slope at the left than at the right. Because the bulge points to the left, we can try to straighten this relationship by transforming assets down the ladder of power and roots. Trial and error suggests the log transformation of assets, after which a component-plus-residual plot for the modified model (Figure 15.10) is unremarkable.

Box-Tidwell constructed-variable plots⁵⁶ also extend straightforwardly to GLMs: When considering the transformation of X_j , simply add the constructed variable $X_j \log_e X_j$ to the model and examine the added-variable plot for the constructed variable. Applied to assets in Ornstein's quasi-Poisson regression, this procedure produces the constructed-variable plot in Figure 15.11, which suggests that evidence for the transformation is spread throughout the data. The coefficient for assets $\times \log_e$ assets in the constructed-variable regression is -0.02177 with a standard error of 0.00371 ; the t -statistic for the constructed variable, $t_0 = -0.02177/0.00371 = -5.874$, therefore indicates strong evidence for the transformation of assets. By comparing the coefficient of assets in the *original* quasi-Poisson regression

⁵⁶See Section 12.5.2.

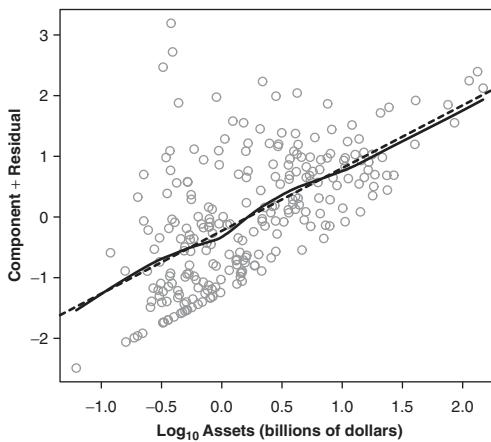


Figure 15.10 Component-plus-residual plot following the log transformation of assets. The lowess fit is for a span of 0.6.

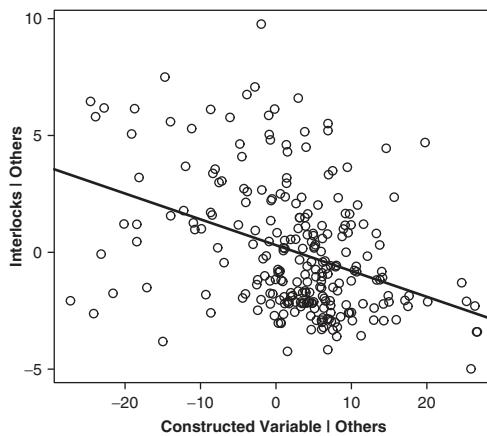


Figure 15.11 Constructed variable plot for the transformation of assets in the interlocking-directorate quasi-Poisson regression.

(0.02085) with the coefficient of the constructed variable, we get the suggested power transformation

$$\tilde{\lambda} = 1 + \frac{-0.02177}{0.02085} = -0.044$$

which is essentially the log transformation, $\lambda = 0$.

Finally, it is worth noting the relationship between the problems of influence and nonlinearity in this example: Observation 1 was influential in the original regression because its very large assets gave it high leverage and because unmodeled nonlinearity put the observation below the erroneously linear fit for assets, pulling the regression surface toward it. Log-transforming assets fixes both these problems.

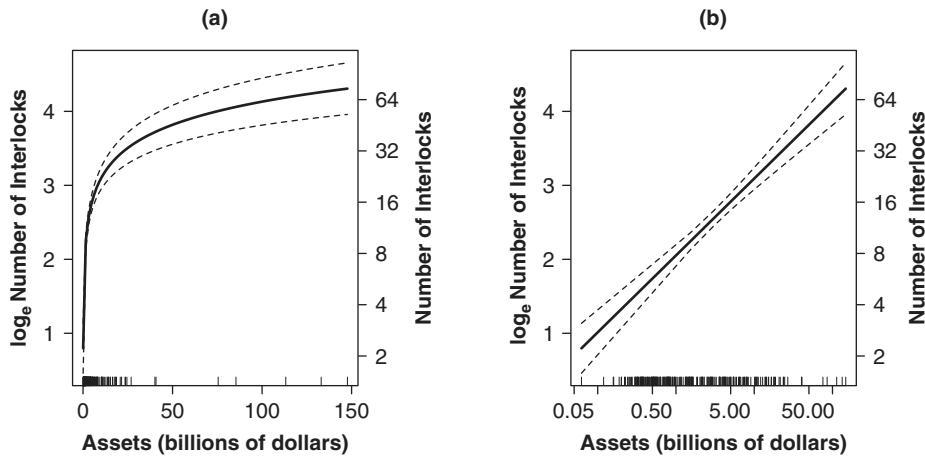


Figure 15.12 Effect displays for assets in the quasi-Poisson regression model in which assets has been log-transformed. Panel (a) plots assets on its “natural” scale, while panel (b) uses a log scale for assets. Rug-plots for assets appear at the bottom of the graphs. The broken lines give pointwise 95% confidence intervals around the estimated effect.

Alternative effect displays for assets in the transformed model are shown in Figure 15.12. Panel (a) in this figure graphs assets on its “natural” scale; on this scale, of course, the fitted partial relationship between log interlocks and assets is nonlinear. Panel (b) uses a log scale for assets, rendering the partial relationship linear.

15.4.3 Collinearity Diagnostics*

I mentioned in Chapter 13 that generalized variance-inflation factors, and consequently individual-coefficient variance-inflation factors, can be computed from the correlation matrix of the estimated regression coefficients.⁵⁷ This result can be applied to generalized linear models, where the correlations among the coefficients are obtained from their covariance matrix, given in Equation 15.20 (page 448).

For example, for the quasi-Poisson model fit to Ornstein’s interlocking-directorate data, with assets log-transformed, I computed the following generalized variance-inflation factors:

Term	GVIF	df	GVIF ^{1/2df}
log(Assets)	2.617	1	1.618
Nation of Control	1.619	3	1.084
Sector	3.718	9	1.076

⁵⁷ See footnote 25 on page 358.

As explained in Section 13.1.2, taking the $2df$ root of the GVIF is analogous to taking the square root of the VIF and makes generalized variance-inflation factors comparable across dimensions. None of these GVIFs are very large.

Most of the standard diagnostics for linear models extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least squares. Available diagnostics include studentized residuals, hat-values, Cook's distances, DFBETA and DFBETAS, added-variable plots, component-plus-residual plots, and the constructed-variable plot for transforming an explanatory variable.

15.5 Analyzing Data From Complex Sample Surveys

The purpose of this section is to introduce basic ideas about analyzing data from complex sample surveys. An in-depth consideration of statistical inference in complex surveys is well beyond the scope of this text, and I refer the reader to the recommended readings for details. While the placement of this material in the present chapter is essentially arbitrary—after all, the considerations raised here apply equally to all of the statistical models considered in this book—generalized linear models provide a suitably rich context in which to develop the topic.

In Chapter 1, I mentioned in passing the distinction between model-based and design-based inference: In *model-based inference*, we seek to draw conclusions (as I argued, at least in the first instance, descriptive conclusions) about the process generating the data. In *design-based inference*, the object is to estimate characteristics of a real population. Suppose, for example, that we are interested in establishing the difference in mean income between employed women and men. If the object of inference is the *real population* of employed Canadians at a particular point in time, then we could in principle compute the mean difference in income between women and men exactly if we had access to a census of the whole population. If, on the other hand, we are interested in the social *process* that generated the population, even a value computed from a census would represent an estimate, inasmuch as that process could have produced a different observed outcome. The methods discussed in this section deal with design-based inference about a real, finite population from which a survey sample is drawn. Consequently, the objects of inference are characteristics such as regression coefficients that could in principle be computed exactly with access to the whole population.

Other than in this section, the methods of statistical inference discussed in this book are appropriate for design-based inference only for *independent random samples*, that is, samples in which the observations are drawn with equal probability from the population and entirely independently of each other. Independence implies that once drawn into the sample, an individual is replaced in the population and therefore could in principle be selected more than once.

In practice, survey samples are never drawn with replacement, and the most basic survey-sampling design is *simple random sampling*, in which individuals are drawn with equal probability from those yet to be selected—that is, equal-probability random sampling *without replacement*. Letting N represent the number of individuals in the population and n the number

in the sample, in simple random sampling, all subsets of n individuals are equally likely to constitute the sample that is selected. Unless the sample size is a substantial fraction of the population size, however, the dependencies induced among observations in a simple random sample are trivially small, and data from such a sample may, as a practical matter, be analyzed as if they constituted an independent random sample. If, alternatively, the *sampling fraction* n/N is larger than about 5%, simple random sampling will be noticeably more efficient than independent random sampling: Standard errors of statistics in a simple random sample are deflated by the factor $\sqrt{1 - n/N}$, called the *finite population correction*. In the extreme, if the entire population is surveyed, then $n = N$, there is no sampling uncertainty, and the finite population correction is 0.

Complex survey samples often employ *unequal probabilities of selection*, *stratification*, *clustering*, and *multistage sampling*, all of which complicate the analysis of the data:

- In both independent random sampling and simple random sampling, each individual in the population has an equal probability n/N of being selected into the sample. This is not true, however, of all random sampling designs, and for a variety of reasons, different individuals may have different probabilities of selection.⁵⁸ For example, in the 2011 Canadian Election Study (CES) campaign-period survey,⁵⁹ individuals in small provinces had by design a higher probability of selection than those in more populous provinces to increase the precision of interprovincial comparisons. Random sampling requires that each individual in the population has a knowable nonzero probability of selection, not that these probabilities are all equal.

In design-based inference, where, as explained, the object is to estimate characteristics of a real population, unequal probabilities of selection are compensated by differential *weighting* in the computation of estimates and their sampling variability: To produce unbiased estimates of population characteristics, sampled observations with a higher probability of selection are down-weighted relative to those with a lower probability of selection, with the weight for each observation inversely proportional to its probability of selection.⁶⁰

Survey weights can also be employed to compensate both for sampling variation and for differential global nonresponse by matching samples to known characteristics of the population obtained, for example, from the census, beyond those characteristics employed to define strata. This procedure is termed *postweighting*.

- In *stratified sampling*, the population is divided into subpopulations, called *strata*, and a sample of predetermined size is selected from each stratum. The strata sample sizes may be proportional to population size, in which case the sampling fractions in the various strata are equal, or they may be disproportional, as in the CES survey. With strata subsamples proportional to population size, a stratified sample is usually modestly more

⁵⁸Some complex survey samples are explicitly designed to ensure that each individual in the population has an equal probability of selection; such designs are termed *self-weighting*.

⁵⁹The 2011 Canadian Election Study campaign-period survey is used for an example later in this section. See Fournier, Cutler, Soroka, and Stolle (2013) and Northrup (2012) for information on the CES.

⁶⁰Sampling weights are often confused with the inverse-variance weights employed to deal with nonconstant error variance in weighted-least-squares regression (as discussed in Section 12.2.2), where high-variance observations are down-weighted relative to low-variance observations. The two sets of weights have distinct purposes, and in general, WLS regression cannot be employed to obtain correct coefficient standard errors from weighted sample-survey data, although WLS usually produces correct point estimates of regression coefficients.

efficient (i.e., produces samples more representative of the population and consequently lower-variance estimates) than a simple random of equivalent size. Stratification induces typically small dependencies among the observations in a sample.

- In *cluster sampling*, individuals (or other sampled units) are selected into the sample in related subgroups, called *clusters*. For example, in the CES survey, a cluster comprises all eligible voters in a household. Individuals in a cluster tend to be more similar to one another than unrelated randomly selected individuals. These intracluster dependencies are typically substantial, and thus a cluster sample that includes more than one individual in each cluster is almost always considerably *less* efficient than a simple random sample of the same size. Clustering may be employed for reasons of cost—for example, to reduce the travel costs associated with face-to-face interviewing, where a cluster may comprise several adjacent households—or to facilitate sample selection—as, for example, in the CES, where the number of eligible individuals in each household is unknown prior to selection of a household cluster into the sample.⁶¹
- In a *multistage sampling design*, there is more than one step involving random selection. The CES campaign-period survey, for example, had a relatively simple two-stage design: In Stage 1, the Canadian population was divided into provincial strata, and random-digit dialing was employed to select a sample of household phone numbers in each stratum. If there was more than one eligible individual in a contacted household, a random procedure was employed in Stage 2 to select one of these individuals. Consequently, individuals' probability of selection into the sample was at the second stage of sampling inversely proportional to the number of eligible voters in their households; the survey weights for the CES take this factor into account, along with disproportional sampling by strata.

To illustrate design-based inference, I will use data drawn from the preelection survey of the 2011 Canadian Election Study. As I have explained, this survey employed a two-stage stratified-sampling design with household clusters selected in the first stage and an eligible individual selected from each sampled household in the second stage. Because only one individual was interviewed in each cluster, dependencies among observations in the sample are small and one could, in my opinion, reasonably analyze the data using either model-based or design-based inference. The principal difference between the two approaches to the CES data is the use of substantially different sampling weights employed in the design-based approach.

The CES survey included a direct question about abortion rights: “Should abortion be banned?” with responses “yes” and “no.” The (unweighted) number of individuals answering this question, along with others used as explanatory variables (see below), was 2231; of these, both the weighted and unweighted percentages answering yes were 18.5%.

I fit two straightforward additive logistic regressions to the data, for which the answer to the abortion-rights question is the response variable: one model employing maximum-likelihood estimates computed assuming independently sampled observations and the other computing estimates and standard errors based on the design of the CES survey.⁶² The explanatory variables for these logistic regressions, selected after some exploration of the data, are

⁶¹Clustering may itself be a key characteristic of data collection and statistical modeling, not simply a complication of sampling: See the discussion of mixed-effects models in Chapter 23 and 24.

⁶²I used the **survey** package for the R statistical computing environment (Lumley, 2010) to perform the design-based computations.

Table 15.10 Model-Based and Design-Based Estimates of Logistic Regressions for Agreement With Banning Abortion, Using the 2011 Canadian Election Study Survey

Coefficient	Model-Based		Design-Based	
	Estimate	Std. Error	Estimate	Std. Error
Intercept	-2.170	0.269	-2.270	0.321
<i>Importance of Religion</i>				
Not very important	0.442	0.310	0.458	0.348
Somewhat important	1.203	0.235	1.327	0.271
Very important	2.977	0.225	3.141	0.262
<i>Gender</i>				
Female	-0.375	0.127	-0.328	0.148
<i>Education</i>				
High school	-0.322	0.194	-0.445	0.238
Some postsecondary	-0.651	0.235	-0.852	0.290
College degree	-0.508	0.199	-0.562	0.240
Bachelor's degree	-0.901	0.208	-0.980	0.250
Postgraduate degree	-0.937	0.266	-0.675	0.309
<i>Urban/Rural Residence</i>				
Urban	-0.306	0.136	-0.283	0.166

- The answer to the question, “In your life, would you say religion is very important, somewhat important, not very important, or not important at all?” Individuals who said that they had no religion were not asked this question and were assigned the response “not important at all.” Four dummy regressors were generated from this variable, with “not important at all” as the baseline category.
- Gender, coded as female or male, with the latter as the baseline category for a dummy regressor.
- Education, with levels less than high school, high school graduate, some postsecondary, community college or technical school graduate, bachelor’s degree, and postgraduate degree, represented by four dummy regressors, with less than high school as the baseline category.
- Urban or rural residence, represented by a dummy regressor with rural residence as the baseline category.

Table 15.10 compares the coefficient estimates and standard errors obtained by the two approaches. The coefficient estimates are reasonably similar, but the design-based standard errors are 12% to 23% larger than the corresponding model-based standard errors. Controlling for the other explanatory variables, agreement with banning abortion increases with the reported importance of religion, declines with education, and is lower for women and for urban residents. Table 15.11 shows Wald chi-square tests for the terms in the logistic-regression models, computed from the coefficient estimates and their estimated variances and covariances.⁶³ As a consequence of their larger standard errors, the *p*-values for the design-based tests are larger than for the

⁶³See Section 15.3.3. I show Wald tests rather than likelihood-ratio tests for the model-based approach for greater comparability to the design-based results, where likelihood-ratio tests are not available.

Table 15.11 Wald Tests for Terms in the Logistic Regressions Fit to the 2011 Canadian Election Study Survey

Term	df	Model-Based		Design-Based	
		Wald Chi-square	p-value	Wald Chi-square	p-value
Importance of Religion	3	311.32	<.0001	253.61	<.0001
Gender	1	8.67	.0032	4.89	.0270
Education	5	25.28	.0001	17.83	.0032
Urban/Rural Residence	1	5.08	.0241	2.90	.0885

model-based tests, and the effect of urban versus rural residence is not statistically significant in the design-based model (by a two-sided test).

Design-based inference in sample surveys has as its object the estimation of characteristics of a real, finite population based on a sample drawn so that all individuals in the population have known, nonzero, though not necessarily equal, probabilities of selection into the sample. When probabilities of selection are unequal, survey weights are employed to obtain unbiased estimates of population values. Survey samples are often characterized by stratification of the population into subpopulations prior to sample selection, by random selection of clusters of related individuals, and by multistage sampling, in which there is more than one step involving random selection.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 15.1. Testing overdispersion: Let $\delta \equiv 1/\omega$ represent the inverse of the scale parameter for the negative-binomial regression model (see Equation 15.4 on page 432). When $\delta = 0$, the negative-binomial model reduces to the Poisson regression model (why?), and consequently a test of $H_0: \delta = 0$ against the one-sided alternative hypothesis $H_a: \delta > 0$ is a test of overdispersion. A Wald test of this hypothesis is straightforward, simply dividing $\hat{\delta}$ by its standard error. We can also compute a likelihood-ratio test contrasting the deviance under the more specific Poisson regression model with that under the more general negative-binomial model. Because the negative-binomial model has one additional parameter, we refer the likelihood-ratio test statistic to a chi-square distribution with 1 degree of freedom; as Cameron and Trivedi (1998, p. 78) explain, however, the usual right-tailed p -value obtained from the chi-square distribution must be halved. Apply this likelihood-ratio test for overdispersion to Ornstein's interlocking-directorate regression.

Exercises 15.2. *Zero-inflated count regression models:

- (a) Show that the mean and variance of the response variable Y_i in the zero-inflated Poisson (ZIP) regression model, given in Equations 15.5 and 15.6 on page 433, are

$$\begin{aligned} E(Y_i) &= (1 - \pi_i)\mu_i \\ V(Y_i) &= (1 - \pi_i)\mu_i(1 + \pi_i\mu_i) \end{aligned}$$

(Hint: Recall that there are two sources of 0s: observations in the first latent class, whose value of Y_i is *necessarily* 0, and observations in the second latent class, whose value *may be* 0. Probability of membership is π_i in the first class and $1 - \pi_i$ in the second.) Show that $V(Y_i) > E(Y_i)$ when $\pi_i > 0$.

- (b) Derive the log-likelihood for the ZIP model, given in Equation 15.7 (page 434).
(c) The *zero-inflated negative-binomial* (ZINB) *regression model* substitutes a negative-binomial GLM for the Poisson-regression submodel of Equations 15.6 on page 433:

$$\begin{aligned} \log_e \mu_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\ p(y_i|x_1, \dots, x_k) &= \frac{\Gamma(y_i + \omega)}{y! \Gamma(\omega)} \times \frac{\mu_i^{y_i} \omega^\omega}{(\mu_i + \omega)^{\mu_i + \omega}} \end{aligned}$$

Show that $E(Y_i) = (1 - \pi_i)\mu_i$ (as in the ZIP model) and that

$$V(Y_i) = (1 - \pi_i)\mu_i[1 + \mu_i(\pi_i + 1/\omega)]$$

When $\pi_i > 0$, the conditional variance is greater in the ZINB model than in the standard negative-binomial GLM, $V(Y_i) = \mu_i + \mu_i^2/\omega$; why? Derive the log-likelihood for the ZINB model. [Hint: Simply substitute the negative-binomial GLM for the Poisson-regression submodel in Equation 15.7 (page 434)].

Exercise 15.3. The usual Pearson chi-square statistic for testing for independence in a two-way contingency table is

$$X_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

where the Y_{ij} are the observed frequencies in the table, and the $\hat{\mu}_{ij}$ are the estimated expected frequencies under independence. The estimated expected frequencies can be computed from the maximum-likelihood estimates for the loglinear model of independence, or they can be computed directly as $\hat{\mu}_{ij} = Y_{i+}Y_{+j}/n$. The likelihood-ratio statistic for testing for independence can also be computed from the estimated expected counts as

$$G_0^2 = 2 \sum_{i=1}^r \sum_{j=1}^c Y_{ij} \log_e \frac{Y_{ij}}{\hat{\mu}_{ij}}$$

Both test statistics have $(r - 1)(c - 1)$ degrees of freedom. The two tests are asymptotically equivalent and usually produce similar results. Applying these formulas to the two-way table for voter turnout and intensity of partisan preference in Table 15.4 (page 435), compute both

test statistics, verifying that the direct formula for G_0^2 produces the same result as given in the text. Do the Pearson and likelihood-ratio tests agree?

Exercise 15.4. *Show that the normal distribution can be written in exponential form as

$$p(y; \theta, \phi) = \exp \left\{ \frac{y\theta - \theta^2/2}{\phi} - \frac{1}{2} \left[\frac{y^2}{\phi} + \log_e(2\pi\phi) \right] \right\}$$

where $\theta = g_c(\mu) = \mu$; $\phi = \sigma^2$; $a(\phi) = \phi$; $b(\theta) = \theta^2/2$; and $c(y, \phi) = -\frac{1}{2}[y^2/\phi + \log_e(2\pi\phi)]$.

Exercise 15.5. *Show that the binomial distribution can be written in exponential form as

$$p(y; \theta, \phi) = \exp \left[\frac{y\theta - \log_e(1 + e^\theta)}{1/n} + \log_e \binom{n}{ny} \right]$$

where $\theta = g_c(\mu) = \log_e[\mu/(1 - \mu)]$, $\phi = 1$; $a(\phi) = 1/n$, $b(\theta) = \log_e(1 + e^\theta)$, and $c(y, \phi) = \log_e \binom{n}{ny}$.

Exercise 15.6. *Using the results given in Table 15.9 (on page 444), verify that the Poisson, gamma, and inverse-Gaussian families can all be written in the common exponential form

$$p(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

Exercise 15.7. *Using the general result that the conditional variance of a distribution in an exponential family is

$$V(Y) = a(\phi) \frac{d^2 b(\theta)}{d\theta^2}$$

and the values of $a(\cdot)$ and $b(\cdot)$ given in Table 15.9 (on page 444), verify that the variances of the Gaussian, binomial, Poisson, gamma, and inverse-Gaussian families are, consecutively, ϕ , $\mu(1 - \mu)/n$, μ , $\phi\mu^2$, and $\phi\mu^3$.

Exercise 15.8. *Show that the derivative of the log-likelihood for an individual observation with respect to the regression coefficients in a GLM can be written as

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij}, \text{ for } j = 0, 1, \dots, k$$

(See Equation 15.17 on page 446.)

Exercise 15.9. *Using the general expression for the residual deviance,

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^n \frac{Y_i[g(Y_i) - g(\hat{\mu}_i)] - b[g(Y_i)] + b[g(\hat{\mu}_i)]}{a_i}$$

show that the deviances for the several exponential families can be written in the following forms:

Family	Residual Deviance
Gaussian	$\sum (Y_i - \hat{\mu}_i)^2$
Binomial	$2 \sum \left[n_i Y_i \log_e \frac{Y_i}{\hat{\mu}_i} + n_i (1 - Y_i) \log_e \frac{1 - Y_i}{1 - \hat{\mu}_i} \right]$
Poisson	$2 \sum \left[Y_i \log_e \frac{Y_i}{\hat{\mu}_i} - (Y_i - \hat{\mu}_i) \right]$
Gamma	$2 \sum \left[-\log_e \frac{Y_i}{\hat{\mu}_i} + \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Inverse-Gaussian	$\sum \frac{(Y_i - \hat{\mu}_i)^2}{Y_i \hat{\mu}_i^2}$

Exercise 15.10. *Using the SLID data, Table 12.1 (on page 313) reports the results of a regression of log wages on sex, the square of education, a quadratic in age, and interactions between sex and education-squared, as well as between sex and the quadratic for age.

- (a) Estimate the age γ_1 at which women attain on average their highest level of wages, controlling for education. Use the delta method to estimate the standard error of $\hat{\gamma}_1$. *Note:* You will need to fit the model yourself to obtain the covariance matrix for the estimated regression coefficients, which is not given in the text.
- (b) Estimate the age γ_2 at which men attain on average their highest level of wages, controlling for education. Use the delta method to estimate the standard error of $\hat{\gamma}_2$.
- (c) Let $\gamma_3 \equiv \gamma_1 - \gamma_2$, the difference between the ages at which men and women attain their highest wage levels. Compute $\hat{\gamma}_3$. Use the delta method to find the standard error of $\hat{\gamma}_3$ and then test the null hypothesis $H_0: \gamma_3 = 0$.

Exercises 15.11. Coefficient quasi-variances: Coefficient quasi-variances for dummy-variable regressors were introduced in Section 7.2.1. Recall that the object is to approximate the standard errors for pairwise *differences* between categories,

$$\text{SE}(C_j - C_{j'}) = \sqrt{\tilde{V}(C_j) + \tilde{V}(C_{j'}) - 2 \times \tilde{C}(C_j, C_{j'})}$$

where C_j and $C_{j'}$ are two dummy-variable coefficients for an m -category polytomous explanatory variable, $\tilde{V}(C_j)$ is the estimated sampling variance of C_j , and $\tilde{C}(C_j, C_{j'})$ is the estimated sampling covariance of C_j and $C_{j'}$. By convention, we take C_m (the coefficient of the baseline category) and its standard error, $\text{SE}(C_m)$, to be 0. We seek coefficient quasi-variances $\tilde{V}(C_j)$, so that

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\tilde{V}(C_j) + \tilde{V}(C_{j'})}$$

for all pairs of coefficients C_j and $C_{j'}$, by minimizing the total log relative error of approximation, $\sum_{j < j'} [\log(\text{RE}_{jj'})]^2$, where

$$\text{RE}_{jj'} \equiv \frac{\tilde{V}(C_j - C_{j'})}{\tilde{V}(C_j - C_{j'})} = \frac{\tilde{V}(C_j) + \tilde{V}(C_{j'})}{\tilde{V}(C_j) + \tilde{V}(C_{j'}) - 2 \times \tilde{C}(C_j, C_{j'})}$$

Firth (2003) cleverly suggests implementing this criterion by fitting a GLM in which the response variable is $Y_{jj'} \equiv \log_e[\hat{V}(C_j - C_{j'})]$ for all unique pairs of categories j and j' ; the linear predictor is $\eta_{jj'} \equiv \beta_j + \beta_{j'}$; the link function is the exponential link, $g(\mu) = \exp(\mu)$ (which is, note, *not* one of the common links in Table 15.1); and the variance function is constant, $V(Y|\eta) = \phi$. The quasi-likelihood estimates of the coefficients β_j are the quasi-variances $\tilde{V}(C_j)$. For example, for the Canadian occupational prestige regression described in Section 7.2.1, where the dummy variables pertain to type of occupation (professional and managerial, white collar, or blue collar), we have

Pair (j, j')	$Y_{jj'} = \log_e[\hat{V}(C_j - C_{j'})]$
Professional, White Collar	$\log_e(2.771^2) = 2.038$
Professional, Blue Collar	$\log_e(3.867^2) = 2.705$
White Collar, Blue Collar	$\log_e(2.514^2) = 1.844$

and model matrix

$$\mathbf{X} = \begin{bmatrix} (\beta_1) & (\beta_2) & (\beta_3) \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

With three unique pairs and three coefficients, we should get a perfect fit: As I mentioned in Section 7.2.1, when there are only three categories, the quasi-variances perfectly recover the estimated variances for pairwise differences in coefficients. Demonstrate that this is the case by fitting the GLM. Some additional comments:

- The computation outlined here is the basis of Firth's **qvcalc** package (described in Firth, 2003) for the R statistical programming environment.
- The computation of quasi-variances applies not only to dummy regressors in linear models but also to all models with a linear predictor for which coefficients and their estimated covariance matrix are available—for example, the GLMs described in this chapter.
- Quasi-variances may be used to approximate the standard error for any linear combination of dummy-variable coefficients, not just for pairwise differences.
- Having found the quasi-variance approximations to a set of standard errors, we can then compute and report the (typically small) maximum relative error of these approximations. Firth and De Menezes (2004) give more general results for the maximum relative error for *any* contrast of coefficients.

Summary

- A generalized linear model (or GLM) consists of three components:
 1. A random component, specifying the conditional distribution of the response variable, Y_i (for the i th of n independently sampled observations), given the values of

the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions.

2. A linear predictor—that is, a linear function of regressors,

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{ik} X_k$$

3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{ik} X_k$$

- A convenient property of distributions in the exponential families is that the conditional variance of Y_i is a function of its mean μ_i and, possibly, a dispersion parameter ϕ . In addition to the familiar Gaussian and binomial families (the latter for proportions), the Poisson family is useful for modeling count data and the gamma and inverse-Gaussian families for modeling positive continuous data, where the conditional variance of Y increases with its expectation.
- GLMs are fit to data by the method of maximum likelihood, providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients.
- The ANOVA for linear models has an analog in the analysis of deviance for GLMs. The residual deviance for a GLM is $D_m \equiv 2(\log_e L_s - \log_e L_m)$, where L_m is the maximized likelihood under the model in question, and L_s is the maximized likelihood under a saturated model. The residual deviance is analogous to the residual sum of squares for a linear model.
- In GLMs for which the dispersion parameter is fixed to 1 (binomial and Poisson GLMs), the likelihood-ratio test statistic is the difference in the residual deviances for nested models and is asymptotically distributed as chi-square under the null hypothesis. For GLMs in which there is a dispersion parameter to estimate (Gaussian, gamma, and inverse-Gaussian GLMs), we can instead compare nested models by an incremental F -test.
- The basic GLM for count data is the Poisson model with log link. Frequently, however, when the response variable is a count, its conditional variance increases more rapidly than its mean, producing a condition termed *overdispersion* and invalidating the use of the Poisson distribution. The quasi-Poisson GLM adds a dispersion parameter to handle overdispersed count data; this model can be estimated by the method of quasi-likelihood. A similar model is based on the negative-binomial distribution, which is not an exponential family. Negative-binomial GLMs can nevertheless be estimated by maximum likelihood. The zero-inflated Poisson regression model may be appropriate when there are more zeroes in the data than is consistent with a Poisson distribution.
- Loglinear models for contingency tables bear a formal resemblance to ANOVA models and can be fit to data as Poisson GLMs with a log link. The loglinear model for a contingency table, however, treats the variables in the table symmetrically—none of the variables is distinguished as a response variable—and consequently the parameters of the model represent the associations among the variables, not the effects of explanatory

variables on a response. When one of the variables is construed as the response, the log-linear model reduces to a binomial or multinomial logit model.

- The Gaussian, binomial, Poisson, gamma, and inverse-Gaussian distributions can all be written in the common linear-exponential form:

$$p(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions that vary from one exponential family to another; $\theta = g_c(\mu)$ is the canonical parameter for the exponential family in question; $g_c(\cdot)$ is the canonical link function; and $\phi > 0$ is a dispersion parameter, which takes on a fixed, known value in some families. It is generally the case that $\mu = E(Y) = b'(\theta)$ and that $V(Y) = a(\phi)b''(\theta)$.

- The maximum-likelihood estimating equations for generalized linear models take the common form

$$\sum_{i=1}^n \frac{Y_i - \mu_i}{a_i v(\mu_i)} \times \frac{d\mu_i}{d\eta_i} \times x_{ij} = 0, \text{ for } j = 0, 1, \dots, k$$

These equations are generally nonlinear and therefore have no general closed-form solution, but they can be solved by iterated weighted least squares (IWLS). The estimating equations for the coefficients do not involve the dispersion parameter, which (for models in which the dispersion is not fixed) then can be estimated as

$$\tilde{\phi} = \frac{1}{n - k - 1} \sum \frac{(Y_i - \hat{\mu}_i)^2}{a_i v(\hat{\mu}_i)}$$

The estimated asymptotic covariance matrix of the coefficients is

$$\widehat{\mathcal{V}}(\mathbf{b}) = \tilde{\phi} (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}$$

where \mathbf{b} is the vector of estimated coefficients and \mathbf{W} is a diagonal matrix of weights from the last IWLS iteration.

- The maximum-likelihood estimating equations, and IWLS estimation, can be applied whenever we can express the transformed mean of Y as a linear function of the X s and can write the conditional variance of Y as a function of its mean and (possibly) a dispersion parameter—even when we do not specify a particular conditional distribution for Y . The resulting quasi-likelihood estimator shares many of the properties of maximum-likelihood estimators.
- The residual deviance for a model is twice the difference in the log-likelihoods for the saturated model, which dedicates one parameter to each observation, and the model in question:

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\mu}}) &\equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})] \\ &= 2 \sum_{i=1}^n \frac{Y_i[g(Y_i) - g(\hat{\mu}_i)] - b[g(Y_i)] + b[g(\hat{\mu}_i)]}{a_i} \end{aligned}$$

Dividing the residual deviance by the estimated dispersion parameter produces the scaled deviance, $D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) \equiv D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\tilde{\phi}$.

- To test the general linear hypothesis $H_0: \mathbf{L}\beta = \mathbf{c}$, where the hypothesis matrix \mathbf{L} has q rows, we can compute the Wald chi-square test statistic

$$Z_0^2 = (\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})$$

with q degrees of freedom. Alternatively, if the dispersion parameter is estimated from the data, we can compute the F -test statistic

$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})' [\mathbf{L}\widehat{\mathcal{V}}(\mathbf{b})\mathbf{L}']^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})}{q}$$

on q and $n - k - 1$ degrees of freedom.

- The delta method may be used to approximate the standard error of a nonlinear function of regression coefficients in a GLM. If $\gamma \equiv f(\beta_0, \beta_1, \dots, \beta_k)$, then

$$\widehat{\mathcal{V}}(\widehat{\gamma}) \approx \sum_{j=0}^k \sum_{j'=0}^k v_{jj'} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_j} \times \frac{\partial \widehat{\gamma}}{\partial \widehat{\beta}_{j'}}$$

- Effect displays for GLMs are based on the fitted values $\widehat{\eta}^* = \mathbf{X}^*\mathbf{b}$, representing a high-order term in the model; that is, \mathbf{X}^* has the same general structure as the model matrix \mathbf{X} , with the explanatory variables in the high-term order ranging over their values in the data, while other explanatory variables are set to typical values. The standard errors of $\widehat{\eta}^*$, given by the square-root diagonal entries of $\mathbf{X}^*\widehat{\mathcal{V}}(\mathbf{b})\mathbf{X}^{*\prime}$, may be used to compute pointwise confidence intervals for the effects.
- Most of the standard diagnostics for linear models extend relatively straightforwardly to GLMs. These extensions typically take advantage of the computation of maximum-likelihood and quasi-likelihood estimates for GLMs by iterated weighted least squares. Available diagnostics include studentized residuals, hat-values, Cook's distances, DFBETA and DFBETAS, added-variable plots, component-plus-residual plots, and the constructed-variable plot for transforming an explanatory variable.
- Design-based inference in sample surveys has as its object the estimation of characteristics of a real, finite population based on a sample drawn so that all individuals in the population have known, nonzero, though not necessarily equal, probabilities of selection into the sample. When probabilities of selection are unequal, survey weights are employed to obtain unbiased estimates of population values. Survey samples are often characterized by stratification of the population into subpopulations prior to sample selection, by random selection of clusters of related individuals, and by multistage sampling in which there is more than one step involving random selection.

Recommended Reading

- McCullagh and Nelder (1989), the “bible” of GLMs, is a rich and interesting—if generally difficult—text.
- Dobson (2001) presents a much briefer overview of generalized linear models at a more moderate level of statistical sophistication.

- Aitkin, Francis, and Hinde's (2005) text, geared to the statistical computer package GLIM for fitting GLMs, is still more accessible.
- A chapter by Firth (1991) is the best brief treatment of generalized linear models that I have read.
- Long (1997) includes an excellent presentation of regression models for count data (though not from the point of view of GLMs); an even more extensive treatment may be found in Cameron and Trivedi (1998).
- Groves et al. (2009) present a wide-ranging overview of survey-research methods, Fuller (2009) describes the details of estimation in complex sample surveys, and Lumley (2010) offers an accessible, compact treatment of the topic, focused on the **survey** package for R but of more general interest. Also see the edited volume by Skinner, Holt, and Smith (1989) on analyzing data from complex sample surveys.

PART V

Extending Linear and Generalized Linear Models

16

Time-Series Regression and Generalized Least Squares*

This part of the book introduces several important extensions of linear least-squares regression and generalized linear models:

- The current chapter describes the application of linear regression models to time-series data in which the errors are correlated over time rather than independent.
- Nonlinear regression, the subject of Chapter 17, fits a specific nonlinear function of the explanatory variables by least squares.
- Chapter 18 develops nonparametric regression analysis, introduced in Chapter 2, which does not assume a specific functional form relating the response variable to the explanatory variables (as do traditional linear, generalized linear, and nonlinear regression models).
- Chapter 19 takes up robust regression analysis, which employs criteria for fitting a linear model that are not as sensitive as least squares to unusual data.

Taken together, the methods in the first four chapters of Part V considerably expand the range of application of regression analysis.

The standard linear model of Chapters 5 through 10 assumes independently distributed errors. The assumption of independence is rarely (if ever) quite right, but it is often a reasonable approximation. When the observations comprise a *time series*, however, dependencies among the errors can be very strong.

In time-series data, a single unit of observation (person, organization, nation, etc.) is tracked over many time periods or points of time.¹ These time periods or time points are usually evenly spaced, at least approximately, and I will assume here that this is the case. Economic statistics for Canada, for example, are reported on a daily, monthly, quarterly, and yearly basis. Crime statistics, likewise, are reported on a yearly basis. Later in this section, we will use yearly time series for the period 1931 to 1968 to examine the relationship between Canadian women's crime rates and fertility, women's labor force participation, women's participation in higher education, and men's crime rates.

It is not generally reasonable to suppose that the errors in a time-series regression are independent: After all, time periods that are close to one another are more likely to be similar than time periods that are relatively remote. This similarity may well extend to the errors, which

¹Temperature, for example, may be recorded at evenly spaced time points. Gross national product is cumulated over the period of a year. Most social data are collected in time periods rather than at time points.

represent (most important) the omitted causes of the response variable. Although the time dependence among the errors may turn out to be negligible, it is unwise to assume a priori that this is the case.

In time-series data, a single individual is tracked over many time periods or points of time. It is not generally reasonable to suppose that the errors in a time-series regression are independent.

16.1 Generalized Least-Squares Estimation

I will first address dependencies among the errors in a very general context. Consider the usual linear model,

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times k+1)}{\mathbf{X}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}$$

Rather than assuming that the errors are independently distributed, however, let us instead assume that

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon})$$

where the order- n matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is symmetric and positive definite. Nonzero off-diagonal entries in the covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ correspond to correlated errors.²

To capture serial dependence among the errors in the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we drop the assumption that the errors are independent of one another; instead, we assume that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon})$, where nonzero off-diagonal entries in the error covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ correspond to correlated errors.

Let us assume unrealistically (and only temporarily) that we know $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$. Then the log-likelihood for the model is³

$$\log_e L(\boldsymbol{\beta}) = -\frac{n}{2} \log_e 2\pi - \frac{1}{2} \log_e (\det \boldsymbol{\Sigma}_{\varepsilon\varepsilon}) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (16.1)$$

It is clear that the log-likelihood is maximized when the *generalized sum of squares* $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.⁴ Differentiating the generalized sum of squares with

²Because of the assumption of normality, dependence implies correlation. Like the standard linear model with independent errors, however, most of the results of this section do not require the assumption of normality. Unequal *diagonal* entries of $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ correspond to unequal error variances, a problem discussed in Section 12.2. Indeed, weighted least-squares regression (Section 12.2.2) is a special case of generalized least-squares estimation, where $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is a diagonal matrix.

³See Exercise 16.1 for this and other results described in this section.

⁴Recall that $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is assumed to be known.

respect to β , setting the partial derivatives to $\mathbf{0}$, and solving for β produces the *generalized least-squares (GLS) estimator*

$$\mathbf{b}_{\text{GLS}} = (\mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{y} \quad (16.2)$$

It is simple to show that the GLS estimator is unbiased, $E(\mathbf{b}_{\text{GLS}}) = \beta$; that its sampling variance is

$$V(\mathbf{b}_{\text{GLS}}) = (\mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1}$$

and that, by an extension of the Gauss-Markov theorem, \mathbf{b}_{GLS} is the minimum-variance linear unbiased estimator of β .⁵ None of these results (with the exception of the one establishing the GLS estimator as the ML estimator) requires the assumption of normality.

Here is another way of thinking about the GLS estimator: Let $\boldsymbol{\Gamma}_{(n \times n)}$ be a “square-root” of $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}$, in the sense that $\boldsymbol{\Gamma}' \boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}$.⁶ From Equation 16.2,

$$\begin{aligned} \mathbf{b}_{\text{GLS}} &= (\mathbf{X}' \boldsymbol{\Gamma}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Gamma}' \mathbf{y} \\ &= (\mathbf{X}^* \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{y}^* \end{aligned}$$

where $\mathbf{X}^* \equiv \boldsymbol{\Gamma} \mathbf{X}$ and $\mathbf{y}^* \equiv \boldsymbol{\Gamma} \mathbf{y}$. Thus, the GLS estimator is the *ordinary-least-squares (OLS)* estimator for the regression of \mathbf{y}^* on \mathbf{X}^* —that is, following the linear transformation of \mathbf{y} and \mathbf{X} using the transformation matrix $\boldsymbol{\Gamma}$.

If the error covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is known, then the maximum-likelihood (ML) estimator of β is the generalized least-squares estimator $\mathbf{b}_{\text{GLS}} = (\mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{y}$. The sampling variance-covariance matrix of \mathbf{b}_{GLS} is $V(\mathbf{b}_{\text{GLS}}) = (\mathbf{X}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1}$. The generalized least-squares estimator can also be expressed as the OLS estimator $(\mathbf{X}^* \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{y}^*$ for the transformed variables $\mathbf{X}^* \equiv \boldsymbol{\Gamma} \mathbf{X}$ and $\mathbf{y}^* \equiv \boldsymbol{\Gamma} \mathbf{y}$, where the transformation matrix $\boldsymbol{\Gamma}$ is a square root of $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}$.

16.2 Serially Correlated Errors

I have, thus far, left the covariance matrix of the errors $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ very general: Because of its symmetry, there are $n(n + 1)/2$ distinct elements in $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$. Without further assumptions concerning the structure of this matrix, we cannot hope to estimate its elements from only n observations if—as is always the case in real applications of time-series regression— $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is not known.

Suppose, however, that the process generating the errors is *stationary*. Stationarity means that the errors all have the same expectation (which, indeed, we have already assumed to be 0), that the errors have a common variance (σ_ε^2), and that the covariance of two errors depends only on their separation in time. Let ε_t denote the error for time period t and ε_{t+s} the error for

⁵The Gauss-Markov theorem is discussed in Section 9.3.2 in the context of ordinary least-squares regression.

⁶Because $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}$ is nonsingular, it is always possible to find a square-root matrix, although the square root is not in general unique; see online Appendix B on matrices and linear algebra.

time period $t + s$ (where s is an integer—positive, negative, or 0). Stationarity implies that, for any t , the covariance between ε_t and ε_{t+s} is

$$C(\varepsilon_t, \varepsilon_{t+s}) = E(\varepsilon_t \varepsilon_{t+s}) = \sigma_\varepsilon^2 \rho_s = C(\varepsilon_t, \varepsilon_{t-s})$$

where ρ_s , called the *autocorrelation* (or *serial correlation*) at lag s , is the correlation between two errors separated by $|s|$ time periods.

The error covariance matrix, then, has the following pattern:

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{P} \quad (16.3)$$

The situation is much improved, but it is not good enough: There are now n distinct parameters to estimate in $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ —that is, σ_ε^2 and $\rho_1, \dots, \rho_{n-1}$ —still too many.

When, more realistically, the error covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is unknown, we need to estimate its contents along with the regression coefficients β . Without restricting its form, however, $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ contains too many distinct elements to estimate directly. Assuming that the errors are generated by a stationary time-series process reduces the number of independent parameters in $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ to n , including the error variance σ_ε^2 and the autocorrelations at various lags, $\rho_1, \dots, \rho_{n-1}$.

16.2.1 The First-Order Autoregressive Process

To proceed, we need to specify a stationary process for the errors that depends on fewer parameters. The process that is by far most commonly used in practice is the *first-order autoregressive process*, abbreviated *AR(1)*:

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t \quad (16.4)$$

where the error in time period t depends directly only on the error in the previous time period, ε_{t-1} , and on a random contemporaneous “shock” ν_t . Unlike the regression errors ε_t , we will assume that the random shocks ν_t are independent of each other (and of ε s from earlier time periods) and that $\nu_t \sim N(0, \sigma_\nu^2)$. Serial correlation in the regression errors, therefore, is wholly generated by the partial dependence of each error on the error of the previous time period. Because of its importance in applications and its simplicity, I will describe the AR(1) process in some detail.

For Equation 16.4 to specify a stationary process, it is necessary that $|\rho| < 1$. Otherwise, the errors will tend to grow without bound. If the process is stationary, and if all errors have 0 expectations and common variance, then

$$\begin{aligned} \sigma_\varepsilon^2 &\equiv V(\varepsilon_t) = E(\varepsilon_t^2) \\ &= V(\varepsilon_{t-1}) = E(\varepsilon_{t-1}^2) \end{aligned}$$

Squaring both sides of Equation 16.4 and taking expectations,

$$\begin{aligned} E(\varepsilon_t^2) &= \rho^2 E(\varepsilon_{t-1}^2) + E(\nu_t^2) + 2\rho E(\varepsilon_{t-1}\nu_t) \\ \sigma_\varepsilon^2 &= \rho^2 \sigma_\varepsilon^2 + \sigma_\nu^2 \end{aligned}$$

because $E(\varepsilon_{t-1}\nu_t) = C(\varepsilon_{t-1}, \nu_t) = 0$. Solving for the variance of the regression errors yields

$$\sigma_\varepsilon^2 = \frac{\sigma_\nu^2}{1 - \rho^2}$$

It is also a simple matter to find the autocorrelation at lag s . For example, at lag 1, we have the *autocovariance*

$$\begin{aligned} C(\varepsilon_t, \varepsilon_{t-1}) &= E(\varepsilon_t \varepsilon_{t-1}) \\ &= E[(\rho \varepsilon_{t-1} + \nu_t) \varepsilon_{t-1}] \\ &= \rho \sigma_\varepsilon^2 \end{aligned}$$

So the autocorrelation at lag 1 is just

$$\begin{aligned} \rho_1 &= \frac{C(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{V(\varepsilon_t) \times V(\varepsilon_{t-1})}} \\ &= \frac{\rho \sigma_\varepsilon^2}{\sigma_\varepsilon^2} \\ &= \rho \end{aligned}$$

Likewise, at lag 2,

$$\begin{aligned} C(\varepsilon_t, \varepsilon_{t-2}) &= E(\varepsilon_t \varepsilon_{t-2}) \\ &= E\{[\rho(\rho \varepsilon_{t-2} + \nu_{t-1}) + \nu_t] \varepsilon_{t-2}\} \\ &= \rho^2 \sigma_\varepsilon^2 \end{aligned}$$

and, therefore, $\rho_2 = \rho^2$.

More generally, for the first-order autoregressive process, $\rho_s = \rho^s$, and because $|\rho| < 1$, the autocorrelations of the errors decay exponentially toward 0 as the lag s gets larger. This behavior is apparent in the examples in Figure 16.1, which shows AR(1) *autocorrelation functions* for $\rho = .9$ and $\rho = -.7$. Note that the autocorrelation at lag 0 is $\rho_0 = 1$.

To reduce the number of parameters in $\Sigma_{\varepsilon\varepsilon}$ further, we can adopt a specific time-series model for the errors. The most commonly employed such model is the first-order autoregressive process $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$, where $|\rho| < 1$ and the random shocks ν_t are independently distributed as $N(0, \sigma_\nu^2)$. Under this specification, two errors, ε_t and ε_{t+s} , separated by s time periods have autocovariance $\rho^s \sigma_\varepsilon^2$ and autocorrelation ρ^s . The variance of the regression errors is $\sigma_\varepsilon^2 = \sigma_\nu^2 / (1 - \rho^2)$.

Some “realizations” of time series generated by Equation 16.4, with $\nu_t \sim N(0, 1)$, are shown in Figure 16.2. In Figure 16.2(a), $\rho = 0$ and, consequently, the ε_t are uncorrelated, a time-series process sometimes termed *white noise*. In Figure 16.2(b), $\rho = .9$; note how values of the series

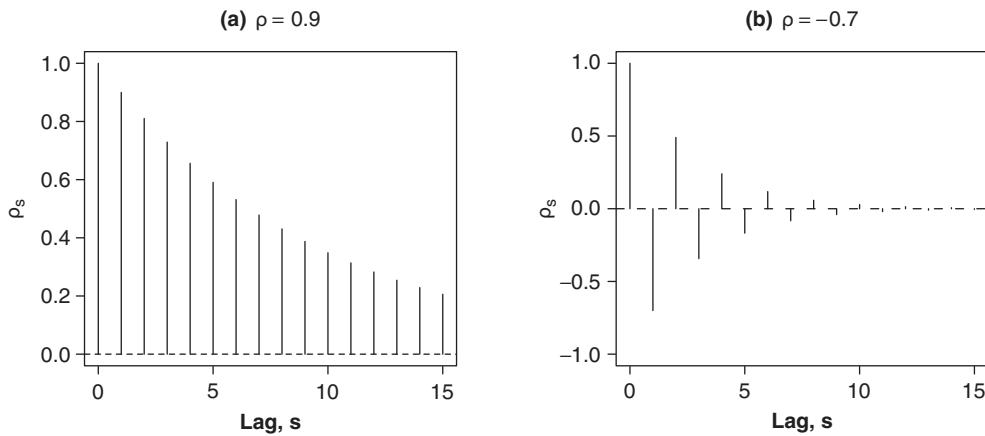


Figure 16.1 Theoretical autocorrelations ρ_s for the first-order autoregressive process $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$, with (a) $\rho = .9$ and (b) $\rho = -.7$.

close to one another tend to be similar. In Figure 16.2(c), $\rho = -.7$; note here how the series tends to bounce from negative to positive values. Negatively autocorrelated series are not common in the social sciences. Finally, Figure 16.2(d) illustrates a nonstationary process, with $\rho = 1.01$.

Figure 16.2(b) also provides some intuitive insight into the problems for estimation posed by autocorrelated errors:

- Because errors that are close in time are likely to be similar, there is much less information in a highly autocorrelated time series than in an independent random sample of the same size. It is, for example, often unproductive to proliferate observations by using more closely spaced time periods (e.g., monthly or quarterly rather than yearly data⁷). To do so will likely increase the autocorrelation of the errors.⁸
- Over a relatively short period of time, a highly autocorrelated series is likely to rise or to fall—that is, show a positive or negative trend. This is true even though the series is stationary and, therefore, will eventually return to its expectation of 0. If, for example, our sample consisted only of the first 50 observations in Figure 16.2(b), then there would be a negative trend in the errors; if our sample consisted of observations 60 to 75, then there would be a positive trend.

⁷Monthly or quarterly data also raise the possibility of “seasonal” effects. One simple approach to seasonal effects is to include dummy regressors for months or quarters. Likewise, dummy regressors for days of the week might be appropriate for some daily time series. More sophisticated approaches to seasonal effects are described in most texts on time-series analysis, such as Harvey (1990, Section 7.6) and Judge, Griffiths, Hill, Lütkepohl, and Lee (1985, Sections 7.2.4 and 7.7.2).

⁸This point is nicely illustrated by considering the sampling variance of the sample mean \bar{Y} . From elementary statistics, we know that the variance of \bar{Y} in an independent random sample of size n is σ^2/n , where σ^2 is the population variance. If instead we sample observations from a first-order autoregressive process with parameter ρ , the variance of \bar{Y} is

$$\frac{\sigma^2}{n} \times \frac{1+\rho}{1-\rho}$$

The sampling variance of \bar{Y} is, therefore, much larger than σ^2/n when the autocorrelation ρ is close to 1. Put another way, the “effective” number of observations is $n(1-\rho)/(1+\rho)$ rather than n . I am grateful to Robert Stine of the University of Pennsylvania for suggesting this illustration.

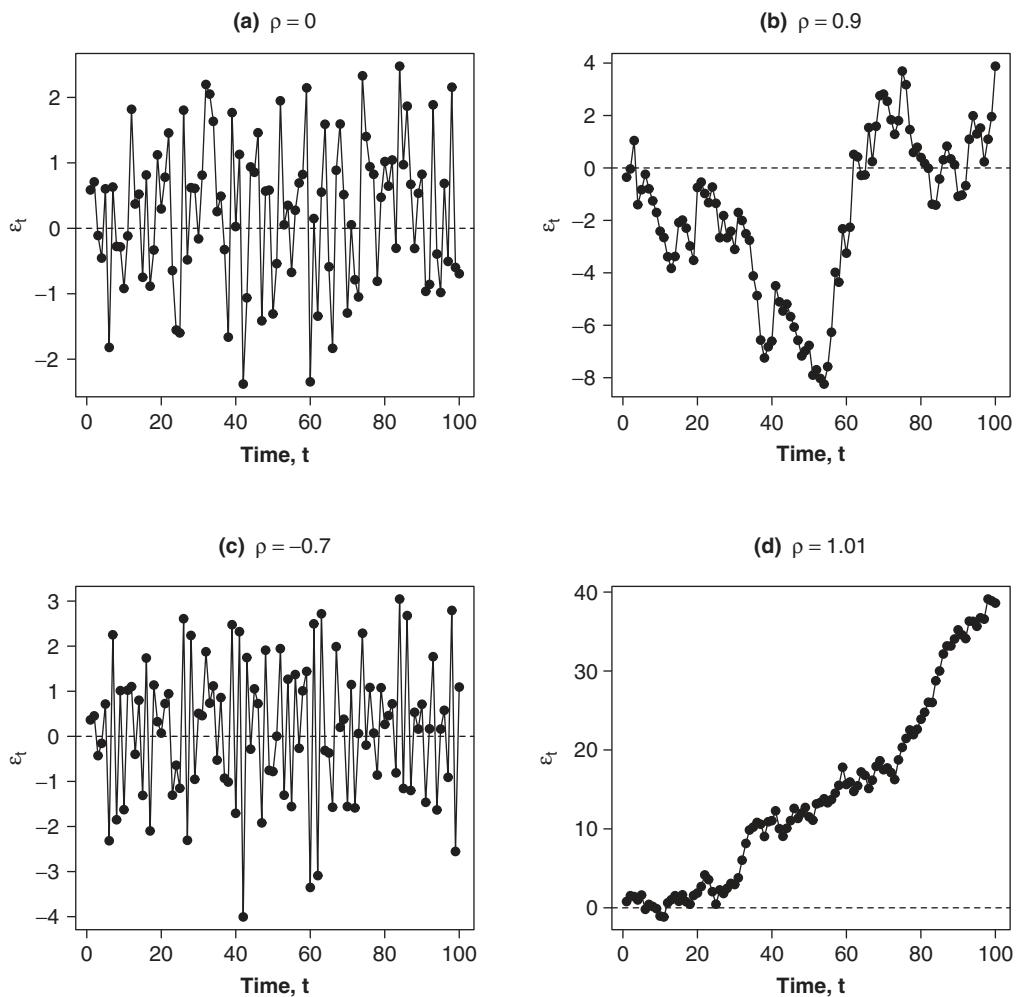


Figure 16.2 Four realizations, each of sample size $n=100$, of the first-order autoregressive process $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$: (a) $\rho = 0$ ("white noise"), (b) $\rho = .9$, (c) $\rho = -.7$, and (d) $\rho = 1.01$ (a nonstationary process). In each case, $\nu_t \sim N(0, 1)$.

Because explanatory variables in a time-series regression also often manifest directional trends, a rise or fall in the errors of a short time series can induce a correlation between an explanatory variable and the errors *in a specific sample*. It is important, however, to understand that there is no implication that the OLS estimates are biased because of correlation between the explanatory variables and the errors: Over many samples, there will sometimes be negative correlations between the errors and the explanatory variables, sometimes positive correlations, and sometimes virtually no correlation. The correlations—sometimes negative, sometimes positive—that occur in specific samples can markedly increase the variance of the OLS estimator, however.⁹

⁹The effect of autocorrelated errors on OLS estimation is explored in Exercise 16.3.

- Finally, because the OLS estimator forces 0 *sample* correlations between the explanatory variables and the residuals (as opposed to the unobserved errors), the sampling variances of the OLS coefficients may be grossly underestimated by $S_E^2(\mathbf{X}'\mathbf{X})^{-1}$. Recall that in a short series, highly autocorrelated errors will often manifest a trend in a *particular* sample.

16.2.2 Higher-Order Autoregressive Processes

The AR(1) process is the simplest member of the family of autoregressive processes. In the p th-order autoregressive process [abbreviated AR(p)], ε_t depends directly on the previous p errors and a random shock ν_t :

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_p \varepsilon_{t-p} + \nu_t \quad (16.5)$$

(where I use ϕ rather than ρ because the autoregression coefficients are no longer correlations).

It is rare in time-series regression to go beyond $p = 2$, that is, the AR(2) process,

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \nu_t \quad (16.6)$$

For this process to be stationary, the roots β of the quadratic equation

$$1 - \phi_1 \beta - \phi_2 \beta^2 = 0$$

must both have modulus exceeding 1.¹⁰

Multiplying Equation 16.6 through by ε_{t-1} and taking expectations produces

$$\begin{aligned} C(\varepsilon_t, \varepsilon_{t-1}) &= \phi_1 E(\varepsilon_{t-1}^2) + \phi_2 E(\varepsilon_{t-1} \varepsilon_{t-2}) \\ &= \phi_1 \sigma_\varepsilon^2 + \phi_2 C(\varepsilon_t, \varepsilon_{t-1}) \end{aligned}$$

because $E(\varepsilon_{t-1}^2) = \sigma_\varepsilon^2$ and $E(\varepsilon_{t-1} \varepsilon_{t-2}) = C(\varepsilon_{t-1}, \varepsilon_{t-2}) = C(\varepsilon_t, \varepsilon_{t-1})$. Solving for the autocovariance,

$$\sigma_1 \equiv C(\varepsilon_t, \varepsilon_{t-1}) = \frac{\phi_1}{1 - \phi_2} \sigma_\varepsilon^2$$

Similarly, for $s > 1$,

$$\begin{aligned} \sigma_s \equiv C(\varepsilon_t, \varepsilon_{t-s}) &= \phi_1 E(\varepsilon_{t-1} \varepsilon_{t-s}) + \phi_2 E(\varepsilon_{t-2} \varepsilon_{t-s}) \\ &= \phi_1 \sigma_{s-1} + \phi_2 \sigma_{s-2} \end{aligned}$$

and thus we can find the autocovariances recursively. For example, for $j = 2$,

$$\begin{aligned} \sigma_2 &= \phi_1 \sigma_1 + \phi_2 \sigma_0 \\ &= \phi_1 \sigma_1 + \phi_2 \sigma_\varepsilon^2 \end{aligned}$$

(where $\sigma_0 = \sigma_\varepsilon^2$), and for $j = 3$,

$$\sigma_3 = \phi_1 \sigma_2 + \phi_2 \sigma_1$$

Autocorrelations for the AR(2) process follow upon division of the autocovariances by σ_ε^2 :

¹⁰In general, the roots can be complex numbers, of the form $\beta = \beta_1 + \beta_2 i$. The modulus of β is $\sqrt{\beta_1^2 + \beta_2^2}$. This stationarity condition generalizes to higher-order AR processes; see, for example, Chatfield (2003, Section 3.2). For an example, see Exercise 16.7.

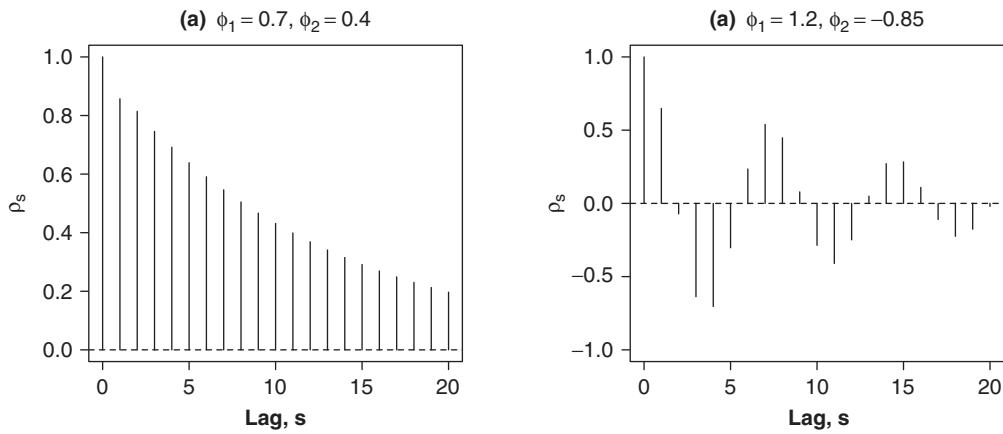


Figure 16.3 Theoretical autocorrelations ρ_s for the AR(2) process $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \nu_t$, with (a) $\phi_1 = 0.7, \phi_2 = 0.4$, and (b) $\phi_1 = 1.2, \phi_2 = -0.85$.

$$\begin{aligned} \text{lag 0: } & \rho_0 = 1 \\ \text{lag 1: } & \rho_1 = \frac{\phi_1}{1 - \phi_2} \\ \text{lag 2: } & \rho_2 = \phi_1 \rho_1 + \phi_2 \\ \text{lag 3: } & \rho_3 = \phi_1 \rho_2 + \phi_2 \rho_1 \\ \text{lag } j > 3: & \rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} \end{aligned}$$

If the process is stationary, then these autocorrelations decay toward 0, although the pattern of decay may be more or less complex depending on the values and signs of the autoregressive parameters ϕ_1 and ϕ_2 . Two examples appear in Figure 16.3.

16.2.3 Moving-Average and Autoregressive-Moving-Average Processes

Although autoregressive processes are the most frequently employed in time-series regression, *moving-average (MA)* and combined *autoregressive-moving-average (ARMA)* processes sometimes can provide simplification. That is, a low-order MA or ARMA process may represent the data as well as a much higher-order AR process.

In the order- q moving-average process [MA(q)], the error at time t depends on the random shock at time t and on the shocks in the previous q time periods:¹¹

¹¹It is common to write the MA(q) model as

$$\varepsilon_t = \nu_t - \theta_1 \nu_{t-1} - \theta_2 \nu_{t-2} - \cdots - \theta_q \nu_{t-q}$$

that is, *subtracting*, rather than adding the terms $\theta_s \nu_{t-s}$, and therefore reversing the signs of the MA parameters θ_s . I believe that the notation that I employ is slightly easier to follow. A similar point applies to the MA terms in ARMA(p, q) models, described below.

$$\varepsilon_t = \nu_t + \theta_1 \nu_{t-1} + \theta_2 \nu_{t-2} + \cdots + \theta_q \nu_{t-q}$$

It is unusual to specify $q > 2$. Applying the same tools that we used to analyze AR processes,¹² we find that the MA(1) process

$$\varepsilon_t = \nu_t + \theta \nu_{t-1}$$

has autocorrelations

$$\begin{aligned}\rho_1 &= \frac{\theta}{1 + \theta^2} \\ \rho_s &= 0, \text{ for } s > 1\end{aligned}$$

For the MA(2) process,

$$\varepsilon_t = \nu_t + \theta_1 \nu_{t-1} + \theta_2 \nu_{t-2}$$

we have,

$$\begin{aligned}\rho_1 &= \frac{\theta_1 + \theta_1 \theta_2}{1 + \theta_1^2 + \theta_2^2} \\ \rho_2 &= \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \\ \rho_s &= 0, \text{ for } s > 2\end{aligned}$$

More generally, in the MA(q) process, $\rho_s = 0$ for $s > q$.

MA processes are stationary without restrictions on the parameters θ_s , but for there to be a one-to-one correspondence between an MA(q) process and a particular autocorrelation function, it is necessary that the process satisfy a condition termed *invertibility*. This condition is closely analogous to the condition for stationarity on the parameters of an AR process, as described above. In particular, for an MA(1) process, we require that $|\theta| < 1$, and for an MA(2) process, we require that the roots of the equation

$$1 + \theta_1 \beta + \theta_2 \beta^2 = 0$$

both have modulus larger than 1.

As its name implies, the autoregressive-moving-average process ARMA(p, q) combines autoregressive and MA components:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_p \varepsilon_{t-p} + \nu_t + \theta_1 \nu_{t-1} + \theta_2 \nu_{t-2} + \cdots + \theta_q \nu_{t-q}$$

These more general ARMA processes are capable of parsimoniously modeling a wider variety of patterns of autocorrelation, but it is rare to go beyond ARMA(1, 1),¹³

$$\varepsilon_t = \phi \varepsilon_{t-1} + \nu_t + \theta \nu_{t-1}$$

This process is stationary if $|\phi| < 1$ and invertible if $|\theta| < 1$.

The autocorrelations for the ARMA(1, 1) process are¹⁴

¹²See Exercise 16.2.

¹³For further details, see, for example, Judge et al. (1985, chaps. 7 and 8) and Chatfield (2003, Section 3.4).

¹⁴See Exercise 16.2.

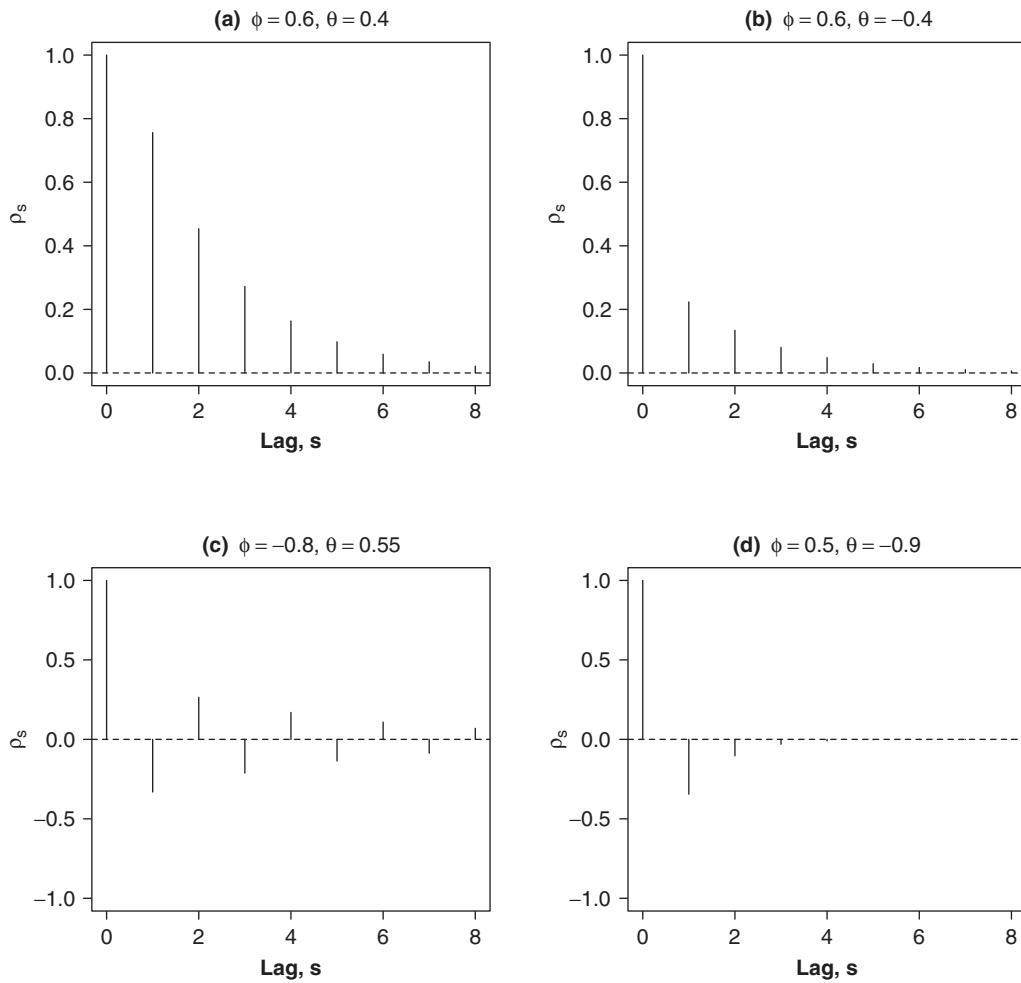


Figure 16.4 Theoretical autocorrelations for the ARMA(1, 1) process, with (a) $\phi = 0.6, \theta = 0.4$; (b) $\phi = 0.6, \theta = -0.4$; (c) $\phi = -0.8, \theta = 0.55$; and (d) $\phi = 0.5, \theta = -0.9$.

$$\rho_1 = \frac{(1 + \phi\theta)(\phi + \theta)}{1 + \theta^2 + 2\phi\theta}$$

$$\rho_s = \phi\rho_{s-1}, \text{ for } s > 1$$

The autocorrelations, consequently, decay exponentially as the lag s grows. Some examples are shown in Figure 16.4.

Higher-order autoregressive processes, moving-average processes, and mixed autoregressive-moving-average processes can be used to model more complex forms of serial dependence in the errors.

16.2.4 Partial Autocorrelations

Let ρ_s^* represent the partial correlation between ε_s and ε_{t-s} “controlling for” $\varepsilon_{t-1}, \dots, \varepsilon_{t-s+1}$.¹⁵ Suppose that ε_t follows an AR(s) process (as given in Equation 16.5 on page 481). Multiplying through successively by $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-s}$, taking expectations, and dividing by the variance σ_ε^2 produces the so-called *Yule-Walker equations*:

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2 \rho_1 + \cdots + \phi_s \rho_{s-1} \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 + \cdots + \phi_s \rho_{s-2} \\ &\vdots \\ \rho_s &= \phi_1 \rho_{s-1} + \phi_2 \rho_{s-2} + \cdots + \phi_s\end{aligned}\tag{16.7}$$

or, in matrix form, $\boldsymbol{\rho} = \mathbf{P}\boldsymbol{\phi}$. Solving for the autoregressive parameters in terms of the autocorrelations yields $\boldsymbol{\phi} = \mathbf{P}^{-1}\boldsymbol{\rho}$. The *partial autocorrelation* ρ_s^* is the last autoregression coefficient, ϕ_s . Setting s in turn to the values $1, 2, \dots$ and forming and solving the resulting sets of Yule-Walker equations produces the partial autocorrelations $\rho_1^*, \rho_2^*, \dots$. It is apparent from this mode of computation that for an AR(p) process, $\rho_s^* = 0$ for $s > p$.

The partial autocorrelations can also be computed for MA and ARMA processes, but rather than falling abruptly to 0, the partial autocorrelations decay exponentially in a more or less complex pattern depending on the order and signs of the coefficients of the MA or ARMA process. Put another way, the partial autocorrelations of an MA process behave much like the autocorrelations of an AR process. Indeed, this link between MA and AR processes is more than superficial: An MA process may be represented as an AR process of infinite order and vice versa.¹⁶

The partial autocorrelations of an AR(p) process fall abruptly to 0 after lag p ; those of MA and ARMA processes decay toward 0 in a pattern determined by the coefficients of the process. The distinctive features of the autocorrelation and partial-autocorrelation functions of AR, MA, and ARMA processes may be used to help select a process to model a particular time series.

16.3 GLS Estimation With Autocorrelated Errors

If the errors follow a first-order autoregressive process, then the covariance matrix of the regression errors, given in general form in Equation 16.3 (page 477), takes the relatively simple form

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon}(\rho, \sigma_\varepsilon^2) = \frac{\sigma_\varepsilon^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}\tag{16.8}$$

¹⁵Partial correlations were introduced in Exercise 5.8.

¹⁶See, for example, Chatfield (2003, Section 3.4).

As the notation implies, the error covariance matrix depends on *only two* parameters: ρ and σ_ν^2 —or, alternatively, on ρ and $\sigma_\varepsilon^2 = \sigma_\nu^2/(1 - \rho^2)$. If we knew the values of these parameters, then we could form $\Sigma_{\varepsilon\varepsilon}$ and proceed directly to GLS estimation.

Recall that GLS estimation can be realized as OLS following a transformation of \mathbf{y} and \mathbf{X} . In the present case (ignoring a constant factor), the transformation matrix is¹⁷

$$\boldsymbol{\Gamma} = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

Then the transformed variables are

$$\mathbf{y}^* = \boldsymbol{\Gamma}\mathbf{y} = \begin{bmatrix} \sqrt{1 - \rho^2}Y_1 \\ Y_2 - \rho Y_1 \\ \vdots \\ Y_n - \rho Y_{n-1} \end{bmatrix} \quad (16.9)$$

and

$$\mathbf{X}^* = \boldsymbol{\Gamma}\mathbf{X} = \begin{bmatrix} \sqrt{1 - \rho^2} & \sqrt{1 - \rho^2}X_{11} & \cdots & \sqrt{1 - \rho^2}X_{1k} \\ 1 - \rho & X_{21} - \rho X_{11} & \cdots & X_{2k} - \rho X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 - \rho & X_{n1} - \rho X_{n-1,1} & \cdots & X_{nk} - \rho X_{n-1,k} \end{bmatrix} \quad (16.10)$$

where the first column of \mathbf{X}^* is the transformed constant regressor.

Except for the first observation, $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_{tj}^* = X_{tj} - \rho X_{t-1,j}$. These transformations have the following intuitive interpretation: Write out the regression equation in scalar form as

$$\begin{aligned} Y_t &= \alpha + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \varepsilon_t \\ &= \alpha + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \rho \varepsilon_{t-1} + \nu_t \end{aligned} \quad (16.11)$$

For the previous observation ($t - 1$), we have, similarly,

$$Y_{t-1} = \alpha + \beta_1 X_{t-1,1} + \cdots + \beta_k X_{t-1,k} + \varepsilon_{t-1} \quad (16.12)$$

Multiplying Equation 16.12 through by ρ and subtracting the result from Equation 16.11 produces

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \alpha(1 - \rho) + \beta_1(X_{t1} - \rho X_{t-1,1}) \\ &\quad + \cdots + \beta_k(X_{tk} - \rho X_{t-1,k}) + \nu_t \\ Y_t^* &= \alpha^* + \beta_1 X_{t1}^* + \cdots + \beta_k X_{tk}^* + \nu_t \quad \text{for } t = 2, \dots, n \end{aligned} \quad (16.13)$$

Because the errors in Equation 16.3 are the ν_t , which are independent of each other and of the X^* s, the transformed equation can legitimately be fit by OLS regression. The only slippage

¹⁷See Exercise 16.4.

here is that the first observation is lost, for there are no data at $t - 1 = 1 - 1 = 0$. Applying OLS to Equation 16.3 is, therefore, not quite the same as GLS.

Qualitatively similar, but more complex, results apply to higher-order AR processes and to MA and ARMA processes for the errors.

16.3.1 Empirical GLS Estimation

All of this presupposes that we know the value of the error autocorrelation ρ . In practice, of course, we need to estimate ρ along with the regression parameters $\alpha, \beta_1, \dots, \beta_k$ and the variance of the random shocks σ_ν^2 (or, alternatively, the variance of the regression errors σ_e^2). One approach to this problem is first to estimate ρ . Then, using the estimate (say $\hat{\rho}$) as if ρ were known, we can calculate GLS estimates and their standard errors—either directly or, equivalently, by OLS following transformation of \mathbf{y} and \mathbf{X} . This approach is called *empirical generalized least squares (EGLS)*.

An especially simple option is to base the estimate of ρ on the lag-1 sample autocorrelation of the residuals from the OLS regression of \mathbf{y} on \mathbf{X} :¹⁸

$$r_1 = \frac{\sum_{t=2}^n E_t E_{t-1}}{\sum_{t=1}^n E_t^2} \quad (16.14)$$

where the E_t are the OLS residuals. The sum in the numerator of Equation 16.14 is over observations $t = 2, \dots, n$ (because E_{t-1} —i.e., E_0 —is unavailable for $t = 1$). Using $\hat{\rho} = r_1$ in Equations 16.9 and 16.10 produces transformed variables from which to calculate the EGLS estimates by OLS regression. The variance of the residuals from this OLS regression estimates σ_ν^2 .

This procedure can be extended to more complex processes for the errors.¹⁹

16.3.2 Maximum-Likelihood Estimation

It is preferable to estimate all of the parameters— ρ, σ_ν^2 , and β —directly and simultaneously, by maximum likelihood, thereby acknowledging the additional uncertainty produced by having to estimate the parameters of the error process—uncertainty that is ignored by the EGLS estimator. We just have to think of the log-likelihood as a function of all of the parameters (adapting Equation 16.1 on page 475):

$$\log_e L(\beta, \rho, \sigma_\nu^2) = -\frac{n}{2} \log_e 2\pi - \frac{1}{2} \log_e (\det \Sigma_{ee}) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \Sigma_{ee}^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

where Σ_{ee} for AR(1) errors is determined by the parameters ρ and σ_ν^2 according to Equation 16.8 (page 485). This approach is, moreover, quite general, because *any* AR, MA, or ARMA process provides an expression for Σ_{ee} as a function of the parameters of the error process. An illustrative application using an AR(2) process for the errors is described in the next section.

¹⁸Although there are other methods to obtain a preliminary estimate of ρ , none holds a particular advantage. For details, see, for example, Judge et al. (1985, Section 8.2.1).

¹⁹See, for example, Judge et al. (1985, Section 8.2).

To apply GLS estimation to a regression model with AR(1) errors, we can first estimate the autocorrelation of the errors from the sample autocorrelation of the OLS residuals: $\hat{\rho} = r_1 = (\sum_{t=2}^n E_t E_{t-1}) / (\sum_{t=1}^n E_t^2)$. We can then use $\hat{\rho}$ to form an estimate of the correlation matrix of the errors or to transform \mathbf{y} and \mathbf{X} . Except for the first observation, these transformations take a very simple form: $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$ and $X_{ij}^* = X_{ij} - \hat{\rho} X_{t-1,j}$. This procedure is called empirical GLS estimation. Empirical GLS can be extended to more complex time-series models for the errors. A better approach, however, is to estimate the parameters of the error process along with the regression coefficients by the method of maximum likelihood.

16.4 Correcting OLS Inference for Autocorrelated Errors

The OLS estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ of the regression coefficients β in the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ is unbiased and consistent even when the errors ε are autocorrelated, with covariance matrix $\Sigma_{\varepsilon\varepsilon}$. When the errors are correlated, however, the covariance matrix of \mathbf{b} is given by

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma_{\varepsilon\varepsilon}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (16.15)$$

rather than by $\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$, as it would be for independent errors with constant variance.²⁰ An alternative approach to estimation and inference in time-series regression is therefore to retain the OLS estimator but to base statistical inference on an estimate of $V(\mathbf{b})$ in Equation 16.15.

This approach was suggested by Newey and West (1987) (among others) and is similar in spirit and form to White's heteroscedasticity-consistent (HC) coefficient covariance-matrix estimator.²¹ The Newey-West estimator is termed a *heteroscedasticity and autocorrelation consistent (HAC)* estimator because it potentially accounts both for nonconstant error variance and for autocorrelated errors. The details of the Newey-West and closely related coefficient covariance-matrix estimators are substantially more complicated than White's HC estimator—in particular, in time-series regression, the off-diagonal elements of $\Sigma_{\varepsilon\varepsilon}$ are generally nonzero—and so I will simply sketch the basic ideas here.²²

Let us rewrite Equation 16.15 as

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\Phi(\mathbf{X}'\mathbf{X})^{-1} \quad (16.16)$$

where $\Phi \equiv \mathbf{X}'\Sigma_{\varepsilon\varepsilon}\mathbf{X}$, and let

$$V_t(\beta) \equiv \mathbf{x}_t(Y_t - \mathbf{x}_t'\beta) = \mathbf{x}_t\varepsilon_t$$

where \mathbf{x}_t' is the t th row of the model matrix \mathbf{X} . At the OLS solution $\beta = \mathbf{b}$,²³

²⁰See Exercise 16.3; in this exercise, it is assumed that the errors are generated by a first-order autoregressive process, but the formula for $V(\mathbf{b})$ in Equation 16.15 is more general.

²¹White's estimator is discussed in Section 12.2.3.

²²See Exercise 16.8 for an application, and Andrews (1991) for a general treatment of the topic.

²³See Exercise 16.8.

$$\sum_{t=1}^n \widehat{\mathbf{V}}_t \equiv \sum V_t(\mathbf{b}) = \mathbf{0} \quad (16.17)$$

The Newey-West and similar estimators of $V(\mathbf{b})$ estimate Φ at the center of the “sandwich” in Equation 16.16 by

$$\widehat{\Phi} = \sum_{t=1}^n \sum_{t'=1}^n w(|t - t'|) \widehat{\mathbf{v}}_t \widehat{\mathbf{v}}'_{t'}$$

where the *weights* $w(|t - t'|)$ decline with $|t - t'|$; in Newey and West’s scheme, for example, the weights decay linearly to some maximum lag L ,

$$w(|t - t'|) = 1 - \frac{|t - t'|}{L + 1}$$

and are 0 thereafter. A common simple choice, traceable to Newey and West (1987), is to take L as $n^{1/4}$ rounded to the next integer.²⁴ Finally, it is also common to “correct” the estimated coefficient covariance matrix for degrees of freedom, producing

$$\widehat{V}(\mathbf{b}) = \frac{n}{n - k - 1} (\mathbf{X}'\mathbf{X})^{-1} \widehat{\Phi} (\mathbf{X}'\mathbf{X})^{-1}$$

where, as usual, there are $k + 1$ coefficients in the coefficient vector β .

An alternative to generalized-least-squares estimation, suggested by Newey and West (1987), is to retain the ordinary least-squares estimator—which is unbiased and consistent in the presence of autocorrelated errors—but to replace the usual OLS coefficient covariance-matrix estimator with a covariance-matrix estimator that is consistent when the errors are autocorrelated (and, incidentally, when the error variance is not constant).

16.5 Diagnosing Serially Correlated Errors

We need to ask whether the data support the hypothesis that the errors are serially correlated, because, in the absence of serially correlated errors, we can legitimately employ OLS estimation. As usual, our key to the behavior of the unobservable errors is the least-squares residuals.²⁵

Figure 16.5, based on data from Fox and Hartnagel (1979), shows a yearly time-series plot of the female indictable-offense conviction rate (FCR) per 100,000 Canadian women aged 15 years and older, for the period 1931 to 1968.²⁶ The conviction rate rose from the mid-1930s until 1940, then declined until the mid-1950s, and subsequently rose again.

²⁴See Newey and West (1994) for a more sophisticated approach to selecting L and Andrews (1991) for alternative weighting schemes.

²⁵We used the least-squares residuals to learn about the errors in the discussion of regression diagnostics in Chapters 11 and 12.

²⁶Because the basis for reporting convictions changed in 1949, the data for the period 1950 to 1968 have been adjusted. The adjustment used here is very slightly different from the one employed by Fox and Hartnagel (1979). Indictable offenses are relatively serious crimes. I am grateful to Timothy Hartnagel of the University of Alberta for helping me to assemble the data for this example.

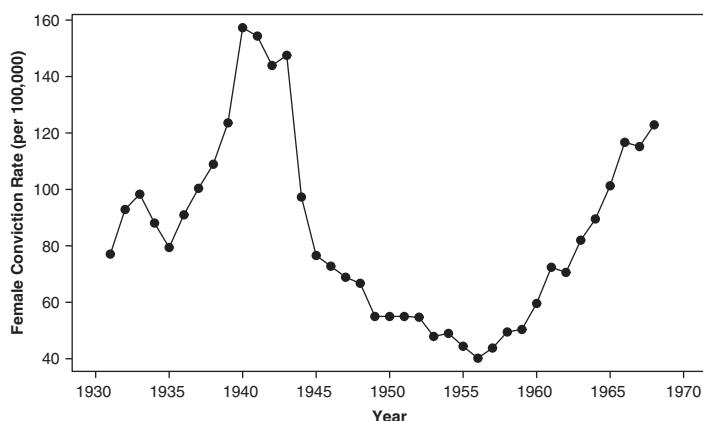


Figure 16.5 Canadian women's indictable-offense conviction rate per 100,000 population, for the period 1931 to 1968.

Fox and Hartnagel were interested in relating variations in women's crime rates to changes in their position within Canadian society. To this end, they regressed women's conviction rate on the following explanatory variables:

- The *total fertility rate (TFR)*—the number of births to an imaginary cohort of 1000 women who live through their childbearing years at current age-specific fertility rates.
- *Women's labor-force participation rate (LFPR)* per 1000 population.
- *Women's postsecondary-degree rate (PSDR)* per 10,000 population.
- *Men's indictable-offense conviction rate (MCR)* per 100,000 population. This explanatory variable was meant to represent factors affecting women's conviction rate that are not specifically included in the model.

The results from the OLS regression of women's conviction rate on these explanatory variables are as follows (with standard errors in parentheses below the estimated coefficients):

$$\begin{aligned} \widehat{\text{FCR}} = & 127.6 - 0.04657 \times \text{TFR} + 0.2534 \times \text{LFPR} \\ & (59.9) \quad (0.00803) \quad (0.1152) \\ & - 0.2120 \times \text{PSDR} + 0.05911 \times \text{MCR} \\ & (0.2115) \quad (0.04515) \end{aligned} \quad (16.18)$$

$R^2 = .6948$

The coefficients are not estimated very precisely—after all, the data set is quite small—and those for PSDR and MCR are not statistically significantly different from 0.²⁷

A useful next step is to plot the residuals against time, as is done for Fox and Hartnagel's regression in Figure 16.6. It is clear from this graph that the residuals are positively autocorrelated, but another problem is apparent as well: The model is not doing a very good job during

²⁷Exercise 16.7 addresses the adequacy of the model.

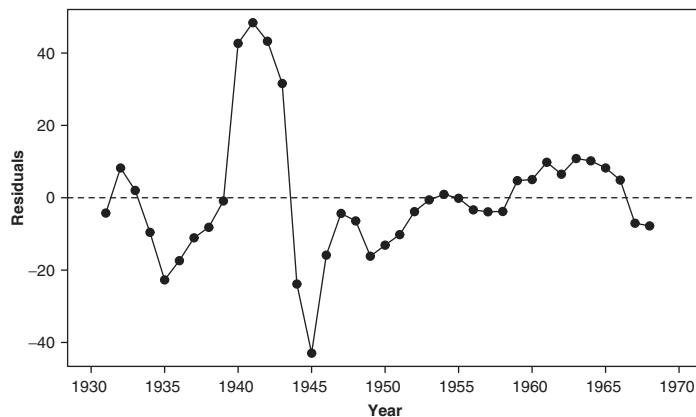


Figure 16.6 Time-series plot of residuals from the OLS regression of female conviction rate on several explanatory variables.

the Second World War, accounting neither for the jump in female crime at the beginning of the war nor for its subsequent decline.

After examining the OLS residuals, we can calculate sample autocorrelations for the residuals:

$$r_s = \frac{\sum_{t=s+1}^n E_t E_{t-s}}{\sum_{t=1}^n E_t^2}$$

for lags $s = 1, 2, \dots, m$, where the maximum lag m should be no larger than about $n/4$. We can also compute sample partial autocorrelations, r_s^* at various lags, using the sample analogs of the Yule-Walker equations (given in Equations 16.7 on page 485). If the residuals were independently distributed (which, recall, they are not—even when the errors are independent²⁸), then the standard error of each r_s and r_s^* would be approximately $1/\sqrt{n}$, and the autocorrelations and partial autocorrelations would be asymptotically normally distributed.²⁹

The pattern of the sample autocorrelations and partial autocorrelations of the residuals can help us to identify a time-series process to use in modeling serial dependency among the errors. If the errors follow a first-order autoregressive process, for example, then the residual correlations should (roughly) decay exponentially toward 0, and only the first partial autocorrelation should be large.

Graphs of the residual autocorrelations and partial autocorrelations (called *correlograms*) for the OLS regression using Fox and Hartnagel's data are shown in Figure 16.7. As a rough guide to the “statistical significance” of the residual autocorrelations and partial autocorrelations, I have placed reference lines in the correlograms at $\pm 2/\sqrt{n}$.³⁰ The pattern is clearly indicative

²⁸See Section 11.8.2.

²⁹See, for example, Chatfield (2003, Section 4.1).

³⁰A further reason for caution in interpreting the correlogram is that there are many sample autocorrelations that are themselves correlated, creating a problem of simultaneous inference.

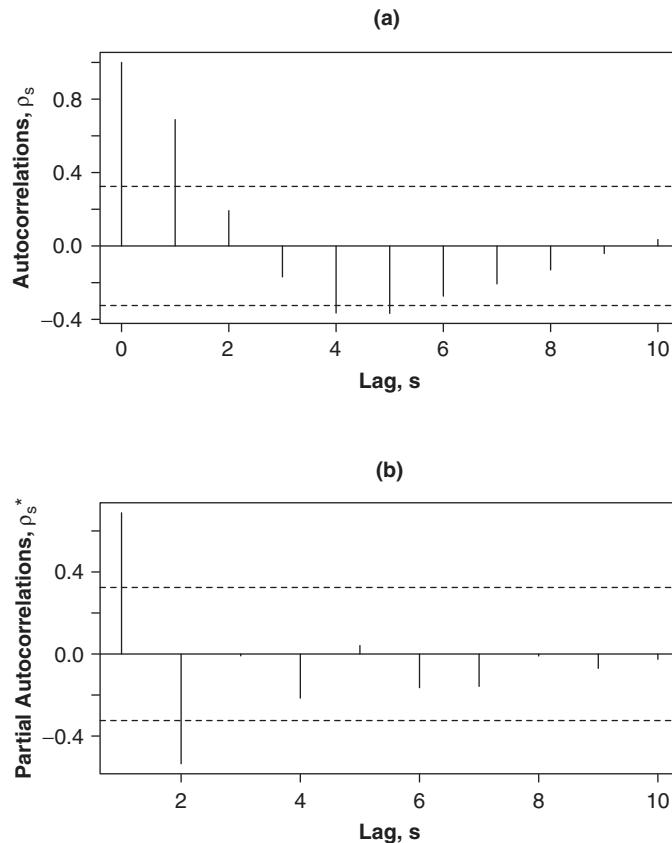


Figure 16.7 Autocorrelations and partial autocorrelations for the residuals from the OLS regression of female conviction rate on several explanatory variables.

of an AR(2) process, with a positive autoregression coefficient at lag 1 and negative coefficient at lag 2.

Because simple tests for autocorrelation based on the r_s and r_s^* are, at best, rough, more precise methods have been proposed. A common approach to testing for autocorrelated errors is due to Durbin and Watson (1950, 1951). Durbin and Watson's test statistic is based on the assumption that the errors follow a first-order autoregressive process and tests the null hypothesis that the autoregression parameter ρ is 0:³¹

$$D \equiv \frac{\sum_{t=2}^n (E_t - E_{t-1})^2}{\sum_{t=1}^n E_t^2}$$

When n is large, $D \approx 2(1 - r_1)$.³² If the null hypothesis is correct, therefore, we expect to observe values of D close to 2; if the null hypothesis is wrong, and the errors are *positively*

³¹The Durbin-Watson statistic can be generalized to lags $s > 1$: $D_s \equiv \sum_{t=s+1}^n (E_t - E_{t-s})^2 / \sum_{t=1}^n E_t^2$. For this and other tests of autocorrelated errors, see, for example, Judge et al. (1985, Section 8.4).

³²See Exercise 16.6.

autocorrelated (i.e., $\rho > 0$), then we expect to observe values of D that are substantially *smaller than 2*. The range of D values is 0 to 4. Although it is applied more broadly, the Durbin-Watson test is most powerful when the true error-generating process is AR(1).

The sampling distribution of the Durbin-Watson statistic is complex and, unfortunately, depends on the configuration of the explanatory variables. Durbin and Watson initially calculated critical values of D for two extreme scenarios: the X configuration producing the smallest critical value and the X configuration producing the largest critical value. This approach leads to an extensive set of tables (because different critical values are required for different combinations of sample size n and number of explanatory variables k) and to ambiguous results if—as is common—the observed value of D falls between the two extreme critical values.³³

Modern statistical software for time-series regression, however, typically calculates a p -value for D based on the X -values in the sample at hand. For Fox and Hartnagel's regression, for example, $D = 0.617$, for which the two-sided p -value is much less than .0001, strongly supporting the conclusion that the errors are autocorrelated.

Reestimating Fox and Hartnagel's model by ML, assuming AR(2) errors, produces the following estimates and coefficient standard errors:

$$\begin{aligned} \widehat{\text{FCR}} &= 83.34 - 0.03999 \times \text{TFR} + 0.2876 \times \text{LFPR} \\ &\quad (59.47) \quad (0.00928) \quad (0.1120) \\ &\quad - 0.2098 \times \text{PSDR} + 0.07569 \times \text{MCR} \\ &\quad (0.2066) \quad (0.03501) \end{aligned} \tag{16.19}$$

$$\begin{aligned} \widehat{\phi}_1 &= 1.068 \\ \widehat{\phi}_2 &= -0.5507 \end{aligned}$$

With the exception of the regression constant,³⁴ neither the coefficients nor their standard errors change much from the OLS results in Equation 16.18.

The following table compares the maximized log-likelihood for three nested models, in which the error process is AR(p), with p set successively to 2 through 0 [where AR(0) corresponds to the initial OLS regression]:

p	Log-likelihood
2	-144.71
1	-149.21
0	-163.50

Because these models are nested, comparing adjacent lines of the table produces likelihood-ratio tests of the hypotheses $H_0^{(2)}: \phi_2 = 0$ and $H_0^{(1)}: \phi_1 = 0 \mid \phi_2 = 0$. The chi-square test statistics, each on one degree of freedom, are $G_2^2 = 2(149.21 - 144.71) = 9.00$ and $G_1^2 = 2(163.50 - 149.21) = 28.58$, for which the p -values are, respectively, .0027 and $\ll .0001$, strongly supporting the inclusion of ϕ_1 and ϕ_2 in the model.

³³Tables of critical values of D were published by Durbin and Watson (1951) and are widely reproduced (e.g., in Harvey, 1990, pp. 362–363).

³⁴The constant is very imprecisely estimated: After all, setting all of the explanatory variables to 0 extrapolates the regression far beyond the observed range of the data.

To diagnose serial correlation in the errors, we can examine correlograms of the OLS residuals, calculating the sample autocorrelations $r_s = (\sum_{t=s+1}^n E_t E_{t-s}) / (\sum_{t=1}^n E_t^2)$ and partial autocorrelations r_s^* for a number of lags $s = 1, 2, \dots, m$. If, for example, the serial dependence of the residuals is well described by a first-order autoregressive process, then the autocorrelations should decay exponentially toward 0, and partial autocorrelations after the first should be negligible. The Durbin-Watson statistic $D \equiv [\sum_{t=2}^n (E_t - E_{t-1})^2] / (\sum_{t=1}^n E_t^2) \approx 2(1 - r_1)$ can be used to test for serial correlation of the errors.

16.6 Concluding Remarks

It is tempting to conclude that the theoretical advantage of GLS regression should mandate its use, especially if the residuals are significantly autocorrelated, but several factors suggest caution:³⁵

- The extent to which GLS estimation is more efficient than OLS estimation and the extent to which the usual formula for OLS standard errors produces misleading results depend on a number of complex factors, including the process generating the errors, the degree of autocorrelation of the errors, and the distribution of the explanatory-variable values.³⁶
- When the errors are highly autocorrelated and follow a first-order autoregressive process and when explanatory-variable values manifest a linear trend, the advantage of GLS estimation can be strongly dependent on retaining the first observation (i.e., using the transformation $\sqrt{1 - \rho^2}$ of the first observation).³⁷ Precisely in these circumstances, the first observation can become influential in the transformed regression, however. The comfort that we derive from GLS may therefore be tenuous. It is, consequently, useful to examine influential-data diagnostics for the *transformed* equation.³⁸ This point extends to other error processes.
- Many of the properties of GLS, EGLS, and ML estimators depend on asymptotic results, but time-series data sets are usually quite small.³⁹ Moreover, long time series raise the possibility that regression relationships—for example, the slope associated with a particular explanatory variable—may themselves change over time. There is no *generally* satisfactory method for detecting such changes, although it may help to plot residuals against time.⁴⁰

³⁵For an elaboration of some of these points, see, for example, the discussion in Judge et al. (1985, chap. 8).

³⁶See Exercise 16.3.

³⁷See Exercise 16.3.

³⁸See Chapter 11.

³⁹Bootstrapping, discussed in Chapter 21, can prove helpful in this context.

⁴⁰This is not to say, however, that there are *no* methods for detecting changes in regression coefficients over time; see, for example, Brown, Durbin, and Evans (1975).

- The performance of a method like GLS estimation may depend crucially on getting the error-generating process right. If the process is not AR(1), for example, basing estimation on this process may cause more harm than good. Real social processes do not, in any event, unfold according to AR, MA, or ARMA processes that, at best, produce reasonable descriptive summaries. We have to be careful not to construe time-series regression models too literally.

Three further cautionary notes are in order:

- Because time-series data often manifest strong trends, the explanatory variables in a time-series regression can be strongly collinear. This problem can be exacerbated when time itself (i.e., the regressor $X_t = t$) is included in the regression to capture a linear trend. The general rationale for employing time as a regressor is to control statistically for omitted factors that change smoothly with time and that are correlated with the explanatory variables included in the model.
- The models discussed in this chapter assume *contemporaneous* effects. That is, all of the variables in the model are measured at time t . It is sometimes reasonable to suppose, however, that the effect of an explanatory variable (say, X_1) will occur after an interval of time has elapsed, for example, after one time period. The time-series regression model would then take the form

$$Y_t = \alpha + \beta_1 X_{t-1,1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \varepsilon_t$$

Aside from the loss of the first observation (because $X_{t-1,1}$ is typically unavailable when $t = 1$ and hence $t - 1 = 0$), specifications of this form pose no new problems. If, however, we do not know in advance that the effect of X_1 is lagged some specific number of time periods and rather want to consider effects at several lags, then autocorrelation of X_1 can induce serious collinearity. Special techniques of estimation (called “distributed lags”⁴¹) exist to deal with this situation, but these methods require that we know in advance something about the form of the lagged effects of X_1 on Y .

- Finally, the methods of this chapter are generally inappropriate (in the presence of auto-correlated errors) when the *response variable* appears as a lagged effect on the right-hand side of the model,⁴² as in

$$Y_t = \alpha + \beta Y_{t-1} + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \varepsilon_t$$

There are several practical and theoretical difficulties that limit the effectiveness and range of application of EGLS and ML estimation of time-series regression models.

⁴¹See, for example, the discussion of distributed lags in Judge et al. (1985, chaps. 9 and 10).

⁴²See, for example, Harvey (1990, chap. 8) for a discussion of regression with a lagged response on the right-hand side of the model.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 16.1. *Generalized least squares: For the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim N_n(\mathbf{0}, \Sigma_{\varepsilon\varepsilon})$, where the error covariance matrix $\Sigma_{\varepsilon\varepsilon}$ is known:

- (a) Show that the log-likelihood for the model is

$$\log_e L(\beta) = -\frac{n}{2} \log_e 2\pi - \frac{1}{2} \log_e (\det \Sigma_{\varepsilon\varepsilon}) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \Sigma_{\varepsilon\varepsilon}^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

(Hint: Use the formula for the multivariate normal distribution.)

- (b) Show that the ML estimator of β is

$$\mathbf{b}_{\text{GLS}} = (\mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{y}$$

and that its sampling variance is

$$V(\mathbf{b}_{\text{GLS}}) = (\mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1}$$

- (c) Prove the Gauss-Markov theorem for the GLS estimator. That is, show that under the assumptions $E(\varepsilon) = \mathbf{0}$ and $V(\varepsilon) = \Sigma_{\varepsilon\varepsilon}$, \mathbf{b}_{GLS} is the minimum-variance linear unbiased estimator of β . (Hints: See Section 9.3.2 and Exercise 12.4 on page 335.)

Exercise 16.2. *Autocorrelations for MA and ARMA processes:

- (a) Show that the MA(1) process $\varepsilon_t = \nu_t + \theta\nu_{t-1}$ has autocorrelations

$$\rho_1 = \frac{\theta}{1 + \theta^2}$$

$$\rho_s = 0, \text{ for } s > 1$$

- (b) Show that the MA(2) process $\varepsilon_t = \nu_t + \theta_1\nu_{t-1} + \theta_2\nu_{t-2}$ has autocorrelations

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

$$\rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

$$\rho_s = 0, \text{ for } s > 2$$

- (c) Show that the ARMA(1, 1) process $\varepsilon_t = \phi\varepsilon_{t-1} + \nu_t + \theta\nu_{t-1}$ has autocorrelations

$$\rho_1 = \frac{(1 + \phi\theta)(\phi + \theta)}{1 + \theta^2 + 2\phi\theta}$$

$$\rho_s = \phi\rho_{s-1}, \text{ for } s > 1$$

(Hint: To find the autocovariance at lag 1, multiply through the equation for the process by ε_{t-1} and take expectations. Divide by the variance of ε_t to get ρ_1 . Repeat this procedure for other lags.)

Exercise 16.3. Autocorrelated errors and OLS estimation: Assume that $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ and that the errors follow an AR(1) process, $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$, where $\nu_t \sim N(0, \sigma_\nu^2)$.

- (a) Show that the OLS estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is unbiased despite the autocorrelation of the errors. [Hint: Recall the proof that $E(\mathbf{b}) = \beta$ from Section 9.3.1.]
- (b) Show that the variance of the OLS estimator is

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma_{\varepsilon\varepsilon}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\Sigma_{\varepsilon\varepsilon}$ is given by Equation 16.8 (page 485).

- (c) Suppose that we fit the simple-regression model $Y_t = \alpha + \beta x_t + \varepsilon_t$; that the x -values are 1, 2, ..., 10; that the errors follow an AR(1) process; and that the variance of the random shocks ν_t is 1. [Recall that the variance of the errors is $\sigma_\varepsilon^2 = \sigma_\nu^2/(1 - \rho^2)$.] Calculate (1) the *true* sampling variance of the OLS estimator; (2) the sampling variance of the OLS estimator according to the usual formula, $\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$, appropriate when the errors are independent; and (3) the sampling variance of the GLS estimator, using the formula

$$V(\mathbf{b}_{\text{GLS}}) = (\mathbf{X}'\Sigma_{\varepsilon\varepsilon}^{-1}\mathbf{X})^{-1}$$

Pay particular attention to the sampling variance of the estimators of the slope β . What do you conclude from these calculations? Perform the calculations assuming successively that $\rho = 0$, $\rho = .5$, and $\rho = .9$.

- (d) *Now suppose that the first observation is dropped and that we perform the regression in (c) using $t = 2, \dots, 10$. Working with the transformed scores $x_t^* = x_t - \rho x_{t-1}$ and again employing the three different values of ρ , find the sampling variance of the resulting estimator of β . How does the efficiency of this estimator compare to that of the true GLS estimator? With the OLS estimator? What would happen if there were many more observations?
- (e) Repeat (c) [and (d)], but with $x_t = (t - 5)^2$ for $t = 1, 2, \dots, 9$.

Exercise 16.4. *Show that the appropriate GLS transformation matrix for AR(1) errors is

$$\Gamma = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

[Hints: First show that

$$\frac{1}{1 - \rho^2} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

is the inverse of

$$\mathbf{P} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

Then show that $\mathbf{P}^{-1} = [1/(1 - \rho^2)]\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$. The constant $1/(1 - \rho^2)$ can be ignored in forming the square-root matrix $\boldsymbol{\Gamma}$. Why?]

Exercise 16.5. *Maximum-likelihood estimation with AR(1) errors: Assume that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and that the errors follow an AR(1) process but that the autoregression parameter ρ and the variance of the random shocks σ_v^2 are unknown. Show that the log-likelihood under this model can be written in the following form:

$$\begin{aligned} \log_e L(\boldsymbol{\beta}, \rho, \sigma_v^2) &= -\frac{n}{2} \log_e 2\pi - \frac{n}{2} \log_e \sigma_v^2 + \frac{1}{2} \log_e (1 - \rho^2) \\ &\quad - \frac{1}{2\sigma_v^2} (\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta})' (\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) \end{aligned}$$

where $\mathbf{y}^* = \boldsymbol{\Gamma}\mathbf{y}$ and $\mathbf{X}^* = \boldsymbol{\Gamma}\mathbf{X}$. [Hints: Start with Equation 16.1 (on page 475) for the log-likelihood of the general model with error covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$. Then use $\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = (1/\sigma_v^2)\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$, noting that $\det \boldsymbol{\Sigma}_{\varepsilon\varepsilon} = (1/\sigma_v^2)^n (\det \boldsymbol{\Gamma})^2$ and that $\det \boldsymbol{\Gamma} = \sqrt{1 - \rho^2}$.]

Exercise 16.6. Show that when n is large, the Durbin-Watson statistic D is approximately equal to $2(1 - r_1)$, where r_1 is the lag-1 autocorrelation of the OLS residuals. (Hint: When n is large, $\sum_{t=1}^n E_t^2 \approx \sum_{t=2}^n E_t^2 \approx \sum_{t=2}^n E_{t-1}^2$)

Exercise 16.7. With reference to Fox and Hartnagel's regression of Canadian women's conviction rates on several explanatory variables:

- (a) Use regression diagnostics, as described in Part III of the text, to explore the adequacy of the preliminary OLS regression fit to these data (Equation 16.18 on page 490). If you detect any problems, try to correct them, and then repeat the subsequent analysis of the data.
- (b) Show that the estimated parameters of the AR(2) process fit to the errors, $\hat{\phi}_1 = 1.068$ and $\hat{\phi}_2 = -0.5507$ (see Equation 16.19 on page 493), correspond to a stationary time-series process. (Hint: Use the quadratic formula to solve the equation $1 - 1.068\beta + 0.5507\beta^2 = 0$, and verify that both roots have modulus greater than 1.)
- (c) Reestimate Fox and Hartnagel's regression with AR(2) errors by EGLS, comparing your results with those produced by the method of maximum likelihood (in Equation 16.19). Obtain estimates of the error autoregression parameters ϕ_1 and ϕ_2 by solving the Yule-Walker equations (see Equations 16.7 on page 485)

$$\begin{aligned} r_1 &= \hat{\phi} + \hat{\phi}_2 r_1 \\ r_2 &= \hat{\phi}_1 r_1 + \hat{\phi}_2 \end{aligned}$$

where r_1 and r_2 are the lag-1 and lag-2 sample autocorrelations of the OLS residuals.

Exercise 16.8. Newey-West coefficient standard errors:

- (a) Explain why (repeating Equation 16.17 from page 489)

$$\sum_{t=1}^n \hat{\mathbf{v}}_t \equiv \sum V_t(\mathbf{b}) = \mathbf{0}$$

where \mathbf{b} is the OLS coefficient vector and $V_t(\mathbf{b}) = \mathbf{x}_t(Y_t - \mathbf{x}_t' \mathbf{b})$. (*Hint:* Show that these are simply the OLS estimating equations.)

- (b) Apply Newey and West's procedure to compute HAC standard errors for the OLS coefficients in Fox and Hartnagel's time-series regression. Do the results differ from those employing the usual OLS standard errors? Do the results differ from those obtained by GLS?

Summary

- In time-series data, a single individual is tracked over many time periods or points of time. It is not generally reasonable to suppose that the errors in a time-series regression are independent.
- To capture serial dependence among the errors in the regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, we drop the assumption that the errors are independent of one another; instead, we assume that $\varepsilon \sim N_n(\mathbf{0}, \Sigma_{\varepsilon\varepsilon})$, where nonzero off-diagonal entries in the error covariance matrix $\Sigma_{\varepsilon\varepsilon}$ correspond to correlated errors.
- If the error covariance matrix $\Sigma_{\varepsilon\varepsilon}$ is known, then the maximum-likelihood estimator of β is the generalized least-squares estimator

$$\mathbf{b}_{\text{GLS}} = (\mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{y}$$

The sampling variance-covariance matrix of \mathbf{b}_{GLS} is

$$V(\mathbf{b}_{\text{GLS}}) = (\mathbf{X}' \Sigma_{\varepsilon\varepsilon}^{-1} \mathbf{X})^{-1}$$

The generalized least-squares estimator can also be expressed as the OLS estimator $(\mathbf{X}^* \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{y}^*$ for the transformed variables $\mathbf{X}^* \equiv \Gamma \mathbf{X}$ and $\mathbf{y}^* \equiv \Gamma \mathbf{y}$, where the transformation matrix Γ is a square root of $\Sigma_{\varepsilon\varepsilon}^{-1}$.

- When, more realistically, the error covariance matrix $\Sigma_{\varepsilon\varepsilon}$ is unknown, we need to estimate its contents along with the regression coefficients β . Without restricting its form, however, $\Sigma_{\varepsilon\varepsilon}$ contains too many distinct elements to estimate directly. Assuming that the errors are generated by a stationary time-series process reduces the number of independent parameters in $\Sigma_{\varepsilon\varepsilon}$ to n , including the error variance σ_ε^2 and the autocorrelations at various lags, $\rho_1, \dots, \rho_{n-1}$.
- To reduce the number of parameters in $\Sigma_{\varepsilon\varepsilon}$ further, we can adopt a specific time-series model for the errors. The most commonly employed such model is the first-order autoregressive process $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$, where $|\rho| < 1$ and the random shocks ν_t are independently distributed as $N(0, \sigma_\nu^2)$. Under this specification, two errors, ε_t and ε_{t+s} , separated by s time periods have autocovariance $\rho^s \sigma_\varepsilon^2$ and autocorrelation ρ^s . The variance of the regression errors is $\sigma_\varepsilon^2 = \sigma_\nu^2 / (1 - \rho^2)$.

- Higher-order autoregressive processes, moving-average processes, and mixed autoregressive-moving-average processes can be used to model more complex forms of serial dependence in the errors.
- The partial autocorrelations of an AR(p) process fall abruptly to 0 after lag p ; those of MA and ARMA processes decay toward 0 in a pattern determined by the coefficients of the process. The distinctive features of the autocorrelation and partial-autocorrelation functions of AR, MA, and ARMA processes may be used to help select a process to model a particular time series.
- To apply GLS estimation to a regression model with AR(1) errors, we can first estimate the autocorrelation of the errors from the sample autocorrelation of the OLS residuals:

$$\hat{\rho} = r_1 = \frac{\sum_{t=2}^n E_t E_{t-1}}{\sum_{t=1}^n E_t^2}$$

We can then use $\hat{\rho}$ to form an estimate of the correlation matrix of the errors or to transform \mathbf{y} and \mathbf{X} . Except for the first observation, these transformations take a very simple form: $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$ and $X_{ij}^* = X_{ij} - \hat{\rho} X_{i-1,j}$. This procedure is called empirical GLS estimation. Empirical GLS can be extended to more complex time-series models for the errors. A better approach, however, is to estimate the parameters of the error process along with the regression coefficients by the method of maximum likelihood.

- An alternative to generalized least-squares estimation, suggested by Newey and West (1987), is to retain the ordinary least-squares estimator—which is unbiased and consistent in the presence of autocorrelated errors—but to replace the usual OLS coefficient covariance-matrix estimator with a covariance-matrix estimator that is consistent when the errors are autocorrelated (and, incidentally, when the error variance is not constant).
- To diagnose serial correlation in the errors, we can examine correlograms of the OLS residuals, calculating the autocorrelations

$$r_s = \frac{\sum_{t=s+1}^n E_t E_{t-s}}{\sum_{t=1}^n E_t^2}$$

and partial autocorrelations r_s^* for a number of lags $s = 1, 2, \dots, m$. If, for example, the serial dependence of the residuals is well described by a first-order autoregressive process, then the autocorrelations should decay exponentially toward 0, and partial autocorrelations after the first should be negligible. The Durbin-Watson statistic

$$D \equiv \frac{\sum_{t=2}^n (E_t - E_{t-1})^2}{\sum_{t=1}^n E_t^2} \approx 2(1 - r_1)$$

can be used to test for serial correlation of the errors.

- Several practical and theoretical difficulties limit the effectiveness and range of application of EGLS and ML estimation of time-series regression models.

Recommended Reading

- Time-series analysis—including, but not restricted to, time-series regression—is a deep and rich topic, well beyond the scope of the discussion in this chapter. A good, relatively brief, general introduction to the subject may be found in Chatfield (2003).

- Most econometric texts include some treatment of time-series regression. The emphasis is typically on a formal understanding of statistical models and methods of estimation rather than on the use of these techniques in data analysis. Wonnacott and Wonnacott (1979), for example, present an insightful, relatively elementary treatment of time-series regression and generalized least squares. Judge et al.'s (1985) presentation of the subject is more encyclopedic, with many references to the literature. An extensive treatment also appears in Harvey (1990). Greene (2003, chaps. 10, 11, and 20) includes a good overview of GLS estimation, regression with autocorrelated errors, and time-series models more generally.
- The current chapter merely scratches the surface of time-series regression models. A compact, accessible, and much more general treatment may be found in Pickup (2014), which includes many examples, most drawn from political science.

17

Nonlinear Regression

As I have explained, there is a distinction between explanatory variables and regressors.¹ The general linear model is linear in the regressors but not necessarily in the explanatory variables that generate these regressors.

In dummy-variable regression, for example, categorical explanatory variables do not appear directly in the model; a single polytomous explanatory variable gives rise to several dummy regressors. Likewise, in polynomial regression, a single quantitative explanatory variable generates several regressors (e.g., linear, quadratic, and cubic terms). Interaction regressors are functions of two or more explanatory variables. We can also transform quantitative explanatory variables or the response variable prior to formulating a linear model.

In its least restrictive form, then, we can write the general linear model as

$$f(Y_i) = \beta_0 f_0(\mathbf{x}'_i) + \beta_1 f_1(\mathbf{x}'_i) + \cdots + \beta_p f_p(\mathbf{x}'_i) + \varepsilon_i$$

$$Y'_i = \beta_0 X'_{i0} + \beta_1 X'_{i1} + \cdots + \beta_p X'_{ip} + \varepsilon_i$$

where

- Y_i is the response variable for the i th observation;
- $\mathbf{x}'_i = (X_{i1}, \dots, X_{ik})$ is a vector of k (not necessarily quantitative) explanatory variables;²
- $\beta_0, \beta_1, \dots, \beta_p$ are parameters to estimate;
- the ε_i are independent and normally distributed errors, with zero expectations and constant variance; and
- the functions $f(\cdot), f_0(\cdot), \dots, f_p(\cdot)$ do not involve unknown parameters.

For the least-squares estimates of the β s to be unique, the regressors X'_0, \dots, X'_p cannot be perfectly collinear. If, as is usually the case, the model includes the constant regressor, then $X'_{i0} \equiv f_0(\mathbf{x}'_i) = 1$. An X'_j can be a function of more than one explanatory variable, encompassing models such as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon \quad (17.1)$$

In an application, the functions $f(\cdot), f_0(\cdot), \dots, f_p(\cdot)$ may be suggested by prior theoretical considerations or by examination of the data, as when we transform an explanatory variable to linearize its relationship to Y .

¹See Chapters 7 and 8.

²If you are not familiar with vector notation, simply think of \mathbf{x}'_i as a *list* of the explanatory variables for the i th observation.

Nonlinear regression models that are linear in the parameters, for example, the quadratic regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$, can be fit by linear least squares.

17.1 Polynomial Regression

Polynomial regression is an important form of nonlinear regression that is accommodated by the general linear model. We have already seen how a quadratic function can be employed to model a U-shaped (or \cap -shaped) relationship.³ Similarly, a cubic function can be used to model a relationship in which the direction of curvature changes. More generally, a polynomial of order p can have $p - 1$ “bends.” Polynomials of degree greater than 3 are rarely employed in data analysis, however.

Polynomial regressors are especially useful when a quantitative explanatory variable is discrete. As we know,⁴ we can capture any—potentially nonlinear—partial relationship between Y and X_j by constructing $m - 1$ dummy regressors to represent the m distinct values of X_j . The powers $X_j, X_j^2, \dots, X_j^{m-1}$ can be thought of as an alternative coding of the discrete variable X_j , providing the same fit to the data as the dummy regressors (in the same sense as dummy coding and deviation coding are equivalent).⁵

We can then “step down” through the powers $X_j^{m-1}, X_j^{m-2}, \dots, X_j^2, X_j^1$, testing the contribution of each term to the model, omitting the term if it proves unnecessary, and refitting the model. We stop dropping terms when one proves to be important. Thus, if the model includes a cubic term, it will also generally include the lower-order quadratic and linear terms.⁶ Even if the relationship between Y and X_j is nonlinear, it is usually possible to represent this relationship with a polynomial of degree less than $m - 1$.⁷

Polynomials in two or more explanatory variables can be used to model interactions between quantitative explanatory variables. Consider, for example, the full quadratic model for two explanatory variables given in Equation 17.1 above. As illustrated in Figure 17.1(a), this model represents a curved surface relating $E(Y)$ to X_1 and X_2 . Of course, certain specific characteristics of the regression surface—such as direction of curvature and monotonicity—depend on the parameters of the model and on the range of values for X_1 and X_2 .⁸ The partial relationships of Y to each of X_1 and X_2 are apparent in the lines drawn on the regression surfaces in Figure 17.1:

³See, for example, the nonlinear partial relationship between log wages and age in the Canadian Survey of Labour and Income Dynamics data, discussed in Section 12.3.

⁴See Section 12.4.

⁵See Section 8.1. *Put another way, the powers $X_j, X_j^2, \dots, X_j^{m-1}$ provide an alternative—and hence equivalent—basis for the subspace spanned by the dummy regressors.

⁶This is an application of the principle of marginality, introduced in Section 7.3.2. If, in a polynomial in X of order p , all lower-order terms are included, the fit of the model is invariant with respect to linear transformations of X —produced, for example, by subtracting the mean from each value, $X_i - \bar{X}$ (i.e., centering X). If lower-order terms are omitted, however, this invariance does not hold.

⁷Also see Exercise 17.2 for a discussion of *orthogonal* polynomial contrasts.

⁸See Exercise 17.3.

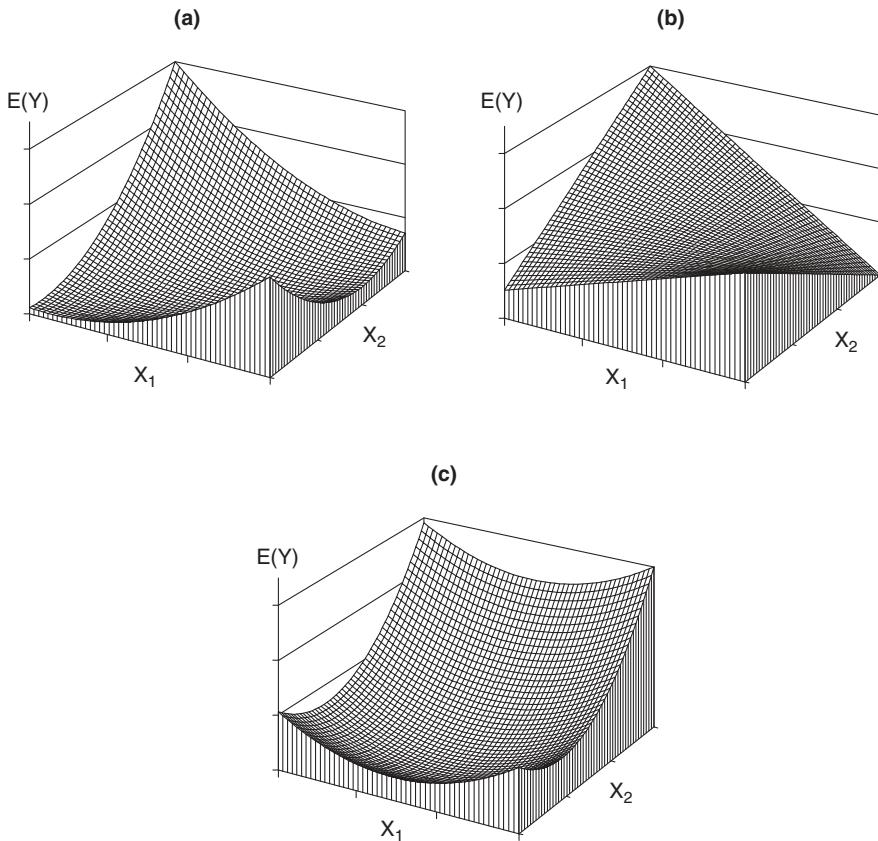


Figure 17.1 The model $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1^2 + \beta_4X_2^2 + \beta_5X_1X_2$, illustrated in (a), represents a curved surface in which the *quadratic* partial relationship of Y to X_1 changes with the value of X_2 (and the quadratic partial relationship of Y to X_2 changes with the value of X_1). The model $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2$ in (b) represents a curved surface in which the slope of the *linear* partial relationship of Y to X_1 changes with the value of X_2 (and the slope of the linear partial relationship of Y to X_2 changes with the value of X_1). The model $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1^2 + \beta_4X_2^2$ in (c) represents a curved surface in which the *quadratic* partial relationship of Y to X_1 is the same at different levels of X_2 (and the quadratic partial relationship of Y to X_2 is the same at different levels of X_1).

- For the full quadratic of Equation 17.1, shown in panel (a), the partial-regression lines are curved, and each quadratic partial relationship changes with the value of the other explanatory variable. Hence, X_1 and X_2 interact in their effect on Y .
- Panel (b) represents the equation

$$E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2$$

Here, the partial relationships are *linear*, but the partial relationship between Y and each X changes with the value of the other X , inducing a bend in the regression surface, representing the interaction between the two explanatory variables.

Table 17.1 Cowles and Davis's Logistic Regression for Volunteering

Coefficient	Estimate	Standard Error
Constant	-2.358	0.501
Sex (Male)	-0.2471	0.1116
Neuroticism	0.1108	0.0376
Extraversion	0.1668	0.0377
Neuroticism×Extraversion	-0.008552	0.002934

- The regression equation

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2$$

is illustrated in panel (c). Here, the partial relationships are nonlinear, but, for example, the quadratic partial relationship of Y to X_1 does not depend on the value at which X_2 is “held constant,” reflecting the absence of interaction between the two explanatory variables.

To illustrate quadratic regression and also to make the point that the method is applicable to *any* statistical model with a linear predictor, I will develop an application in logistic regression, taken from research by Cowles and Davis (1987) on volunteering for psychological experiments. The response variable in the study is dichotomous: whether or not each of 1421 subjects volunteered to participate in an experiment.⁹

The authors modeled the data in a logistic regression of volunteering on the factor sex, the personality dimensions neuroticism and extraversion, and the product of neuroticism and extraversion. The personality variables each can take on integer values between 0 and 24. Preliminary work suggests that sex does not interact with the personality dimensions and that squared terms in neuroticism and extraversion are not needed,¹⁰ producing the estimated coefficients and standard errors in Table 17.1.

An effect display for the neuroticism-by-extraversion interaction is shown in Figure 17.2, setting the dummy regressor for sex to 0.45—the proportion of men in the data set (the dummy variable for sex is coded 1 for men and 0 for women). The graph shows the relationship between volunteering and extraversion at representative values of neuroticism. At low levels of neuroticism, volunteering rises with extraversion; this relationship becomes weaker as neuroticism grows, and at the highest level of neuroticism, the relationship between volunteering and extraversion is negative. Because the fitted values are plotted on the logit scale, the partial regression lines relating volunteering to extraversion are straight [and are produced by slicing the regression surface in the direction of extraversion; cf. Figure 17.1(b)]; note that the lines meet at a point—a characteristic of a model of this structure.¹¹

⁹These data were used in Fox (1987). I am grateful to Michael Cowles and Caroline Davis of York University for making the data available.

¹⁰See Exercise 17.4.

¹¹See Exercise 17.5.

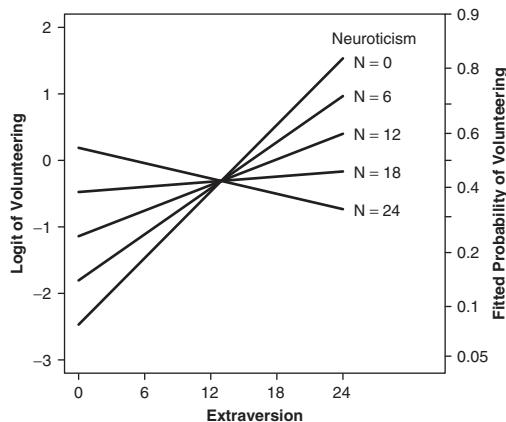


Figure 17.2 Fitted logit and probability of volunteering as a function of neuroticism and extraversion, from Cowles and Davis's logistic regression of volunteering on sex, neuroticism, extraversion, and the product of neuroticism and extraversion. To construct this effect display, the dummy regressor for sex was set to 0.45, which is the proportion of men in the data set.

17.1.1 A Closer Look at Quadratic Surfaces*

The essential structure of nonlinear models is often clarified by differentiating the model with respect to each explanatory variable.¹² Differentiating the equation for the full quadratic model (repeating Equation 17.1),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

produces

$$\begin{aligned}\frac{\partial E(Y)}{\partial X_1} &= \beta_1 + 2\beta_3 X_1 + \beta_5 X_2 \\ \frac{\partial E(Y)}{\partial X_2} &= \beta_2 + 2\beta_4 X_2 + \beta_5 X_1\end{aligned}$$

The slope of the partial relationship between Y and X_1 , therefore, depends not only on the level of X_1 but also on the specific value at which X_2 is held constant—indicating that X_1 and X_2 interact in affecting Y . Moreover, the shape of the partial relationship between Y and X_1 is quadratic, fixing the value of X_2 . Because of the symmetry of the model, similar statements apply to the partial relationship between Y and X_2 , holding X_1 constant.

In contrast, although the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

¹²See Exercise 17.1.

also represents a curved surface [an illustration appears in Figure 17.1(b)], the *slices* of this surface in the direction of each explanatory variable, holding the other constant, are, as I noted, linear:

$$\begin{aligned}\frac{\partial E(Y)}{\partial X_1} &= \beta_1 + \beta_3 X_2 \\ \frac{\partial E(Y)}{\partial X_2} &= \beta_2 + \beta_3 X_1\end{aligned}$$

Thus, for example, the slope of the relationship between Y and X_1 is different *at different levels* of X_2 , but *at each fixed level* of X_2 , the relationship between Y and X_1 is linear.

Finally, the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \varepsilon$$

[illustrated in Figure 17.1(c)] represents a curved surface in which the quadratic partial relationship between Y and each of the explanatory variables is invariant across the levels of the other explanatory variable:

$$\begin{aligned}\frac{\partial E(Y)}{\partial X_1} &= \beta_1 + 2\beta_3 X_1 \\ \frac{\partial E(Y)}{\partial X_2} &= \beta_2 + 2\beta_4 X_2\end{aligned}$$

17.2 Piece-wise Polynomials and Regression Splines

A potential problem with polynomial-regression fits is that they can be highly nonlocal: Data in one region, including outlying values, can seriously affect the fit in another region. Moreover, polynomials are inappropriate for regressions that approach an asymptote. As illustrated in Figure 17.3, these problems can be especially acute with high-degree polynomials: Here, I have fit a fifth-degree polynomial to the United Nations (UN) infant-mortality data.¹³ The other regression curve drawn on this plot is for a natural regression spline (described later in this section) with 6 degrees of freedom—the same number of degrees of freedom as for the polynomial. It is clear that the regression spline does a much better job of following the pattern of the data.

As an alternative to a global polynomial regression, we can partition the data into bins, fitting a different polynomial regression in each bin—a generalization of the idea of binning and averaging (described in Chapter 2). Indeed, as shown in Figure 17.4(a), binning and averaging can be thought of as fitting a degree-0 polynomial in each bin. The data in this graph, and in the other panels of Figure 17.4, were artificially generated.¹⁴

In Figure 17.4(b), a least-squares line—that is, a degree-1 polynomial—is fit in each bin. Finally, in Figure 17.4(c), a line is fit in each bin, but the lines are constrained to be continuous at the bin boundaries. Continuity can be imposed by fitting the model¹⁵

¹³These data were introduced in Chapter 3. In Section 4.3, we discovered that the relationship between infant mortality and gross domestic product (GDP) can be rendered nearly linear by log-transforming both variables. The regression-spline fit in Figure 17.3 is reasonably similar to the lowess fit in Figure 3.14 (on page 45).

¹⁴The data were generated according to the regression equation $Y = \frac{1}{5}X + \cos(X + 1) + \varepsilon$, where $\cos(X + 1)$ is evaluated with X measured in radians and where the errors were sampled from $N(0, 0.5^2)$. The 50 X -values were drawn from a uniform distribution on the interval $[0, 10]$. This example was inspired by a similar one in Hastie, Tibshirani, and Friedman (2009, pp. 118–119).

¹⁵See Exercise 17.6 for this and other results in this section.

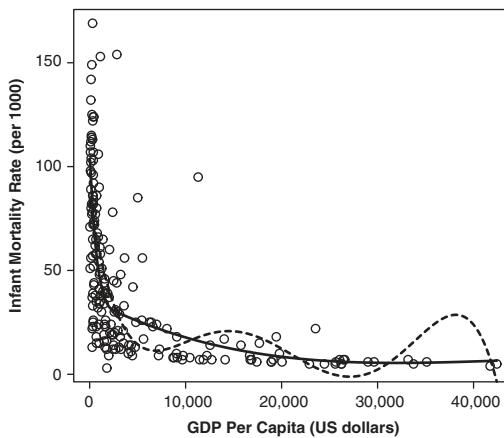


Figure 17.3 Infant-mortality rate by GDP per capita. The broken line is for an order-5 polynomial fit by least squares; the solid line is for a natural regression spline with 6 degrees of freedom, also fit by least squares.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

where $X_{i1} \equiv X_i$;

$$X_{i2} \equiv \begin{cases} 0 & \text{for } X_i \leq k_1 \\ X_i - k_1 & \text{for } X_i > k_1 \end{cases}$$

and

$$X_{i3} \equiv \begin{cases} 0 & \text{for } X_i \leq k_2 \\ X_i - k_2 & \text{for } X_i > k_2 \end{cases}$$

The points at which the lines join, at $X = k_1$ and $X = k_2$, are called *knots*.

This approach can be generalized to higher-order polynomials, as illustrated for cubic polynomials in Figure 17.5 (using the same artificial data as in Figure 17.4). In panel (a), a separate cubic is fit to each bin, and consequently the fit is discontinuous at the bin boundaries. In panel (b), the cubics are constrained to join at the knots, by fitting the regression

$$\begin{aligned} Y_i = & \alpha + \beta_{11} X_{i1} + \beta_{12} X_{i1}^2 + \beta_{13} X_{i1}^3 + \beta_{21} X_{i2} + \beta_{22} X_{i2}^2 + \beta_{23} X_{i2}^3 \\ & + \beta_{31} X_{i3} + \beta_{32} X_{i3}^2 + \beta_{33} X_{i3}^3 + \varepsilon_i \end{aligned}$$

where X_1 , X_2 , and X_3 are defined as above. In Figure 17.5(c), the cubic regressions are further constrained to have equal slopes at the knots by omitting the second and third linear terms from the model:

$$Y_i = \alpha + \beta_{11} X_{i1} + \beta_{12} X_{i1}^2 + \beta_{13} X_{i1}^3 + \beta_{22} X_{i2}^2 + \beta_{23} X_{i2}^3 + \beta_{32} X_{i3}^2 + \beta_{33} X_{i3}^3 + \varepsilon_i$$

Finally, in Figure 17.5(d), not only the slopes but also the curvature of the regressions are matched at the knots, fitting the equation¹⁶

¹⁶*The slope is the first derivative of the regression function, while the curvature depends on the second derivative. Thus, not only the regression curve but also its first and second derivatives are continuous at the knots in Figure 17.5(d).

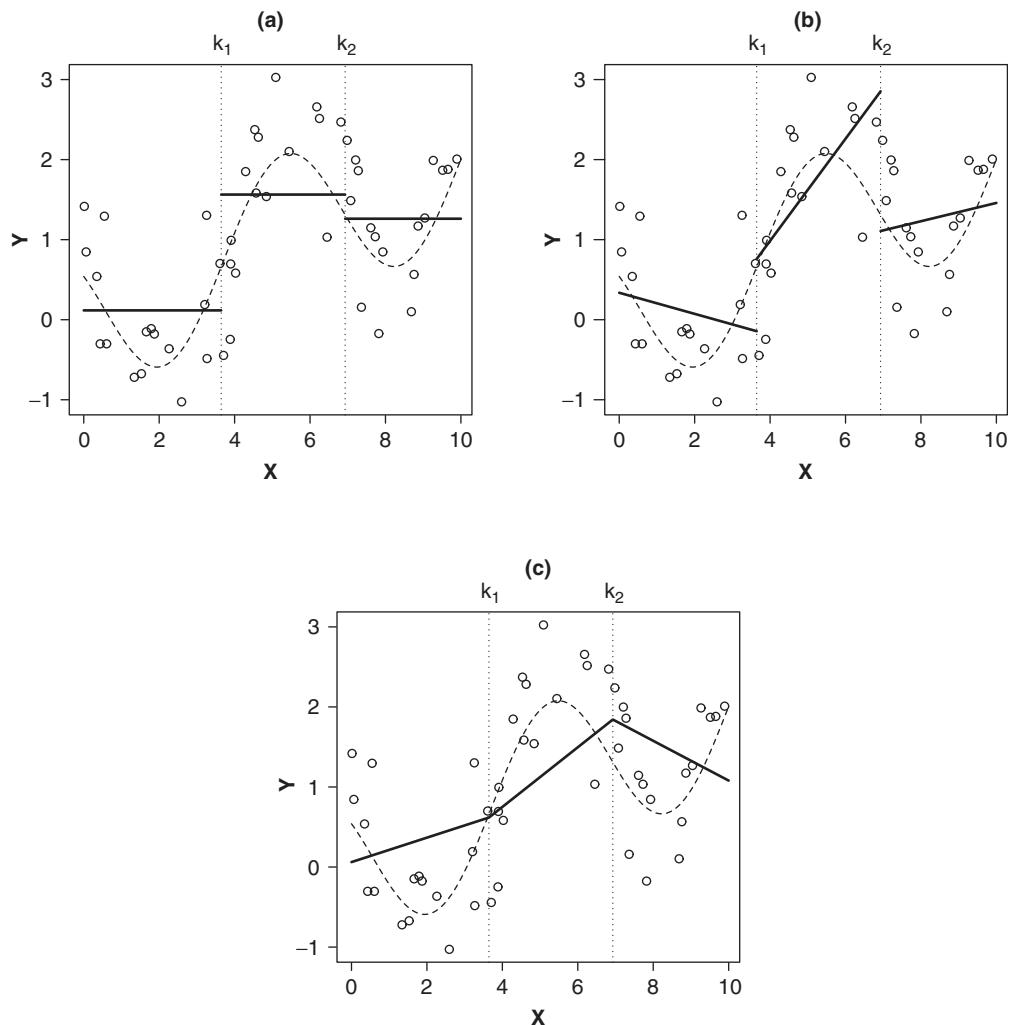


Figure 17.4 (a) Piece-wise constant, (b) piece-wise discontinuous-linear, and (c) piece-wise continuous-linear fits to artificially generated data. The data are binned at the values $X = k_1$ and $X = k_2$. The broken line in each graph is the “true” regression function used to generate the data.

$$Y_i = \alpha + \beta_{11}X_{i1} + \beta_{12}X_{i1}^2 + \beta_{13}X_{i1}^3 + \beta_{23}X_{i2}^3 + \beta_{33}X_{i3}^3 + \varepsilon_i \quad (17.2)$$

Note that as the regression curve is progressively constrained in this manner, it grows smoother.

Equation 17.2 is called a *cubic regression spline*.¹⁷ More generally, if we fit a cubic regression spline with k knots, dividing the data into $k + 1$ bins, the resulting regression model uses

¹⁷A *spline* is a flexible tool used in drafting to draw smooth, continuous curves. *Spline functions* in mathematics are piece-wise continuous and smooth polynomials traditionally used for interpolation.

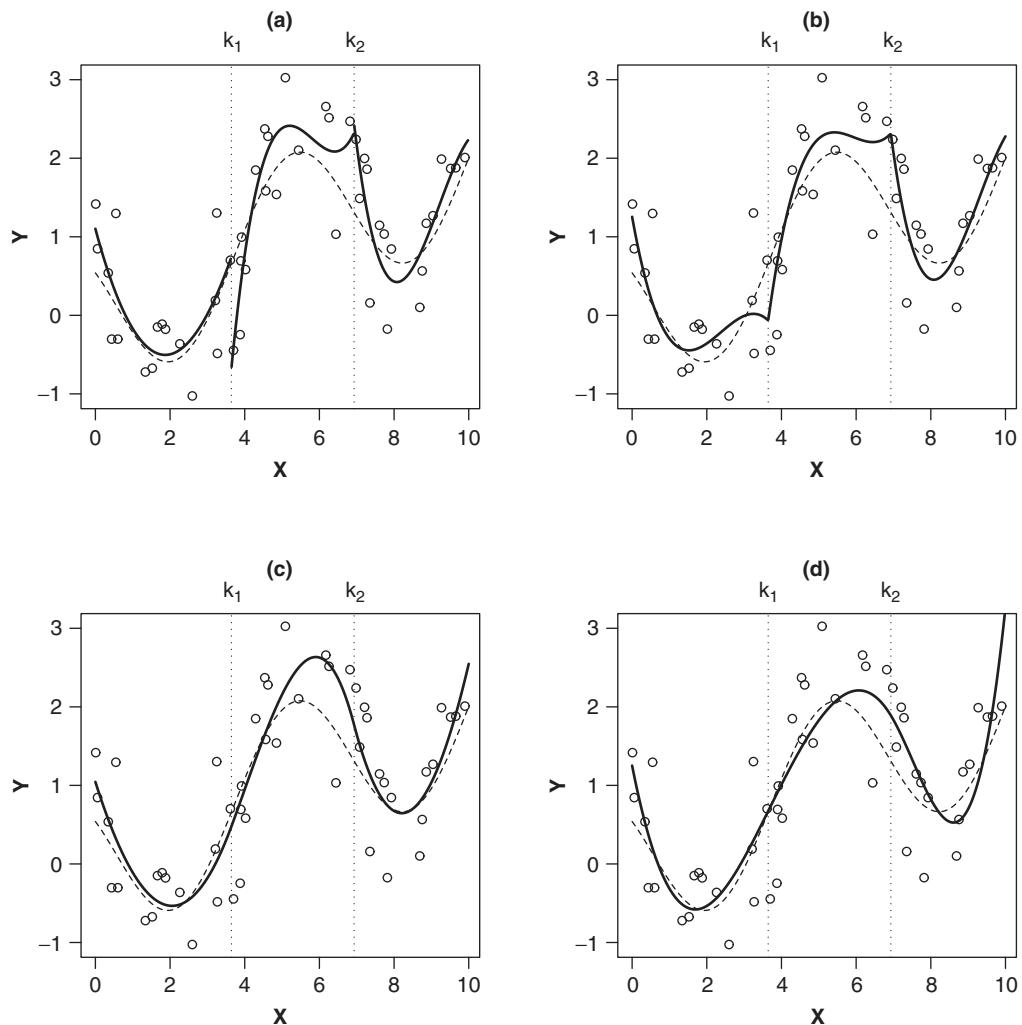


Figure 17.5 Piece-wise cubic fits to artificially generated data: (a) discontinuous, (b) continuous, (c) continuous with continuous slopes, and (d) continuous with continuous slopes and curvature. The broken line in each graph is the “true” regression function used to generate the data.

$k + 4$ degrees of freedom (i.e., has $k + 4$ parameters): 1 for the constant; 3 for the linear, quadratic, and cubic terms in the first bin; and k for the additional cubic terms, one for each remaining bin. In practice, the regressors in the generalization of Equation 17.2 can become highly correlated, and it is therefore advantageous to fit the model with an alternative, but equivalent, set of less correlated regressors.¹⁸

^{18*}In the language of Chapters 9 and 10, we select a different basis for the subspace spanned by the cubic-regression-spline regressors. The *B-spline* basis is frequently used in practice; the details are beyond this discussion, but see, for example, Hastie et al., (2009, pp. 160–161). Because of their behavior near the boundaries of the data, B-splines produce a slightly different fit than the simple regression splines described in this section.

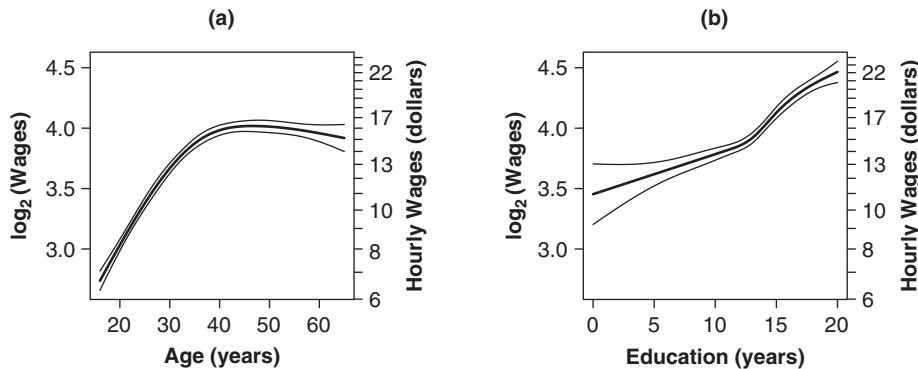


Figure 17.6 Effect displays for age and education in the regression of log wages on these variables and sex. Age and education are represented by natural splines, each with $k = 4$ knots. The lighter lines give pointwise 95% confidence envelopes around the fits.

I have yet to address how to select the number and placement of the knots in fitting a regression spline to data. This is a key point, because with the values of the knots fixed, a regression spline is a linear model and, as such, provides a fully parametric fit to the data. The issue is less critical, however, than it may appear: It almost always suffices to use 4 or 5 knots and to distribute the knots at evenly spaced quantiles of the explanatory variable: Thus, for example, 4 knots divide the data into 5 bins and can be placed at the 20th, 40th, 60th, and 80th percentiles of X . Moreover, this strategy extends in a straightforward manner to *multiple* regression analysis, where individual quantitative explanatory variables in the model can be represented by regression splines.

Cubic regression splines can behave erratically near the boundaries of the data. A simple fix is to constrain the regression function to be linear at and beyond the left and right boundaries. These additional constraints produce what is termed a *natural cubic regression spline*; the boundary constraints account for 2 degrees of freedom, and therefore with k knots, a natural regression spline uses only $k + 2$ degrees of freedom, counting the regression constant.¹⁹

An illustration, using data from the Canadian Survey of Labour and Income Dynamics, is shown in Figure 17.6. The effect displays in this graph are for the regression of log wages on years of education, years of age, and a dummy regressor for sex (the effect of which is not shown). Age and education are modeled using natural splines, each with $k = 4$ knots (and, hence, 5 degrees of freedom each, *not* counting the regression constant).²⁰

Spline regression models are a bridge between the linear and generalized linear models of Parts II and IV of the text and the nonparametric regression models to be discussed in Chapter 18: Regression splines fit comfortably into the *parametric* linear-predictor toolbox of linear and generalized linear models; yet, like nonparametric-regression models, they are flexible and therefore can conform to local characteristics of the data.

¹⁹The natural cubic regression spline can be written as a regression equation in much the same manner as the cubic regression spline of Equation 17.2. The formulation of the regressors is complicated, however; see Hastie et al. (2009, Section 5.2.1).

²⁰Compare these effect displays with those in Figure 12.8 (page 313), where I used a quadratic specification for the age effect and squared education to linearize the regression.

Regression splines are piece-wise cubic polynomials that are continuous at join-points, called knots, and that are constrained to have equal slopes and curvature on either side of a knot. Although fully parametric, regression splines generally do a good job of responding to local aspects of the data and can be incorporated as building blocks into linear and generalized linear models.

17.3 Transformable Nonlinearity

As explained in the previous section, linear statistical models effectively encompass models that are linear in the parameters, even if they are nonlinear in the variables. The forms of nonlinear relationships that can be expressed as linear models are, therefore, very diverse. In certain circumstances, however, theory dictates that we fit models that are nonlinear in their parameters. This is a relatively rare necessity in the social sciences, primarily because our theories are seldom mathematically concrete, although nonlinear models arise in some areas of demography, economics, and psychology and occasionally in sociology, political science, and other social sciences.

Some models that are nonlinear in the parameters can be transformed into linear models and, consequently, can be fit to data by linear least squares. A model of this type is the so-called gravity model of migration, employed in human geography (see Abler, Adams, & Gould, 1971, pp. 221–233). Let Y_{ij} represent the number of migrants moving from city i to city j , let D_{ij} represent the geographical distance between these cities, and let P_i and P_j represent their respective populations.

The gravity model of migration is built in rough analogy to the Newtonian formula for gravitational attraction between two objects, where population plays the role of mass and migration the role of gravity. The analogy is loose, in part because gravitational attraction is *symmetric*, while there are *two* migration streams of generally *different* sizes between a pair of cities: one from city i to city j and the other from j to i .

The gravity model is given by the equation

$$\begin{aligned} Y_{ij} &= \alpha \frac{P_i^\beta P_j^\gamma}{D_{ij}^\delta} \varepsilon_{ij} \\ &= \tilde{Y}_{ij} \varepsilon_{ij} \end{aligned} \tag{17.3}$$

where α , β , γ , and δ are unknown parameters to be estimated from the data, and ε_{ij} is a necessarily positive multiplicative error term that reflects the imperfect determination of migration by distance and population size. When ε_{ij} is 1, Y_{ij} is equal to its “predicted” value \tilde{Y}_{ij} , given by the systematic part of the model; when ε_{ij} is less than 1, Y_{ij} is smaller than \tilde{Y}_{ij} ; and when ε_{ij} is greater than 1, Y_{ij} exceeds \tilde{Y}_{ij} .²¹ I will say more about the error presently.

²¹Because of the multiplicative form of the gravity model, \tilde{Y}_{ij} is not $E(Y_{ij})$ —hence the use of the term *predicted* rather than *expected* value.

Although the gravity model (17.3) is nonlinear in its parameters, it can be transformed into a linear equation by taking logs:²²

$$\begin{aligned}\log Y_{ij} &= \log \alpha + \beta \log P_i + \gamma \log P_j - \delta \log D_{ij} + \log \varepsilon_{ij} \\ Y'_{ij} &= \alpha' + \beta P'_i + \gamma P'_j + \delta D'_{ij} + \varepsilon'_{ij}\end{aligned}\quad (17.4)$$

where

$$\begin{aligned}\alpha' &\equiv \log \alpha \\ P'_i &\equiv \log P_i \\ P'_j &\equiv \log P_j \\ D'_{ij} &\equiv -\log D_{ij} \\ \varepsilon'_{ij} &\equiv \log \varepsilon_{ij}\end{aligned}$$

If we can make the usual linear-model assumptions about the transformed errors ε'_{ij} , then we are justified in fitting the transformed model (17.4) by linear least squares. In the gravity model, it is probably unrealistic to assume that the transformed errors are independent, because individual cities are involved in many different migration streams. A particularly attractive city, for example, might have positive errors for each of its in-migration streams and negative errors for each of its out-migration streams.²³

Our ability to linearize the model given in Equation 17.3 by a log transformation depends on the multiplicative errors in this model. The multiplicative error specifies that the general magnitude of the difference between Y_{ij} and \tilde{Y}_{ij} is proportional to the size of the latter: The model tends to make larger absolute errors in predicting large migration streams than in predicting small ones. This assumption appears reasonable here. In most cases, we would prefer to specify a form of error—additive or multiplicative—that leads to a simple statistical analysis—supposing, of course, that the specification is sensible. A subsequent analysis of residuals permits us to subject these assumptions to scrutiny.

Another form of multiplicative model is

$$\begin{aligned}Y_i &= \alpha \exp(\beta_1 X_{i1}) \exp(\beta_2 X_{i2}) \cdots \exp(\beta_k X_{ik}) \varepsilon_i \\ &= \alpha \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \varepsilon_i\end{aligned}\quad (17.5)$$

Taking logs produces the linear equation

$$Y'_i = \alpha' + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon'_i$$

with

$$\begin{aligned}Y'_i &\equiv \log_e Y_i \\ \alpha' &\equiv \log_e \alpha \\ \varepsilon'_i &\equiv \log_e \varepsilon_i\end{aligned}$$

²²The log transformation requires that $Y_{ij}, \alpha, P_i, P_j, D_{ij}$, and ε_{ij} are all positive, as is the case for the gravity model of migration.

²³See Exercise 17.7 for an illustrative application of the gravity model.

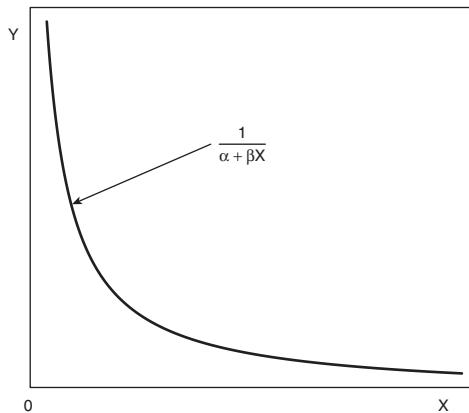


Figure 17.7 The model $Y = 1/(\alpha + \beta X_i + \varepsilon_i)$, for $X > 0$ and $\beta > 0$.

In Equation 17.5, the impact on Y of increasing X_j by one unit is proportional to the level of Y . The effect of rescaling Y by taking logs is to eliminate interaction among the X s. A similar result is, at times, achievable empirically through other power transformations of Y .²⁴

Some nonlinear models can be rendered linear by a transformation. For example, the multiplicative gravity model of migration,

$$Y_{ij} = \alpha \frac{P_i^\beta P_j^\gamma}{D_{ij}^\delta} \varepsilon_{ij}$$

(where Y_{ij} is the number of migrants moving from location i to location j , P_i is the population at location i , D_{ij} is the distance separating the two locations, and ε_{ij} is a multiplicative error term), can be linearized by taking logs.

Multiplicative models provide the most common instance of transformable nonlinearity, but there are also other models to which this approach is applicable. Consider, for example, the model

$$Y_i = \frac{1}{\alpha + \beta X_i + \varepsilon_i} \quad (17.6)$$

where ε_i is a random error satisfying the standard assumptions. Then, if we take $Y'_i \equiv 1/Y_i$, we can rewrite the model as the linear equation $Y'_i = \alpha + \beta X_i + \varepsilon_i$. This model is illustrated in Figure 17.7 (for a positive β and positive values of X).²⁵

²⁴The use of power transformations to promote additivity was pioneered by Tukey (1949); also see, for example, Emerson and Hoaglin (1983).

²⁵The graph shows the systematic part of the model, $\tilde{Y} = 1/(\alpha + \beta X)$, but because the transformation of Y that linearizes the model is not a linear transformation, the curve does not give the expectation of Y .

17.4 Nonlinear Least Squares*

Models that are nonlinear in the parameters and that cannot be rendered linear by a transformation are called *essentially nonlinear*. The *general nonlinear model* is given by the equation

$$Y_i = f(\boldsymbol{\beta}, \mathbf{x}'_i) + \varepsilon_i \quad (17.7)$$

in which

- Y_i is the response-variable value for the i th of n observations,
- $\boldsymbol{\beta}$ is a vector of p parameters to be estimated from the data,
- \mathbf{x}'_i is a row vector of scores for observation i on the k explanatory variables (some of which may be qualitative), and
- ε_i is the error for the i th observation.

It is convenient to write the model in matrix form for the full sample of n observations as

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\beta}, \mathbf{X}) + \boldsymbol{\varepsilon}$$

I will assume, as in the general linear model, that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.²⁶

An illustrative essentially nonlinear model is the logistic population-growth model (Shryock, Siegel, & Associates, 1973, pp. 382–385):

$$Y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 X_i)} + \varepsilon_i \quad (17.8)$$

where Y_i is population size, and X_i is time; for equally spaced observations, it is conventional to take $X_i = i - 1$, and so $X = 0, 1, 2, \dots$. Because the logistic growth model is fit to time-series data, the assumption of independent errors is problematic: It may well be the case that the errors are *autocorrelated*—that is, that errors close in time tend to be similar.²⁷ The additive form of the error is also questionable here, for errors may well grow larger in magnitude as population size increases.²⁸ Despite these potential difficulties, the logistic population-growth model can provide a useful, if gross and preliminary, representation of the data.

Under the assumption of independent and normally distributed errors, with 0 expectations and common variance, the general nonlinear model (Equation 17.7) has likelihood

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{i=1}^n [Y_i - f(\boldsymbol{\beta}, \mathbf{x}'_i)]^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} S(\boldsymbol{\beta})\right] \end{aligned}$$

where $S(\boldsymbol{\beta})$ is the sum-of-squares function

²⁶For multiplicative errors, we can put the model in the form of Equation 17.7 by taking logs.

²⁷See the discussion of time-series regression in Chapter 16.

²⁸See Exercise 17.9.

$$S(\beta) \equiv \sum_{i=1}^n [Y_i - f(\beta, \mathbf{x}'_i)]^2$$

As for the general linear model, we therefore maximize the likelihood by minimizing the sum of squared errors $S(\beta)$.

To derive estimating equations for the nonlinear model, we need to differentiate $S(\beta)$, obtaining

$$\frac{\partial S(\beta)}{\partial \beta} = -2 \sum [Y_i - f(\beta, \mathbf{x}'_i)] \frac{\partial f(\beta, \mathbf{x}'_i)}{\partial \beta}$$

Setting these partial derivatives to $\mathbf{0}$ and replacing the unknown parameters β with the estimator \mathbf{b} produces the *nonlinear least-squares* estimating equations. It is convenient to write the estimating equations in matrix form as

$$[\mathbf{F}(\mathbf{b}, \mathbf{X})]' [\mathbf{y} - \mathbf{f}(\mathbf{b}, \mathbf{x})] = \mathbf{0} \quad (17.9)$$

where $\mathbf{F}_{(n \times p)}(\mathbf{b}, \mathbf{X})$ is the matrix of derivatives, with i, j th entry

$$F_{ij} \equiv \frac{\partial f(\mathbf{b}, \mathbf{x}'_i)}{\partial b_j}$$

The solution \mathbf{b} of Equation 17.9 is the maximum-likelihood estimate of β . If there is more than one root to the estimating equations, then we choose the solution associated with the smallest residual sum of squares $S(\mathbf{b})$.

Nonlinear models of the form $Y_i = f(\beta, \mathbf{x}'_i) + \varepsilon_i$ can be estimated by nonlinear least squares, finding the value of \mathbf{b} that minimizes $S(\mathbf{b}) = \sum_{i=1}^n [Y_i - f(\mathbf{b}, \mathbf{x}'_i)]^2$.

17.4.1 Minimizing the Residual Sum of Squares

Because the estimating equations (17.9) arising from a nonlinear model are, in general, themselves nonlinear, their solution is often difficult. It is, for this reason, unusual to obtain nonlinear least-squares estimates by explicitly solving the estimating equations. Instead, it is more common to work directly with the sum-of-squares function.

There are several practical methods for obtaining nonlinear least-squares estimates. I will pursue in some detail a technique called *steepest descent*. Although the method of steepest descent usually performs poorly relative to alternative procedures, the rationale of the method is simple. Furthermore, many general aspects of nonlinear least-squares calculations can be explained clearly for the steepest-descent procedure. Because of the practical limitations of steepest descent, however, I will also briefly describe two superior procedures—the *Gauss-Newton method* and the *Marquardt method*—without developing their rationales.²⁹

The method of steepest descent, like other methods for calculating nonlinear least-squares estimates, begins with a vector $\mathbf{b}^{(0)}$ of initial estimates. These initial estimates can be obtained

²⁹See the recommended readings at the end of the chapter for details.

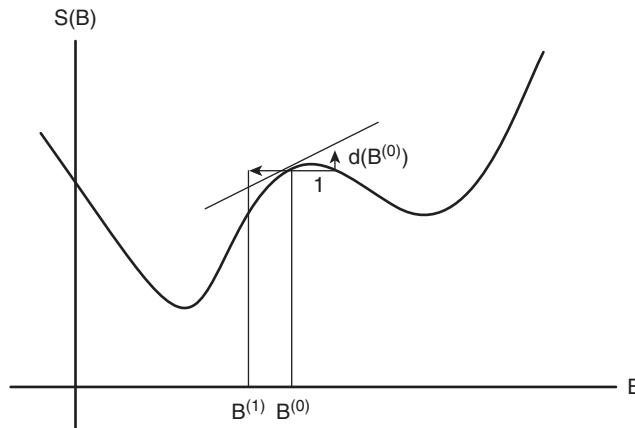


Figure 17.8 The method of steepest descent for one estimated parameter, B . Because the slope of the sum-of-squares function $S(B)$ is positive above the initial estimate $B^{(0)}$, the first step is to the *left*, to $B^{(1)}$.

in a variety of ways. We can, for example, choose p “typical” observations, substitute their values into the model given in Equation 17.7 (page 515), and solve the resulting system of p nonlinear equations for the p parameters.

Alternatively, we can select a set of reasonable trial values for each parameter, find the residual sum of squares for every combination of trial values, and pick as initial estimates the combination associated with the smallest residual sum of squares. It is sometimes possible to choose initial estimates on the basis of prior research, hypothesis, or substantive knowledge of the process being modeled.

It is unfortunate that the choice of starting values for the parameter estimates may prove consequential: Iterative methods such as steepest descent generally converge to a solution more quickly for initial values that are close to the final values, and even more important, the sum-of-squares function $S(\mathbf{b})$ may have local minima different from the global minimum (as illustrated in Figure 17.8).

Let us denote the *gradient* (i.e., derivative) vector for the sum-of-squares function as

$$\mathbf{d}(\mathbf{b}) = \frac{\partial S(\mathbf{b})}{\partial \mathbf{b}}$$

The vector $\mathbf{d}(\mathbf{b}^{(0)})$ gives the direction of maximum *increase* of the sum-of-squares function from the initial point $\{\mathbf{b}^{(0)}, S(\mathbf{b}^{(0)})\}$; the *negative* of this vector, $-\mathbf{d}(\mathbf{b}^{(0)})$, therefore, gives the *direction of steepest descent*. Figure 17.8 illustrates these relations for the simple case of one parameter, where we can move either left or right from the initial estimate $B^{(0)}$.

If we move in the direction of steepest descent, then we can find a new estimated parameter vector

$$\mathbf{b}^{(1)} = \mathbf{b}^{(0)} - M_0 \mathbf{d}(\mathbf{b}^{(0)})$$

for which $S(\mathbf{b}^{(1)}) < S(\mathbf{b}^{(0)})$: Because $S(\mathbf{b})$ is, by definition, decreasing in the direction of steepest descent, unless we are already at a minimum, we can *always* choose a number M_0 small

enough to improve the residual sum of squares. We can, for instance, first try $M_0 = 1$; if this choice does not lead to a decrease in $S(\mathbf{b})$, then we can take $M_0 = \frac{1}{2}$ and so on.

Our new estimate $\mathbf{b}^{(1)}$ can subsequently be improved in the same manner, by finding

$$\mathbf{b}^{(2)} = \mathbf{b}^{(1)} - M_1 \mathbf{d}(\mathbf{b}^{(1)})$$

so that $S(\mathbf{b}^{(2)}) < S(\mathbf{b}^{(1)})$. This procedure continues iteratively until it converges on a solution \mathbf{b} —that is, until the changes in $S(\mathbf{b}^{(l)})$ and $\mathbf{b}^{(l)}$ from one iteration to the next are negligible. In practice, however, the method of steepest descent often converges painfully slowly and at times falls prey to other computational difficulties.

At each iteration, we need to compute the gradient vector $\mathbf{d}(\mathbf{b})$ for the current value of $\mathbf{b} = \mathbf{b}^{(l)}$. From our previous work in this section, we have

$$\begin{aligned} -\mathbf{d}(\mathbf{b}) &= 2[\mathbf{F}(\mathbf{b}, \mathbf{X})]'[\mathbf{y} - \mathbf{f}(\mathbf{b}, \mathbf{x})] \\ &= 2 \sum_{i=1}^n \left[\frac{\partial f(\mathbf{b}, \mathbf{x}'_i)}{\partial \mathbf{b}} \right] [Y_i - f(\mathbf{b}, \mathbf{x}'_i)] \end{aligned} \quad (17.10)$$

The partial derivatives $\partial f(\mathbf{b}, \mathbf{x}'_i)/\partial B_j$ either can be supplied analytically (which is generally preferable) or can be evaluated numerically [i.e., approximated by finding the slope of $f(\mathbf{b}, \mathbf{x}'_i)$ in a small interval around the current value of B_j]. For example, for the logistic growth model (Equation 17.8 on page 515) discussed earlier in this section, the analytic derivatives are

$$\begin{aligned} \frac{\partial f(\mathbf{b}, X_i)}{\partial B_1} &= [1 + \exp(B_2 + B_3 X_i)]^{-1} \\ \frac{\partial f(\mathbf{b}, X_i)}{\partial B_2} &= -B_1 [1 + \exp(B_2 + B_3 X_i)]^{-2} \exp(B_2 + B_3 X_i) \\ \frac{\partial f(\mathbf{b}, X_i)}{\partial B_3} &= -B_1 [1 + \exp(B_2 + B_3 X_i)]^{-2} \exp(B_2 + B_3 X_i) X_i \end{aligned}$$

In the method of steepest descent, we take

$$\mathbf{b}^{(l+1)} = \mathbf{b}^{(l)} + M_l \mathbf{F}'_l \mathbf{e}^{(l)}$$

where $\mathbf{F}_l \equiv \mathbf{F}(\mathbf{b}^{(l)}, \mathbf{X})$ and $\mathbf{e}^{(l)} = \mathbf{y} - \mathbf{f}(\mathbf{b}^{(l)}, \mathbf{X})$ (and the constant 2 in Equation 17.10 is absorbed into M_l). The Gauss-Newton method, in contrast, calculates

$$\mathbf{b}^{(l+1)} = \mathbf{b}^{(l)} + M_l (\mathbf{F}'_l \mathbf{F}_l)^{-1} \mathbf{F}'_l \mathbf{e}^{(l)}$$

As for steepest descent, the step-size M_l is selected so that $S(\mathbf{b}^{(l+1)}) < S(\mathbf{b}^{(l)})$; we first try $M_l = 1$, then $M_l = \frac{1}{2}$, and so on. The direction chosen in the Gauss-Newton procedure is based on a first-order Taylor-series expansion of $S(\mathbf{b})$ around $S(\mathbf{b}^{(l)})$.

In the Marquardt procedure,

$$\mathbf{b}^{(l+1)} = \mathbf{b}^{(l)} + (\mathbf{F}'_l \mathbf{F}_l + M_l \mathbf{I}_p)^{-1} \mathbf{F}'_l \mathbf{e}^{(l)}$$

Initially, M_0 is set to some small number, such as 10^{-8} . If $S(\mathbf{b}^{(l+1)}) < S(\mathbf{b}^{(l)})$, then we accept the new value of $\mathbf{b}^{(l+1)}$ and proceed to the next iteration, with $M_{l+1} = M_l/10$; if, however, $S(\mathbf{b}^{(l+1)}) > S(\mathbf{b}^{(l)})$, then we increase M_l by a factor of 10 and try again. When M is small, the Marquardt procedure is similar to Gauss-Newton; as M grows larger, Marquardt approaches steepest descent. Marquardt's method is thus an adaptive compromise between the other two approaches.

Estimated asymptotic sampling covariances for the parameter estimates can be obtained by the maximum-likelihood approach and are given by³⁰

$$\widehat{\mathcal{V}}(\mathbf{b}) = S_E^2 \{ [\mathbf{F}(\mathbf{b}, \mathbf{X})]' \mathbf{F}(\mathbf{b}, \mathbf{X}) \}^{-1} \quad (17.11)$$

We can estimate the error variance from the residuals $\mathbf{e} = \mathbf{y} - \mathbf{f}(\mathbf{b}, \mathbf{X})$, according to the formula³¹

$$S_E^2 = \frac{\mathbf{e}' \mathbf{e}}{n - p}$$

Note the similarity of Equation 17.11 to the familiar linear least-squares result, $\widehat{\mathcal{V}}(\mathbf{b}) = S_E^2 (\mathbf{X}' \mathbf{X})^{-1}$. Indeed, $\mathbf{F}(\mathbf{b}, \mathbf{X}) = \mathbf{X}$ for the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$.

Iterative methods for finding the nonlinear least-squares estimates include the method of steepest descent and, more practically, the Gauss-Newton and Marquardt methods. Estimated asymptotic covariances for the coefficients are given by

$$\widehat{\mathcal{V}}(\mathbf{b}) = S_E^2 \{ [\mathbf{F}(\mathbf{b}, \mathbf{X})]' \mathbf{F}(\mathbf{b}, \mathbf{X}) \}^{-1}$$

where $\mathbf{F}(\mathbf{b}, \mathbf{X})$ is the matrix of derivatives, with i,j th entry $\partial f(\mathbf{b}, \mathbf{x}_i') / \partial B_j$, and $S_E^2 = \sum E_i^2 / (n - p)$ is the estimated error variance.

17.4.2 An Illustration: U.S. Population Growth

Decennial population data for the United States appear in Table 17.2 for the period from 1790 to 2010; the data are plotted in Figure 17.9(a). Let us fit the logistic growth model (Equation 17.8 on page 515) to these data using nonlinear least squares.

The parameter β_1 of the logistic growth model gives the asymptote that expected population approaches as time increases. In 2010, when $Y = 308.746$ (million), population did not appear to be near an asymptote;³² so as not to extrapolate too far beyond the data, I will arbitrarily set $B_1^{(0)} = 350$. At time $X_1 = 0$, we have

$$Y_1 = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 0)} + \varepsilon_1 \quad (17.12)$$

Ignoring the error, using $B_1^{(0)} = 350$, and substituting the observed value of $Y_1 = 3.929$ into Equation 17.12, we get $\exp(B_2^{(0)}) = (350/3.929) - 1$, or $B_2^{(0)} = \log_e 88.081 = 4.478 \approx 4.5$. At time $X_2 = 1$,

$$Y_2 = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 1)} + \varepsilon_2$$

³⁰See Bard (1974, pp. 176–179).

³¹Alternatively, we can use the maximum-likelihood estimator of the error variance, without the correction for “degrees of freedom,” $\hat{\sigma}_e^2 = \mathbf{e}' \mathbf{e} / n$.

³²That population in 2010 did not appear to be near an asymptote suggests that we might not need to fit an asymptotic growth model to the data; see Exercise 17.10.

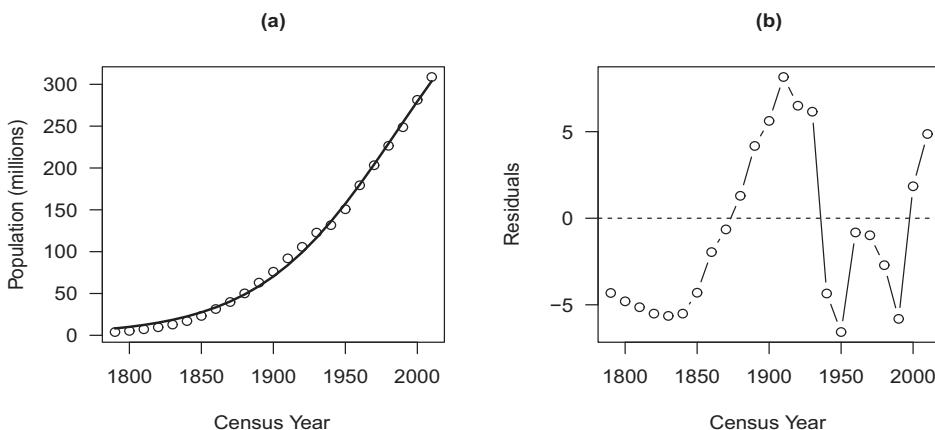


Figure 17.9 Panel (a) shows the population of the United States from 1790 through 2010; the line represents the fitted logistic growth model. Residuals from the logistic growth model are plotted against time in panel (b).

Table 17.2 Population of the United States, in Millions, 1790–2010

Year	Population	Year	Population
1790	3.929	1900	75.995
1800	5.308	1910	91.972
1810	7.240	1920	105.711
1820	9.638	1930	122.775
1830	12.866	1940	131.669
1840	17.069	1950	150.697
1850	23.192	1960	179.323
1860	31.443	1970	203.302
1870	39.818	1980	226.542
1880	50.156	1990	248.718
1890	62.948	2000	281.425
		2010	308.746

SOURCE: U.S. Bureau of the Census (2006, 2011).

Again ignoring the error, and making appropriate substitutions, $\exp(4.5 + B_3^{(0)}) = (350/5.308) - 1$, or $B_3^{(0)} = \log_e 64.938 - 4.5 = -0.327 \approx -0.3$.

The Gauss-Newton iterations based on these start values are shown in Table 17.3. Asymptotic standard errors for the coefficients also appear in this table and indicate that (with the exception of the population asymptote β_1) the parameters are estimated precisely. Although the logistic model captures the major trend in U.S. population growth, the residuals from the

Table 17.3 Gauss-Newton Iterations for the Logistic Growth Model Fit to the U.S. Population Data

Iteration	Residual Sum of Squares	Coefficients		
		B_1	B_2	B_3
0	13,374.54	350.0	4.5	-0.3
1	5,378.14	387.65	3.7000	-0.20173
:				
5	512.77	487.66	4.0628	-0.20753
Final	512.77	487.65	4.0629	-0.20754
	Standard error	35.60	0.0630	0.00886

NOTE: Asymptotic standard errors are given below the final coefficient estimates.

least-squares fit [plotted against time in Figure 17.9(b)] suggest that the error variance is not constant and that the residuals are autocorrelated.³³ Note as well the large drop in the residual for 1940 and the large increase for 1960 (when Alaska and Hawaii were first included in the population count) and again in 2000.

An example of an essentially nonlinear model, which requires nonlinear least squares, is the logistic population-growth model

$$Y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 X_i)} + \varepsilon_i$$

where Y_i is population size, and X_i is time.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 17.1. *Interpreting effects in nonlinear models (based on Stolzenberg, 1979): For simplicity, disregard the error and let Y represent the systematic part of the response variable. Suppose that Y is a function of two explanatory variables, $Y = f(X_1, X_2)$.

- The *metric effect* of X_1 on Y is defined as the partial derivative $\partial Y / \partial X_1$.
- The *effect of proportional change* in X_1 on Y is defined as $X_1 (\partial Y / \partial X_1)$.
- The *instantaneous rate of return* of Y with respect to X_1 is $(\partial Y / \partial X_1) / Y$.
- The *point elasticity* of Y with respect to X_1 is $(\partial Y / \partial X_1)(X_1 / Y)$.

³³See Exercise 17.8.

Find each of these four measures of the effect of X_1 in each of the following models. Which measure yields the simplest result in each case? How can the several measures be interpreted? How would you fit each model to data, assuming convenient forms for the errors [e.g., additive errors for models (a), (b), and (c)]?

- (a) $Y = \alpha + \beta_1 X_1 + \beta_2 X_2.$
- (b) $Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2.$
- (c) $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$
- (d) $Y = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2).$
- (e) $Y = \alpha X_1^{\beta_1} X_2^{\beta_2}.$

Exercise 17.2. *Orthogonal polynomial contrasts: The polynomial regressors X, X^2, \dots, X^{m-1} generated to represent a quantitative, discrete X with values $1, 2, \dots, m$ are substantially correlated. It is convenient (but by no means essential) to remove these correlations. Suppose that there are equal numbers of observations in the different levels of X , so that it suffices to make the columns of the row basis of the model matrix for X orthogonal. Working with the row basis, begin by subtracting the mean from X , calling the result X^* . *Centering* X in this manner makes X^* orthogonal to the constant regressor $\mathbf{1}$. (Why?) X^2 can be made orthogonal to the constant and X^* by projecting the X^2 vector onto the subspace generated by $\mathbf{1}$ and X^* ; call the residual from this projection X^{*2} . The remaining columns X^{*3}, \dots, X^{*m-1} of the new row basis are formed in a similar manner, each orthogonal to the preceding ones.

- (a) Show that the orthogonal polynomial contrasts $\mathbf{1}, X^*, \dots, X^{*m-1}$ span the same subspace as the original polynomial regressors $\mathbf{1}, X, \dots, X^{m-1}$.
- (b) Show that the incremental sum of squares for each orthogonal contrast $X^*, X^{*2}, \dots, X^{*m-1}$ is the same as the step-down sum of squares for the corresponding regressor among the original (correlated) polynomial terms, X^{m-1}, \dots, X^2, X . (*Hint:* Remember that $X^*, X^{*2}, \dots, X^{*m-1}$ are uncorrelated.)
- (c) What, then, is the advantage of orthogonal polynomial contrasts?
- (d) Can the same approach be applied to a continuous quantitative explanatory variable—defining, for example, a quadratic component that is orthogonal to the linear component and a cubic component orthogonal to both the linear and quadratic components?

Exercise 17.3. Working with the full quadratic regression model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

and a three-dimensional graphics program, draw pictures of the regression surface (similar to Figure 17.1 on page 504) for various values of the parameters β_1, \dots, β_5 and various values of the explanatory variables X_1 and X_2 . You will derive a better impression of the flexibility of this model if you experiment freely, but the following suggestions may prove useful:

- Try both positive and negative values of the parameters.
- Try cases in which there are both positive and negative X s, as well as cases in which the X s are all positive.
- Try setting some of the parameters to 0.

Exercises 17.4. Cowles and Davis's logistic regression of volunteering on sex, neuroticism, extraversion and the product of neuroticism and extraversion is discussed in Section 17.1. Show that (a) interactions between sex and the other factors and (b) squared terms in neuroticism and extraversion are not required in this model.

Exercise 17.5. *Show that in the model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

the lines for the regression of Y on X_1 at various fixed values of X_2 all cross at a point (as in Figure 17.2 for Cowles and Davis's logistic regression).

Exercise 17.6. *Properties of piece-wise fits and regression splines:

- (a) For the regression equation

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

where $X_{il} \equiv X_i$;

$$X_{i2} \equiv \begin{cases} 0 & \text{for } X_i \leq k_1 \\ X_i - k_1 & \text{for } X_i > k_1 \end{cases}$$

and

$$X_{i3} \equiv \begin{cases} 0 & \text{for } X_i \leq k_2 \\ X_i - k_2 & \text{for } X_i > k_2 \end{cases}$$

show that the regression lines in the three bins are continuous (i.e., join at knots on the bin boundaries at $X = k_1$ and $X = k_2$).

- (b) Similarly, show that the cubics in the regression equation

$$\begin{aligned} Y_i = & \alpha + \beta_{11} X_{i1} + \beta_{12} X_{i1}^2 + \beta_{13} X_{i1}^3 + \beta_{21} X_{i2} + \beta_{22} X_{i2}^2 + \beta_{23} X_{i2}^3 \\ & + \beta_{31} X_{i3} + \beta_{32} X_{i3}^2 + \beta_{33} X_{i3}^3 + \varepsilon_i \end{aligned}$$

are continuous, that the cubics in the regression equation

$$Y_i = \alpha + \beta_{11} X_{i1} + \beta_{12} X_{i1}^2 + \beta_{13} X_{i1}^3 + \beta_{22} X_{i2}^2 + \beta_{23} X_{i2}^3 + \beta_{32} X_{i3}^2 + \beta_{33} X_{i3}^3 + \varepsilon_i$$

are both continuous and have continuous slopes at the knots, and that the cubics in the regression-spline equation

$$Y_i = \alpha + \beta_{11} X_{i1} + \beta_{12} X_{i1}^2 + \beta_{13} X_{i1}^3 + \beta_{23} X_{i2}^3 + \beta_{33} X_{i3}^3 + \varepsilon_i$$

join at the knots, have continuous slopes, and have continuous curvature.

Exercise 17.7. Table 17.4 reports interprovincial migration in Canada for the period 1966 to 1971. Also shown in this table are the 1966 and 1971 provincial populations. Table 17.5 gives road distances among the major cities in the 10 provinces. Averaging the 1966 and 1971 population figures, fit the gravity model of migration (Equation 17.3 on page 512) to the interprovincial migration data. Display the residuals from the fitted model in a 10×10 table. Can you account for the pattern of residuals? How might the model be modified to provide a more

Table 17.4 Canadian Interprovincial Migration and Population for the Period 1966-1971

1971 Residence	1966 Residence							AB	BC
	NL	PE	NS	NB	QC	ON	MB		
Newfoundland	255	2380	1140	2145	6295	215	185	425	425
Prince Edward Island	340	1975	1310	755	3060	400	95	185	330
Nova Scotia	3340	2185	8310	6090	18,805	1825	840	2000	2490
New Brunswick	1740	1335	7635	9315	12,455	1405	480	1130	1195
Quebec	2235	635	4350	7905	48,370	4630	1515	3305	4740
Ontario	17,860	3570	25,730	18,550	99,430	23,785	11,805	17,655	21,205
Manitoba	680	265	1655	1355	4330	18,245	16,365	7190	6310
Saskatchewan	280	125	620	495	1570	6845	9425	10,580	6090
Alberta	805	505	3300	2150	7750	23,550	17,410	41,910	27,765
British Columbia	1455	600	6075	3115	16,740	47,395	26,910	29,920	58,915
1966 Population	493,396	108,535	756,039	616,788	5,780,845	6,960,870	963,066	955,344	1,463,203
1971 Population	522,104	111,641	788,960	534,557	6,027,764	7,703,106	988,247	926,242	1,627,874
									1,873,674 2,184,621

SOURCES: Statistics Canada (1971, Vol. 1, Part 2, Table 32) and Department of Mines and Technical Surveys (1962).

Table 17.5 Road Distances in Miles Among Major Canadian Cities

City	NL	PE	NS	NB	QC	ON	MB	SK	AB	BC
St. John's, NL	0	924	952	1119	1641	1996	3159	3542	4059	4838
Charlottetown, PE	924	0	164	252	774	1129	2293	2675	3192	3972
Halifax, NS	952	164	0	310	832	1187	2351	2733	3250	4029
Fredericton, NB	1119	252	310	0	522	877	2041	2423	2940	3719
Montreal, QC	1641	774	832	522	0	355	1519	1901	2418	3197
Toronto, ON	1996	1129	1187	877	355	0	1380	1763	2281	3059
Winnipeg, MB	3159	2293	2351	2041	1519	1380	0	382	899	1679
Regina, SK	3542	2675	2733	2423	1901	1763	382	0	517	1297
Edmonton, AB	4059	3192	3250	2940	2418	2281	899	517	0	987
Vancouver, BC	4838	3972	4029	3719	3197	3059	1679	1297	987	0

SOURCE: Canada (1962).

satisfactory fit to the data? (Why can't we simply use dummy regressors to incorporate province effects for the source and destination provinces?)

Exercise 17.8. Calculate the autocorrelation for the residuals from the logistic growth model fit to the U.S. population data in Section 17.4.2. Recalling the discussion of autocorrelated errors in *linear* regression (in Chapter 16), does autocorrelation appear to be a serious problem here?

Exercise 17.9 Using nonlinear least squares, refit the logistic growth model to the U.S. population data (given in Table 17.2) assuming multiplicative rather than additive errors:

$$Y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 X_i)} \varepsilon_i$$

Which form of the model appears more adequate for these data?

Exercise 17.10. As mentioned in Section 17.4.2, the population of the United States in the year 2010 did not seem to be near an asymptote. As an alternative to the logistic growth model, we might entertain the exponential growth model for the period 1790 to 2010; assuming multiplicative errors, this model takes the form

$$Y_i = \alpha \exp(\beta X_i) \varepsilon_i$$

where, as in the text, Y_i is population, and $X_i = 0, 1, \dots, 21$ is time. Because of the multiplicative errors, this model can be transformed to linearity by taking the log of both sides:

$$\log_e Y_i = \alpha' + \beta X_i + \varepsilon'_i$$

where $\alpha' \equiv \log_e \alpha$ and $\varepsilon'_i \equiv \log_e \varepsilon_i$. Fit the exponential growth model to the data by linear least-squares regression and graph the fit as in Figure 17.9. [Hint: transform the fitted values back to the original population scale as $\exp(\hat{Y}_i)$]. Plot fitted values for the exponential growth model against those for the logistic growth model. Which model appears to do a better job of representing the data?

Exercise 17.11. Recall the Box-Tidwell regression model,

$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$

In Section 12.5.2, I described a procedure for fitting the Box-Tidwell model that relies on constructed variables. I applied this procedure to data from the Canadian Survey of Labour and Income Dynamics to fit the model

$$\log_2 \text{wages} = \alpha + \beta_1 \text{age}^{\gamma_1} + \beta_2 \text{education}^{\gamma_2} + \beta_3 \text{male} + \varepsilon$$

where *male* is a dummy regressor coded 1 for men and 0 for women. Fit this model to the data by general nonlinear least squares. Are there any advantages to using general nonlinear least squares in place of Box and Tidwell's procedure? Any disadvantages? Can the two approaches be combined?

Summary

- Nonlinear regression models that are linear in the parameters, for example, the quadratic regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

can be fit by linear least squares.

- Regression splines are piece-wise cubic polynomials that are continuous at join-points, called knots, and that are constrained to have equal slopes and curvature on either side of a knot. Although fully parametric, regression splines generally do a good job of responding to local aspects of the data and can be incorporated as building blocks into linear and generalized linear models.
- Some nonlinear models can be rendered linear by a transformation. For example, the multiplicative gravity model of migration,

$$Y_{ij} = \alpha \frac{P_i^\beta P_j^\gamma}{D_{ij}^\delta} \varepsilon_{ij}$$

(where Y_{ij} is the number of migrants moving from location i to location j , P_i is the population at location i , D_{ij} is the distance separating the two locations, and ε_{ij} is a multiplicative error term) can be linearized by taking logs.

- More generally, nonlinear models of the form $Y_i = f(\boldsymbol{\beta}, \mathbf{x}'_i) + \varepsilon_i$ (in which $\boldsymbol{\beta}$ is a vector of p parameters to be estimated, and \mathbf{x}'_i is a vector of explanatory-variable values) can be estimated by nonlinear least squares, finding the value of \mathbf{b} that minimizes

$$S(\mathbf{b}) = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n [Y_i - f(\mathbf{b}, \mathbf{x}'_i)]^2$$

- Iterative methods for finding the nonlinear least-squares estimates include the method of steepest descent and, more practically, the Gauss-Newton and Marquardt methods. Estimated asymptotic covariances for the coefficients are given by

$$\widehat{\mathcal{V}}(\mathbf{b}) = S_E^2 \{[\mathbf{F}(\mathbf{b}, \mathbf{X})]' \mathbf{F}(\mathbf{b}, \mathbf{X})\}^{-1}$$

where $\mathbf{F}(\mathbf{b}, \mathbf{X})$ is the matrix of derivatives, with i,j th entry $\partial f(\mathbf{b}, \mathbf{x}'_i) / \partial B_j$, and $S_E^2 = \sum E_i^2 / (n - p)$ is the estimated error variance.

- An example of an essentially nonlinear model, which requires nonlinear least squares, is the logistic population-growth model,

$$Y_i = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 X_i)} + \varepsilon_i$$

where Y_i is population size and X_i is time.

Recommended Reading

- Hastie et al. (2009, chap. 5) provide a rigorous treatment of regression splines, explaining their relationship to smoothing splines for nonparametric regression.³⁴ Regression splines also figure prominently in Harrell (2001).
- Further discussion of nonlinear least squares can be found in many sources, including Gallant (1975), Draper and Smith (1998, chap. 24), Greene (2003, chap. 9), Bard (1974), and Bates and Watts (1988) in rough order of increasing detail and difficulty.
- Draper and Smith (1998, chaps. 12 and 22) also discuss polynomial and orthogonal-polynomial regression models.

³⁴Nonparametric regression is taken up in the next chapter.

18

Nonparametric Regression

The essential idea of *nonparametric-regression analysis* was introduced in Chapter 2: to examine the conditional distribution of the response variable—or some aspect of that distribution, such as its center—as a function of one or more explanatory variables, without assuming in advance what form that function takes. This chapter elaborates that simple idea, developing methods of nonparametric simple and multiple regression for quantitative response variables, along with generalized nonparametric-regression models for categorical responses, for count data, and for non-normal quantitative response variables. Taken together, these methods provide a more flexible alternative to the *parametric* linear, generalized linear, and non-linear regression models described in the earlier chapters of the book.

18.1 Nonparametric Simple Regression: Scatterplot Smoothing

This section presents several methods of nonparametric simple regression: kernel regression; local-polynomial regression, which generalizes kernel regression; and smoothing splines. Because kernel regression and local-polynomial regression are mathematically simpler than smoothing splines, I will emphasize these methods in developing some of the statistical theory underlying nonparametric regression.

Nonparametric simple regression is useful in its own right and for its extension to and use in nonparametric multiple-regression and additive-regression models. A principal application of nonparametric simple regression is to examine the relationship between two quantitative variables in a scatterplot, and these methods are therefore often called “scatterplot smoothers.” Indeed, in the preceding chapters, I have often used the lowess local-regression smoother to facilitate the interpretation of scatterplots.

18.1.1 Kernel Regression

Kernel regression generalizes the simple local-averaging method of nonparametric regression described in Chapter 2.¹ Suppose that we wish to estimate the regression function $Y = f(x) + \varepsilon$ at a particular value of the explanatory variable, $X = x_0$. As in linear and

¹See Section 2.3. It would be useful to reread that section now and generally to review the material in Chapter 2.

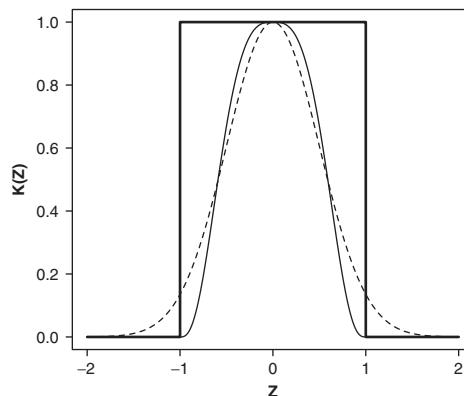


Figure 18.1 Tricube (light solid line), normal (broken line), and rectangular (heavy solid line) kernel functions. The normal kernel is rescaled to facilitate comparison.

nonlinear regression models, we will assume that the error ε is normally and independently distributed with an expectation of 0 and constant variance σ_ε^2 .² As well, although the regression function $f(\cdot)$ is left unspecified, we will assume that it is smooth and continuous.³

The basic idea of kernel regression is that in estimating $f(x_0)$, it is desirable to give greater weight to observations that are close to the focal x_0 and less weight to those that are remote. Let $z_i \equiv (x_i - x_0)/h$ denote the scaled, signed distance between the X -value for the i th observation and the focal x_0 . As I will explain shortly, the scale factor h , called the *bandwidth* of the kernel estimator, plays a role similar to the window width of a local average and controls the smoothness of the kernel estimator.

We need a *kernel function* $K(z)$ that attaches greatest weight to observations that are close to the focal x_0 and then falls off symmetrically and smoothly as $|z|$ grows.⁴ Given these characteristics, the specific choice of a kernel function is not critical. Having calculated weights $w_i = K[(x_i - x_0)/h]$, we proceed to compute a fitted value at x_0 by weighted local averaging of the Y values:

$$\hat{f}(x_0) = \hat{Y}|_{x_0} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

Two popular choices of kernel functions, illustrated in Figure 18.1, are the *Gaussian* or *normal kernel* and the *tricube kernel*:

²For simplicity of exposition, I will treat the explanatory variables in this chapter as fixed rather than random. As in linear regression, this stipulation can be relaxed by assuming that the errors are independent of the explanatory variables. See Chapter 6 and Section 9.6.

³This assumption of smoothness is common to *all* of the methods of nonparametric regression considered in the chapter. There are methods of nonparametric regression that can, for example, deal with discontinuities (such as wavelet regression—see, e.g., Nason & Silverman, 2000), but they are beyond the scope of the current discussion.

⁴Kernel functions were introduced in Section 3.1.2 in connection with nonparametric density estimation, which may be thought of as a univariate analog of kernel and local-polynomial regression.

- The Gaussian kernel is simply the standard-normal density function,

$$K_N(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Here, the bandwidth h is the standard deviation of a normal distribution centered at x_0 . Observations at distances greater than $2h$ from the focal value therefore receive nearly 0 weight, because the normal density is small beyond 2 standard deviations from the mean.

- The tricube kernel is

$$K_T(z) = \begin{cases} (1 - |z|^3)^3 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geq 1 \end{cases}$$

For the tricube kernel, h is the half-width of a window centered at the focal x_0 . Observations that fall outside of the window receive 0 weight.

- Using a *rectangular kernel* (also shown in Figure 18.1)

$$K_R(z) = \begin{cases} 1 & \text{for } |z| < 1 \\ 0 & \text{for } |z| \geq 1 \end{cases}$$

gives equal weight to each observation in a window of half-width h centered at x_0 and 0 weight to observations outside of this window, producing an *unweighted* local average.⁵

I have implicitly assumed that the bandwidth h is *fixed*, but the kernel estimator is easily adapted to *nearest-neighbor* bandwidths, which include a constant number or proportion of the data. The adaptation is simplest for kernel functions, like the tricube kernel, that fall to 0: Simply adjust $h(x)$ so that a constant number of observations m is included in the window. The fraction m/n is called the *span* of the kernel smoother. It is common to evaluate the kernel estimator either at a number of values evenly distributed across the range of X or at the ordered observations $x_{(i)}$.

Nearest-neighbor kernel estimation is illustrated in Figure 18.2 for the relationship between the prestige and income levels of 102 Canadian occupations in 1971.⁶ Panel (a) shows a neighborhood containing 40 observations centered on the 80th ordered X -value. Panel (b) shows the tricube weight function defined on the window; the bandwidth $h[x_{(80)}]$ is selected so that the window accommodates the 40 nearest neighbors of the focal $x_{(80)}$. Thus, the span of the smoother is $40/102 \approx 0.4$. Panel (c) shows the locally weighted average, $\hat{Y}_{(80)} = \hat{Y}|x_{(80)}$; note that this is the fitted value associated with $x_{(80)}$, *not* the 80th ordered fitted value. Finally, panel (d) connects the fitted values above the $x_{(i)}$ to obtain the kernel estimate of the regression of prestige on income. In comparison to the local-average regression (Figure 2.8 on page 24), the kernel estimate is smoother, but it still exhibits artificial flattening at the boundaries (called *boundary bias*). Varying the span of the kernel estimator controls the smoothness of the estimated regression function: Larger spans produce smoother results. A simple approach to

⁵This is the local-averaging estimator described in Section 2.3.

⁶The Canadian occupational prestige data set was introduced in Chapter 2.

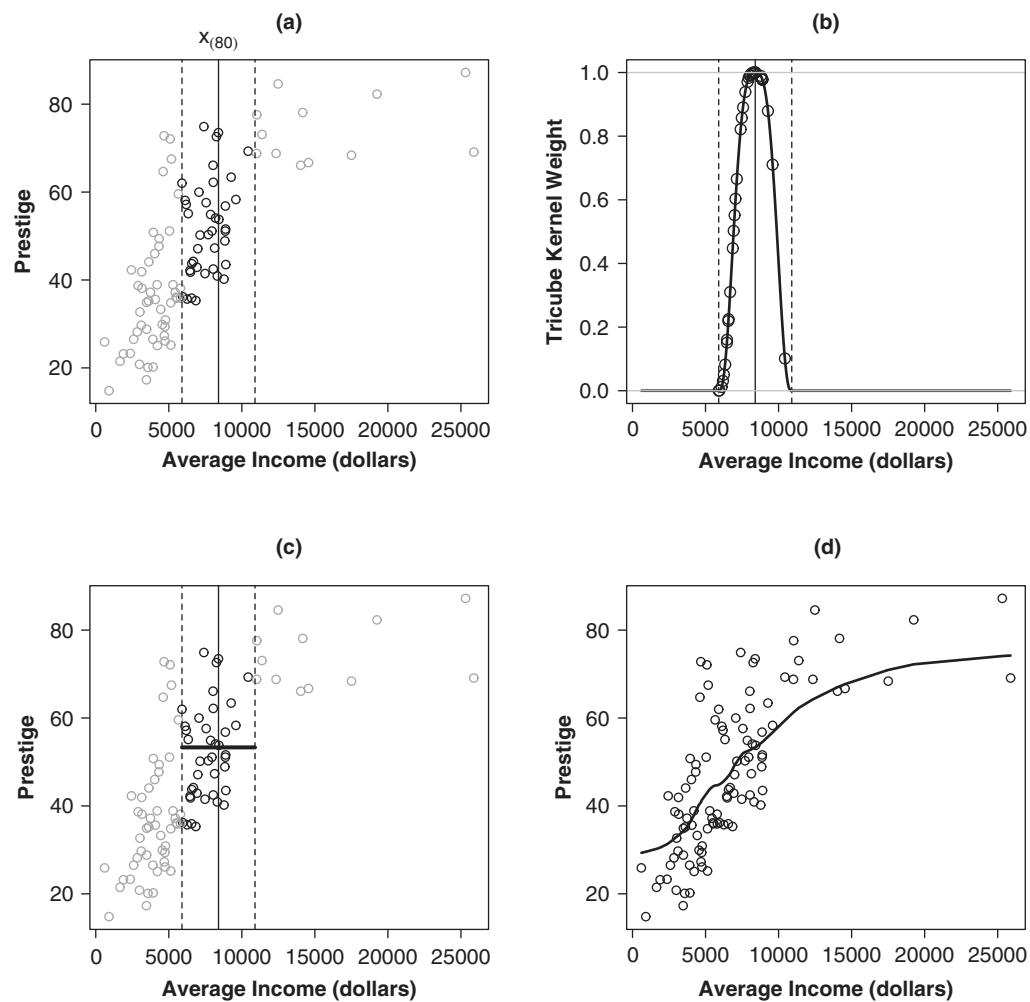


Figure 18.2 The kernel estimator applied to the Canadian occupational prestige data: (a) a window containing the $m = 40$ nearest X neighbors of the focal value $x_{(80)}$; (b) the tricube weight function and weights for observations within the window; (c) the weighted average $\hat{Y}_{(80)}$ of the Y values in the window; (d) the nonparametric regression line connecting the locally weighted averages centered at each $x_{(i)}$.

selecting the span is to pick the smallest value that produces an acceptably smooth fit to the data.⁷

⁷Choice of bandwidth or span is discussed in more detail in connection with local polynomial regression in the next section. See Exercise 18.1 for the effect on the kernel estimator of varying the span in the regression of occupational prestige on income.

Kernel regression estimates the regression function at a focal value x_0 of the explanatory variable by weighted local averaging of Y :

$$\hat{f}(x_0) = \hat{Y}|_{x_0} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

The weights are provided by a kernel function, $w_i = K[(x_i - x_0)/h]$, which takes on its largest value at $K(0)$ and falls symmetrically toward 0 as $|(x_i - x_0)/h|$ grows. Observations close to the focal x_0 therefore receive greatest weight. The kernel estimator is evaluated at representative focal values of X or at the ordered X -values, $x_{(i)}$. The bandwidth h of the kernel estimator can be fixed or can be adjusted to include a fixed proportion of the data, called the span of the kernel estimate. The larger the span, the smoother the kernel regression.

18.1.2 Local-Polynomial Regression

Local-polynomial regression corrects some of the deficiencies of kernel estimation. It provides a generally adequate method of nonparametric regression that extends straightforwardly to multiple regression, additive regression, and generalized nonparametric regression (as described later in this chapter).

We are familiar with polynomial regression,⁸ where a p th-degree polynomial in an explanatory variable X ,

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

is fit to data; $p = 1$ corresponds to a linear fit, $p = 2$ to a quadratic fit, and so on. Fitting a constant (i.e., the mean) corresponds to $p = 0$.

Local-polynomial regression extends kernel estimation to a polynomial fit at the focal value x_0 , using local kernel weights, $w_i = K[(x_i - x_0)/h]$. The resulting weighted least-squares (WLS) regression⁹ fits the equation

$$Y_i = A + B_1(x_i - x_0) + B_2(x_i - x_0)^2 + \cdots + B_p(x_i - x_0)^p + E_i$$

to minimize the weighted residual sum of squares, $\sum_{i=1}^n w_i E_i^2$. Once the WLS solution is obtained, the fitted value at the focal x_0 is just $\hat{Y}|_{x_0} = A$. As in kernel regression, this procedure is repeated for representative focal values of X or at the observations x_i .

Also as in kernel regression, we can employ a fixed bandwidth or adjust the bandwidth to include a fixed proportion—or span—of nearest neighbors to the focal value x_0 . Nearest-neighbor local-polynomial regression is often called *lowess* (an acronym for *locally weighted scatterplot smoother*, alternatively rendered as *loess*, for *local regression*)—a term with which we are already familiar.

Selecting $p = 1$ produces a local-linear fit, the most common case. The “tilt” of the local-linear fit promises reduced bias in comparison to the kernel estimator of the previous section,

⁸See Section 17.1.

⁹Weighted-least-squares regression is developed in Section 12.2.2. Centering at x_0 , by employing $x_i - x_0$, is convenient (but inessential) in that the fitted value at x_0 is then simply the intercept A (see below).

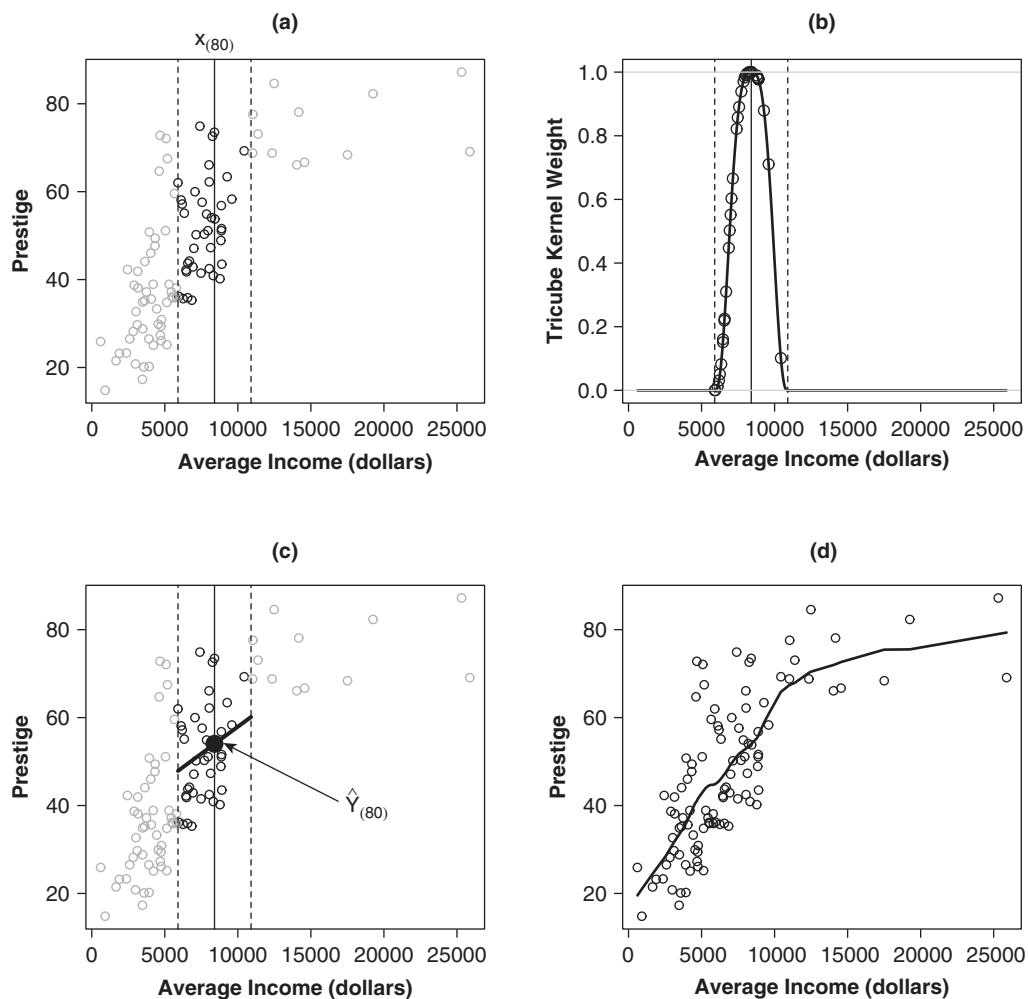


Figure 18.3 Nearest-neighbor local linear regression of prestige on income. The window in (a) includes the $m = 40$ nearest neighbors of the focal value $x_{(80)}$. The tricube weights for this window are shown in (b) and the locally weighted least-squares line in (c), producing the fitted value $\hat{Y}_{(80)}$. Fitted values for all the observations are connected in (d) to produce the nonparametric local-polynomial regression line.

which corresponds to $p = 0$. This advantage is most apparent at the boundaries, where the kernel estimator tends to flatten. The values $p = 2$ or $p = 3$, local quadratic or cubic fits, produce more flexible regressions. Greater flexibility has the potential to reduce bias further, but flexibility also entails the cost of greater sampling variation. There is, it turns out, a theoretical advantage to odd-order local polynomials, so $p = 1$ is generally preferred to $p = 0$ and $p = 3$ to $p = 2$. These issues are explored below.

Figure 18.3 illustrates the computation of a local-linear-regression fit to the Canadian occupational prestige data, using the tricube kernel function and nearest-neighbor bandwidths. Panel (a) shows a window accommodating the 40 nearest neighbors of the focal value $x_{(80)}$, corresponding

to a span of $40/102 \approx 0.4$. Panel (b) shows the tricube weight function defined on this window. The locally weighted linear fit appears in panel (c). Fitted values calculated at each observed X -value are connected in panel (d). There is no flattening of the fitted regression function at the boundaries, as there was for kernel estimation (cf. Figure 18.2 on page 531).

Local-polynomial regression extends kernel estimation to a polynomial fit at the focal value x_0 , using local kernel weights, $w_i = K[(x_i - x_0)/h]$. The resulting WLS regression fits the equation

$$Y_i = A + B_1(x_i - x_0) + B_2(x_i - x_0)^2 + \cdots + B_p(x_i - x_0)^p + E_i$$

to minimize the weighted residual sum of squares, $\sum_{i=1}^n w_i E_i^2$. The fitted value at the focal x_0 is just $\hat{Y}|_{x_0} = A$. This procedure is repeated for representative focal values of X , or at the observations x_i . We can employ a fixed bandwidth or adjust the bandwidth for a fixed span. Nearest-neighbor local-polynomial regression is often called lowess (or loess).

Selecting the Span

I will assume nearest-neighbor bandwidths, so bandwidth choice is equivalent to selecting the span of the local-regression smoother. I will also assume a locally linear fit. The methods of this section generalize in an obvious manner to fixed-bandwidth and higher-order polynomial smoothers.

A generally effective approach to selecting the span is guided trial and error. The span $s = 0.5$ is often a good point of departure. If the fitted regression looks too rough, then try increasing the span; if it looks smooth, then see if the span can be decreased without making the fit too rough. We want the *smallest* value of s that provides a smooth fit.

The terms *smooth* and *rough* are admittedly subjective, and a sense of what I mean here is probably best conveyed by example. An illustration, for the Canadian occupational prestige data, appears in Figure 18.4. For these data, selecting $s = 0.5$ or $s = 0.7$ appears to provide a reasonable compromise between smoothness and fidelity to the data.

More sophisticated methods for selecting the span will be described presently.¹⁰ The visual approach usually works very well, however, and visual trial and error should be performed even if more sophisticated approaches are used to provide an initial value of s .

A generally effective visual approach to selecting the span in local-polynomial regression is guided trial and error. The span $s = 0.5$ is often a good point of departure. If the fitted regression looks too rough, then try increasing the span; if it looks smooth, then see if the span can be decreased without making the fit too rough. We want the smallest value of s that provides a smooth fit.

¹⁰Also see Exercise 18.2.

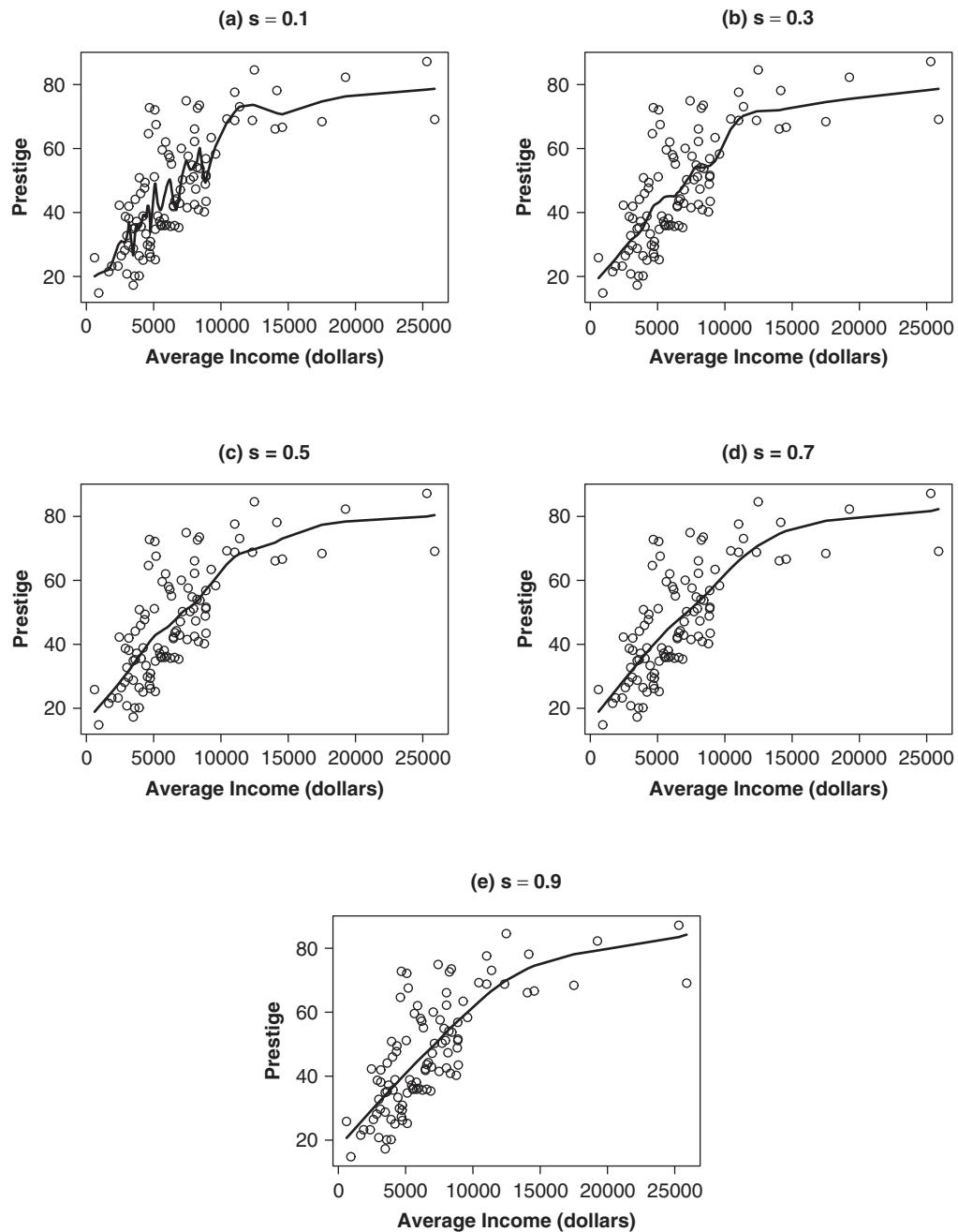


Figure 18.4 Nearest-neighbor local linear regression of prestige on income, for several values of the span s . The value $s = 0.5$ or 0.7 appears to reasonably balance smoothness with fidelity to the data.

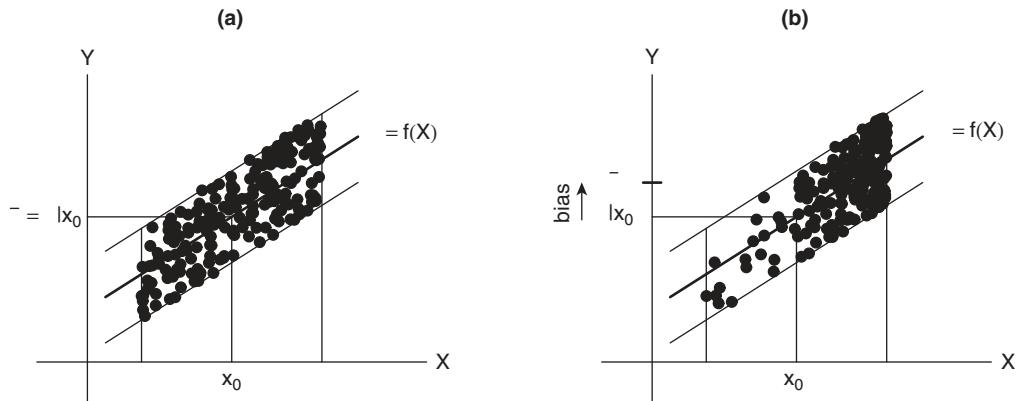


Figure 18.5 (a) When the relationship is linear in the neighborhood of the focal x_0 and the observations are symmetrically distributed around x_0 , both the kernel estimator (which estimates $\bar{\mu}$) and the local-linear estimator (which, because the relationship is linear in the window, directly estimates $\mu|x_0$) are unbiased. (b) When the regression is linear in the neighborhood, but the observations are *not* symmetrically distributed around x_0 , the local-linear estimator is *still* unbiased, but the kernel estimator is biased.

Statistical Issues in Local Regression*

I will again assume local-linear regression. The results in this section extend to local-polynomial fits of higher degree, but the linear case is simpler.

Figure 18.5 demonstrates why the locally linear estimator has a bias advantage in comparison to the kernel estimator. In both panels (a) and (b), the true regression function (given by the heavy line) is linear in the neighborhood of the focal value x_0 .

- In panel (a), the X -values in the window are symmetrically distributed around the focal x_0 at the center of the window. As a consequence, the weighted average $\bar{\mu}$ of the Y s in the window (or, indeed, the simple average of the Y s in the window) provides an unbiased estimate of $\mu|x_0 \equiv E(Y|x_0)$; the local regression line *also* provides an unbiased estimate of $\mu|x_0$ because it estimates the true local regression function.
- In panel (b), in contrast, there are relatively more observations at the right of the window. Because the true regression function has a positive slope in the window, $\bar{\mu}$ exceeds $\mu|x_0$ —that is, the kernel estimator is biased. The local-linear regression, however, *still* estimates the true regression function and therefore provides an unbiased estimate of $\mu|x_0$. The boundaries are regions in which the observations are asymmetrically distributed around the focal x_0 , accounting for the boundary bias of the kernel estimator, but the point is more general.

Of course, if the true regression in the window is *nonlinear*, then both the kernel estimate and the locally linear estimate will usually be biased, if to varying degrees.¹¹ The conclusion to be drawn from these pictures is that *the bias of the kernel estimate depends on the distribution of X -values, while the bias of the locally linear estimate does not*. Because the locally linear

¹¹It is possible that $\bar{\mu} = \mu|x_0$ by good fortune, but this is an unusual occurrence.

estimate can adapt to a “tilt” in the true regression function, it generally has smaller bias when the X -values are unevenly distributed and at the boundaries of the data. Because the kernel and locally linear estimators have the same asymptotic variance, the smaller bias of the locally linear estimator translates into smaller mean-squared error.

These conclusions generalize to local-polynomial regressions of even degree p and odd degree $p + 1$ (e.g., $p = 2$ and $p + 1 = 3$): Asymptotically, the bias of the odd member of the pair is independent of the distribution of X -values, while the bias of the even member is not. The bias of the odd member of the pair is generally smaller than that of the even member, while the variance is the same. Asymptotically, therefore, the odd member of the pair (e.g., the local cubic estimator) has a smaller mean-squared error than the even member (e.g., the local quadratic estimator).

A Closer Look at the Bandwidth of the Local-Regression Smoother*

As the bandwidth h of the local-regression estimator decreases, the bias of the estimator decreases and its sampling variance increases. Suppose that we evaluate the local regression at the focal value x_0 :

- At one extreme, $h = 0$ and only observations with X -values *exactly equal to* x_0 contribute to the local fit. In this case, it is not possible to fit a unique local regression line, but we could still find the fitted value at x_0 as the average Y value for $X = x_0$; if there are no tied values of x_0 , then the fit is exact, $\hat{Y}_0 = Y_0$, and the local-regression estimator simply joins the points in the scatterplot. Because $E(Y|x_0) = \mu|x_0$, the bias of the estimator is 0; its variance—equal to the conditional variance σ_ε^2 of an individual observation—is large, however.
- At the other extreme, $h = \infty$. Then, the scaled distances of explanatory-variable values x_i from the focal x_0 , that is, $x_i = (x_i - x_0)/h$, are all 0, and the weights $w_i = K(z_i)$ are all equal to the maximum (e.g., 1 for the tricube kernel function). With equal weights for all the observations, the fit is no longer local. In effect, we fit a global least-squares line to the data. Now the bias is large (unless, of course, the true regression *really is* globally linear), but the sample-to-sample variance of the fit is small.

The bottom line is the mean-squared error of the estimator,

$$\begin{aligned} \text{MSE}(\hat{Y}|x_0) &\equiv E[(\hat{Y}|x_0 - \mu|x_0)^2] \\ &= E\left\{\left[\hat{Y}|x_0 - E(\hat{Y}|x_0)\right]^2\right\} + \left[E(\hat{Y}|x_0) - \mu|x_0\right]^2 \end{aligned}$$

which is the sum of variance and squared bias. We seek the bandwidth h^* at x_0 that minimizes the mean-squared error (MSE), providing an optimal trade-off of bias against variance (see below). Of course, we need to repeat this process at each focal value of X for which $f(x) = \mu|x$ is to be estimated, adjusting the bandwidth as necessary to minimize MSE.

The expectation and variance of the local-linear smoother at the focal value x_0 are

$$\begin{aligned} E(\hat{Y}|x_0) &\approx f(x_0) + \frac{h^2}{2}s_K^2f''(x_0) \\ V(\hat{Y}|x_0) &\approx \frac{\sigma_\varepsilon^2a_K^2}{nhp_X(x_0)} \end{aligned} \tag{18.1}$$

where (as before)

- $\hat{Y}|x_0 \equiv \hat{f}(x_0)$ is the fitted value at $X = x_0$;
- $\sigma_\varepsilon^2 = V(\varepsilon)$ is the variance of the errors, that is, the conditional (constant) variance of Y around the true regression function;
- h is the bandwidth;
- n is the sample size;

and

- $f''(x_0)$ is the second derivative of the true regression function at the focal x_0 (indicative of the curvature of the regression function, that is, the rapidity with which the slope of the regression function is changing at x_0);
- $p_X(x_0)$ is the probability density for the distribution of X at x_0 (large values of which, therefore, indicate an x_0 near which many observations will be made);¹²
- s_K^2 and a_K^2 are positive constants that depend on the kernel function.¹³

The bias at x_0 is

$$\text{bias}(\hat{Y}|x_0) \equiv E(\hat{Y}|x_0) - f(x_0) \approx \frac{h^2}{2} s_K^2 f''(x_0)$$

The bias of the estimator is large, therefore, when the bandwidth h and curvature $f''(x_0)$ of the regression function are large. In contrast, the variance of the estimator (from Equations 18.1) is large when the error variance σ_ε^2 is large, when the sample size n is small, when the bandwidth h is small, and where data are sparse [i.e., $p_X(x_0)$ is small].¹⁴

Because making h larger increases the bias but decreases the variance, bias and variance, as usual, work at cross-purposes. The value of h that minimizes the MSE—the sum of squared bias and variance—at x_0 is

$$h^*(x_0) = \left[\frac{a_K^2}{s_K^4} \times \frac{\sigma_\varepsilon^2}{np_X(x_0)[f''(x_0)]^2} \right]^{\frac{1}{5}} \quad (18.2)$$

Note that where the curvature $f''(x_0)$ is 0, the optimal bandwidth $h^*(x_0)$ is infinite, suggesting a globally linear fit to the data.¹⁵ Nearest-neighbor bandwidths, which employ a fixed span,

¹²In contrast to the rest of the presentation, here the explanatory variable X is treated as a random variable.

¹³These formulas are derived in Bowman and Azzalini (1997, pp. 72–73). The two constants are

$$s_K^2 = \int z^2 K(z) dZ$$

$$a_K^2 = \int [K(z)]^2 dZ$$

If the kernel $K(z)$ is a probability density function symmetric around 0, such as the standard-normal distribution, then s_K^2 is the variance of this distribution. For the standard-normal kernel, for example, $s_K^2 = 1$ and $a_K^2 = 0.282$. For the triangle kernel (which is *not* a density function), $s_K^2 = 1/6$ and $a_K^2 = 0.949$.

¹⁴The expected effective sample size contributing to the estimate at $x = x_0$ is proportional to $nhp_X(x_0)$, the denominator of the variance.

¹⁵See Exercise 18.4 for an illustration of these points.

adjust for the factor $np_X(x_0)$ but do not take account of the local curvature of the regression function.

To assess the overall accuracy of the nearest-neighbor local-regression estimator, we need some way of cumulating mean-squared error over observed X -values. One way of doing so is to calculate the *average squared error* (ASE):

$$\text{ASE}(s) = \frac{\sum_{i=1}^n [\hat{Y}_i(s) - \mu_i]^2}{n} \quad (18.3)$$

where $\mu_i \equiv E(Y|x_i)$ is the “true” expected value of the response for the i th observation, and $\hat{Y}_i(s)$ is the i th fitted value for span s . Some points to note are the following:

- The squared error is evaluated at the observed X -values and then averaged over the n observations.
- The ASE is calculated for a *particular* set of data, not as an expectation with respect to repeated sampling.¹⁶
- To calculate the ASE requires knowledge of the true regression function, and the ASE therefore cannot be used in practice to select the span. The cross-validation function, described in the next section, estimates the ASE.

The bias and variance of the local-linear estimator at the focal value x_0 are both a function of the bandwidth h , as well as of properties of the data and the kernel function:

$$\begin{aligned}\text{bias}(\hat{Y}|x_0) &\approx \frac{h^2}{2} s_K^2 f''(x_0) \\ V(\hat{Y}|x_0) &\approx \frac{\sigma_e^2 a_K^2}{nhp_X(x_0)}\end{aligned}$$

where s_K^2 and a_K^2 are constants that depend on the kernel function, $f''(x_0)$ is the second derivative (“curvature”) of the regression function at x_0 , and $p_X(x_0)$ is the probability-density of X -values at x_0 . We would ideally like to choose the value of h at each focal value that minimizes the mean-squared error of estimation—that is, the sum of squared bias and variance.

Selecting the Span by Cross-Validation

A conceptually appealing, but complex, approach to bandwidth selection is formally to estimate the optimal bandwidth h^* . We need to estimate $h^*(x_0)$ for each value x_0 of X at which $\hat{Y}|x$ is to be evaluated or to estimate an optimal average value to be used with the fixed-bandwidth estimator. A similar approach is applicable to the nearest-neighbor local-regression estimator.¹⁷

¹⁶See Exercise 18.5 for an illustration.

¹⁷*The so-called plug-in estimate of h^* proceeds by estimating its components— σ_e^2 , $f''(x_0)$, and $p_X(x_0)$; we need not estimate the other quantities in Equation 18.2 (on page 538), because the sample size n is known, and the constants a_K^2 and s_K^4 can be calculated from the kernel function. Estimating σ_e^2 and $f''(x_0)$ requires a preliminary estimate of the regression function.

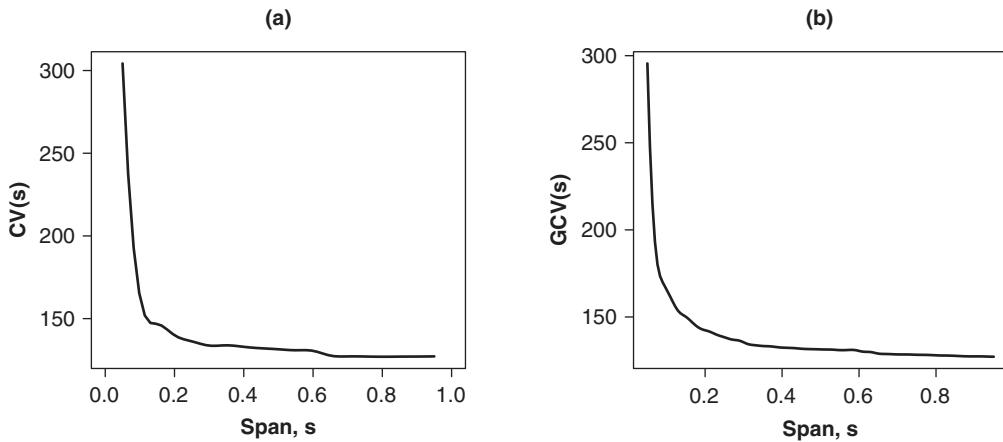


Figure 18.6 (a) Cross-validation function and (b) generalized cross-validation function for the local-linear regression of prestige on income.

A simpler approach, applicable to both the fixed-bandwidth and the nearest-neighbor estimators, is to estimate the optimal bandwidth or span by *cross-validation*.¹⁸ I will consider the nearest-neighbor estimator; the development for the fixed-bandwidth estimator is similar. In cross-validation, we evaluate the regression function at the observations x_i .

The key idea in cross-validation is to *omit* the i th observation from the local regression at the focal value x_i . We denote the resulting estimate of $E(Y|x_i)$ as $\hat{Y}_{-i}|x_i$. Omitting the i th observation makes the fitted value $\hat{Y}_{-i}|x_i$ independent of the observed value Y_i .

The *cross-validation function* is

$$CV(s) \equiv \frac{\sum_{i=1}^n [\hat{Y}_{-i}(s) - Y_i]^2}{n}$$

where $\hat{Y}_{-i}(s)$ is $\hat{Y}_{-i}|x_i$ for span s . The object is to find the value of s that minimizes $CV(s)$. In practice, we can compute $CV(s)$ for a range of values of s .

Figure 18.6(a) shows $CV(s)$ for the regression of occupational prestige on income. In this case, the cross-validation function provides little specific help in selecting the span, suggesting simply that s should be relatively large. Compare this with the value $s \approx .6$ that we arrived at by visual trial and error.

The cross-validation function $CV(s)$ can be costly to compute because it requires refitting the model n times for each candidate value of the span s .¹⁹ For this reason, approximations have been proposed, one of which is termed *generalized cross-validation*, abbreviated *GCV* (Wahba, 1985).²⁰ In the present context, the GCV criterion is

¹⁸The more general use of cross-validation for model selection is discussed in Chapter 22.

¹⁹In the current context, we typically want to evaluate the local regression at each observation anyway, but computational short-cuts (such as interpolation for closely spaced values of X) make the point valid. Moreover, the computational burden imposed by cross-validation extends to other contexts as well.

²⁰The GCV criterion also exhibits a desirable invariance property that the CV criterion does not share. See, for example, Wood (2006, Section 4.5.3).

$$\text{GCV}(s) \equiv \frac{n \times \text{RSS}(s)}{[df_{\text{res}}(s)]^2} \quad (18.4)$$

where $\text{RSS}(s)$ is the residual sum of squares and $df_{\text{res}}(s)$ the “equivalent” residual degrees of freedom for the local-regression model with span s .²¹ The GCV function for the example appears in Figure 18.6(b). In this case, $\text{GCV}(s)$ provides an excellent approximation to $\text{CV}(s)$.

The cross-validation function

$$\text{CV}(s) \equiv \frac{\sum_{i=1}^n [\hat{Y}_{-i}(s) - Y_i]^2}{n}$$

can be used to select the span s in local-polynomial regression, picking s to minimize $\text{CV}(s)$. The fitted value at each observation $\hat{Y}_{-i}(s)$ is computed from a local regression that omits that observation. Because the cross-validation function $\text{CV}(s)$ can be costly to compute, approximations such as generalized cross-validation have been proposed. The GCV criterion is

$$\text{GCV}(s) = \frac{n \times \text{RSS}(s)}{[df_{\text{res}}(s)]^2}$$

where $\text{RSS}(s)$ is the residual sum of squares and $df_{\text{res}}(s)$ the “equivalent” residual degrees of freedom for the local-regression smoother with span s .

A Closer Look at Cross-Validation* The cross-validation function is a kind of estimate of the mean (i.e., expected) ASE at the observed X s,²²

$$\text{MASE}(s) \equiv E \left\{ \frac{\sum_{i=1}^n [\hat{Y}_i(s) - \mu_i]^2}{n} \right\}$$

Because of the independence of \hat{Y}_{-i} and Y_i , the expectation of $\text{CV}(s)$ is

$$\begin{aligned} E[\text{CV}(s)] &= \frac{\sum_{i=1}^n E[\hat{Y}_{-i}(s) - Y_i]^2}{n} \\ &\approx \text{MASE}(s) + \sigma_\varepsilon^2 \end{aligned}$$

²¹See below (page 546) for an explanation of degrees of freedom in local-polynomial regression.

²²Alternatively, rather than averaging over the observed X -values, we could integrate over the probability-density of X , producing the mean integrated square error (MISE):

$$\text{MISE}(s) = \int \{E[\hat{Y}|x(s)] - \mu|x\}\rho(x)dX$$

We can think of MASE as a discrete version of MISE.

The substitution of Y_i for μ_i increases the expectation of $\text{CV}(s)$ by σ_ε^2 , but because σ_ε^2 is a constant, the value of s that minimizes $E[\text{CV}(s)]$ is (approximately) the value that minimizes $\text{MASE}(s)$.

To understand why it is important in this context to omit the i th observation in calculating the fit at the i th observation, consider what would happen were we not to do this. Then, setting the span to 0 would minimize the estimated MASE, because (in the absence of tied X -values) the local-regression estimator simply interpolates the observed data: The fitted and observed values are equal, and $\widehat{\text{MASE}}(0) = 0$.

Although cross-validation is often a useful method for selecting the span in local-polynomial regression, it should be appreciated that $\text{CV}(s)$ is only an estimate and is therefore subject to sampling variation. Particularly in small samples, this variability can be substantial. Moreover, the approximations to the expectation and variance of the local-regression estimator in Equation 18.1 (page 537) are asymptotic, and in small samples, $\text{CV}(s)$ often tends to provide values of s that are too small.

Statistical Inference for Local-Polynomial Regression

In parametric regression—for example, linear least-squares regression—the central objects of estimation are the regression coefficients. Statistical inference naturally focuses on these coefficients, typically taking the form of confidence intervals or hypothesis tests.²³ In nonparametric regression, in contrast, there are *no* regression coefficients. Instead, the central object of estimation is the regression function, and inference focuses on the regression function *directly*.

Many applications of simple nonparametric regression have as their goal visual smoothing of a scatterplot. In these instances, statistical inference is at best of secondary interest. Inference becomes more prominent in nonparametric multiple regression.²⁴

The current section takes up several aspects of statistical inference for local-polynomial regression with one explanatory variable. I start by explaining how to construct an approximate confidence envelope for the regression function. Then, I present a simple approach to hypothesis testing, based on an analogy to procedures for testing hypotheses in linear least-squares regression. The statistical theory behind these relatively simple methods is subsequently examined.

A general caveat concerns the selection of the span s : Because s is typically selected on examination of the data—either visually or by employing a criterion such as CV or GCV—the validity of classical statistical inference is compromised. The methods of this section are therefore best regarded as rough guides.²⁵

Confidence Envelopes Consider the local-polynomial estimate $\widehat{f}(x) = \widehat{Y}|x$ of the regression function $f(x)$. For notational convenience, I assume that the regression function is evaluated at the observed X -values, x_1, x_2, \dots, x_n , although the line of reasoning to be developed here is more general.

²³See Chapter 6.

²⁴See Section 18.2.

²⁵Issues of model selection are addressed more generally in Chapter 22.

The fitted value $\hat{Y}_i = \hat{Y}|x_i$ results from a locally weighted least-squares regression of Y on X . This fitted value is therefore a weighted sum of the observations:²⁶

$$\hat{Y}_i = \sum_{j=1}^n s_{ij} Y_j \quad (18.5)$$

where the weights s_{ij} are functions of the X -values. For the tricube weight function, for example, s_{ij} is 0 for any observations outside the neighborhood of the focal x_i . Because (by assumption) the Y 's are independently distributed, with common conditional variance $V(Y|X = x_i) = V(Y_i) = \sigma_\varepsilon^2$, the sampling variance of the fitted value \hat{Y}_i is

$$V(\hat{Y}_i) = \sigma_\varepsilon^2 \sum_{j=1}^n s_{ij}^2$$

To apply this result, we require an estimate of σ_ε^2 . In linear least-squares simple regression, we estimate the error variance as

$$S_E^2 = \frac{\sum E_i^2}{n - 2}$$

where $E_i = Y_i - \hat{Y}_i$ is the residual for observation i , and $n - 2$ is the degrees of freedom associated with the residual sum of squares. We “lose” 2 degrees of freedom as a consequence of estimating two regression parameters—the intercept α and the slope β .²⁷

We can calculate residuals in nonparametric regression in the same manner—that is, $E_i = Y_i - \hat{Y}_i$, where, of course, the fitted value \hat{Y}_i is from the nonparametric regression. To complete the analogy, we require the *equivalent number of parameters* or *equivalent degrees of freedom* for the model, df_{mod} , from which we can obtain the equivalent residual degrees of freedom, $df_{\text{res}} = n - df_{\text{mod}}$. Then, the estimated error variance is

$$S_E^2 = \frac{\sum E_i^2}{df_{\text{res}}}$$

and the estimated variance of the fitted value \hat{Y}_i is

$$\hat{V}(\hat{Y}_i) = S_E^2 \sum_{j=1}^n s_{ij}^2 \quad (18.6)$$

Assuming normally distributed errors, or a sufficiently large sample, a 95% confidence interval for $E(Y|x_i) = f(x_i)$ is approximately

$$\hat{Y}_i \pm 2 \sqrt{\hat{V}(\hat{Y}_i)} \quad (18.7)$$

Putting the confidence intervals together for $X = x_1, x_2, \dots, x_n$ produces a pointwise 95% confidence band or confidence envelope for the regression function.

An example, employing the local-linear regression of prestige on income in the Canadian occupational prestige data (with span $s = 0.6$), appears in Figure 18.7. Here, $df_{\text{mod}} = 5.006$, and $S_E^2 = 12,004.72/(102 - 5.006) = 123.77$. The nonparametric-regression smooth therefore

²⁶See the starred material in this section for this and other results.

²⁷See Section 18.2.

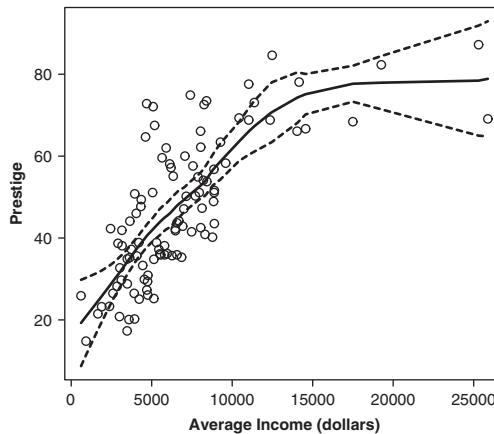


Figure 18.7 Local-linear regression of occupational prestige on income, showing an approximate pointwise 95% confidence envelope. The span of the smoother is $s = 0.6$.

uses the equivalent of about five parameters—roughly the same as a fourth-degree polynomial. The fit to the data, however, can differ substantially from that of fourth-degree polynomial, which is much less sensitive to local characteristics of the regression function.²⁸

Although this procedure for constructing a confidence band has the virtue of simplicity, it is not quite correct, due to the bias in $\hat{Y}|x$ as an estimate of $E(Y|x)$. If we have chosen the span and degree of the local-polynomial estimator judiciously, however, the bias should be small. Bias in $\hat{Y}|x$ has the following consequences:

- S_E^2 is biased upward, tending to overstate the error variance and making the confidence interval too wide.²⁹
- The confidence interval is on average centered in the wrong location.

These errors tend to offset each other. Because $\hat{Y}|x$ is biased, it is more accurate to describe the envelope around the sample regression constructed according to Equation 18.7 as a “variability band” rather than as a confidence band.³⁰

The fitted values in a local-polynomial regression are linear functions of the observations, $\hat{Y}_i = \sum_{j=1}^n s_{ij} Y_j$. Estimating the error variance as $S_E^2 = \sum E_i^2 / df_{\text{res}}$, where df_{res} is the equivalent residual degrees of freedom for the model, the estimated variance of a fitted value is $\hat{V}(\hat{Y}_i) = S_E^2 \sum_{j=1}^n s_{ij}^2$. An approximate 95% pointwise confidence band around the regression curve evaluated at the fitted values may be formed as $\hat{Y}_i \pm 2\sqrt{\hat{V}(\hat{Y}_i)}$.

²⁸See Exercise 18.7.

²⁹Bowman and Azzalini (1997, Section 4.3) consider alternative approaches to estimating the error variance σ_ε^2 .

³⁰We could, moreover, make the same point about confidence intervals for fitted values in *linear* least-squares regression, when the assumption of linearity is not exactly correct. Indeed, the bias in estimates of $E(Y|x)$ is likely to be *less* in nonparametric regression than in linear regression.

Hypothesis Tests In linear least-squares regression, F -tests of hypotheses are formulated by comparing alternative nested models. To say that two models are nested means that one, the more specific model, is a special case of the other, more general model.³¹ For example, in linear least-squares simple regression, the F -statistic

$$F_0 = \frac{\frac{\text{TSS} - \text{RSS}}{\text{RSS}}}{\frac{n-2}{n-2}}$$

with 1 and $n-2$ degrees of freedom tests the hypothesis of no linear relationship between Y and X . Here, the total sum of squares, $\text{TSS} = \sum (Y_i - \bar{Y})^2$, is the variation in Y associated with the null model of no relationship, $Y_i = \alpha + \varepsilon_i$, and the residual sum of squares, $\text{RSS} = \sum (Y_i - \hat{Y}_i)^2$, represents the variation in Y conditional on the linear relationship between Y and X , based on residuals from the model $Y_i = \alpha + \beta x_i + \varepsilon_i$. Because the null model is a special case of the linear model, with $\beta = 0$, the two models are nested.

An analogous, but more general, F -test of no relationship for the nonparametric-regression model is

$$F_0 = \frac{\frac{\text{TSS} - \text{RSS}}{\text{RSS}}}{\frac{df_{\text{mod}} - 1}{df_{\text{res}}}} \quad (18.8)$$

with $df_{\text{mod}} - 1$ and $df_{\text{res}} = n - df_{\text{mod}}$ degrees of freedom. Here RSS is the residual sum of squares for the nonparametric-regression model. Applied to the local-linear regression of prestige on income, where $n = 102$, $\text{TSS} = 29,895.43$, $\text{RSS} = 12,004.72$, and $df_{\text{mod}} = 5.006$, we have

$$F_0 = \frac{\frac{29,895.43 - 12,004.72}{12,004.72}}{\frac{5.006 - 1}{102 - 5.006}} = 36.08$$

with $5.006 - 1 = 4.006$ and $102 - 5.006 = 96.994$ degrees of freedom. The resulting p -value is much smaller than .0001, casting strong doubt on the null hypothesis of no relationship between prestige and income of occupations.

A test of nonlinearity is simply constructed by contrasting the nonparametric-regression model with the linear simple-regression model.³² The models are properly nested because a linear relationship is a special case of a general, potentially nonlinear, relationship. Denoting the residual sum of squares from the linear model as RSS_0 and the residual sum of squares from the more general nonparametric-regression model as RSS_1 , we have

$$F_0 = \frac{\frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1}}{\frac{df_{\text{mod}} - 2}{df_{\text{res}}}}$$

³¹See Section 6.2.2.

³²Cf. the test for “lack of fit” described in Section 12.4.1.

with $df_{\text{mod}} - 2$ and $df_{\text{res}} = n - df_{\text{mod}}$ degrees of freedom. This test is constructed according to the rule that the most general model—here the nonparametric-regression model—is employed for estimating the error variance in the denominator of the F -statistic, $S_E^2 = \text{RSS}_1 / df_{\text{res}}$. For the regression of occupational prestige on income, $\text{RSS}_0 = 14,616.17$, $\text{RSS}_1 = 12,004.72$, and $df_{\text{mod}} = 5.006$; thus,

$$F_0 = \frac{\frac{14,616.17 - 12,004.72}{5.006 - 2}}{\frac{12,004.72}{102 - 5.006}} = 7.02$$

with $5.006 - 2 = 3.006$ and $102 - 5.006 = 96.994$ degrees of freedom. The corresponding p -value, .0003, suggests that the relationship between the two variables is significantly nonlinear.

Approximate incremental F -tests for hypotheses in local-polynomial regression are formulated by contrasting nested models, in analogy to similar tests for linear models fit by least squares. For example, to test the hypothesis of no relationship in the nonparametric-regression model, we can compute the F -test statistic

$$F_0 = \frac{\frac{\text{TSS} - \text{RSS}}{df_{\text{mod}} - 1}}{\frac{\text{RSS}}{df_{\text{res}}}}$$

where df_{mod} and $df_{\text{res}} = n - df_{\text{mod}}$ are respectively the equivalent degrees of freedom for the regression model and for error, and RSS is the residual sum of squares for the model.

Similarly, to test for nonlinearity, we can contrast the fitted nonparametric-regression model with a linear model, computing

$$F_0 = \frac{\frac{\text{RSS}_0 - \text{RSS}_1}{df_{\text{mod}} - 2}}{\frac{\text{RSS}_1}{df_{\text{res}}}}$$

where RSS_0 is the residual sum of squares for the linear regression and RSS_1 the residual sum of squares for the more general nonparametric regression.

Degrees of Freedom* As noted, the fitted values \hat{Y}_i in local-polynomial regression are weighted sums of the observed Y values (repeating Equation 18.5 on page 543):

$$\hat{Y}_i = \sum_{j=1}^n s_{ij} Y_j$$

Let us collect the weights s_{ij} into the *smoother matrix*

$$\mathbf{S}_{(n \times n)} \equiv \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1i} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2i} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & s_{i2} & \cdots & s_{ii} & \cdots & s_{in} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{ni} & \cdots & s_{nn} \end{bmatrix}$$

Then,

$$\hat{\mathbf{y}}_{(n \times 1)} = \mathbf{S}_{(n \times n)} \mathbf{y}_{(n \times 1)}$$

where $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n]'$ is the vector of fitted values, and $\mathbf{y} = [y_1, y_2, \dots, y_n]'$ is the vector of observed response values.

The covariance matrix of the fitted values is

$$V(\hat{\mathbf{y}}) = \mathbf{S}V(\mathbf{y})\mathbf{S}' = \sigma_\varepsilon^2 \mathbf{S}\mathbf{S}' \quad (18.9)$$

This result follows from the assumptions that the conditional variance of Y_i is constant (σ_ε^2) and that the observations are independent, implying that $V(\mathbf{y}) = \sigma_\varepsilon^2 \mathbf{I}_n$. Equation 18.6 (page 543) for the variance of \hat{Y}_i is just an expansion of the i th diagonal entry of $V(\hat{\mathbf{y}})$.

The smoother matrix \mathbf{S} is analogous to the hat-matrix $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in linear least-squares regression, where \mathbf{X} is the model matrix for the linear model.³³ The residuals in linear least-squares regression are

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

The corresponding expression in local regression is

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{S})\mathbf{y}$$

To determine the smoother matrix \mathbf{S} , recall that \hat{Y}_i results from a locally weighted polynomial regression of Y on X :

$$Y_j = A_i + B_{1i}(x_j - x_i) + B_{2i}(x_j - x_i)^2 + \cdots + B_{pi}(x_j - x_i)^p + E_{ji}$$

where the weights $w_{ji} = K[(x_j - x_i)/h]$ decline with distance from the focal x_i . The local-regression coefficients are chosen to minimize $\sum_{j=1}^n w_{ji}E_{ji}^2$. The fitted value \hat{Y}_i is just the regression constant A_i . In matrix form, the local regression is

$$\mathbf{y} = \mathbf{X}_i \mathbf{b}_i + \mathbf{e}_i$$

The model matrix \mathbf{X}_i contains the regressors in the local-regression equation (including an initial column of 1s for the constant), and the coefficient vector \mathbf{b}_i contains the regression coefficients.

Define the diagonal matrix $\mathbf{W}_i \equiv \text{diag}\{\sqrt{w_{ji}}\}$ of square-root kernel weights. Then, the local-regression coefficients are

³³See Section 11.8.

$$\mathbf{b}_i = (\mathbf{X}'_i \mathbf{W}_i \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{W}_i \mathbf{y}$$

and the i th row of the smoother matrix is the first row of $(\mathbf{X}'_i \mathbf{W}_i \mathbf{X}_i)^{-1} \mathbf{X}'_i \mathbf{W}_i$ (i.e., the row that determines the constant, $A_i = \hat{Y}_i$). To construct \mathbf{S} , we need to repeat this procedure for $i = 1, 2, \dots, n$.

In linear least-squares regression, the degrees of freedom for the model can be defined in a variety of equivalent ways. Most directly, assuming that the model matrix \mathbf{X} is of full column rank, the degrees of freedom for the model are equal to the number of regressors $k + 1$ (including the regression intercept). The degrees of freedom for the model are also equal to the following:

- the rank and trace of the hat matrix, \mathbf{H} ,
- the trace of \mathbf{HH}' , and
- the trace of $2\mathbf{H} - \mathbf{HH}'$.

These alternative expressions follow from the fact that the hat-matrix is symmetric and idempotent—that is, $\mathbf{H} = \mathbf{H}'$ and $\mathbf{H} = \mathbf{HH}$. The degrees of freedom for error in least-squares linear regression are

$$df_{\text{res}} = \text{rank}(\mathbf{I}_n - \mathbf{H}) = \text{trace}(\mathbf{I}_n - \mathbf{H}) = n - \text{trace}(\mathbf{H}) = n - k - 1$$

because $\mathbf{I}_n - \mathbf{H}$ projects \mathbf{y} onto the orthogonal complement of the column space of \mathbf{X} to obtain the residuals: $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$.³⁴

Analogous degrees of freedom for the local-regression model are obtained by substituting the smoother matrix \mathbf{S} for the hat-matrix \mathbf{H} . The analogy is not perfect, however, and in general $\text{trace}(\mathbf{S}) \neq \text{trace}(\mathbf{SS}') \neq \text{trace}(2\mathbf{S} - \mathbf{SS}')$

- Defining $df_{\text{mod}} = \text{trace}(\mathbf{S})$ is an attractive choice because it is easy to calculate.
- In a linear model, the degrees of freedom for the model are equal to the sum of variances of the fitted values divided by the error variance,

$$\frac{\sum_{i=1}^n V(\hat{Y}_i)}{\sigma_e^2} = k + 1$$

In the current context (from Equation 18.9),

$$\frac{\sum_{i=1}^n V(\hat{Y}_i)}{\sigma_e^2} = \text{trace}(\mathbf{SS}')$$

motivating the definition, $df_{\text{mod}} = \text{trace}(\mathbf{SS}')$.

- The expectation of the residual sum of squares in local-polynomial regression is³⁵

$$E(\text{RSS}) = \sigma_e^2[n - \text{trace}(2\mathbf{S} - \mathbf{SS}')] + \text{bias}^2$$

where $\text{bias}^2 = \sum_{i=1}^n [E(\hat{Y}_i) - f(x_i)]^2$ is the cumulative bias in the local regression evaluated at the observed X -values. If the bias is negligible, then $\text{RSS}/[n - \text{trace}(2\mathbf{S} - \mathbf{SS}')] = \sigma_e^2$ is an estimator of the error variance σ_e^2 , suggesting that $n - \text{trace}(2\mathbf{S} - \mathbf{SS}')$ is a suitable definition of the degrees of freedom for error and that $df_{\text{mod}} = \text{trace}(2\mathbf{S} - \mathbf{SS}')$. This last

³⁴The vector geometry of linear least-squares regression is developed in Chapter 10.

³⁵See Hastie and Tibshirani (1990, Sections 3.4–3.5).

definition is possibly the most attractive theoretically, but it is relatively difficult to compute.³⁶

The smoother matrix \mathbf{S} in nonparametric local-polynomial regression plays a role analogous to the hat-matrix \mathbf{H} in linear least-squares regression. Like the hat-matrix, the smoother matrix linearly transforms the observations into the fitted values: $\hat{\mathbf{y}} = \mathbf{Sy}$. Pursuing this analogy, the equivalent degrees of freedom for the nonparametric-regression model can variously be defined as $df_{\text{mod}} = \text{trace}(\mathbf{S})$, $\text{trace}(\mathbf{SS'})$, or $\text{trace}(2\mathbf{S} - \mathbf{SS'})$.

18.1.3 Smoothing Splines*

In contrast with regression splines—which are *parametric* regression models³⁷—*smoothing splines* arise as the solution to the following *nonparametric*-regression problem: Find the function $\hat{f}(x)$ with two continuous derivatives that minimizes the *penalized sum of squares*,

$$\text{SS}^*(h) = \sum_{i=1}^n [Y_i - f(x_i)]^2 + h \int_{x_{\min}}^{x_{\max}} [f''(x)]^2 dx \quad (18.10)$$

where h is a smoothing constant, analogous to the bandwidth of a kernel or local-polynomial estimator.

- The first term in Equation 18.10 is the residual sum of squares.
- The second term is a *roughness penalty*, which is large when the integrated second derivative of the regression function $f''(x)$ is large—that is, when $f(x)$ is rough. The endpoints of the integral enclose the data: $x_{\min} < x_{(1)}$ and $x_{\max} > x_{(n)}$.
- At one extreme, if the smoothing constant is set at $h = 0$ (and if all the X -values are distinct), then $\hat{f}(x)$ simply interpolates the data.
- At the other extreme, if h is very large, then \hat{f} will be selected so that $\hat{f}''(x)$ is everywhere 0, which implies a *globally linear* least-squares fit to the data.

It turns out, surprisingly and elegantly, that the function $\hat{f}(x)$ that minimizes Equation 18.10 is a natural cubic spline with knots at the distinct observed values of X . Although this result seems to imply that n parameters are required (when all X -values are distinct), the roughness penalty imposes additional constraints on the solution, typically reducing the equivalent number of parameters for the smoothing spline considerably and preventing $\hat{f}(x)$ from interpolating

³⁶Hastie and Tibshirani (1990, Section 3.5) demonstrate a simple relationship between $\text{trace}(2\mathbf{S} - \mathbf{SS'})$ and $\text{trace}(\mathbf{S})$ that allows the latter to be used to approximate the former. The software used for most of the examples in the current chapter (the **gam** package for R: Hastie & Tibshirani, 1990; Hastie, 1992) takes this approach. Further discussion of these issues may be found in Hastie and Tibshirani (1990, Section 3.5) and in Cleveland, Grosse, and Shyu (1992, Section 8.4.1). Hastie and Tibshirani (1990, Sections 3.8– 3.9) show how incremental F -tests can be made more precise by adjusting the degrees of freedom used in finding p -values. Similar procedures can be applied to improve the performance of confidence bands for the regression curve, using the t -distribution in the calculation of margins of error.

³⁷See Section 17.2.

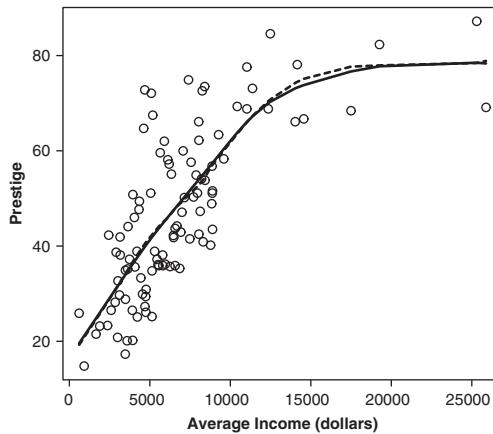


Figure 18.8 Nonparametric regression of occupational prestige on income, using a smoothing spline (solid line) and local-linear regression (broken line), both with five equivalent parameters.

the data. Indeed, it is common practice to select the smoothing constant h *indirectly* by setting the equivalent number of parameters for the smoother.

An illustration, for the regression of occupational prestige on income, appears in Figure 18.8, comparing a smoothing spline with a local-linear fit employing the same equivalent number of parameters (degrees of freedom). The two fits are nearly identical.

Precisely the same considerations arise in the selection of h for smoothing splines as in the selection of the bandwidth or span for local-polynomial smoothers: We can proceed, for example, by visual trial and error or by cross-validation or generalized cross-validation.

Smoothing splines offer certain small advantages in comparison to local-polynomial smoothers. Both smoothers are linear, in the sense that they can be written in the form $\hat{y} = \mathbf{S}y$ for a suitably defined smoother matrix \mathbf{S} . The smoother matrix for smoothing splines is slightly better behaved, however, and if smoothing splines are employed as building blocks of an additive regression model, then the backfitting algorithm that can be used to fit this model is guaranteed to converge, a property that does not hold for the local-polynomial smoother.³⁸ On the negative side, smoothing splines are more difficult to generalize to multiple regression.³⁹

18.2 Nonparametric Multiple Regression

18.2.1 Local-Polynomial Multiple Regression

Local-polynomial regression extends straightforwardly from simple to multiple regression. The method also has an intuitively appealing rationale, and it is relatively simple to implement.

³⁸Additive regression models and the backfitting algorithm are described in Section 18.2.2.

³⁹This is not to say that spline-based methods for multiple regression are either impossible or unattractive, just that their development is relatively complex. See, for example, Green and Silverman (1994) or Wood (2006).

Moreover, local-polynomial regression generalizes easily to binary and other non-normal data.⁴⁰

Kernel Weights in Multiple Regression

As a formal matter, it is simple to extend the local-polynomial estimator to several explanatory variables. To obtain a fitted value $\hat{Y}|\mathbf{x}$ at the focal point $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})'$ in the space of the explanatory variables, we perform a weighted-least-squares polynomial regression of Y on the X s, emphasizing observations close to the focal point.⁴¹

- A local-linear fit therefore takes the following form:

$$Y_i = A + B_1(x_{i1} - x_{01}) + B_2(x_{i2} - x_{02}) + \dots + B_k(x_{ik} - x_{0k}) + E_i$$

- For $k = 2$ explanatory variables, a local-quadratic fit takes the form

$$\begin{aligned} Y_i = & A + B_1(x_{i1} - x_{01}) + B_2(x_{i2} - x_{02}) + B_{11}(x_{i1} - x_{01})^2 + B_{22}(x_{i2} - x_{02})^2 \\ & + B_{12}(x_{i1} - x_{01})(x_{i2} - x_{02}) + E_i \end{aligned}$$

including linear, quadratic, and cross-product terms for the X s. When there are several explanatory variables, the number of terms in the local-quadratic regression grows large. As a consequence, we will not consider cubic or higher-order polynomials, which contain even more terms.

In either the linear or quadratic case, we find local-regression coefficients by minimizing the weighted residual sum of squares $\sum_{i=1}^n w_i E_i^2$ for suitably defined weights w_i . The fitted value at the focal point in the X -space is then the regression constant, $\hat{Y}|\mathbf{x}_0 = A$.

Figure 18.9 shows the scatterplot for two variables, X_1 and X_2 , sampled from a bivariate-normal distribution with means $\mu_1 = \mu_2 = 20$, standard deviations $\sigma_1 = 8$ and $\sigma_2 = 4$, and covariance $\sigma_{12} = 25$ [producing the correlation $\rho_{12} = 25/(8 \times 4) = .78$]. As illustrated in this figure, there are two straightforward ways to extend kernel weighting to local-polynomial multiple regression:

1. Calculate *marginal weights* separately for each X , and then take the product of the marginal weights. That is, for the j th explanatory variable and observation i , calculate the marginal kernel weight

$$w_{ij} = K\left(\frac{x_{ij} - x_{0j}}{h_j}\right)$$

where x_{0j} is the focal value for explanatory variable X_j , and h_j is the marginal bandwidth for this variable. As in local-polynomial simple regression, we can use a fixed bandwidth or we can adjust the bandwidth to include a constant number of nearest neighbors of x_{0j} . Having found marginal weights for the k explanatory variables, the final weight attributed to observation i in the local regression is their product:

⁴⁰See Section 18.3.

⁴¹If you are unfamiliar with vector notation, think of \mathbf{x} simply as the collection of values of the explanatory variables.

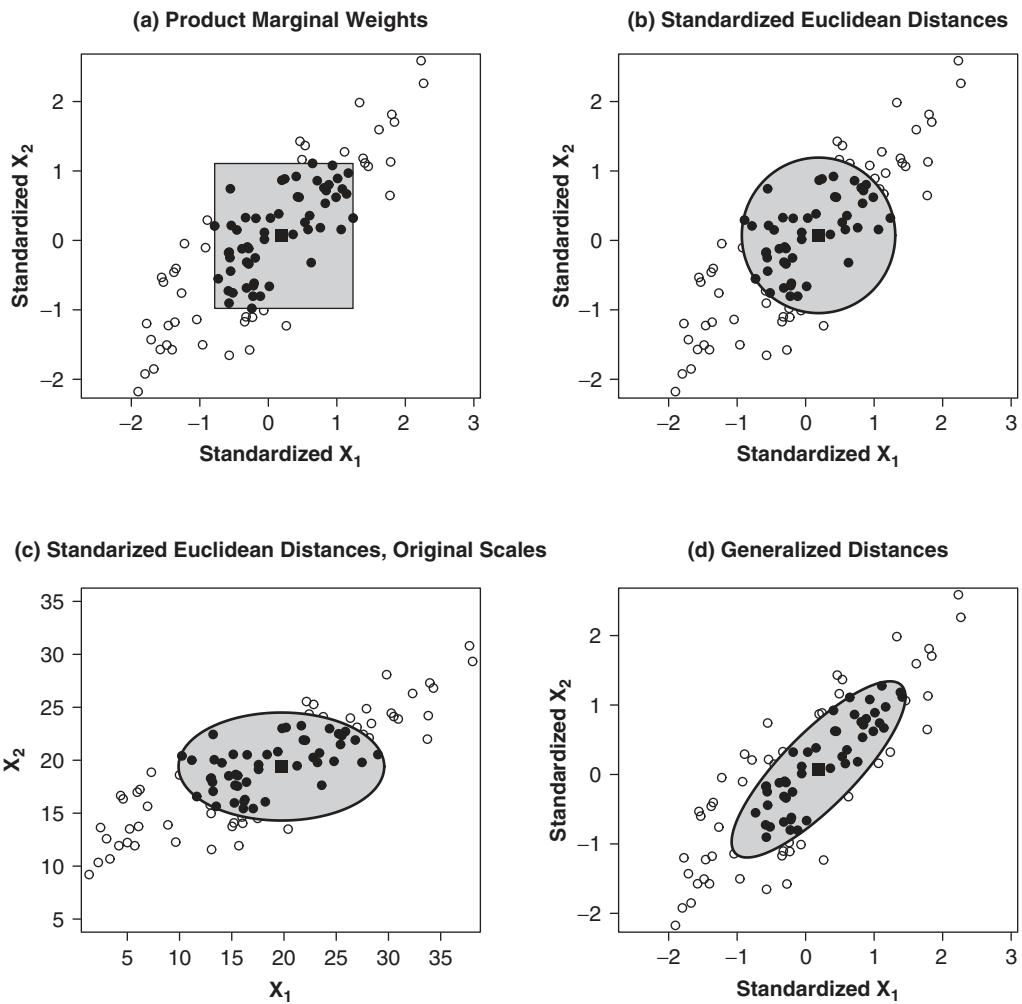


Figure 18.9 Defining neighborhoods for local-polynomial multiple regression: In each case, the focal point is marked by a black square at the center of the bivariate window. (a) Product-marginal weights, each for span = 0.7; (b) standardized Euclidean distances, span = 0.5; (c) standardized Euclidean distances plotted on the *unstandardized* scales for the two variables; (d) generalized distances, span = 0.5.

$$w_i = w_{i1}w_{i2} \cdots w_{ik}$$

Product-marginal weights define a rectangular neighborhood around the focal \mathbf{x}_0 . Figure 18.9(a) shows such a neighborhood for the artificially generated data.⁴²

⁴²The explanatory variables in Figure 18.9(a) are standardized for comparability with the other parts of the figure (see below). Standardization does not affect the product-marginal weights.

2. Measure the distance $D(\mathbf{x}_i, \mathbf{x}_0)$ in the X -space between the explanatory-variable values \mathbf{x}_i for observation i and the focal \mathbf{x}_0 . Then, kernel weights can be calculated directly from these distances,

$$w_i = K \left[\frac{D(\mathbf{x}_i, \mathbf{x}_0)}{h} \right]$$

As before, the bandwidth h can either be fixed or adjusted to include a constant number of nearest neighbors of the focal point. There is, however, more than one way to define distances between points in the X space:

- *Simple Euclidean distance*:

$$D_E(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^k (x_{ij} - x_{0j})^2}$$

Euclidean distances only make sense when the X s are measured in the same units, and even in this case, we may prefer another approach. An obvious application of Euclidean distance is to spatially distributed data, where the two explanatory variables X_1 and X_2 represent coordinates on a map, and the regression surface traces how the average value of Y changes spatially—a “topographical” map where altitude represents the average level of the response.

- *Scaled Euclidean distance*: Scaled distances adjust each X by a measure of dispersion to make values of the explanatory variables roughly comparable. We could use a robust measure of spread, such as the median absolute deviation from the median or the interquartile range, but the standard deviation is typically used. It is also common to center the X s by subtracting the mean from each value; centering does not affect distances, however. The first step, then, is to standardize the X s,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where \bar{x}_j and s_j are respectively the mean and standard deviation of X_j . The scaled Euclidean distance between an observation \mathbf{x}_i and the focal point \mathbf{x}_0 is

$$D_S(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^k (z_{ij} - z_{0j})^2}$$

This is the most common approach to defining distances.

For two X s, scaled Euclidean distances generate a circular neighborhood around the focal point in the standardized X space [see Figure 18.9(b)]. Plotted in the original, *unscaled* X -space, the neighborhood is elliptical, with axes parallel to the X_1 and X_2 axes [Figure 18.9(c)].

- *Generalized distance*: *Generalized distances adjust not only for the dispersion of the X s but also for their correlational structure:

$$D_G(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{(\mathbf{x}_i - \mathbf{x}_0)' \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{x}_0)}$$

where \mathbf{V} is the covariance matrix of the X s, perhaps estimated robustly.⁴³ Figure 18.9(d) illustrates generalized distances for $k = 2$ explanatory variables. Here, the neighborhood around the focal point is elliptical, with axes reflecting the correlation between the X s.

As mentioned, simple Euclidean distances do not make sense unless the explanatory variables are on the same scale. Beyond that point, the choice of product marginal weights, weights based on scaled Euclidean distances, or weights based on generalized distances usually does not make a great deal of difference.

Generalizing local-polynomial regression to multiple regression is conceptually and computationally straightforward. For example, to obtain the fitted value for a local-linear regression at the focal point $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})'$ in the space of the explanatory variables, we perform a weighted-least-squares regression of Y on the X s,

$$Y_i = A + B_1(x_{i1} - x_{01}) + B_2(x_{i2} - x_{02}) + \dots + B_k(x_{ik} - x_{0k}) + E_i$$

emphasizing observations close to the focal point by minimizing the weighted residual sum of squares, $\sum_{i=1}^n w_i E_i^2$. The fitted value at the focal point in the X -space is then $\hat{Y}|\mathbf{x}_0 = A$. The weights w_i can be computed in several ways, including by multiplying marginal kernel weights for the several explanatory variables or by basing kernel weights on one or another measure of distance between the focal \mathbf{x}_0 and the observed X -values, \mathbf{x}_i . Given a distance measure $D(\mathbf{x}_i, \mathbf{x}_0)$, the kernel weights are calculated as $w_i = K[D(\mathbf{x}_i, \mathbf{x}_0)/h]$.

Span Selection, Statistical Inference, and Order Selection

Methods of span selection for local-polynomial multiple regression are essentially the same as the methods for simple regression discussed in Section 18.1.2; they are, briefly:

- *Visual Trial and Error:* We can vary the span and examine the resulting regression surface, balancing smoothness against detail. We seek the smallest span that produces a smooth regression surface.
- *Cross-Validation:* For a given span s , we fit the model omitting each observation in turn, obtaining a fitted value $\hat{Y}_{-i}(s) = \hat{Y}|\mathbf{x}_i$ at the omitted observation. Then, we select the span that minimizes the cross-validation function

$$\text{CV}(s) = \frac{\sum_{i=1}^n [\hat{Y}_{-i}(s) - Y_i]^2}{n}$$

or the generalized cross-validation function

$$\text{GCV}(s) = \frac{n \times \text{RSS}(s)}{[df_{\text{res}}(s)]^2}$$

⁴³Methods such as M estimation, to be introduced in Chapter 19 on robust regression, can be adapted to estimate the mean vector and covariance matrix for a vector of variables.

It is, in addition, possible to derive an expression for the mean-square error of estimation in local-polynomial multiple regression.⁴⁴ One could in principle proceed to estimate the MSE and to select the span that minimizes the estimate. As far as I know, this more complex approach has not been implemented for multiple regression.

Inference for local-polynomial multiple regression also closely parallels local-polynomial simple regression. At each observation \mathbf{x}_i , the fitted value $\hat{Y}_i = \hat{\mathbf{Y}}|\mathbf{x}_i$ results from a weighted-least-squares regression and is therefore a linear function of the response,

$$\hat{Y}_i = \sum_{j=1}^n s_{ij} Y_j$$

- *Degrees of Freedom:* *As in local-polynomial simple regression, equivalent degrees of freedom for the model come from the smoother matrix \mathbf{S} , where

$$\begin{matrix} \hat{\mathbf{y}} \\ (n \times 1) \end{matrix} = \begin{matrix} \mathbf{S} \\ (n \times n) \end{matrix} \begin{matrix} \mathbf{y} \\ (n \times 1) \end{matrix}$$

and are variously defined as $df_{\text{mod}} = \text{trace}(\mathbf{S})$, $\text{trace}(\mathbf{SS}')$, or $\text{trace}(2\mathbf{S} - \mathbf{SS}')$.

- *Error Variance:* The error variance σ_e^2 can be estimated as

$$S_E^2 = \frac{\sum E_i^2}{df_{\text{res}}}$$

where the $E_i = Y_i - \hat{Y}_i$ are the residuals from the model, and $df_{\text{res}} = n - df_{\text{mod}}$.

- *Confidence Intervals:* The estimated variance of the fitted value \hat{Y}_i at \mathbf{x}_i is

$$\hat{V}(\hat{Y}_i) = S_E^2 \sum_{j=1}^n s_{ij}^2$$

Then, an approximate 95% confidence interval for the population regression surface above \mathbf{x}_i is

$$\hat{Y}_i \pm 2\sqrt{\hat{V}(\hat{Y}_i)}$$

- *Hypothesis Tests:* Incremental F -tests can be formulated by fitting alternative models to the data and comparing residual sums of squares and degrees of freedom. For example, to test for the effect of a particular explanatory variable X_j , we can omit the variable from the model, taking care to adjust the span to reflect the reduced dimensionality of the regression problem.⁴⁵ Let RSS_1 represent the residual sum of squares for the full model, which has df_1 equivalent degrees of freedom, and RSS_0 represent the residual sum of squares for the model omitting the j th explanatory variable, which has df_0 degrees of freedom. Then, under the null hypothesis that Y has no partial relationship to X_j ,

⁴⁴See Fan and Gijbels (1996, Section 7.8) and Simonoff (1996, Section 5.7) for the local-linear case.

⁴⁵That is, if the span for the multiple regression is s and there are k explanatory variables, then (by appealing, e.g., to product-marginal weighting) the “span per explanatory variable” is $\sqrt[k]{s}$. Therefore, if one X is dropped from the model, the span should be adjusted to $s^{(k-1)/k}$. For example, for $k = 2$ and $s = 0.25$, on dropping one X from the model, the adjusted span becomes $0.25^{1/2} = 0.5$.

$$F_0 = \frac{\frac{\text{RSS}_0 - \text{RSS}_1}{df_1 - df_0}}{\frac{\text{RSS}_1}{df_{\text{res}}}}$$

follows an approximate F -distribution with $df_1 - df_0$ and $df_{\text{res}} = n - df_1$ degrees of freedom. In general, and as usual, we use the most complete model fit to the data for the error-variance estimate in the denominator of the incremental F -statistic.

As explained previously, because of proliferation of terms, it is typical to consider only local-linear (Order 1) and quadratic (Order 2) regressions. A local-quadratic fit is indicated if the curvature of the regression surface changes too quickly to be captured adequately by the local-linear estimator. To a certain extent, however, the order of the local regressions can be traded off against their span, because a local-linear regression can be made more flexible by reducing the span. To decide between the local-linear and quadratic fits, we can compare them visually, or we can perform an incremental F -test of the hypothesis that the additional terms in the local quadratic model are necessary.

Methods for selecting the span in local-polynomial multiple regression are much the same as in local-polynomial simple regression: We can proceed visually by trial and error or apply a criterion such as $\text{CV}(s)$ or $\text{GCV}(s)$. Similarly, approximate pointwise confidence limits for the fitted regression can be calculated as in local-polynomial simple regression, as can incremental F -tests comparing nested models.

Obstacles to Nonparametric Multiple Regression

Although, as a formal matter, it is therefore simple to extend local-polynomial estimation to multiple regression, there are two flies in the ointment:

1. *The “curse of dimensionality”*:⁴⁶ As the number of explanatory variables increases, the number of points “near” a focal point tends to decline rapidly. To include a fixed number of observations in the local fit as the number of X s grows therefore requires making the neighborhood around the focal point less and less local. A general assumption of local-polynomial regression is that observations close in the X -space to the focal \mathbf{x}_0 are informative about $f(\mathbf{x}_0)$; increasing the size of the neighborhood around the focal point therefore potentially decreases the quality of the estimate of $f(\mathbf{x}_0)$ by inflating the bias of the estimate.

The problem is illustrated in Figure 18.10 for $k = 2$ explanatory variables. This figure represents a “best-case” scenario, where the X s are independent and uniformly distributed. As we have seen, neighborhoods constructed by product-marginal weighting correspond to rectangular (here, square) regions in the graph. Neighborhoods defined by distance from a focal point correspond to circular (more generally, if the distances are scaled, elliptical) regions in the graph. To include half the observations in a square

⁴⁶The curse of dimensionality was introduced in Section 2.2.

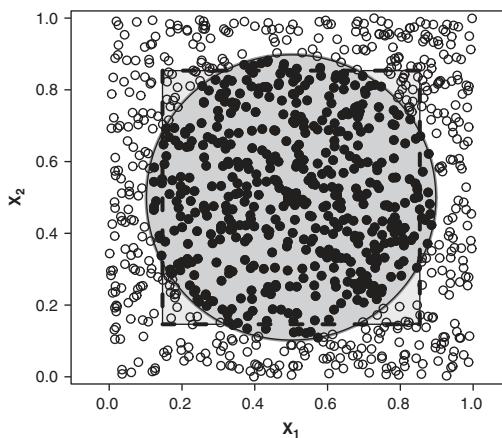


Figure 18.10 The “curse of dimensionality”: 1,000 observations for independent, uniformly distributed random variables X_1 and X_2 . The 500 nearest neighbors of the focal point $\mathbf{x}_0 = (0.5, 0.5)'$ are highlighted, along with the circle (of diameter ≈ 0.8) that encloses them. Also shown is the square centered on \mathbf{x}_0 (with sides $= \sqrt{1/2}$) enclosing about half the data.

neighborhood centered on a focal \mathbf{x}_0 , we need to define marginal neighborhoods for each of X_1 and X_2 that include roughly $\sqrt{1/2} \approx 0.71$ of the data; for $k = 10$ explanatory variables, the marginal neighborhoods corresponding to a hyper-cube that encloses half the observations would each include about $\sqrt[10]{1/2} \approx 0.93$ of the data. A circular neighborhood in two dimensions enclosing half the data has diameter $2\sqrt{0.5/\pi} \approx 0.8$ along each axis; the diameter of the hyper-sphere enclosing half the data also grows with dimensionality, but the formula is too complicated to warrant presentation here.

2. *Difficulties of interpretation:* Because nonparametric regression does not provide an equation relating the average response to the explanatory variables, we must display the response surface graphically. This is no problem, of course, when there is only one X , because the scatterplot relating Y to X is two-dimensional, and the regression “surface” is just a curve. When there are two X s, the scatterplot is three-dimensional, and the regression surface is two-dimensional. Here, we can represent the regression surface in an isometric or perspective plot, as a contour plot, or by slicing the surface. These strategies are illustrated in the example developed immediately below. As I will explain, there are obstacles to extending graphical displays of the regression surface beyond two or three explanatory variables.

These problems motivate the additive regression model, described in Section 18.2.2.

The curse of dimensionality and the difficulty of visualizing high-dimensional surfaces limit the practical application of unrestricted nonparametric multiple regression when there are more than a very small number of explanatory variables.

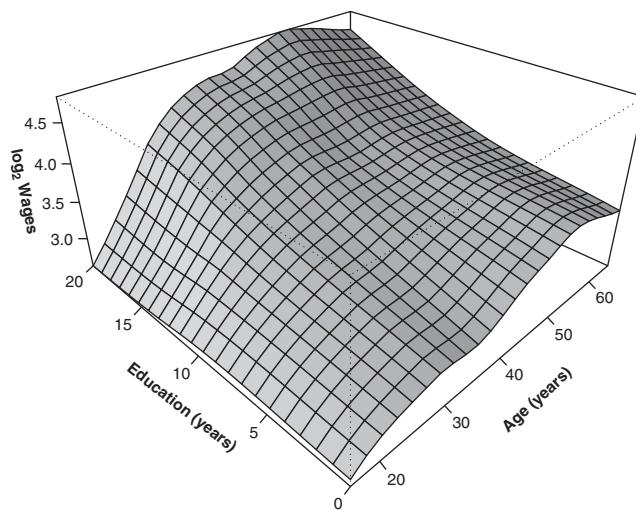


Figure 18.11 Perspective plot for the local-linear regression of log wages on age and education. The span of the local regression is $s = 0.25$.

An Illustration: Data From the Survey of Labour and Income Dynamics

To illustrate local-polynomial multiple regression, I return to data from the Statistics Canada Survey of Labour and Income Dynamics, regressing the log (base 2) of the $n = 3997$ respondents' composite hourly wage rate on their age and years of education.⁴⁷ I selected a span of 0.25 for a local-linear regression after examining the generalized cross-validation criterion.⁴⁸ Figures 18.11 to 18.14 show three graphical representations of the local-linear fit:

- Figure 18.11 is a *perspective plot* (perspective projection) of the fitted regression surface. I find it relatively easy to visualize the general relationship of wages to age and education but hard to make precise visual judgments: I can see that at fixed levels of age, wages generally rise with education (though not at the youngest age levels—see below); likewise, wages first rise and then fall somewhat with age at fixed levels of education. But it is difficult to discern, for example, the fitted value of log wages for a 40-year-old individual with 10 years of education. Perspective plots are even more effective when they can be dynamically rotated on a computer, allowing us to view the regression surface from different angles and conveying a greater sense of depth.
- Figure 18.12 is a *contour plot* of the data, showing “iso-log-wages” lines for combinations of values of age and education. I find it difficult to visualize the regression surface from a contour plot (perhaps hikers and mountain climbers do better), but it is relatively easy to see, for example, that our hypothetical 40-year-old with 10 years of education has fitted log wages of about 3.8 (i.e., fitted wages of $2^{3.8} = \$13.93$ per hour).

⁴⁷We previously encountered these data in Chapter 12, where I dealt with nonlinearity in the regression of log wages on age and education by specifying a quadratic in age and transforming education, and in Chapter 17, where I fit regression splines in the two explanatory variables.

⁴⁸The GCV criterion is lowest between about $s = 0.1$ and $s = 0.2$ and rises very gradually thereafter. Using, for example, $s = 0.15$ produces a regression surface that looks rough, so I opted for somewhat more smoothing.

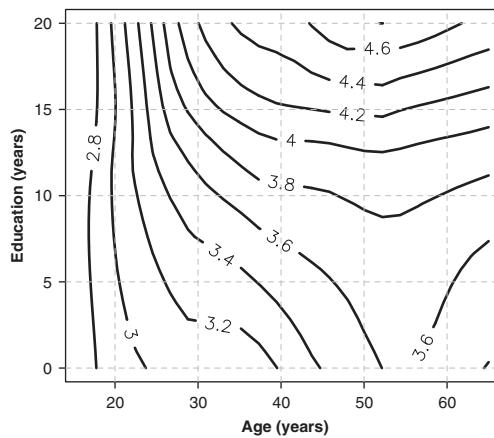


Figure 18.12 Contour plot for the local-liner regression of log-wages on age and education.

- Figure 18.13 is a *conditioning plot* or “coplot,”⁴⁹ showing the fitted relationship between log wages and age for several levels of education. The levels at which education is “held constant” are given in each panel of the figure, which shows the fit at a particular level of education. The lines in the panels of the coplot are lines on the regression surface in the direction of income (fixing education) in Figure 18.11 but displayed two-dimensionally. The broken lines in Figure 18.13 give pointwise 95% confidence envelopes around the fitted regression surface. The confidence envelopes are wide where data are sparse—for example, for 0 years of education. That the shape of the partial relationship of log wages to age varies somewhat with education is indicative of interaction between age and education in affecting income. Figure 18.14 shows a similar coplot displaying the fitted relationship between log wages and education controlling for age. Note that at age 16, the higher levels of education are not possible (a point that could also have been made with respect to the preceding coplot), making the fit in this region a meaningless extrapolation beyond the data. It is useful to display both coplots because both partial relationships are of interest. Again, a small amount of interaction between education and age is apparent in the fit (after all, interaction is a symmetric concept). As well, the degree of nonlinearity in the partial relationships of log wages to education at fixed age levels appears slight in most of the partial plots.

Is log wages significantly related to both age and education? We can answer this question by dropping each explanatory variable in turn and noting the increase in the residual sum of squares. Because the span for the local-linear multiple-regression fit is $s = 0.25$, the corresponding simple-regression models use spans of $s = \sqrt{0.25} = 0.5$.⁵⁰

⁴⁹See Section 3.3.4.

⁵⁰The heuristic here is as follows: In product-marginal kernel weighting of uniformly distributed data, marginal spans of 0.5 produce a neighborhood including roughly $0.5^2 = 0.25$ of the data. This rough reasoning is also supported by the degrees of freedom for the models in the table in Equation 18.11: The model with the age effect and an intercept has 4.3 degrees of freedom, and the model with the education effect and an intercept has 4.9 degrees of freedom. Therefore, a comparable model that allows age and education to interact should have roughly $4.3 \times 4.9 = 21.1$ degrees of freedom—close to the 18.6 degrees of freedom for the local-linear multiple regression with span 0.25.

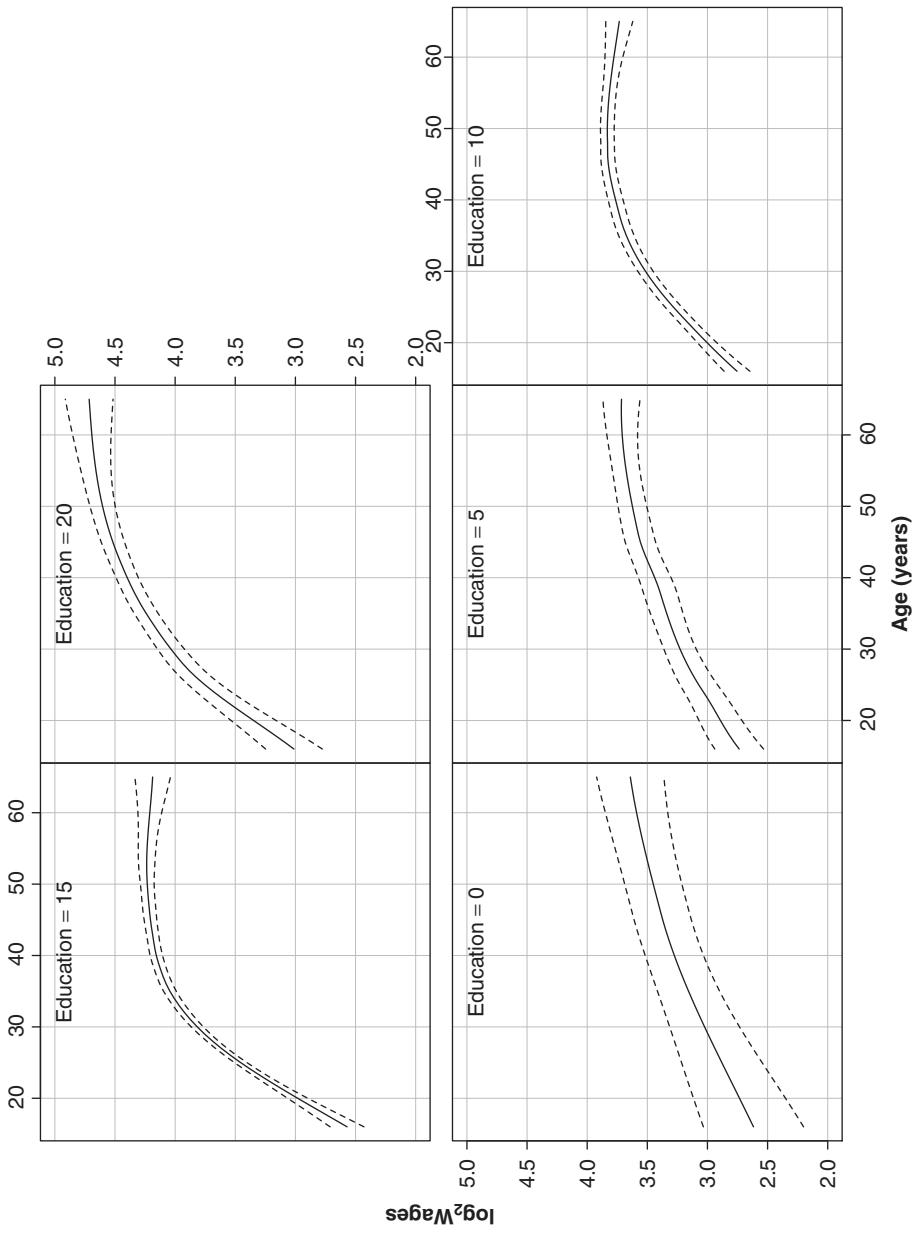


Figure 18.13 Conditioning plot showing the relationship between log-wages and age for various levels of education. The broken lines give a pointwise 95% confidence envelope around the fit.

Model	df_{mod}	RSS	
Age and Education	18.6	1348.592	
Age (alone)	4.3	1547.237	
Education (alone)	4.9	1880.918	(18.11)

F-tests for age and education are as follows:

$$F_{\text{Age}|\text{Education}} = \frac{\frac{1880.918 - 1348.592}{18.6 - 4.9}}{\frac{1348.592}{3997 - 18.6}} = 114.63$$

$$F_{\text{Education}|\text{Age}} = \frac{\frac{1547.237 - 1348.592}{18.6 - 4.3}}{\frac{1348.592}{3997 - 18.6}} = 40.98$$

$F_{\text{Age}|\text{Education}}$, for example, is to be read as the incremental *F*-statistic for age “after” education. These *F*-statistics have, respectively, 13.7 and 3978.4 degrees of freedom, and 14.3 and 3978.4 degrees of freedom. Both *p*-values are close to 0, supporting the partial relationship of log wages to both age and education.

Extension of these displays beyond two or three explanatory variables presents difficulties:

- Perspective plots and contour plots cannot easily be generalized to more than two explanatory variables: Although three-dimensional contour plots can be constructed, they are very difficult to understand, in my opinion, and higher-dimensional contour plots are out of the question.
- One can draw two-dimensional perspective or contour plots for two explanatory variables at fixed combinations of values of other explanatory variables, but the resulting displays are usually confusing.
- Coplots can be usefully constructed for three explanatory variables by arranging combinations of values of two of the variables in a rectangular array and focusing on the fitted relationship of the response to the third explanatory variable. A complete set of coplots rotates the role of the third variable, producing three such displays.
- Coplots can in principle be extended to *any* number of explanatory variables by focusing on each variable in turn, but the resulting proliferation of graphs quickly gets unwieldy.

When there are two explanatory variables, the fitted nonparametric-regression surface can be visualized in a three-dimensional perspective plot, in a contour plot, or in coplots for each explanatory variable at fixed levels of the other variable. Coplots can be generalized to three or more explanatory variables but quickly become unwieldy.

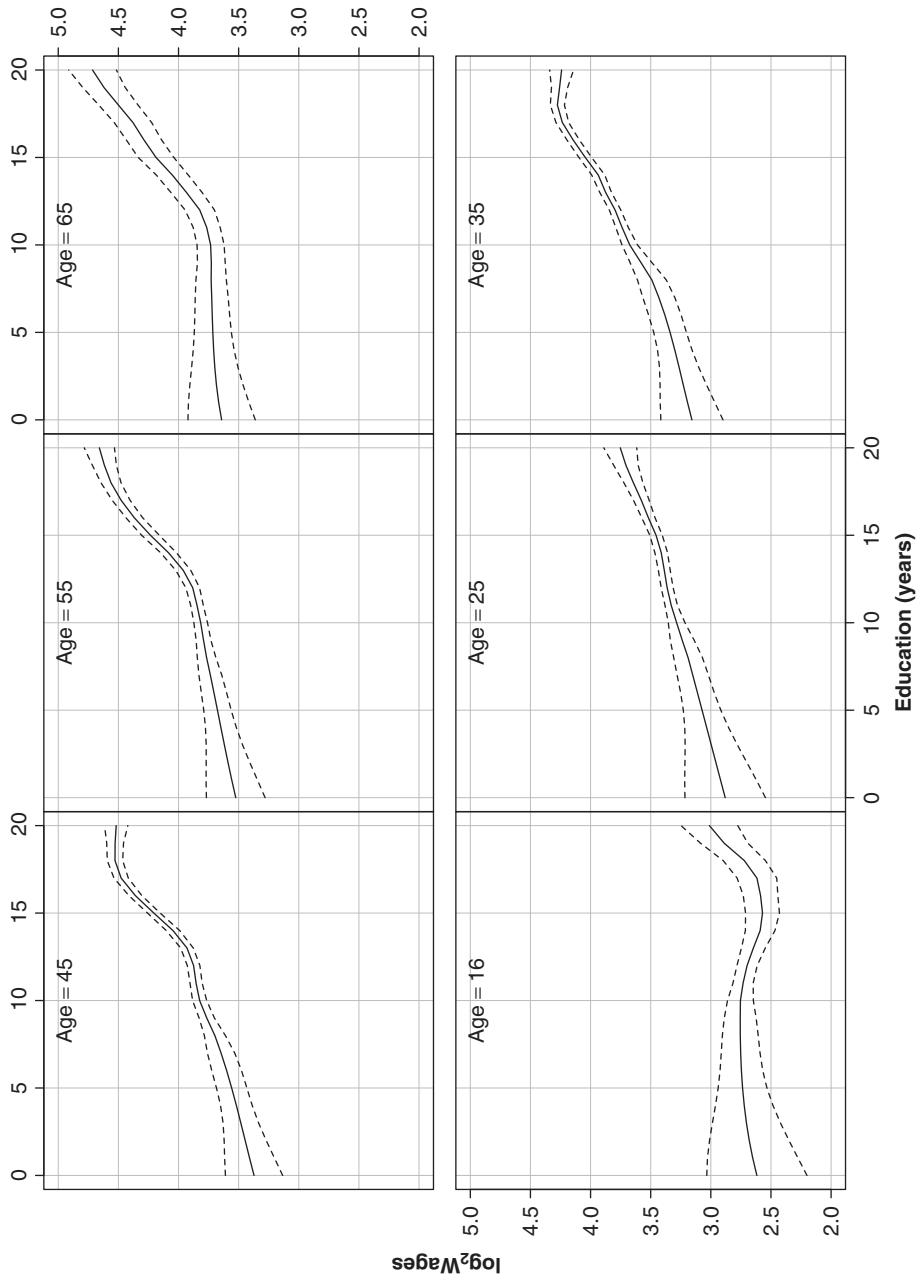


Figure 18.14 Conditioning plot showing the relationship between \log_2 wages and education for various levels of age. The broken lines give a pointwise 95% confidence envelope around the fit.

18.2.2 Additive Regression Models

In unrestricted nonparametric multiple regression, we model the conditional average value of Y as a general, smooth function of several X s,

$$E(Y|x_1, x_2, \dots, x_k) = f(x_1, x_2, \dots, x_k)$$

In linear-regression analysis, in contrast, the average value of the response variable is modeled as a linear function of the explanatory variables,

$$E(Y|x_1, x_2, \dots, x_k) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Like the linear model, the *additive regression model* specifies that the average value of Y is the sum of separate terms for each explanatory variable, but these terms are merely assumed to be smooth functions of the X s:

$$E(Y|x_1, x_2, \dots, x_k) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k)$$

Because it excludes interactions among the X s, the additive regression model is more restrictive than the general nonparametric-regression model but more flexible than the standard linear-regression model.

An advantage of the additive regression model in comparison to the general nonparametric-regression model is that the additive model reduces to a series of two-dimensional partial-regression problems. This is true both in the computational sense and, even more important, with respect to interpretation:

- Because each partial-regression problem is two-dimensional, we can estimate the partial relationship between Y and X_j by using a suitable scatterplot smoother, such as local-polynomial regression or a smoothing spline. We need somehow to remove the effects of the other explanatory variables, however—we cannot simply smooth the marginal scatterplot of Y on X_j *ignoring* the other X s. Details are given later in this section.
- A two-dimensional plot suffices to examine the estimated partial-regression function \hat{f}_j relating Y to X_j holding the other X s constant. Interpretation of additive regression models is therefore relatively simple—assuming that the additive model adequately captures the dependence of Y on the X s.

The additive regression model

$$E(Y|x_1, x_2, \dots, x_k) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k)$$

expresses the average value of the response variable as the sum of smooth functions of several explanatory variables. The additive model is therefore more restrictive than the general nonparametric-regression model but more flexible than the linear-regression model.

Figure 18.15 shows estimated partial-regression functions for the additive regression of log wages on age and education in the SLID data. Each partial-regression function was fit by a nearest-neighbor local-linear smoother, using span $s = 0.5$. The points in each graph are *partial*

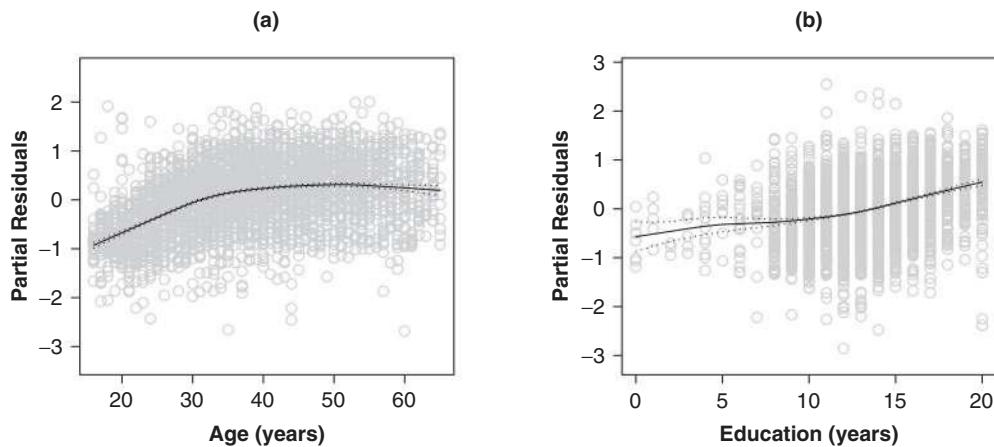


Figure 18.15 Plots of the estimated partial-regression functions for the additive regression of log wages on (a) age and (b) education. Each partial regression uses a local-linear smoother with span $s = 0.5$. The points in the graphs represent partial residuals for each explanatory variable. The broken lines give pointwise 95% confidence envelopes for the partial fits.

residuals for the corresponding explanatory variable, removing the effect of the other explanatory variable. The broken lines mark off pointwise 95% confidence envelopes for the partial fits (both of which are very precisely estimated in this moderately large data set).

The component-plus-residual plot, which graphs partial residuals against an explanatory variable, is a standard diagnostic for nonlinearity in regression.⁵¹ The additive model extends the notion of partial residuals by subtracting the potentially *nonlinear* fits for the other X s from the response; for example, for X_1 ,

$$E_{i[1]} = Y_i - A - \hat{f}_2(x_{i2}) - \cdots - \hat{f}_k(x_{ik})$$

Then, $E_{i[1]}$ is smoothed against X_1 to estimate f_1 . (To apply this idea, we need an estimate A of α and estimates of the *other* partial-regression functions, f_2 through f_k —see below.)

Figure 18.16 is a three-dimensional perspective plot of the fitted additive-regression surface relating log wages to age and education. Slices of this surface in the direction of age (i.e., holding education constant at various values) are all parallel; likewise, slices in the direction of education (holding age constant) are parallel: This is the essence of the additive model, ruling out interaction between the explanatory variables. Because all the slices in a given direction are parallel, we need only view *one* of them edge-on, as in Figure 18.15. Compare the additive-regression surface with the fit of the unrestricted nonparametric-regression model in Figure 18.11 (on page 558).

⁵¹See Section 12.3.1 for a discussion of component-plus-residual plots. The notation for partial residuals here differs from that used in Section 12.3.1, however, by denoting the explanatory variable in question by a bracketed subscript rather than a parenthetical superscript. In the development of additive regression below, I will use a parenthetical superscript for an iteration counter.

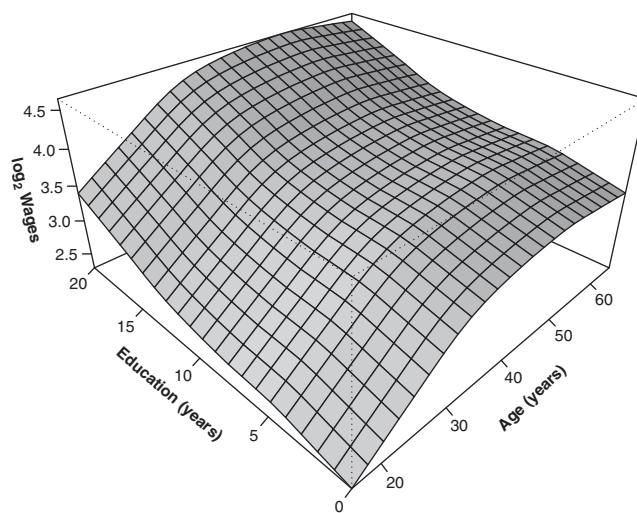


Figure 18.16 Perspective plot of the fitted additive regression of log wages on age and education.

Is anything lost in moving from the general nonparametric-regression model to the more restrictive additive model? Residual sums of squares and equivalent numbers of parameters (degrees of freedom) for the two models are as follows:

Model	df_{mod}	RSS
General	18.6	1348.592
Additive	8.2	1377.801

An approximate F -test comparing the two models is

$$F_0 = \frac{\frac{1377.801 - 1348.592}{18.6 - 8.2}}{\frac{1348.592}{3997 - 18.6}} = 8.29$$

with 10.4 and 3978.4 degrees of freedom, for which $p \ll .0001$. There is, therefore, strong evidence of lack of fit for the additive model; nevertheless, the additive model may be a reasonable, if simplified, summary of the data: The proportion of variation accounted for by the additive model is only slightly smaller than for the general model,

$$\begin{aligned} \text{General: } R^2 &= 1 - \frac{1348.592}{2104.738} = 0.3593 \\ \text{Additive: } R^2 &= 1 - \frac{1377.801}{2104.738} = 0.3454 \end{aligned}$$

where 2104.738 is the total sum of squares for log wages.⁵²

⁵²Issues of model selection are discussed in a general context in Chapter 22.

To test the contribution of each explanatory variable to the additive model, we compare the full additive model with models omitting each variable in turn:

Model	df_{mod}	RSS
Additive	8.2	1377.801
Age (alone)	4.3	1547.237
Education (alone)	4.9	1880.918

Then,

$$F_{\text{Age}|\text{Education}} = \frac{\frac{1880.918 - 1377.801}{8.2 - 4.9}}{\frac{1348.592}{3997 - 18.6}} = 449.76$$

$$F_{\text{Education}|\text{Age}} = \frac{\frac{1547.237 - 1377.801}{8.2 - 4.3}}{\frac{1348.592}{3997 - 18.6}} = 128.16$$

with, respectively, 3.3 and 3978.4, and 3.9 and 3978.4 degrees of freedom; both F -statistics have p -values close to 0. Because there are only two explanatory variables, the second and third models in the table are the same as we employed to test the contribution of each explanatory variable to the *general* nonparametric-regression model (see the table in Equation 18.11 on page 561). The error-variance estimate in the denominator of the F -statistic is based on the largest model that we fit to the data—the general nonparametric-regression model of the preceding section.

Fitting the Additive Regression Model

For simplicity, consider the case of two explanatory variables, as in the regression of log wages on age and education; the generalization to several explanatory variables is immediate (as is described subsequently):

$$Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i$$

Suppose, unrealistically, that the partial-regression function f_2 is known but that f_1 is not. Rearranging the regression equation,

$$Y_i - f_2(x_{i2}) = \alpha + f_1(x_{i1}) + \varepsilon_i$$

So smoothing $Y_i - f_2(x_{i2})$ against x_{i1} will produce an estimate of $\alpha + f_1(x_{i1})$.

The regression constant α is a bit of a nuisance here. We could absorb α into one of the partial-regression functions. Or—somewhat more gracefully—we could force the partial-regression functions evaluated at the observed x_{ij} s to sum to 0; in this case, α becomes the unconditional expectation of Y , estimated by \bar{Y} . Then, we estimate f_1 by smoothing $Y_i - \bar{Y} - f_2(x_{i2})$ against x_{i1} . Of course, in a real application, neither f_1 nor f_2 is known.

- Let us start, then, with preliminary estimates of the partial-regression functions, denoted $\hat{f}_1^{(0)}$ and $\hat{f}_2^{(0)}$, based on the *linear* least-squares regression of Y on the X s:

$$Y_i - \bar{Y} = B_1(x_{i1} - \bar{x}_1) + B_2(x_{i2} - \bar{x}_2) + E_i$$

[The parenthetical superscript (0) indicates that these are “Step 0” estimates in an iterative (repetitive) process of estimation.] Then,

$$\begin{aligned}\hat{f}_1^{(0)}(x_{i1}) &= B_1(x_{i2} - \bar{x}_2) \\ \hat{f}_2^{(0)}(x_{i2}) &= B_2(x_{i2} - \bar{x}_2)\end{aligned}$$

Expressing the variables as deviations from their means ensures that the partial-regression functions sum to 0.

- Form the partial residual

$$\begin{aligned}E_{i[1]}^{(1)} &= Y_i - \bar{Y} - B_2(x_{i2} - \bar{x}_2) \\ &= E_i + B_1(x_{i1} - \bar{x}_1)\end{aligned}$$

which removes from Y its linear relationship to X_2 but retains the linear relationship between Y and X_1 , possibly along with a nonlinear relationship in the least-squares residuals E_i .⁵³ Smoothing $E_{i[1]}^{(1)}$ against X_{i1} provides a new estimate $\hat{f}_1^{(1)}$ of f_1 . [The parenthetical superscript (1) in $E_{i[1]}^{(1)}$ and $\hat{f}_1^{(1)}$ indicates that these quantities pertain to iteration 1; the bracketed subscript [1] in $E_{i[1]}^{(1)}$ indicates that these are the partial residuals for the *first* explanatory variable, X_1 .]

- Using the updated estimate $\hat{f}_1^{(1)}$, form partial residuals for X_2 :

$$E_{i[2]}^{(1)} = Y_i - \bar{Y} - \hat{f}_1^{(1)}(x_{i1})$$

Smoothing $E_{i[2]}^{(1)}$ against x_{i2} yields a new estimate $\hat{f}_2^{(1)}$ of f_2 .

- The new estimate $\hat{f}_2^{(1)}$, in turn, is used to calculate updated partial residuals $E_{i[1]}^{(2)}$ for X_1 , which, when smoothed against x_{i1} , produce the updated estimate $\hat{f}_1^{(2)}$ of f_1 . This iterative process, called *backfitting*, continues until the estimated partial-regression functions stabilize.

The additive regression model can be fit to data by the method of backfitting, which iteratively smooths the partial residuals for each explanatory variable by using current estimates of the regression functions for other explanatory variables.

Some Statistical Details*

More on Backfitting Backfitting implicitly solves the following set of estimating equations:

⁵³These are simply the familiar partial residuals used for component-plus-residual plots in linear regression (described in Section 12.3.1).

$$\begin{bmatrix} 1 & \mathbf{0}'_n & \mathbf{0}'_n & \cdots & \mathbf{0}'_n \\ \mathbf{0}_n & \mathbf{I}_n & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{0}_n & \mathbf{S}_2 & \mathbf{I}_n & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_n & \mathbf{S}_k & \mathbf{S}_k & \cdots & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} A \\ \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_k \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \mathbf{1}'_n \mathbf{y} \\ \mathbf{S}_1 \mathbf{y} \\ \mathbf{S}_2 \mathbf{y} \\ \vdots \\ \mathbf{S}_k \mathbf{y} \end{bmatrix} \quad (18.12)$$

$$\mathbf{S}_{[(kn+1) \times (kn+1)]} \hat{\mathbf{f}}_{[(kn+1) \times 1]} = \mathbf{Q}_{[(kn+1) \times n]} \mathbf{y}_{(n \times 1)}$$

where

- A is the estimate of the regression intercept, α .
- $\mathbf{0}_n$ is an $n \times 1$ column-vector of 0s, and thus $\mathbf{0}'_n$ is a $1 \times n$ row-vector of 0s.
- $\mathbf{1}_n$ is an $n \times 1$ vector of 1s.
- \mathbf{I}_n is the order- n identity matrix.
- \mathbf{S}_j is the smoother matrix for the j th explanatory variable.⁵⁴
- $\hat{\mathbf{f}}_j = \{\hat{f}_j(x_{ij})\}$ is the $n \times 1$ vector of partial-regression estimates for the j th explanatory variable, evaluated at the observed values, x_{ij} .

The first estimating equation simply specifies that $A = \frac{1}{n} \mathbf{1}'_n \mathbf{y} = \bar{Y}$. The remaining matrix equations, composing the rows of Equation 18.12, are each of the form

$$\hat{\mathbf{f}}_j + \mathbf{S}_j \sum_{r \neq j} \hat{\mathbf{f}}_r = \mathbf{S}_j \mathbf{y}$$

Solving for $\hat{\mathbf{f}}_j$, the fitted partial-regression function is the smoothed partial residual:

$$\hat{\mathbf{f}}_j = \mathbf{S}_j \left(\mathbf{y} - \sum_{r \neq j} \hat{\mathbf{f}}_r \right)$$

The estimating equations (18.12) are a system of $kn + 1$ linear equations in an equal number of unknowns. As long as the composite smoother matrix \mathbf{S} is nonsingular—which would normally be the case—these equations have the explicit solution

$$\hat{\mathbf{f}} = \mathbf{S}^{-1} \mathbf{Q} \mathbf{y} = \mathbf{R} \mathbf{y} \quad (18.13)$$

(defining $\mathbf{R} \equiv \mathbf{S}^{-1} \mathbf{Q}$). The size of this system of equations, however, makes it impractical to solve it directly by inverting \mathbf{S} . Backfitting is a practical, iterative procedure for solving the estimating equations.

Statistical Inference It is apparent from Equation 18.13 that the fitted partial-regression functions are linear functions of the response variable. Focusing on the fit for the j th explanatory variable, therefore,

$$V(\hat{\mathbf{f}}_j) = \mathbf{R}_j V(\mathbf{y}) \mathbf{R}'_j = \sigma^2 \mathbf{R}_j \mathbf{R}'_j$$

where \mathbf{R}_j comprises the rows of \mathbf{R} that produce $\hat{\mathbf{f}}_j$.

⁵⁴The smoother matrix was introduced on page 548.

To apply this result, we require an estimate of the error variance (to be addressed presently). A more immediate obstacle is that we must compute \mathbf{R}_j , which is difficult to obtain directly. Notice that \mathbf{R}_j , which takes into account relationships among the X s, is different from the smoother matrix \mathbf{S}_j , which depends *only* on the j th X . A simple expedient, which works reasonably well if the explanatory variables are not strongly related, is simply to use \mathbf{S}_j in place of \mathbf{R}_j . To construct a confidence envelope for the fit, we require only the *variances* of the elements of $\hat{\mathbf{f}}_j$, which, in turn depend only on the *diagonal* entries of \mathbf{S}_j , and so the burden of computation is not onerous.⁵⁵

To estimate the error variance σ_e^2 , we need the degrees of freedom for error. Any of the approaches described previously could be adapted here,⁵⁶ substituting the matrix \mathbf{R} from the solution of the estimating equations for the smoother matrix \mathbf{S} . For example, working from the expectation of the residual sum of squares produces

$$df_{\text{res}} = n - \text{trace}(2\mathbf{R} - \mathbf{RR}')$$

Then, the estimated error variance is $S_E^2 = \text{RSS}/(n - df_{\text{res}})$.

Because, as mentioned, finding \mathbf{R} is computationally demanding, a simpler, if rougher, solution is to take the degrees of freedom for each explanatory variable as $df_j = \text{trace}(2\mathbf{S}_j - \mathbf{S}_j\mathbf{S}'_j) - 1$ or even as $df_j = \text{trace}(\mathbf{S}_j) - 1$. Then, define $df_{\text{res}} = n - \sum_{j=1}^k df_j - 1$. Note that 1 is subtracted from the degrees of freedom for each explanatory variable because of the constraint that the partial-regression function for the variable sums to 0, and 1 is subtracted from the residual degrees of freedom to account for the constant α in the model.

F -tests for the contributions of the several explanatory variables are based on incremental sums of squares and differences in degrees of freedom. The incremental sum of squares for X_j is easily found:

$$\text{SS}_j = \text{RSS}_{-j} - \text{RSS}$$

where RSS is the residual sum of squares for the full model, and RSS_{-j} is the residual sum of squares for the model deleting the j th explanatory variable. The degrees of freedom for the effect of X_j are then

$$df_j = \text{trace}(2\mathbf{R} - \mathbf{RR}') - \text{trace}(2\mathbf{R}_{-j} - \mathbf{R}_{-j}\mathbf{R}'_{-j})$$

where \mathbf{R}_{-j} comes from the solution of the estimating equations in the *absence* of variable j . Alternatively, df_j can be approximated, as above.

Semiparametric Models and Models With Interactions

This section develops two straightforward relatives of additive regression models:

1. *Semiparametric models* are additive regression models in which some terms enter non-parametrically while others enter linearly. These models are therefore hybrids of the additive regression model and the linear regression model.
2. Models in which some of the explanatory variables are permitted to interact, for example in pairwise fashion.

⁵⁵Hastie and Tibshirani (1990, Section 5.4.4) suggest a more sophisticated procedure to calculate the \mathbf{R}_j .

⁵⁶See Section 18.1.2.

It is also possible to combine these strategies, so that some terms enter linearly, others additively, and still others are permitted to interact.

The semiparametric regression model is written

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + f_{r+1}(x_{i,r+1}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

where the errors ε_i are, as usual, assumed to be independently and normally distributed with constant variance. The first r regressors, therefore, enter the model linearly, while the partial relationships of Y to the remaining $k - r$ explanatory variables are simply assumed to be smooth. The semiparametric model can be estimated by backfitting. At each iteration, all of the linear terms can be estimated in a single step: Form partial residuals that remove the current estimates of the nonparametric terms, and then regress these partial residuals on X_1, \dots, X_r to obtain updated estimates of the β s.

The semiparametric model is applicable whenever there is reason to believe that one or more X 's enter the regression linearly:

- In rare instances, there may be prior reasons for believing that this is the case, or examination of the data might suggest a linear relationship, perhaps after transforming an X .⁵⁷
- More commonly, if some of the X 's are dummy regressors—representing the effects of one or more categorical explanatory variables—then it is natural to enter the dummy regressors as linear terms.⁵⁸
- Finally, we can test for nonlinearity by contrasting two models, one of which treats an explanatory variable nonparametrically and the other linearly. For example, to test for nonlinearity in the partial relationship between Y and X_1 , we contrast the additive model

$$Y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

with the semiparametric model

$$Y_i = \alpha + \beta_1 x_{i1} + f_2(x_{i2}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

To illustrate this last procedure, let us return to the SLID data, fitting three models for the regression of log income on age and education:

	<i>Model</i>	<i>df_{mod}</i>	<i>RSS</i>
1	Additive	8.2	1377.801
2	Age linear	5.9	1523.883
3	Education linear	5.3	1390.481

Model 1 is the additive regression model (fit previously); Model 2 is a semiparametric model containing a linear term for age and a nonparametric term for education; Model 3 is a semiparametric model with a linear term for education and a nonparametric term for age.

Contrasting Models 1 and 2 produces a test for nonlinearity in the partial relationship of log income to age—that is, a test of the null hypothesis that this partial relationship is linear; contrasting Models 1 and 3 produces a test for nonlinearity in the relationship of log income to education:

⁵⁷Transformations for linearity are discussed in Chapters 4 and 12.

⁵⁸Dummy regressors are introduced in Chapter 7.

$$F_{\text{Age}(\text{nonlinear})} = \frac{\frac{1523.883 - 1377.801}{8.2 - 5.9}}{\frac{1348.592}{3997 - 18.6}} = 187.37$$

$$F_{\text{Education}(\text{nonlinear})} = \frac{\frac{1390.481 - 1377.801}{8.2 - 5.3}}{\frac{1348.592}{3997 - 18.6}} = 12.90$$

The first of these F -test statistics has 2.3 and 3978.4 degrees of freedom, for which $p \approx 0$; the second has 2.9 and 3978.4 degrees of freedom, for which $p \ll .0001$. Once again, the estimated error variance in the denominator of these F -statistics comes from the general nonparametric-regression model, which is the largest model that we have entertained. There is, therefore, reliable evidence of nonlinearity in both partial relationships, but the nonlinearity in the partial relationship of log wages to education is not great, as we can see in Figure 18.15 (page 564) and by comparing the proportion of variation accounted for by the three models:

$$\text{Additive: } R^2 = 1 - \frac{1377.801}{2104.738} = 0.3454$$

$$\text{Age linear: } R^2 = 1 - \frac{1523.883}{2104.738} = 0.2760$$

$$\text{Education linear: } R^2 = 1 - \frac{1390.481}{2104.738} = 0.3394$$

While semiparametric regression models make the additive model more restrictive, incorporating interactions makes the model more flexible. For example, the following model permits interaction (nonadditivity) in the partial relationship of Y to X_1 and X_2 :

$$Y_i = \alpha + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

Once again, this model can be estimated by backfitting, employing a *multiple-regression* smoother (such as local-polynomial multiple regression) to estimate f_{12} . Contrasting this model with the more restrictive additive model produces an incremental F -test for the interaction between X_1 and X_2 . This strategy can, in principle, be extended to models with higher-order interactions—for example, $f_{123}(x_{i1}, x_{i2}, x_{i3})$ —but the curse of dimensionality and difficulty of interpretation limit the utility of such models.

Semiparametric models are additive regression models in which some terms enter nonparametrically while others enter linearly:

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + f_{r+1}(x_{i,r+1}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

Linear terms may be used, for example, to incorporate dummy regressors in the model. Interactions may be included in an otherwise additive regression model by employing a multiple-regression smoother for interacting explanatory variables, such as X_1 and X_2 in the model

$$Y_i = \alpha + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) + \cdots + f_k(x_{ik}) + \varepsilon_i$$

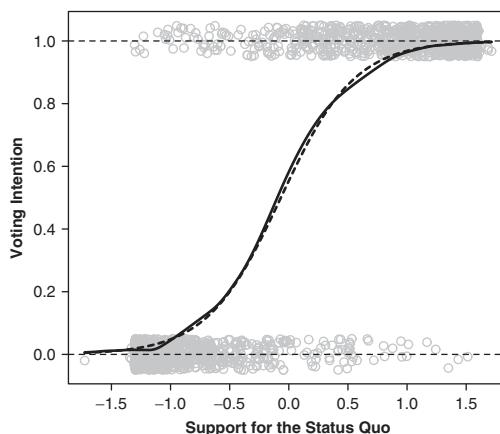


Figure 18.17 Scatterplot of voting intention in the Chilean plebiscite (1 = yes, 0 = no) by support for the status quo. The points are vertically jittered to minimize overplotting. The broken line shows the fit of a linear logistic regression; the solid line shows the fit of a local-linear logistic regression with a span of 0.3.

18.3 Generalized Nonparametric Regression

Generalized nonparametric regression bears the same relationship to the nonparametric-regression models discussed previously in the chapter that generalized linear models bear to linear models,⁵⁹ expanding the range of application of nonparametric regression to a wide variety of response variables. I consider two kinds of generalized nonparametric-regression models in this section: unconstrained generalized nonparametric regression fit by local-likelihood estimation, an extension of the local-polynomial regression models described in Sections 18.1.2 and 18.2.1, and generalized additive models, which are extensions of the additive regression models described in Section 18.2.2.

18.3.1 Local Likelihood Estimation*

Figure 18.17 illustrates generalized regression (and, incidentally, shows why scatterplot smoothing is particularly helpful for dichotomous responses). This graph displays data from a survey conducted prior to the 1989 Chilean plebiscite, where the response variable represents voting intention (1 = yes, 0 = no), and the explanatory variable is a scale indicating support for the status quo (i.e., support for the policies of the military government of Augusto Pinochet, who was then in power).⁶⁰

The points in Figure 18.17 are “jittered” vertically to minimize overplotting. The summary curves on the graph, however, are fit to the unjittered data. Two fitted regressions are shown:

1. The broken line shows the linear logistic regression of voting intention on support for the status quo. The relationship appears to be positive, as expected. Fitted values

⁵⁹Generalized linear models are the subject of Part IV of the text.

⁶⁰A yes vote was a vote to extend military rule in Chile. The Chilean plebiscite data were introduced in Section 14.1.

between 0 and 1 are interpretable as the estimated proportion of *yes* voters at various levels of support for the status quo.

2. The solid line shows a local-linear logistic regression of the kind to be described in this section. The fit in this case is very similar to that of the linear logistic regression, lending credibility to the latter.

Generalized linear models are typically estimated by the method of maximum likelihood.⁶¹ The log-likelihood for these models takes the general form

$$\log_e L = \sum_{i=1}^n l(\mu_i, \phi; Y_i)$$

where the Y_i are the observations on the response variable, ϕ is a dispersion parameter (which takes on fixed values for some generalized linear models, such as binomial and Poisson-regression models, where $\phi = 1$), and

$$\mu_i \equiv E(Y_i) = g^{-1}(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

Here, $g^{-1}(\cdot)$ is the inverse of the link function. For example, for a binary logistic-regression model, the components of the log-likelihood are

$$l(\mu_i; Y_i) = Y_i \log_e \mu_i + (1 - Y_i) \log_e (1 - \mu_i)$$

and the expected value of Y is

$$\begin{aligned} \mu_i &= g^{-1}(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \\ &= \frac{1}{1 + \exp[-(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]} \end{aligned}$$

The maximum-likelihood estimates of the parameters are the values $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$ that maximize $\log_e L$.

In *local-polynomial generalized nonparametric regression*, we estimate the regression function at some set of focal values of the explanatory variables. For simplicity, suppose that there is one X , that the response variable is dichotomous, and that we want to estimate $\mu|x_0$ at the focal value x_0 . We can perform a logistic polynomial regression of the form

$$\log_e \frac{\mu_i}{1 - \mu_i} = \alpha + \beta_1(x_i - x_0) + \beta_2(x_i - x_0)^2 + \cdots + \beta_p(x_i - x_0)^p$$

maximizing the *weighted log-likelihood*

$$\log_e L_w = \sum_{i=1}^n w_i l(\mu_i; Y_i)$$

where the $w_i = K[(x_i - x_0)/h]$ are kernel weights. Then, $\hat{\mu}|x_0 = g^{-1}(\hat{\alpha})$.

To trace the estimated regression curve, as in Figure 18.17, we repeat this procedure for representative values of X or at the observed x_i . As in local-polynomial least-squares regression, the window half-width h can be fixed or adjusted to include a fixed number of nearest neighbors of the focal x_0 .

⁶¹See Chapters 14 and 15. You may wish to review this material prior to proceeding.

Other characteristics of local-polynomial regression generalize readily as well. For example, as in a generalized linear model, the residual deviance under a generalized local-regression model is twice the difference between the log-likelihood for a saturated model that fits the response perfectly and for the model in question:⁶²

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) \equiv 2[\log_e L(\mathbf{y}, \phi; \mathbf{y}) - \log_e L(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})]$$

where $\mathbf{y} \equiv \{Y_i\}$ is the response vector, and $\hat{\boldsymbol{\mu}} \equiv \{\hat{\mu}_i\}$ is the vector of fitted values—that is, the local generalized regression evaluated at the observed X -values. Note that this log-likelihood is *not* the weighted likelihood used to get individual fitted values but is rather based on those fitted values.

Approximate hypothesis tests can be formulated from the deviance and equivalent degrees of freedom (obtaining the latter, e.g., from the trace of \mathbf{SS}'), much as in a generalized linear model. Similarly, the GCV criterion (Equation 18.4 on page 541) becomes

$$\text{GCV}(s) \equiv \frac{n \times D(s)}{[df_{\text{res}}(s)]^2} \quad (18.14)$$

where D is the deviance and s is the span of the local generalized-regression smoother.⁶³

Applied to the Chilean plebiscite data, the GCV criterion suggested a span of 0.13, which produced a fitted regression curve that looked too rough. I adjusted the span visually to $s = 0.3$. The deviance and equivalent degrees of freedom associated with the local logistic-regression model for $s = 0.3$ are, respectively, $D = 746.33$ and $df_{\text{mod}} = 6.2$. The deviance and degrees of freedom for the *linear* logistic-regression model, in comparison, are $D = 752.59$ and $df_{\text{mod}} = 2$. A likelihood-ratio chi-square test for lack of fit in the linear logistic-regression model is therefore $G_0^2 = 752.59 - 746.33 = 6.26$ on $6.2 - 2 = 4.2$ degrees of freedom, for which $p = .20$, suggesting that this model fits the data adequately.

The extension of this approach to multiple regression is straightforward, although the curse of dimensionality and the difficulty of interpreting higher-dimensional fits are no less a problem than in local least-squares regression.

The method of local likelihood can be used to fit generalized local-polynomial nonparametric-regression models, where, as in generalized linear models, the conditional distribution of the response variable can be a member of an exponential family—such as a binomial or Poisson distribution. Statistical inference can then be based on the deviance and equivalent degrees of freedom for the model, formulating likelihood-ratio chi-square or F -tests as in a generalized linear model. Other aspects of local-polynomial nonparametric regression also generalize readily, such as the selection of the span by generalized cross-validation.

⁶²See Section 15.3.2.

⁶³In generalized regression models for distributional families in which the dispersion is fixed, such as the binomial or Poisson families, an alternative is to minimize the *unbiased risk estimator (UBRE)* criterion. See Wood (2006, Section 4.5).

18.3.2 Generalized Additive Models

The *generalized additive model* (or *GAM*) replaces the parametric terms in the generalized linear model with smooth terms in the explanatory variables:

$$\eta_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

where the *additive predictor* η_i plays a role analogous to the linear predictor in a generalized linear model. Local likelihood (described in the preceding section), however, cannot be easily employed to estimate the generalized additive model. An alternative is to adapt the method of *iterated weighted least squares (IWLS)*, which is typically used to obtain maximum-likelihood estimates for generalized linear models.⁶⁴

To keep the level of difficulty relatively low, I will focus on binary logistic regression. Results for other generalized regression models follow a similar pattern. To estimate the additive logistic-regression model,

$$\log_e \frac{\mu_i}{1 - \mu_i} = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

IWLS estimation can be combined with backfitting (introduced in Section 18.2.2):

- Pick starting values of the regression constant and the partial-regression functions, such as

$$\begin{aligned}\alpha^{(0)} &= \log_e \frac{\sum Y_i}{n - \sum Y_i} \\ \text{all } f_j^{(0)}(x_{ij}) &= 0\end{aligned}$$

- Using these initial values, calculate working-response values,

$$Z_i^{(0)} = \eta_i^{(0)} + \frac{Y_i - \mu_i^{(0)}}{\mu_i^{(0)}(1 - \mu_i^{(0)})}$$

and weights,

$$W_i^{(0)} = \mu_i^{(0)}(1 - \mu_i^{(0)})$$

using the additive predictor (in place of the linear predictor of a generalized linear model):

$$\begin{aligned}\eta_i^{(0)} &= \alpha^{(0)} + f_1^{(0)}(x_{i1}) + f_2^{(0)}(x_{i2}) + \cdots + f_k^{(0)}(x_{ik}) \\ \mu_i^{(0)} &= \frac{1}{1 + \exp(-\eta_i^{(0)})}\end{aligned}$$

⁶⁴See Section 15.3.2. For a different approach to estimating GAMs, see Wood (2006).

3. Find new values $\alpha^{(1)}$ and $f_1^{(1)}, \dots, f_k^{(1)}$ by applying the backfitting procedure to the weighted additive regression of $Z^{(0)}$ on the X s, using the $W_i^{(0)}$ as weights.
4. Return to Step 2 to compute new working-response values and weights based on the updated values $\alpha^{(1)}$ and $f_1^{(1)}, \dots, f_k^{(1)}$. Repeat this procedure until the estimates stabilize, producing $\hat{\alpha}$ and $\hat{f}_1, \dots, \hat{f}_k$.

Notice that this estimation procedure is *doubly* iterative, because each backfitting step (Step 3) requires iteration.

Statistical Inference

Once again, I will concentrate on binary logistic regression, with similar results applying to other generalized additive models.

After the IWLS-backfitting procedure converges, the fitted values on the scale of the additive predictor η can be written as linear functions of the working-response values,

$$\hat{\eta}_i = r_{i1}Z_1 + r_{i2}Z_2 + \dots + r_{in}Z_n = \sum_{j=1}^n r_{ij}Z_j$$

The working response Z_j has estimated asymptotic variance $1/[\hat{\mu}_j(1 - \hat{\mu}_j)]$, and because the observations are asymptotically independent, the estimated asymptotic variance of $\hat{\eta}_i$ is⁶⁵

$$\hat{\mathcal{V}}(\hat{\eta}_i) = \sum_{j=1}^n \frac{r_{ij}^2}{\hat{\mu}_j(1 - \hat{\mu}_j)}$$

An approximate pointwise 95% confidence band for the fitted regression surface follows as

$$\hat{\eta}_i \pm 2\sqrt{\hat{\mathcal{V}}(\hat{\eta}_i)}$$

If desired, the endpoints of the confidence band can be transformed to the probability scale by using the inverse of the logit link, $\mu = 1/[1 + \exp(-\eta)]$. Approximate confidence bands can also be constructed for the individual partial-regression functions, f_j .

The deviance for a generalized additive model can be calculated in the usual manner; for a binary logit model, this is

$$D(\boldsymbol{\mu}; \mathbf{y}) = -2 \sum_{i=1}^n [Y_i \log_e \hat{\mu}_i + (1 - Y_i) \log_e (1 - \hat{\mu}_i)]$$

with degrees of freedom equal to n minus the equivalent number of parameters in the model.

⁶⁵There are complications here: The working response is itself a function of the fitted values,

$$Z_i = \hat{\eta}_i + \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i(1 - \hat{\mu}_i)}$$

and, unlike in the additive regression model, the coefficients r_{ij} for transforming the working response depend on the observed Y s. The results given here hold asymptotically, however. See Hastie and Tibshirani (1990, Section 6.8.2).

The generalized additive model (or GAM) replaces the parametric terms in the generalized linear model with smooth terms in the explanatory variables:

$$\eta_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

where the additive predictor η_i plays the same role as the linear predictor in a generalized linear model. GAMs can be fit to data by combining the backfitting algorithm used for additive regression models with the iterated weighted-least-squares algorithm for fitting generalized linear models. Approximate pointwise confidence intervals around the fitted regression surface and statistical tests based on the deviance of the fitted model follow in a straightforward manner.

An Illustration: Labor Force Participation in the SLID

To illustrate generalized additive modeling, I will examine young married women's labor force participation using data from the Canadian Survey of Labour and Income Dynamics (SLID).⁶⁶ The response variable is dichotomous: whether or not the woman worked outside the home at some point during the year preceding the survey. The explanatory variables include a factor representing region (Atlantic Canada, Quebec, Ontario, the prairie provinces, and British Columbia); factors representing the presence in the woman's household of children between 0 and 4 years of age and between 5 and 9 years of age; the woman's after-tax family income, excluding her own income; and the woman's years of education.

I fit a semiparametric generalized additive model to these data, including dummy regressors for region and the two presence-of-children factors, as well as local-linear terms for family income and education, choosing the span for each of these terms by simultaneous generalized cross-validation. That is, I performed a "grid search" over combinations of values of the spans for these two explanatory variables, selecting the combination of spans—as it turned out, $s = 0.7$ in both cases—that together minimize the GCV criterion (given in Equation 18.4 on page 541). The resulting model uses the equivalent of 3.4 degrees of freedom for the family income effect and 2.9 degrees of freedom for the education effect. The coefficients for the region and presence-of-children terms are similar to those of a linear logistic regression fit to these data (see Table 14.4 on page 386) and so are not given here.

Partial plots for family income and education are shown in Figure 18.18. Even discounting the rise at the right—where the very small number of observations (see the rug-plot at the bottom of the graph) is reflected in very broad confidence limits—the partial plot for family income in panel (a) of this figure suggests some nonlinearity, with labor force participation first rising slightly and then falling with family income. Similarly discounting the region at the left of panel (b), where data are sparse, the partial plot suggests modest nonlinearity in the partial relationship between labor force participation and education.

To test these impressions, I fit three models to the data: the initial semiparametric generalized additive model just described and two other semiparametric models in which income and

⁶⁶This example was used to illustrate linear logistic regression in Section 14.1.4, where the data are described in more detail.

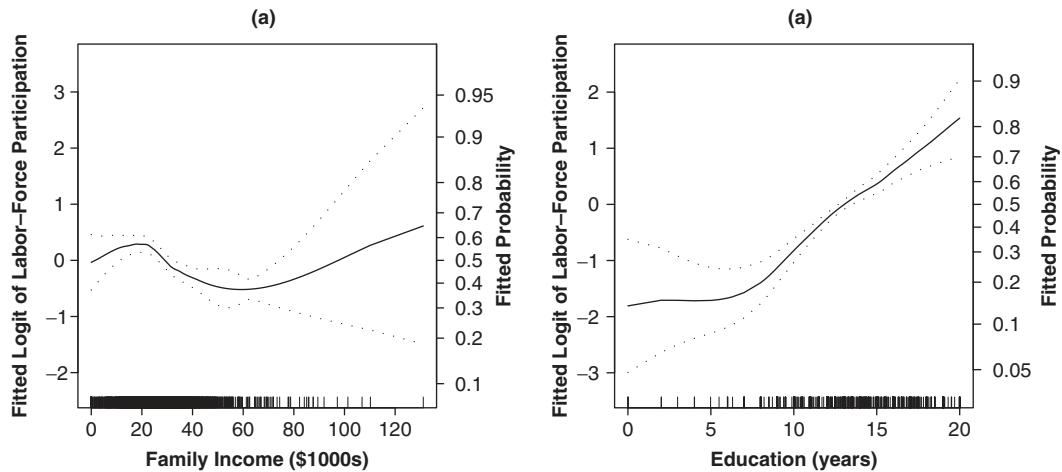


Figure 18.18 Partial plots for family income and education in the generalized semiparametric regression of young married women's labor force participation on region, presence of children 0 to 4 years, presence of children 5 to 9 years, family income, and education. The dotted lines in each plot give approximate pointwise 95% confidence envelopes around the fit. The rug-plot at the bottom of each graph shows the distribution of the corresponding explanatory variable.

education, in turn, enter linearly, producing the following equivalent degrees of freedom and deviances:

	Model	df_{mod}	Residual Deviance
1	Initial model	13.3	1787.007
2	Income linear	10.9	1800.406
3	Education linear	11.4	1797.112

Likelihood-ratio chi-square tests for nonlinearity in the family income and education effects are as follows:

$$G^2_{\text{Income}(\text{nonlinear})} = 1800.406 - 1787.007 = 13.40$$

$$G^2_{\text{Education}(\text{nonlinear})} = 1797.112 - 1787.007 = 10.11$$

on, respectively, 2.4 and 1.9 degrees of freedom, for which $p = .0020$ and $p = .010$. Both tests are therefore statistically significant, but there is stronger evidence of nonlinearity in the relationship between labor force participation and family income.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 18.1. Vary the span of the kernel estimator for the regression of prestige on income in the Canadian occupational prestige data. Does $s = 0.4$ appear to be a reasonable choice?

Exercise 18.2. Selecting the span by smoothing residuals: A complementary visual approach to selecting the span in local-polynomial regression is to find the residuals from the fit from the local regression, $E_i = Y_i - \hat{Y}_i$, and to smooth the residuals against the x_i . If the data have been oversmoothed, then there will be a systematic relationship between the average residual and X ; if the fit does not oversmooth the data, then the average residual will be approximately 0 regardless of the value of X . We seek the *largest* value of s that yields residuals that are unrelated to X . Apply this approach to the regression of prestige on income in the Canadian occupational prestige data by smoothing the residuals from local-linear regressions with various spans. An examination of the scatterplots in Figure 18.4 suggested picking $s \approx 0.6$. Is this choice supported by smoothing the residuals?

Exercise 18.3. Comparing the kernel and local-linear estimators: To illustrate the reduced bias of the local-linear estimator in comparison to the kernel estimator, generate $n = 100$ observations of artificial data according to the cubic regression equation

$$Y = 100 - 5\left(\frac{x}{10} - 5\right) + \left(\frac{x}{10} - 5\right)^3 + \varepsilon \quad (18.15)$$

where the X -values are sampled from the uniform distribution $X \sim U(0, 100)$, and the errors are sampled from the normal distribution $\varepsilon \sim N(0, 20^2)$. Draw a scatterplot of the data showing the true regression line $E(Y) = 100 - 5(x/10 - 5) + (x/10 - 5)^3$. Then, use both kernel regression and local-linear regression to estimate the regression of Y on X , in each case adjusting the span to produce a smooth regression curve. Which estimator has less bias? Why?⁶⁷ Save the data from this exercise, or generate the data in a manner that can be replicated.

Exercise 18.4. *Bias, variance, and MSE as a function of bandwidth: Consider the artificial regression function introduced in the preceding exercise. Using Equation 18.1 (page 537), write down expressions for the expected value and variance of the local-linear estimator as a function of the bandwidth h of the estimator. Employing these results, compute the variance, bias, and mean-squared error of the local-linear estimator at the focal value $x_0 = 10$ as a function of h , allowing h to range between 1 and 20. What value of h produces the smallest MSE? Does this agree with the optimal bandwidth $h^*(10)$ from Equation 18.2? Then, using Equation 18.2, graph the optimal bandwidth $h^*(x_0)$ as a function of the focal value x_0 , allowing x_0 to range between 0 and 100. Relate the resulting function to the regression function.

Exercise 18.5. *Employing the artificial data generated in Exercise 18.3, use Equation 18.3 (on page 539) to compute the average squared error (ASE) of the local-linear regression estimator for various spans between $s = 0.05$ and $s = 0.95$, drawing a graph of $ASE(s)$ versus s . What span produces the smallest ASE? Does this confirm your visual selection of the span of the local-linear estimator in Exercise 18.3?

Exercise 18.6. *Continuing with the artificial data from Exercise 18.3, graph the cross-validation function $CV(s)$ and generalized cross-validation function $GCV(s)$ as a function of span, letting the span range between $s = 0.05$ and $s = 0.95$.

⁶⁷I am using the term *bias* slightly loosely here, because we are examining the performance of each of these estimators for a *particular* sample, rather than averaged over *all* samples, but the point is nevertheless valid.

- (a) Compare the shape of $CV(s)$ with the average squared error $ASE(s)$ function from the preceding exercise. Are the shapes similar?
- (b) Now compare the level of $CV(s)$ to that of $ASE(s)$. Are the levels different? Why?
- (c) Does $GCV(s)$ do a good job of approximating $CV(s)$?
- (d) Do $CV(s)$ and $GCV(s)$ provide useful guidance for selecting the span in this problem?

Exercise 18.7. Comparing polynomial and local regression:

- (a) The local-linear regression of prestige on income with span $s = 0.6$ (in Figure 18.7 on page 543) has 5.006 equivalent degrees of freedom, very close to the number of degrees of freedom for a global fourth-order polynomial. Fit a fourth-order polynomial to these data and compare the resulting regression curve with the local-linear regression.
- (b) Now, consider the local-linear regression of infant mortality on GDP per capita for 193 nations shown in Figure 3.14 (page 45), which is for a span of $s = 0.5$ and which has 5.9 equivalent degrees of freedom. Fit a fifth-order polynomial to these data and compare the fitted regression curve to the local-linear regression.
- (c) What do you conclude from these two examples?

Exercise 18.8. Equivalent kernels: One way of comparing linear smoothers like local-polynomial estimators and smoothing splines is to think of them as variants of the kernel estimator, where fitted values arise as weighted averages of observed response values. This approach is illustrated in Figure 18.19, which shows *equivalent kernel weights* at two focal X -values in the Canadian occupational prestige data: One value, $x_{(5)}$, is near the boundary of the data; the other, $x_{(60)}$, is closer to the middle of the data. The figure shows tricube-kernel weights [panels (a) and (b)], along with the equivalent kernel weights for the local-linear estimator with span = 0.6 (or five equivalent parameters) [panels (c) and (d)] and the smoothing spline with five equivalent parameters [in panels (e) and (f)]. Compare and contrast the equivalent kernel weights for the three estimators. Are there any properties of the equivalent kernels for the local-linear and smoothing-spline estimators that you find surprising?

Summary

- Kernel regression estimates the regression function at a focal value x_0 of the explanatory variable by weighted local averaging of Y :

$$\widehat{f}(x_0) = \widehat{Y}|_{x_0} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

The weights are provided by a kernel function, $w_i = K[(x_i - x_0)/h]$, which takes on its largest value at $K(0)$ and falls symmetrically toward 0 as $|(x_i - x_0)/h|$ grows. Observations close to the focal x_0 therefore receive greatest weight. The kernel estimator is evaluated at representative focal values of X or at the ordered X -values, $x_{(i)}$. The bandwidth h of the kernel estimator can be fixed or can be adjusted to include a fixed proportion of the data, called the span of the kernel estimate. The larger the span, the smoother the kernel regression.

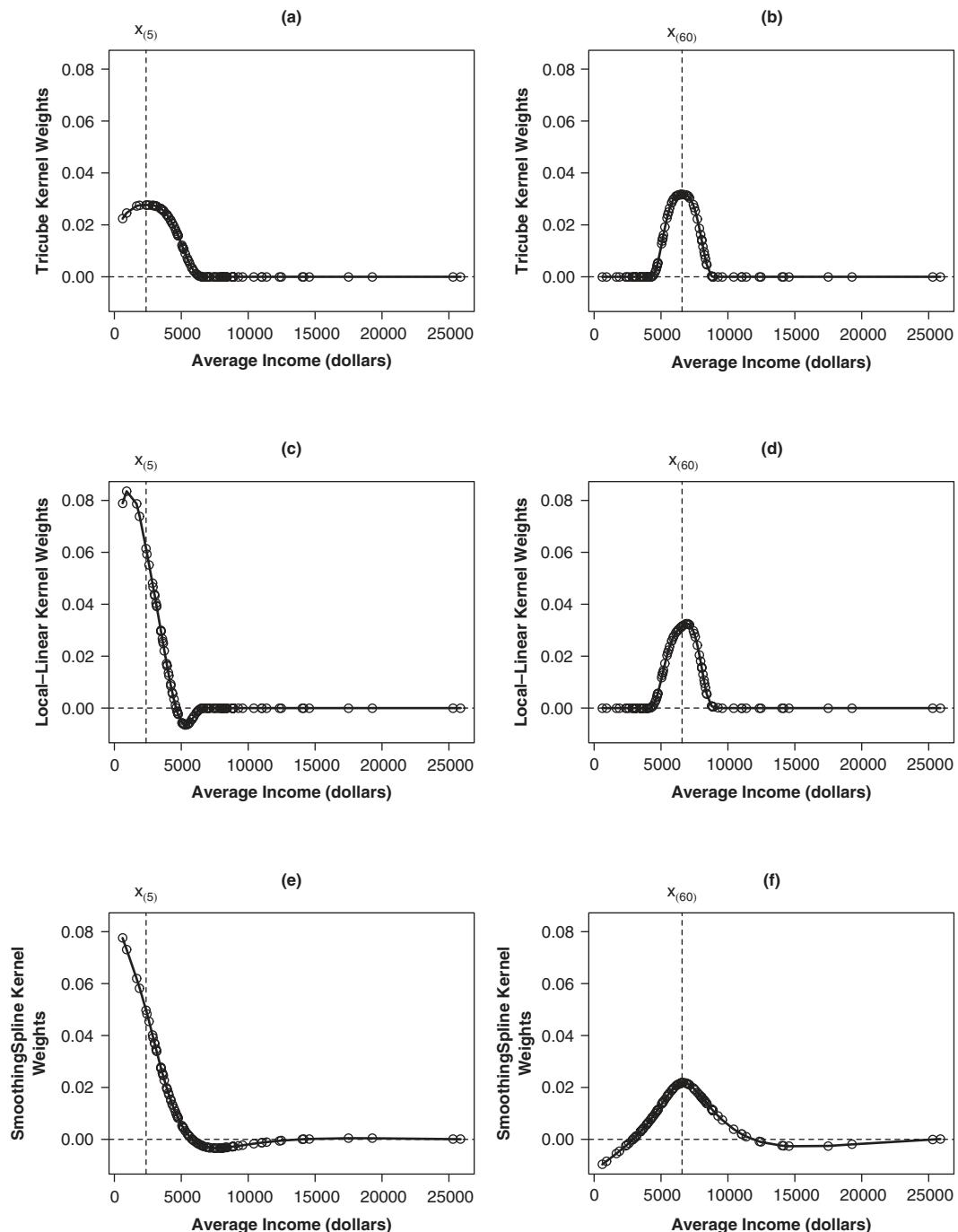


Figure 18.19 Equivalent kernels for three nonparametric estimators of the regression of occupational prestige on income: (a) and (b) nearest-neighbor tricube kernel estimator, with span = 0.6; (c) and (d) nearest-neighbor local-linear estimator, with span = 0.6 (five equivalent parameters); and (e) and (f) smoothing spline, with five equivalent parameters. The focal point, marked at the top of each graph, is $x_{(5)}$ in (a), (c), and (e) and $x_{(60)}$ in (b), (d), and (f).

- Local-polynomial regression extends kernel estimation to a polynomial fit at the focal value x_0 , using local kernel weights, $w_i = K[(x_i - x_0)/h]$. The resulting WLS regression fits the equation

$$Y_i = A + B_1(x_i - x_0) + B_2(x_i - x_0)^2 + \cdots + B_p(x_i - x_0)^p + E_i$$

to minimize the weighted residual sum of squares, $\sum_{i=1}^n w_i E_i^2$. The fitted value at the focal x_0 is just $\hat{Y}|_{x_0} = A$. This procedure is repeated for representative focal values of X or at the observations x_i . We can employ a fixed bandwidth or adjust the bandwidth for a fixed span. Nearest-neighbor local-polynomial regression is often called lowess (or loess).

- A generally effective visual approach to selecting the span in local-polynomial regression is guided trial and error. The span $s = 0.5$ is often a good point of departure. If the fitted regression looks too rough, then try increasing the span; if it looks smooth, then see if the span can be decreased without making the fit too rough. We want the smallest value of s that provides a smooth fit.
- The bias and variance of the local-linear estimator at the focal value x_0 are both a function of the bandwidth h , as well as of properties of the data and the kernel function:

$$\begin{aligned}\text{bias}(\hat{Y}|_{x_0}) &\approx \frac{h^2}{2} s_K^2 f''(x_0) \\ V(\hat{Y}|_{x_0}) &\approx \frac{\sigma_e^2 a_K^2}{n h p_X(x_0)}\end{aligned}$$

where s_K^2 and a_K^2 are constants that depend on the kernel function, $f''(x_0)$ is the second derivative (“curvature”) of the regression function at x_0 , and $p_X(x_0)$ is the probability-density of X -values at x_0 . We would ideally like to choose the value of h at each focal value that minimizes the mean-squared error of estimation—that is, the sum of squared bias and variance.

- The cross-validation function

$$\text{CV}(s) = \frac{\sum_{i=1}^n [\hat{Y}_{-i}(s) - Y_i]^2}{n}$$

can be used to select the span s in local-polynomial regression, picking s to minimize $\text{CV}(s)$. The fitted value at each observation $\hat{Y}_{-i}(s)$ is computed from a local regression that omits that observation. Because the cross-validation function $\text{CV}(s)$ can be costly to compute, approximations such as generalized cross-validation have been proposed. The GCV criterion is

$$\text{GCV}(s) = \frac{n \times \text{RSS}(s)}{[df_{\text{res}}(s)]^2}$$

where $\text{RSS}(s)$ is the residual sum of squares and $df_{\text{res}}(s)$ the “equivalent” residual degrees of freedom for the local-regression smoother with span s .

- The fitted values in a local-polynomial regression are linear functions of the observations, $\hat{Y}_i = \sum_{j=1}^n s_{ij} Y_j$. Estimating the error variance as $S_E^2 = \sum E_i^2 / df_{\text{res}}$, where df_{res} is the equivalent residual degrees of freedom for the model, the estimated variance of a fitted

value is $\widehat{V}(\widehat{Y}_i) = S_E^2 \sum_{j=1}^n s_{ij}^2$. An approximate 95% pointwise confidence band around the regression curve evaluated at the fitted values may be formed as $\widehat{Y}_i \pm 2\sqrt{\widehat{V}(\widehat{Y}_i)}$.

- Approximate incremental F -tests for hypotheses in local-polynomial regression are formulated by contrasting nested models, in analogy to similar tests for linear models fit by least squares. For example, to test the hypothesis of no relationship in the nonparametric-regression model, we can compute the F -test statistic

$$F_0 = \frac{\frac{\text{TSS} - \text{RSS}}{df_{\text{mod}} - 1}}{\frac{\text{RSS}}{df_{\text{res}}}}$$

where df_{mod} and $df_{\text{res}} = n - df_{\text{mod}}$ are respectively the equivalent degrees of freedom for the regression model and for error, and RSS is the residual sum of squares for the model. Similarly, to test for nonlinearity, we can contrast the fitted nonparametric-regression model with a linear model, computing

$$F_0 = \frac{\frac{\text{RSS}_0 - \text{RSS}_1}{df_{\text{mod}} - 2}}{\frac{\text{RSS}_1}{df_{\text{res}}}}$$

where RSS_0 is the residual sum of squares for the linear regression and RSS_1 the residual sum of squares for the more general nonparametric regression.

- The smoother matrix S in nonparametric local-polynomial regression plays a role analogous to the hat-matrix H in linear least-squares regression. Like the hat-matrix, the smoother matrix linearly transforms the observations into the fitted values: $\widehat{y} = Sy$. Pursuing this analogy, the equivalent degrees of freedom for the nonparametric-regression model can variously be defined as $df_{\text{mod}} = \text{trace}(S)$, $\text{trace}(SS')$, or $\text{trace}(2S - SS')$.
- Generalizing local-polynomial regression to multiple regression is conceptually and computationally straightforward. For example, to obtain the fitted value for a local-linear regression at the focal point $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})'$ in the space of the explanatory variables, we perform a weighted-least-squares regression of Y on the X s,

$$Y_i = A + B_1(x_{i1} - x_{01}) + B_2(x_{i2} - x_{02}) + \cdots + B_k(x_{ik} - x_{0k}) + E_i$$

emphasizing observations close to the focal point by minimizing the weighted residual sum of squares, $\sum_{i=1}^n w_i E_i^2$. The fitted value at the focal point in the X -space is then $\widehat{Y}|\mathbf{x}_0 = A$. The weights w_i can be computed in several ways, including by multiplying marginal kernel weights for the several explanatory variables or by basing kernel weights on one or another measure of distance between the focal \mathbf{x}_0 and the observed X -values, \mathbf{x}_i . Given a distance measure $D(\mathbf{x}_i, \mathbf{x}_0)$, the kernel weights are calculated as $w_i = K[D(\mathbf{x}_i, \mathbf{x}_0)/h]$.

- Methods for selecting the span in local-polynomial multiple regression are much the same as in local-polynomial simple regression: We can proceed visually by trial and error or apply a criterion such as $\text{CV}(s)$ or $\text{GCV}(s)$. Similarly, approximate pointwise

confidence limits for the fitted regression can be calculated as in local-polynomial simple regression, as can incremental F -tests comparing nested models.

- The curse of dimensionality and the difficulty of visualizing high-dimensional surfaces limit the practical application of unrestricted nonparametric multiple regression when there are more than a very small number of explanatory variables.
- When there are two explanatory variables, the fitted nonparametric-regression surface can be visualized in a three-dimensional perspective plot, in a contour plot, or in coplots for each explanatory variable at fixed levels of the other variable. Coplots can be generalized to three or more explanatory variables but quickly become unwieldy.
- The additive regression model

$$E(Y|x_1, x_2, \dots, x_k) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k)$$

expresses the average value of the response variable as the sum of smooth functions of several explanatory variables. The additive model is therefore more restrictive than the general nonparametric-regression model but more flexible than the linear-regression model.

- The additive regression model can be fit to data by the method of backfitting, which iteratively smooths the partial residuals for each explanatory variable using current estimates of the regression functions for other explanatory variables.
- Semiparametric models are additive regression models in which some terms enter nonparametrically while others enter linearly:

$$Y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_r x_{ir} + f_{r+1}(x_{i,r+1}) + \dots + f_k(x_{ik}) + \varepsilon_i$$

Linear terms may be used, for example, to incorporate dummy regressors in the model. Interactions may be included in an otherwise additive regression model by employing a multiple-regression smoother for interacting explanatory variables, such as X_1 and X_2 in the model

$$Y_i = \alpha + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) + \dots + f_k(x_{ik}) + \varepsilon_i$$

- The method of local likelihood can be used to fit generalized local-polynomial nonparametric-regression models, where, as in generalized linear models, the conditional distribution of the response variable can be a member of an exponential family—such as a binomial or Poisson distribution. Statistical inference can then be based on the deviance and equivalent degrees of freedom for the model, formulating likelihood-ratio chi-square or F -tests as in a generalized linear model. Other aspects of local-polynomial nonparametric regression also generalize readily, such as the selection of the span by generalized cross-validation.
- The generalized additive model (or GAM) replaces the parametric terms in the generalized linear model with smooth terms in the explanatory variables:

$$\eta_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_k(x_{ik})$$

where the additive predictor η_i plays the same role as the linear predictor in a generalized linear model. GAMs can be fit to data by combining the backfitting algorithm used

for additive regression models with the iterated weighted-least-squares algorithm for fitting generalized linear models. Approximate pointwise confidence intervals around the fitted regression surface and statistical tests based on the deviance of the fitted model follow in a straightforward manner.

Recommended Reading

There are many fine sources on nonparametric regression and smoothing.

- Hastie and Tibshirani's (1990) text on generalized additive models includes a wealth of valuable material. Most of the book is leisurely paced and broadly accessible, with many effective examples. As a preliminary to generalized additive models, Hastie and Tibshirani include a fine treatment of scatterplot smoothing.
- Wood (2006) also presents an excellent and wide-ranging treatment of generalized additive models that stresses smoothing splines and automatic selection of smoothing parameters.
- A briefer presentation by Hastie of generalized additive models appears in an edited book (Chambers & Hastie, 1992) on statistical modeling in the S computing environment (also implemented in R). This book includes a chapter by Cleveland, Grosse, and Shyu on local regression models.
- Cleveland's (1993) text on data visualization presents information on local regression in two and more dimensions.
- Härdle (1991) gives an overview of nonparametric regression, stressing kernel smoothers for bivariate scatterplots.
- Additional details may be found in Fan and Gijbels (1996), Simonoff (1996), and Bowman and Azzalini (1997).
- Much of the exposition in the current chapter was adapted from Fox (2000a, 2000b), which presents the topic in greater detail (and which omits some newer material).

19

Robust Regression*

The efficiency of least-squares regression is seriously impaired by heavy-tailed error distributions; in particular, least squares is vulnerable to outlying observations at high-leverage points.¹ One response to this problem is to employ diagnostics for high-leverage, influential, and outlying data; if unusual data are discovered, then these can be corrected, removed, or otherwise accommodated.

Robust estimation is an alternative approach to outliers and the heavy-tailed error distributions that tend to generate them. Properly formulated, robust estimators are almost as efficient as least squares when the error distribution is normal and much more efficient when the errors are heavy tailed. Robust estimators hold their efficiency well because they are resistant to outliers. Rather than simply discarding discrepant data, however, robust estimation (as we will see) down-weights them.

Much of the chapter is devoted to a particular strategy of robust estimation, termed *M estimation*, due originally to Huber (1964). I also describe two other approaches to robust estimation: *bounded-influence regression* and *quantile regression*. Finally, I briefly present robust estimators for generalized linear models.

19.1 M Estimation

19.1.1 Estimating Location

Although our proper interest is in robust estimation of linear models, it is helpful to narrow our focus initially to a simpler setting: robust estimation of *location*—that is, estimation of the center of a distribution. Let us, then, begin our exploration of robust estimation with the minimal linear model

$$Y_i = \mu + \varepsilon_i$$

where the observations Y_i are independently sampled from some symmetric distribution with center μ (and hence the errors ε_i are independently and symmetrically distributed around 0).²

If the distribution from which the observations are drawn is normal, then the sample mean $\hat{\mu} = \bar{Y}$ is the maximally efficient estimator of μ , producing the fitted model

¹See Chapter 11.

²In the absence of symmetry, what we mean by the center of the distribution becomes ambiguous.

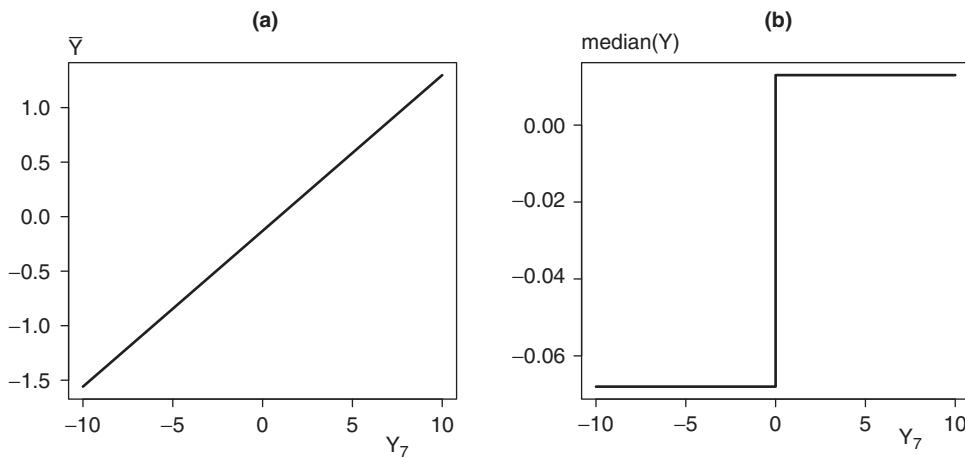


Figure 19.1 The influence functions for the mean (a) and median (b) for the sample $Y_1 = -0.068$, $Y_2 = -1.282$, $Y_3 = 0.013$, $Y_4 = 0.141$, $Y_5 = -0.980$, $Y_6 = 1.263$. The influence function for the median is bounded, while that for the mean is not. Note that the vertical axes for the two graphs have different scales.

$$Y_i = \bar{Y} + E_i$$

The mean minimizes the least-squares *objective function*:

$$\sum_{i=1}^n \rho_{\text{LS}}(E_i) = \sum_{i=1}^n \rho_{\text{LS}}(Y_i - \hat{\mu}) \equiv \sum_{i=1}^n (Y_i - \hat{\mu})^2$$

The mean, however, is very sensitive to outliers, as is simply demonstrated: I drew a sample of six observations from the standard-normal distribution, obtaining

$$\begin{aligned} Y_1 &= -0.068 & Y_2 &= -1.282 & Y_3 &= 0.013 \\ Y_4 &= 0.141 & Y_5 &= -0.980 & Y_6 &= 1.263 \end{aligned}$$

The mean of these six values is $\bar{Y} = -0.152$. Now, imagine adding a seventh observation, Y_7 , allowing it to take on all possible values from -10 to $+10$ (or, with greater imagination, from $-\infty$ to $+\infty$). The result, called the *influence function* of the mean, is graphed in Figure 19.1(a). It is apparent from this figure that as the discrepant seventh observation grows more extreme, the sample mean chases it.

The shape of the influence function for the mean follows from the derivative of the least-squares objective function with respect to E :

$$\psi_{\text{LS}}(E) \equiv \rho'_{\text{LS}}(E) = 2E$$

Influence, therefore, is proportional to the residual E . It is convenient to redefine the least-squares objective function as $\rho_{\text{LS}}(E) \equiv \frac{1}{2}E^2$, so that $\psi_{\text{LS}}(E) = E$.

Now consider the sample median as an estimator of μ . The median minimizes the *least-absolute-values* (LAV) objective function:³

³See Exercise 19.1.

$$\sum_{i=1}^n \rho_{\text{LAV}}(E_i) = \sum_{i=1}^n \rho_{\text{LAV}}(Y_i - \hat{\mu}) \equiv \sum_{i=1}^n |Y_i - \hat{\mu}|$$

As a result, the median is much more resistant than the mean to outliers. The influence function of the median for the illustrative sample is shown in Figure 19.1(b). In contrast to the mean, the influence of a discrepant observation on the median is *bounded*. Once again, the derivative of the objective function gives the shape of the influence function:⁴

$$\psi_{\text{LAV}}(E) \equiv \rho'_{\text{LAV}}(E) = \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}$$

Although the median is more resistant than the mean to outliers, it is less efficient than the mean if the distribution of Y is normal. When $Y \sim N(\mu, \sigma^2)$, the sampling variance of the mean is σ^2/n , while the variance of the median is $\pi\sigma^2/2n$: That is, $\pi/2 \approx 1.57$ times as large as for the mean. Other objective functions combine resistance to outliers with greater robustness of efficiency. Estimators that can be expressed as minimizing an objective function $\sum_{i=1}^n \rho(E)$ are called *M estimators*.⁵

Two common choices of objective functions are the *Huber* and Tukey's *biweight* (or *bisquare*) functions:

- The Huber objective function is a compromise between least squares and least absolute values, behaving like least squares in the center and like least absolute values in the tails:

$$\rho_H(E) = \begin{cases} \frac{1}{2}E^2 & \text{for } |E| \leq k \\ k|E| - \frac{1}{2}k^2 & \text{for } |E| > k \end{cases}$$

The Huber objective function ρ_H and its derivative, the influence function ψ_H , are graphed in Figure 19.2:⁶

$$\psi_H(E) = \begin{cases} k & \text{for } E > k \\ E & \text{for } |E| \leq k \\ -k & \text{for } E < -k \end{cases}$$

The value k , which defines the center and tails, is called a *tuning constant*.

It is most natural to express the tuning constant as a multiple of the *scale* (i.e., the spread) of the variable Y , that is, to take $k = cS$, where S is a measure of scale. The sample standard deviation is a poor measure of scale in this context because it is even more affected than the mean by outliers. A common robust measure of scale is the *median absolute deviation* (MAD):

$$\text{MAD} \equiv \text{median}|Y_i - \hat{\mu}|$$

⁴Strictly speaking, the derivative of ρ_{LAV} is undefined at $E = 0$, but setting $\psi_{\text{LAV}}(0) \equiv 0$ is convenient.

⁵Estimators that can be written in this form can be thought of as generalizations of maximum-likelihood estimators, hence the term *M estimator*. The maximum-likelihood estimator is produced by taking $\rho_{\text{ML}}(y - \mu) \equiv -\log_e p(y - \mu)$ for an appropriate probability or probability density function $p(\cdot)$.

⁶My terminology here is loose but convenient: Strictly speaking, the ψ -function is not the influence function, but it has the same shape as the influence function.

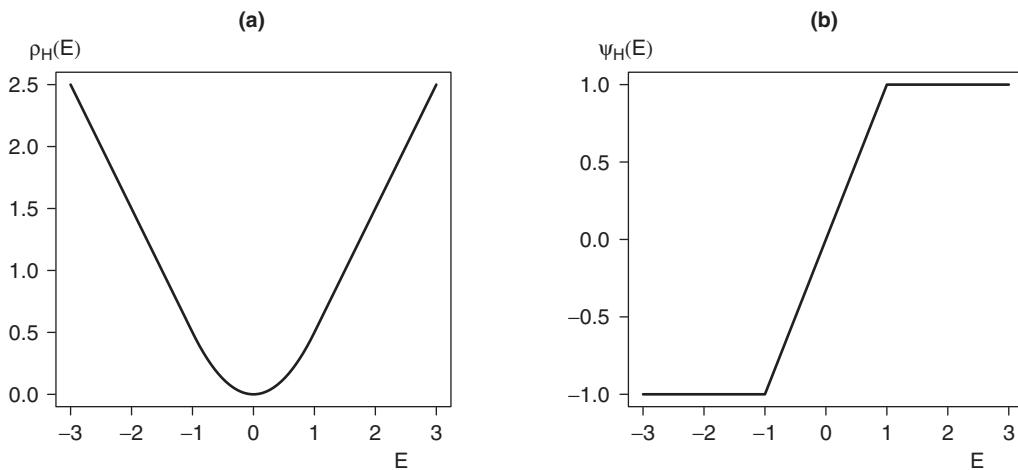


Figure 19.2 Huber objective function ρ_H (a) and “influence function” ψ_H (b). To calibrate these graphs, the tuning constant is set to $k = 1$. (See the text for a discussion of the tuning constant.)

The estimate $\hat{\mu}$ can be taken, at least initially, as the median value of Y . We can then define $S \equiv \text{MAD}/0.6745$, which ensures that S estimates the standard deviation σ when the population is normal. Using $k = 1.345S$ (i.e., $1.345/0.6745 \approx 2$ MADs) produces 95% efficiency relative to the sample mean when the population is normal, along with considerable resistance to outliers when it is not. A smaller tuning constant can be employed for more resistance.

- The biweight (or bisquare) objective function levels off at very large residuals:⁷

$$\rho_{\text{BW}}(E) = \begin{cases} \frac{k^2}{6} \left\{ 1 - \left[1 - \left(\frac{E}{k} \right)^2 \right]^3 \right\} & \text{for } |E| \leq k \\ \frac{k^2}{6} & \text{for } |E| > k \end{cases}$$

The influence function for the biweight estimator, therefore, “redescends” to 0, *completely discounting* observations that are sufficiently discrepant:

$$\psi_{\text{BW}}(E) = \begin{cases} E \left[1 - \left(\frac{E}{k} \right)^2 \right]^2 & \text{for } |E| \leq k \\ 0 & \text{for } |E| > k \end{cases}$$

The functions ρ_{BW} and ψ_{BW} are graphed in Figure 19.3. Using $k = 4.685S$ (i.e., $4.685/0.6745 \approx 7$ MADs) produces 95% efficiency when sampling from a normal population.

⁷The term bisquare applies literally to the ψ -function and to the weight function (hence biweight) to be introduced presently—not to the objective function.

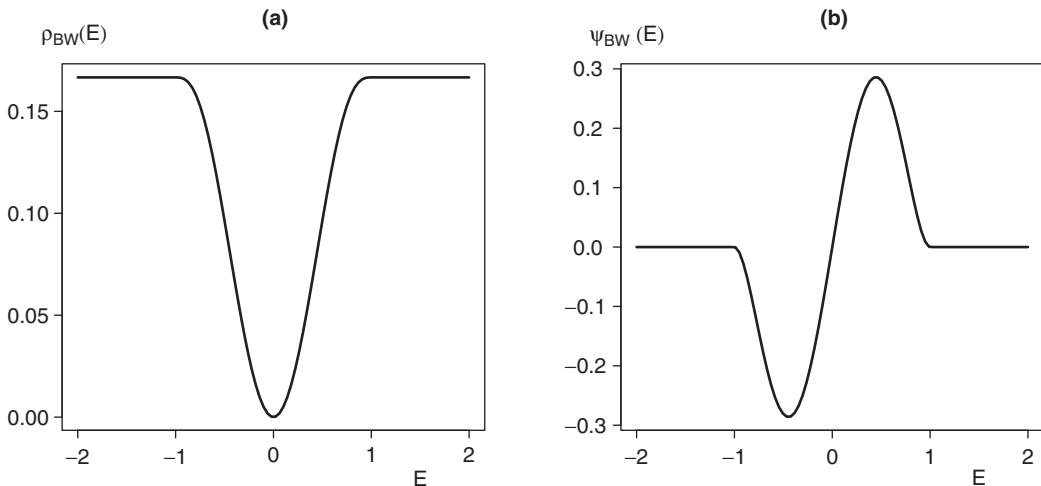


Figure 19.3 Biweight objective function ρ_{BW} (a) and “influence function” ψ_{BW} (b). To calibrate these graphs, the tuning constant is set to $k = 1$. The influence function “redescends” to 0 when $|E|$ is large.

Robust M estimators of location, for the parameter μ in the simple model $Y_i = \mu + \varepsilon_i$, minimize the objective function $\sum_{i=1}^n \rho(E_i) = \sum_{i=1}^n \rho(Y_i - \hat{\mu})$, selecting $\rho(\cdot)$ so that the estimator is relatively unaffected by outlying values. Two common choices of objective function are the Huber and the biweight (or bisquare). The sensitivity of an M estimator to individual observations is expressed by the influence function of the estimator, which has the same shape as the derivative of the objective function, $\psi(E) \equiv \rho'(E)$.

Calculation of M estimators usually requires an iterative procedure (although iteration is not necessary for the mean and median, which, as we have seen, fit into the M estimation framework). An estimating equation for $\hat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\hat{\mu}$) to 0, obtaining

$$\sum_{i=1}^n \psi(Y_i - \hat{\mu}) = 0 \quad (19.1)$$

There are several general approaches to solving Equation 19.1; probably the most straightforward, and the simplest to implement computationally, is to reweight the mean iteratively—a special case of *iterative weighted least squares* (*IWLS*):⁸

⁸In Chapter 15, we employed IWLS estimation for generalized linear models. The method is also called *iteratively reweighted least squares* (*IRLS*).

Table 19.1 Weight Functions $w(E) = \psi(E)/E$ for Several M Estimators

Estimator	Weight Function $w(E)$
Least squares	1
Least absolute values	$1/ E $ (for $E \neq 0$)
Huber	$\begin{cases} 1 & \text{for } E \leq k \\ k/ E & \text{for } E > k \end{cases}$
Bisquare (biweight)	$\begin{cases} \left[1 - \left(\frac{E}{k}\right)^2\right]^2 & \text{for } E \leq k \\ 0 & \text{for } E > k \end{cases}$

1. Define the weight function $w(E) = \psi(E)/E$. Then, the estimating equation becomes

$$\sum_{i=1}^n (Y_i - \hat{\mu}) w_i = 0 \quad (19.2)$$

where

$$w_i \equiv w(Y_i - \hat{\mu})$$

The solution of Equation 19.2 is the weighted mean

$$\hat{\mu} = \frac{\sum w_i Y_i}{\sum w_i}$$

The weight functions corresponding to the least-squares, LAV, Huber, and bisquare objective functions are shown in Table 19.1 and graphed in Figure 19.4. The least-squares weight function accords equal weight to each observation, while the bisquare gives 0 weight to observations that are sufficiently outlying; the LAV and Huber weight functions descend toward 0 but never quite reach it.

2. Select an initial estimate $\hat{\mu}^{(0)}$, such as the median of the Y values.⁹ Using $\hat{\mu}^{(0)}$, calculate an initial estimate of scale $S^{(0)}$ and initial weights $w_i^{(0)} = w(Y_i - \hat{\mu}^{(0)})$. Set the iteration counter $l = 0$. The scale is required to calculate the tuning constant $k = cS$ (for prespecified c).
3. At each iteration l , calculate $\hat{\mu}^{(l)} = \sum w_i^{(l-1)} Y_i / \sum w_i^{(l-1)}$. Stop when the change in $\hat{\mu}^{(l)}$ is negligible from one iteration to the next.

An estimating equation for $\hat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\hat{\mu}$) to 0, obtaining $\sum_{i=1}^n \psi(Y_i - \hat{\mu}) = 0$. The simplest procedure for solving this estimating equation is by iteratively reweighted means. Defining the weight function as $w(E) \equiv \psi(E)/E$, the estimating equation becomes $\sum_{i=1}^n (Y_i - \hat{\mu}) w_i = 0$, from which $\hat{\mu} = \sum w_i Y_i / \sum w_i$. Starting with an initial estimate $\hat{\mu}^{(0)}$, initial weights are calculated, and the value of $\hat{\mu}$ is updated. This procedure continues iteratively until the value of $\hat{\mu}$ converges.

⁹Because the estimating equation for redescending M estimators, such as the bisquare, can have more than one root, the selection of an initial estimate might be consequential.

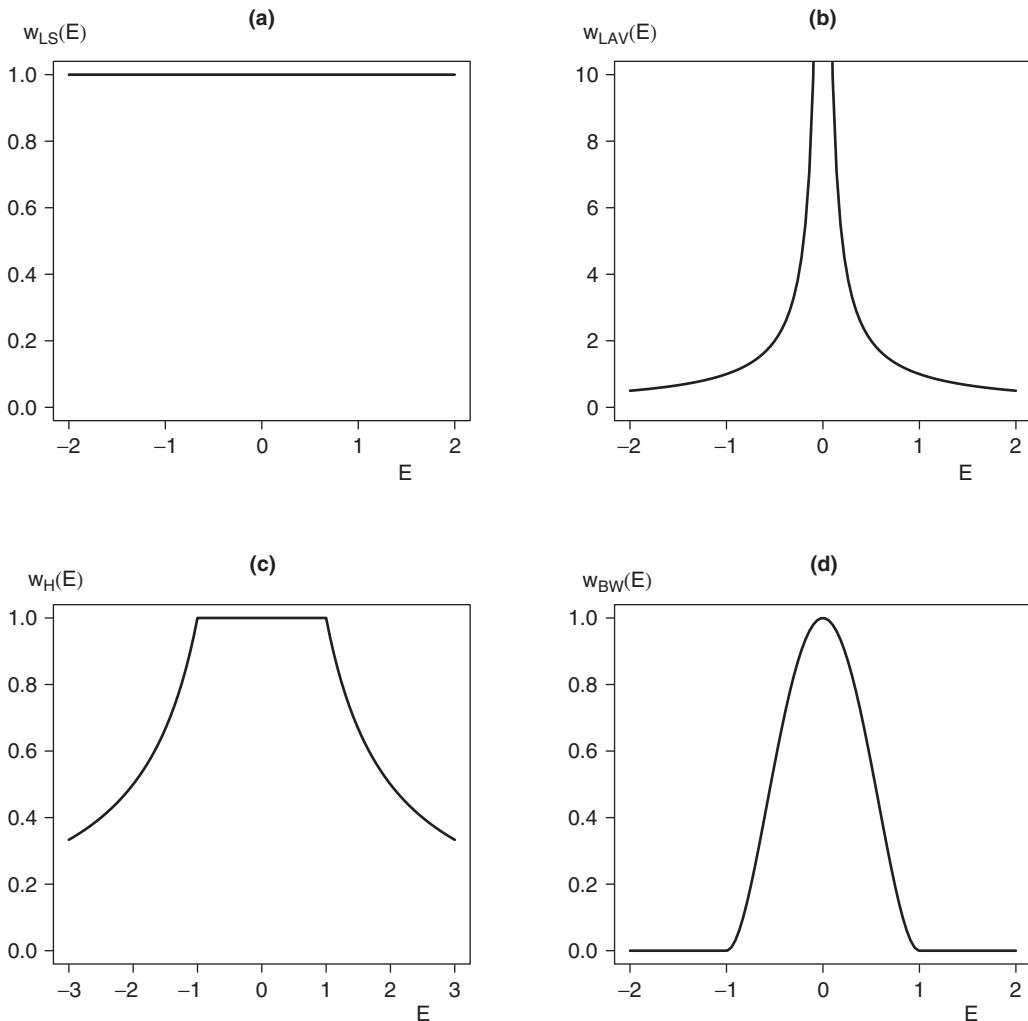


Figure 19.4 Weight functions $w(E)$ for the (a) least-squares, (b) least-absolute-values, (c) Huber, and (d) biweight estimators. The tuning constants for the Huber and biweight estimators are taken as $k = 1$. Note that the vertical axis in the graph for the LAV estimator and the horizontal axis in the graph for the Huber estimator are different from the others.

19.1.2 M Estimation in Regression

With the exception of one significant caveat, to be addressed in the next section, the generalization of M estimators to regression is immediate. We now wish to estimate the linear model

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \\ &= \underset{(1 \times k+1)}{\mathbf{x}'_i} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} + \varepsilon_i \end{aligned}$$

The estimated model is

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + \cdots + B_k X_{ik} + E_i \\ &= \mathbf{x}'_i \mathbf{b} + E_i \end{aligned}$$

The general M estimator minimizes the objective function

$$\sum_{i=1}^n \rho(E_i) = \sum_{i=1}^n \rho(Y_i - \mathbf{x}'_i \mathbf{b})$$

Differentiating the objective function and setting the derivative to $\mathbf{0}$ produces

$$\sum_{i=1}^n \psi(Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}_i = \mathbf{0} \quad (19.3)$$

which is a system of $k+1$ estimating equations in the $k+1$ elements of \mathbf{b} .

M estimation for the regression model $Y_i = \mathbf{x}'_i \beta + \varepsilon_i$ is a direct extension of M estimation of location: We seek to minimize an objective function of the regression residuals, $\sum_{i=1}^n \rho(E_i) = \sum_{i=1}^n \rho(Y_i - \mathbf{x}'_i \mathbf{b})$. Differentiating the objective function and setting the derivatives to 0 produces the estimating equations $\sum_{i=1}^n \psi(Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}_i = \mathbf{0}$.

Using the weight function $w(E) \equiv \psi(E)/E$ and letting $w_i \equiv w(E_i)$, the estimating equations become

$$\sum_{i=1}^n w_i (Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}_i = \mathbf{0}$$

The solution to these estimating equations minimizes the weighted sum of squares $\sum w_i E_i^2$.¹⁰ Because the weights depend on the residuals, the estimated coefficients depend on the weights, and the residuals depend on the estimated coefficients, an iterative solution is required. The IWLS algorithm for regression is as follows:

1. Select initial estimates $\mathbf{b}^{(0)}$ and set the iteration counter $l = 0$. Using the initial estimates, find residuals $E_i^{(0)} = Y_i - \mathbf{x}'_i \mathbf{b}^{(0)}$, and from these, calculate the estimated scale of the residuals $S^{(0)}$ and the weights $w_i^{(0)} = w(E_i^{(0)})$.
2. At each iteration l , solve the estimating equations using the current weights, minimizing $\sum w_i^{(l-1)} E_i^2$ to obtain $\mathbf{b}^{(l)}$. The solution is conveniently expressed as

$$\mathbf{b}^{(l)} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}$$

where the model matrix $\mathbf{X}_{(n \times k+1)}$ has \mathbf{x}'_i as its i th row, and $\mathbf{W}_{(n \times n)} \equiv \text{diag}\{w_i^{(l-1)}\}$.

Continue until $\mathbf{b}^{(l)} - \mathbf{b}^{(l-1)} \approx \mathbf{0}$.¹¹

¹⁰See the discussion of weighted-least-squares regression in Section 12.2.2.

¹¹As in the location problem, it is possible that the estimating equations for a redescending estimator have more than one root. If you use the bisquare estimator, for example, it is prudent to pick a good start value, such as provided by the Huber estimator.

Table 19.2 *M* Estimates for Duncan's Regression of Occupational Prestige on Income and Education for 45 U.S. Occupations

Estimator	Coefficient		
	Constant	Income	Education
Least squares	-6.065	0.5987	0.5458
Least squares*	-6.409	0.8674	0.3322
Least absolute values	-6.408	0.7477	0.4587
Huber	-7.111	0.7014	0.4854
Bisquare (biweight)	-7.412	0.7902	0.4186

NOTE: The estimator marked "Least squares*" omits ministers and railroad conductors.

Using the weight function, the estimating equations can be written as

$$\sum_{i=1}^n w_i(Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}'_i = \mathbf{0}$$

The solution of the estimating equations then follows by weighted least squares:

$$\mathbf{b} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}$$

where \mathbf{W} is the diagonal matrix of weights. The method of iterated weighted least squares starts with initial estimates $\mathbf{b}^{(0)}$, calculates initial residuals from these estimates, and calculates initial weights from the residuals. The weights are used to update the parameter estimates, and the procedure is iterated until it converges.

The asymptotic covariance matrix of the M estimator is given by

$$\mathcal{V}(\mathbf{b}) = \frac{E(\psi^2)}{[E(\psi')]^2} (\mathbf{X}' \mathbf{X})^{-1}$$

Using $\sum [\psi(E_i)]^2/n$ to estimate $E(\psi^2)$ and $[\sum \psi'(E_i)/n]^2$ to estimate $[E(\psi')]^2$ produces the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$. Research suggests, however, that these sampling variances are not to be trusted unless the sample size is large.¹²

To illustrate M estimation, recall Duncan's regression of occupational prestige on income and education. In our previous analysis of these data, we discovered two influential observations: *ministers* and *railroad conductors*.¹³ Another observation, *reporters*, has a relatively large residual but is not influential; still another observation, *railroad engineers*, is at a high-leverage point but is not discrepant. Table 19.2 summarizes the results of estimating Duncan's regression using four M estimators, including ordinary least squares. (The least-squares

¹²See Li (1985, pp. 300–301). For an alternative approach that may have better small-sample properties, see Street, Carroll, and Ruppert (1988). Also see the discussion of bootstrap methods in Chapter 21.

¹³See Chapter 11.

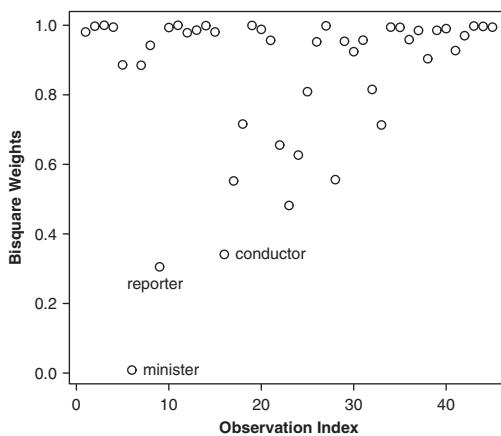


Figure 19.5 Final weights for the bisquare estimator applied to Duncan's regression of occupational prestige on income and education.

estimates obtained after deleting *ministers* and *railroad conductors* are also shown for comparison.) The three robust estimators produce quite similar results, with a larger income coefficient and smaller education coefficient than least squares. The redescending bisquare estimator is most different from least squares (and most similar to least squares after removing the two discrepant observations).

Figure 19.5 shows the final weights for the bisquare estimator applied to Duncan's data. *Railroad conductors*, *reporters*, and (especially) *ministers* have comparatively small weights, although some other occupations are down-weighted as well. Rather than simply regarding robust regression as a procedure for automatically down-weighting outliers, the method can often be used effectively, as here, to identify outlying observations.

19.2 Bounded-Influence Regression

The flies in the ointment of M estimation in regression are high-leverage outliers. In the location problem, M estimators such as the Huber and the bisquare bounded the influence of individual discrepant observations, but this is not the case in regression—if we admit the possibility of X -values with high leverage. High-leverage observations can force small residuals even when these observations depart from the pattern of the rest of the data.¹⁴

A key concept in assessing influence is the *breakdown point* of an estimator: The breakdown point is the fraction of arbitrarily “bad” data that the estimator can tolerate without being affected to an arbitrarily large extent. In the location problem, for example, the mean has a breakdown point of 0, because a *single* bad observation can change the mean by an arbitrary amount. The median, in contrast, has a breakdown point of 50%, because fully half the data can be bad without causing the median to become completely unstuck.¹⁵ It is disquieting that in regression analysis, *all* M estimators have breakdown points of 0.

¹⁴For an illustration of this phenomenon, see Exercise 19.3.

¹⁵See Exercise 19.2.

There are regression estimators, however, that have breakdown points of nearly 50%. One such *bounded-influence* estimator is *least-trimmed-squares* (LTS) regression.¹⁶

Return to the fitted regression model $Y_i = \mathbf{x}'_i \mathbf{b} + E_i$, ordering the squared residuals from smallest to largest:¹⁷ $(E^2)_{(1)}, (E^2)_{(2)}, \dots, (E^2)_{(n)}$. Then, select \mathbf{b} to minimize the sum of the smaller “half” of the squared residuals—that is,

$$\sum_{i=1}^m (E^2)_{(i)} \quad (19.4)$$

for $m = \lfloor (n+k+2)/2 \rfloor$ (where the “floor” brackets indicate rounding down to the next smallest integer).

The LTS criterion is easily stated, but the LTS estimate is not so easily computed. One approach is to consider all subsets of observations of size $k+1$ for which the vectors \mathbf{x}'_i are distinct. Let the $(k+1) \times (k+1)$ model matrix for a particular such subset be represented as \mathbf{X}^* . Because the rows of \mathbf{X}^* are all different, it is almost surely the case that the matrix \mathbf{X}^* is of full rank, and we can compute the regression coefficients for this subset as $\mathbf{b}^* = \mathbf{X}^{*-1} \mathbf{y}^*$ (where \mathbf{y}^* contains the corresponding entries of the response vector).¹⁸ For each such subset, we compute the LTS criterion in Equation 19.4 and take as the LTS estimator \mathbf{b}_{LTS} the value of \mathbf{b}^* that minimizes this criterion.

If there are no repeated rows in the model matrix \mathbf{X} , then the number of subsets of observations of size $k+1$ is

$$\binom{n}{k+1} = \frac{n!}{(n-k-1)!(k+1)!}$$

which is a very large number unless n is small. Even with highly efficient computational methods, it quickly becomes impractical, therefore, to find the LTS estimator by this approach. But we can compute a close approximation to \mathbf{b}_{LTS} by randomly sampling many (but not unmanageably many) subsets of observations and minimizing the LTS criterion over the sampled subsets.

In the case of Duncan’s occupational prestige regression, it is feasible to compute *all* subsets of size $k+1 = 3$ of the $n = 45$ observations, of which there are

$$\binom{45}{3} = \frac{45!}{42!3!} = 14,190$$

The LTS estimates, it turns out, are similar to the bisquare estimates given in the previous section (Table 19.2 on page 594):

$$\widehat{\text{Prestige}} = -5.764 + 0.8023 \times \text{Income} + 0.4098 \times \text{Education}$$

¹⁶The LTS estimator, the *MM* estimator introduced below, and other bounded-influence estimators in regression are described in detail by Rousseeuw and Leroy (1987).

¹⁷Lest the notation appear confusing, note that it is the *squared* residuals E_i^2 that are ordered from smallest to largest, *not* the residuals E_i themselves.

¹⁸See Exercise 19.4.

Unlike the M estimator of location, the M estimator in regression is vulnerable to high-leverage observations. Bounded-influence estimators limit the impact of high-leverage observations. One such bounded-influence estimator is LTS, which selects the regression coefficients to minimize the smaller “half” of the squared residuals, $\sum_{i=1}^m (E^2)_{(i)}$ (where $m = \lfloor (n + k + 2)/2 \rfloor$). The LTS estimator can be computed by calculating the regression coefficients for all subsets of observations of size $k + 1$ and selecting the regression coefficients from the subset that minimizes the LTS criterion. If there are too many such subsets, then a manageable number can be sampled randomly.

LTS and other bounded-influence estimators are not a panacea for linear-model estimation, because they can give unreasonable results for some data configurations.¹⁹ As well, the LTS estimator has much lower efficiency than the M estimators that we considered if the errors are in fact normal.

The latter problem can be addressed by combining bounded-influence estimation with M estimation, producing a so-called MM estimator, which retains the high breakdown point of the bounded-influence estimator and the high efficiency under normality of the M estimator. The MM estimator uses a bounded-influence estimator for start values in the computation of an M estimate and also to estimate the scale of the errors. For example, starting with the LTS estimator of the Duncan regression and following with the bisquare estimator yields the MM estimates

$$\widehat{\text{Prestige}} = -7.490 + 0.8391 \times \text{Income} + 0.3935 \times \text{Education}$$

The MM estimator combines the high breakdown point of bounded-influence regression with the high efficiency of M estimation for normally distributed errors. The MM estimator uses start values and a scale estimate obtained from a preliminary bounded-influence regression.

19.3 Quantile Regression

Quantile regression, due to Koenker and Bassett (1978), is a conceptually straightforward generalization of LAV regression. As I have explained, LAV regression estimates the conditional median (i.e., 50th percentile) of the response variable as a function of the explanatory variables. Quantile regression extends this approach to estimating other conditional quantiles of the response, such as the quartiles.

The LAV criterion in linear regression is written most directly as

$$\sum_{i=1}^n \rho_{\text{LAV}}(Y_i - \mathbf{x}'_i \mathbf{b}) \equiv \sum_{i=1}^n |Y_i - \mathbf{x}'_i \mathbf{b}|$$

The LAV estimator, \mathbf{b}_{LAV} , is the value of \mathbf{b} that minimizes this criterion. An equivalent expression, the motivation for which will become clear presently, is

¹⁹See Stefanski (1991).

$$\sum_{i=1}^n \rho_{\text{LAV}}(Y_i - \mathbf{x}'_i \mathbf{b}) = 0.5 \times \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) < 0} |Y_i - \mathbf{x}'_i \mathbf{b}| + 0.5 \times \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) > 0} |Y_i - \mathbf{x}'_i \mathbf{b}|$$

that is, the LAV criterion consists of two components: The first component includes observations producing negative residuals and the second, observations producing positive residuals; residuals in these two classes are weighted *equally*.

Koenker and Bassett show that estimating the conditional q quantile (where $0 < q < 1$) is equivalent to minimizing

$$\sum_{i=1}^n \rho_q(Y_i - \mathbf{x}'_i \mathbf{b}) = q \times \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) < 0} |Y_i - \mathbf{x}'_i \mathbf{b}| + (1 - q) \times \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) > 0} |Y_i - \mathbf{x}'_i \mathbf{b}|$$

(i.e., a sum of *differentially weighted* negative and positive residuals) and that, furthermore, finding the value $\mathbf{b} = \mathbf{b}_q$ that minimizes this criterion is a straightforward linear programming problem.²⁰ They proceed to derive the asymptotic covariance matrix of the estimated quantile regression coefficients as²¹

$$V(\mathbf{b}_q) = \sigma_q^2 (\mathbf{X}' \mathbf{X})^{-1}$$

where

$$\sigma_q^2 \equiv \frac{q(1-q)}{p[P^{-1}(q)]}$$

Here, $p(\cdot)$ is the probability density function for the error distribution, and $P^{-1}(\cdot)$ is the quantile function for the errors (supposing, as may not be the case, that the errors are identically distributed). Thus, $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.²² Note that σ_q^2 plays the same role as the error variance σ_e^2 does in the formula for the covariance matrix of the least-squares estimates.²³ In applications, σ_q^2 is estimated from the distribution of the residuals.

Quantile regression estimates a linear model for the conditional quantile q of the response variable by minimizing the criterion

$$\sum_{i=1}^n \rho_q(Y_i - \mathbf{x}'_i \mathbf{b}) = \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) < 0} q \times |Y_i - \mathbf{x}'_i \mathbf{b}| + \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) > 0} (1 - q) \times |Y_i - \mathbf{x}'_i \mathbf{b}|$$

The asymptotic covariance matrix of the quantile regression estimator \mathbf{b}_q is $V(\mathbf{b}_q) = \sigma_q^2 (\mathbf{X}' \mathbf{X})^{-1}$, where $\sigma_q^2 \equiv q(1-q)/\{p[P^{-1}(q)]\}$ and $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.

²⁰Linear programming is a common type of optimization problem, for which there are well-understood and efficient methods. See, for example, Gass (2003).

²¹Koenker and Bassett (1978) also give exact finite-sample results, but these are too computationally demanding to prove useful in practice. An alternative to using the asymptotic standard errors is to base inference for quantile regression on the bootstrap, as described in Chapter 21.

²²See the formula for the standard error of an order statistic given in Equation 3.4 (page 39).

²³See Section 9.3.1.

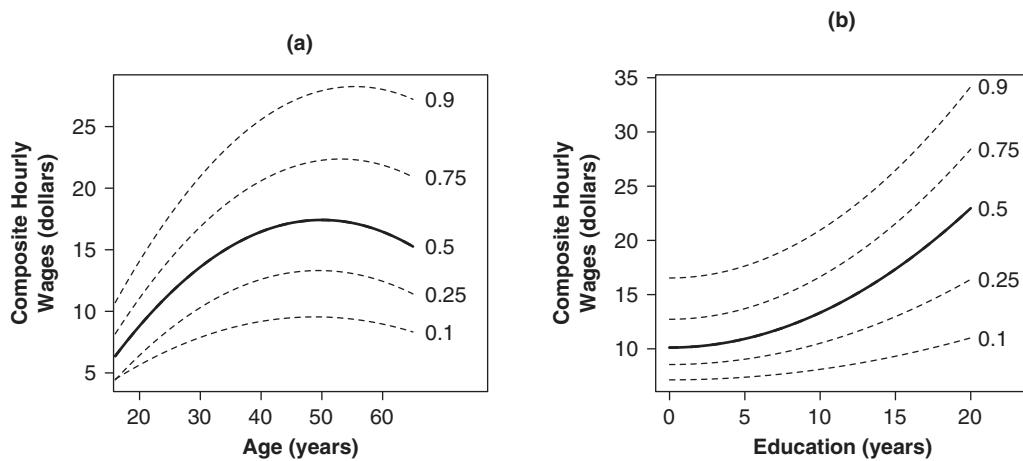


Figure 19.6 Effect displays for (a) age and (b) education in the quantile regression of wages on these variables and sex in the SLID data. In each case, the conditional .1, .25, .5, .75, and .9 quantiles are estimated. To construct each effect display, the other quantitative explanatory variable is set to its median value in the data, and the dummy regressor for sex is set to 0.5.

Figure 19.6 illustrates quantile regression by applying it to data from the Canadian Survey of Labour and Income Dynamics (SLID). I previously fit a model in which the log of the composite hourly wage rate for individuals in the sample was regressed on a dummy variable for sex, a quadratic in age, and the square of education.²⁴ For example, the estimated regression equation for the conditional median (i.e., the .5 quantile) is

$$\begin{aligned} \widehat{\text{Median Wages}} = & -13.44 + 3.066 \times \text{Male} + 0.9564 \times \text{Age} \\ & (0.69) \quad (0.216) \quad (0.0485) \\ & - 0.009567 \times \text{Age}^2 + 0.03213 \times \text{Education}^2 \\ & (0.000679) \quad (0.00157) \end{aligned}$$

Asymptotic standard errors are in parentheses below the coefficients. Note in Figure 19.6 how the regression quantiles spread apart at higher values of age and education, where the median level of wages is relatively high, and how, for the most part, the conditional distribution of wages is positively skewed, with the upper quantiles more spread out than the lower ones. These characteristics, recall, motivated the log transformation of wages in least-squares regressions for these data.

Quantile regression is an attractive method not only because of its robustness relative to least squares but also because of its simple interpretation and because of its focus on the *whole*

²⁴See Section 12.3. There is, however, a subtle change here: We were careful on log-transforming wages to make sure that the form in which age and education entered the model adequately captured the dependence of the conditional mean response on these explanatory variables. Because the log transformation is not linear, a quadratic in age and the square of education may not be appropriate for the conditional median of *untransformed* wages. I could, of course, compute instead the quantile regression for *log* wages (and I invite the reader to do so), but I wanted to illustrate how quantile regression can reveal asymmetry and nonconstant spread in the conditional distribution of the response.

conditional distribution of the response. Moreover, quantile regression extends naturally beyond linear models, for example, to nonlinear regression and to nonparametric regression.²⁵

19.4 Robust Estimation of Generalized Linear Models

The maximum-likelihood or quasi-likelihood estimating equations for a generalized linear model can be written as

$$\sum_{i=1}^n \frac{1}{a_i} (Y_i - \mu_i) \begin{pmatrix} \mathbf{x}_i \\ (k+1 \times 1) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ (k+1 \times 1) \end{pmatrix} \quad (19.5)$$

where Y_i is the response variable for the i th of n observations, $\mu_i = g^{-1}(\mathbf{x}'_i \boldsymbol{\beta})$ is the conditional expectation of the response given the values of the regressors \mathbf{x}'_i for observation i , $\boldsymbol{\beta}$ is the parameter vector of $k + 1$ regression coefficients to be estimated, and $g^{-1}(\cdot)$ is the inverse of the link function (i.e., the mean function) for the model. The constants a_i depend on the distributional family used to model the response; for example, for the Gaussian family $a_i = 1$, while for the binomial family $a_i = 1/n_i$ (the inverse of the number of binomial trials).²⁶ Because it depends directly on the difference between the observed and fitted response, the maximum-likelihood or quasi-likelihood estimator based on these estimating equations is generally not robust.

Cantoni and Ronchetti (2001) suggest replacing Equation 19.5 by estimating equations of the form

$$\sum_{i=1}^n \psi(Y_i, \mu_i) = \mathbf{0} \quad (19.6)$$

where $\psi(\cdot)$ is selected to produce high resistance to outliers. Equation 19.6 is a generalization of the estimating equations for M estimators in linear regression (Equation 19.3 on page 593). Bounded influence is achieved by down-weighting high-leverage observations: The weights employed are the product of (1) weights measuring the discrepancy between the observed and fitted response and (2) weights accounting for the leverage of the observations.²⁷ The details are beyond the scope of this presentation and are developed in Cantoni and Ronchetti's (2001) study.

Because the response variable in a binomial generalized linear model is bounded by 0 and 1, it is rare (but not impossible) to find highly influential observations in a logit or probit regression. GLMs for count data and for non-normal continuous responses are another matter, and robust estimation for these models is potentially useful. My limited experience with Cantoni

²⁵See Koenker (2005).

²⁶See Section 15.3.

²⁷A simple choice of leverage-based weights is $\sqrt{1 - h_i}$, where h_i is the hat-value for the i th observation (see Section 11.2). A higher breakdown point can be achieved, however, by using a robust covariance matrix for the X s to judge the unusualness of observations in the X -space; using a robust covariance matrix is not sensible when the model matrix includes dummy regressors or other contrasts. Applied to linear regression, bounded-influence estimators using weights based on the product of leverage and discrepancy are called *GM* (generalized-*M*) estimators (see Rousseeuw & Leroy, 1987, Chapter 1).

and Ronchetti's estimator, however, suggests that it is not entirely reliable for detecting and discounting influential data.²⁸

Robust bounded-influence estimators for generalized linear models can be obtained by replacing the usual maximum-likelihood or quasi-likelihood estimating equations for GLMs by $\sum_{i=1}^n \psi(Y_i, \mu_i) = \mathbf{0}$, where $\psi(\cdot)$ is selected to produce high resistance to outliers. Bounded influence is achieved by down-weighting observations that have large residuals or large leverage.

19.5 Concluding Remarks

A final caution concerning robust regression: Robust estimation is not a substitute for close examination of the data. Although robust estimators can cope with heavy-tailed error distributions and outliers, they cannot correct nonlinearity, for example. Indeed, one use of robust estimation is to employ it as a routine diagnostic for unusual data in small- to medium-size samples, comparing the results obtained for a robust estimator with those of least-squares regression and investigating when the two estimators produce substantially different estimates (see, e.g., the discussion of the final weights for the Duncan regression displayed in Figure 19.5 on page 595).

As I have pointed out with respect to quantile regression, robust estimators can be extended to other settings. For example, it is a simple matter, and indeed common, to employ M estimator "robustness weights" in local-polynomial nonparametric regression, multiplying these weights by the neighborhood weights for the usual local-polynomial estimator, thereby rendering the local-polynomial estimator resistant to outliers.²⁹

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 19.1. *Prove that the median minimizes the least-absolute-values objective function:

$$\sum_{i=1}^n \rho_{\text{LAV}}(E_i) = \sum_{i=1}^n |Y_i - \hat{\mu}|$$

Exercise 19.2. Breakdown: Consider the contrived data set

$$\begin{aligned} Y_1 &= -0.068 & Y_2 &= -1.282 & Y_3 &= 0.013 & Y_4 &= 0.141 \\ Y_5 &= -0.980 \end{aligned}$$

²⁸See, for example, Exercise 19.5.

²⁹See Chapter 18.

(an adaptation of the data used to construct Figure 19.1). Show that more than two values must be changed to influence the median of the five values to an arbitrary degree. (Try, e.g., to make the first two values progressively and simultaneously larger, graphing the median of the altered data set against the common value of Y_1 and Y_2 ; then, do the same for the first three observations.)

Exercise 19.3. The following contrived data set (discussed in Chapter 3) is from Anscombe (1973):

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

- (a) Graph the data and confirm that the third observation is an outlier. Find the least-squares regression of Y on X , and plot the least-squares line on the graph.
- (b) Fit a robust regression to the data using the bisquare or Huber M estimator. Plot the fitted regression line on the graph. Is the robust regression affected by the outlier?
- (c) Omitting the third observation $\{13, 12.74\}$, the line through the rest of the data has the equation $Y = 4 + 0.345X$, and the residual of the third observation from this line is 4.24. (Verify these facts.) Generate equally discrepant observations at X -values of 23 and 33 by substituting these values successively into the equation $Y = 4 + 0.345X + 4.24$. Call the resulting Y values Y'_3 and Y''_3 . Redo parts (a) and (b), replacing the third observation with the point $\{23, Y'_3\}$. Then, replace the third observation with the point $\{33, Y''_3\}$. What happens?
- (d) Repeat part (c) using the LTS bounded-influence estimator. Do it again with the MM estimator.

Exercise 19.4. Computing the LTS estimator: Why is it almost surely the case that the $(k+1) \times (k+1)$ matrix \mathbf{X}^* , with rows selected from among those of the complete model matrix \mathbf{X} , is of full rank when all its rows are different? (Put another way, how is it possible that \mathbf{X}^* would *not* be of full rank?) Thinking in terms of the $(k+1)$ -dimensional scatterplot of Y against X_1, \dots, X_k , what does the hyperplane defined by $\mathbf{b}^* = \mathbf{X}^{*-1}\mathbf{y}^*$ represent?

Exercise 19.5. In Chapter 15, I fit a Poisson regression of number of interlocks on assets, nation of control, and sector for Ornstein's Canadian interlocking-directorate data. The results from this regression are given in Table 15.3 (page 428). Influential-data diagnostics (see, e.g.,

Figure 15.7 on page 456) suggest that the first observation in the data set is quite influential; in particular, the coefficient of assets changes considerably when the first observation is removed. Perform a robust Poisson regression for this model. How do the results compare to removing the first observation from the data set? (Recall, however, that the influence of the first observation depends on unmodeled nonlinearity in the relationship between interlocks and assets—a problem that I ultimately addressed in Chapter 15 by log-transforming assets.)

Summary

- Robust M estimators of location, for the parameter μ in the simple model $Y_i = \mu + \varepsilon_i$, minimize the objective function

$$\sum_{i=1}^n \rho(E_i) = \sum_{i=1}^n \rho(Y_i - \hat{\mu})$$

selecting $\rho(\cdot)$ so that the estimator is relatively unaffected by outlying values. Two common choices of objective function are the Huber and the biweight (or bisquare).

- The sensitivity of an M estimator to individual observations is expressed by the influence function of the estimator, which has the same shape as the derivative of the objective function, $\psi(E) \equiv \rho'(E)$.
- An estimating equation for $\hat{\mu}$ is obtained by setting the derivative of the objective function (with respect to $\hat{\mu}$) to 0, obtaining $\sum_{i=1}^n \psi(Y_i - \hat{\mu}) = 0$. The simplest procedure for solving this estimating equation is by iteratively reweighted means. Defining the weight function as $w(E) \equiv \psi(E)/E$, the estimating equation becomes $\sum_{i=1}^n (Y_i - \hat{\mu})w_i = 0$, from which $\hat{\mu} = \sum w_i Y_i / \sum w_i$. Starting with an initial estimate $\hat{\mu}^{(0)}$, initial weights are calculated, and the value of $\hat{\mu}$ is updated. This procedure continues iteratively until the value of $\hat{\mu}$ converges.
- M estimation for the regression model $Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$ is a direct extension of M estimation of location: We seek to minimize an objective function of the regression residuals:

$$\sum_{i=1}^n \rho(E_i) = \sum_{i=1}^n \rho(Y_i - \mathbf{x}'_i \mathbf{b})$$

Differentiating the objective function and setting the derivatives to 0 produces the estimating equations

$$\sum_{i=1}^n \psi(Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}'_i = \mathbf{0}$$

- Using the weight function, the estimating equations can be written as

$$\sum_{i=1}^n w_i (Y_i - \mathbf{x}'_i \mathbf{b}) \mathbf{x}'_i = \mathbf{0}$$

The solution of the estimating equations then follows by weighted least squares:

$$\mathbf{b} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}$$

where \mathbf{W} is the diagonal matrix of weights. The method of iterated weighted least squares starts with initial estimates $\mathbf{b}^{(0)}$, calculates initial residuals from these estimates, and calculates initial weights from the residuals. The weights are used to update the parameter estimates, and the procedure is iterated until it converges.

- Unlike the M estimator of location, the M estimator in regression is vulnerable to high-leverage observations. Bounded-influence estimators limit the effect of high-leverage observations. One such bounded-influence estimator is LTS, which selects the regression coefficients to minimize the smaller “half” of the squared residuals $\sum_{i=1}^m (E^2)_{(i)}$ (where $m = \lfloor (n + k + 2)/2 \rfloor$). The LTS estimator can be computed by calculating the regression coefficients for all subsets of observations of size $k + 1$ and selecting the regression coefficients from the subset that minimizes the LTS criterion. If there are too many such subsets, then a manageable number can be sampled randomly.
- The MM estimator combines the high breakdown point of bounded-influence regression with the high efficiency of M estimation for normally distributed errors. The MM estimator uses start values and a scale estimate obtained from a preliminary bounded-influence regression.
- Quantile regression estimates a linear model for the conditional quantile q of the response variable by minimizing the criterion

$$\sum_{i=1}^n \rho_q(Y_i - \mathbf{x}'_i \mathbf{b}) = \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) < 0} q \times |Y_i - \mathbf{x}'_i \mathbf{b}| + \sum_{i: (Y_i - \mathbf{x}'_i \mathbf{b}) > 0} (1 - q) \times |Y_i - \mathbf{x}'_i \mathbf{b}|$$

The asymptotic covariance matrix of the quantile regression estimator \mathbf{b}_q is $V(\mathbf{b}_q) = \sigma_q^2 (\mathbf{X}' \mathbf{X})^{-1}$, where $\sigma_q^2 \equiv q(1 - q)/\{p[P^{-1}(q)]\}$, and $p[P^{-1}(q)]$ is the density at the q quantile of the error distribution.

- Robust bounded-influence estimators for generalized linear models can be obtained by replacing the usual maximum-likelihood or quasi-likelihood estimating equations for GLMs by $\sum_{i=1}^n \psi(Y_i, \mu_i) = \mathbf{0}$, where $\psi(\cdot)$ is selected to produce high resistance to outliers. Bounded influence is achieved by down-weighting observations that have large residuals or large leverage.

Recommended Reading

- In a volume on robust and exploratory methods, edited by Hoaglin, Mosteller, and Tukey (1983), Goodall (1983) presents a high-quality, readable treatment of M estimators of location.
- A fine chapter by Li on M estimators for regression appears in a companion volume (Hoaglin, Mosteller, & Tukey, 1985).
- Another good source on M estimators is Wu (1985).
- Rousseeuw and Leroy’s (1987) book on robust regression and outlier detection emphasizes bounded-influence, high-breakdown estimators.
- Andersen (2007) presents a broad and largely accessible overview of methods of robust regression, including a discussion of robust estimation for generalized linear models.
- Koenker (2005) offers an extensive treatment of quantile regression by the originator of the method.

20

Missing Data in Regression Models

Missing data are a regrettably common feature of data sets in the social sciences. Despite this fact, almost all statistical methods in widespread use, including the methods introduced in the previous chapters of this book, assume that the data in hand are complete.

The current chapter provides a basic introduction to modern methods for handling missing data. The first section of the chapter draws some basic distinctions concerning the processes that generate missing data. The second section briefly describes traditional methods for coping with missing data and explains why they are problematic. The third section shows how the method of maximum likelihood (ML) can be used to estimate the parameters of statistical models in the presence of missing data. The fourth section introduces multiple imputation of missing data—a general, flexible, and convenient method for dealing with missing data that can perform well in certain circumstances. The final section of the chapter introduces methods for handling selection bias and censored data, which are special kinds of missing data.

Data may be missing for a variety of reasons:

- In survey research, for example, certain respondents may be unreachable or may refuse to participate in the survey, giving rise to *global* or *unit nonresponse*.
- Alternatively, again in survey research, some respondents may not know the answers to specific questions or may refuse to respond to them, giving rise to *item nonresponse*.
- Missing data may also be produced by errors in data collection—as when an interviewer fails to ask a question of a survey respondent—or in data processing.
- In some cases, missing data are built into the design of a study, as when particular questions in a survey are asked only of a random subset of respondents.
- It is sometimes the case that data values in a study are *censored*. The most common example of censored data occurs in *survival analysis* (also called *event-history analysis*, *duration analysis*, or *failure-time analysis*), which concerns the timing of events. In a prototypical biomedical application, subjects in a clinical trial are followed for a fixed period of time, and their survival times are recorded at their deaths. Some subjects, however, happily live beyond the termination of the study, and their survival times are therefore censored. Survival analysis is beyond the scope of this book,¹ but censored data can occur in other contexts as well—as, for example, in an exam with a fixed number of questions where it is not possible to score fewer than 0 nor more than the total number of questions correct, no matter how little or much an individual knows.

¹There are many texts on survival analysis. For example, see Allison (2014) for a brief introduction to survival analysis or Hosmer and Lemeshow (1999) for a more extensive treatment.

Missing data, in the sense that is developed in this chapter, should be distinguished from data that are *conditionally undefined*. A survey respondent who has no children, for example, cannot report their ages. Conditionally undefined data do not threaten the representativeness of a sample as truly missing data do. Sometimes, however, the distinction between missing and conditionally undefined data is not entirely clear-cut: Voters in a postelection survey who did not vote cannot be asked for whom they voted, but they could be (and may not have been) asked whether and for whom they had a preference. Similarly, some respondents asked to state an opinion on an issue may not have an opinion. Are these data missing or simply nonexistent?

It is important to realize at the outset that there is no magic cure for missing data, and it is generally impossible to proceed in a principled manner without making at least partly unverifiable assumptions about the process that gives rise to the missing information. As King Lear said, “Nothing will come of nothing” (although he applied this insight unwisely).

20.1 Missing Data Basics

Rubin (1976) introduced some key distinctions concerning missing data.² Let the matrix $\mathbf{X}_{(n \times p)}$ represent the complete data for a sample of n observations on p variables.³ Some of the entries of \mathbf{X} , denoted by \mathbf{X}_{mis} , are missing, and the remaining entries, \mathbf{X}_{obs} , are observed.⁴

- Missing data are said to be *missing completely at random (MCAR)* if the missing data (and hence the observed data) can be regarded as a simple random sample of the complete data. Put alternatively, the probability that a data value is missing, termed *missingness*, is unrelated to the data value itself or to any other value, missing or observed, in the data set.
- If, however, missingness is related to the observed data but—conditioning on the observed data—not to the missing data, then missing data are said to be *missing at random (MAR)*. In a survey, for example, certain individuals may refuse to report their income, and these people may even differ systematically in income from the sample as a whole. Nevertheless, if the observations are independently sampled, so that one respondent’s decision to withhold information about income is independent of others’ responses, and if, *conditional on* the information that the respondent does provide (e.g., education, occupation), failure to provide information on income is independent of income itself, then the data are MAR. MCAR is a stronger condition—and a special case—of MAR.
- Finally, if missingness is related to the missing values themselves—that is, if the probability that a data value is missing depends on missing data (including, and indeed usually, the data value itself), even when the information in the observed data is taken into account—then missing data are said to be *missing not at random (MNAR)*. For example, if conditional on all the observed data, individuals with higher incomes are more likely than others to withhold information about their incomes, then the missing income data are MNAR.

²Although Rubin’s terminology is potentially confusing, it is in common use and has guided most subsequent work on missing data by statisticians. It would therefore be a mistake, I think, to introduce different terms for these concepts.

³If you are unfamiliar with matrix notation, simply think of the matrix \mathbf{X} as a rectangular table of data, with the observations given by the n rows of the table and the variables by the p columns.

⁴Despite the notation, \mathbf{X}_{mis} and \mathbf{X}_{obs} are not really matrices; they are, rather, subsets of the complete data matrix \mathbf{X} . Together, \mathbf{X}_{mis} and \mathbf{X}_{obs} comprise \mathbf{X} .

These distinctions are important because they affect the manner in which missing data can be properly handled. In particular, if the data are MCAR or MAR, then it is not necessary to model the process that generates the missing data to accommodate the missing data. When data are MCAR or MAR, the “mechanism” that produces the missing data is therefore *ignorable*. In contrast, when data are MNAR, the missing-data mechanism is *nonignorable*, and it becomes necessary to model this mechanism to deal with the missing data in a valid manner.

Except in some special situations, it is not possible to know whether data are MCAR, MAR, or MNAR. We may be able to show that missingness on some variable in a data set is related to observed data on one or more other variables, in which case we can rule out MCAR, but the converse is not the case—that is, demonstrating that missingness in a variable is not related to observed data in other variables does not *prove* that the missing data are MCAR (because, e.g., nonrespondents in a survey may be differentiated from respondents in some *unobserved* manner). If, on the other hand, a survey question is asked of a random subset of respondents, then data are MCAR by design of the study.

Missing data are missing completely at random (MCAR) if the missing data can be regarded as a simple random sample of the complete data. If missingness is related to the observed data but not to the missing data (conditional on the observed data), then data are missing at random (MAR). If missingness is related to the missing values themselves, even when the information in the observed data is taken into account, then data are missing not at random (MNAR). When data are MCAR or MAR, the process that produces missing data is ignorable, in the sense that valid methods exist to deal with the missing data without explicitly modeling the process that generates them. In contrast, when data are MNAR, the process producing missing data is nonignorable and must be modeled. Except in special situations, it is not possible to know whether data are MCAR, MAR, or MNAR.

20.1.1 An Illustration

To clarify these distinctions, let us consider the following example (adapted from Little & Rubin, 1990): We have a data set with $n = 250$ observations and two variables. The first variable, X_1 , is completely observed, but some of the observations on X_2 are missing. This pattern—where one variable has missing data and all others (in this instance, *one* other variable) are completely observed—is called *univariate missing data*. Univariate missing data are especially easy to handle. For example, while general patterns of missing data may require iterative techniques (as described later in this chapter), univariate missing data do not. Nevertheless, we will get a great deal of mileage out of this simple example.

For concreteness, suppose that the complete data are sampled from a bivariate-normal distribution with means $\mu_1 = 10$, $\mu_2 = 20$, variances $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, and covariance $\sigma_{12} = 8$.⁵ The population correlation between X_1 and X_2 is therefore $\rho_{12} = 8/\sqrt{9 \times 16} = 2/3$; the slope for the regression of X_1 on X_2 is $\beta_{12} = 8/16 = 1/2$, and the slope for the regression of X_2 on X_1 is $\beta_{21} = 8/9 \approx 0.889$.

⁵The bivariate-normal distribution is described in online Appendix D on probability and estimation.

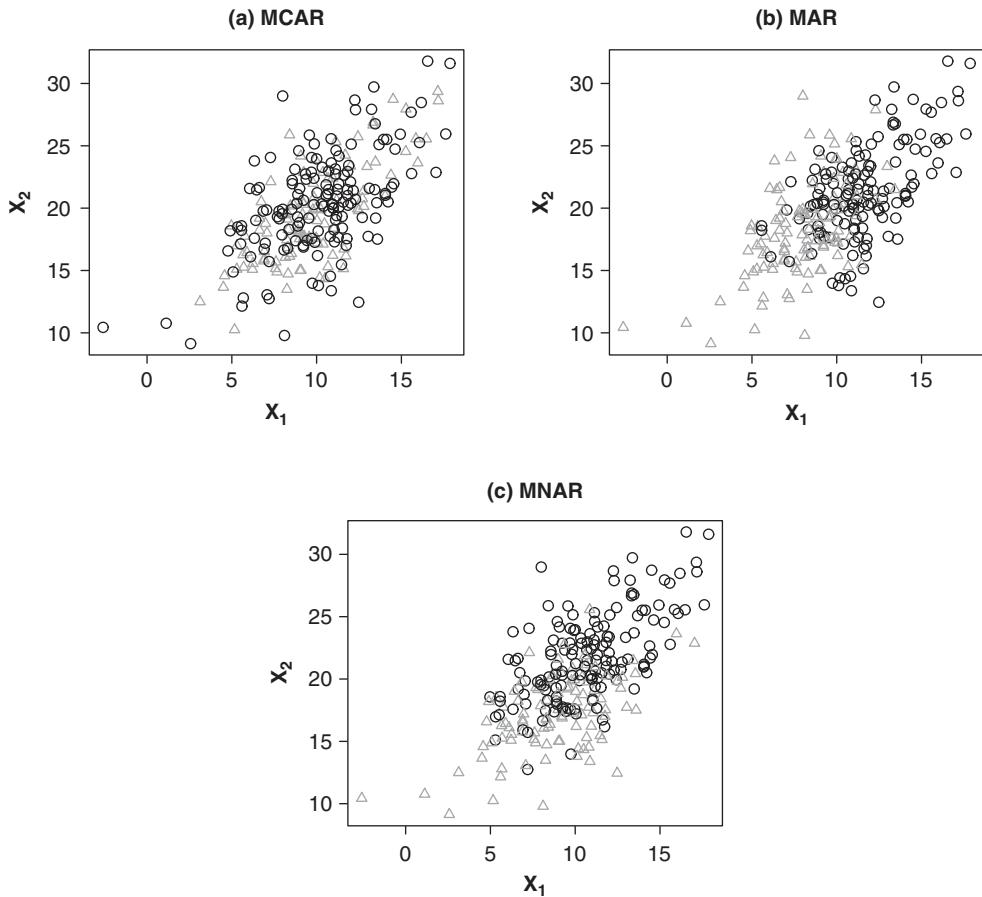


Figure 20.1 The 250 observations in each scatterplot were sampled from a bivariate-normal distribution; in each case, the observations shown as gray triangles have missing data on X_2 . In panel (a), the 100 observations with missing data were sampled at random, and the missing data on X_2 are therefore missing completely at random (MCAR). In (b), the probability that an observation has a missing value on X_2 is related to its value on X_1 , and so the missing data on X_2 are missing at random (MAR). In (c), the probability that an observation has a missing value on X_2 is related to its value on X_2 , and so the missing data on X_2 are missing not at random (MNAR).

Consider the following three mechanisms for generating missing data in the sample of 250 observations:

1. One hundred of the observations on X_2 are selected at random and set to missing. This situation is illustrated in Figure 20.1(a), where the data points represented by black circles are fully observed, and those represented by gray triangles are missing X_2 . Here, the missing values of X_2 are MCAR, and the subset of valid observations is a simple random sample of the full data set.

2. In Figure 20.1(b), an observation’s missingness on X_2 is related to its (observed) value of X_1 :

$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{2}{3}(X_{i1} - 10)\right]} \quad (20.1)$$

We recognize Equation 20.1 as a logistic-regression equation,⁶ with the probability that X_2 is missing declining as X_1 grows larger. The regression coefficients were calibrated so that approximately 100 observations will have missing data on X_2 (and for the sample in Figure 20.1(b), there are, as it turned out, 109 missing values produced by simulating the missing-data-generating process). X_1 and X_2 are positively correlated, and consequently, there are relatively fewer small values of X_2 in the observed data than in the complete data; moreover, if we look only at the observations with valid data on *both* X_1 and X_2 , this subset of observations also has relatively few small values of X_1 . Because X_1 , recall, is fully observed, the missing data on X_2 are MAR.

3. In Figure 20.1(c), an observation’s missingness on X_2 is related to the (potentially unobserved) value of X_2 itself:

$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{1}{2}(X_{i2} - 20)\right]} \quad (20.2)$$

For our data set, the simulation of this process produced exactly 100 observations with missing data on X_2 . Here, too, and indeed more directly, there are relatively few small values of X_2 (and, incidentally, if we exclude the observations with missing data on X_2 , of X_1 also). Because missingness on X_2 depends directly on the value of X_2 , the missing data are MNAR.

As mentioned, except in those relatively rare instances where missing data are built into the design of a study, it is not possible to verify from the data whether they are MCAR or even MAR—that is, whether the missing-data mechanism is ignorable. Indeed, it is fair to say that missing data are almost always MNAR. Nevertheless, if we can argue plausibly that the departure from MAR is likely small, then dealing with missing data becomes a much more tractable problem. Furthermore, unless we are willing to discard the data, we have to proceed in *some* manner. Rather than requiring perfection, which is probably unattainable, we may have to settle for a solution that simply gets us closer to the truth.

20.2 Traditional Approaches to Missing Data

In evaluating missing-data methods, there are three general questions to answer:

1. Does the method provide *consistent estimates* of population parameters, or does it introduce systematic biases into the results?
2. Does the method provide *valid statistical inferences*, or are confidence intervals and p -values distorted?

⁶See Section 14.1.

3. Does the method use the observed data *efficiently* or does it profligately discard information?

The answers to these questions depend partly on the methods themselves, partly on the nature of the process generating missing data, and partly on the statistics of interest.

There are many ad hoc methods that have been proposed for dealing with missing data; I will briefly describe several of the most common here and will explain why, in which respects, and under what circumstances they are problematic. This discussion is far from complete, however: For example, I have omitted discussion of methods based on reweighting the data.⁷

Complete-case analysis (also called *listwise* or *casewise deletion* of missing data), probably the most widely used approach, simply ignores observations with any missing data on the variables included in the analysis. Complete-case analysis has its advantages: It is simple to implement, provides consistent estimates and valid inferences when the data are missing completely at random, and provides consistent estimates of regression coefficients and valid inferences when missingness on all the variables in a regression does not depend on the response variable (even if data are not MCAR). Because it discards some valid data, however, complete-case analysis generally does not use the information in the data efficiently. This problem can become acute when there are many variables, each with some missing data. For example, suppose each of 10 variables is missing 5% of observations and that missingness in different variables is independent.⁸ Then, we would expect only $100 \times .95^{10} \approx 60\%$ of the observations to be completely observed. Furthermore, when data are MAR or MNAR, complete-case analysis usually provides biased results and invalid inferences.

Available-case analysis (also called *pairwise deletion* of missing data) uses all nonmissing observations to compute each statistic of interest. In a least-squares regression analysis, for example, the regression coefficients can be calculated from the means, variances, and covariances of the variables (or, equivalently, from their means, variances, and correlations). To apply available-case analysis to least-squares regression, each mean and variance is calculated from all observations with valid data for a variable and the covariance of two variables from all observations that have valid data for both.⁹ Available case analysis appears to use more information than complete-case analysis, but in certain instances, this is an illusion: That is, estimators based on available cases can be *less* efficient than those based on complete cases.¹⁰ Moreover, by basing different statistics on different subsets of the data, available-case analysis can lead to nonsensical results, such as covariances that are inconsistent with one another or correlations outside the range from -1 to $+1$.¹¹ Finally, except in simple applications, such as linear least-squares regression, it is not obvious how to apply the available-case approach.

⁷As a general matter, relatively simple weighting schemes can reduce bias in estimates but do not provide valid inferences. See, for example, Little and Rubin (2002, Section 3.3).

⁸This is not a generally realistic condition: Missingness on different variables is probably positively associated, producing a result not quite as dismal as the one described here. The general point is valid, however: With many variables subject to missing data, there are typically many fewer complete cases than valid observations on individual variables.

⁹This description is slightly ambiguous: In computing the covariance, for example, do we use the means for each variable computed from all valid data for that variable or (as is more common and as I have done in the example reported below) recompute the means for each pair using observations with valid data for both variables in the pair?

¹⁰An example is estimating the difference between the means of two highly correlated variables (as in a paired *t*-test): See Little and Rubin (1990, pp. 378–380).

¹¹See Exercise 20.1.

Several methods attempt to fill in missing data, replacing missing values with plausible *imputed* values. The resulting completed data set is then analyzed using standard methods. One such approach, termed *unconditional mean imputation* (or *mean substitution*) replaces each missing value with the mean of the observed data for the variable in question. Although mean imputation preserves the means of variables, it makes their distributions less variable and tends to weaken relationships between variables. One consequence is that mean imputation generally yields biased regression coefficients and invalid inferences even when data are MCAR. In addition, by treating the missing data as if they were observed, mean imputation exaggerates the effective size of the data set, further distorting statistical inference—a deficiency that it shares with other simple imputation methods.

A more sophisticated approach, called *conditional-mean imputation*, replaces missing data with predicted values obtained, for example, from a regression equation (in which case the method is also called *regression imputation*). Using available data, we regress each variable with missing data on other variables in the data set; the resulting regression equation is used to produce predicted values that replace the missing data.¹² A problem with regression imputation is that the imputed observations tend to be less variable than real data because they lack residual variation; another problem is that we have failed to account for uncertainty in the estimation of the regression coefficients used to obtain the imputed values. The first of these problems can be addressed, for example, by adding a randomly sampled residual to each filled-in value. The second problem leads naturally to Bayesian multiple imputation of missing values, described below.¹³ Regression imputation improves on unconditional mean imputation, but it is far from a perfect technique, generally providing biased estimates and invalid inferences even for missing data that are MCAR.

I applied several methods of handling missing data to the artificial data sets graphed in Figure 20.1 (page 608) and described in the preceding section. The results are shown in Table 20.1. Recall that the data for this example were sampled from a bivariate-normal distribution (with parameters shown at the top of the table). Statistics for the complete data set of $n = 250$ observations are also shown (near the top of the table). Some of the results—for example, the equivalence of complete-case analysis, available-case analysis, and mean imputation for the slope coefficient B_{12} of the regression of X_1 (the completely observed variable) on X_2 —are peculiar to univariate missing data.¹⁴ Other characteristics are more general, such as the reasonable results produced by complete-case analysis when missingness does not depend on the response variable (i.e., for the coefficient β_{12} when data are MCAR or, for this example, MNAR, and for the coefficient β_{21} when, again for this example, data are MAR). Note that ML estimation and multiple imputation are the only methods that provide uniformly good results for *all* parameters in *both* the MCAR and MAR data sets.

To illustrate further the properties of the various missing-data methods, I conducted a small simulation study, drawing 1000 samples from the bivariate-normal distribution described above, producing from each sample a data set in which missing data were MAR, and applying complete-case analysis, unconditional-mean imputation, regression imputation, and Bayesian

¹²Because the predictor variables in each of these auxiliary regressions may themselves have missing data, the implementation of regression imputation can be complicated, requiring us to fit different regression equations for different patterns of missing information. The basic idea, however, is straightforward.

¹³See Section 20.4.

¹⁴See Exercise 20.2.

Table 20.1 Parameter Estimates Obtained by Several Methods of Handling Missing Data Under Different Conditions

	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}	ρ_{12}	β_{12}	β_{21}
Parameter	10.000	20.000	9.000	16.000	8.000	.667	0.500	0.889
<i>Complete data (n = 250)</i>								
Estimates	10.002	19.976	9.432	16.731	8.114	.646	0.485	0.860
<i>MCAR data set</i>								
Complete cases	10.210	20.400	9.768	17.114	7.673	.593	0.448	0.785
Available cases	10.002	20.400	9.432	17.114	7.673	.604	0.448	0.813
Mean imputation	10.002	20.400	9.432	10.241	4.591	.467	0.448	0.487
Regression imputation	10.002	20.237	9.432	12.454	7.409	.683	0.595	0.785
Maximum likelihood	10.002	20.237	9.394	16.809	7.379	.587	0.439	0.785
Multiple imputation	10.002	20.269	9.432	16.754	7.415	.590	0.443	0.786
<i>MAR data set</i>								
Complete cases	11.615	21.349	6.291	14.247	5.456	.576	0.383	0.867
Available cases	10.002	21.349	9.432	14.247	5.456	.508	0.383	0.578
Mean imputation	10.002	21.385	9.432	8.010	3.068	.353	0.383	0.325
Regression imputation	10.002	19.950	9.432	12.443	8.179	.755	0.657	0.867
Maximum likelihood	10.002	20.000	9.394	17.044	8.103	.640	0.475	0.863
Multiple imputation	10.002	19.914	9.432	17.493	8.342	.649	0.477	0.884
<i>MNAR data set</i>								
Complete cases	10.811	21.833	8.238	12.823	6.389	.622	0.498	0.776
Available cases	10.002	21.833	9.432	12.823	6.389	.581	0.498	0.677
Mean imputation	10.002	21.833	9.432	7.673	3.823	.449	0.498	0.405
Regression imputation	10.002	21.206	9.432	10.381	7.315	.739	0.705	0.776
Maximum likelihood	10.002	17.891	9.394	9.840	5.421	.564	0.551	0.577
Multiple imputation	10.002	21.257	9.432	13.167	7.154	.642	0.543	0.758

NOTES: The data were sampled from a bivariate-normal distribution with means, variances, and covariance as shown. The ML and multiple-imputation methods are described later in the chapter.

multiple imputation to each data set. The results are given in Table 20.2.¹⁵ To simplify the table, I have not shown results for available-case analysis or for ML estimation (which produces results similar to those for multiple imputation). In addition, I have focused on the means and regression coefficients, which are the parameters that are usually of most direct interest.

Table 20.2 shows not only the average parameter estimates for each method (in the upper panel), which are useful for assessing bias, but also the RMSE of each estimator (i.e., the square root of the mean-square error, expressing the efficiency of the estimator), as well as (in the lower panel) the coverage and average interval width of nominally 95% confidence intervals for each method. If a confidence interval is valid, then the coverage should be close to .95. The results generally support the observations that I made above, and in particular, the only method that does uniformly well for all parameters—producing unbiased estimates, valid confidence intervals, and relatively efficient estimates—is multiple imputation.

¹⁵Similar but more extensive simulations appear in Schafer and Graham (2002). Also see Exercise 20.3.

Table 20.2 Mean Parameter Estimates and Confidence Interval Coverage for a Simulation Experiment With Data Missing at Random (MAR)

Parameter	Complete Cases	Mean Imputation	Regression Imputation	Multiple Imputation
<i>Mean parameter estimate (RMSE)</i>				
$\mu_1 = 10$	11.476 (1.489)	10.001 (0.189)	10.001 (0.189)	10.001 (0.189)
$\mu_2 = 20$	21.222 (1.355)	21.322 (1.355)	20.008 (0.326)	20.008 (0.344)
$\beta_{12} = 0.5$	0.391 (0.117)	0.391 (0.117)	0.645 (0.151)	0.498 (0.041)
$\beta_{21} = 0.889$	0.891 (0.100)	0.353 (0.538)	0.891 (0.100)	0.890 (0.106)
<i>Confidence-interval coverage (mean interval width)</i>				
μ_1	0 (0.792)	.951 (0.750)	.951 (0.750)	.951 (0.746)
μ_2	.005 (1.194)	0 (0.711)	.823 (0.881)	.947 (1.451)
β_{12}	.304 (0.174)	.629 (0.246)	.037 (0.140)	.955 (0.175)
β_{21}	.953 (0.396)	0 (0.220)	.661 (0.191)	.939 (0.463)

NOTES: The root-mean-square error (RMSE) of the parameter estimates is shown in parentheses below the mean estimates; the mean width of the confidence intervals is shown in parentheses below the coverage. Confidence intervals were constructed at a nominal level of .95.

Traditional methods of handling missing data include complete-case analysis, available-case analysis, and unconditional and conditional mean imputation. Complete-case analysis produces consistent estimates and valid statistical inferences when data are MCAR (and in certain other special circumstances), but even in this advantageous situation, it does not use information in the sample efficiently. The other traditional methods suffer from more serious problems.

20.3 Maximum-Likelihood Estimation for Data Missing at Random*

The method of maximum likelihood can be applied to parameter estimation in the presence of missing data. Doing so requires making assumptions about the distribution of the complete data

and about the process producing missing data. If the assumptions hold, then the resulting ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency.¹⁶

Let $p(\mathbf{X}; \boldsymbol{\theta}) = p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta})$ represent the joint probability density for the complete data \mathbf{X} , which as before, is composed of observed and missing components denoted, respectively, as \mathbf{X}_{obs} and \mathbf{X}_{mis} . The vector $\boldsymbol{\theta}$ contains the unknown parameters on which the complete-data distribution depends. For example, if the variables in \mathbf{X} are multivariately normally distributed (a case that I will examine presently), then $\boldsymbol{\theta}$ includes the population means and covariances among the variables.

In a seminal paper on statistical methods for missing data—the same paper in which he introduced distinctions among data that are MCAR, MAR, and MNAR—Rubin (1976) showed that the ML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained from the marginal distribution of the observed data, *if missing data are missing at random*. In the general case that I am considering here, we can find the marginal distribution for the observed data by integrating over the missing data, producing

$$p(\mathbf{X}_{\text{obs}}; \boldsymbol{\theta}) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Moreover, as I will explain shortly, it is, as a practical matter, possible to find $\hat{\boldsymbol{\theta}}$ in the general case by iterative techniques.¹⁷ As usual, the likelihood function $L(\boldsymbol{\theta}; \mathbf{X}_{\text{obs}})$ is the same as the probability density function for the data but treats the observed data as fixed and the unknown parameters as variable. Once we have found the ML parameter estimates $\hat{\boldsymbol{\theta}}$, we can proceed with statistical inference in the usual manner; for example, we can compute likelihood-ratio tests of nested models and construct Wald tests or confidence intervals for the elements of $\boldsymbol{\theta}$ based on estimated asymptotic variances for $\hat{\boldsymbol{\theta}}$ obtained from the inverse of the observed information matrix

$$\mathcal{I}(\boldsymbol{\theta}; \mathbf{X}_{\text{obs}}) = - \frac{\partial^2 \log_e L(\boldsymbol{\theta}; \mathbf{X}_{\text{obs}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

Consider, for example, bivariately normally distributed variables X_1 and X_2 ; as in the previous section, X_1 is completely observed in a sample of n observations, but X_2 has $m < n$ observations missing at random, which for notational convenience, I will take as the first m observations.¹⁸ Then, from the univariate-normal distribution,

$$p_1(x_{i1}; \mu_1, \sigma_1^2) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left[-\frac{(x_{i1} - \mu_1)^2}{2\sigma_1^2} \right]$$

is the marginal probability density for observation i on variable X_1 , and from the bivariate-normal distribution,

$$p_{12}(x_{i1}, x_{i2}; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}) = \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \quad (20.3)$$

¹⁶For a general introduction to the method of maximum likelihood, see online Appendix D on probability and estimation.

¹⁷See Section 20.3.1 on the expectation-maximization (EM) algorithm.

¹⁸This is the univariate pattern of missing data employed in the examples of the preceding sections.

is the joint probability density for observation i on variables X_1 and X_2 . In Equation 20.3, $\mathbf{x}_i \equiv (x_{i1}, x_{i2})'$ is a vector giving a pair of values for X_{i1} and X_{i2} , $\boldsymbol{\mu} \equiv (\mu_1, \mu_2)'$ is the vector of means for the two variables, and

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

is their covariance matrix. Using results in Little and Rubin (1990, pp. 382–383; 2002, chap. 7), the log-likelihood for the observed data is

$$\begin{aligned} \log_e L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}) &= \sum_{i=1}^m \log_e p_1(x_{i1}; \mu_1, \sigma_1^2) \\ &\quad + \sum_{i=m+1}^n \log_e p_{12}(x_{i1}, x_{i2}; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}) \end{aligned} \quad (20.4)$$

The log-likelihood in Equation 20.4 can easily be maximized numerically, but there is also a simple analytic solution. The statistics

$$\begin{aligned} \bar{X}_1^* &\equiv \frac{\sum_{i=m+1}^n X_{i1}}{n-m} \\ \bar{X}_2^* &\equiv \frac{\sum_{i=m+1}^n X_{i2}}{n-m} \\ S_1^{2*} &\equiv \frac{\sum_{i=m+1}^n (X_{i1} - \bar{X}_1^*)^2}{n-m} \\ S_2^{2*} &\equiv \frac{\sum_{i=m+1}^n (X_{i2} - \bar{X}_2^*)^2}{n-m} \\ S_{12}^* &\equiv \frac{\sum_{i=m+1}^n (X_{i1} - \bar{X}_1^*)(X_{i2} - \bar{X}_2^*)}{n-m} \end{aligned} \quad (20.5)$$

are the means, variances, and covariance for the two variables computed from the $n - m$ complete cases, and

$$\begin{aligned} \bar{X}_1 &\equiv \frac{\sum_{i=1}^n X_{i1}}{n} \\ S_1^2 &\equiv \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2}{n} \end{aligned}$$

are the mean and variance of X_1 computed from all n available cases.¹⁹ The ML estimators of the parameters of the bivariate-normal model are

¹⁹Note that the denominators for the variances and covariance are the number of observations, $n - m$ or n , rather than degrees of freedom $n - m - 1$ or $n - 1$. Recall that ML estimators of variance are biased but consistent. (See online Appendix D on probability and estimation.)

$$\begin{aligned}
 \hat{\mu}_1 &= \bar{X}_1 \\
 \hat{\mu}_2 &= \bar{X}_2^* + \frac{S_{12}^*}{S_1^{2*}} (\bar{X}_1 - \bar{X}_1^*) \\
 \hat{\sigma}_1^2 &= S_1^2 \\
 \hat{\sigma}_2^2 &= S_2^{2*} + \left(\frac{S_{12}^*}{S_1^{2*}} \right)^2 (S_1^2 - S_1^{2*}) \\
 \hat{\sigma}_{12} &= S_{12}^* \left(\frac{S_1^2}{S_1^{2*}} \right)
 \end{aligned} \tag{20.6}$$

Thus, the ML estimates combine information from the complete-case and available-case statistics.²⁰

The method of ML can be applied to parameter estimation in the presence of missing data. If the assumptions made concerning the distribution of the complete data and the process generating missing data hold, then ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency. When data are MAR, the ML estimate $\hat{\theta}$ of the parameters θ of the complete-data distribution can be obtained from the marginal distribution of the observed data, integrating over the missing data:

$$p(\mathbf{X}_{\text{obs}}; \theta) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \theta) d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Once we have found the ML parameter estimates, we can proceed with statistical inference in the usual manner, for example, computing likelihood-ratio tests of nested models and constructing Wald tests or confidence intervals.

20.3.1 The EM Algorithm

Arbitrary patterns of missing data do not yield simple expressions for the log-likelihood (such as in Equation 20.4 on page 615 for a univariate missing-data pattern in bivariate-normal data) no closed-form equations for the ML estimates (such as in Equation 20.6). The *expectation-maximization (EM)* algorithm, due to Dempster, Laird, and Rubin (1977), is a general iterative method for finding ML estimates in the presence of arbitrary patterns of missing data. Although the EM algorithm is broadly applicable, generally easy to implement, and effective, it has the disadvantage that it does not produce the information matrix and therefore does not yield standard errors for the estimated parameters. The version of the EM algorithm that I will describe here is for ignorable missing data (and is adapted from Little & Rubin, 2002, chaps. 8 and 11). The algorithm can also be applied to problems for which data are MNAR and hence are nonignorable.²¹

²⁰See Exercise 20.4 for further interpretation of the ML estimators in Equation 20.6.

²¹See, for example, Little and Rubin (2002, chap. 15).

As before, let \mathbf{X} represent the complete data, composed of the observed data \mathbf{X}_{obs} and the missing data \mathbf{X}_{mis} . The likelihood based on the complete data is $L(\boldsymbol{\theta}; \mathbf{X})$, where recall, $\boldsymbol{\theta}$ contains the parameters for the distribution of \mathbf{X} . Let $\boldsymbol{\theta}^{(l)}$ represent the parameter estimates at the l th iteration of the EM algorithm. Starting values $\boldsymbol{\theta}^{(0)}$ may be obtained from the complete cases, for example. Each iteration of the EM algorithm comprises two steps: an *E (expectation) step* and an *M (maximization) step*. Hence the name “EM.”

- In the E step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right] = \int \log_e L(\boldsymbol{\theta}; \mathbf{X}) p\left(\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(l)}\right) d\mathbf{X}_{\text{mis}}$$

- In the M step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right]$; these are the parameter estimates for the next iteration.

When the parameter values stop changing from one iteration to the next (to an acceptable tolerance), they converge to the ML estimates $\hat{\boldsymbol{\theta}}$.

Suppose, for example, that the complete data \mathbf{X} , consisting of n observations on p variables, is multivariately normally distributed, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sums and sums of squares and cross-products of the variables are a set of sufficient statistics for these parameters:

$$\begin{aligned} T_j &\equiv \sum_{i=1}^n X_{ij} \text{ for } j = 1, \dots, p \\ T_{jj'} &\equiv \sum_{i=1}^n X_{ij} X_{ij'} \text{ for } j, j' = 1, \dots, p \end{aligned}$$

If we have access to the complete data, then the ML estimates of the parameters could be computed from the sufficient statistics:

$$\begin{aligned} \hat{\mu}_j &= \frac{T_j}{n} \\ \hat{\sigma}_{jj'} &= \frac{T_{jj'}}{n} - \hat{\mu}_j \hat{\mu}_{j'} \end{aligned}$$

(where the estimated variance of X_j is $\hat{\sigma}_j^2 = \hat{\sigma}_{jj}$).

Now, imagine that some of the data in \mathbf{X} are MAR but in an arbitrary pattern. Then, in the E step, we find expected sums and sums of products by filling in the missing data with their conditional expected values, given the observed data and current estimates of the parameters. That is,

$$\begin{aligned} E\left(T_j | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) &= \sum_{i=1}^n X_{ij}^{(l)} \\ E\left(T_{jj'} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}\right) &= \sum_{i=1}^n \left(X_{ij}^{(l)} X_{ij'}^{(l)} + C_{ijj'}^{(l)}\right) \end{aligned}$$

where

$$X_{ij}^{(l)} = \begin{cases} X_{ij} & \text{if } X_{ij} \text{ is observed} \\ E(X_{ij} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}) & \text{if } X_{ij} \text{ is missing} \end{cases}$$

and

$$C_{ijj'}^{(l)} = \begin{cases} 0 & \text{if either } X_{ij} \text{ or } X_{ij'} \text{ is observed} \\ C(X_{ij}, X_{ij'} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l-1)}, \boldsymbol{\Sigma}^{(l-1)}) & \text{if both } X_{ij} \text{ and } X_{ij'} \text{ are missing} \end{cases} \quad (20.7)$$

Finally, $E(X_{ij} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)})$ is obtained as the fitted value from the regression of X_j on the other X s, using the current estimates $\boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\Sigma}^{(l)}$ to obtain the regression coefficients, and $C(X_{ij}, X_{ij'} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)})$ is the covariance of the fitted values for X_{ij} and $X_{ij'}$ obtained from the multivariate regression of X_j and $X_{j'}$ on the other X s, again at current values of the parameters.²²

Once we have the expected sums and sums of cross-products, the M step is straightforward:

$$\begin{aligned} \mu_j^{(l)} &= \frac{\sum_{i=1}^n X_{ij}^{(l)}}{n} \\ \sigma_{jj'}^{(l)} &= \frac{\sum_{i=1}^n (X_{ij}^{(l)} X_{ij'}^{(l)} + C_{ijj'}^{(l)})}{n} - \mu_j^{(l)} \mu_{j'}^{(l)} \end{aligned}$$

Consider the comparatively simple case of bivariate-normal data where the variable X_1 is completely observed and the first m of n observations on X_2 are missing. Take as starting values the means, variances, and covariance computed from the $n - m$ complete cases (given in Equation 20.5 on page 615). Then, because X_1 is completely observed,

$$\begin{aligned} E(T_1 | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \sum_{i=1}^n X_{i1} \\ E(T_{11} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \sum_{i=1}^n X_{i1}^2 \end{aligned} \quad (20.8)$$

and, for sums involving X_2 , which has m missing values,

$$\begin{aligned} E(T_2 | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \sum_{i=1}^m \hat{X}_{i2} + \sum_{i=m+1}^n X_{i2} \\ E(T_{22} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \sum_{i=1}^m (\hat{X}_{i2}^2 + S_{2|1}^{2(0)}) + \sum_{i=m+1}^n X_{i2}^2 \\ E(T_{12} | \mathbf{X}_{\text{obs}}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \sum_{i=1}^m (X_{i1} \hat{X}_{i2}^2) + \sum_{i=m+1}^n X_{i1} X_{i2} \end{aligned} \quad (20.9)$$

where \hat{X}_{i2} is the fitted value from the complete-case regression of X_2 on X_1 , and $S_{2|1}^{2(0)}$ is the residual variance from this regression. The M-step estimates computed from these expectations are just the ML estimates previously given in Equation 20.6.²³ That is, in the simple case of monotone missing data, the EM algorithm converges to the ML estimates in a single iteration.

²²In multivariate regression, there is more than one response variable. In the current context, the role of the response variables is played by X_j and $X_{j'}$. See Section 9.5 and Exercise 20.5.

²³See Exercise 20.6.

The EM algorithm is a general iterative procedure for finding ML estimates—but not their standard errors—in the presence of arbitrary patterns of missing data. When data are MAR, iteration l of the EM algorithm consists of two steps: (1) In the E (expectation) step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right] = \int \log_e L(\boldsymbol{\theta}; \mathbf{X}) p\left(\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(l)}\right) d\mathbf{X}_{\text{mis}}$$

(2) In the M (maximization) step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E\left[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}\right]$; these are the parameter estimates for the next iteration. At convergence, the EM algorithm produces the ML estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

20.4 Bayesian Multiple Imputation

Bayesian multiple imputation (abbreviated as *MI*) is a flexible and general method for dealing with missing data that are MAR. Like ML estimation, multiple imputation begins with a specification of the distribution of the complete data (assumed to be known except for a set of parameters to be estimated from the data).

The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing *several* values for each missing value, each imputed value drawn from the *predictive distribution* of the missing data and, therefore, producing not one but several completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets. Parameters of interest are estimated along with their standard errors for each imputed data set. Estimated parameters are then averaged across completed data sets; standard errors are also combined across imputed data sets, taking into account the variation among the estimates in the several data sets, thereby capturing the added uncertainty due to having to impute the missing data.

A multivariate-normal model for the complete data is both relatively simple and useful in applications. Indeed, because the model assumed to describe the complete data is used just to obtain imputed values for the missing data, it turns out that the method of multiple imputation is usually not terribly sensitive to the assumption of multivariate normality.²⁴

Suppose that X_1 and X_2 are bivariately normally distributed and that, as in previously developed examples, there is a univariate pattern of missing data, with X_1 completely observed and m of the n observations on X_2 MAR. For convenience, and again as before, let us order the data so that the missing observations on X_2 are the first m observations. Let $A_{2|1}^*$ and $B_{2|1}^*$ represent the intercept and slope for the complete-case least-squares regression of X_2 on X_1 .²⁵ In regression imputation, recall, we replace the missing values with the fitted values

²⁴See, for example, Schafer (1997, chap. 5). As described in Section 20.4.3, however, there are some pitfalls to be avoided.

²⁵The results of the preceding section imply that $A_{2|1}^*$ and $B_{2|1}^*$ are the ML estimators of $\alpha_{2|1}$ and $\beta_{2|1}$. See Exercise 20.6.

$$\widehat{X}_{i2} = A_{2|1}^* + B_{2|1}^* X_{i1} \quad (20.10)$$

Recall as well that a defect of this procedure is that it ignores residual variation in X_2 conditional on X_1 . A more sophisticated version of regression imputation adds a randomly generated residual to the fitted value, taking the imputed value as $\widehat{X}_{i2} + E_{i2|1}$, where $E_{i2|1}$ is drawn randomly from the normal distribution $N(0, S_{2|1}^{*2})$, and where

$$S_{2|1}^{*2} \equiv \frac{\sum_{i=m+1}^n (X_{i2} - \widehat{X}_{i2})^2}{n-m}$$

is the ML estimator of the residual variance of X_2 given X_1 (based on the $n-m$ complete cases).

There is still a problem, however: The fitted values and generated residuals on which the imputations are based fail to take into account the fact that the regression coefficients $A_{2|1}^*$ and $B_{2|1}^*$ and the residual variance $S_{2|1}^{*2}$ are themselves *estimates* that are subject to sampling variation. MI draws values of the regression parameters and the error variance—let us call these values $\tilde{\alpha}_{2|1}$, $\tilde{\beta}_{2|1}$, and $\tilde{\sigma}_{2|1}^2$ —from the *posterior distribution* of the parameters, typically assuming a *noninformative prior distribution*.²⁶

As Little and Rubin (1990, pp. 386–387) explain, we may proceed as follows:

- Given a random draw Z^2 from the chi-square distribution with $n-m-2$ degrees of freedom, find

$$\tilde{\sigma}_{2|1}^2 \equiv \frac{\sum_{i=m+1}^n (X_{i2} - \widehat{X}_{i2})^2}{Z^2}$$

- With $\tilde{\sigma}_{2|1}^2$ in hand, draw a random slope $\tilde{\beta}_{2|1}$ from the normal distribution

$$N\left(B_{2|1}^*, \frac{\tilde{\sigma}_{2|1}^2}{[(n-m)S_1^2]^2}\right)$$

Here, $S_1^2 \equiv \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2/n$ is the ML estimate of the variance of X_1 , and $\bar{X}_1 \equiv \sum_{i=1}^n X_{i1}/n$ is the ML estimate of the mean of X_1 , based on all n cases.

- Using the previously obtained values of $\tilde{\sigma}_{2|1}^2$ and $\tilde{\beta}_{2|1}$, draw a random intercept $\tilde{\alpha}_{2|1}$ from the normal distribution

$$N\left(\hat{\mu}_2 - \tilde{\beta}_{2|1}\bar{X}_1, \frac{\tilde{\sigma}_{2|1}^2}{(n-m)^2}\right)$$

where $\hat{\mu}_2$ is the ML estimate of the mean of X_2 (given in Equation 20.6 on page 616).

- Finally, replace the missing values in X_2 by

$$\widetilde{X}_{i2} \equiv \tilde{\alpha}_{2|1} + \tilde{\beta}_{2|1} X_{i1} + \widetilde{E}_i$$

where \widetilde{E}_i is sampled from $N(0, \tilde{\sigma}_{2|1}^2)$.

²⁶Think of the posterior distribution of the parameters as capturing our uncertainty about the values of the parameters. Basic concepts of Bayesian statistical inference, including the notions of prior and posterior distributions, are described in online Appendix D on probability and estimation.

In multiple imputation, this procedure is repeated g times, producing g completed data sets.

More generally, we have a complete data set \mathbf{X} comprising n cases and p multivariately normally distributed variables; some of the entries of \mathbf{X} are MAR in an arbitrary pattern. In this more general case, there is no fully adequate closed-form procedure for sampling from the predictive distribution of the data to impute missing values. Instead, simulation methods must be employed to obtain imputations. Two such methods are data augmentation (described in Schafer, 1997) and importance sampling (described in King, Honaker, Joseph, & Scheve, 2001). General descriptions of these methods are beyond the scope of this chapter.²⁷

Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (2001) and van Buuren and Oudshoorn (1999) (also see van Buuren, 2012) suggest a simpler approach that cycles iteratively through a set of regression equations for the variables containing missing data. The formal properties of this approach have not been established, although it appears to work well in practice.²⁸ Multiple imputation can be extended beyond the multivariate-normal distribution to other models for the complete data, such as the multinomial distribution for a set of categorical variables, and mixed multinomial-normal models for data sets containing both quantitative and categorical data.²⁹

20.4.1 Inference for Individual Coefficients

Having obtained g completed data sets, imagine that we have analyzed the data sets in parallel, producing g sets of regression coefficients, $B_0^{(l)}, B_1^{(l)}, \dots, B_k^{(l)}$ for $l = 1, \dots, g$ (where, for notational convenience, I have represented the regression constant as B_0). We also find the coefficient standard errors, $SE(B_0^{(l)}), SE(B_1^{(l)}), \dots, SE(B_k^{(l)})$, computed in the usual manner for each completed data set. Rubin (1987) provides simple rules for combining information across multiple imputations of the missing data, rules that are valid as long as the sample size is sufficiently large for the separate estimates to be approximately normally distributed. The context here is quite general: The regression coefficients and their standard errors might be produced by linear least-squares regression, but they might also be produced by ML estimation of a logistic-regression model, by nonlinear least squares, or by *any* parametric method of regression analysis.

Point estimates of the population regression coefficients are obtained by averaging across imputations:

$$\tilde{\beta}_j = \frac{\sum_{l=1}^g B_j^{(l)}}{g} \quad (20.11)$$

The standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients:

²⁷Multiple imputation by data augmentation is implemented in Schafer's software, available for SAS, S-PLUS, R, and in stand-alone programs. Multiple imputation by importance sampling is implemented in King's software, available for R and in a stand-alone program.

²⁸This approach is implemented in the IVEware (imputation and variance estimation) software for SAS, as a stand-alone program; in the MICE (multivariate imputation by chained equations) software for S-PLUS and R, as well as in a stand-alone program; and in the **mi** package for R (Su, Gelman, Hill, & Yajima, 2011). Access to convenient software for multiple imputation is important because the method is computationally intensive.

²⁹See, for example, Schafer (1997, chaps. 7–9).

$$\widetilde{SE}(\tilde{\beta}_j) \equiv \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}} \quad (20.12)$$

where the within-imputation component is

$$V_j^{(W)} \equiv \frac{\sum_{l=1}^g SE^2(B_j^{(l)})}{g}$$

and the between-imputation component is

$$V_j^{(B)} \equiv \frac{\sum_{l=1}^g (B_j^{(l)} - \tilde{\beta}_j)^2}{g-1}$$

Inference based on $\tilde{\beta}_j$ and $\widetilde{SE}(\tilde{\beta}_j)$ uses the t -distribution, with degrees of freedom determined by

$$df_j = (g-1) \left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}} \right)^2$$

For example, to construct a 95% confidence interval for β_j ,

$$\beta_j = \tilde{\beta}_j \pm t_{0.025, df_j} \widetilde{SE}(\tilde{\beta}_j)$$

Let γ_j denote the relative amount of information about the parameter β_j that is missing. This is not quite the same as the fraction of observations that are missing on the explanatory variable X_j because, unless X_j is uncorrelated with the other variables in the data set, there will be information in the data relevant to imputing the missing values and because data missing on one variable influence all the regression estimates. The *estimated rate of missing information* is

$$\hat{\gamma}_j = \frac{R_j}{R_j + 1} \quad (20.13)$$

where

$$R_j \equiv \frac{g+1}{g} \times \frac{V_j^{(B)}}{V_j^{(W)}}$$

The efficiency of the multiple-imputation estimator relative to the maximally efficient ML estimator—that is, the ratio of sampling variances of the ML estimator to the MI estimator—is $RE(\tilde{\beta}_j) = g/(g + \gamma_j)$. If the number of imputations g is infinite, MI is therefore as efficient as ML, but even when the rate of missing information is quite high and the number of imputations modest, the relative efficiency of the MI estimator hardly suffers. Suppose, for example, that $\gamma_j = 0.5$ (a high rate of missing information) and that $g = 5$; then $RE(\tilde{\beta}_j) = 5/(5 + 0.5) = 0.91$. Expressed on the scale of the standard error of $\tilde{\beta}_j$, which is proportional to the length of the confidence interval for β_j , we have $\sqrt{RE(\tilde{\beta}_j)} = 0.95$.³⁰

³⁰See Exercise 20.7.

Bayesian multiple imputation (MI) is a flexible and general method for dealing with data that are missing at random. The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing g values for each missing value, drawing each imputed value from the predictive distribution of the missing data (a process that usually requires simulation) and therefore producing not one but g completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets.

- According to Rubin's rules, MI estimates (e.g., of a population regression coefficient β_j) are obtained by averaging over the imputed data sets:

$$\tilde{\beta}_j = \frac{\sum_{l=1}^g B_j^{(l)}}{g}$$

where $B_j^{(l)}$ is the estimate of β_j from imputed data set l .

- Standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients,

$$\widetilde{SE}(\tilde{\beta}_j) = \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

where the within-imputation component is

$$V_j^{(W)} = \frac{\sum_{l=1}^g SE^2(B_j^{(l)})}{g}$$

and the between-imputation component is

$$V_j^{(B)} = \frac{\sum_{l=1}^g (B_j^{(l)} - \tilde{\beta}_j)^2}{g-1}$$

Here, $SE(B_j^{(l)})$ is the standard error of B_j computed in the usual manner for the l th imputed data set.

- Inference based on $\tilde{\beta}_j$ and $\widetilde{SE}(\tilde{\beta}_j)$ uses the t -distribution, with degrees of freedom determined by

$$df_j = (g-1) \left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}} \right)^2$$

Inference for several coefficients proceeds in a similar, if more complex, manner.

20.4.2 Inference for Several Coefficients*

The generalization of Rubin's rules to simultaneous tests or confidence regions for several coefficients entails some complications.³¹ Suppose that we wish to test the hypothesis $H_0: \beta_1 = \beta_0$, where β_1 is a subset of $s > 1$ of the $k + 1$ elements of the parameter vector β ; typically, this would be the hypothesis $H_0: \beta_1 = \mathbf{0}$. Were it not for the missing data, we could base the hypothesis test on the Wald chi-square statistic,

$$Z_0^2 = (\mathbf{b}_1 - \beta_0)' \hat{\mathcal{V}}^{-1}(\mathbf{b}_1)(\mathbf{b}_1 - \beta_0)$$

where the vector \mathbf{b}_1 contains the estimated coefficients and $\hat{\mathcal{V}}(\mathbf{b}_1)$ is the estimated asymptotic covariance matrix of \mathbf{b}_1 .³²

In the present context, we have estimates for several completed data sets in which the missing data have been imputed, and so we first average the estimates, obtaining

$$\tilde{\beta}_1 \equiv \frac{1}{g} \sum_{l=1}^g \mathbf{b}_1^{(l)}$$

Then we compute the between- and within-imputation components of the covariance matrix of these estimates:

$$\begin{aligned} \mathbf{V}^{(W)} &\equiv \frac{1}{g} \sum_{l=1}^g \hat{\mathcal{V}}\left(\mathbf{b}_1^{(g)}\right) \\ \mathbf{V}^{(B)} &\equiv \frac{1}{g-1} \sum_{l=1}^g \left(\mathbf{b}_1^{(g)} - \tilde{\beta}_1\right) \left(\mathbf{b}_1^{(g)} - \tilde{\beta}_1\right)' \end{aligned}$$

In analogy to the single-coefficient case, we could compute the total covariance matrix

$$\mathbf{V} \equiv \mathbf{V}^{(W)} + \frac{g+1}{g} \mathbf{V}^{(B)}$$

Basing a test on \mathbf{V} , however, turns out to be complicated.

Instead, simplification of the problem leads to the test statistic

$$F_0 \equiv \frac{(\tilde{\beta}_1 - \beta_0)' (\mathbf{V}^{(W)})^{-1} (\tilde{\beta}_1 - \beta_0)'}{s(1+R)}$$

where

$$R \equiv \frac{g+1}{g} \times \frac{\text{trace}\left[\mathbf{V}^{(B)} (\mathbf{V}^{(W)})^{-1}\right]}{s}$$

The test statistic F_0 follows an approximate F -distribution, with s degrees of freedom in the numerator and denominator given by

³¹The results that I give here, and alternative procedures, are explained in greater detail in Rubin (1987, chaps. 3 and 4) and in Schafer (1997, Section 4.3.3).

³²See, for example, the discussion of Wald tests for generalized linear models in Section 15.3.3.

$$df = \begin{cases} 4 + [s(g - 1) - 4] \left[1 + \frac{1}{R} \times \frac{s(g - 1) - 2}{s(g - 1)} \right] & \text{when } s(g - 1) > 4 \\ \frac{1}{2}(g - 1)(s + 1) \left(1 + \frac{1}{R} \right)^2 & \text{when } s(g - 1) \leq 4 \end{cases}$$

20.4.3 Practical Considerations

Although the multivariate-normal model can prove remarkably useful in providing multiple imputations even when the data are not normally distributed, multiple imputation cannot preserve features of the data that are not represented in the imputation model. How essential it is to preserve particular features of the data depends on the statistical model used to analyze the multiply imputed data sets. It is therefore important in formulating an imputation model to ensure that the imputation model is consistent with the intended analysis. The following points should assist in this endeavor:

- *Try to include variables in the imputation model that make the assumption of ignorability reasonable.* Think of imputation as a pure prediction problem, not as a statistical model subject to substantive interpretation. If we are able to do a good job of predicting missing values (and missingness), then the assumption that data are MAR is more credible. Finding variables that are highly correlated with a variable that has missing data, but for which data are available, therefore, will likely improve the quality of imputations, as will variables that are related to missingness. In particular, it is perfectly acceptable, and indeed desirable, to include variables in the imputation model that *are not used* in the subsequent statistical analysis, alongside the variables that are used in the data analysis.³³ There is also nothing wrong with using the variable that is ultimately to be treated as a response to help impute missing data in variables that are to be treated as explanatory variables. To reiterate, the model used for imputation is essentially a prediction model—not a model to be interpreted substantively.
- *If possible, transform variables to approximate normality.*³⁴ After the imputed data are obtained, the variables can be transformed back to their original scales, if desired, prior to analyzing the completed data sets.
- *Adjust the imputed data to resemble the original data.* For example, imputed values of an integer-valued variable can be rounded to the nearest integer. Ordinal variables can be handled by providing integer codes and then rounding the imputed values to integers. Occasional negative imputed values of a nonnegative variable can be set to 0. Imputed values of a 0/1 dummy variable can be set to 0 if less than or equal to 0.5 and to 1 if greater than 0.5. These steps may not be necessary to analyze the imputed data, but they should not hurt in any event.
- *Make sure that the imputation model captures relevant features of the data.* What is relevant depends on the use to which the imputed data will be put. For example, the multivariate-normal distribution ensures that regressions of one variable on others are linear

³³See Collins, Schafer, and Kam (2001), who present evidence supporting what they term an *inclusive strategy* for formulating imputation models.

³⁴The material in Section 4.2 on transformations for symmetry and in Section 4.6 on Box-Cox transformations for multivariate normality is particularly relevant here.

and additive. Using the multivariate-normal distribution for imputations, therefore, will not preserve *nonlinear* relationships and *interactions* among the variables, unless we make special provision for these features of the data.

Suppose, for example, that we are interested in modeling the potential interaction between gender and education in determining income. Because gender is likely completely observed, but there may well be missing data on both education and income, we could divide the data set into two parts based on gender, obtaining multiply imputed data sets separately for each part and combining them in our analysis of the completed data sets. This approach runs into problems, however, if we find it necessary to divide the data set into too many parts or if the categorical variable or variables used to partition the data are themselves not completely observed.

Allison (2002) suggests forming interaction regressors and polynomial regressors as part of the data set to which the imputation model is applied. The imputed interaction and polynomial regressors are then used in the analysis of the completed data sets. Although such variables are not normally distributed, there is some evidence that multiple imputation based on the multivariate-normal model nevertheless works well in these circumstances.

Although, as explained, the multivariate-normal model can be used to impute a dummy regressor for a dichotomous factor, it is not obvious how to proceed with a polytomous factor. Allison (2002) proposes the following procedure: For an m -category factor, select an arbitrary baseline category (say the last), and code $m - 1$ dummy regressors,³⁵ including these dummy variables in the multiple-imputation process. From the imputed values for the i th observation in the l th imputation, $D_{i1}^{(l)}, D_{i2}^{(l)}, \dots, D_{i,m-1}^{(l)}$, compute $D_{im}^{(l)} = 1 - \sum_{j=1}^{m-1} D_{ij}^{(l)}$. Assign the i th observation to the category $(1, 2, \dots, m)$ for which $D_{ij}^{(l)}$ is largest.

Multiple imputation based on the multivariate-normal distribution can be remarkably effective in practice, even when the data are not normally distributed. To apply multiple imputation effectively, however, it is important to include variables in the imputation model that make the assumption of ignorable missingness reasonable; to transform variables to approximate normality, if possible; to adjust the imputed data so that they resemble the original data; and to make sure that the imputation model captures features of the data, such as nonlinearities and interactions, to be used in subsequent data analysis.

20.4.4 Example: A Regression Model for Infant Mortality

Figure 20.2 (repeating Figure 3.14 from page 45) shows the relationship between infant mortality (number of infant deaths per 1000 live births) and gross domestic product per capita (in U.S. dollars) for 193 nations, part of a larger data set of 207 countries compiled by the United

³⁵See Chapter 7 for a general discussion of dummy-variable regressors.

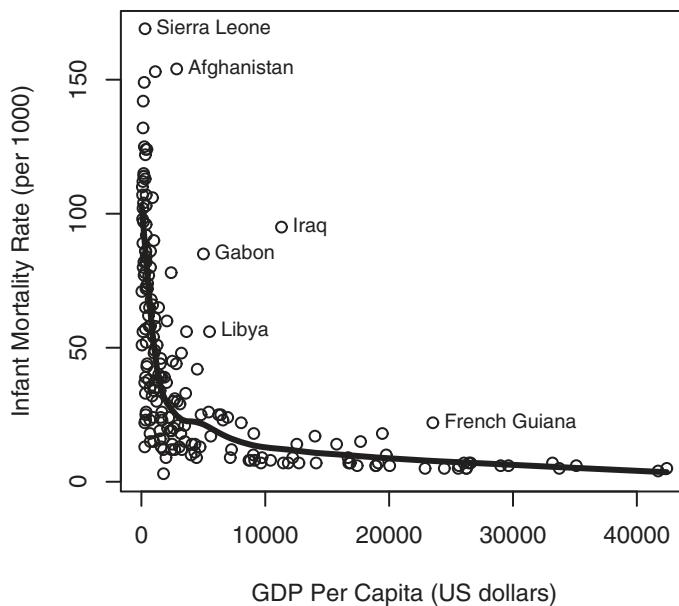


Figure 20.2 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

Nations. The amount of missing data in Figure 20.2 is therefore relatively small, comprising only about 7% of the cases.

Let us now consider the regression of infant mortality not only on GDP per capita but also on the percentage of married women practicing contraception and the average number of years of education for women. To linearize the regression, I log-transformed both infant mortality and GDP.³⁶ A complete-case analysis includes only 62 of the 207 countries and produces the results shown in the upper panel of Table 20.3.

The number of observations with missing data for each of the variables in the analysis is as follows:

Infant Mortality	GDP	Contraception	Female Education
6	10	63	131

There are, however, other variables in the full data set that are highly correlated with contraception and female education, such as the total fertility rate and the illiteracy rate for women. I decided to base imputations on a multivariate-normal model with the four variables in the regression plus the total fertility rate, the expectation of life for women, the percentage of women engaged in economic activity outside the home, and the illiteracy rate for women. Preliminary examination of the data suggested that the multivariate-normal model could be made more appropriate for the data by transforming several of these variables. In particular—as in the regression model in Table 20.3—I log-transformed infant mortality and GDP. I also took the square root of the total fertility rate; cubed female expectation of life, after subtracting

³⁶See Exercise 20.8.

Table 20.3 Estimated Coefficients and Standard Errors for the Regression of Infant Mortality on GDP Per Capita, Percentage Using Contraception, and Average Female Education, for 207 Nations (62 Complete Cases)

	Intercept	$\log_e GDP$	Contraception	Female Education
<i>Complete-case analysis</i>				
Coefficient, B_j	6.88	-0.294	-0.0113	-0.0770
SE(B_j)	(0.29)	(0.058)	(0.0042)	(0.0338)
<i>Multiple-imputation analysis</i>				
Coefficient, $\tilde{\beta}_j$	6.57	-0.234	-0.00953	-0.105
SE($\tilde{\beta}_j$)	(0.18)	(0.049)	(0.00294)	(0.033)
Missing Information, $\hat{\gamma}_j$	0.20	0.61	0.41	0.69

Table 20.4 Means and Standard Deviations of Variables in the Infant Mortality Regression, Complete-Case, and Maximum-Likelihood Estimates

	$\log_e \text{Infant Mortality}$	$\log_e GDP$	Contraception	Female Education
<i>Estimates based on Complete Cases</i>				
Mean	3.041	8.151	50.90	11.30
SD	(1.051)	(1.703)	(23.17)	(3.55)
<i>Maximum-Likelihood Estimates</i>				
Mean	3.300	7.586	44.36	10.16
SD	(1.022)	(1.682)	(24.01)	(3.51)

a start of 35 from each value; and took the 1/4 power of female illiteracy. The resulting data set did not look quite multivariate-normal, but several of the variables were more symmetrically distributed than before.

To get a sense of the possible influence of missing data on conclusions drawn from the data, I computed the complete-case estimates of the means and standard deviations of the four variables to be used in the regression, along with ML estimates, obtained by the EM algorithm applied to the eight variables to be used in the imputation model. These results are given in Table 20.4. As one might expect, the means for the complete cases show lower average infant mortality, higher GDP per capita, higher rates of contraception, and a higher level of female education than the ML estimates assuming ignorable missing data; the two sets of standard deviations, however, are quite similar.

Using Schafer's data augmentation method and employing the multivariate-normal model, I obtained imputations for 10 completed data sets.³⁷ Then, applying Equations 20.11, 20.12, and

³⁷Data augmentation employs a *Markov-chain Monte-Carlo (MCMC)* method to sample from the predictive distribution of the data. Using Schafer's *textrbfnorm* package for the R statistical computing environment for these computations, I set the number of steps for the data augmentation algorithm to 20. Technical aspects of the data augmentation algorithm are discussed in Schafer (1997) and, in less detail, in Allison (2002).

20.13 (on pages 621–622), I computed the estimated coefficients, standard errors, and estimated rate of missing information for each coefficient, shown in the lower panel of Table 20.3. With the exception of the female education coefficient, the standard errors from the multiple-imputation analysis are noticeably smaller than those from the complete-case analysis. In addition, the coefficients for GDP and female education differ between the two analyses by about one standard error; the coefficients for contraception, in contrast, are very similar. Finally, the rates of missing information for the three slope coefficients are all large. Because 10 imputations were employed, however, the square-root relative efficiency of the estimated coefficients based on the multiply imputed data is at worst $\sqrt{10/(10 + 0.69)} = 0.97$.

20.5 Selection Bias and Censoring

When missing data are not ignorable (i.e., MNAR), consistent estimation of regression models requires an explicit auxiliary model for the missingness mechanism. Accommodating nonignorable missing data is an intrinsically risky venture because the resulting regression estimates can be very sensitive to the specifics of the model assumed to generate the missing data.

This section introduces two models in wide use for data that are MNAR: Heckman's model to overcome selection bias in regression and the so-called tobit model (and related models) for a censored response variable in regression. Before examining these models, however, it is useful to develop some basic ideas concerning truncated- and censored-normal distributions.

20.5.1 Truncated- and Censored-Normal Distributions

The distinction between *truncation* and *censoring* is illustrated in Figure 20.3. In each case, there is an unobserved variable ξ that follows the standard-normal distribution, $N(0, 1)$. The observed variable Z in panel (a) *truncates* this distribution *on the left* by suppressing all values of ξ below $\xi = -0.75$; that is, there are no observations below the truncation point. The density function $p(z)$ of Z still must enclose an area of 1, and so this density is given by

$$p(z) = \frac{\phi(z)}{1 - \Phi(-0.75)} = \frac{\phi(z)}{\Phi(0.75)} \text{ for } z \geq -0.75$$

where $\phi(\cdot)$ is the density function and $\Phi(\cdot)$ the cumulative distribution function of the standard-normal distribution. In panel (b), where the distribution of ξ is *left-censored* rather than truncated,

$$Z = \begin{cases} -0.75 & \text{for } \xi \leq -0.75 \\ \xi & \text{for } \xi > -0.75 \end{cases}$$

Consequently, $\Pr(Z = -0.75) = \Phi(-0.75)$, that is, the area to the left of -0.75 under the standard-normal density function $\phi(\cdot)$.

It will be useful to have expressions for the mean and variance of a truncated-normal distribution. Suppose now that ξ is normally distributed with an *arbitrary* mean μ and variance

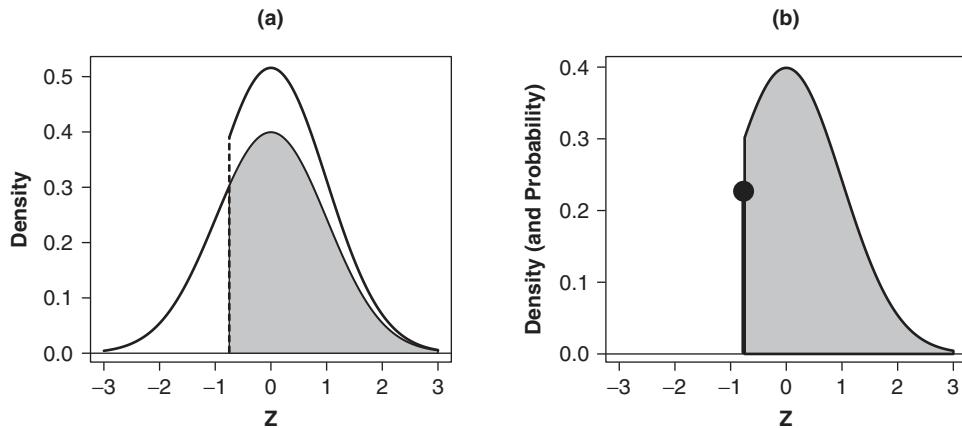


Figure 20.3 (a) Truncated- and (b) censored-normal distributions. In both cases, the underlying distribution is standard normal, $N(0,1)$. In (a), there are no values of Z observed below $Z = -0.75$, and the remaining density is rescaled (see the upper curve) to an area of 1. In (b), values below -0.75 are set to $Z = -0.75$; the probability of observing this value is represented by the “spike” topped by a circle.

σ^2 —that is, $\xi \sim N(\mu, \sigma^2)$ —and that this distribution is left-truncated at the *threshold* a , giving rise to the observable variable Y . Then, the mean and variance of Y are³⁸

$$\begin{aligned} E(Y) &= E(\xi | \xi \geq a) = \mu + \sigma m(z_a) \\ V(Y) &= V(\xi | \xi \geq a) = \sigma^2 [1 - d(z_a)] \end{aligned} \quad (20.14)$$

where

$$\begin{aligned} z_a &\equiv \frac{a - \mu}{\sigma} \\ m(z_a) &\equiv \frac{\phi(z_a)}{1 - \Phi(z_a)} = \frac{\phi(z_a)}{\Phi(-z_a)} \\ d(z_a) &\equiv m(z_a)[m(z_a) - z_a] \end{aligned} \quad (20.15)$$

The quantity $m(z_a)$, called the *inverse Mills ratio*, is a function of the standardized threshold; it will figure prominently in the remainder of this section. As a general matter, the mean of the left-truncated variable Y exceeds that of ξ by an amount that depends on the standardized threshold and the standard deviation σ of the untruncated distribution; similarly, the variance of Y is smaller than the variance of ξ by a factor dependent on the standardized threshold.³⁹ The inverse Mills ratio is graphed against z_a in Figure 20.4: As the threshold moves to the right, the relationship between the inverse Mills ratio and z_a becomes more linear.

³⁸The derivation of these results, and of some other results in this section, is beyond the level of the text, even in starred material or exercises. See Johnson, Kotz, and Balakrishnan (1994) and Kotz, Balakrishnan, and Johnson (1994).

³⁹See Exercise 29.4 for the mean and variance of a right-truncated normal variable.

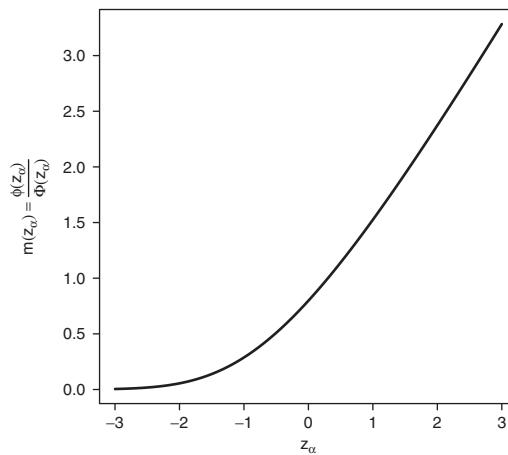


Figure 20.4 The inverse Mills ratio $m(z_\alpha)$ as a function of the standardized threshold z_α .

The expectation and variance of a censored-normal variable follow straightforwardly. Suppose that $\xi \sim N(\mu, \sigma^2)$ is left-censored at $\xi = a$, so that

$$Y = \begin{cases} a & \text{for } \xi \leq a \\ \xi & \text{for } \xi > a \end{cases}$$

Then,⁴⁰

$$\begin{aligned} E(Y) &= a\Phi(z_a) + [\mu + \sigma m(z_a)][1 - \Phi(z_a)] \\ V(Y) &= \sigma^2[1 - \Phi(z_a)]\left\{1 - d(z_a) + [z_a - m(z_a)]^2\Phi(z_a)\right\} \end{aligned} \quad (20.16)$$

A variable can be truncated or censored at the right as well as at the left or can be truncated or censored at both ends simultaneously (the latter is termed *interval censoring*). The analysis of right-censored or interval-censored data is essentially similar to the analysis of left-censored data, making adjustments to the formulas in Equations 20.16.⁴¹

Finally, suppose that the unobservable variables ξ and ζ follow a bivariate-normal distribution, with means μ_ξ and μ_ζ , variances σ_ξ^2 and σ_ζ^2 , and correlation ρ (so that the covariance of ξ and ζ is $\sigma_{\xi\zeta} = \rho\sigma_\xi\sigma_\zeta$). Imagine that, as before, Y is a truncated version of ξ , but now the truncation depends *not* on the value of ξ *itself* but rather on that of ζ , so that $Y = \xi$ when $\zeta \geq a$ and Y is unobserved when $\zeta < a$. This process is called *incidental truncation* or *selection*. The mean and variance for the incidentally truncated variable Y are

$$\begin{aligned} E(Y) &= E(\xi | \zeta \geq a) = \mu_\xi + \sigma_\xi \rho m(z_a) \\ V(Y) &= V(\xi | \zeta \geq a) = \sigma_\xi^2[1 - \rho^2 d(z_a)] \end{aligned} \quad (20.17)$$

where $z_a \equiv (a - \mu_\zeta)/\sigma_\zeta$ and $m(\cdot)$ and $d(\cdot)$ are defined as in Equations 20.15. The effect of incidental truncation depends, therefore, not only on the standardized threshold z_a but also on

⁴⁰See Exercise 20.10.

⁴¹See Exercise 20.11.

the correlation ρ between the latent variables ξ and ζ . For example, if these variables are positively correlated, then $E(Y) > E(\xi)$ and $V(Y) < V(\xi)$.

The distribution of a variable is truncated when values below or above a threshold (or outside a particular range) are unobserved. The distribution of a variable is censored when values below or above a threshold (or outside a particular range) are set equal to the threshold. The distribution of a variable is incidentally censored if its value is unobserved when another variable is below or above a threshold (or outside a particular range). Simple formulas exist for the mean and variance of truncated- and censored-normal distributions and for the mean and variance of an incidentally truncated variable in a bivariate-normal distribution.

20.5.2 Heckman's Selection-Regression Model

The model and methods of estimation described in this section originated with James Heckman (e.g., Heckman, 1974, 1976), whose work on selection bias won him a Nobel Prize in economics. Heckman's selection-regression model consists of two parts:

1. A *regression equation* for a *latent response variable* ξ :

$$\begin{aligned}\xi_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \\ &= \eta_i + \varepsilon_i\end{aligned}\tag{20.18}$$

2. A *selection equation* that determines whether or not ξ is observed:

$$\begin{aligned}\zeta_i &= \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i \\ &= \psi_i + \delta_i\end{aligned}\tag{20.19}$$

where the *observed response variable*

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

The explanatory variables in Equation 20.19 (i.e., the Z s) are intended to predict missingness; they need not be the same as the explanatory variables used in the regression equation of principal interest (Equation 20.18), but in applications, there is usually considerable overlap between the Z s and the X s. The observed response, for example, might represent earnings for married women, which is missing when they are not in the paid labor force; the latent variable would then represent a notional “potential earnings.” An example based on this idea is developed below.

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $\sigma_\varepsilon^2 \equiv V(\varepsilon_i)$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$. Errors for

different observations are assumed to be independent. Equation 20.19, together with the assumption that the errors δ are normally distributed, specifies a *probit model* for nonmissingness.⁴²

As we will see presently, estimating the regression equation (Equation 20.18) just for complete cases—simply omitting observations for which Y is missing—generally produces inconsistent estimates of the regression coefficients. In addition, because of the correlation of the two error variables, the missing data are not ignorable, and so it would be inappropriate, for example, to generate multiple imputations of the missing values of Y as if they were MAR.

Restricting our attention to the complete cases,

$$\begin{aligned} E(Y_i|\zeta_i > 0) &= \eta_i + E(\varepsilon_i | \zeta_i > 0) \\ &= \eta_i + E(\varepsilon_i | \delta_i > -\psi_i) \end{aligned}$$

The conditional expectation of the error ε_i follows from Equations 20.17 for the incidentally truncated bivariate-normal distribution:⁴³

$$E(\varepsilon_i | \delta_i > -\psi_i) = \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i)$$

Therefore,

$$\begin{aligned} E(Y_i | \zeta_i > 0) &= \eta_i + \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i) \\ &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i \end{aligned}$$

where $\beta_\lambda \equiv \sigma_\varepsilon \rho_{\varepsilon\delta}$ and $\lambda_i \equiv m(-\psi_i)$.

Letting $\nu_i \equiv Y_i - E(Y_i | \zeta_i > 0)$, we can write the regression equation for the complete cases as

$$(Y_i | \zeta_i > 0) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i + \nu_i$$

Regressing Y on the X s using only the complete cases omits the explanatory variable λ_i , which is the inverse Mills ratio based on the negative of the linear predictor ψ_i from the selection equation (Equation 20.19). Ignoring the missingness mechanism, therefore, can be conceptualized as a kind of omitted-variable specification error.⁴⁴ If the errors from the regression and selection equations are uncorrelated (i.e., $\rho_{\varepsilon\delta} = 0$), then $\beta_\lambda = 0$, and ignoring λ_i is inconsequential. Similarly, if λ_i were uncorrelated with the X s, then we could ignore it without threatening the consistency of the least-squares estimators of the regression coefficients. Uncorrelation of λ_i and the X s is unlikely, however: The selection and regression equations typically contain many of the same explanatory variables, and unless the degree of selection is low, the inverse Mills ratio is nearly a linear function of the linear predictor (recall Figure 20.4 on page 631). Indeed, *high* correlation between λ_i and the X s can make consistent estimation of the regression coefficients (by the methods described immediately below) unstable. Note, as well, that the variance of the errors ν_i is not constant.⁴⁵

There are two common strategies for estimating Heckman's regression-selection model: direct application of ML estimation and employing an estimate of λ_i as an auxiliary regressor.

⁴²For a general treatment of probit regression, see Section 14.1. Recall that we can arbitrarily set the threshold above which Y is observed and below which it is missing to 0 and the error variance δ to 1, to fix the origin and scale of the latent variable.

⁴³See Exercise 20.12.

⁴⁴See Sections 6.3 and 9.7.

⁴⁵The variance of ν_i follows from Equations 20.17 for the variance of an incidentally truncated variable in the bivariate-normal distribution: See Exercise 20.13.

- *ML Estimation*^{*}: Let $\beta \equiv (\alpha, \beta_1, \dots, \beta_k)'$ be the vector of regression coefficients in the regression equation (Equation 20.18); let $\gamma \equiv (\gamma_0, \gamma_1, \dots, \gamma_p)'$ be the vector of regression coefficients in the selection equation (Equation 20.19); let $\mathbf{x}'_i \equiv (1, X_{i1}, \dots, X_{ik})$ be the i th row of the model matrix for the regression equation; and let $\mathbf{z}'_i \equiv (1, Z_{i1}, \dots, Z_{ip})$ be the i th row of the model matrix for the selection equation. For notational convenience, order the data so that the missing observations on Y are the first m of n observations. Then the log-likelihood for Heckman's model can be formulated as follows:⁴⁶

$$\begin{aligned} \log_e L(\beta, \gamma, \sigma_\varepsilon^2, \rho_{\varepsilon\delta}) &= \sum_{i=1}^m \log_e \Phi(\mathbf{z}'_i \gamma) \\ &+ \sum_{i=m+1}^n \log_e \left[\frac{1}{\sigma_\varepsilon} \phi\left(\frac{Y_i - \mathbf{x}'_i \beta}{\sigma_\varepsilon}\right) \Phi\left(\frac{\mathbf{z}'_i \gamma + \rho_{\varepsilon\delta} \frac{Y_i - \mathbf{x}'_i \beta}{\sigma_\varepsilon}}{\sqrt{\frac{1 - \rho_{\varepsilon\delta}}{\sigma_\varepsilon}}}\right) \right] \end{aligned} \quad (20.20)$$

This log-likelihood can be maximized numerically.

- *Two-Step Estimation*: Heckman (1979) also proposed a simple and widely used two-step procedure for estimating his regression-selection model.

Step 1: Define the dichotomous response variable

$$W_i = \begin{cases} 1 & \text{if } Y_i \text{ is observed} \\ 0 & \text{if } Y_i \text{ is missing} \end{cases}$$

Perform a probit regression of W_i on the Z s, estimating the γ s in the usual manner by ML,⁴⁷ and finding fitted values on the probit scale,

$$\hat{\psi}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_{i1} + \hat{\gamma}_2 Z_{i2} + \dots + \hat{\gamma}_p Z_{ip}$$

$\hat{\psi}_i$ is simply the estimated linear predictor from the probit model. For each observation, compute the estimated inverse Mills ratio

$$\hat{\lambda}_i = m(-\hat{\psi}_i) = \frac{\phi(-\hat{\psi}_i)}{1 - \Phi(-\hat{\psi}_i)} = \frac{\phi(\hat{\psi}_i)}{\Phi(\hat{\psi}_i)}$$

Step 2: Use $\hat{\lambda}$ as an auxiliary regressor in the least-squares regression of Y_i on the X s for the complete cases,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \beta_{\hat{\lambda}} \hat{\lambda}_i + \nu_i^*, \quad (20.21)$$

for $i = m + 1, \dots, n$

This least-squares regression provides consistent estimates of the regression coefficients, $\alpha, \beta_1, \beta_2, \dots, \beta_k$. The heteroscedasticity of the errors, however, requires an adjustment to the usual OLS standard errors.⁴⁸

⁴⁶See Exercise 20.14.

⁴⁷As described in Chapters 14 and 15.

⁴⁸See Exercise 20.15.

To illustrate the application of Heckman's selection-regression model, I will return to the Canadian Survey of Labour and Income Dynamics (the SLID),⁴⁹ examining the relationship between women's earnings and their education, age, and the region in which they reside, restricting attention to married women between the ages of 18 and 65. Earnings is represented by the women's composite hourly wage rate, which is missing if they are not in the paid labor force. Preliminary examination of the data suggested regressing the log of composite hourly wages on the square of years of education and a quadratic in age, along with four dummy regressors for five regions of Canada (the Atlantic provinces, Quebec, Ontario, the prairie provinces, and British Columbia, taking the Atlantic provinces as the baseline category).

Of the 6427 women in the sample, 3936 were in the paid labor force.⁵⁰ Because women whose potential earnings are relatively low may very well be less likely to work outside the home, there is a potential for selection bias if we simply ignore the 2491 women who are not in the labor force, causing us to underestimate the effects of the explanatory variables on (potential) earnings.⁵¹

I formulated a selection model in which labor-force participation is regressed on region dummy variables; dummy regressors for the presence in the household of children 0 to 4 and 5 to 9 years of age; family income less the woman's own income, if any (in thousands of dollars); education (in years); and a quadratic in age (in years). The results are shown in Table 20.5. At the left of the table are ordinary least-squares estimates *ignoring* the selection process. The table also shows two-step and ML estimates for the Heckman model, both for the regression equation and for the selection equation. For the two-step estimation procedure, the selection equation was estimated in a preliminary probit regression.

In this application, the two-step/probit and ML estimates are very similar and are not terribly different from the OLS estimates based on the complete cases. Moreover, the ML estimate of the correlation between the errors of the regression and selection equations is fairly small: $\hat{\rho}_{\varepsilon\delta} = .320$. The degree of collinearity induced by the introduction of the inverse Mills ratio regressor in the second step of the two-step procedure is not serious, as shown in Table 20.6, which compares generalized variance inflation factors for the model as estimated by OLS and Heckman's two-step procedure.⁵²

⁴⁹In Chapter 12, I used the SLID for a regression of earnings on sex, age, and education. In Chapter 14, the SLID provided data for a logistic regression of young married women's labor force participation on region, presence of children, family income, and education.

⁵⁰The complete SLID sample of married women between 18 and 65 years of age consists of 6900 respondents. I omitted the relatively small number of observations (comprising about 7% of the sample) with missing data on variables other than earnings.

⁵¹If, however, our goal is to *describe* the regression of earnings on the explanatory variables for those who are in the paid labor force, then an analysis based on women who have earnings should be perfectly fine, as long as we are careful to ensure the descriptive accuracy of the model—for example, by using component-plus-residual plots to check for nonlinearity (see Chapter 12).

⁵²Generalized variance-inflation factors (GVIFs), introduced in Section 13.1.2, are appropriate for terms in a model that have more than 1 degree of freedom, such as the region and age terms in this model. When a term has 1 degree of freedom, the GVIF reduces to the usual variance inflation factor (VIF, also discussed in Chapter 13). Taking the $1/(2df)$ power of the GVIF makes values roughly comparable across different degrees of freedom. Treating the linear and quadratic components of the age term as a set is important here because otherwise there would be artifactual collinearity induced by the high correlation between Age and Age².

Table 20.5 Least-Squares, Heckman Two-Step, and Heckman ML Estimates for the Regression of Women's Composite Hourly Wages on Region, Education, and Age

	OLS		Two-Step/Probit		ML	
	Estimate	SE	Estimate	SE	Estimate	SE
Coefficient	Regression Equation					
Constant	1.10	0.15	0.442	0.227	0.755	0.177
Quebec	0.223	0.031	0.205	0.033	0.214	0.032
Ontario	0.303	0.026	0.332	0.028	0.319	0.027
Prairies	0.126	0.027	0.147	0.029	0.137	0.027
B.C.	0.371	0.036	0.392	0.038	0.382	0.037
Education ²	0.00442	0.00013	0.00492	0.00018	0.00469	0.00014
Age	0.0687	0.0074	0.0917	0.0096	0.0807	0.0081
Age ²	-0.000717	0.000088	-0.00105	0.00012	-0.000892	0.000099
Inv. Mills Ratio			0.361	0.088		
Selection Equation						
Constant		-1.46	0.30	-1.44	0.30	
Quebec		-0.0665	0.0533	-0.0674	0.0533	
Ontario		0.193	0.048	0.194	0.048	
Prairies		0.117	0.049	0.117	0.049	
B.C.		0.145	0.067	0.150	0.067	
Children 0–4		-0.414	0.050	-0.439	0.049	
Children 5–9		-0.261	0.043	-0.251	0.042	
Family Income		-0.00399	0.00097	-0.00475	0.00097	
Education		0.0815	0.0061	0.0817	0.0060	
Age		0.0878	0.0135	0.0882	0.0134	
Age ²		-0.00145	0.00015	-0.00145	0.00015	

As is seldom a bad idea, I will leave the last word on Heckman-type adjustments for selection bias to John Tukey (1986), who states,⁵³

I think that an important point that we have to come back to at intervals is that knowledge always comes from a combination of data and assumptions. If the assumptions are too important, many of us get unhappy. I think one thing we were told in this last discussion was that all the formal ways that have been found for attacking this problem ended up being very dependent upon these assumptions. Therefore, people like me have to be very uncomfortable about the results. (p. 58)

⁵³Tukey made these comments about a paper delivered by Heckman and Robb (1986) to a symposium on statistical methods for self-selected samples (collected in a volume edited by Wainer, 1986). The models introduced by Heckman and Robb are not the same as Heckman's selection-regression model discussed in this section, but they are similarly motivated and structured.

Table 20.6 Generalized Variance Inflation Factors for Terms in the OLS and Heckman Two-Step Regression of Log Hourly Wages on Region, Education, and Age

Term	df	GVIF ^{1/(2df)}	
		OLS Estimates	Heckman Two-Step Estimates
Region	4	1.003	1.024
Education ²	1	1.025	1.402
Age (quadratic)	2	1.012	1.347
Inverse Mills Ratio	1	–	2.202

Heckman's regression model consists of two parts:

1. A regression equation for a latent response variable ξ :

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

2. A selection equation that determines whether or not ξ is observed:

$$\zeta_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i = \psi_i + \delta_i$$

where the observed response variable

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $V(\varepsilon_i) = \sigma_\varepsilon^2$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$ and that errors for different observations are independent. Heckman's model can be consistently estimated by ML or by a two-step procedure. In the first step of the two-step procedure, the selection equation is estimated as a probit model; in the second step, the regression equation is estimated by OLS after incorporating the auxiliary regressor $\hat{\lambda}_i = \phi(\hat{\psi}_i)/\Phi(\hat{\psi}_i)$, where $\hat{\psi}_i$ is the fitted value from the first-step probit equation, $\phi(\cdot)$ is the density function of the standard-normal distribution, and $\Phi(\cdot)$ is the distribution function of the standard-normal distribution.

20.5.3 Censored-Regression Models

When the response variable Y in a regression is censored, values of Y cannot be observed outside a certain range—say, the interval (a, b) . We can detect, however, whether an observation falls below the lower threshold a or above the upper threshold b , and consequently, we have *some* information about the censored values.

Let us assume, in particular, that the *latent response variable* ξ is linearly related to the regressors X_1, X_2, \dots, X_k , so that

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad (20.22)$$

and that the other assumptions of the normal-regression model hold: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly, however, but instead we collect data on the *censored response variable* Y , where

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases} \quad (20.23)$$

Equations 20.22 and 20.23 define the *censored-regression model*. A model of this type was first proposed by James Tobin (1958), for data censored to the left at 0—that is, for $a = 0$ and $b = \infty$. Left-censored regression models are called *tobit models*, in honor of Tobin (another Nobel Prize winner in economics). The censored-regression model can be estimated by the method of ML.⁵⁴

*Rewriting the regression equation in vector form for compactness as $\xi_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$, the log-likelihood for the censored-regression model in Equations 20.22 and 20.23 is

$$\begin{aligned} \log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = & \sum_{Y_i=a} \log_e \Phi\left(\frac{a - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma_\varepsilon}\right) + \sum_{a < Y_i < b} \log_e \left[\frac{1}{\sigma_\varepsilon} \phi\left(\frac{Y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma_\varepsilon}\right) \right] \\ & + \sum_{Y_i=b} \log_e \Phi\left(\frac{\mathbf{x}'_i \boldsymbol{\beta} - b}{\sigma_\varepsilon}\right) \end{aligned}$$

The log-likelihood therefore comprises terms for left-censored, fully observed, and right-censored observations.⁵⁵

For an example of censored regression, I turn once again to the Canadian SLID data. We last encountered the SLID in the previous section, where the earnings of married women were regressed on region, education, and age. I employed Heckman's selection-regression model because earnings were unavailable for women who were not in the paid labor force. I will now develop a similar example in which the response variable is hours worked in the year preceding the survey. This variable is left-censored at the value 0, producing a classic tobit regression model.⁵⁶ The explanatory variables are region, the presence in the household of children 0 to 4 and 5 to 9 years old, family income less the woman's own income (if any), education, and a quadratic in age. The SLID data set includes 6340 respondents with valid data on the variables employed in this example.

Preliminary examination of the data suggested a square-root transformation of hours worked. This transformation does not, of course, serve to spread out the values of the response variable for the 31% of respondents who reported 0 hours worked—that is, the transformed response for all censored observations is $\sqrt{0} = 0$. OLS and ML tobit estimates for the regression model are shown in Table 20.7. The OLS estimates are consistently smaller in magnitude than the corresponding tobit estimates.⁵⁷

⁵⁴An alternative is to employ Heckman's two-step procedure, described in the preceding section.

⁵⁵Estimation is facilitated by reparameterization. See, for example, Greene (2003, Section 22.3.3).

⁵⁶The latent response variable therefore represents "propensity" to work outside the home; presumably, if that propensity is above the threshold 0, we observe positive hours worked.

⁵⁷See Exercise 20.16.

Table 20.7 OLS and ML Tobit Estimates for the Regression of Square-Root Hours Worked on Several Explanatory Variables

Coefficient	OLS		Tobit	
	Estimate	SE	Estimate	SE
Constant	-20.3	3.8	-58.7	5.4
Quebec	-0.745	0.710	-1.58	1.01
Ontario	3.55	0.63	5.02	0.89
Prairies	3.64	0.65	5.36	0.91
B.C.	2.09	0.88	3.73	1.23
Children 0–4 (present)	-6.56	0.65	-8.63	0.91
Children 5–9 (present)	-5.05	0.56	-6.91	0.79
Family Income (\$1000s)	-0.0977	0.0128	-0.139	0.018
Education (years)	1.29	0.08	1.87	0.11
Age (years)	2.32	0.17	3.84	0.25
Age ²	-0.0321	0.0019	-0.0529	0.0028

In the censored-regression model, the latent response variable ξ is linearly related to the regressors X_1, X_2, \dots, X_k :

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly but instead collect data on the censored response variable Y :

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases}$$

When Y is left-censored at 0 (i.e., $a = 0$ and $b = \infty$), the censored-regression model is called a tobit model in honor of James Tobin. The censored-regression model can be estimated by maximum likelihood.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 20.1. Consider the following contrived data set for the variables X_1, X_2 , and X_3 , where the question marks indicate missing data:

X_1	X_2	X_3
1	1	?
1	?	1
−1	−1	?
−1	?	−1
?	1	−1
?	−1	1
5	?	?

- (a) Using available cases (and recomputing the means and standard deviations for each pair of variables), find the pairwise correlations among the three variables and explain why the correlations are not consistent with each other.
- (b) Compute the correlation between X_1 and X_2 using means and standard deviations computed *separately* from the valid observations for each variable. What do you find?
- (c) *Show that the available-case correlation matrix among the variables X_1 , X_2 , and X_3 is not positive semidefinite.

Exercise 20.2. *In univariate missing data, where there are missing values for only one variable in a data set, some of the apparently distinct methods for handling missing data produce identical results for certain statistics. Consider Table 20.1 on page 612, for example, where data are missing on the variable X_2 but not on X_1 . Note that the complete-case, available-case, and mean-imputation estimates of the slope β_{12} for the regression of X_1 on X_2 are identical. Prove that this is no accident. Are there any other apparent agreements between or among methods in the table? If so, can you determine whether they are coincidences?

Exercise 20.3. *Duplicate the small simulation study reported in Table 20.2 on page 613, comparing several methods of handling univariate missing data that are MAR. Then repeat the study for missing data that are MCAR and for missing data that are MNAR (generated as in Figure 20.1 on page 608). What do you conclude? *Note:* This is not a *conceptually* difficult project, but it is potentially time-consuming; it also requires some programming skills and statistical software that can generate and analyze simulated data.

Exercise 20.4. *Equation 20.6 (on page 616) gives the ML estimators for the parameters μ_1 , μ_2 , σ_1^2 , σ_2^2 , and σ_{12} in the bivariate-normal model with some observations on X_2 missing at random but X_1 completely observed. The interpretation of $\hat{\mu}_1$ and $\hat{\sigma}_1^2$ is straightforward: They are the available-case mean and variance for X_1 . Noting that S_{12}^*/S_1^{2*} is the complete-case slope for the regression of X_2 on X_1 , offer interpretations for the other ML estimators.

Exercise 20.5. *Multivariate linear regression fits the model

$$\mathbf{Y}_{(n \times m)} = \mathbf{X}_{(n \times k+1)} \mathbf{B}_{(k+1 \times m)} + \mathbf{E}_{(n \times m)}$$

where \mathbf{Y} is a matrix of response variables; \mathbf{X} is a model matrix (just as in the *univariate* linear model); \mathbf{B} is a matrix of regression coefficients, one column per response variable; and \mathbf{E} is a

matrix of errors. The least-squares estimator of \mathbf{B} is $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (equivalent to what one would get from separate least squares regressions of each Y on the X s). See Section 9.5 for a discussion of the multivariate linear model.

- (a) Show how $\widehat{\mathbf{B}}$ can be computed from the means of the variables, $\widehat{\mu}_Y$ and $\widehat{\mu}_X$, and from their covariances, $\widehat{\Sigma}_{XX}$ and $\widehat{\Sigma}_{XY}$ (among the X s and between the X s and Y s, respectively).
- (b) The fitted values from the multivariate regression are $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}$. It follows that the fitted values \widehat{Y}_{ij} and $\widehat{Y}_{ij'}$ for the i th observation on response variables j and j' are both linear combinations of the i th row of the model matrix, \mathbf{x}_i' . Use this fact to find an expression for the covariance of \widehat{Y}_{ij} and $\widehat{Y}_{ij'}$.
- (c) Show how this result can be used in Equation 20.7 (on page 618), which applies the EM algorithm to multivariate-normal data with missing values.

Exercise 20.6. *Consider once again the case of univariate missing data MAR for two bivariately normal variables, where the first variable, X_1 , is completely observed, and m observations (for convenience, the first m) on the second variable, X_2 , are missing.

- (a) Let $A_{2|1}^*$ and $B_{2|1}^*$ represent the intercept and slope for the complete-case least-squares regression of X_2 on X_1 . Show that $A_{2|1}^*$ and $B_{2|1}^*$ are the ML estimators of $\alpha_{2|1}$ and $\beta_{2|1}$. (*Hint:* Use Equations 20.6 giving the ML estimators of μ_1 , μ_2 , σ_1^2 , σ_2^2 , and σ_{12} .)
- (b) Show that the M step from the first iteration of the EM algorithm (see Equations 20.8 and 20.9 on page 618 for the E step) produces the ML estimates (given in Equations 20.6 on page 616). That is, demonstrate that the EM algorithm converges in a single iteration.

Exercise 20.7. As explained in Section 20.4.1, the efficiency of the multiple-imputation estimator of a coefficient $\widetilde{\beta}_j$ relative to the ML estimator $\widehat{\beta}_j$ is $RE(\widetilde{\beta}_j) = g/(g + \gamma_j)$, where g is the number of imputations employed and γ_j is the rate of missing information for coefficient β_j . The square root of $RE(\widetilde{\beta}_j)$ expresses relative efficiency on the coefficient standard-error scale. Compute $RE(\widetilde{\beta}_j)$ and $\sqrt{RE(\widetilde{\beta}_j)}$ for combinations of values of $g = 1, 2, 3, 5, 10, 20$, and 100 , and $\gamma_j = .05, .1, .2, .5, .9$, and $.99$. What do you conclude about the number of imputations required for efficient inference?

Exercise 20.8. Examine the United Nations data on infant mortality and other variables for 207 countries, discussed in Section 20.4.4.

- (a) Perform a complete-case linear least-squares regression of infant mortality on GDP per capita, percentage using contraception, and female education. Does it appear reasonable to log-transform infant mortality and GDP to linearize this regression? What about contraception and education?
- (b) *Examine a scatterplot matrix (Section 3.3.1) for the variables used in the imputation example. What do you find? Then apply the multivariate Box-Cox procedure described in Section 4.6 to these variables. Remember first to subtract 35 from female expectation of life (why?). Do the results that you obtain support the transformations

employed in the text? Apply the transformations and reexamine the data. Do they appear more nearly normal?

Exercise 20.9. Truncated normal distributions:

- (a) Suppose that $\xi \sim N(0, 1)$. Using Equations 20.14 (page 630) for the mean and variance of a left-truncated normal distribution, calculate the mean and variance of $\xi | \xi > a$ for each of $a = -2, -1, 0, 1$, and 2 .
- (b) *Find similar formulas for the mean and variance of a *right*-truncated normal distribution. What happens to the mean and variance as the threshold moves to the left?

Exercise 20.10. *Suppose that $\xi \sim N(\mu, \sigma^2)$ is left-censored at $\xi = a$, so that

$$Y = \begin{cases} a & \text{for } \xi \leq a \\ \xi & \text{for } \xi > a \end{cases}$$

Using Equations 20.14 (on page 630) for the *truncated* normal distribution, show that (repeating Equations 20.16 on page 631)

$$\begin{aligned} E(Y) &= a\Phi(z_a) + [\mu + \sigma m(z_a)][1 - \Phi(z_a)] \\ V(Y) &= \sigma^2[1 - \Phi(z_a)]\left\{1 - d(z_a) + [z_a - m(z_a)]^2\Phi(z_a)\right\} \end{aligned}$$

Exercise 20.11. *Equations 20.16 (on page 631) give formulas for the mean and variance of a left-censored normally distributed variable. (These formulas are also shown in the preceding exercise.) Derive similar formulas for (a) a right-censored and (b) an interval-censored normally distributed variable.

Exercise 20.12. *Using Equations 20.17 (page 631) for the incidentally truncated bivariate-normal distribution, show that the expectation of the error ε_i in the Heckman regression model (Equations 20.18 and 20.19 on page 632) conditional on Y being observed is

$$E(\varepsilon_i | \zeta_i > 0) = E(\varepsilon_i | \delta_i > -\psi_i) = \sigma_\varepsilon \rho_{\varepsilon\delta} m(-\psi_i)$$

Exercise 20.13. *As explained in the text, the Heckman regression model (Equations 20.18 and 20.19, page 632) implies that

$$(Y_i | \zeta_i > 0) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \beta_\lambda \lambda_i + \nu_i$$

where $\beta_\lambda \equiv \sigma_\varepsilon \rho_{\varepsilon\delta}$, $\lambda_i \equiv m(-\psi_i)$, and

$$\psi_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip}$$

Show that the errors ν_i are heteroscedastic, with variance

$$V(\nu_i) = \sigma_\varepsilon^2 [1 - \rho_{\varepsilon\delta}^2 \lambda_i (\lambda_i + \psi_i)]$$

where σ_ε^2 is the error variance in the regression equation (Equation 20.18), and $\rho_{\varepsilon\delta}$ is the correlation between the errors of the regression and selection equations. (*Hint:* See Equations 20.17

on page 631 for the variance of an incidentally truncated variable in a bivariate-normal distribution.)

Exercise 20.14. *The log-likelihood for the Heckman regression-selection model is given in Equation 20.20 (page 634). Derive this expression. (*Hint:* The first sum in the log-likelihood, for the observations for which Y is missing, is of the log-probability that each such Y_i is missing; the second sum is of the log of the probability density at the observed values of Y_i times the probability that each such value is observed.)

Exercise 20.15. *Explain how White's coefficient-variance estimator (see Section 12.2.3), which is used to correct the covariance matrix of OLS regression coefficients for heteroscedasticity, can be employed to obtain consistent coefficient standard errors for the two-step estimator of Heckman's regression-selection model—the second step of which entails an OLS regression with heteroscedastic errors (Equation 20.21 on page 634). (*Hint:* Refer to Exercise 20.13 for the variance of the errors in the second-step OLS regression.)

Exercise 20.16. Greene (2003 p. 768) remarks that the ML estimates $\hat{\beta}_j$ of the regression coefficients in a censored-regression model are often approximately equal to the OLS estimates B_j divided by the proportion P of *uncensored* observations; that is, $\hat{\beta}_j \approx B_j/P$. Does this pattern hold for the hours-worked regression in Table 20.7 (page 639), where $P = .69$?

Summary

- Missing data are missing completely at random (MCAR) if they can be regarded as a simple random sample of the complete data. If missingness is related to the observed data but not to the missing data (conditional on the observed data), then data are missing at random (MAR). If missingness is related to the missing values themselves, even when the information in the observed data is taken into account, then data are missing not at random (MNAR). When data are MCAR or MAR, the process that produces missing data is ignorable, in the sense that valid methods exist to deal with the missing data without explicitly modeling the process that generates them. In contrast, when data are MNAR, the process producing missing data is nonignorable and must be modeled. Except in special situations, it is not possible to know whether data are MCAR, MAR, or MNAR.
- Traditional methods of handling missing data include complete-case analysis, available-case analysis, and unconditional and conditional mean imputation. Complete-case analysis produces consistent estimates and valid statistical inferences when data are MCAR (and in certain other special circumstances), but even in this advantageous situation, it does not use information in the sample efficiently. The other traditional methods suffer from more serious problems.
- The method of maximum likelihood (ML) can be applied to parameter estimation in the presence of missing data. If the assumptions made concerning the distribution of the complete data and the process generating missing data hold, then ML estimates have their usual optimal properties, such as consistency and asymptotic efficiency. When data are MAR, the ML estimate $\hat{\theta}$ of the parameters θ of the complete-data distribution can be obtained from the marginal distribution of the observed data, by integrating over the missing data,

$$p(\mathbf{X}_{\text{obs}}; \boldsymbol{\theta}) = \int p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) d\mathbf{X}_{\text{mis}}$$

Although it may be difficult to apply this result directly, simplification is possible in certain cases. Once we have found the ML parameter estimates, we can proceed with statistical inference in the usual manner, for example, by computing likelihood-ratio tests of nested models and constructing Wald tests or confidence intervals.

- The EM algorithm is a general iterative procedure for finding ML estimates—but not their standard errors—in the presence of arbitrary patterns of missing data. When data are MAR, iteration l of the EM algorithm consists of two steps: (1) In the E (expectation) step, we find the expectation of the complete-data log-likelihood, integrating over the missing data, given the observed data and the current estimates of the parameters:

$$E[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}] = \int \log_e L(\boldsymbol{\theta}; \mathbf{X}) p(\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \boldsymbol{\theta}^{(l)}) d\mathbf{X}_{\text{mis}}$$

(2) In the M (maximization) step, we find the values $\boldsymbol{\theta}^{(l+1)}$ of $\boldsymbol{\theta}$ that maximize the expected log-likelihood $E[\log_e L(\boldsymbol{\theta}; \mathbf{X}) | \boldsymbol{\theta}^{(l)}]$; these are the parameter estimates for the next iteration. At convergence, the EM algorithm produces the ML estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

- Bayesian multiple imputation (MI) is a flexible and general method for dealing with data that are missing at random. The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing g values for each missing value, drawing each imputed value from the predictive distribution of the missing data (a process that usually requires simulation), and therefore producing not one but g completed data sets. Standard methods of statistical analysis are then applied in parallel to the completed data sets.
 - According to Rubin's rules, MI estimates (e.g., of a population regression coefficient β_j) are obtained by averaging over the imputed data sets:

$$\tilde{\beta}_j = \frac{\sum_{l=1}^g B_j^{(l)}}{g}$$

where $B_j^{(l)}$ is the estimate of β_j from imputed data set l .

- Standard errors of the estimated coefficients are obtained by combining information about within- and between-imputation variation in the coefficients,

$$\widetilde{SE}(\tilde{\beta}_j) = \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

where the within-imputation component is

$$V_j^{(W)} = \frac{\sum_{l=1}^g \text{SE}^2(B_j^{(l)})}{g}$$

and the between-imputation component is

$$V_j^{(B)} = \frac{\sum_{l=1}^g (B_j^{(l)} - \tilde{\beta}_j)^2}{g - 1}$$

Here, $\text{SE}(B_j^{(l)})$ is the standard error of B_j computed in the usual manner for the l th imputed data set.

- Inference based on $\hat{\beta}_j$ and $\text{SE}(\hat{\beta}_j)$ uses the t -distribution, with degrees of freedom determined by

$$df_j = (g - 1) \left(1 + \frac{g}{g + 1} \times \frac{V_j^{(W)}}{V_j^{(B)}} \right)^2$$

Inference for several coefficients proceeds in a similar, if more complex, manner.

- Multiple imputation based on the multivariate-normal distribution can be remarkably effective in practice, even when the data are not normally distributed. To apply multiple imputation effectively, however, it is important to include variables in the imputation model that make the assumption of ignorable missingness reasonable; to transform variables to approximate normality, if possible; to adjust the imputed data so that they resemble the original data; and to make sure that the imputation model captures features of the data, such as nonlinearities and interactions, to be used in subsequent data analysis.
- The distribution of a variable is truncated when values below or above a threshold (or outside a particular range) are unobserved. The distribution of a variable is censored when values below or above a threshold (or outside a particular range) are set equal to the threshold. The distribution of a variable is incidentally censored if its value is unobserved when another variable is below or above a threshold (or outside a particular range). Simple formulas exist for the mean and variance of truncated- and censored-normal distributions and for the mean and variance of an incidentally truncated variable in a bivariate-normal distribution.
- Heckman's regression model consists of two parts:

1. A regression equation for a latent response variable ξ ,

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

2. A selection equation that determines whether or not ξ is observed,

$$\zeta_i = \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + \cdots + \gamma_p Z_{ip} + \delta_i = \psi_i + \delta_i$$

where the observed response variable

$$Y_i = \begin{cases} \text{missing} & \text{for } \zeta_i \leq 0 \\ \xi_i & \text{for } \zeta_i > 0 \end{cases}$$

It is assumed that the two error variables ε_i and δ_i follow a bivariate-normal distribution with means $E(\varepsilon_i) = E(\delta_i) = 0$, variances $V(\varepsilon_i) = \sigma_\varepsilon^2$ and $V(\delta_i) = 1$, and correlation $\rho_{\varepsilon\delta}$, and that errors for different observations are independent.

Heckman's model can be consistently estimated by ML or by a two-step procedure. In the first step of the two-step procedure, the selection equation is estimated as a

probit model; in the second step, the regression equation is estimated by OLS after incorporating the auxiliary regressor $\hat{\lambda}_i = \phi(\hat{\psi}_i)/\Phi(\hat{\psi}_i)$, where $\hat{\psi}_i$ is the fitted value from the first-step probit equation, $\phi(\cdot)$ is the density function of the standard-normal distribution, and $\Phi(\cdot)$ is the distribution function of the standard-normal distribution.

- In the censored-regression model, the latent response variable ξ is linearly related to the regressors X_1, X_2, \dots, X_k :

$$\xi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. We cannot observe ξ directly but instead collect data on the censored response variable Y ,

$$Y_i = \begin{cases} a & \text{for } \xi_i \leq a \\ \xi_i & \text{for } a < \xi_i < b \\ b & \text{for } \xi_i \geq b \end{cases}$$

When Y is left-censored at 0 (i.e., $a = 0$ and $b = \infty$), the censored-regression model is called a tobit model in honor of James Tobin. The censored-regression model can be estimated by ML.

Recommended Reading

- Little and Rubin (2002), central figures in the recent development of more adequate methods for handling missing data, present a wide-ranging and largely accessible overview of the field. A briefer treatment by the same authors appears in Little and Rubin (1990).
- Another fine, if mathematically more demanding, book on handling missing data is Schafer (1997). Also see the overview paper by Schafer and Graham (2002).
- van Buuren (2012) is an accessible, book-length presentation of the simpler chained-equations approach to multiple imputation of missing data.
- Allison's (2002) monograph on missing data is clear, comprehensive, and directed to social scientists (as is the paper by King et al., 2001).
- The edited volume by Wainer (1986) on sample-selection issues contrasts the points of view of statisticians and econometricians—in particular in an exchange between John Tukey and James Heckman. Also see the paper by Stolzenberg and Relles (1997) and the review paper by Winship and Mare (1992).

21

Bootstrapping Regression Models

Bootstrapping is a nonparametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results.¹ Bootstrapping offers a number of advantages:

- The bootstrap is quite general, although there are some cases in which it fails.
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically.
- It is relatively simple to apply the bootstrap to complex data collection plans (such as many complex sample surveys).

21.1 Bootstrapping Basics

My principal aim is to explain how to bootstrap regression models (broadly construed to include generalized linear models, etc.), but the topic is best introduced in a simpler context: Suppose that we draw an independent random sample from a large population.² For concreteness and simplicity, imagine that we sample four working, married couples, determining in each case the husband's and wife's income, as recorded in Table 21.1. I will focus on the difference in incomes between husbands and wives, denoted as Y_i for the i th couple.

We want to estimate the mean difference in income between husbands and wives in the population. Please bear with me as I review some basic statistical theory: A point estimate of this population mean difference μ is the sample mean,

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{6 - 3 + 5 + 3}{4} = 2.75$$

Elementary statistical theory tells us that the standard deviation of the sampling distribution of sample means is $SD(\bar{Y}) = \sigma/\sqrt{n}$, where σ is the population standard deviation of Y .

¹The term *bootstrapping*, coined by Efron (1979), refers to using the sample to learn about the sampling distribution of a statistic without reference to external assumptions—as in “pulling oneself up by one’s bootstraps.”

²Recall from Section 15.5 that in an *independent random sample*, each element of the population can be selected more than once. In a *simple random sample*, in contrast, once an element is selected into the sample, it is removed from the population, so that sampling is done “without replacement.” When the population is very large in comparison to the sample (say, at least 20 times as large), the distinction between independent and simple random sampling becomes inconsequential.

Table 21.1 Contrived “Sample” of Four Married Couples, Showing Husbands’ and Wives’ Incomes in Thousands of Dollars

Observation	Husband’s Income	Wife’s Income	Difference Y_i
1	34	28	6
2	24	27	-3
3	50	45	5
4	54	51	3

If we knew σ , and if Y were normally distributed, then a 95% confidence interval for μ would be

$$\mu = \bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

where $z_{.025} = 1.96$ is the standard normal value with a probability of .025 to the right. If Y is *not* normally distributed in the population, then this result applies asymptotically. Of course, the asymptotics are cold comfort when $n = 4$.

In a real application, we do not know σ . The usual estimator of σ is the sample standard deviation,

$$S = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

from which the standard error of the mean (i.e., the *estimated* standard deviation of \bar{Y}) is $SE(\bar{Y}) = S/\sqrt{n}$. If the population is normally distributed, then we can take account of the added uncertainty associated with estimating the standard deviation of the mean by substituting the heavier-tailed t -distribution for the normal distribution, producing the 95% confidence interval

$$\mu = \bar{Y} \pm t_{n-1,.025} \frac{S}{\sqrt{n}}$$

Here, $t_{n-1,.025}$ is the critical value of t with $n - 1$ degrees of freedom and a right-tail probability of .025.

In the present case, $S = 4.031$, $SE(\bar{Y}) = 4.031/\sqrt{4} = 2.015$, and $t_{3,.025} = 3.182$. The 95% confidence interval for the population mean is thus

$$\mu = 2.75 \pm 3.182 \times 2.015 = 2.75 \pm 6.41$$

or, equivalently,

$$-3.66 < \mu < 9.16$$

As one would expect, this confidence interval—which is based on only four observations—is very wide and includes 0. It is, unfortunately, hard to be sure that the population is reasonably close to normally distributed when we have such a small sample, and so the t -interval may not be valid.³

³To say that a confidence interval is “valid” means that it has the stated coverage. That is, a 95% confidence interval is valid if it is constructed according to a procedure that encloses the population mean in 95% of samples.

Bootstrapping begins by using the distribution of data values in the sample (here, $Y_1 = 6, Y_2 = -3, Y_3 = 5, Y_4 = 3$) to *estimate* the distribution of Y in the population.⁴ That is, we define the random variable Y^* with distribution⁵

y^*	$p^*(y^*)$
6	.25
-3	.25
5	.25
3	.25

from which

$$E^*(Y^*) = \sum_{\text{all } y^*} y^* p(y^*) = 2.75 = \bar{Y}$$

and

$$\begin{aligned} V^*(Y^*) &= \sum [y^* - E^*(Y^*)]^2 p(y^*) \\ &= 12.187 = \frac{3}{4} S^2 = \frac{n-1}{n} S^2 \end{aligned}$$

Thus, the expectation of Y^* is just the sample mean of Y , and the variance of Y^* is [except for the factor $(n-1)/n$, which is trivial in larger samples] the sample variance of Y .

We next mimic sampling from the original population by treating the sample as if it were the population, enumerating all possible samples of size $n = 4$ from the probability distribution of Y^* . In the present case, each *bootstrap sample* selects four values *with replacement* from among the four values of the original sample. There are, therefore, $4^4 = 256$ different bootstrap samples,⁶ each selected with probability 1/256. A few of the 256 samples are shown in Table 21.2. Because the four observations in each bootstrap sample are chosen with replacement, particular bootstrap samples usually have repeated observations from the original sample. Indeed, of the illustrative bootstrap samples shown in Table 21.2, only sample 100 does *not* have repeated observations.

Let us denote the b th bootstrap sample⁷ as $\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, Y_{b3}^*, Y_{b4}^*]'$, or more generally, $\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, \dots, Y_{bn}^*]'$, where $b = 1, 2, \dots, n^n$. For each such bootstrap sample, we calculate the mean,

⁴An alternative would be to resample from a distribution given by a nonparametric density estimate (see, e.g., Silverman & Young, 1987). Typically, however, little if anything is gained by using a more complex estimate of the population distribution. Moreover, the simpler method explained here generalizes more readily to more complex situations in which the population is multivariate or not simply characterized by a distribution.

⁵The asterisks on $p^*(\cdot)$, E^* , and V^* remind us that this probability distribution, expectation, and variance are conditional on the specific sample in hand. Were we to select another sample, the values of Y_1, Y_2, Y_3 , and Y_4 would change and—along with them—the probability distribution of Y^* , its expectation, and variance.

⁶Many of the 256 samples have the same elements but in different order—for example, [6, 3, 5, 3] and [3, 5, 6, 3]. We could enumerate the unique samples without respect to order and find the probability of each, but it is simpler to work with the 256 orderings because each ordering has equal probability.

⁷If vector notation is unfamiliar, then think of \mathbf{y}_b^* simply as a list of the bootstrap observations Y_{bi}^* for sample b .

Table 21.2 A Few of the 256 Bootstrap Samples for the Data Set [6, -3, 5, 3], and the Corresponding Bootstrap Means, \bar{Y}_b^*

Bootstrap Sample		b	Y_{b1}^*	Y_{b2}^*	Y_{b3}^*	Y_{b4}^*	\bar{Y}_b^*
		1	6	6	6	6	6.00
		2	6	6	6	-3	3.75
		3	6	6	6	5	5.75
		\vdots	\vdots				\vdots
		100	-3	5	6	3	2.75
		101	-3	5	-3	6	1.25
		\vdots	\vdots				\vdots
		255	3	3	3	5	3.50
		256	3	3	3	3	3.00

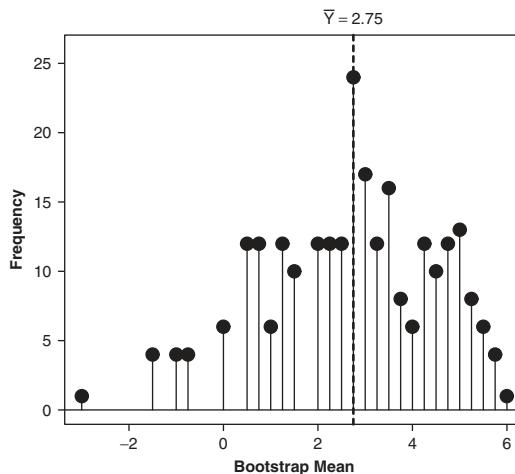


Figure 21.1 Graph of the 256 bootstrap means from the sample [6, -3, 5, 3]. The broken vertical line gives the mean of the original sample, $\bar{Y} = 2.75$, which is also the mean of the 256 bootstrap means.

$$\bar{Y}_b^* = \frac{\sum_{i=1}^n Y_{bi}^*}{n}$$

The sampling distribution of the 256 bootstrap means is shown in Figure 21.1.

The mean of the 256 bootstrap sample means is just the original sample mean, $\bar{Y} = 2.75$. The standard deviation of the bootstrap means is

$$\begin{aligned} \text{SD}^*(\bar{Y}^*) &= \sqrt{\frac{\sum_{b=1}^n (\bar{Y}_b^* - \bar{Y})^2}{n}} \\ &= 1.745 \end{aligned}$$

We divide here by n^n rather than by $n^n - 1$ because the distribution of the $n^n = 256$ bootstrap sample means (Figure 21.1) is known, *not* estimated. The standard deviation of the bootstrap means is nearly equal to the usual standard error of the sample mean; the slight slippage is due to the factor $\sqrt{n/(n-1)}$, which is typically negligible (though not when $n = 4$):⁸

$$\text{SE}(\bar{Y}) = \sqrt{\frac{n}{n-1}} \text{SD}^*(\bar{Y}^*)$$

$$2.015 = \sqrt{\frac{4}{3}} \times 1.745$$

This precise relationship between the usual formula for the standard error and the bootstrap standard deviation is peculiar to *linear statistics* (i.e., linear functions of the data) like the mean. For the mean, then, the bootstrap standard deviation is just a more complicated way to calculate what we already know, but

- bootstrapping might still provide more accurate confidence intervals, as I will explain presently, and
- bootstrapping can be applied to *nonlinear* statistics for which we do not have standard-error formulas or for which only asymptotic standard errors are available.

Bootstrapping exploits the following central analogy:

**The population is to the sample
as
the sample is to the bootstrap samples.**

Consequently,

- the *bootstrap observations* Y_{bi}^* are analogous to the *original observations* Y_i ,
- the *bootstrap mean* \bar{Y}_b^* is analogous to the *mean of the original sample* \bar{Y} ,
- the *mean of the original sample* \bar{Y} is analogous to the (unknown) *population mean* μ , and
- the *distribution of the bootstrap sample means* is analogous to the (unknown) *sampling distribution of means* for samples of size n drawn from the original population.

Bootstrapping uses the sample data to estimate relevant characteristics of the population. The sampling distribution of a statistic is then constructed empirically by resampling from the sample. The resampling procedure is designed to parallel the process by which sample observations were drawn from the population. For example, if the data represent an independent random sample of size n (or a simple random sample of size n from a much larger population), then each bootstrap sample selects n observations with replacement from the original sample. The key bootstrap analogy is the following: *The population is to the sample as the sample is to the bootstrap samples.*

⁸See Exercise 21.1.

Table 21.3 Contrived “Sample” of 10 Married Couples, Showing Husbands’ and Wives’ Incomes in Thousands of Dollars

Observation	Husband’s Income	Wife’s Income	Difference Y_i
1	34	28	6
2	24	27	-3
3	50	45	5
4	54	51	3
5	34	28	6
6	29	19	10
7	31	20	11
8	32	40	-8
9	40	33	7
10	34	25	9

The bootstrapping calculations that we have undertaken thus far depend on very small sample size, because the number of bootstrap samples (n^n) quickly becomes unmanageable: Even for samples as small as $n = 10$, it is impractical to enumerate all the $10^{10} = 10$ billion bootstrap samples. Consider the “data” shown in Table 21.3, an extension of the previous example. The mean and standard deviation of the differences in income Y are $\bar{Y} = 4.6$ and $S = 5.948$. Thus, the standard error of the sample mean is $SE(\bar{Y}) = 5.948/\sqrt{10} = 1.881$.

Although we cannot (as a practical matter) enumerate *all* the 10^{10} bootstrap samples, it is easy to draw at random a large number of bootstrap samples. To estimate the standard deviation of a statistic (here, the mean)—that is, to get a bootstrap standard error—100 or 200 bootstrap samples should be more than sufficient. To find a confidence interval, we will need a larger number of bootstrap samples, say 1000 or 2000.⁹

A practical bootstrapping procedure, therefore, is as follows:

1. Let r denote the number of *bootstrap replications*—that is, the number of bootstrap samples to be selected.
2. For each bootstrap sample $b = 1, \dots, r$, randomly draw n observations $Y_{b1}^*, Y_{b2}^*, \dots, Y_{bn}^*$ with replacement from among the n sample values, and calculate the bootstrap sample mean,

$$\bar{Y}_b^* = \frac{\sum_{i=1}^n Y_{bi}^*}{n}$$

⁹Results presented by Efron and Tibshirani (1993, chap. 19) suggest that basing bootstrap confidence intervals on 1000 bootstrap samples generally provides accurate results, and using 2000 bootstrap replications should be very safe.

Table 21.4 A Few of the $r = 2000$ Bootstrap Samples Drawn From the Data Set $[6, -3, 5, 3, 6, 10, 11, -8, 7, 9]$ and the Corresponding Bootstrap Means, \bar{Y}_b^*

b	Y_{b1}^*	Y_{b2}^*	Y_{b3}^*	Y_{b4}^*	Y_{b5}^*	Y_{b6}^*	Y_{b7}^*	Y_{b8}^*	Y_{b9}^*	Y_{b10}^*	\bar{Y}_b^*
1	6	10	6	5	-8	9	9	6	11	3	5.7
2	9	9	7	7	3	3	-3	-3	-8	6	3.0
3	9	-3	6	5	10	6	10	10	10	6	6.9
:	:										:
1999	6	9	6	3	11	6	6	7	3	9	6.6
2000	7	6	7	3	10	6	9	3	10	6	6.7

3. From the r bootstrap samples, estimate the standard deviation of the bootstrap means:¹⁰

$$\text{SE}^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^r (\bar{Y}_b^* - \bar{\bar{Y}}^*)^2}{r-1}}$$

where

$$\bar{\bar{Y}}^* = \frac{\sum_{b=1}^r \bar{Y}_b^*}{r}$$

is the mean of the bootstrap means. We can, if we wish, “correct” $\text{SE}^*(\bar{Y}^*)$ for degrees of freedom, multiplying by $\sqrt{n/(n-1)}$.

To illustrate this procedure, I drew $r = 2000$ bootstrap samples, each of size $n = 10$, from the “data” given in Table 21.3, calculating the mean, \bar{Y}_b^* , for each sample. A few of the 2000 bootstrap replications are shown in Table 21.4, and the distribution of bootstrap means is graphed in Figure 21.2.

We know from statistical theory that were we to enumerate all the 10^{10} bootstrap samples (or, alternatively, to sample infinitely from the population of bootstrap samples), the average bootstrap mean would be $E^*(\bar{Y}^*) = \bar{Y} = 4.6$, and the standard deviation of the bootstrap means would be

$$\text{SE}^*(\bar{Y}^*) = \text{SE}(\bar{Y}) \sqrt{\frac{n-1}{n}} = 1.881 \sqrt{\frac{9}{10}} = 1.784$$

For the 2000 bootstrap samples that I selected, $\bar{\bar{Y}}^* = 4.693$ and $\text{SE}(\bar{Y}^*) = 1.750$ —both quite close to the theoretical values.

The bootstrapping procedure described in this section can be generalized to derive the empirical sampling distribution for an estimator $\hat{\theta}$ of the parameter θ :

¹⁰It is important to distinguish between the “ideal” bootstrap estimate of the standard deviation of the mean, $\text{SD}^*(\bar{Y}^*)$, which is based on all n^n bootstrap samples, and the estimate of this quantity, $\text{SE}^*(\bar{Y}^*)$, which is based on r randomly selected bootstrap samples. By making r large enough, we seek to ensure that $\text{SE}^*(\bar{Y}^*)$ is close to $\text{SD}^*(\bar{Y}^*)$. Even $\text{SD}^*(\bar{Y}^*) = \text{SE}(\bar{Y})$ is an imperfect estimate of the true standard deviation of the sample mean $\text{SD}(\bar{Y})$, however, because it is based on a particular sample of size n drawn from the original population.

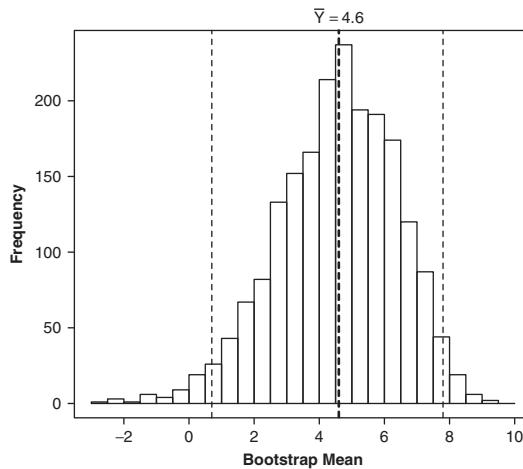


Figure 21.2 Histogram of $r = 2000$ bootstrap means, produced by resampling from the “sample” $[6, -3, 5, 3, 6, 10, 11, -8, 7, 9]$. The heavier broken vertical line gives the sample mean, $\bar{Y} = 4.6$; the lighter broken vertical lines give the boundaries of the 95% percentile confidence interval for the population mean μ based on the 2000 bootstrap samples. The procedure for constructing this confidence interval is described in the next section.

1. Specify the data collection scheme \mathcal{S} that gives rise to the observed sample when applied to the population:¹¹

$$\mathcal{S}(\text{Population}) \Rightarrow \text{Sample}$$

The estimator $\hat{\theta}$ is some function $\mathcal{S}(\cdot)$ of the observed sample. In the preceding example, the data collection procedure is independent random sampling from a large population.

2. Using the observed sample data as a “stand-in” for the population, replicate the data collection procedure, producing r bootstrap samples:

$$\mathcal{S}(\text{Sample}) \left\{ \begin{array}{l} \Rightarrow \text{Bootstrap sample}_1 \\ \Rightarrow \text{Bootstrap sample}_2 \\ \vdots \\ \Rightarrow \text{Bootstrap sample}_r \end{array} \right.$$

3. For each bootstrap sample, calculate the estimate $\hat{\theta}_b^* = \mathcal{S}(\text{Bootstrap sample}_b)$.
4. Use the distribution of the $\hat{\theta}_b^*$ s to estimate properties of the sampling distribution of $\hat{\theta}$. For example, the bootstrap standard error of $\hat{\theta}$ is $\text{SE}^*(\hat{\theta}^*)$ (i.e., the standard deviation of the r bootstrap replications $\hat{\theta}_b^*$):¹²

¹¹The “population” can be real—the population of working married couples—or hypothetical—the population of conceivable replications of an experiment. What is important in the present context is that the sampling procedure can be described concretely.

¹²We may want to apply the correction factor $\sqrt{n/(n-1)}$.

$$\text{SE}^*(\hat{\theta}^*) \equiv \sqrt{\frac{\sum_{b=1}^r (\hat{\theta}_b^* - \bar{\theta}^*)^2}{r-1}}$$

where

$$\bar{\theta}^* \equiv \frac{\sum_{b=1}^r \hat{\theta}_b^*}{r}$$

21.2 Bootstrap Confidence Intervals

21.2.1 Normal-Theory Intervals

Most statistics, including sample means, are asymptotically normally distributed; in large samples, we can therefore use the bootstrap standard error, along with the normal distribution, to produce a $100(1 - a)\%$ confidence interval for θ based on the estimator $\hat{\theta}$:

$$\theta = \hat{\theta} \pm z_{a/2} \text{SE}^*(\hat{\theta}^*) \quad (21.1)$$

In Equation 21.1, $z_{a/2}$ is the standard normal value with probability $a/2$ to the right. This approach will work well if the bootstrap sampling distribution of the estimator is approximately normal, and so it is advisable to examine a normal quantile-comparison plot of the bootstrap distribution.

There is no advantage to calculating normal-theory bootstrap confidence intervals for linear statistics like the mean, because in this case, the ideal bootstrap standard deviation of the statistic and the standard error based directly on the sample coincide. Using bootstrap resampling in this setting just makes for extra work and introduces an additional small random component into standard errors.

Having produced r bootstrap replicates $\hat{\theta}_b^*$ of an estimator $\hat{\theta}$, the bootstrap standard error is the standard deviation of the bootstrap replicates: $\text{SE}^*(\hat{\theta}^*) = \sqrt{\sum_{b=1}^r (\hat{\theta}_b^* - \bar{\theta}^*)^2 / (r-1)}$, where $\bar{\theta}^*$ is the mean of the $\hat{\theta}_b^*$. In large samples, where we can rely on the normality of $\hat{\theta}$, a 95% confidence interval for θ is given by $\hat{\theta} \pm 1.96 \text{SE}^*(\hat{\theta}^*)$.

21.2.2 Percentile Intervals

Another very simple approach is to use the quantiles of the bootstrap sampling distribution of the estimator to establish the end points of a confidence interval *nonparametrically*. Let $\hat{\theta}_{(b)}^*$ represent the ordered bootstrap estimates, and suppose that we want to construct a $(100 - a)\%$ confidence interval. If the number of bootstrap replications r is large (as it should be to construct a

percentile interval), then the $a/2$ and $1 - a/2$ quantiles of $\hat{\theta}_b^*$ are approximately $\hat{\theta}_{(\text{lower})}^*$ and $\hat{\theta}_{(\text{upper})}^*$, where lower = $ra/2$ and upper = $r(1 - a/2)$. If lower and upper are not integers, then we can interpolate between adjacent ordered values $\hat{\theta}_{(b)}^*$ or round off to the nearest integer.

A nonparametric confidence interval for θ can be constructed from the quantiles of the bootstrap sampling distribution of $\hat{\theta}^*$. The 95% percentile interval is $\hat{\theta}_{(\text{lower})}^* < \theta < \hat{\theta}_{(\text{upper})}^*$, where the $\hat{\theta}_{(b)}^*$ are the r ordered bootstrap replicates; lower = $.025 \times r$ and upper = $.975 \times r$.

A 95% confidence interval for the $r = 2000$ resampled means in Figure 21.2, for example, is constructed as follows:

$$\begin{aligned}\text{lower} &= 2000(.05/2) = 50 \\ \text{upper} &= 2000(1 - .05/2) = 1950 \\ \bar{Y}_{(50)}^* &= 0.7 \\ \bar{Y}_{(1950)}^* &= 7.8 \\ 0.7 < \mu < 7.8\end{aligned}$$

The endpoints of this interval are marked in Figure 21.2. Because of the skew of the bootstrap distribution, the percentile interval is not quite symmetric around $\bar{Y} = 4.6$. By way of comparison, the standard t -interval for the mean of the original sample of 10 observations is

$$\begin{aligned}\mu &= \bar{Y} \pm t_{9,.025} \text{SE}(\bar{Y}) \\ &= 4.6 \pm 2.262 \times 1.881 \\ &= 4.6 \pm 4.255 \\ 0.345 < \mu < 8.855\end{aligned}$$

In this case, the standard interval is a bit wider than the percentile interval, especially at the top.

21.2.3 Improved Bootstrap Intervals

I will briefly describe an adjustment to percentile intervals that improves their accuracy.¹³ As before, we want to produce a $100(1 - a)\%$ confidence interval for θ having computed the sample estimate $\hat{\theta}$ and bootstrap replicates $\hat{\theta}_b^*$, $b = 1, \dots, r$. We require $z_{a/2}$, the unit-normal value with probability $a/2$ to the right, and two “correction factors,” Z and A , defined in the following manner:

¹³The interval described here is called a “bias-corrected, accelerated” (or BC_a) percentile interval. Details can be found in Efron and Tibshirani (1993, chap. 14); also see Stine (1990) for a discussion of different procedures for constructing bootstrap confidence intervals.

- Calculate

$$Z \equiv \Phi^{-1} \left[\frac{\sum_{b=1}^r (\widehat{\theta}_b^* < \widehat{\theta})}{r} \right]$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard-normal distribution function (i.e., the standard-normal quantile function), and $\#(\widehat{\theta}_b^* < \widehat{\theta})/r$ is the proportion of bootstrap replicates below the estimate $\widehat{\theta}$. If the bootstrap sampling distribution is symmetric and if $\widehat{\theta}$ is unbiased, then this proportion will be close to .5, and the “correction factor” Z will be close to 0.

- Let $\widehat{\theta}_{(-i)}$ represent the value of $\widehat{\theta}$ produced when the i th observation is deleted from the sample;¹⁴ there are n of these quantities. Let $\bar{\theta}$ represent the average of the $\widehat{\theta}_{(-i)}$; that is, $\bar{\theta} \equiv \sum_{i=1}^n \widehat{\theta}_{(-i)}/n$. Then calculate

$$A \equiv \frac{\sum_{i=1}^n (\bar{\theta} - \widehat{\theta}_{(-i)})^3}{6 \left[\sum_{i=1}^n (\bar{\theta} - \widehat{\theta}_{(-i)})^2 \right]^{3/2}} \quad (21.2)$$

With the correction factors Z and A in hand, compute

$$\begin{aligned} A_1 &\equiv \Phi \left[Z + \frac{Z - z_{a/2}}{1 - A(Z - z_{a/2})} \right] \\ A_2 &\equiv \Phi \left[Z + \frac{Z + z_{a/2}}{1 - A(Z + z_{a/2})} \right] \end{aligned}$$

where $\Phi(\cdot)$ is the standard-normal cumulative distribution function. When the correction factors Z and A are both 0, $A_1 = \Phi(-z_{a/2}) = a/2$, and $A_2 = \Phi(z_{a/2}) = 1 - a/2$. The values A_1 and A_2 are used to locate the endpoints of the corrected percentile confidence interval. In particular, the corrected interval is

$$\widehat{\theta}_{(\text{lower}^*)}^* < \theta < \widehat{\theta}_{(\text{upper}^*)}^*$$

where $\text{lower}^* = rA_1$ and $\text{upper}^* = rA_2$ (rounding or interpolating as required).

The lower and upper bounds of percentile confidence intervals can be corrected to improve the accuracy of these intervals.

Applying this procedure to the “data” in Table 21.3, we have $z_{.05/2} = 1.96$ for a 95% confidence interval. There are 926 bootstrapped means below $\bar{Y} = 4.6$, and so $Z = \Phi^{-1}(926/2000) = -0.09288$. The $\bar{Y}_{(-i)}$ are 4.444, 5.444, ..., 4.111; the mean of these

¹⁴The $\widehat{\theta}_{(-i)}$ are called the *jackknife values* of the statistic $\widehat{\theta}$. The jackknife values can also be used as an alternative to the bootstrap to find a nonparametric confidence interval for θ . See Exercise 21.2.

values is $\bar{\bar{Y}} = \bar{Y} = 4.6$,¹⁵ and (from Equation 21.2) $A = -0.05630$. Using the correction factors z and A ,

$$\begin{aligned} A_1 &= \Phi \left\{ -0.09288 + \frac{-0.09288 - 1.96}{1 - [-0.05630(-0.09288 - 1.96)]} \right\} \\ &= \Phi(-2.414) = 0.007889 \\ A_2 &= \Phi \left\{ -0.09288 + \frac{-0.09288 + 1.96}{1 - [-0.05630(-0.09288 + 1.96)]} \right\} \\ &= \Phi(1.597) = 0.9449 \end{aligned}$$

Multiplying by r , we have $2000 \times 0.007889 \approx 16$ and $2000 \times 0.9449 \approx 1890$, from which

$$\begin{aligned} \bar{Y}_{(16)}^* < \mu < \bar{Y}_{(1890)}^* \\ -0.4 < \mu < 7.3 \end{aligned} \tag{21.3}$$

Unlike the other confidence intervals that we have calculated for the “sample” of 10 differences in income between husbands and wives, the interval given in Equation 21.3 includes 0.

21.3 Bootstrapping Regression Models

The procedures of the previous section can be easily extended to regression models. The most straightforward approach is to collect the response-variable value and regressors for each observation,

$$\mathbf{z}'_i \equiv [Y_i, X_{i1}, \dots, X_{ik}]$$

Then the observations $\mathbf{z}'_1, \mathbf{z}'_2, \dots, \mathbf{z}'_n$ can be resampled, and the regression estimator computed for each of the resulting bootstrap samples, $\mathbf{z}'_{b1}', \mathbf{z}'_{b2}', \dots, \mathbf{z}'_{bn}'$, producing r sets of bootstrap regression coefficients, $\mathbf{b}_b^* = [A_b^*, B_{b1}^*, \dots, B_{bk}^*]'$. The methods of the previous section can be applied to compute standard errors or confidence intervals for the regression estimates.

Directly resampling the observations \mathbf{z}'_i implicitly treats the regressors X_1, \dots, X_k as *random* rather than *fixed*. We may want to treat the X s as fixed (if, e.g., the data derive from an experimental design). In the case of linear regression, for example,

1. Estimate the regression coefficients A, B_1, \dots, B_k for the original sample, and calculate the fitted value and residual for each observation:

$$\begin{aligned} \hat{Y}_i &= A + B_1 x_{i1} + \dots + B_k x_{ik} \\ E_i &= Y_i - \hat{Y}_i \end{aligned}$$

2. Select bootstrap samples of the *residuals*, $\mathbf{e}_b^* = [E_{b1}^*, E_{b2}^*, \dots, E_{bn}^*]'$, and from these, calculate bootstrapped Y values, $\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, \dots, Y_{bn}^*]'$, where $Y_{bi}^* = \hat{Y}_i + E_{bi}^*$.
3. Regress the bootstrapped Y values on the *fixed* X -values to obtain bootstrap regression coefficients.

¹⁵The average of the jackknifed estimates is not, in general, the same as the estimate calculated for the full sample, but this is the case for the jackknifed sample means. See Exercise 21.2.

- *If, for example, estimates are calculated by least-squares regression, then $\mathbf{b}_b^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_b^*$ for $b = 1, \dots, r$.
4. The resampled $\mathbf{b}_b^* = [A_b^*, B_{b1}^*, \dots, B_{bk}^*]'$ can be used in the usual manner to construct bootstrap standard errors and confidence intervals for the regression coefficients.

Bootstrapping with fixed X draws an analogy between the fitted value \hat{Y} in the sample and the conditional expectation of Y in the population, as well as between the residual E in the sample and the error ε in the population. Although no assumption is made about the *shape* of the error distribution, the bootstrapping procedure, by constructing the Y_{bi}^* according to the linear model, implicitly assumes that the functional form of the model is correct.

Furthermore, by resampling residuals and randomly reattaching them to fitted values, the procedure implicitly assumes that the errors are *identically distributed*. If, for example, the true errors have nonconstant variance, then this property will *not* be reflected in the resampled residuals. Likewise, the unique impact of a high-leverage outlier will be lost to the resampling.¹⁶

Regression models and similar statistical models can be bootstrapped by (1) treating the regressors as random and selecting bootstrap samples directly from the observations $\mathbf{z}'_i = [Y_i, X_{i1}, \dots, X_{ik}]$, or (2) treating the regressors as fixed and resampling from the residuals E_i of the fitted regression model. In the latter instance, bootstrap observations are constructed as $Y_{bi}^* = \hat{Y}_i + E_{bi}^*$, where the \hat{Y}_i are the fitted values from the original regression, and the E_{bi}^* are the resampled residuals for the b th bootstrap sample. In each bootstrap sample, the Y_{bi}^* are then regressed on the original X s. A disadvantage of fixed- X resampling is that the procedure implicitly assumes that the functional form of the regression model fit to the data is correct and that the errors are identically distributed.

To illustrate bootstrapping regression coefficients, I will use Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations.¹⁷ The Huber M estimator applied to Duncan's regression produces the following fit, with asymptotic standard errors shown in parentheses beneath each coefficient:¹⁸

$$\widehat{\text{Prestige}} = -7.289 + 0.7104 \text{ Income} + 0.4819 \text{ Education}$$

$$(3.588) \quad (0.1005) \quad (0.0825)$$

Using random- X resampling, I drew $r = 2000$ bootstrap samples, calculating the Huber estimator for each bootstrap sample. The results of this computationally intensive procedure are summarized in Table 21.5. The distributions of the bootstrapped regression coefficients for income and education are graphed in Figure 21.3(a) and (b), along with the percentile confidence intervals for these coefficients. Figure 21.3(c) shows a scatterplot of the bootstrapped coefficients

¹⁶For these reasons, random- X resampling may be preferable even if the X -values are best conceived as fixed. See Exercise 21.3.

¹⁷These data were discussed in Chapter 19 on robust regression and at several other points in this text.

¹⁸ M estimation is a method of robust regression described in Section 19.1.

Table 21.5 Statistics for $r = 2000$ Bootstrapped Huber Regressions Applied to Duncan's Occupational Prestige Data

	Coefficient		
	Constant	Income	Education
Average bootstrap estimate	-7.001	0.6903	0.4918
Bootstrap standard error	3.165	0.1798	0.1417
Asymptotic standard error	3.588	0.1005	0.0825
Normal-theory interval	(-13.423, -1.018)	(0.3603, 1.0650)	(0.2013, 0.7569)
Percentile interval	(-13.150, -0.577)	(0.3205, 1.0331)	(0.2030, 0.7852)
Adjusted percentile interval	(-12.935, -0.361)	(0.2421, 0.9575)	(0.2511, 0.8356)

NOTES: Three bootstrap confidence intervals are shown for each coefficient. Asymptotic standard errors are also shown for comparison.

for income and education, which gives a sense of the covariation of the two estimates; it is clear that the income and education coefficients are strongly negatively correlated.¹⁹

The bootstrap standard errors of the income and education coefficients are much larger than the asymptotic standard errors, underscoring the inadequacy of the latter in small samples. The simple normal-theory confidence intervals based on the bootstrap standard errors (and formed as the estimated coefficients ± 1.96 standard errors) are reasonably similar to the percentile intervals for the income and education coefficients; the percentile intervals differ slightly from the adjusted percentile intervals. Comparing the average bootstrap coefficients \bar{A}^* , \bar{B}_1^* , and \bar{B}_2^* with the corresponding estimates A , B_1 , and B_2 suggests that there is little, if any, bias in the Huber estimates.²⁰

21.4 Bootstrap Hypothesis Tests*

In addition to providing standard errors and confidence intervals, the bootstrap can also be used to test statistical hypotheses. The application of the bootstrap to hypothesis testing is more or less obvious for individual coefficients because a bootstrap confidence interval can be used to test the hypothesis that the corresponding parameter is equal to any specific value (typically 0 for a regression coefficient).

More generally, let $T \equiv t(\mathbf{z})$ represent a test statistic, written as a function of the sample \mathbf{z} . The contents of \mathbf{z} vary by context. In regression analysis, for example, \mathbf{z} is the $n \times k + 1$ matrix $[\mathbf{y}, \mathbf{X}]$ containing the response variable and the regressors.

For concreteness, suppose that T is the Wald-like test statistic for the omnibus null hypothesis $H_0: \beta_1 = \dots = \beta_k = 0$ in a robust regression, calculated using the estimated asymptotic covariance matrix for the regression coefficients. That is, let \mathbf{V}_{11} contain the rows and $(k \times k)$

¹⁹The negative correlation of the coefficients reflects the *positive* correlation between income and education (see Section 9.4.4). The hint of bimodality in the distribution of the income coefficient suggests the possible presence of influential observations. See the discussion of Duncan's regression in Section 4.6.

²⁰For the use of the bootstrap to estimate bias, see Exercise 21.4.

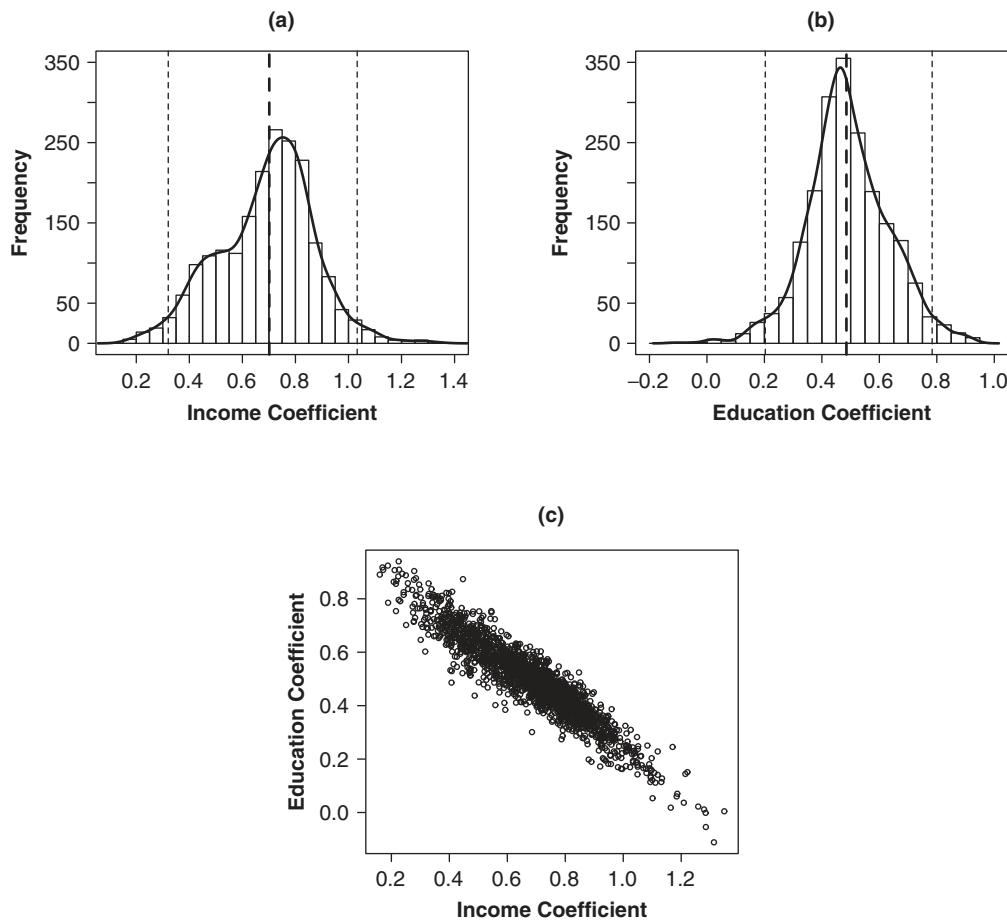


Figure 21.3 Panels (a) and (b) show histograms and kernel density estimates for the $r = 2000$ bootstrap replicates of the income and education coefficients in Duncan's occupational prestige regression. The regression model was fit by M estimation using the Huber weight function. Panel (c) shows a scatterplot of the income and education coefficients for the 2000 bootstrap samples.

columns of the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$ that pertain to the k slope coefficients $\mathbf{b}_1 = [B_1, \dots, B_k]'$. We can write the null hypothesis as $H_0: \beta_1 = \mathbf{0}$. Then the test statistic is

$$T = \mathbf{b}'_1 \mathbf{V}_{11}^{-1} \mathbf{b}_1$$

We could compare the obtained value of this statistic to the quantiles of χ^2_k , but we are loath to do so because we do not trust the asymptotics. We can, instead, construct the sampling distribution of the test statistic nonparametrically, using the bootstrap.

Let $T_b^* \equiv t(\mathbf{z}_b^*)$ represent the test statistic calculated for the b th bootstrap sample, \mathbf{z}_b^* . We have to be careful to draw a proper analogy here: Because the original-sample estimates play the role of the regression parameters in the bootstrap “population” (i.e., the original sample), the

bootstrap analog of the null hypothesis—to be used with each bootstrap sample—is $H_0: \beta_1 = B_1, \dots, \beta_k = B_k$. The bootstrapped test statistic is, therefore,

$$T_b^* = (\mathbf{b}_{b1}^* - \mathbf{b}_1)' \mathbf{V}_{b,11}^{*-1} (\mathbf{b}_{b1}^* - \mathbf{b}_1)$$

Having obtained r bootstrap replications of the test statistic, the bootstrap estimate of the p -value for H_0 is simply²¹

$$\hat{p}^* = \frac{\#_{b=1}^r (T_b^* \geq T)}{r}$$

Note that for this chi-square-like test, the p -value is entirely from the upper tail of the distribution of the bootstrapped test statistics.

Bootstrap hypothesis tests proceed by constructing an empirical sampling distribution for the test statistic. If T represents the test statistic computed for the original sample, and T_b^* is the test statistic for the b th of r bootstrap samples, then (for a chi-square-like test statistic) the p -value for the test is $\#(T_b^* \geq T)/r$.

21.5 Bootstrapping Complex Sampling Designs

One of the great virtues of the bootstrap is that it can be applied in a natural manner to more complex sampling designs.²² If, for example, the population is divided into S strata, with n_s observations drawn from stratum s , then bootstrap samples can be constructed by resampling n_s observations with replacement from the s th stratum. Likewise, if observations are drawn into the sample in *clusters* rather than individually, then the bootstrap should resample clusters rather than individuals. We can still calculate estimates and test statistics in the usual manner using the bootstrap to assess sampling variation in place of the standard formulas, which are appropriate for independent random samples but not for complex survey samples.

When different observations are selected for the sample with unequal probabilities, it is common to take account of this fact by differentially weighting the observations in inverse proportion to their probability of selection.²³ Thus, for example, in calculating the (weighted) sample mean of a variable Y , we take

$$\bar{Y}^{(w)} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

and to calculate the (weighted) correlation of X and Y , we take

²¹There is a subtle point here: We use the sample estimate \mathbf{b}_1 in place of the hypothesized parameter $\beta_1^{(0)}$ to calculate the bootstrapped test statistic T_b^* *regardless* of the hypothesis that we are testing—because in the central bootstrap analogy \mathbf{b}_1 stands in for β_1 (and the bootstrapped sampling distribution of the test statistic is computed under the assumption that the hypothesis is *true*). See Exercise 21.5 for an application of this test to Duncan’s regression.

²²Analytic methods for statistical inference in complex surveys are described briefly in Section 15.5.

²³These “case weights” are to be distinguished from the variance weights used in weighted least-squares regression (see Section 12.2.2). Survey case weights are described in Section 15.5.

$$r_{XY}^{(w)} = \frac{\sum w_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum w_i(X_i - \bar{X})^2][\sum w_i(Y_i - \bar{Y})^2]}}$$

Other statistical formulas can be adjusted analogously.²⁴

The case weights are often scaled so that $\sum w_i = n$, but simply incorporating the weights in the usual formulas for standard errors does not produce correct results. Once more, the bootstrap provides a straightforward solution: Draw bootstrap samples in which the probability of inclusion is proportional to the probability of inclusion in the original sample, and calculate bootstrap replicates of the statistics of interest using the case weights.

The essential “trick” of using the bootstrap in these (and other) instances is to resample from the data in the same way as the original sample was drawn from the population. Statistics are calculated for each bootstrap replication in the same manner as for the original sample.

The bootstrap can be applied to many complex sampling designs (involving, e.g., stratification, clustering, and case weighting) by resampling from the sample data in the same manner as the original sample was selected from the population.

Social scientists frequently analyze data from complex sampling designs as if they originate from independent random samples (even though there are often nonnegligible dependencies among the observations) or employ ad hoc adjustments (e.g., by weighting). A tacit defense of common practice is that to take account of the dependencies in complex sampling designs is too difficult. The bootstrap provides a simple solution.²⁵

21.6 Concluding Remarks

If the bootstrap is so simple and of such broad application, why isn’t it used more in the social sciences? Beyond the problem of lack of familiarity (which surely can be remedied), there are, I believe, three serious obstacles to increased use of the bootstrap:

1. Common practice—such as relying on asymptotic results in small samples or treating dependent data as if they were independent—usually *understates* sampling variation and makes results look stronger than they really are. Researchers are understandably reluctant to report honest standard errors when the usual calculations indicate greater precision. It is best, however, not to fool yourself, regardless of what you think about fooling others.
2. Although the conceptual basis of the bootstrap is intuitively simple and although the calculations are straightforward, to apply the bootstrap, it is necessary to write or find suitable statistical software. There is some bootstrapping software available, but the nature of the bootstrap—which adapts resampling to the data collection plan and

²⁴See Exercise 21.6.

²⁵Alternatively, we can use sampling-variance estimates that are appropriate to complex survey samples, as described in Section 15.5.

statistics employed in an investigation—apparently precludes full generality and makes it difficult to use traditional statistical computer packages. After all, researchers are not tediously going to draw 2000 samples from their data unless a computer program can fully automate the process. This impediment is much less acute in programmable statistical computing environments.²⁶

3. Even with good software, the bootstrap is computationally intensive. This barrier to bootstrapping is more apparent than real, however. Computational speed is central to the exploratory stages of data analysis: When the outcome of one of many small steps immediately affects the next, rapid results are important. This is why a responsive computing environment is especially useful for regression diagnostics, for example. It is not nearly as important to calculate standard errors and p -values quickly. With powerful, yet relatively inexpensive, desktop computers, there is nothing to preclude the machine from cranking away unattended for a few hours (although that is rarely necessary—a few minutes is more typical). The time and effort involved in a bootstrap calculation are usually small compared with the totality of a research investigation—and are a small price to pay for accurate and realistic inference.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 21.1. *Show that the mean of the n^n bootstrap means is the sample mean

$$E^*(\bar{Y}^*) = \frac{\sum_{b=1}^{n^n} \bar{Y}_b^*}{n^n} = \bar{Y}$$

and that the standard deviation (standard error) of the bootstrap means is

$$\text{SE}^*(\bar{Y}^*) = \sqrt{\frac{\sum_{b=1}^{n^n} (\bar{Y}_b^* - \bar{Y})^2}{n^n}} = \frac{S}{\sqrt{n-1}}$$

where $S = \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)}$ is the sample standard deviation. (*Hint:* Exploit the fact that the mean is a linear function of the observations.)

Exercise 21.2. The jackknife: The “jackknife” (suggested for estimation of standard errors by Tukey, 1958) is an alternative to the bootstrap that requires less computation, but that often does not perform as well and is not quite as general. Efron and Tibshirani (1993, chap. 11) show that the jackknife is an approximation to the bootstrap. Here is a brief description of the jackknife for the estimator $\hat{\theta}$ of a parameter θ :

1. Divide the sample into m independent groups. In most instances (unless the sample size is very large), we take $m = n$, in which case each observation constitutes a “group.” If the data originate from a cluster sample, then the observations in a cluster should be kept together.

²⁶See, for example, the bootstrapping software for the S and R statistical computing environments described by Efron and Tibshirani (1993, appendix) and by Davison and Hinkley (1997, chap. 11). General bootstrapping facilities are also provided in the Stata programming environment.

2. Recalculate the estimator omitting the j th group, $j = 1, \dots, m$, denoting the resulting value of the estimator as $\hat{\theta}_{(-j)}$. The *pseudo-value* associated with the j th group is defined as $\hat{\theta}_j^* \equiv m\hat{\theta} - (m-1)\hat{\theta}_{(-j)}$.
3. The average of the pseudo-values, $\hat{\theta}^* \equiv (\sum_{j=1}^m \hat{\theta}_j^*)/m$, is the jackknifed estimate of θ . A jackknifed $100(1-a)\%$ confidence interval for θ is given by

$$\theta = \hat{\theta}^* \pm t_{a/2, m-1} \frac{S^*}{\sqrt{n}}$$

where $t_{a/2, m-1}$ is the critical value of t with probability $a/2$ to the right for $m-1$ degrees of freedom, and $S^* \equiv \sqrt{\sum_{j=1}^m (\hat{\theta}_j^* - \hat{\theta}^*)^2 / (m-1)}$ is the standard deviation of the pseudo-values.

- (a) *Show that when the jackknife procedure is applied to the mean with $m = n$, the pseudo-values are just the original observations, $\hat{\theta}_i^* = Y_i$; the jackknifed estimate $\hat{\theta}^*$ is, therefore, the sample mean \bar{Y} ; and the jackknifed confidence interval is the same as the usual t confidence interval.
- (b) Demonstrate the results in part (a) numerically for the contrived “data” in Table 21.3. (These results are peculiar to linear statistics like the mean.)
- (c) Find jackknifed confidence intervals for the Huber M estimator of Duncan’s regression of occupational prestige on income and education. Compare these intervals with the bootstrap and normal-theory intervals given in Table 21.5.

Exercise 21.3. Random versus fixed resampling in regression:

- (a) Recall (from Chapter 2) Davis’s data on measured and reported weight for 101 women engaged in regular exercise. Bootstrap the least-squares regression of reported weight on measured weight, drawing $r = 1000$ bootstrap samples using (1) random- X resampling and (2) fixed- X resampling. In each case, plot a histogram (and, if you wish, a density estimate) of the 1000 bootstrap slopes, and calculate the bootstrap estimate of standard error for the slope. How does the influential outlier in this regression affect random resampling? How does it affect fixed resampling?
- (b) Randomly construct a data set of 100 observations according to the regression model $Y_i = 5 + 2x_i + \varepsilon_i$, where $x_i = 1, 2, \dots, 100$, and the errors are independent (but seriously heteroscedastic), with $\varepsilon_i \sim N(0, x_i^2)$. As in (a), bootstrap the least-squares regression of Y on x , using (1) random resampling and (2) fixed resampling. In each case, plot the bootstrap distribution of the slope coefficient, and calculate the bootstrap estimate of standard error for this coefficient. Compare the results for random and fixed resampling. For a few of the bootstrap samples, plot the least-squares residuals against the fitted values. How do these plots differ for fixed versus random resampling?
- (c) Why might random resampling be preferred in these contexts, even if (as is *not* the case for Davis’s data) the X -values are best conceived as fixed?

Exercise 21.4. Bootstrap estimates of bias: The bootstrap can be used to estimate the bias of an estimator $\hat{\theta}$ of a parameter θ , simply by comparing the mean of the bootstrap distribution $\bar{\theta}^*$ (which stands in for the expectation of the estimator) with the sample estimate $\hat{\theta}$ (which stands

in for the parameter); that is, $\widehat{\text{bias}} = \bar{\theta}^* - \hat{\theta}$. (Further discussion and more sophisticated methods are described in Efron & Tibshirani, 1993, chap. 10.) Employ this approach to estimate the bias of the maximum-likelihood estimator of the variance, $\hat{\sigma}^2 = \sum (Y_i - \bar{Y})^2/n$, for a sample of $n = 10$ observations drawn from the normal distribution $N(0, 100)$. Use $r = 500$ bootstrap replications. How close is the bootstrap bias estimate to the theoretical value $-\sigma^2/n = -100/10 = -10$?

Exercise 21.5. *Test the omnibus null hypothesis $H_0: \beta_1 = \beta_2 = 0$ for the Huber M estimator in Duncan's regression of occupational prestige on income and education.

- (a) Base the test on the estimated asymptotic covariance matrix of the coefficients.
- (b) Use the bootstrap approach described in Section 21.4.

Exercise 21.6. Case weights:

- (a) *Show how case weights can be used to "adjust" the usual formulas for the least-squares coefficients and their covariance matrix. How do these case-weighted formulas compare with those for weighted-least-squares regression (discussed in Section 12.2.2.)?
- (b) Using data from a sample survey that employed disproportional sampling and for which case weights are supplied, estimate a least-squares regression (1) ignoring the case weights, (2) using the case weights to estimate both the regression coefficients and their standard errors (rescaling the case weights, if necessary, so that they sum to the sample size), and (3) using the case weights but estimating coefficient standard errors with the bootstrap. Compare the estimates and standard errors obtained in (1), (2), and (3).

Exercise 21.7. *Bootstrapping time-series regression: Bootstrapping can be adapted to time-series regression but, as in the case of fixed- X resampling, the procedure makes strong use of the model fit to the data—in particular, the manner in which serial dependency in the data is modeled. Suppose that the errors in the linear model $y = \mathbf{X}\beta + \varepsilon$ follow a first-order autoregressive process (see Chapter 16), $\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i$; the ν_i are independently and identically distributed with 0 expectations and common variance σ_ν^2 . Suppose further that we use the method of maximum likelihood to obtain estimates $\hat{\rho}$ and $\hat{\beta}$. From the residuals $e = y - \mathbf{X}\hat{\beta}$, we can estimate ν_i as $V_i = E_i - \hat{\rho}E_{i-1}$ for $i = 2, \dots, n$; by convention, we take $V_1 = E_1$. Then, for each bootstrap replication, we sample n -values with replacement from the V_i ; call them $V_{b1}^*, V_{b2}^*, \dots, V_{bn}^*$. Using these values, we construct residuals $E_{b1}^* = V_{b1}^*$ and $E_{bi}^* = \hat{\rho}E_{b,i-1}^* + V_{bi}^*$ for $i = 2, \dots, n$; and from these residuals and the original fitted values $\hat{Y}_i = \mathbf{x}'_i\hat{\beta}$, we construct bootstrapped Y -values, $Y_{bi}^* = \hat{Y}_i + E_{bi}^*$. The Y_{bi}^* are used along with the original \mathbf{x}'_i to obtain bootstrap replicates $\hat{\beta}_b^*$ of the ML coefficient estimates. (Why are the \mathbf{x}'_i treated as fixed?) Employ this procedure to compute standard errors of the coefficient estimates in the time-series regression for the Canadian women's crime rate data (discussed in Chapter 16), using an AR(1) process for the errors. Compare the bootstrap standard errors with the usual asymptotic standard errors. Which standard errors do you prefer? Why? Then describe a bootstrap procedure for a time-series regression model with AR(2) errors, and apply this procedure to the Canadian women's crime rate regression.

Summary

- Bootstrapping is a broadly applicable, nonparametric approach to statistical inference that substitutes intensive computation for more traditional distributional assumptions and asymptotic results. The bootstrap can be used to derive accurate standard errors, confidence intervals, and hypothesis tests for most statistics.
- Bootstrapping uses the sample data to estimate relevant characteristics of the population. The sampling distribution of a statistic is then constructed empirically by resampling from the sample. The resampling procedure is designed to parallel the process by which sample observations were drawn from the population. For example, if the data represent an independent random sample of size n (or a simple random sample of size n from a much larger population), then each bootstrap sample selects n observations with replacement from the original sample. The key bootstrap analogy is the following: *The population is to the sample as the sample is to the bootstrap samples.*
- Having produced r bootstrap replicates $\hat{\theta}_b^*$ of an estimator $\hat{\theta}$, the bootstrap standard error is the standard deviation of the bootstrap replicates:

$$\text{SE}^*(\hat{\theta}^*) = \sqrt{\frac{\sum_{b=1}^r (\hat{\theta}_b^* - \bar{\theta}^*)^2}{r - 1}}$$

where $\bar{\theta}^*$ is the mean of the $\hat{\theta}_b^*$. In large samples, where we can rely on the normality of $\hat{\theta}$, a 95% confidence interval for θ is given by $\hat{\theta} \pm 1.96 \text{SE}^*(\hat{\theta}^*)$.

- A nonparametric confidence interval for θ can be constructed from the quantiles of the bootstrap sampling distribution of $\hat{\theta}^*$. The 95% percentile interval is $\hat{\theta}_{(\text{lower})}^* < \theta < \hat{\theta}_{(\text{upper})}^*$, where the $\hat{\theta}_{(b)}^*$ are the r ordered bootstrap replicates; lower = $.025 \times r$ and upper = $.975 \times r$.
- The lower and upper bounds of percentile confidence intervals can be corrected to improve the accuracy of these intervals.
- Regression models can be bootstrapped by (1) treating the regressors as random and selecting bootstrap samples directly from the observations $\mathbf{z}'_i = [Y_i, X_{i1}, \dots, X_{ik}]$, or (2) treating the regressors as fixed and resampling from the residuals E_i of the fitted regression model. In the latter instance, bootstrap observations are constructed as $Y_{bi}^* = \hat{Y}_i + E_{bi}^*$, where the \hat{Y}_i are the fitted values from the original regression, and the E_{bi}^* are the resampled residuals for the b th bootstrap sample. In each bootstrap sample, the Y_{bi}^* are then regressed on the original X s. A disadvantage of fixed- X resampling is that the procedure implicitly assumes that the regression model fit to the data is correct and that the errors are identically distributed.
- Bootstrap hypothesis tests proceed by constructing an empirical sampling distribution for the test statistic. If T represents the test statistic computed for the original sample and T_b^* is the test statistic for the b th of r bootstrap samples, then (for a chi-square-like test statistic) the p -value for the test is $\#(T_b^* \geq T)/r$.
- The bootstrap can be applied to many complex sampling designs (involving, e.g., stratification, clustering, and case weighting) by resampling from the sample data in the same manner as the original sample was selected from the population.

Recommended Reading

Bootstrapping is a rich topic; the presentation in this chapter has stressed computational procedures at the expense of a detailed account of statistical properties and limitations.

- Although Efron and Tibshirani's (1993) book on the bootstrap contains some relatively advanced material, most of the exposition requires only modest statistical background and is eminently readable.
- Davison and Hinkley (1997) is another statistically sophisticated, comprehensive treatment of bootstrapping.
- A briefer source on bootstrapping, addressed to social scientists, is Stine (1990), which includes a fine discussion of the rationale of bootstrap confidence intervals.
- Young's (1994) paper and the commentary that follows it focus on practical difficulties in applying the bootstrap.

22

Model Selection, Averaging, and Validation

This chapter addresses practical issues in building statistical models. The first section of the chapter discusses criteria for selecting among statistical models—criteria that move beyond hypothesis tests for terms in a model.

The second section deals with an alternative approach, termed *model averaging*, that combines information from different statistical models fit to the same data.

Model validation, which is described in the third section of the chapter, provides a simple basis for honest statistical inference when—as is typically the case in a careful investigation—we need to examine the data to formulate a descriptively adequate statistical model. In validation, the data are divided at random into two parts: One part is used for data exploration and model formulation (including, possibly, model selection); the second part is used to evaluate the model, thus preserving the integrity of statistical inferences.

22.1 Model Selection

I have touched, in passing, on issues of model selection at several points in the text, often simplifying a model after preliminary statistical hypothesis tests.¹ Issues of model search extend beyond the selection of explanatory variables or terms to include in a regression model to questions such as the removal of outliers and variable transformations. The strategy of basing model selection on hypothesis tests is problematic for a number of reasons (largely familiar from elementary statistics):

- *Simultaneous inference:* If we are testing many terms simultaneously, the probability of rejecting one or more true null hypotheses by chance (i.e., the probability of committing a Type I error) is larger—possibly much larger—than the level of any individual test.²
- *The fallacy of affirming the consequent:* Failing to reject a null hypothesis is to be distinguished from demonstrating that the null hypothesis is supported by the data. For example, the power of the test may be weak. It is important to distinguish, therefore, a small coefficient that is precisely estimated (where, e.g., the confidence interval for the coefficient is narrow and includes 0) from a coefficient that is imprecisely estimated (where, e.g., the size of the estimated coefficient may be large, but the confidence interval includes zero). Eliminating an imprecisely estimated term from a model can

¹In addition, the Bayesian information criterion (BIC) was used for model selection in Section 13.2.2, and cross-validation and generalized cross-validation were discussed in the context of nonparametric regression in Chapter 18.

²See Exercise 22.1.

seriously bias other estimates if the population coefficient is large and the variable in question is strongly related to others in the model.³

- *The impact of large samples on hypothesis tests:* In the social sciences, we rarely expect a null hypothesis to be *exactly* correct, and given a sufficiently large sample, a false hypothesis—even one that is nearly correct—will be rejected with high probability. Thus, hypothesis testing may lead us to include terms in a model because they are “statistically significant” even when they are trivially small.
- *Exaggerated precision:* Coefficient standard errors computed after model selection tend to overstate the precision of estimation when terms correlated with those retained in the model are eliminated. Consequently, confidence intervals are artificially narrow and *p*-values artificially small.

There are several general strategies for addressing these concerns:

- *Using alternative model-selection criteria:* One approach—namely, to employ a criterion other than statistical significance to decide questions of model selection—is the principal subject of this section.
- *Compensating for simultaneous inference:* We can seek to compensate for simultaneous inference by employing a Bonferroni adjustment, for example, or by holding back some of our data to validate a statistical model selected by another approach.⁴
- *Avoiding model selection:* Still another strategy, which preserves the integrity of classical statistical inference, is to specify and interpret a maximally complex and flexible model without seeking to simplify it.⁵ Although this is a defensible approach, it encounters two general difficulties, in my opinion: (1) We are often not in a position to specify a fully adequate model, even a maximally complex one, prior to examining the data, and (2) retaining what prove to be unnecessary terms in a model contradicts a common goal of statistical data analysis, which is permissible simplification (possibly begging the question of what is “permissible”).
- *Model averaging:* Rather than selecting a single model and discarding all others, model-averaging techniques seek to account for model uncertainty by weighting contending models according to their relative degree of support from the data.⁶

It is problematic to use statistical hypothesis tests for model selection. Doing so leads to issues of simultaneous inference, can produce biased results, tends to yield complicated models in large samples, and exaggerates the precision of results.

³See the discussions of “specification errors” in Sections 6.3 and 9.7.

⁴Model validation is the subject of Section 22.3. For an example of the use of Bonferroni adjustment in model selection, see Foster and Stine (2004).

⁵This strategy is advocated, for example, by Harrell (2001).

⁶See Section 22.2.

22.1.1 Model Selection Criteria

Model selection is conceptually simplest when our goal is *prediction*—that is, the development of a regression model that will predict new data as accurately as possible. Although predictive accuracy is a desirable characteristic in any statistical model, most interesting statistical problems in the social sciences are not pure prediction problems, and a more typical objective is to use a statistical model for substantive interpretation, data summary, and explanation.

When our goal is prediction, we need not be concerned about the consequences for interpretation of eliminating an explanatory variable that is highly correlated with one included in the model, because the excluded variable would typically lend little additional predictive power to the model. This conclusion assumes that the configuration of explanatory variables will be similar in the data for which predictions are desired as in the data used to calibrate the model. Likewise, where the aim is prediction, we need have no qualms about including among the *predictor* variables (note, in this context, not “explanatory” variables) *symptoms* (i.e., effects) of the response variable. Indeed, the inclusion of symptoms as predictors is standard in areas such as differential diagnosis in medicine.

Because of the current expansion of computer power and the availability of very large data sets (e.g., in genomics and in “big data” collected on the Internet), model selection problems in the context of prediction are receiving a great deal of attention in statistics: Once an epithet, “data mining” is now a topic of serious study.⁷

This section describes several criteria that have been suggested for selecting among competing statistical models.⁸ I assume that we have n observations on a response variable Y and associated predictors, X s, producing a set of m contending statistical models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$ for Y .

Corrected R²

The *squared multiple correlation “corrected”* (or “adjusted”) for degrees of freedom is an intuitively reasonable criterion for comparing linear-regression models with different numbers of parameters.⁹ Suppose that model M_j is one of the models under consideration. If M_j has s_j regression coefficients (including the regression constant) and is fit to a data set with n observations, then the corrected R^2 for the model is

$$\begin{aligned}\tilde{R}_j^2 &\equiv 1 - \frac{S_E^{(j)2}}{S_Y^2} \\ &= 1 - \frac{n-1}{n-s_j} \times \frac{\text{RSS}_j}{\text{TSS}}\end{aligned}$$

where RSS_j is the residual sum of squares under the model, $\text{TSS} = \sum (Y_i - \bar{Y})^2$ is the total sum of squares for the response variable Y , and $S_E^{(j)2} = \text{RSS}_j/(n-s_j)$ is the estimated error variance. Consequently, models with relatively large numbers of parameters are penalized for

⁷See, for example, Hastie, Tibshirani, and Friedman (2009).

⁸This presentation is by no means exhaustive. For example, I have omitted a promising information-theoretic approach described in Stine (2004).

⁹The corrected R^2 was introduced in Section 5.2.3.

their lack of parsimony. Beyond this intuitive rationale, however, there is no deep justification for using \tilde{R}^2 as a model selection criterion.

Mallows's C_p Statistic

One approach to subset selection in least-squares regression is based on the total (normed) mean-squared error (MSE) of estimating the expected values of the response, $E(Y_i)$, from the fitted values, \hat{Y}_i —that is, using the fitted regression to estimate the population regression surface over the observed X s:

$$\begin{aligned}\gamma_j &\equiv \frac{1}{\sigma_e^2} \sum_{i=1}^n \text{MSE}\left(\hat{Y}_i^{(j)}\right) \\ &= \frac{1}{\sigma_e^2} \sum_{i=1}^n \left\{ V\left(\hat{Y}_i^{(j)}\right) + \left[E\left(\hat{Y}_i^{(j)}\right) - E(Y_i) \right]^2 \right\}\end{aligned}\quad (22.1)$$

where the fitted values $\hat{Y}_i^{(j)}$ are based on model M_j , which contains $s_j \leq k + 1$ regressors (counting the constant, which is always included in the model) and where k is the number of regressors (less the constant) in the largest model under consideration. Using the error in estimating $E(Y)$ as a criterion for model quality is reasonable if the goal is literally to predict Y from the X s and if new observations on the X s for which predictions are required will be similar to those included in the data. An implicit assumption is that the full model, with all $k + 1$ regressors, accurately captures the dependence of Y on the X s.

The term $\left[E\left(\hat{Y}_i^{(j)}\right) - E(Y_i) \right]^2$ in Equation 22.1 represents the squared bias of $\hat{Y}_i^{(j)}$ as an estimator of the population regression surface $E(Y_i)$. When collinear regressors are deleted from the model, for example, the variance of the fitted value, $V\left(\hat{Y}_i^{(j)}\right)$, will usually decrease, but—depending on the configuration of data points and the true β s for the deleted regressors—bias may be introduced into the fitted values. Because the prediction MSE is the sum of variance and squared bias, the essential question is whether the decrease in variance offsets any increase in bias.

Mallows's C_p statistic (Mallows, 1973) estimates γ_j as

$$\begin{aligned}C_{p_j} &\equiv \frac{\sum E_i^{(j)2}}{S_E^2} + 2s_j - n \\ &= (k + 1 - s_j)(F_j - 1) + s_j\end{aligned}$$

where the residuals $E_i^{(j)}$ are from model M_j ; the error variance estimate S_E^2 is based on the *full* model fit to the data, containing all $k + 1$ regressors; and F_j is the incremental F -statistic for testing the hypothesis that the regressors omitted from model M_j have population coefficients of 0.¹⁰ If this hypothesis is true, then $E(F_j) \approx 1$, and thus $E(C_{p_j}) \approx s_j$. A good model, therefore, has C_{p_j} close to or below s_j . As well, minimizing C_p for models of a given size minimizes the sum of squared residuals and thus maximizes R^2 . For the full model, C_p necessarily equals $k + 1$.

¹⁰See Exercise 22.2.

Cross-Validation and Generalized Cross-Validation

We previously encountered cross-validation in Chapter 18 on nonparametric regression as a method for selecting the smoothing parameter in a local-polynomial or smoothing-spline regression model. Cross-validation can be applied more generally to model selection.

As before, suppose that model M_j is one of m models under consideration. In *leave-one-out cross-validation*, we fit the model n times, omitting the i th observation at step i and using the resulting fitted model to obtain a predicted value for the omitted observation, $\hat{Y}_{-i}^{(j)}$. The *cross-validation criterion* estimates the mean-squared prediction error for model M_j as

$$\text{CV}_j \equiv \frac{\sum_{i=1}^n (\hat{Y}_{-i}^{(j)} - Y_i)^2}{n} \quad (22.2)$$

We prefer the model with the smallest value of CV_j .¹¹

In linear least-squares regression, there are efficient procedures for computing the leave-one-out fitted values $\hat{Y}_{-i}^{(j)}$ that do not require literally refitting the model.¹² In other applications, however, leave-one-out cross-validation can be computationally expensive. An alternative is to divide the data into a relatively small number of subsets (e.g., 10) of roughly equal size and to fit the model omitting each subset in turn, obtaining fitted values for all observations in the omitted subset. With p subsets, this method is termed *p -fold cross-validation*. The cross-validation criterion is calculated as in Equation 22.2, using the fitted values from the p omitted subsets. Still another possibility is to approximate CV by the *generalized cross-validation criterion*

$$\text{GCV}_j \equiv \frac{n \times \text{RSS}_j}{df_{\text{res}_j}^2}$$

where RSS_j is the residual sum of squares and $df_{\text{res}_j} = n - s_j$ are the residual degrees of freedom for model M_j —an approach similar to that taken in the adjusted R^2 .¹³

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)

The *Akaike information criterion (AIC)* and the *Bayesian information criterion (BIC)*, also called *Schwarz's Bayesian criterion* (Schwarz, 1978), are currently the most commonly used model selection criteria beyond classical hypothesis tests. Both are members of a more general family of *penalized* model-fit statistics (let us call them “*IC”), applicable to regression models fit by maximum likelihood, that take the form

$${}^*\text{IC}_j = -2 \log_e L(\hat{\theta}_j) + cs_j$$

where $L(\hat{\theta}_j)$ is the maximized likelihood under model M_j ; θ_j is the vector of parameters of the model,¹⁴ including, for example, regression coefficients and an error variance or dispersion

¹¹The numerator of $\text{CV}(j)$, that is, $\sum_{i=1}^n (\hat{Y}_{-i}^{(j)} - Y_i)^2$, is called the *prediction sum of squares* (or PRESS).

¹²Recall the discussion of deletion diagnostics in Chapter 11.

¹³But see Exercise 22.3.

¹⁴If you are not familiar with vector notation, simply think of θ_j as a list of the parameters in the model; for example, in a linear regression model with normal errors, θ_j contains the regression coefficients, $\alpha^{(j)}, \beta_1^{(j)}, \dots, \beta_k^{(j)}$ and the error variance $\sigma_\epsilon^{(j)2}$.

parameter [and $\hat{\theta}_j$ is the vector of maximum-likelihood estimates (MLEs) of the parameters]; s_j is the number of parameters in θ_j ; and c is a constant that differs from one model selection criterion to another. The first term, $-2 \log_e L(\hat{\theta}_j)$, is the residual deviance under the model (or differs from the residual deviance by a constant); for a linear model with normal errors, it is simply the residual sum of squares. The *magnitude* of *IC is not generally interpretable, but *differences* between values for different models are of interest, and the model with the smallest *IC is the one that receives most support from the data.

The AIC and BIC are defined as follows:

$$\begin{aligned} \text{AIC}_j &\equiv -2 \log_e L(\hat{\theta}_j) + 2s_j \\ \text{BIC}_j &\equiv -2 \log_e L(\hat{\theta}_j) + s_j \log_e n \end{aligned}$$

For example, in a linear model with normal errors, the MLE of the regression coefficients is the least-squares estimator, and the MLE of the error variance is $\hat{\sigma}_\varepsilon^{(j)2} = (\sum E_i^{(j)2})/n$, where the $E_i^{(j)}$ are the least-squares residuals for model M_j ;¹⁵ then,

$$\begin{aligned} \text{AIC}_j &= n \log_e \hat{\sigma}_\varepsilon^{(j)2} + 2s_j \\ \text{BIC}_j &= n \log_e \hat{\sigma}_\varepsilon^{(j)2} + s_j \log_e n \end{aligned}$$

The lack-of-parsimony penalty for the BIC grows with the sample size, while that for the AIC does not. The penalty for the BIC is also larger than that for the AIC (when $n \geq 8$), and the BIC therefore tends to nominate models with fewer parameters. Although the AIC and BIC are often justified by vague appeals to parsimony, both statistics are based on deeper statistical considerations, to which I now turn.¹⁶

Model selection criteria, some applicable to regression models fit by least squares and others more general:

- The squared multiple correlation adjusted for degrees of freedom,

$$\tilde{R}^2 = 1 - \frac{n-1}{n-s} \times \frac{\text{RSS}}{\text{TSS}}$$

where n is the number of observations, s is the number of regression coefficients in the model, RSS is the residual sum of squares under the model, and TSS is the total sum of squares.

- Mallows's C_p statistic,

$$C_p = (k+1-s)(F-1) + s$$

where k is the number of predictors in the full model fit to the data and F is the incremental F -statistic for the hypothesis that the $k+1-s$ predictors excluded from the model are 0. A good model has C_p close to or below s .

¹⁵See Section 9.3.3.

¹⁶The exposition of the AIC is adapted from Burnham and Anderson (2004) and of the BIC from Raftery (1995).

- The cross-validation criterion,

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

where \hat{Y}_{-i} is the fitted value for observation i obtained when the model is fit with observation i omitted.

- The generalized cross-validation criterion,

$$\text{GCV} = \frac{n \times \text{RSS}}{df_{\text{res}}^2}$$

where df_{res} is the residual degrees of freedom under the model.

- The Akaike information criterion (AIC),

$$\text{AIC} = -2 \log_e L(\hat{\theta}) + 2s$$

where $\log_e L(\hat{\theta})$ is the maximized log-likelihood under the model (and θ is the parameter vector for the model).

- The Bayesian information criterion (BIC),

$$\text{BIC} = -2 \log_e L(\hat{\theta}) + s \log_e n$$

For both the AIC and the BIC, the model with the smallest value is the one most supported by the data.

A Closer Look at the AIC* Let $p(\mathbf{y})$ represent the “true” probability distribution or density function for the response vector \mathbf{y} in a regression model. The response \mathbf{y} can be quite general—certainly including all the models fit by the method of maximum likelihood in this book.¹⁷ The “true model” generating the data need not be among the models that we are comparing, and, indeed, we do not have to commit ourselves to the existence of a true model: The probability distribution of the data could be generated by a complex process that cannot be captured precisely by a statistical model.¹⁸

Imagine, as before, that we have a set of m statistical models under consideration, each with parameters to be estimated from the data, θ_j for model M_j , and implying the probability distribution $p_j(\mathbf{y}|\theta_j)$ for the data, which can be thought of as an approximation to the true distribution $p(\mathbf{y})$ of \mathbf{y} .¹⁹ The “best” model is the one that provides the most accurate approximation.

Kullback-Leibler information is a measure of the “distance” between two distributions, representing the information “lost” when the second distribution is used to approximate the first. The AIC applies Kullback-Leibler information to the difference between $p(\mathbf{y})$ and each $p_j(\mathbf{y}|\theta_j)$:

¹⁷The range of application of the AIC (and the BIC) is wider still—to virtually any class of statistical models fit by maximum likelihood.

¹⁸This notion is consistent with the view of statistical models presented in Chapter 1.

¹⁹ $p_j(\cdot)$ is subscripted here by the index of the model because we wish to consider the distribution of the data under different models.

$$\begin{aligned}
\mathcal{I}(p, p_j) &\equiv \int_{\text{all } \mathbf{y}} p(\mathbf{y}) \log_e \frac{p(\mathbf{y})}{p_j(\mathbf{y}|\boldsymbol{\theta}_j)} d\mathbf{y} \\
&= \int_{\text{all } \mathbf{y}} p(\mathbf{y}) \log_e p(\mathbf{y}) d\mathbf{y} - \int_{\text{all } \mathbf{y}} p(\mathbf{y}) \log_e p_j(\mathbf{y}|\boldsymbol{\theta}_j) d\mathbf{y} \\
&= E_p[\log_e p(\mathbf{y})] - E_p[\log_e p_j(\mathbf{y}|\boldsymbol{\theta}_j)] \\
&= \phi - E_p[\log_e p_j(\mathbf{y}|\boldsymbol{\theta}_j)]
\end{aligned} \tag{22.3}$$

The object, then, is to find the model M_j that minimizes the information loss. The value $\phi \equiv E_p[\log_e p(\mathbf{y})]$, in the last line of Equation 22.3, is a constant that does not depend on the model and is therefore irrelevant to model comparisons; the expectation in the term $E_p[\log_e p_j(\mathbf{y}|\boldsymbol{\theta}_j)]$ is with respect to the *true* probability distribution $p(\mathbf{y})$.

The AIC focuses on the quantity

$$E_y E_{y^*} \left\{ \log_e p_j \left[\mathbf{y}^* | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right] \right\}$$

Here, \mathbf{y}^* is a notional *second*, independently selected sample of values of the response variable (though, in an application, we have only the sample \mathbf{y}), $\hat{\boldsymbol{\theta}}_j(\mathbf{y})$ is the maximum-likelihood estimator of $\boldsymbol{\theta}_j$ based on the *original* sample \mathbf{y} , and the expectation is taken with respect to both samples. The quantity $E_{y^*} \left\{ \log_e p_j \left[\mathbf{y}^* | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right] \right\}$ is similar to $E_p[\log_e p_j(\mathbf{y}|\boldsymbol{\theta}_j)]$, substituting the MLE $\hat{\boldsymbol{\theta}}_j(\mathbf{y})$ for $\boldsymbol{\theta}_j$, and $\log_e p_j \left[\mathbf{y}^* | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right]$ is the new-sample log-likelihood under the model M_j evaluated at the MLE for the original sample. The maximized log-likelihood $\log_e L(\hat{\boldsymbol{\theta}}_j|\mathbf{y}) = \log_e p_j \left[\mathbf{y} | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right]$ is an upwardly biased estimate of $E_y E_{y^*} \left\{ \log_e p_j \left[\mathbf{y}^* | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right] \right\}$, with asymptotic bias approximately equal to the number of parameters s_j in $\boldsymbol{\theta}_j$. This is an intuitively reasonable result because we expect the *new-sample* (i.e., predicted) log-likelihood $\log_e p_j \left[\mathbf{y}^* | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right]$ to be smaller than the maximized log-likelihood $\log_e p_j \left[\mathbf{y} | \hat{\boldsymbol{\theta}}_j(\mathbf{y}) \right]$ for the original sample [for which $\hat{\boldsymbol{\theta}}_j(\mathbf{y})$ is the optimal value].

For a large sample and a distribution $p_j(\cdot)$ that is close to $p(\cdot)$, therefore,

$$\widehat{E}_{\hat{\boldsymbol{\theta}}_j} [\mathcal{I}(p, \hat{p}_j)] \approx \phi - \log_e L(\hat{\boldsymbol{\theta}}_j|\mathbf{y}) + s_j$$

where $\widehat{E}_{\hat{\boldsymbol{\theta}}_j} [\mathcal{I}(p, \hat{p}_j)]$ is the estimated expected Kullback-Leibler information loss, and \hat{p}_j represents $p_j(\cdot | \boldsymbol{\theta}_j)$ evaluated at $\boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_j$. The constant ϕ is not estimable but, as noted, it does not figure in model comparisons. The AIC, therefore, which is used for model comparison, ignores ϕ and is defined as

$$\text{AIC}_j \equiv -2 \log_e L(\hat{\boldsymbol{\theta}}_j|\mathbf{y}) + 2s_j$$

The factor 2 is inessential but puts the AIC on the same scale as the deviance.

An improvement on the AIC, called the *bias-corrected AIC*, or AIC_c , reduces small-sample bias:

$$\text{AIC}_{c_j} \equiv -2 \log_e L(\hat{\boldsymbol{\theta}}_j|\mathbf{y}) + 2s_j + \frac{2s_j(s_j + 1)}{n - s_j - 1} \tag{22.4}$$

The correction (i.e., the last term in Equation 22.4) gets smaller as the ratio of sample size to number of parameters grows and is negligible when n/s_j is large (say more than about 40). As Burnham and Anderson (2004) suggest, however, one could simply use AIC_c in all applications.

The AIC is based on the Kullback-Leibler information comparing the true distribution of the data $p(\mathbf{y})$ to the distribution of the data $p_j(\mathbf{y}|\boldsymbol{\theta}_j)$ under a particular model M_j .

A Closer Look at the BIC* The BIC has its origin in Bayesian hypothesis testing, which compares the relative weight of evidence for each of two competing hypotheses. I will broadly sketch the rationale for the BIC here.²⁰ Suppose, as before, that we are considering a set of m models for the response variable \mathbf{y} and that model M_j has parameter vector $\boldsymbol{\theta}_j$ with s_j elements. The probability or probability density for \mathbf{y} under model M_j given the values of the parameters is $p_j(\mathbf{y}|\boldsymbol{\theta}_j)$. Let $p_j(\boldsymbol{\theta}_j)$ represent the prior distribution for $\boldsymbol{\theta}_j$. Then, the marginal distribution of \mathbf{y} under the model M_j is

$$p_j(\mathbf{y}) = \int_{\text{all } \boldsymbol{\theta}_j} p_j(\mathbf{y}|\boldsymbol{\theta}_j)p_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j$$

and the posterior distribution of $\boldsymbol{\theta}_j$ is

$$p_j(\boldsymbol{\theta}_j|\mathbf{y}) = \frac{p_j(\mathbf{y}|\boldsymbol{\theta}_j)p_j(\boldsymbol{\theta}_j)}{p_j(\mathbf{y})}$$

Let us focus initially on two of the models, M_1 and M_2 , and assume that one of these is the “correct” model for the data.²¹ The posterior probability that M_1 is the correct model is

$$p(M_1|\mathbf{y}) = \frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y}|M_1)p(M_1) + p(\mathbf{y}|M_2)p(M_2)}$$

Here, $p(M_j)$ is the prior probability assigned to model M_j , and $p(\mathbf{y}|M_j)$ is the *marginal probability of the data* under model M_j (also called the *predictive probability of the data*):

$$p(\mathbf{y}|M_j) = \int_{\text{all } \boldsymbol{\theta}_j} p_j(\mathbf{y}|\boldsymbol{\theta}_j)p_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j \quad (22.5)$$

A direct formula for $p(M_2|\mathbf{y})$ is similar, but because there are just two models under consideration, it is also the case that $p(M_2|\mathbf{y}) = 1 - p(M_1|\mathbf{y})$.

After observing the data, the relative support for model M_2 versus M_1 is given by the *posterior odds*

$$\frac{p(M_2|\mathbf{y})}{p(M_1|\mathbf{y})} = \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)} \times \frac{p(M_2)}{p(M_1)}$$

The posterior odds are, therefore, the product of two terms: the ratio of marginal probabilities of the data under the competing models, $p(\mathbf{y}|M_2)/p(\mathbf{y}|M_1)$, called the *Bayes factor* for model

²⁰This section assumes an acquaintance with the general principles of Bayesian statistical inference, which are described in online Appendix D on probability and estimation.

²¹It is possible to develop the BIC by making an argument based on accuracy of out-of-sample prediction without assuming that one of the models is the “true” model for the data. See, for example, Kass and Raftery (1995).

M_2 versus M_1 , and $p(M_2)/p(M_1)$, the ratio of prior probabilities for the models. It seems fair to set equal prior probabilities, $p(M_1) = p(M_2)$,²² in which case the posterior odds are simply the Bayes factor.

An important point concerning the posterior odds is that there are *two* prior distributions to consider: (1) the prior probabilities $p(M_j)$ for the *models* under consideration and (2) the prior distribution $p_j(\boldsymbol{\theta}_j)$ for the *parameters* in each model, on which the marginal probability of the data under the model depends (Equation 22.5). As mentioned, it seems evenhanded to accord the various models equal prior probability, at least in the absence of a convincing argument to the contrary, but the priors $p_j(\boldsymbol{\theta}_j)$ on the parameters are another question entirely. In *Bayesian estimation*, the importance of the prior distribution on the parameters fades as the sample size grows; thus, unless the sample size is small and there is a sound basis for specific prior beliefs, the argument for so-called noninformative or vague priors can be compelling. This is not the case, however, in *Bayesian hypothesis testing*, where the prior distribution on the parameters of each model affects the marginal probability of the data, and through it the Bayes factor, *even in large samples*.

The BIC is an approximation to the Bayes factor, employing a particular choice of prior distribution on the parameters of each model (see below).²³ It is convenient to introduce the function

$$f(\boldsymbol{\theta}_j) \equiv \log_e [p_j(\mathbf{y}|\boldsymbol{\theta}_j)p_j(\boldsymbol{\theta}_j)] \quad (22.6)$$

which is the log of the integrand in Equation 22.5. Let $\tilde{\boldsymbol{\theta}}_j$ represent the value of the parameter vector that maximizes $f(\boldsymbol{\theta}_j)$ for the observed data \mathbf{y} . A second-order Taylor-series expansion of $f(\boldsymbol{\theta}_j)$ around $\tilde{\boldsymbol{\theta}}_j$ is²⁴

$$\begin{aligned} f(\boldsymbol{\theta}_j) &\approx f(\tilde{\boldsymbol{\theta}}_j) + (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)' \frac{\partial f(\tilde{\boldsymbol{\theta}}_j)}{\partial \boldsymbol{\theta}_j} + \frac{1}{2}(\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)' \frac{\partial^2 f(\tilde{\boldsymbol{\theta}}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}'_j} (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j) \\ &\approx f(\tilde{\boldsymbol{\theta}}_j) + \frac{1}{2}(\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)' \frac{\partial^2 f(\tilde{\boldsymbol{\theta}}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}'_j} (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j) \end{aligned}$$

The second term in the expansion vanishes because the first-order partial derivatives $\partial f(\tilde{\boldsymbol{\theta}}_j)/\partial \tilde{\boldsymbol{\theta}}_j$ are $\mathbf{0}$ at the maximum of $f(\boldsymbol{\theta}_j)$. Given sufficient data, we expect $\tilde{\boldsymbol{\theta}}_j$ to be close to $\boldsymbol{\theta}_j$ and expect that the likelihood $p_j(\mathbf{y}|\boldsymbol{\theta}_j)$ will decline rapidly as $\boldsymbol{\theta}_j$ departs from $\tilde{\boldsymbol{\theta}}_j$. Under these circumstances, the marginal probability of the data (from Equation 22.5) is approximately

$$p(\mathbf{y}|M_j) \approx \exp[f(\tilde{\boldsymbol{\theta}}_j)] \int \exp\left[\frac{1}{2}(\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)' \frac{\partial^2 f(\tilde{\boldsymbol{\theta}}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}'_j} (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)\right] d\boldsymbol{\theta}_j \quad (22.7)$$

A clever trick facilitates the evaluation of the integral in Equation (22.7). With the exception of the absence of the multiplicative factor $(2\pi)^{-s_j/2} (\det \tilde{\Sigma})^{-1/2}$, where

$$\tilde{\Sigma} \equiv - \left[\frac{\partial^2 f(\tilde{\boldsymbol{\theta}}_j)}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}'_j} \right]^{-1}$$

²²With only two models under consideration, we therefore have $p(M_1) = p(M_2) = 1/2$.

²³The development here is quite dense, even for starred material; the reader may wish to skip to the key result given in Equation 22.11 (on page 680).

²⁴If the sample size is sufficiently large, then higher-order terms in the Taylor expansion should be negligible.

the integrand in this equation looks like the formula of the multivariate-normal density, with θ_j playing the role of the vector random variable, $\tilde{\theta}_j$ the role of the mean vector, and $\tilde{\Sigma}$ the role of the covariance matrix;²⁵ note that this is simply an *analogy* that will help us evaluate the integral in Equation 22.7. Because the multivariate-normal density integrates to 1, the integral evaluates to the inverse of the missing constant, $(2\pi)^{s_j/2} (\det \tilde{\Sigma})^{1/2}$. Consequently,

$$p(\mathbf{y}|M_j) \approx \exp[f(\tilde{\theta}_j)] (2\pi)^{s_j/2} (\det \tilde{\Sigma})^{1/2}$$

and (using Equation 22.6)

$$\begin{aligned} \log_e p(\mathbf{y}|M_j) &\approx f(\tilde{\theta}_j) + \frac{s_j}{2} \log_e 2\pi + \frac{1}{2} \log_e (\det \tilde{\Sigma}) \\ &\approx \log_e p_j(\mathbf{y}|\tilde{\theta}_j) + \log_e p_j(\tilde{\theta}_j) + \frac{s_j}{2} \log_e 2\pi + \frac{1}{2} \log_e (\det \tilde{\Sigma}) \end{aligned} \quad (22.8)$$

If the sample size is large, then we would expect the posterior mode $\hat{\theta}_j$ to be close to the maximum-likelihood estimator $\tilde{\theta}_j$ of θ_j . Substituting $\hat{\theta}_j$ for $\tilde{\theta}_j$,

$$\begin{aligned} \tilde{\Sigma}^{-1} \approx \hat{\Sigma}^{-1} &= -\frac{\partial f(\hat{\theta}_j)}{\partial \hat{\theta}_j \partial \hat{\theta}'_j} \\ &= -n \times E_{\mathbf{y}} \left[\frac{\partial^2 \log_e p(Y|\theta_j)}{\partial \theta_j \partial \theta'_j} \middle| \theta_j = \hat{\theta}_j \right] \\ &= n \times \mathcal{I}(\hat{\theta}_j) \end{aligned}$$

The matrix

$$\mathcal{I}(\hat{\theta}_j) \equiv -E_{\mathbf{y}} \left[\frac{\partial^2 \log_e p(Y|\theta_j)}{\partial \theta_j \partial \theta'_j} \middle| \theta_j = \hat{\theta}_j \right]$$

is the expected Fisher information associated with a single observation Y on the response variable. Noting that in a large sample $\det \hat{\Sigma} \approx -[n^{s_j} \det \mathcal{I}(\hat{\theta}_j)]^{-1}$ and substituting this approximation into Equation 22.8 gives

$$\log_e p(\mathbf{y}|M_j) \approx \log_e p_j(\mathbf{y}|\hat{\theta}_j) + \log_e p_j(\hat{\theta}_j) + \frac{s_j}{2} \log_e 2\pi - \frac{s_j}{2} \log_e n - \frac{1}{2} \log_e [\det \mathcal{I}(\hat{\theta}_j)] \quad (22.9)$$

The BIC uses the *unit-information prior distribution* $\theta_j \sim N_{s_j}[\hat{\theta}_j, \mathcal{I}(\hat{\theta}_j)]$ —quite a diffuse prior centered on the MLE of θ_j ; under this prior,²⁶

$$\log_e p_j(\hat{\theta}_j) = -\frac{s_j}{2} \log_e 2\pi + \frac{1}{2} \log_e [\det \mathcal{I}(\hat{\theta}_j)]$$

Substituting this result into Equation 22.9 produces

$$\log_e p(\mathbf{y}|M_j) \approx \log_e p_j(\mathbf{y}|\hat{\theta}_j) - \frac{s_j}{2} \log_e n \quad (22.10)$$

On the basis of the preceding work, the log-Bayes factor for model M_2 relative to model M_1 can then be approximated as

²⁵See online Appendix D on probability and estimation for a discussion of the multivariate-normal distribution.

²⁶It is also possible to construe the BIC as an approximation to the Bayes factor under an *unspecified* prior, but then the quality of the approximation can be much worse. See, for example, Raftery (1995).

Table 22.1 Relative Support for Model M_2 Versus M_1 as a Function of Differences in BIC

Difference in BIC	Bayes Factor	$p(M_2 \mathbf{y})$	Evidence for M_2
0–2	1–3	.50–.75	“Weak”
2–6	3–20	.75–.95	“Positive”
6–10	20–150	.95–.99	“Strong”
>10	>150	>.99	“Conclusive”

SOURCE: Adapted from Raftery (1995, Table 6).

$$\log_e \frac{p(\mathbf{y}|M_2)}{p(\mathbf{y}|M_1)} \approx \log_e p_2(\mathbf{y}|\hat{\boldsymbol{\theta}}_2) - \log_e p_1(\mathbf{y}|\hat{\boldsymbol{\theta}}_1) - \frac{1}{2}(s_2 - s_1) \log_e n \quad (22.11)$$

Recall that the choice of the unit-information prior to obtain this approximation is not necessarily benign: Different priors produce different Bayes factors.²⁷ Moreover, several approximations were made in arriving at this result, and for some classes of models, more accurate approximations are available.²⁸

The BIC for model M_j is defined as

$$\text{BIC}_j \equiv -2 \log_e p_j(\mathbf{y}|\hat{\boldsymbol{\theta}}_j) + s_j \log_e n$$

Given this definition, twice the log-Bayes factor for any pair of models M_j and $M_{j'}$ is approximated by the difference in their BICs:

$$2 \times \log_e \frac{p(\mathbf{y}|M_j)}{p(\mathbf{y}|M_{j'})} \approx \text{BIC}_{j'} - \text{BIC}_j \quad (22.12)$$

Under the unit-information prior, the difference in BIC therefore expresses the relative support in the data for model M_j versus $M_{j'}$, and the model with the smallest BIC is the one that receives most support from the data. A BIC difference of 0, for example, is equivalent to a Bayes factor of $\exp(\frac{1}{2} \times 0) = 1$ —that is, equal support in the data for the two models; if these are the only models under consideration (and if the prior probabilities for the two models are equal), therefore, we would have posterior probabilities $p(M_2|\mathbf{y}) = p(M_1|\mathbf{y}) = \frac{1}{2}$. Similarly, a BIC of 2 is equivalent to a Bayes factor of $\exp(\frac{1}{2} \times 2) \approx 2.718$ in favor of model M_2 , or $p(M_2|\mathbf{y}) \approx .73$ and $p(M_1|\mathbf{y}) \approx .27$ —that is, relatively weak evidence in favor of M_2 . Table 22.1, adapted from Raftery (1995), extends these interpretations to various differences in BIC.²⁹

Like classical testing, then, the BIC is based on the notion of a statistical hypothesis test. What, then, accounts for the difference between the two approaches, and, in particular, why does the BIC tend to prefer more parsimonious models? Part of the difference between the BIC and classical testing lies in the role of prior distributions for the parameters of the models in the formulation of the BIC, but even more fundamentally, the two kinds of tests treat evidence

²⁷Burnham and Anderson (2004) show, for example, that the AIC can be derived as an approximation to the log of the Bayes factor using a prior different from the unit-information prior. Consequently, the choice of AIC or BIC as a model selection criterion cannot simply be construed as a contest between “frequentist” and Bayesian approaches to the problem of model selection.

²⁸See, for example, the results given in Raftery (1996) for generalized linear models and the general discussion in Kass and Raftery (1995).

²⁹See Exercise 22.4.

differently. Suppose, for example, that we test model M_2 versus M_1 , where M_2 is nested within M_1 (as is the case when M_2 is derived from M_1 by setting certain parameters to 0). In this instance, the classical test is of the null hypothesis that the parameter restrictions on M_1 producing M_2 are correct, against the alternative that they are wrong, and the two models play an *asymmetric* role in the formulation of the test: The p -value for the null hypothesis is the probability of obtaining data *as extreme as or more extreme than* the observed data assuming the truth of M_2 . In the Bayesian test (to which the BIC is an approximation), the two models play a *symmetric* role, with the Bayes factor weighing the relative strength of evidence for the models *in the observed data*; data more extreme than those observed do not figure in the test, lending greater support to the null hypothesis than it has in the classical test.

The BIC has its basis in Bayesian hypothesis testing, comparing the degree of support in the data for two models. The BIC is an approximation to twice the log of the Bayes factor comparing a particular model to the saturated model, where the Bayes factor is the ratio of the marginal probability of the data under the two models. When the prior probabilities for the two models are the same, the posterior odds for the models are equal to the Bayes factor. Differences in BIC approximate twice the log of the Bayes factor comparing two models to each other. The BIC approximation to the Bayes factor is accurate for a particular choice of prior distribution over the parameters of the models, called the unit-information prior, but may not be accurate for other priors. Differences in BIC of about 6 or more represent strong evidence in favor of the model with the smaller BIC.

22.1.2 An Illustration: Baseball Salaries

To illustrate model selection, I will use data on major-league baseball players' salaries from the 1987 season, excluding pitchers and restricting attention to players who were active during the 1986 season.³⁰ In addition to the player's name and the team for which he played at the beginning of the 1987 season, the data source also included the player's annual salary (in thousands of dollars) at the start of the 1987 season and information on number of times at bat (AB), number of hits (H), number of home runs (HR), number of runs scored (R), number of runs batted in (RBI), and number of walks (bases on balls, BB), both for the 1986 season and during the player's career; the player's number of put-outs (PO), assists (A), and errors (E) during the 1986 season; the player's position (or positions) in the field during the 1986 season; and the player's number of years in the major leagues.³¹

³⁰The data set originated in a 1988 poster session sponsored by the Statistical Graphics Section of the American Statistical Association and was used, for example, by Friendly (2002) in a paper on graphical display of correlation matrices. The version used here has a number of errors corrected.

My apologies to readers who are unfamiliar with baseball: Even a superficial explanation of that subtle sport would require more space than the rest of the chapter. I expect that the general sense of the example will be clear even if nuances are missed.

³¹The abbreviations (e.g., AB for at-bats) are standard. There was, in addition, information on the player's team and the division and league (i.e., National or American) in which he played. I decided not to use this information in predicting salary, because I thought that it would weaken interest in the example: One could argue that playing for a high-paying team is a reflection of a player's earning potential.

From these variables, I derived several additional potential predictors of salary: the player's 1986 and career batting average (AVG—i.e., number of hits divided by number of at-bats), 1986 and career on-base percentage ($OBP = 100 \times [\text{hits} + \text{walks}] / [\text{at-bats} + \text{walks}]$), and the numbers of at-bats, hits, home runs, runs scored, and runs batted in recorded per year over the player's career (e.g., number of career home runs divided by number of years in the majors). Rather than create 24 dummy variables for the 25 positions and combinations of positions that appear in the data set, I created four 0/1 dummy variables, coded 1 for players who consistently played second base or shortstop (i.e., middle infielders, MI), catcher (C), center field (CF), or designated hitter (DH). Middle infield, catcher, and center field are generally considered high-skill positions; designated hitters (a role available only in the American League) bat but do not play the field. After 3 years in the major leagues, players are eligible for salary arbitration, and after 6 years they are eligible for free agency (i.e., can negotiate a contract with any team). I consequently created two 0/1 dummy variables, one coded 1 for players with between 3 and 5 years of major-league experience and the other coded 1 for players with 6 or more years in the majors.³²

Preliminary examination of the data suggested log-transforming salary (the response variable), number of years in the majors, and career at-bats. I also decided to drop one player (Pete Rose) from the data set because of his high leverage in the regressions.³³ These modeling decisions could be made a formal part of the model selection process, but to do so would further complicate an already complicated example.

The data set to be analyzed includes 262 players and 33 variables. A linear least-squares regression of log salary on the 32 predictors accounts for most of the variation in the response variable, $R^2 = .861$, but as one might expect, the regression proves difficult to interpret. There are several “statistically significant” regression coefficients (for BB, MI, C, career AB, career H, career R, career BB, and eligibility for free agency), but the degree of collinearity is very high, with variance-inflation factors topping out at more than 500 (for career H) and a condition number of 133.³⁴

Figure 22.1 shows the predictors in the “best” model of each size, selected according to the BIC.³⁵ Table 22.2 includes all models, regardless of size, within 2 of the minimum BIC, displaying the signs of the coefficients for the predictors in each model, the model R^2 , and the difference in BIC compared to the “best” model.³⁶ An entry of 0 indicates that the corresponding predictor does not appear in the model; only predictors appearing in at least one of the 13 models are shown. The coefficient signs, it turns out, are consistent across models in the table, but not all signs make substantive sense: Why, for example, should number of career hits, which is present in all 13 models, have a negative coefficient, controlling for the other predictors in the

³²A disclaimer: This is not a serious investigation of baseball salaries. Such an investigation would take into account additional information about the players' situations, such as whether they were free agents prior to the 1987 season. Moreover, if prediction is the goal, salary in the previous season is obviously relevant. Finally, it was later established that during this period, baseball owners colluded illegally to limit the salaries of free agents.

³³Fans of baseball will find this decision ironic: Pete Rose, baseball's all-time hits leader, was banned for life from the sport because of his gambling activities.

³⁴Variance-inflation factors and the condition number are described in Section 13.1. Two of the “significant” predictors—career H and career BB—have unexpectedly negative coefficients.

³⁵See Exercise 22.5 for the application of other model selection criteria to the baseball data.

³⁶Some pairs of models (e.g., 1 and 5) with the same number of predictors and the same R^2 have slightly different BIC values. The apparent discrepancy is due to rounding of R^2 to three decimal places.

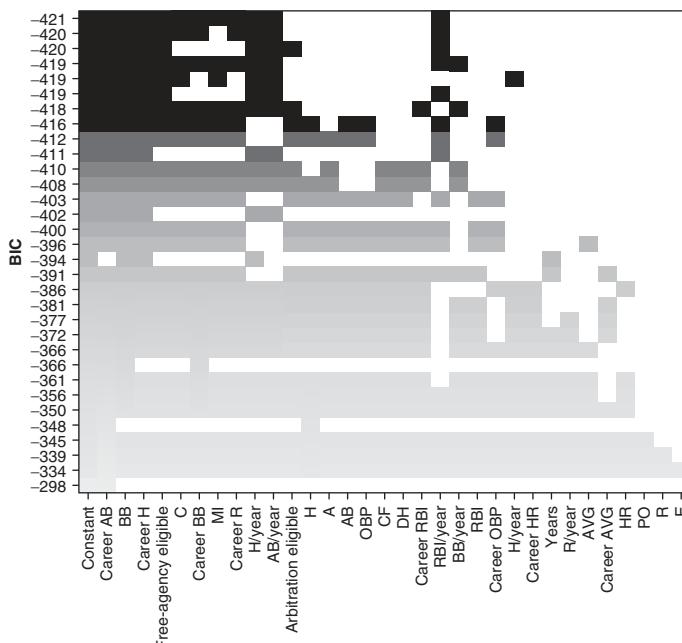


Figure 22.1 BIC for the “best” model of each size fit to the baseball salary data. The models are sorted by BIC, with the lowest BIC at the top; the variables are sorted by the number of models in which they appear. The squares show which variables are included in each model, with darker squares indicating lower values of BIC (i.e., “better” models).

models? It is necessary to remind ourselves that the goal here is to select models that produce accurate predictions.

This example is elaborated in the following section.

22.1.3 Comments on Model Selection

I have stressed the point that automatic model selection methods—stepwise and optimal-subset regression, for example—attend to the predictive adequacy of regression models and are blind to their substantive interpretability. I believe that in most instances, researchers would be better served by judiciously selecting the setting within which to conduct an investigation, thinking carefully about the social process under study and the questions to be put to the data, and focusing on a relatively small number of explanatory variables, with the level of detail growing with the sample size.

Model selection criteria such as the AIC and BIC are not limited to comparing models selected by automatic methods, however, and one of the currently popular applications of the BIC is to justify the removal of small but “statistically significant” terms in regression models fit to large samples of data. Though largely benign, I believe that this practice slightly misses

Table 22.2 The “Best” Models Fit to the Baseball Salary Data According to the BIC

Predictor	Model												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Free-agency eligible	+	+	+	+	+	+	+	+	+	+	+	+	+
H/year	+	+	+	+	+	+	+	+	+	+	+	+	+
AB/year	-	-	-	-	-	-	-	-	-	-	-	-	-
Career H	-	-	-	-	-	-	-	-	-	-	-	-	-
Career AB	+	+	+	+	+	+	+	+	+	+	+	+	+
BB	+	+	+	+	+	0	+	0	+	0	+	+	0
C	+	+	+	+	0	+	0	+	0	0	+	+	0
Career BB	-	-	-	0	-	0	-	0	-	0	-	-	0
Career R	+	+	+	0	+	0	+	0	+	0	+	+	0
RBI/year	+	+	0	+	0	0	0	0	0	0	0	+	+
Career RBI	0	0	+	0	+	0	+	0	+	0	+	0	0
Arbitration	0	0	0	+	0	+	0	0	+	0	0	+	0
BB/year	0	0	0	0	+	0	+	0	0	0	+	0	0
MI	+	0	0	0	0	0	0	0	+	0	0	0	0
HR/year	0	0	0	0	0	0	0	0	0	+	0	0	0
H	0	0	0	0	0	0	0	0	0	0	+	0	0
Number of predictors	11	10	10	8	11	8	12	12	8	9	11	11	7
R^2	.844	.841	.841	.833	.844	.833	.847	.847	.833	.837	.843	.843	.829
BIC–BIC _{min}	0.00	0.18	0.65	1.01	1.14	1.27	1.29	1.33	1.69	1.71	1.86	1.91	1.96

NOTE: All of the models are within 2 of the smallest BIC. The table shows the sign of each coefficient; 0 indicates that the predictor is not in the model. Only predictors entering at least one of the models are shown.

an essential point: Researchers should feel free to remove “statistically significant” terms from a statistical model based on the substantive judgment that these terms are too small to be of interest. It may be nice that the BIC supports this judgment, but that is by no means essential, and in very large samples, even the BIC may point toward models that are unnecessarily complex for summarizing data cogently.

Although I have drawn a clear distinction between prediction and interpretation, in some cases, the line between the two is blurred. It is common, for example, for interest to center on one or a small number of explanatory variables; other explanatory variables are regarded as statistical “controls”—causally prior variables included in the analysis to avoid spurious results.³⁷ In this setting, one might be tempted to specify a model in which the explanatory variables of primary interest are *necessarily* included but for which other explanatory variables are selected by an automatic procedure. Unfortunately, this approach is flawed: Control variables that are highly correlated with the focal explanatory variables will likely be excluded from the model, and it is precisely the exclusion of these variables that raises the specter of spuriousness. In most cases, therefore, if simply controlling for *all* the variables thought to be important turns out to be impractical, then the data are probably insufficiently informative to answer the questions posed by the researcher.

When the focus is on interpretation rather than prediction, researchers should feel free to simplify a statistical model on the basis of substantive considerations, even if that means removing small but “statistically significant” terms from the model. Penalized model selection criteria, such as the BIC, often provide an unnecessary excuse for doing so.

22.2 Model Averaging*

Model selection, described in the preceding section, implies uncertainty about the “best” statistical model for the data.³⁸ Often there are several—or even many—models that are roughly equally supported by the data, providing small justification for choosing among them. Uncertainty can arise from other sources as well, such as the selection of transformations of the response or explanatory variables in a regression or the removal of outliers. *Model averaging* seeks to acknowledge model uncertainty explicitly by combining information from competing statistical models rather than discarding all models but one. As I will argue at the end of this section, I believe that model averaging (like model selection) is most sensible when the goal of a statistical investigation is prediction.

I will describe an approach to model averaging based on the BIC.³⁹ As previously explained, under the unit-information prior, the difference in BIC for two competing models—say models

³⁷See the discussion of specification errors in Sections 6.3 and 9.7.

³⁸This section is starred not because of its difficulty but because it depends on starred material in Section 22.1.1.

³⁹Bayesian model averaging based on the BIC is described in several sources, such as Kass and Raftery (1995), Raftery (1995), and Hoeting, Madigan, Raftery, and Volinsky (1999). The exposition in this section is close to Raftery (1995). There are other approaches to model averaging. See, for example, Exercise 22.6 for model averaging based on the AIC.

M_1 and M_2 —approximates twice the log Bayes factor for the two models;⁴⁰ using the notation of Section 22.1.1,

$$\text{BIC}_2 - \text{BIC}_1 \approx 2 \times \log_e \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Consequently, if the prior probabilities for the two models are equal and if attention is restricted to these models, then the posterior probability for model M_1 is⁴¹

$$p(M_1|\mathbf{y}) \approx \frac{\exp(-\frac{1}{2}\text{BIC}_1)}{\exp(-\frac{1}{2}\text{BIC}_1) + \exp(-\frac{1}{2}\text{BIC}_2)}$$

The extension to a set of models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$ is immediate:

$$p(M_j|\mathbf{y}) \approx p_j \equiv \frac{\exp(-\frac{1}{2}\text{BIC}_j)}{\sum_{j'=1}^m \exp(-\frac{1}{2}\text{BIC}_{j'})} \quad (22.13)$$

The approximate posterior probabilities p_j can be used to determine the strength of evidence for including a particular predictor, say X_ℓ , in the model and for estimating model “outputs” such as coefficients and predicted values. The posterior probability that the coefficient β_ℓ of X_ℓ is not 0 is

$$\Pr(\beta_\ell \neq 0|\mathbf{y}) \approx \sum_{j:M_j \in \mathcal{A}_\ell} p_j$$

where \mathcal{A}_ℓ is the subset of models \mathcal{M} that include the predictor X_ℓ . Restricting attention to this subset of models, the posterior distribution of β_ℓ assuming that the coefficient is not 0 is

$$p(\beta_\ell|\mathbf{y}, \beta_\ell \neq 0) \approx \sum_{j:M_j \in \mathcal{A}_\ell} p(\beta_\ell|\mathbf{y}, M_j)p'_j$$

where

$$p'_j \equiv \frac{p_j}{\sum_{j':M_{j'} \in \mathcal{A}_\ell} p_{j'}}$$

Likewise, conditional on $\beta_\ell \neq 0$, the posterior mean and variance of β_ℓ can be approximated as

$$\begin{aligned} E(\beta_\ell|\mathbf{y}, \beta_\ell \neq 0) &\approx \tilde{\beta}_\ell \equiv \sum_{j:M_j \in \mathcal{A}_\ell} p'_j \hat{\beta}_\ell^{(j)} \\ V(\beta_\ell|\mathbf{y}, \beta_\ell \neq 0) &\approx \sum_{j:M_j \in \mathcal{A}_\ell} p'_j \left[\tilde{V}(\hat{\beta}_\ell^{(j)}) + \hat{\beta}_\ell^{(j)2} \right] - \tilde{\beta}_\ell^2 \end{aligned}$$

where $\hat{\beta}_\ell^{(j)}$ is the MLE of β_ℓ and $\tilde{V}(\hat{\beta}_\ell^{(j)})$ is the estimated sampling variance of this coefficient in model M_j .

A practical obstacle to applying these results is the possibly very large number of candidate models in \mathcal{M} . For the baseball salary regression, for example, where there are $k = 32$

⁴⁰See Equation 22.12 on page 680.

⁴¹See Exercise 22.7.

predictors, the number of models is $m = 2^{32} \approx 4.3 \times 10^9$ or about 4 billion! Most of these models, however, have posterior probabilities very close to 0. To deal with this problem, Madigan and Raftery (1994) suggest excluding from consideration (1) models with BIC more than 6 units higher than the smallest BIC (i.e., with posterior odds relative to the model most supported by the data of about 1/20 or smaller) and (2) models that have a more probable model nested within them (i.e., models for which eliminating one or more terms produces a model with a smaller BIC). Madigan and Raftery call this rule “Occam’s window.”⁴² Posterior probabilities are computed according to Equation 22.13 but excluding models falling outside the window. Evidence suggests that applying only the first part of the rule tends to produce more accurate predictions; in this case, the window of acceptable models is termed symmetric rather than strict. Efficient methods exist for locating the subset of models in Occam’s window without enumerating and fitting all possible models.

22.2.1 Application to the Baseball Salary Data

I applied Bayesian model averaging to the baseball salary regression, with the results given in Table 22.3, using 175 models falling in the symmetric Occam’s window that encompasses all models with BIC within 6 of the “best” model. The regression intercept was included in all the models. The best 13 of these models appeared in Table 22.2 (on page 684).

Many variables conventionally used to measure players’ performance (such as career batting average and career on-base percentage) have very low probabilities of inclusion in the model. Figure 22.2 shows the posterior distribution of the regression coefficients for the nine predictors that have probability of inclusion in the model greater than .5. The vertical line visible in some of the graphs shows the probability that the corresponding coefficient is 0. Two of the coefficients (for career hits and free-agency eligibility) have clearly bimodal posterior distributions.

22.2.2 Comments on Model Averaging

By combining information from many models and thereby avoiding what is typically the illusion of a single “best” predictive model, model averaging holds out the promise of more accurate predictions. Indeed, one can average *predictions* directly, not just regression coefficients.

Nevertheless, because the meaning of a partial regression coefficient depends on the *other* explanatory variables in the model [what Mosteller and Tukey (1977, especially chap. 13) termed the “stock” of explanatory variables in the regression], model-averaged regression coefficients can be difficult to interpret. This point is drawn into focus when the distribution of a coefficient across models is bi- or multimodal (as was the case for at least two of the coefficients in Figure 22.2, for example), but the point is more general. Although a proponent of

⁴²“Occam’s razor,” due to the English philosopher William of Occam (1285–1329), is an early and famous expression of the principle of parsimony. The principle appeared frequently in Occam’s writings, including in the form, “Plurality is not to be assumed without necessity” (see Moody, 1972).

Table 22.3 Probability of Inclusion in the Model and Posterior Expectation and Standard Deviation if Nonzero for the Predictors in the Baseball Salary Data

Predictor	$Pr(\beta_\ell \neq 0 \mathbf{y})$	$E(\beta_\ell \mathbf{y}, \beta_\ell \neq 0)$	$SD(\beta_\ell \mathbf{y}, \beta_\ell \neq 0)$
Constant	—	-0.586	0.671
Career AB	1.000	0.831	0.129
Career H	1.000	-0.00109	0.00041
Free-agency eligible	1.000	0.403	0.201
AB/year	.985	-0.00717	0.00151
H/year	.985	0.0259	0.0052
BB	.920	0.00659	0.00260
C	.680	0.145	0.120
Career BB	.636	-0.000680	0.000648
Career R	.518	0.000816	0.000884
RBI/year	.450	0.00412	0.00488
Arbitration eligible	.416	0.113	0.155
MI	.408	0.0739	0.1093
Career RBI	.397	0.000424	0.000563
BB/year	.232	0.00267	0.00553
HR/year	.123	0.00220	0.00616
DH	.109	-0.0232	0.0773
A	.089	-0.0000531	0.0001902
R/year	.073	0.000693	0.002741
H	.067	0.000302	0.000188
CareerHR	.046	0.0000667	0.0004581
CF	.032	-0.00424	0.02770
AB	.022	-0.0000356	0.0005514
PO	.021	0.00000278	0.00002306
HR	.019	-0.000129	0.001087
R	.015	0.0000713	0.0006055
OBP	.010	-0.000629	0.007083
Career OBP	.009	0.000429	0.005555
RBI	.007	-0.0000111	0.0001894
Career AVG	.007	0.0559	0.6650
E	.004	-0.0000155	0.0003528
AVG	.000	—	—
Years	.000	—	—

model averaging might well reply that this kind of ambiguity is simply the reflection of model uncertainty, mechanically averaging regression coefficients is not a substitute for thinking about the substantive content of a regression equation.

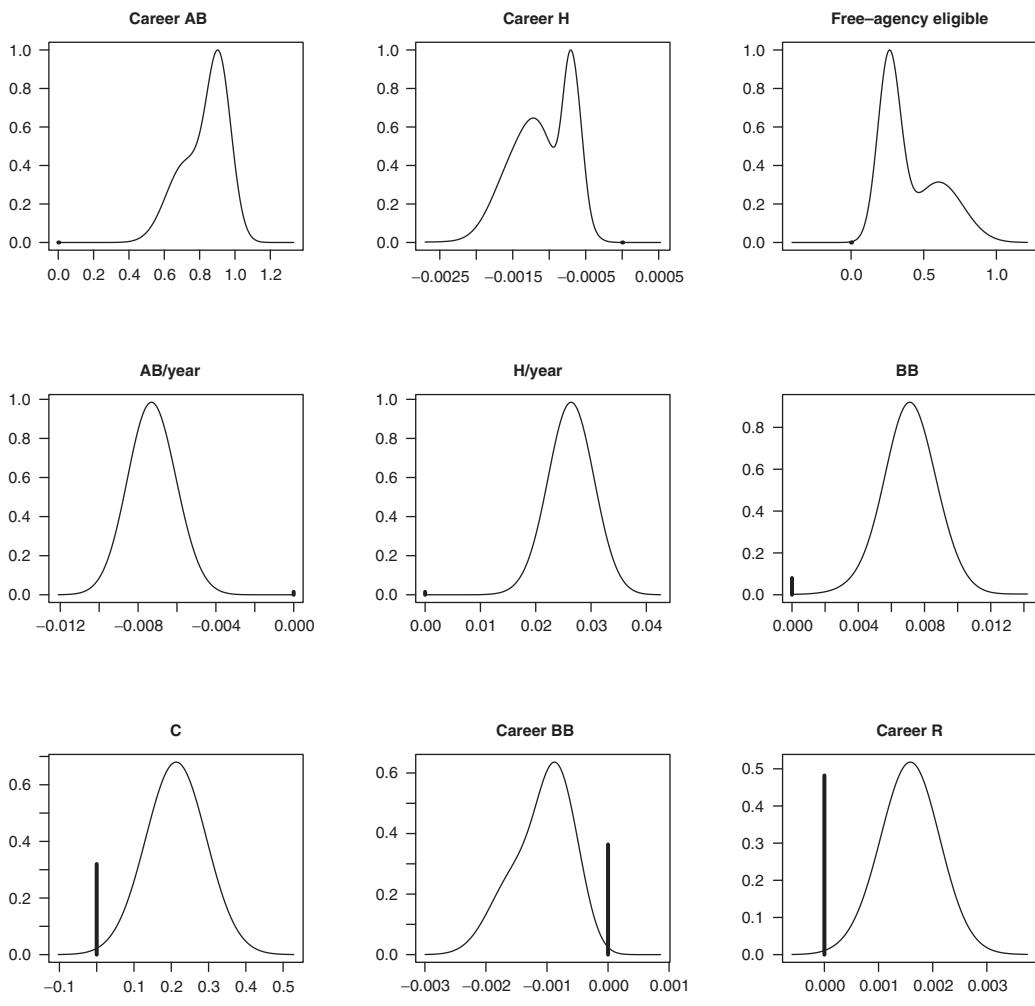


Figure 22.2 Posterior distributions for coefficients in the baseball salary regression. Only coefficients with probability of inclusion in the model exceeding .5 are shown. The vertical line visible in some panels represents the probability that the corresponding coefficient is 0; the vertical axis is scaled so that the highest point of the density curve is the probability that the coefficient is *nonzero*.

The posterior probability for each model M_j in a set of models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$ can be approximated using the BIC:

$$p(M_j | \mathbf{y}) \approx p_j = \frac{\exp(-\frac{1}{2}\text{BIC}_j)}{\sum_{j'=1}^m \exp(-\frac{1}{2}\text{BIC}_{j'})}$$

Then, for a model output such as the regression coefficient β_ℓ of the predictor X_ℓ , the posterior probability that β_ℓ is not 0 is

$$\Pr(\beta_\ell \neq 0 | \mathbf{y}) \approx \sum_{j: M_j \in \mathcal{A}_\ell} p_j$$

where \mathcal{A}_ℓ is the subset of models that include the predictor X_ℓ . Conditional on $\beta_\ell \neq 0$, the posterior mean and variance of β_ℓ are approximately

$$E(\beta_\ell | \mathbf{y}, \beta_\ell \neq 0) \approx \tilde{\beta}_\ell = \sum_{j: M_j \in \mathcal{A}_\ell} p'_j \hat{\beta}_\ell^{(j)}$$

$$V(\beta_\ell | \mathbf{y}, \beta_\ell \neq 0) \approx \sum_{j: M_j \in \mathcal{A}_\ell} p'_j \left[\hat{V}\left(\hat{\beta}_\ell^{(j)}\right) + \hat{\beta}_\ell^{(j)2} \right] - \tilde{\beta}_\ell^2$$

where $\hat{\beta}_\ell^{(j)}$ is the MLE of β_ℓ , $\hat{V}\left(\hat{\beta}_\ell^{(j)}\right)$ is the estimated sampling variance of this coefficient in model M_j , and $p'_j = p_j / \sum_{j': M_j' \in \mathcal{A}_\ell} p_{j'}$. Because the number of candidate models can be extremely large, there is an advantage to restricting attention only to models with relatively large posterior probabilities, such as those with BIC within 6 of the “best” model; these models are said to fall within “Occam’s window.” Because the meaning of a partial regression coefficient depends on the other explanatory variables in the model, however, model-averaged regression coefficients can be difficult to interpret.

22.3 Model Validation

In *model validation*, part of the data (called the “training” or “exploratory” subsample) is used to specify a statistical model, which is then evaluated using the other part of the data (the “validation” or “confirmatory” subsample). Cross-validation, already discussed, is an application of this very simple—but powerful—idea, where the roles of training and validation subsamples are interchanged or rotated.⁴³

I have stressed the importance of descriptive adequacy in statistical modeling, and—in support of this goal—I have described a variety of methods for screening data and for evaluating and, if necessary, modifying statistical models. This process of data exploration, model fitting, model criticism, and model respecification is often iterative, requiring several failed attempts before an adequate description of the data is achieved. In the process, variables may be dropped from the model, terms such as interactions may be incorporated or deleted, variables may be transformed, and unusual data may be corrected, removed, or otherwise accommodated.

The outcome should be a model that more accurately reflects the principal characteristics of the data at hand, but the risk of iterative modeling is that we will capitalize on chance—overfitting the data and overstating the strength of our results. The same risk inheres in the model selection and model-averaging strategies described in this chapter (although the use of penalized model selection criteria such as the AIC and BIC at least partly mitigates this risk). It is obviously problematic to employ the same data both to explore and to validate a statistical

⁴³The term *cross-validation* often is also used for the validation procedure described in the current section, but I believe that it makes more semantic sense to reserve that term for applications in which the roles of training and validation samples are reversed or rotated—hence “crossed.”

model, but the apparent alternative of analyzing data blindly, simply to preserve the “purity” of classical statistical inference, is surely worse.

An ideal solution to this dilemma would be to collect new data with which to validate a model, but this solution is often impractical. Lack of funds or other constraints may preclude the collection of new data, and, in certain circumstances—for example, the examination of historical records—it is impossible even in principle to collect new data.

Model validation simulates the collection of new data by randomly dividing the data that we have in hand into two, possibly equal, parts—the first part to be used for exploration and model formulation, the second for checking the adequacy of the model, formal estimation, and testing. This is such a simple idea that it hardly requires detailed explanation. Perhaps the only subtle point is that the division of the data into exploratory and validation subsamples can exploit the sampling structure of the data. If, for example, the data are collected in a social survey employing a stratified sample, each stratum can be randomly divided between the two subsamples; of course, methods of analysis appropriate to a stratified sample should be employed.⁴⁴

When the same data are employed for selecting a statistical model and for drawing statistical inferences based on the model, the integrity of the inferences is compromised. Validation is a general strategy for protecting the accuracy of statistical inference when—as is typically the case—it is not possible to collect new data with which to assess the model. In model validation, the data at hand are divided at random into two subsamples: a training subsample, which is used to select a statistical model for the data, and a validation subsample, which is used for formal statistical inference.

22.3.1 An Illustration: Refugee Appeals

To illustrate model validation, I will present an abbreviated account of some research that I conducted on the Canadian refugee determination process, employing data that were collected and described by Greene and Shaffer (1992).⁴⁵ Greene and Shaffer’s data pertain to decisions of the Canadian Federal Court of Appeal on cases filed by claimants who were refused refugee status by the Immigration and Refugee Board. The court either granted or denied leave to appeal the board’s decision, and a single judge (rather than the usual tribunal) heard the request for leave to appeal.

During the period of the study, the 10 judges who adjudicated these cases varied widely in the rates with which they approved requests for leave to appeal negative decisions of the Refugee Board—with approval rates ranging from 13% to 56% of cases. The cases were not assigned randomly to judges, however, but rather were heard on a rotating basis. Although it seems unlikely that this procedure would introduce systematic differences into the leave requests processed by different judges, it is conceivable that this is the case. In defending the fairness of the existing procedure, the Crown therefore contended that it was insufficient

⁴⁴See Section 15.5 for a discussion of analyzing data from complex survey samples.

⁴⁵Also see Section 1.2. The analysis of the refugee data reported in the current section uses a larger subset of cases.

Table 22.4 Wald Tests for Terms in the Linear-Logit Model for the Canadian Refugee Data Training and Validation Subsamples

Term	df	Subsample					
		Training		Validation			
		Model 1	Model 2	Model 3			
Country success	1	14.55	.0001	15.72	.0001	25.64 <.0001	
High-refugee country	1	0.09	.77				
Region	4	6.47	.17				
Latin America	1	4.44	.035	4.98	.026	0.58 .45	
Time period	4	9.24	.055				
Linear	1	6.46	.011	5.74	.017	0.98 .32	
Quadratic	1	1.73	.19				
Cubic	1	0.16	.69				
Quartic	1	0.19	.67				
Judge	9	29.75	.0005	29.67	.0005	37.45 <.0001	

NOTE: Z^2 is the Wald test statistic.

simply to demonstrate large and statistically significant differences among the rates of approval for the 10 judges.

To determine whether systematic differences among the cases heard by the judges could account for differences in their judgments, I controlled statistically for several factors that—it was suggested—might influence the decisions, including the following:

1. the rate of success of leave applications from the applicant's country,
2. whether or not this country was identified as a "high refugee producing" country,
3. the region of the world in which the applicant's country is located (Latin America, Europe, Africa, the Middle East, or Asia and the Pacific Islands), and
4. the date of the applicant's case.

These explanatory variables were included in a logistic regression, along with dummy variables identifying the judge who heard the case. The response variable was whether or not leave to appeal was granted. Prior to constructing the logistic regression model, the roughly 800 cases meeting the criteria for inclusion in the study were randomly divided into training and validation subsamples. The data in the training subsample were carefully examined, and several variations of the analysis were undertaken. For example, the date of the case was treated both as a quantitative variable with a linear effect and categorically, divided into five quarters (the period of the study was slightly in excess of 1 year).

Wald tests for terms in two of the models fit to the data in the exploratory subsample are shown in Table 22.4.⁴⁶ Model 1 contains two explanatory variables—national rate of success

⁴⁶Were I to do this analysis now, I would prefer likelihood-ratio to Wald tests.

and judge—that are highly statistically significant, high-refugee country has a small and non-significant coefficient, and the region and time-period effects are relatively small but approach statistical significance. Examination of the region coefficients suggested that applicants from Latin America might be treated differently from those from other regions; likewise, the resolution of time period into orthogonal polynomial components suggested a possible linear effect of time. Model 2 is a final model for the exploratory subsample, incorporating a dummy regressor for Latin America and a linear trend over the five time periods, both of which appear to be statistically significant.

The last columns of Table 22.4 (for Model 3) show the result of refitting Model 2 to the data from the validation subsample. The national rate of success and judge are both still highly statistically significant, but neither the coefficient for Latin America nor the linear time trend proves to be statistically significant. That these latter two coefficients appeared to be statistically significant in the exploratory subsample illustrates the risk of selecting and testing a model on the same data. Most notably, however, differences among judges (not shown) are essentially the same before and after controlling for the other explanatory variables in the analysis.

22.3.2 Comments on Model Validation

Like the bootstrap (described in the preceding chapter), model validation is a good, simple, broadly applicable procedure that is rarely used in social research.⁴⁷ I believe that researchers resist the idea of dividing their data in half. In very small samples, division of the data is usually not practical. Even in samples of moderate size, however (such as the refugee-appeal data discussed in the previous section), halving the sample size makes it more difficult to find “statistically significant” results.

Yet, if statistical inference is to be more than an incantation spoken over the data, it is necessary to conduct research honestly. This is not to say that procedures of inference cannot be approximate—simplifying abstraction of some sort is unavoidable—but it is easy to introduce substantial errors of inference when the same data are used both to formulate and to test a statistical model.

Model validation is not a panacea for these problems, but it goes a long way toward solving them. Issues such as variable selection and choice of transformation are neatly handled by validation. Problems such as influential data are less easily dealt with, because these problems are particular to specific observations: That we locate an outlier in the training subsample, for example, does not imply that an outlier is present in the validation subsample. The reverse could be true as well, of course. We can, however, use the distribution of residuals in the training subsample to help us decide whether to use a method of estimation in the validation subsample that is resistant to unusual data or to adopt a rule for rejecting outliers.

⁴⁷Barnard (1974, p. 133) put it nicely: “The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number where it is actually used.”

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 22.1. Variable selection with randomly generated “noise” (adapted from Freedman, 1983):

- (a) Sampling from the standard normal distribution, independently generate 500 observations for 101 variables. Call the first of these variables the response variable Y and the other variables the predictors X_1, X_2, \dots, X_{100} . Perform a linear least-squares regression of Y on X_1, X_2, \dots, X_{100} . Are any of the individual regression coefficients “statistically significant”? Is the omnibus F -statistic for the regression “statistically significant”? Is this what you expected to observe? (*Hint:* What are the “true” values of the regression coefficients $\beta_1, \beta_2, \dots, \beta_{100}$?)
- (b) Retain the three predictors in part (a) that have the largest absolute t -values, regressing Y *only* on these variables. Are the individual coefficients “statistically significant”? What about the omnibus F ? What happens to the p -values compared to part (a)?
- (c) Using any method of variable selection (stepwise regression or subset regression with any criterion), find the “best” model with three explanatory variables. Obtain the individual t -statistics and omnibus F for this model. How do these tests compare to those in part (a)?
- (d) Using the methods of model selection discussed in this chapter, find the “best” model for these data. How does that model compare to the true model that generated the data?
- (e) Validation: Generate a new set of 500 observations as in part (a), and use that new data set to validate the models that you selected in parts (b), (c), and (d). What do you conclude?
- (f) Repeat the entire experiment several times.

Exercise 22.2. *Prove that Mallows’s C_p statistic,

$$C_{p_j} = \frac{\text{RSS}_j}{S_E^2} + 2s_j - n$$

can also be written

$$C_{p_j} = (k + 1 - s_j)(F_j - 1) + s_j$$

where RSS_j is the residual sum of squares for model M_j ; s_j is the number of parameters (including the constant) in model M_j ; n is the number of observations; S_E^2 is the usual estimate of error variance for the full model, which has k coefficients (excluding the constant); and F_j is the incremental F -statistic for testing the null hypothesis that the $k + 1 - s_j$ coefficients missing from model M_j are 0.

Exercise 22.3. Both the adjusted R^2 ,

$$\tilde{R}^2 = 1 - \frac{n-1}{n-s} \times \frac{\text{RSS}}{\text{TSS}}$$

and the generalized cross-validation criterion

$$\text{GCV} = \frac{n \times \text{RSS}}{(n - s)^2}$$

penalize models that have large numbers of predictors. (Here, n is the number of observations, s the number of parameters in the model, RSS the residual sum of squares under the model, and TSS the total sum of squares.) Do these two criteria necessarily rank a set of models in the same order? That is, if one model has a larger \tilde{R}^2 than another, does it necessarily also have a smaller GCV?

Exercise 22.4. Show that the differences in BIC values given in the first column of Table 22.1 (page 680) correspond roughly to the Bayes factors and posterior model probabilities given in columns 2 and 3 of the table.

Exercise 22.5. Perform model selection for the baseball salary regression using a criterion or criteria different from the BIC, examining the “best” model of each size, and the “best” 10 or 15 models regardless of size. Are the models similar to those nominated by the BIC? Why did you obtain these results?

Exercise 22.6. *Burnham and Anderson (2004) suggest the following procedure for model averaging based on the AIC: Let AIC_{\min} represent the smallest AIC among a set of models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$, and let $\text{AIC}_j^* \equiv \text{AIC}_j - \text{AIC}_{\min}$. Then *Akaike model weights* are given by

$$w_j \equiv \frac{\exp\left(-\frac{1}{2}\text{AIC}_j^*\right)}{\sum_{j'=1}^m \exp\left(-\frac{1}{2}\text{AIC}_{j'}^*\right)}$$

Model-averaged regression coefficients and their sampling variances are defined using these weights:

$$\begin{aligned} \tilde{\beta}_\ell &\equiv \sum_{j=1}^m w_j \hat{\beta}_\ell^{(j)} \\ \tilde{V}(\tilde{\beta}_\ell) &\equiv \left[\sum_{j=1}^m w_j \sqrt{\hat{V}(\hat{\beta}_\ell^{(j)}) + (\hat{\beta}_\ell^{(j)} - \tilde{\beta}_\ell)^2} \right]^2 \end{aligned}$$

where $\hat{\beta}_\ell^{(j)}$ is the MLE of β_ℓ and $\hat{V}(\hat{\beta}_\ell^{(j)})$ is the estimated sampling variance of this coefficient in model M_j .

- (a) How does this procedure compare with model averaging based on the BIC, as described in Section 22.2?
- (b) Restricting attention to the subset of models with AIC within 6 of the “best” model (i.e., applying the idea of Occam’s window to the AIC), find the model-averaged regression coefficients and their estimated variances for the baseball salary regression. Burnham and Anderson indicate that *model averaging* based on the AIC and BIC tends to produce results that are more similar than *model selection* based on the two criteria. Does that hold true here? Note: Following Burnham and Anderson, include 0 coefficients (with variances of 0) in the averages.

Exercise 22.7. *Show that when there are only two models M_1 and M_2 under consideration, the approximate posterior probability of the first model is

$$p(M_1|\mathbf{y}) \approx \frac{\exp(-\frac{1}{2}\text{BIC}_1)}{\exp(-\frac{1}{2}\text{BIC}_1) + \exp(-\frac{1}{2}\text{BIC}_2)}$$

Extend this result to the posterior probability of model M_j in the set of models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$:

$$p(M_j|\mathbf{y}) \approx \frac{\exp(-\frac{1}{2}\text{BIC}_j)}{\sum_{j'=1}^m \exp(-\frac{1}{2}\text{BIC}_{j'})}$$

Summary

- It is problematic to use statistical hypothesis tests for model selection. Doing so leads to issues of simultaneous inference, can produce biased results, tends to yield complicated models in large samples, and exaggerates the precision of results.
- Several criteria are employed for comparing statistical models with differing numbers of parameters, some applicable to regression models fit by least squares and others more general:

- The squared multiple correlation adjusted for degrees of freedom,

$$\tilde{R}^2 = 1 - \frac{n-1}{n-s} \times \frac{\text{RSS}}{\text{TSS}}$$

where n is the number of observations, s is the number of regression coefficients in the model, RSS is the residual sum of squares under the model, and TSS is the total sum of squares for the response variable.

- Mallows's C_p statistic,

$$C_p = (k+1-s)(F-1) + s$$

where k is the number of predictors in the full model fit to the data and F the incremental F -statistic for the hypothesis that the $k+1-s$ predictors excluded from the model are 0. C_p is an estimate of the total MSE of prediction for the model. A good model has C_p close to or below s .

- The cross-validation criterion,

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

where \hat{Y}_{-i} is the fitted value for observation i obtained when the model is fit with observation i omitted.

- The generalized cross-validation criterion,

$$\text{GCV} = \frac{n \times \text{RSS}}{df_{\text{res}}^2}$$

where df_{res} is the residual degrees of freedom under the model.

- The AIC,

$$\text{AIC} = -2 \log_e L(\hat{\theta}) + 2s$$

where $\log_e L(\hat{\theta})$ is the maximized log-likelihood under the model (and θ is the parameter vector for the model).

- The BIC,

$$\text{BIC} = -2 \log_e L(\hat{\theta}) + s \log_e n$$

For both the AIC and the BIC, the model with the smallest value is the one most supported by the data.

- The AIC is based on the Kullback-Leibler information comparing the true distribution of the data $p(\mathbf{y})$ to the distribution of the data $p_j(\mathbf{y}|\theta_j)$ under a particular model M_j .
- The BIC has its basis in Bayesian hypothesis testing, comparing the degree of support in the data for two models. The BIC is an approximation to twice the log of the Bayes factor comparing a particular model to the saturated model, where the Bayes factor is the ratio of the marginal probability of the data under the two models. When the prior probabilities for the two models are the same, the posterior odds for the models are equal to the Bayes factor. Differences in BIC approximate twice the log of the Bayes factor comparing two models to each other. The BIC approximation to the Bayes factor is accurate for a particular choice of prior distribution over the parameters of the models, called the unit-information prior, but may not be accurate for other priors. Differences in BIC of about 6 or more represent strong evidence in favor of the model with the smaller BIC.
- When the focus is on interpretation rather than prediction, researchers should feel free to simplify a statistical model on the basis of substantive considerations, even if that means removing small but statistically significant terms from the model. Penalized model selection criteria, such as the BIC, often provide an unnecessary excuse for doing so.
- The posterior probability for each model M_j in a set of models $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$ can be approximated using the BIC:

$$p(M_j|\mathbf{y}) \approx p_j = \frac{\exp(-\frac{1}{2}\text{BIC}_j)}{\sum_{j'=1}^m \exp(-\frac{1}{2}\text{BIC}_{j'})}$$

Then, for a model output such as the regression coefficient β_ℓ of the predictor X_ℓ , the posterior probability that β_ℓ is not 0 is

$$\Pr(\beta_\ell \neq 0|\mathbf{y}) \approx \sum_{j:M_j \in \mathcal{A}_\ell} p_j$$

where \mathcal{A}_ℓ is the subset of models that include the predictor X_ℓ . Conditional on $\beta_\ell \neq 0$, the posterior mean and variance of β_ℓ are approximately

$$\begin{aligned} E(\beta_\ell|\mathbf{y}, \beta_\ell \neq 0) &\approx \tilde{\beta}_\ell = \sum_{j:M_j \in \mathcal{A}_\ell} p'_j \hat{\beta}_\ell^{(j)} \\ V(\beta_\ell|\mathbf{y}, \beta_\ell \neq 0) &\approx \sum_{j:M_j \in \mathcal{A}_\ell} p'_j \left[\hat{V}(\hat{\beta}_\ell^{(j)}) + \hat{\beta}_\ell^{(j)2} \right] - \tilde{\beta}_\ell^2 \end{aligned}$$

where $\widehat{\beta}_\ell^{(j)}$ is the MLE of β_ℓ , $\widehat{V}(\widehat{\beta}_\ell^{(j)})$ is the estimated sampling variance of this coefficient in model M_j , and $p'_j = p_j / \sum_{j': M_{j'} \in \mathcal{A}_\ell} p_{j'}$. Because the number of candidate models can be extremely large, there is an advantage to restricting attention only to models with relatively large posterior probabilities, such as those with BIC within 6 of the best model; these models are said to fall within Occam's window. Because the meaning of a partial regression coefficient depends on the other explanatory variables in the model, however, model-averaged regression coefficients can be difficult to interpret.

- When the same data are employed for selecting a statistical model and for drawing statistical inferences based on the model, the integrity of the inferences is compromised. Validation is a general strategy for protecting the accuracy of statistical inference when—as is typically the case—it is not possible to collect new data with which to assess the model. In model validation, the data at hand are divided at random into two subsamples: a training subsample, which is used to select a statistical model for the data, and a validation subsample, which is used for formal statistical inference.

Recommended Reading

- There is a vast literature in statistics on automatic methods of model selection, most recently under the rubric of “data mining.” Hastie, Tibshirani, and Friedmann (2009), all of whom have made important contributions in this area, provide a broad overview.
- Several interesting papers on model selection appeared in the November 2004 issue of *Sociological Methods and Research* (Volume 33, Number 2). See, in particular, Burnham and Anderson’s paper on the AIC and Stine’s paper on information-theoretic methods of model selection.
- Burnham and Anderson (1998) present an extended exposition of the use of the AIC in model selection and its roots in information theory.
- Raftery (1995) provides a largely accessible introduction to the BIC; Kass and Raftery (1995) cover much of the same ground but provide greater statistical detail. These papers also discuss Bayesian model averaging.
- Weakliem (1999) presents a critique of the use of the BIC in model selection; his paper is followed by commentary from several authors, including Raftery.
- Bailey, Harding, and Smith (1989) give an overview of model validation. Some more detail can be found in Mosteller and Tukey (1968, 1977).

PART VI

Mixed-Effects Models

23

Linear Mixed-Effects Models for Hierarchical and Longitudinal Data

Recall the standard linear model¹

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ \varepsilon_i, \varepsilon_{i'} &\text{ independent for } i \neq i' \end{aligned}$$

Here, I use β_1 in preference to α for the regression constant to simplify the notation later in this chapter. The standard linear model has one *random effect*, the error term ε_i , and one *variance component*, $\sigma_\varepsilon^2 = V(\varepsilon_i)$. When the assumptions of the linear model hold, *ordinary least-squares (OLS)* regression provides maximum-likelihood estimates of the regression coefficients, B_1, B_2, \dots, B_p . The MLE of the error variance σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum E_i^2}{n}$$

where the residual E_i is

$$E_i = Y_i - (B_1 + B_2 X_{i2} + \cdots + B_p X_{ip})$$

Because $\hat{\sigma}_\varepsilon^2$ is a biased estimator of σ_ε^2 , recall that we usually prefer the unbiased estimator

$$S_E^2 = \frac{\sum E_i^2}{n-p}$$

with the difference between $\hat{\sigma}_\varepsilon^2$ and S_E^2 vanishing as the sample size n grows.²

The standard linear model and OLS regression are generally inappropriate for dependent observations. In Chapter 16, I pursued an alternative to OLS regression when the observations are ordered in time, and in Section 15.5, I discussed briefly the complications arising in complex survey-sampling designs, partly because of dependencies that are induced among observations. One pattern of dependency is *clustering*, where the observations are divided into related subgroups (*clusters*). Clustered data arise in many contexts (including, incidentally, complex survey samples), the two most common of which are *hierarchical data* and *longitudinal data*.

¹The standard linear model is developed in Part II of the text.

²These basic results for the linear model are developed in Section 9.3.

The current chapter deals with linear models for hierarchical and longitudinal data when there is a quantitative response variable. The following chapter develops generalized linear mixed-effects models for non-normal response variables and fundamentally nonlinear mixed-effects models for quantitative responses.

23.1 Hierarchical and Longitudinal Data

Hierarchical data arise when data collection takes place at two or more *levels*, one *nested* within the other. Some examples of hierarchical data:

- Schools are sampled from a population of schools, and then students are sampled within each school. Here there are two levels, with schools at the *higher level* (“Level 2”), and students nested within schools at the *lower level* (“Level 1”).
- Schools are sampled from a population of schools, classrooms are sampled within each school, and data are collected on students within classrooms. Here there are three levels—schools, classrooms, students—with students at the lowest level.
- Individuals are sampled within nations, a two-level hierarchy with nations at Level 2 and individuals at Level 1.³
- Individuals are sampled within communities within nations, a three-level hierarchy.
- Patients are sampled within physicians (two levels).
- Patients are sampled within physicians within hospitals (three levels).

There are also *nonnested* multilevel data—for example, high school students who each have multiple teachers. Such situations give rise to mixed-effects models with *crossed random effects*, a topic not pursued in this text.⁴

Longitudinal data are collected when individuals (or other multiple units of observation) are followed over time. Some examples of longitudinal data:

- Annual data on vocabulary growth among children
- Biannual data on weight preoccupation and exercise among adolescent girls
- Data collected at irregular intervals on recovery of IQ among coma patients
- Annual data on employment and income for a sample of adult Canadians

In all of these cases of hierarchical and longitudinal data, it is unreasonable to assume that observations within the same higher-level unit, or longitudinal observations within the same individual, are independent of one another. Longitudinal data also raise the possibility of serially correlated errors,⁵ in addition to dependency due to clustering. *Linear mixed-effect models*, the subject of this chapter, take account of dependencies in hierarchical, longitudinal, and other dependent data. Unlike the standard linear model, linear mixed-effect models include more than one source of random variation—that is, more than one random effect.

³It is unlikely that the nations in a study would literally be sampled from a larger population of nations, an issue that I address briefly below.

⁴But see the references given at the end of the chapter.

⁵Serially correlated errors were discussed in the context of time-series regression in Chapter 16.

Clustered data commonly arise in two contexts: hierarchical data, in which lower-level units, such as individual students, are nested within higher-level units, such as schools, and longitudinal data, in which individuals (or other multiple units of observation) are followed over time. In both cases, observations within a cluster—lower-level units within higher-level units or different measurement occasions for the same individual—cannot reasonably be treated as statistically independent. Mixed-effect models take account of dependencies in hierarchical, longitudinal, and other dependent data.

An important general point about mixed-effects models is that clustering should not simply be construed as a nuisance: Clustered data often allow us to address questions that cannot be answered effectively with completely independent observations—such as questions concerning *trajectories* of individual change over time or the *contextual effects* of characteristics of higher-level units (such as schools) on lower-level units (such as individual students).

Mixed-effects models have been developed in a variety of disciplines, with varying names and terminology: *random-effects models* (statistics, econometrics), *variance and covariance-component models* (statistics), *hierarchical linear models* (education), *multilevel models* (sociology), *contextual-effects models* (sociology, political science), *random-coefficient models* (econometrics), and *repeated-measures models* (statistics, psychology).⁶ Mixed-effects models also have a long history, dating to Fisher's (1925) and Yates's (1935) work on “split-plot” agricultural experiments. What distinguishes modern mixed models from their predecessors, however, is generality—for example, the ability to accommodate irregular and missing observations.

23.2 The Linear Mixed-Effects Model

This section introduces a very general *linear mixed* (or *mixed-effects*) *model* (abbreviated *LMM*), which I will adapt to particular circumstances. Please do not be put off by the complexity of the model: Not all parts of the model are used for particular applications, and I will presently introduce a variety of relatively simple instances of linear mixed models.

The *Laird-Ware form* of the linear mixed model (so called because it was introduced by Laird and Ware, 1982) is as follows:

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij} + \delta_{1i} Z_{1ij} + \cdots + \delta_{qi} Z_{qij} + \varepsilon_{ij} \\ \delta_{ki} &\sim N(0, \psi_k^2), C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'} \end{aligned} \tag{23.1}$$

$\delta_{ki}, \delta_{k'i}$ are independent for $i \neq i'$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 \lambda_{ij}), C(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_\varepsilon^2 \lambda_{ij'}$$

$\varepsilon_{ij}, \varepsilon_{ij'}$ are independent for $i \neq i'$

$\delta_{ki}, \varepsilon_{ij}$ are independent for all i, i', k, j (including $i = i'$)

⁶There is variation not only in terminology in different disciplines but also in estimation strategies. For example, rather than the likelihood-based methods emphasized in this and the next chapter, econometricians often use approaches based on generalized least squares (see Section 16.1) to estimate random-coefficient regression models, and psychologists often employ traditional approaches based on univariate and multivariate analysis of variance.

where

- Y_{ij} is the value of the response variable for the j th of n_i observations in the i th of m groups or clusters.
- $\beta_1, \beta_2, \dots, \beta_p$ are the *fixed-effect coefficients*, which are identical for all groups.
- X_{2ij}, \dots, X_{pij} are the *fixed-effect regressors* for observation j in group i ; there is also implicitly a constant regressor, $X_{1ij} = 1$.
- $\delta_{1i}, \dots, \delta_{qi}$ are the *random-effect coefficients* for group i , assumed to be (multivariately) normally distributed and independent of the random effects of other groups. The random effects, therefore, vary by group. The δ_{ki} are thought of as random variables, not as parameters, and are similar in this respect to the errors ε_{ij} : Were the study repeated, different clusters would be sampled (at least in principle) and thus the random effects would change. For example, if the clusters represent schools, with individual students observed within their schools, a replication of the study could, and likely would, sample different schools. Arguably, mixed-effects models may also apply even when the identity of the clusters (e.g., nations) would not change on replication of the study, if we can reasonably regard cluster effects as the outcome of a partly random process.⁷ I use the Greek letter δ to symbolize the random effects because, though they are random variables, the random effects are not directly observable.⁸
- Z_{1ij}, \dots, Z_{qij} are the *random-effect regressors*. The Z s are almost always a subset of the X s and may include *all* of the X s. When, as is frequently the case, there is a random intercept term, $Z_{1ij} = 1$.
- ψ_k^2 are the variances and $\psi_{kk'}$ the covariances among the random effects, assumed to be constant across groups. In some applications, the ψ s are parametrized in terms of a smaller number of fundamental parameters, as we will see later in this chapter.
- ε_{ij} is the *error* for observation j in group i . The errors for group i are assumed to be (multivariately) normally distributed and independent of errors in other groups and of the random effects.
- $\sigma_\varepsilon^2 \lambda_{ijj'}$ are the covariances among the errors in group i . Generally, the $\lambda_{ijj'}$ are parametrized in terms of a few basic parameters, and their specific form depends on context. When observations are sampled independently within groups and are assumed to have constant error variance (as is typical in hierarchical models), $\lambda_{ijj} = 1$, $\lambda_{ijj'} = 0$ (for $j \neq j'$), and thus the only free parameter to estimate is the common individual-level error variance, σ_ε^2 . If, alternatively, the observations in a “group” represent longitudinal data on a single individual, then the structure of the λ s may be specified to capture serial (i.e., over-time) dependencies among the errors.⁹

There are, then, two properties that distinguish the linear mixed model from the standard linear model: (1) There are structured cluster-level random effects δ_{ki} in the linear mixed model, in addition to the individual-level errors ε_{ij} , and (2) the mixed model can accommodate certain forms of nonconstant error variance and dependencies among the errors.

⁷This is implicitly the point of view that we take when we try to model Level 1 random effects as a function of Level 2 characteristics, as in Section 23.3.4.

⁸Although notation in the literature on mixed models is not entirely standardized, more commonly the lowercase Roman letter b is used to represent random effects.

⁹See Section 23.3.4.

The linear mixed-effects model (LMM) is applicable both to hierarchical and longitudinal data; in Laird-Ware form, the model is written

$$\begin{aligned}
 Y_{ij} &= \beta_1 + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij} + \delta_{1i} Z_{1ij} + \cdots + \delta_{qi} Z_{qij} + \varepsilon_{ij} \\
 \delta_{ki} &\sim N(0, \psi_k^2), C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'} \\
 \delta_{ki}, \delta_{k'i} &\text{ are independent for } i \neq i' \\
 \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2 \lambda_{ij}), C(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_\varepsilon^2 \lambda_{ijj'} \\
 \varepsilon_{ij}, \varepsilon_{ij'} &\text{ are independent for } i \neq i' \\
 \delta_{ki}, \varepsilon_{ij} &\text{ are independent for all } i, i', k, j \text{ (including } i = i')
 \end{aligned}$$

Here, Y_{ij} is the value of the response variable for the j th of n_i observations in the i th of m clusters, the β s are fixed-effect coefficients, the X s are fixed-effect regressors, the δ s are random-effect coefficients, the Z s are random-effect regressors, and the ε s are errors for individuals within clusters. The ψ s and λ s, which capture the dependencies among the random effects and errors within clusters, are typically expressed in terms of a small number of fundamental variance- and covariance-component parameters.

23.3 Modeling Hierarchical Data

Applications of mixed models to hierarchical data have become common in the social sciences and nowhere more so than in research on education. I will restrict myself to two-level models (students within schools, in the example developed below), but three or more levels (e.g., students within classrooms within schools) can also be handled through an extension of the Laird-Ware model. The two-level case, however, will allow me to develop the essential ideas.

The following example is borrowed from Raudenbush and Bryk (2012) and has been used by others as well. There are disadvantages as well as advantages to employing such a well-worn data set, but we will learn something about the data that apparently has not been noticed before.

The data are from the 1982 “High School and Beyond” survey and pertain to 7185 U.S. high school students from 160 schools—about 45 students on average per school. Seventy of the high schools are Catholic schools and 90 are public schools. An object of the data analysis is to determine how students’ math achievement scores on a standard test are related to their family socioeconomic status (SES).

I will entertain the possibility that the general level of math achievement and the relationship between achievement and SES vary among schools. If there is evidence of variation among schools, I will examine whether this variation is systematically related to school characteristics—specifically, whether the school is a public school or a Catholic school and the average SES of students in the school.

Just because we intend to fit a mixed-effects model to the data does not mean that we should forget data craft, and a good point of departure is to examine the relationship between math achievement and SES separately for each school. One hundred sixty schools are too many to look at individually, so I sampled 18 Catholic school and 18 public schools at random. Scatterplots of math achievement by SES for the sampled schools are in Figure 23.1.

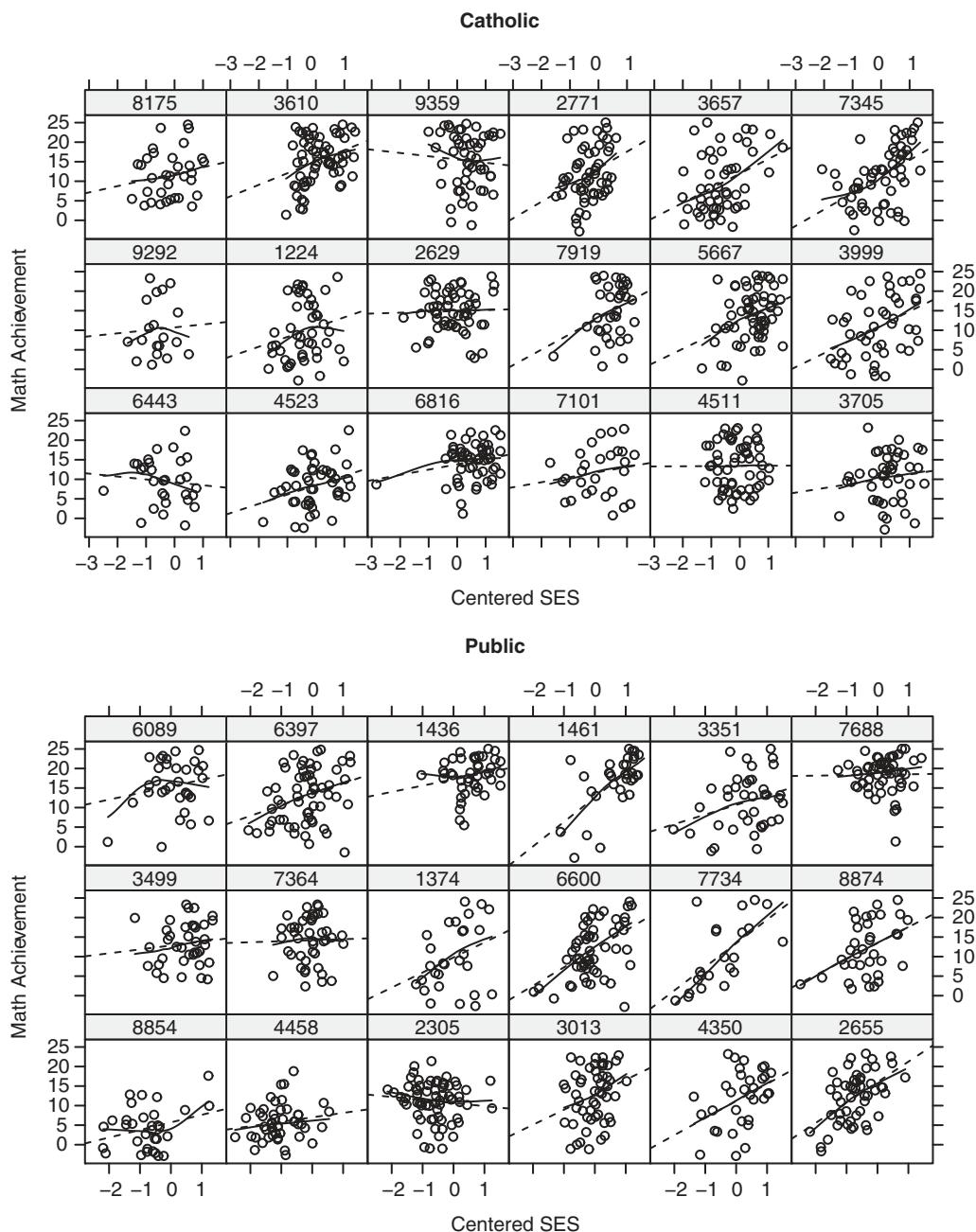


Figure 23.1 Math achievement by SES for students in 18 randomly selected Catholic high schools (upper panel) and 18 randomly selected public high schools (lower panel). SES is centered at the mean of each school. The broken line in each panel is the least-squares line, the solid line is for a nonparametric-regression smooth.

In each panel, the broken line is the linear least-squares fit to the data, while the solid line gives a nonparametric-regression fit.¹⁰ The number at the top of each panel is the ID number of the school. Taking into account the modest number of students in each school, the linear regressions seem to do a reasonable job of summarizing the relationship between math achievement and SES within schools. Although there is substantial variation in the regression lines among schools, there also seems to be a systematic difference between Catholic and public schools: The lines for the public schools appear to have steeper slopes on average.

SES in these scatterplots is expressed as deviations from the school mean SES. That is, the average SES for students in a particular school is subtracted from each individual student's SES. *Centering* SES in this manner makes the within-school intercept from the regression of math achievement on SES equal to the average math achievement score in the school. In the i th school, we have the regression equation¹¹

$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}(\text{ses}_{ij} - \bar{\text{ses}}_i) + \varepsilon_{ij} \quad (23.2)$$

where

$$\bar{\text{ses}}_i = \frac{\sum_{j=1}^{n_i} \text{ses}_{ij}}{n_i}$$

Then the least-squares estimate of the intercept is

$$\hat{\alpha}_{0i} = \overline{\text{mathach}_i} = \frac{\sum_{j=1}^{n_i} \text{mathach}_{ij}}{n_i}$$

A more general point is that it is helpful for interpretation of hierarchical (and other!) models to scale the explanatory variables so that the parameters of the model represent quantities of interest.¹²

Having satisfied myself that linear regressions reasonably represent the within-school relationship between math achievement and SES, I fit this model by least squares to the data from each of the 160 schools. Here are two displays of the least-squares regression coefficients: Figure 23.2 shows 95% confidence intervals for the intercept and slope estimates separately for Catholic and public schools, while Figure 23.3 shows boxplots of the intercepts and slopes for Catholic and public schools, facilitating the comparison between the two categories of schools. It is apparent that the individual slopes and intercepts are not estimated very precisely, not surprising given the relatively small number of students in each school, and there is also a great deal of variation in the regression coefficients from school to school. On average, however, Catholic schools have larger intercepts (i.e., a higher average level of math achievement) and lower slopes (i.e., less of a relationship between math achievement and SES) than public schools do.

¹⁰Methods of nonparametric regression are described in Chapter 18.

¹¹As in Chapter 8 on analysis of variance, a dot in a subscript indicates averaging over the corresponding index; thus, $\bar{\text{ses}}_i$ averages over individuals j in the i th school.

¹²See Section 23.7 for a discussion for centering explanatory variables in mixed-effects models and related subtle issues.

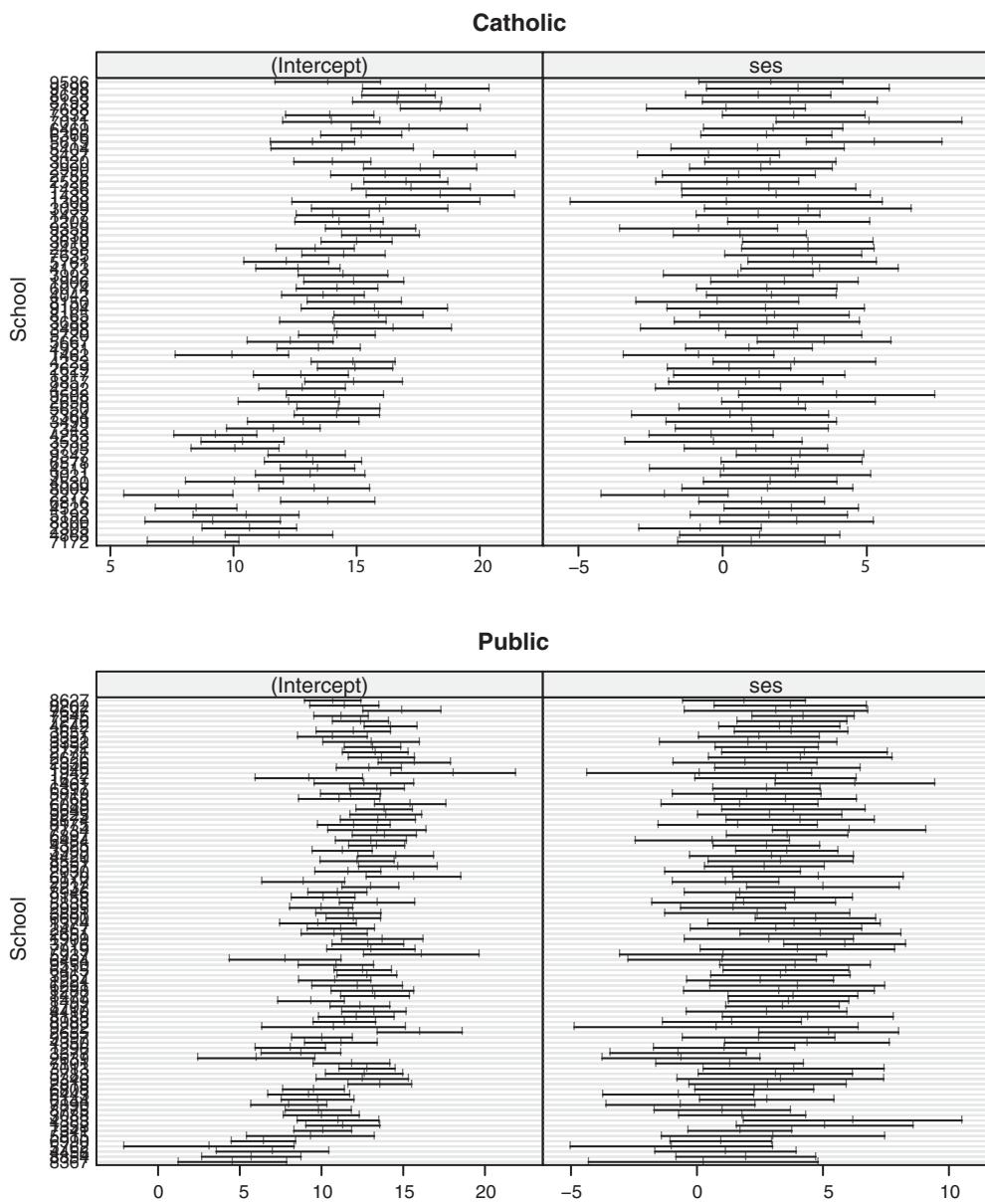


Figure 23.2 Ninety-five percent confidence intervals for least-squares intercepts (left) and slopes (right) for the within-school regressions of math achievement on school-centered SES: 70 Catholic high schools (top) and 90 public high schools (bottom). In comparing coefficients for Catholic and public schools, note that the scales of the coefficients in the top and bottom panels are not the same.

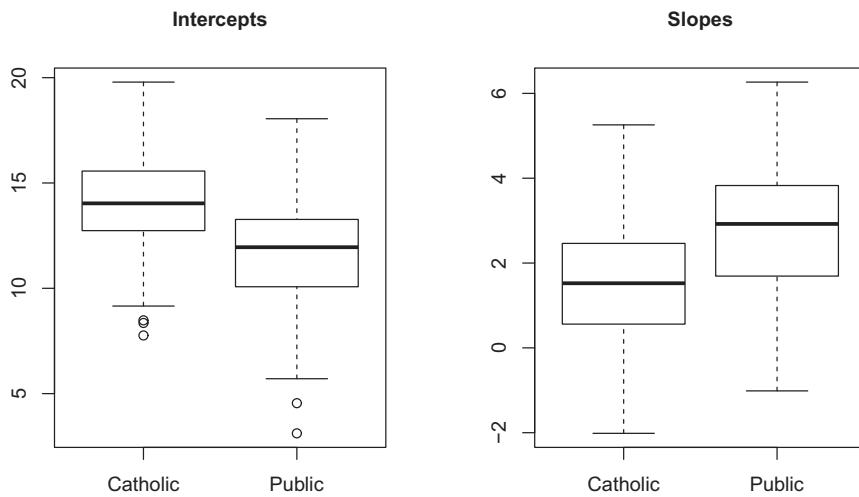


Figure 23.3 Boxplots of within-school coefficients for the least-squares regression of math achievement on school-centered SES, for 70 Catholic and 90 public schools: intercepts (left) and slopes (right).

Type of school—Catholic or public—is a *contextual variable*, characteristic of the higher-level units in the study, schools; the effect, if any, of type of school on individual students’ math achievement is a *contextual effect*. The average level of SES in each school is also a characteristic of the school, but it is derived from the lower-level units—the individual students. Average SES is therefore a *compositional variable*; a compositional variable can also have a contextual effect, in this instance distinct from the effect of students’ individual family SES.¹³ Figure 23.4 shows the relationship between the within-school intercepts and slopes and mean school SES. There is a moderately strong and reasonably linear relationship between the within-school intercepts (i.e., average math achievement) and the average level of SES in the schools. The slopes, however, are weakly and, apparently, nonlinearly related to average SES. As far as I know, despite the ubiquity of the High School and Beyond data in the literature on mixed-effects models, the nonlinear relationship between slopes and mean SES has not been noticed previously.

23.3.1. Formulating a Mixed Model

We already have a Level 1 model relating individual students’ math achievement to their families’ socioeconomic status (from Equation 23.2):

¹³The distinction between contextual and compositional variables is not standard, and both kinds of variables are often called “contextual variables.” I believe, however, that it is conceptually useful to make this distinction.

Moreover, the contextual effect of a compositional variable—understood, in our example, as the expected difference in achievement for students with the same individual SES in two schools that differ by 1 in mean SES—is not estimated by the coefficient of school-mean SES in the model in which individual-student SES is centered at the school mean: See Section 23.7 on centering for an explanation of this and related points.

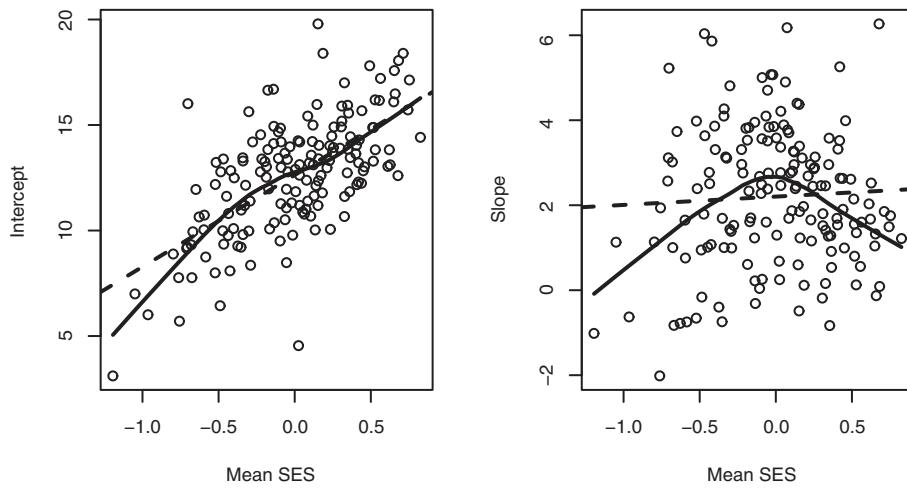


Figure 23.4 Within-school intercepts (left) and slopes (right) by school-mean SES. In each panel, the broken line is the linear least-squares fit, and the solid line is from a nonparametric regression.

$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$

where, for notational compactness, $\text{cses}_{ij} \equiv \text{ses}_{ij} - \bar{\text{ses}}_i$. (i.e., school-centered SES).

A Level 2 model relates the coefficients in the Level 1 model to characteristics of the schools. Our exploration of the data (in Figure 23.4) suggests the following Level 2 model:

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\bar{\text{ses}}_i + \gamma_{02}\text{sector}_i + \omega_{0i} \\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\bar{\text{ses}}_i + \gamma_{12}\bar{\text{ses}}_i^2 + \gamma_{13}\text{sector}_i + \omega_{1i}\end{aligned}\quad (23.3)$$

where sector is a dummy regressor, coded 1 (say) for Catholic schools and 0 for public schools. The linear specification for the Level 1 intercepts α_{0i} and quadratic specification for the Level 1 slopes α_{1i} follow from the patterns in Figure 23.4.

Substituting the school-level equation into the individual-level equation produces the *combined* or *composite model*:

$$\begin{aligned}\text{mathach}_{ij} &= (\gamma_{00} + \gamma_{01}\bar{\text{ses}}_i + \gamma_{02}\text{sector}_i + \omega_{0i}) \\ &\quad + (\gamma_{10} + \gamma_{11}\bar{\text{ses}}_i + \gamma_{12}\bar{\text{ses}}_i^2 + \gamma_{13}\text{sector}_i + \omega_{1i}) \text{cses}_{ij} + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{01}\bar{\text{ses}}_i + \gamma_{02}\text{sector}_i + \gamma_{10}\text{cses}_{ij} \\ &\quad + \gamma_{11}\bar{\text{ses}}_i \times \text{cses}_{ij} + \gamma_{12}\bar{\text{ses}}_i^2 \times \text{cses}_{ij} + \gamma_{13}\text{sector}_i \times \text{cses}_{ij} \\ &\quad + \omega_{0i} + \omega_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}\quad (23.4)$$

Notice that the coefficients of the contextual and compositional variables in the Level 2 equations of the hierarchical form of the mixed model (e.g., $\gamma_{13}\text{sector}_i$ in Equations 23.3) appear as coefficients of *cross-level interactions* in the composite form of the model (e.g., $\gamma_{13}\text{sector}_i \times \text{cses}_{ij}$ in Equation 23.4). Except for notation, this is a mixed model in Laird-Ware

form, as we can see by replacing γ s with β s, ω s with δ s, and the various regressors with X s and Z s:

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_6 X_{6ij} + \beta_7 X_{7ij} + \delta_{1i} + \delta_{2i} Z_{2ij} + \varepsilon_{ij}$$

For example, $X_{4ij} = Z_{2ij} = \text{cses}_{ij}$.

All of the regressors in the Laird-Ware form of the model carry subscripts i for schools and j for individuals within schools, even when the explanatory variable in question is constant within schools. Thus, for example, $X_{2ij} = \bar{\text{ses}}_i$. (and so all individuals in the same school share a common value of school-mean SES). There is both a data management issue here and a conceptual point: With respect to data management, software that fits the Laird-Ware form of the model requires that Level 2 explanatory variables (here sector and school-mean SES, which are characteristics of schools) appear in the Level 1 (i.e., student) data set. The conceptual point is that the model can incorporate contextual effects of contextual variables such as sector and compositional variables like school-mean SES—characteristics of the Level 2 units can influence the Level 1 response variable.

In modeling hierarchical data, it is often natural to formulate an individual-level model within clusters and then to treat the coefficients of that model as random effects that appear as the responses in a higher-level model. The models at the two levels can be combined as an LMM in Laird-Ware form. Contextual variables describe higher-level units; compositional variables also describe higher-level units but are derived from lower-level units (e.g., by averaging).

Rather than proceeding directly with this relatively complicated model, let us first investigate some simpler mixed-effects models derived from it.¹⁴

23.3.2 Random-Effects One-Way Analysis of Variance

Consider the following Level 1 and Level 2 models:

$$\begin{aligned}\text{mathach}_{ij} &= \alpha_{0i} + \varepsilon_{ij} \\ \alpha_{0i} &= \gamma_{00} + \omega_{0i}\end{aligned}$$

The combined model is

$$\text{mathach}_{ij} = \gamma_{00} + \omega_{0i} + \varepsilon_{ij}$$

or, in Laird-Ware form,

$$Y_{ij} = \beta_1 + \delta_{1i} + \varepsilon_{ij}$$

This is a *random-effects one-way ANOVA model* with one *fixed effect*, β_1 , representing the general population mean of math achievement, and two *random effects*, δ_{1i} , representing the deviation of math achievement in school i from the general mean—that is, the mean math

¹⁴I adopt the general strategy of exposition here, if not the details, from Raudenbush and Bryk (2012, chap. 4).

achievement for all students in school i , not just those students sampled, is $\mu_i = \beta_1 + \delta_{1i}$, and ε_{ij} , representing the deviation of individual j 's math achievement in school i from the school mean, $\varepsilon_{ij} = Y_{ij} - \mu_i$. Moreover, two observations, Y_{ij} and $Y_{ij'}$, in the same school i are not independent because they share the random effect, δ_{1i} .

There are also two *variance components* for this model: $\psi_1^2 \equiv V(\delta_{1i})$ is the variance among school means, and $\sigma_\varepsilon^2 \equiv V(\varepsilon_{ij})$ is the variance among individuals in the same school. The random effect δ_{1i} and the errors ε_{ij} are assumed to be independent, and therefore variation in math scores among individuals can be decomposed into these two variance components:

$$V(Y_{ij}) = E[(\delta_{1i} + \varepsilon_{ij})^2] = \psi_1^2 + \sigma_\varepsilon^2$$

because $E(\delta_{1i}) = E(\varepsilon_{ij}) = 0$, and hence $E(Y_{ij}) = \beta_1$.

The *intraclass correlation coefficient* ρ is the proportion of variation in individuals' scores due to differences among schools:

$$\rho \equiv \frac{\psi_1^2}{V(Y_{ij})} = \frac{\psi_1^2}{\psi_1^2 + \sigma_\varepsilon^2}$$

The intraclass correlation can also be interpreted as the correlation between the math scores of two individuals selected at random from the *same* school. That is,

$$\begin{aligned} C(Y_{ij}, Y_{ij'}) &= E[(\delta_{1i} + \varepsilon_{ij})(\delta_{1i'} + \varepsilon_{ij'})] = E(\delta_{1i}^2) = \psi_1^2 \\ V(Y_{ij}) &= V(Y_{ij'}) = \psi_1^2 + \sigma_\varepsilon^2 \\ \text{Cor}(Y_{ij}, Y_{ij'}) &= \frac{C(Y_{ij}, Y_{ij'})}{\sqrt{V(Y_{ij}) \times V(Y_{ij'})}} = \frac{\psi_1^2}{\psi_1^2 + \sigma_\varepsilon^2} = \rho \end{aligned}$$

There are two common methods for estimating linear mixed-effects models:¹⁵

- *Full maximum-likelihood (ML) estimation* maximizes the likelihood with respect to all of the parameters of the model simultaneously (i.e., both the fixed-effects parameters and the variance components).
- *Restricted (or residual) maximum-likelihood (REML) estimation* integrates the fixed effects out of the likelihood and estimates the variance components; given the resulting estimates of the variance components, estimates of the fixed effects are recovered.

A disadvantage of ML estimates of variance components is that they are biased downward in small samples (much as the ML estimate of the error variance in the standard linear model is biased downward). The REML estimates, in contrast, correct for loss of degrees of freedom due to estimating the fixed effects. The difference between the ML and REML estimates can be important when the number of clusters (i.e., Level 2 units) is small.

LMMs can be estimated by maximum likelihood (ML) or by restricted maximum likelihood (REML). When the number of clusters is small, REML tends to produce less biased estimates of variance components.

¹⁵Estimation is discussed in greater detail in Section 23.9.1.

ML and REML estimates for the random-effects one-way ANOVA model in the current example, where there are 160 schools (Level 2 units), are nearly identical:

Parameter	ML Estimate	REML Estimate
β_1	12.637	12.637
ψ_1	2.925	2.935
σ_ε	6.257	6.257

Here and elsewhere in the chapter, I show the standard deviations in preference to the variances of the random effects. It is generally simpler to interpret the standard deviations because they are on the scale of the response variable—in this case, math achievement.

The estimated intraclass correlation coefficient in this example is

$$\hat{\rho} = \frac{2.935^2}{2.935^2 + 6.257^2} = .180$$

and so 18% of the variation in students' math achievement scores is “attributable” to differences among schools. The estimated intercept, $\hat{\beta}_1 = 12.637$, represents the average level of math achievement in the population of schools.

23.3.3 Random-Coefficients Regression Model

Having established that there is variation among schools in students' math achievement, let us introduce school-centered SES into the Level 1 model as an explanatory variable,

$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$

and allow for random intercepts *and* slopes in the Level 2 model:

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \omega_{0i} \\ \alpha_{1i} &= \gamma_{10} + \omega_{1i}\end{aligned}$$

The combined model is now

$$\begin{aligned}\text{mathach}_{ij} &= (\gamma_{00} + \omega_{0i}) + (\gamma_{10} + \omega_{1i}) \text{cses}_{ij} + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{10}\text{cses}_{ij} + \omega_{0i} + \omega_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}$$

In Laird-Ware form,

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \delta_{1i} + \delta_{2i} Z_{2ij} + \varepsilon_{ij} \quad (23.5)$$

This model is a *random-coefficients regression model*. The fixed-effects coefficients β_1 and β_2 represent, respectively, the average within-schools population intercept and slope. Moreover, because SES is centered within schools, the intercept β_1 represents the general level of math achievement in the population (in the sense of the average within-school mean).

The model has four *variance-covariance components*:

- $\psi_1^2 \equiv V(\delta_{1i})$ is the variance among school intercepts (i.e., school means, because SES is school-centered).

- $\psi_2^2 \equiv V(\delta_{2i})$ is the variance among within-school slopes.
- $\psi_{12} \equiv C(\delta_{1i}, \delta_{2i})$ is the covariance between within-school intercepts and slopes.
- $\sigma_\varepsilon^2 \equiv V(\varepsilon_{ij})$ is the error variance around the within-school regressions.

The *composite error* for individual j in school i is

$$\zeta_{ij} \equiv \delta_{1i} + \delta_{2i}Z_{2ij} + \varepsilon_{ij}$$

The variance of the composite error is

$$V(\zeta_{ij}) = E(\zeta_{ij}^2) = E[(\delta_{1i} + \delta_{2i}Z_{2ij} + \varepsilon_{ij})^2] = \psi_1^2 + Z_{2ij}^2\psi_2^2 + 2Z_{2ij}\psi_{12} + \sigma_\varepsilon^2 \quad (23.6)$$

and the covariance of the composite errors for two individuals j and j' in the same school is

$$\begin{aligned} C(\zeta_{ij}, \zeta_{ij'}) &= E(\zeta_{ij} \times \zeta_{ij'}) = E[(\delta_{1i} + \delta_{2i}Z_{2ij} + \varepsilon_{ij})(\delta_{1i} + \delta_{2i}Z_{2ij'} + \varepsilon_{ij'})] \\ &= \psi_1^2 + Z_{2ij}Z_{2ij'}\psi_2^2 + (Z_{2ij} + Z_{2ij'})\psi_{12} \end{aligned} \quad (23.7)$$

The composite errors consequently have nonconstant variance, and errors for individuals in the *same* school are correlated. But the composite errors for two individuals in *different* schools are independent.

ML and REML estimates for the model are as follows:

Parameter	ML Estimate	Std. Error	REML Estimate	Std. Error
β_1	12.636	0.244	12.636	0.245
β_2	2.193	0.128	2.193	0.128
ψ_1	2.936		2.946	
ψ_2	0.824		0.833	
ψ_{12}	0.041		0.042	
σ_ε	6.058		6.058	

Again, the ML and REML estimates are very close.

I have shown standard errors only for the fixed effects. Standard errors for variance and covariance components can be obtained in the usual manner from the inverse of the information matrix,¹⁶ but tests and confidence intervals based on these standard errors tend to be inaccurate. We can, however, test variance and covariance components by a likelihood-ratio test, contrasting the (restricted) log-likelihood for the fitted model with that for a model removing the random effects in question. For example, for the current model (Model 1), removing $\delta_{12}Z_{2ij}$ from the model (producing Model 0) implies that the SES slope is identical across schools. Notice that removing $\delta_{12}Z_{2ij}$ from the model eliminates *two* variance-covariance parameters, ψ_2^2 and

¹⁶See online Appendix D for an introduction to maximum-likelihood estimation.

ψ_{12} . A likelihood-ratio chi-square test (based on the REML estimates) for these parameters on 2 degrees of freedom suggests that they should not be omitted from the model:¹⁷

$$\begin{aligned}\log_e L_1 &= -23,357.12 \\ \log_e L_0 &= -23,362.00 \\ G_0^2 &= 2(\log_e L_1 - \log_e L_0) = 9.76, df = 2, p = .008\end{aligned}$$

Cautionary Remark

Because REML estimates are calculated integrating out the fixed effects,¹⁸ one cannot legitimately perform likelihood-ratio tests across models with *different* fixed effects when the models are estimated by REML. Likelihood-ratio tests for variance-covariance components across nested models with *identical* fixed effects are perfectly fine, however. A common source of estimation difficulties in mixed models is the specification of overly complex random effects. Because interest usually centers on the fixed effects, it often pays to try to simplify the random-effect part of the model.

23.3.4 Coefficients-as-Outcomes Model

The *regression-coefficients-as-outcomes model* introduces explanatory variables at Level 2 to account for variation among the Level 1 regression coefficients. This returns us to the model that I originally formulated for the math achievement data: at Level 1,

$$\text{mathach}_{ij} = \alpha_{0i} + \alpha_{1i}\text{cses}_{ij} + \varepsilon_{ij}$$

and, at Level 2,

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_i + \gamma_{02}\text{sector}_i + \omega_{0i} \\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\overline{\text{ses}}_i + \gamma_{12}\overline{\text{ses}}_i^2 + \gamma_{13}\text{sector}_i + \omega_{1i}\end{aligned}\tag{23.8}$$

The combined model is

$$\begin{aligned}\text{mathach}_{ij} &= \gamma_{00} + \gamma_{01}\overline{\text{ses}}_i + \gamma_{02}\text{sector}_i \\ &\quad + \gamma_{10}\text{cses}_{ij} + \gamma_{11}\overline{\text{ses}}_i \times \text{cses}_{ij} + \gamma_{12}\overline{\text{ses}}_i^2 \times \text{cses}_{ij} + \gamma_{13}\text{sector}_i \times \text{cses}_{ij} \\ &\quad + \omega_{0i} + \omega_{1i}\text{cses}_{ij} + \varepsilon_{ij}\end{aligned}\tag{23.9}$$

or, in Laird-Ware form,

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_5 X_{5ij} + \beta_6 X_{6ij} + \beta_7 X_{7ij} + \delta_{1i} + \delta_{2i} Z_{2ij} + \varepsilon_{ij}\tag{23.10}$$

¹⁷A more careful formulation of this test takes into account the fact that the null hypothesis places the variance component $\psi_2^2 = 0$ on a boundary of the parameter space—of course, a variance cannot be negative. In contrast, the covariance component in the null hypothesis, $\psi_{12} = 0$, is *not* on a boundary of the parameter space. Under these circumstances, the distribution of the likelihood-ratio test statistic G_0^2 under H_0 is a mixture of chi-square distributions, not simply χ_2^2 . I will examine the computation of *p*-values for tests of variance and covariance components more carefully in Section 23.6.

¹⁸If you are unfamiliar with calculus, think of “integrating out” the fixed effects, as summing over the possible values of the fixed effects, with the values weighted by their likelihood.

This model has more fixed effects than the preceding random-coefficients regression model (Equation 23.5) but the same random effects and variance-covariance components: $\psi_1^2 \equiv V(\delta_{1i})$, $\psi_2^2 \equiv V(\delta_{2i})$, $\psi_{12} \equiv C(\delta_{1i}, \delta_{2i})$, and $\sigma_\varepsilon^2 \equiv V(\varepsilon_{ij})$.

After fitting this model to the data by REML, let us check whether random intercepts and slopes are still required:¹⁹

Model	Omitting	$\log_e L$
1	—	-23,247.70
2	ψ_1^2, ψ_{12}	-23,357.86
3	ψ_2^2, ψ_{12}	-23,247.93

The test for random intercepts, contrasting Models 1 and 2, is highly statistically significant,

$$G_0^2 = 2[(-23,247.70) - (-23,357.86)] = 220.32, df = 2, p \approx 0$$

but the test for random slopes is not,

$$G_0^2 = 2[(-23,247.70) - (-23,247.93)] = 0.46, df = 2, p = .80$$

Apparently, the Level 2 explanatory variables do a sufficiently good job of accounting for differences in slopes among schools that the variance component for slopes, and with it, the covariance component for slopes and intercepts, are no longer needed.

Refitting the model removing $\delta_{i2}Z_{2ij}$ produces the following REML estimates:

Parameter	Term	REML Estimate	Std. Error
β_1	intercept	12.128	0.199
β_2	$\bar{\text{ses}}_i$	5.337	0.369
β_3	sector_i	1.225	0.306
β_4	cses_{ij}	3.140	0.166
β_5	$\bar{\text{ses}}_i \times \text{cses}_{ij}$	0.755	0.308
β_6	$\bar{\text{ses}}_i^2 \times \text{cses}_{ij}$	-1.647	0.575
β_7	$\text{sector}_i \times \text{cses}_{ij}$	-1.516	0.237
ψ_1	SD(intercept)	1.542	
σ_ε	SD(ε_{ij})	6.060	

More often than not, primary interest inheres in the fixed effects—that is, the average effects of the explanatory variables across Level 1 units. The fixed-effects estimates, all of which exceed twice their standard errors and hence are statistically significant, have the following interpretations:²⁰

¹⁹Again, because $\psi_1^2 = 0$ and $\psi_2^2 = 0$ are on the boundaries of the parameter space, the p -values reported here are not quite correct; see Section 23.6 for a procedure to compute more accurate p -values in this situation.

²⁰See Sections 23.5 and 23.9.2 for more careful consideration of Wald tests for the fixed effects in an LMM.

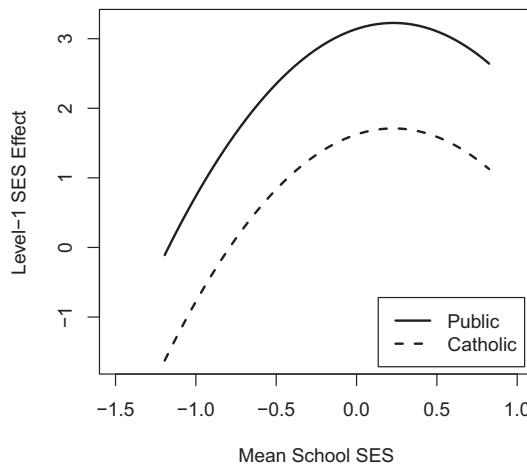


Figure 23.5 The Level 1 effect $\hat{\alpha}_1$ of school-centered SES as a function of type of school (Catholic or public) and mean school SES.

- $\hat{\beta}_1 = 12.128$ is the estimated general level of students' math achievement in public schools (where the dummy regressor for sector is coded 0) at mean school SES. The interpretation of this coefficient depends on the fact that \bar{ses}_i (school SES) is itself centered to a mean of 0 across schools.
- $\hat{\beta}_2 = 5.337$ is the estimated increase in students' mean math achievement associated with a 1-unit increase in school SES.
- $\hat{\beta}_3 = 1.225$ is the estimated difference in students' mean math achievement between Catholic and public schools at fixed levels of school SES. The fixed-effects coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, therefore, describe the *between-schools* regression of mean math achievement on school characteristics.
- Figure 23.4 shows how the coefficients $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$, and $\hat{\beta}_7$ combine to produce the Level 1 (i.e., *within-school*) coefficient for SES.²¹ At fixed levels of school SES, individual SES is more positively related to math achievement in public than in Catholic schools. The maximum positive effect of individual SES is in schools with a slightly higher than average SES level; the effect declines at low and high levels of school SES, and—for Catholic schools—becomes negative at the lowest levels of school SES.

An alternative and possibly more intuitive representation of the fitted model is shown in Figure 23.6, which graphs the estimated within-school regression of math achievement on school-centered SES for Catholic and public schools at three levels of school SES: -0.7 (the approximate 5th percentile of school SES among the 160 schools), 0 (the median), and 0.7 (the 95th percentile). The equations of these regression lines are derived by substituting the three levels of school SES into the Level 2 model of Equation 23.8, along with 0 (public) or 1

²¹From Equations 23.8, 23.9, and 23.10.

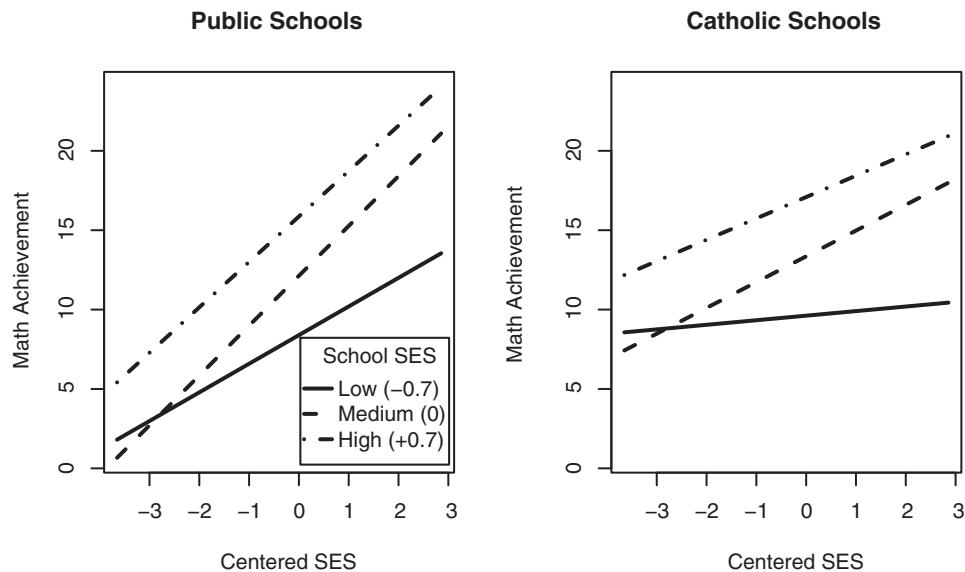


Figure 23.6 Fitted within-school regressions of math achievement on centered SES for public and Catholic schools at three levels of mean school SES.

(Catholic) for the sector dummy regressor.²² Figure 23.6 constitutes an “effect display” for the fixed effects of type of school, individual SES, and mean school SES.²³

The coefficients-as-outcomes model relates regression coefficients of lower-level units within clusters to characteristics of the clusters. Simpler hierarchical models include the random-effects one-way ANOVA model, in which each cluster has its own mean, treated as a random effect, and the random-coefficients regression model, in which several regression coefficients can vary randomly across clusters. In the random-effects one-way ANOVA model, the intraclass correlation measures the proportion of individual-level variation that is due to differences among clusters, $\rho = \psi_1^2 / (\psi_1^2 + \sigma_\epsilon^2)$, where ψ_1^2 is the variance component for clusters and σ_ϵ^2 is the error variance.

23.4 Modeling Longitudinal Data

In most respects, modeling longitudinal data—where there are multiple observations on individuals who are followed over time—is similar to modeling hierarchical data: We can think of individuals as analogous to Level 2 units and measurement occasions as analogous to Level 1

²²See Exercise 23.1.

²³Effect displays for linear models were introduced in Section 7.3.4.

units. Just as it is generally unreasonable to suppose in hierarchical data that observations for individuals in the same Level 2 unit are independent, so it is generally unreasonable to suppose that observations taken on different occasions for the same individual are independent.

An additional complication in longitudinal data is that it may no longer be reasonable to assume that the Level 1 errors ε_{ij} are independent of each other, because observations taken close in time on the same individual may well be more similar than observations farther apart in time. When this happens, we say that the errors are *autocorrelated*. The linear mixed model introduced in Section 23.2 makes provision for autocorrelated Level 1 errors. We encountered a similar phenomenon in time-series regression.²⁴ We can think of longitudinal data as comprising time series for each of a number of individuals, and indeed longitudinal data are sometimes described as *cross-sectional time series*. That said, the number of measurement occasions in longitudinal data is typically much smaller than in time-series data.

In composing a mixed model for longitudinal data, we can work either with the hierarchical form of the model or with the composite (Laird-Ware) form. Consider the following example, drawn from work by Davis, Blackmore, Katzman, and Fox (2005) on the exercise histories of 138 teenaged girls who were hospitalized for eating disorders and of 93 “control” subjects who did not suffer from eating disorders.²⁵ There are several observations for each subject, but because the girls were hospitalized at different ages, the number of observations and the age at the last observation vary. The data include the subject’s age, in years, at the time of observation, along with the amount of exercise in which the subject engaged, expressed as hours per week. All but the last observation for each subject were collected retrospectively at intervals of 2 years, starting at age 8. The age at the last observation is recorded to the nearest day.

It is of interest here to determine the typical trajectory of exercise over time and to establish whether this trajectory differs between eating-disordered and control subjects. Preliminary examination of the data suggests a log transformation of exercise, but because about 12% of the data values are 0, it is necessary to add a small constant to the data before taking logs. I used $5/60 = 1/12$ (i.e., 5 minutes). Figure 23.7 reveals that the original exercise scores are highly skewed, but that the log-transformed scores are much more symmetrically distributed. An alternative to transformation of Y would be to fit a model that takes explicit account of the nonnegative character of the response variable or that accounts explicitly for the 0s in the data.²⁶

Figure 23.8 shows the exercise trajectories for 20 randomly selected control subjects and 20 randomly selected patients. The small number of observations per subject and the substantial irregular intrasubject variation make it hard to draw conclusions, but there appears to be a more consistent pattern of increasing exercise among patients than among the controls. With so few observations per subject, and without clear evidence that it is inappropriate, we should be loath to fit a within-subject model more complicated than a linear trend.

A linear “growth curve” characterizing subject i ’s trajectory suggests the Level 1 model

$$\log\text{-exercise}_{ij} = \alpha_{0i} + \alpha_{1i}(\text{age}_{ij} - 8) + \varepsilon_{ij}$$

²⁴See Chapter 16.

²⁵I am grateful to Caroline Davis and Elizabeth Blackmore of York University, Toronto, for providing the data for this example. The analysis here is simplified from the original source, which took into account the age of onset of eating disorder for the patients: See the data analysis exercises for this chapter.

²⁶For example, we might extend to mixed-effects models the approach to censored data described in Section 20.5.

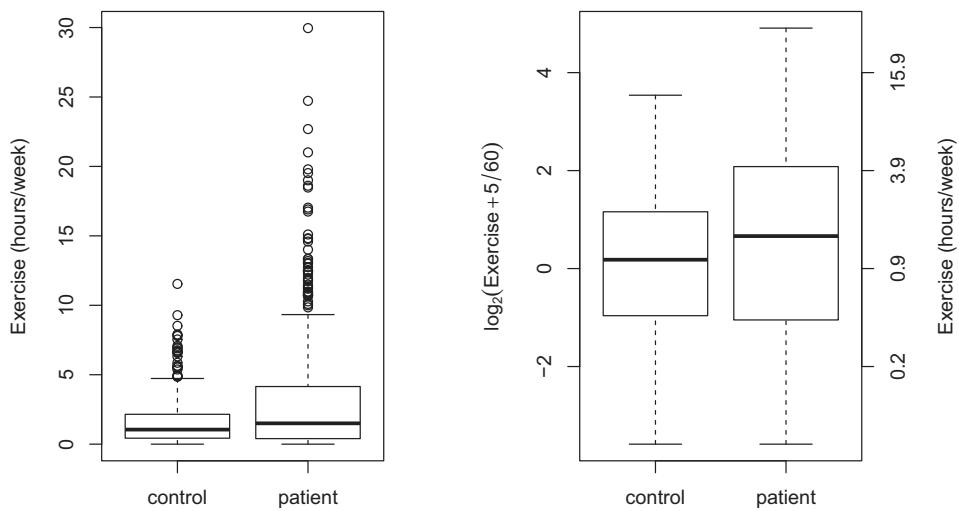


Figure 23.7 Boxplots of exercise (left) and log exercise (right) for controls and patients, for measurements taken on all occasions. Logs are to the base 2, and 5 minutes was added to the exercise times to avoid 0 values.

I have subtracted 8 from age, and so α_{0i} represents the level of exercise at 8 years of age—the start of the study. I fit this model by least-squares regression separately to the data for each subject. Because of the small number of observations per subject, we should not expect very good estimates of the within-subject regression coefficients. Indeed, one of the advantages of mixed models is that they can provide improved estimates of the within-subject coefficients (the random effects) by pooling information across subjects.²⁷ Boxplots of the resulting regression coefficients are shown in Figure 23.9. As expected, there is a great deal of variation in both the intercepts and the slopes. The median intercepts are similar for patients and controls, but there is somewhat more variation among patients. The slopes are higher on average for patients than for controls, for whom the median slope is close to 0.

Our interest in detecting differences in exercise histories between subjects and controls suggests the Level 2 model

$$\begin{aligned}\alpha_{0i} &= \gamma_{00} + \gamma_{01}\text{group}_i + \omega_{0i} \\ \alpha_{1i} &= \gamma_{10} + \gamma_{11}\text{group}_i + \omega_{1i}\end{aligned}$$

²⁷Pooled “estimates” of the random effects provide so-called *empirical best-linear-unbiased predictors* (or EBLUPs) of these coefficients. I put “estimates” in quotation marks because random effects are not formally parameters.

In the present example, it is possible to compute the least-squares estimates of the regression coefficients separately for each subject only because all subjects are observed at least twice. More generally, it may not be possible to fit the Level 1 model directly to the data for each individual. The mixed-effects model can nevertheless use (“pool”) the information from all subjects to provide “estimates” of the random effects even for individuals with insufficient data to estimate the regression coefficients directly by least squares. For more on BLUPs and EBLUPs, see Section 23.8.

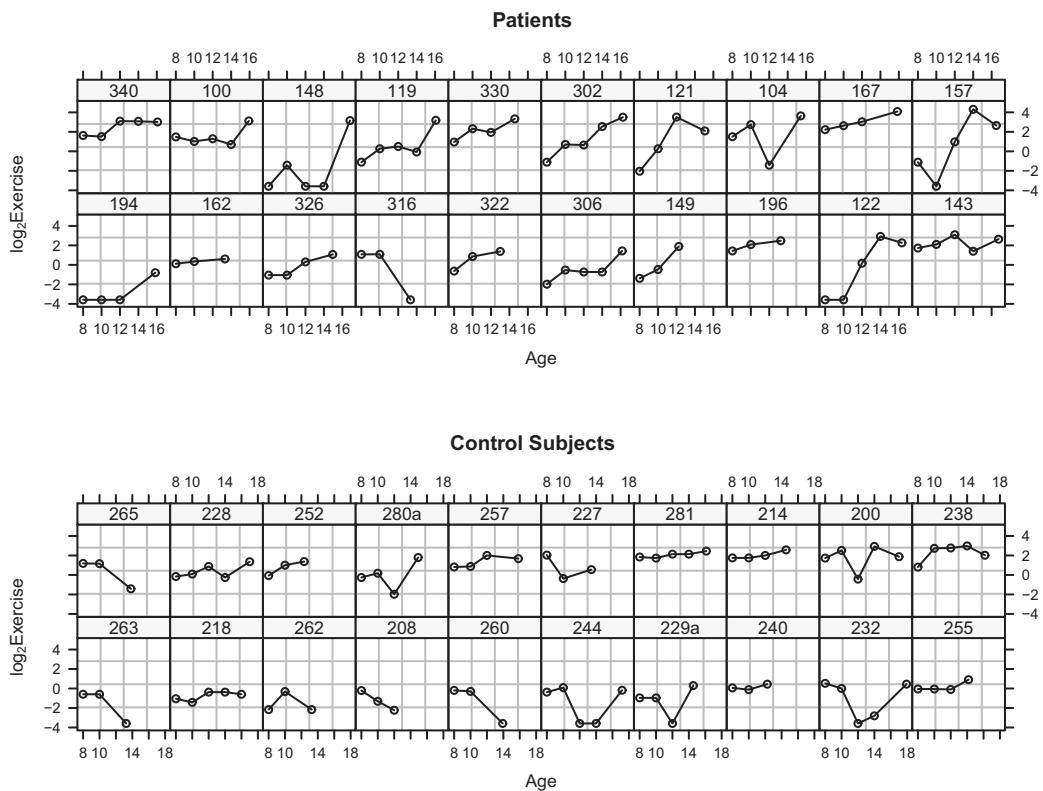


Figure 23.8 Exercise trajectories for 20 randomly selected patients (top) and 20 randomly selected controls (bottom).

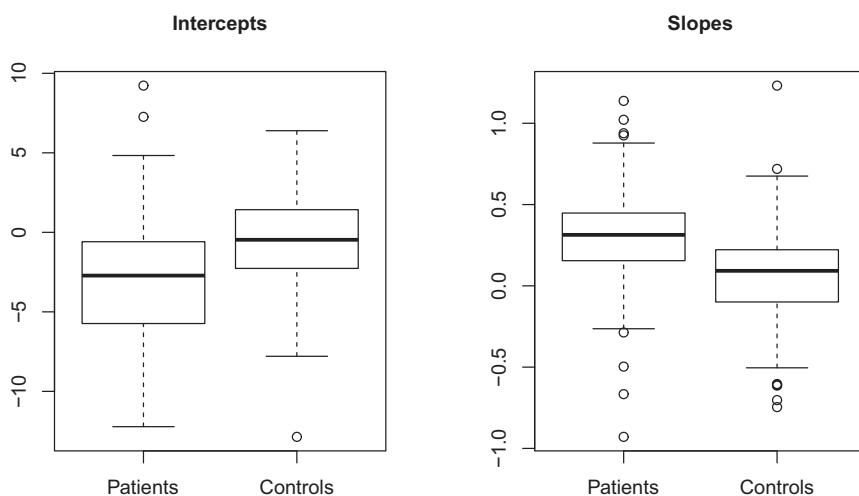


Figure 23.9 Coefficients for the within-subject regressions of log exercise on age, for patients and control subjects: intercepts (left) and slopes (right).

where group is a dummy variable coded 1 for subjects and 0 for controls. Substituting the Level 2 model into the Level 1 model produces the combined model

$$\begin{aligned}\text{log-exercise}_{ij} &= (\gamma_{00} + \gamma_{01}\text{group}_i + \omega_{0i}) + (\gamma_{10} + \gamma_{11}\text{group}_i + \omega_{1i})(\text{age}_{ij} - 8) + \varepsilon_{ij} \\ &= \gamma_{00} + \gamma_{01}\text{group}_i + \gamma_{10}(\text{age}_{ij} - 8) \\ &\quad + \gamma_{11}\text{group}_i \times (\text{age}_{ij} - 8) + \omega_{0i} + \omega_{1i}(\text{age}_{ij} - 8) + \varepsilon_{ij}\end{aligned}$$

that is, the Laird-Ware model

$$Y_{ij} = \beta_1 + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \delta_{1i} + \delta_{2i} Z_{2ij} + \varepsilon_{ij} \quad (23.11)$$

Fitting this model to the data by REML produces the following estimates of the fixed effects and variance-covariance components:

Parameter	Term	REML Estimate	Std. Error
β_1	intercept	-0.2760	0.1824
β_2	group _i	-0.3540	0.2353
β_3	age _{ij} - 8	0.0640	0.0314
β_4	group _i × (age _{ij} - 8)	0.2399	0.0394
ψ_1	SD(intercept)	1.4435	
ψ_2	SD(age _{ij} - 8)	0.1648	
ψ_{12}	C(intercept, age _{ij} - 8)	-0.0668	
σ_ε	SD(ε_{ij})	1.2441	

Letting Model 1 represent Equation 23.11, let us test whether random intercepts or random slopes can be omitted from the model:²⁸

Model	Omitting	REML log _e L
1	—	-1807.07
2	ψ_1^2, ψ_{12} (random intercepts)	-1911.04
3	ψ_2^2, ψ_{12} (random slopes)	-1816.13

Both likelihood-ratio tests are highly statistically significant (particularly the one for random intercepts), suggesting that both random intercepts and random slopes are required:

For $H_0: \psi_1^2 = 0, \psi_{12} = 0, G_0^2 = 2[-1807.07 - (-1911.04)] = 207.94, df = 2, p \approx 0$

For $H_0: \psi_2^2 = 0, \psi_{12} = 0, G_0^2 = 2[-1807.07 - (-1816.13)] = 18.12, df = 2, p = .0001$

The model that I have fit to the Davis et al. data (Equation 23.11) assumes independent Level 1 errors, ε_{ij} . The *composite errors*, $\zeta_{ij} = \delta_{1i} + \delta_{2i} Z_{2ij} + \varepsilon_{ij}$, are *correlated* within individuals, however, as I previously established for mixed models applied to hierarchical data. In the

²⁸As in the preceding section, because testing that variance components are 0 places parameters on the boundary of the parameter space, the *p*-values reported here are not quite correct. See Section 23.6 for more careful tests of variance components.

current context, Z_{2ij} is the time of observation (i.e., age minus 8 years), and the variance and covariances of the composite errors are (from Equations 23.6 and 23.7 on page 713)

$$\begin{aligned} V(\zeta_{ij}) &= \psi_1^2 + Z_{2ij}^2 \psi_2^2 + 2Z_{2ij}\psi_{12} + \sigma_\varepsilon^2 \\ C(\zeta_{ij}, \zeta_{ij'}) &= \psi_1^2 + Z_{2ij}Z_{2ij'}\psi_2^2 + (Z_{2ij} + Z_{2ij'})\psi_{12} \end{aligned}$$

The observations are not taken at entirely regular intervals (in that age at the last observation varies and is given to the nearest day), but assume that we have observations for subject i taken at $Z_{2i1} = 0$, $Z_{2i2} = 2$, $Z_{2i3} = 4$, and $Z_{2i4} = 6$ (i.e., at 8, 10, 12, and 14 years of age). Then the estimated covariance matrix for the composite errors is

$$\widehat{V}(\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) = \begin{bmatrix} 3.631 & 1.950 & 1.816 & 1.683 \\ 1.950 & 3.473 & 1.900 & 1.875 \\ 1.816 & 1.900 & 3.532 & 2.068 \\ 1.683 & 1.875 & 2.068 & 3.808 \end{bmatrix}$$

and the correlations of the composite errors are

$$\widehat{\text{Cor}}(\zeta_{i1}, \zeta_{i2}, \zeta_{i3}, \zeta_{i4}) = \begin{bmatrix} 1.0 & .549 & .507 & .453 \\ .549 & 1.0 & .543 & .516 \\ .507 & .543 & 1.0 & .564 \\ .453 & .516 & .564 & 1.0 \end{bmatrix}$$

The correlations across composite errors are moderately high, and the pattern is what we would expect: The correlations decline with the time separation between occasions. This pattern of declining correlations, however, does not *have* to hold: The correlations depend on the values of the estimated ψ s.

Recall that the linear mixed model allows for correlated Level 1 errors ε_{ij} within individuals. From Equations 23.1 (page 702),

$$\begin{aligned} V(\varepsilon_{ij}) &= \sigma_\varepsilon^2 \lambda_{ijj} \\ C(\varepsilon_{ij}, \varepsilon_{ij'}) &= \sigma_\varepsilon^2 \lambda_{ijj'} \end{aligned}$$

For a model with correlated individual-level errors to be estimable, however, the $\lambda_{ijj'}$ cannot consist of independent parameters; instead, these values are expressed in terms of a much smaller number of fundamental parameters.

For example, for equally spaced occasions, a very common model for the intraindividual errors is the *first-order autoregressive [or AR(1)] process*:²⁹

$$\varepsilon_{ij} = \phi \varepsilon_{i,j-1} + v_{ij}$$

where $v_{ij} \sim N(0, \sigma_v^2)$, $|\phi| < 1$, and v_{ij} and $v_{ij'}$ are independent for $j \neq j'$. Then the *autocorrelation* between two errors for the same individual one time period apart (i.e., at lag 1) is $\rho(1) = \phi$, and the autocorrelation between two errors s time periods apart (at lag s) is $\rho(s) = \phi^{|s|}$.

The occasions of measurement for the Davis et al. data are not all equally spaced, however. For data such as these, a frequently useful model is the *continuous first-order autoregressive process*, with the property that

²⁹The AR(1) process is introduced in the context of time-series regression in Section 16.2.1.

$$\text{Cor}(\varepsilon_{it}, \varepsilon_{i,t+s}) = \rho(s) = \phi^{|s|}$$

and where the time interval between observations, s , need not be an integer.

In modeling longitudinal data, it is often sensible to allow for serial correlation in the errors. A common error-generating process for equally spaced observations is the first-order autoregressive process, AR(1), $\varepsilon_{ij} = \phi \varepsilon_{i,j-1} + v_{ij}$, where $v_{ij} \sim N(0, \sigma_v^2)$, $|\phi| < 1$, and v_{ij} and $v_{ij'}$ are independent for $j \neq j'$. In the AR(1) process, the autocorrelation between two errors s time periods apart is $\rho(s) = \phi^{|s|}$. When the errors are unequally spaced, we may instead specify a continuous first-order autoregressive process, for which similarly $\text{Cor}(\varepsilon_{it}, \varepsilon_{i,t+s}) = \rho(s) = \phi^{|s|}$, but where the time interval between observations, s , need not be an integer.

I tried to fit the same mixed-effects model to the data as before (Equation 23.11), except allowing for first-order autoregressive Level 1 errors. The estimation process did not converge, however, and a close inspection suggests that the model has redundant parameters.³⁰ I then fit two additional models, retaining autocorrelated within-subject errors but omitting in turn random slopes and random intercepts. These models are not nested, so they cannot be compared via likelihood-ratio tests, but we can still compare the fit of the models to the data:

<i>Model</i>	<i>Description</i>	<i>REML Log-Likelihood</i>	<i>df</i>	<i>BIC</i>	<i>AIC</i>
1	Independent within-subject errors, random intercepts and slopes	-1807.1	8	3668.9	3630.1
2	Correlated within-subject errors, random intercepts	-1795.5	7	3638.9	360.0
3	Correlated within-subject errors, random slopes	-1802.3	7	3653.8	3619.8

Thus, the random-intercept model with autocorrelated within-subject errors (Model 2) produces the best fit to the data, according to both the AIC and BIC.³¹ Trading off parameters for the dependence of the within-subject errors against random effects is a common pattern: In this case, all three models produce similar estimates of the fixed effects.³²

Estimates for a final model, incorporating random intercepts and autocorrelated errors, are as follows:

³⁰Overly elaborate random-effects models often produce convergence problems because of effectively redundant parameters. In the current example, as I showed, the mixed model with independent intraindividual errors produces composite errors with a reasonable pattern of declining correlation over time, roughly similar to an AR(1) process.

³¹The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were introduced in Section 22.1.1.

³²It is often happily the case that estimates of fixed effects are relatively insensitive to the details of specification of variance and covariance components.

Parameter	Term	REML Estimate	Std. Error
β_1	intercept	-0.3070	0.1895
β_2	group _i	-0.2838	0.2447
β_3	age _{ij} - 8	0.0728	0.0317
β_4	group _i × (age _{ij} - 8)	0.2274	0.0397
ψ_1	SD(intercept)	1.1497	
σ_ϵ	SD(ε_{ij})	1.5288	
ϕ	error autocorrelation at lag1	0.6312	

Comparing the coefficients to their standard errors, the fixed-effect slope for the control group ($\hat{\beta}_3$)—that is, the average slope for individuals in this group—is statistically significant, and the difference in slopes between the patient group and the controls ($\hat{\beta}_4$) is highly statistically significant. In contrast, the initial difference between the groups (i.e., $\hat{\beta}_2$, the estimated difference at age 8) is nonsignificant.³³ An effect plot for the fixed effects, translating back from log exercise to the exercise scale (and subtracting the 5 minutes that were added prior to taking logs), appears in Figure 23.10.

23.5 Wald Tests for Fixed Effects

As I have explained, it is inappropriate to perform likelihood-ratio tests for fixed effects when the model is fit by REML. It is sometimes recommended that ML be used in place of REML to facilitate tests of fixed effects, but when there are relatively few Level 2 units, ML estimates can be substantially biased.

An alternative is to perform Wald tests of fixed effects, which do not require fitting and contrasting alternative models. For an individual coefficient, for example, we can compute the test statistic $Z_{0k} = \hat{\beta}_k / \text{SE}(\hat{\beta}_k)$ for the null hypothesis $H_0: \beta_k = 0$, referring the obtained value of Z_{0k} to the standard normal distribution. I performed such tests implicitly when I contrasted estimated coefficients to their standard errors in the examples developed above: Coefficients that exceeded twice their standard errors in magnitude were deemed “statistically significant” at the .05 level for a two-sided test. When there are relatively few Level 2 units, however, the distribution of the Wald statistic Z_{0k} may be sufficiently far from normal to render the resulting p -value inaccurate. Similarly, more complex Wald chi-square tests on several degrees of freedom may also yield inaccurate p -values.³⁴

We could attempt a straightforward substitution of Wald t and F tests for normal and chi-square tests, as we did for the linear model,³⁵ but these tests, naively constructed, may run up against two difficulties: (1) Straightforwardly computed coefficient standard errors can have substantial downward bias, and (2) straightforwardly computed denominator degrees of freedom are too large. The first problem can be addressed by a method suggested by Kenward and Roger (1997) for bias-corrected standard errors, while the second problem can be addressed by

³³The next section examines Wald tests for fixed-effects coefficients more carefully.

³⁴See Section 23.9.2 for Wald tests of general linear hypotheses in the LMM.

³⁵See Section 9.4.

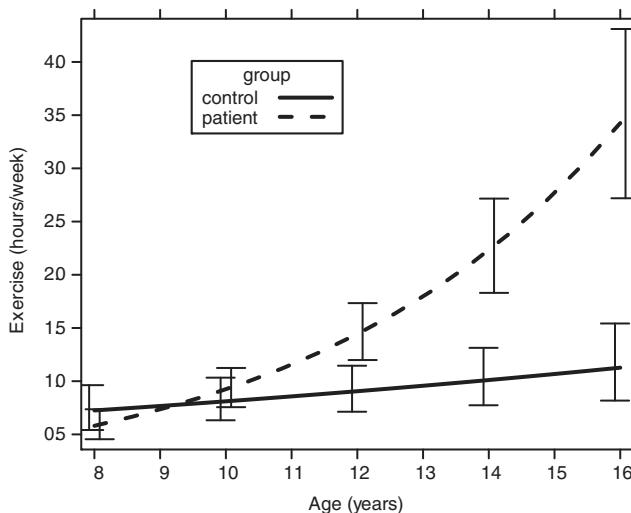


Figure 23.10 Fitted exercise as a function of age and group, with average trajectories based on the fixed effects. The vertical bars show approximate pointwise 95% confidence intervals around the fits at selected ages. The bars are displaced slightly horizontally to avoid overplotting.

applying Satterthwaite's (1946) method for determining approximate degrees of freedom. These solutions are commonly implemented in combination in computer software for mixed-effects models.³⁶

In the case of the mixed-effects model fit to the High School and Beyond data,³⁷ for example, the Kenward and Roger adjusted coefficient standard errors are virtually identical to the naive standard errors, and degrees of freedom for Wald t statistics computed for all coefficients are sufficiently large (at worst 154, approximately equal to the number of schools, 160) that statistical inferences are unaffected. This outcome is typical of applications that have a large number of Level 2 units.

When there are relatively few Level 2 units, naively computed Wald t and F tests, confidence intervals, and confidence regions for fixed-effects coefficients estimated by REML can be inaccurate. Inference based on Wald statistics can be rendered more accurate by employing the Kenward-Roger adjustment to coefficient standard errors and Satterthwaite degrees of freedom.

³⁶ Yet another, more computationally intensive, approach to statistical inference for both fixed effects and variance-covariance components is to use bootstrapping (introduced in Chapter 21).

³⁷ See Section 23.3; for the linear mixed model fit to Davis et al.'s longitudinal data, see the data analysis exercises for the chapter.

23.6 Likelihood-Ratio Tests of Variance and Covariance Components

I remarked in passing that likelihood-ratio tests for variance and covariance parameters should take account of the fact that the null hypothesis for a variance parameter specifies a value (i.e., 0) at the boundary of the parameter space.³⁸ Suppose that the null hypothesis is simply $H_0: \psi_1^2 = 0$, that is, that the random effects δ_{1i} are all 0 in a model that includes just this one random-effect term (in addition to the errors ε_{ij}). We cannot observe a negative estimate of ψ_1^2 , and so the usual p -value obtained from $\Pr(\chi_1^2 > G_0^2)$ must be halved, where G_0^2 is the likelihood-ratio test statistic for the hypothesis. In the rare circumstance that $\hat{\psi}_1^2 = 0$, and thus $G_0^2 = 0$, we take p -value = 1.

The other common situation is one in which removing a random-effect term δ_{ki} from the model removes not only the corresponding variance component ψ_k^2 but all variance/covariance parameters $\psi_{kk'}$, $k' = 1, \dots, q$, where, recall, q is the number of random effects in the model. The p -value for the hypothesis can then be computed as

$$p = \frac{\Pr(\chi_q^2 > G_0^2) + \Pr(\chi_{q-1}^2 > G_0^2)}{2}$$

For the High School and Beyond data, for example, I entertained the null hypothesis $H_0: \psi_2^2 = \psi_{12} = 0$, produced by removing the random effects δ_{2i} from the model but retaining δ_{1i} . Here $q = 2$, and I computed $G_0^2 = 9.76$; thus,

$$p = \frac{\Pr(\chi_2^2 > 9.76) + \Pr(\chi_1^2 > 9.76)}{2} = \frac{.00760 + .00178}{2} = .0047$$

which is smaller than the p -value that we would report (.0076) using χ_2^2 .

When the LMM is estimated by REML, it is inappropriate to use likelihood-ratio tests that compare models that differ in their fixed effects, even when the fixed effects for the two models are nested. We can, however, perform likelihood-ratio tests for variance and covariance components, as long as we are careful to take account of the fact that the null value of 0 for a variance parameter is on the boundary of the parameter space. If we delete one of q random effects from the model, that removes a variance component and $q - 1$ covariance components. The p -value for the resulting likelihood-ratio test statistic is computed as $p = [\Pr(\chi_q^2 > G_0^2) + \Pr(\chi_{q-1}^2 > G_0^2)]/2$.

³⁸See footnotes 17 (page 714), 19 (page 715), and 28 (page 721). I invite the reader to recompute p -values for the likelihood-ratio tests of variance-covariance reported earlier in this chapter.

23.7 Centering Explanatory Variables, Contextual Effects, and Fixed-Effects Models

I will focus on the *random-intercept regression model* with a single quantitative explanatory variable X . The conclusions that we draw from this simple setting are, however, more general.³⁹

Let us initially consider the following models:

1. X centered at the group means:

$$Y_{ij} = \beta_1^{(1)} + \delta_{1i}^{(1)} + \beta_2^{(1)}(X_{ij} - \bar{X}_{i\cdot}) + \varepsilon_{ij}$$

2. X uncentered:

$$Y_{ij} = \beta_1^{(2)} + \delta_{1i}^{(2)} + \beta_2^{(2)}X_{ij} + \varepsilon_{ij}$$

3. X centered at its overall mean:

$$Y_{ij} = \beta_1^{(3)} + \delta_{1i}^{(3)} + \beta_2^{(3)}(X_{ij} - \bar{X}) + \varepsilon_{ij}$$

4. X centered at the mean of the group means:

$$Y_{ij} = \beta_1^{(4)} + \delta_{1i}^{(4)} + \beta_2^{(4)}(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$

where

$$\bar{X}_{i\cdot} \equiv \frac{\sum_j X_{ij}}{n_i}$$

is the mean of X in group i ,

$$\bar{X} \equiv \frac{\sum_i \sum_j X_{ij}}{n} = \frac{\sum_i n_i \bar{X}_{i\cdot}}{n}$$

is the overall mean of X , and

$$\bar{X}_{..} = \frac{\sum_i \bar{X}_{i\cdot}}{m}$$

is the mean of the group means. Each of these models has two variance components: $V(\delta_{1i}) = \psi_1^2$ and $V(\varepsilon_{ij}) = \sigma_\varepsilon^2$. To make the models less abstract, we can appeal to the example in Section 23.3, imagining that the observations are for students grouped by school, that Y_{ij} is the math achievement score for the j th student in the i th school, and that X_{ij} is the student's family socioeconomic status.⁴⁰

The last three of these models define fixed origins for X . In Model 2, X is uncentered; in Model 3, X is centered at the overall mean for all students; and in Model 4, it is centered at the mean of the school means, which generally differs from the overall mean when there are unequal numbers of students in the various schools. These three models are observationally

³⁹As I explain below, the interpretation of the parameters of these models depends on the fact that the expectation of the random effects for the intercept is 0. This would be true for random slopes as well—the expected deviation from the fixed-effect slope parameter is 0.

⁴⁰Raudenbush and Bryk (2012, chap. 5) present a more extensive discussion of these issues using the High School and Beyond Data as an example.

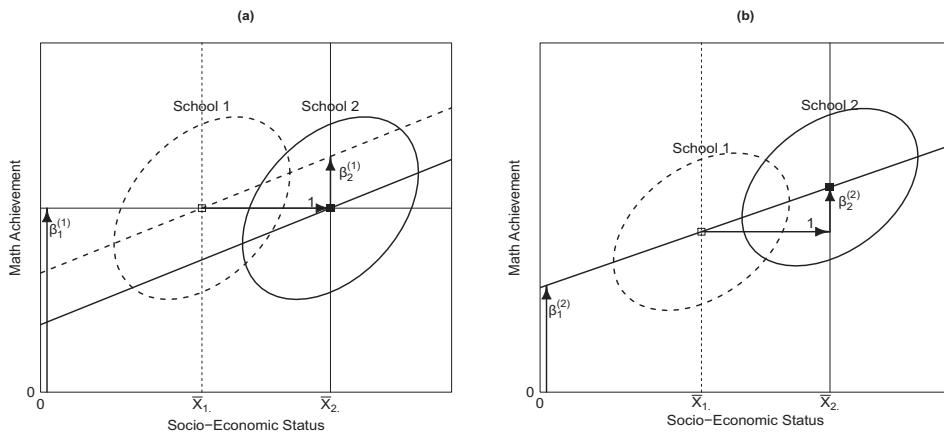


Figure 23.11 Effect of centering on the random-intercept regression model. In panel (a), socio-economic status X is centered at the group means (Model 1), while in panel (b), X is uncentered (Model 2). The expected within-school population data are shown as ellipses for two schools that differ by 1 in their average X -values.

equivalent: They produce the same fit to the data (e.g., the same SES slope, the same maximized likelihood, and the same fitted values). The models are simply reparameterizations of one another: $\beta_2^{(2)} = \beta_2^{(3)} = \beta_2^{(4)}$ and, for example, $\beta_1^{(2)} = \beta_1^{(3)} - \beta_2^{(3)}\bar{X}$. This result is likely unsurprising, because we are accustomed to models being fundamentally invariant with respect to the origin of an explanatory variable. Because the three models are equivalent, I will limit further consideration to Model 2, where X is uncentered.

In contrast, the first model, with X centered differently in each group at the group mean, is *not* observationally equivalent to the other models. I expect that this result is surprising.⁴¹ Consider, however, the different meaning of the parameters in Models 1 and 2: In Model 1, the expected advantage (or disadvantage) of student j in school i over student j' in school i' depends on the students' location relative to their school means and is $\beta_2^{(1)}[(X_{ij} - \bar{X}_i) - (X_{i'j'} - \bar{X}_{i'})]$. In Model 2, in contrast, the expected advantage of student j in school i over student j' in school i' depends on the difference in the two students' family SES and is $\beta_2^{(2)}(X_{ij} - X_{i'j'})$.⁴² Unless the two schools have the same mean SES, $\bar{X}_i = \bar{X}_{i'}$, the models have distinct implications for the expected difference in math achievement between the students.

This result is illustrated in Figure 23.11, for two schools, 1 and 2, differing by 1 in their mean X -values; the graphs are drawn setting the random effects equal to their expectation, $\delta_{11} = \delta_{12} = 0$. Figure 23.11(a) illustrates Model 1, with the X -values centered at the school means, while Figure 23.11(b) illustrates Model 2, with X uncentered. The population-data

⁴¹I know that the result was surprising to me when I first encountered it. I also recall a discussion with one of the early developers of generalized linear mixed models in which he too found the lack of invariance with respect to within-group centering disquieting.

⁴²The difference in scores between the students will also depend on the random effects δ_{1i} and $\delta_{1i'}$ for the two schools and on the errors ε_{ij} and $\varepsilon_{i'j'}$ for the two students, but as mentioned, the *expectations* of the random effects and errors are 0.

ellipses in these graphs represent the *expected* data in each school.⁴³ Were we to take account of the random effects (i.e., if δ_{11} and δ_{12} were nonzero), the data ellipses would each be randomly displaced vertically, but the two models would still not be equivalent.

The situation changes fundamentally if we add the compositional variable $\bar{X}_{i\cdot}$ to each model, obtaining

1. X centered at the group means:

$$Y_{ij} = \beta_1^{(1)} + \delta_{1i}^{(1)} + \beta_2^{(1)}(X_{ij} - \bar{X}_{i\cdot}) + \beta_3^{(1)}\bar{X}_{i\cdot} + \varepsilon_{ij}$$

2. X uncentered:

$$Y_{ij} = \beta_1^{(2)} + \delta_{1i}^{(2)} + \beta_2^{(2)}X_{ij} + \beta_3^{(2)}\bar{X}_{i\cdot} + \varepsilon_{ij}$$

3. X centered at its overall mean:

$$Y_{ij} = \beta_1^{(3)} + \delta_{1i}^{(3)} + \beta_2^{(3)}(X_{ij} - \bar{X}) + \beta_3^{(3)}\bar{X}_{i\cdot} + \varepsilon_{ij}$$

4. X centered at the mean of the group means:

$$Y_{ij} = \beta_1^{(4)} + \delta_{1i}^{(4)} + \beta_2^{(4)}(X_{ij} - \bar{X}_{..}) + \beta_3^{(4)}\bar{X}_{i\cdot} + \varepsilon_{ij}$$

Now all four models are equivalent, including Model 1 with X centered at the group means. The meaning of the coefficient β_3 of the compositional variable $\bar{X}_{i\cdot}$ varies, however; this coefficient is identical for Models 2, 3, and 4 but takes on a different value in Model 1. Again, I will compare Models 1 and 2. In Model 1, the expected advantage (or disadvantage) of student j in school i over student j' in school i' is

$$\begin{aligned} & \beta_2^{(1)}[(X_{ij} - \bar{X}_{i\cdot}) - (X_{i'j'} - \bar{X}_{i'\cdot})] + \beta_3^{(1)}(\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}) \\ &= \beta_2^{(1)}(X_{ij} - X_{i'j'}) + (\beta_3^{(1)} - \beta_2^{(1)})(\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}) \end{aligned} \quad (23.12)$$

while in Model 2, this expected advantage is

$$\beta_2^{(2)}(X_{ij} - X_{i'j'}) + \beta_3^{(2)}(\bar{X}_{i\cdot} - \bar{X}_{i'\cdot}) \quad (23.13)$$

Thus, $\beta_2^{(2)} = \beta_2^{(1)}$ and $\beta_3^{(2)} = \beta_3^{(1)} - \beta_2^{(1)}$. Equations 23.12 and 23.13 suggest that $\beta_3^{(2)}$ (not $\beta_3^{(1)}$) is interpretable as the *contextual effect* of the compositional variable school SES on individual students' math achievement—that is, the expected advantage in math achievement of a student whose school is 1 unit of SES higher than that of another student at the same level of individual SES. This situation is illustrated in Figure 23.12. The coefficient $\beta_3^{(1)}$ of $\bar{X}_{i\cdot}$ in Model 1 is the *sum* of the contextual effect of $\bar{X}_{i\cdot}$ and the individual-level effect of X . At the risk of some terminological confusion, we might term $\beta_3^{(1)} = \beta_2^{(2)} + \beta_3^{(2)}$ the *compositional effect* of X : That is, $\beta_3^{(1)}$ represents the difference in means $\bar{Y}_2 - \bar{Y}_1$ between the two schools, which reflects not only the effect of the compositional variable $\bar{X}_{i\cdot}$ for students of equal family SES (i.e., $\beta_2^{(2)}$) but also the fact that students in School 2 are on average 1 unit higher in individual SES than those in School 1 (i.e., $\beta_3^{(2)}$).

⁴³In Section 9.4.4, I used the data ellipse to visualize the *sample* standard deviations and covariance for two variables, as well as the least-squares regression of one variable on the other. Here, the same representation is applied to the *population* standard deviations and covariance. As in sample data, the population regression line goes through the points of vertical tangency to the ellipse.

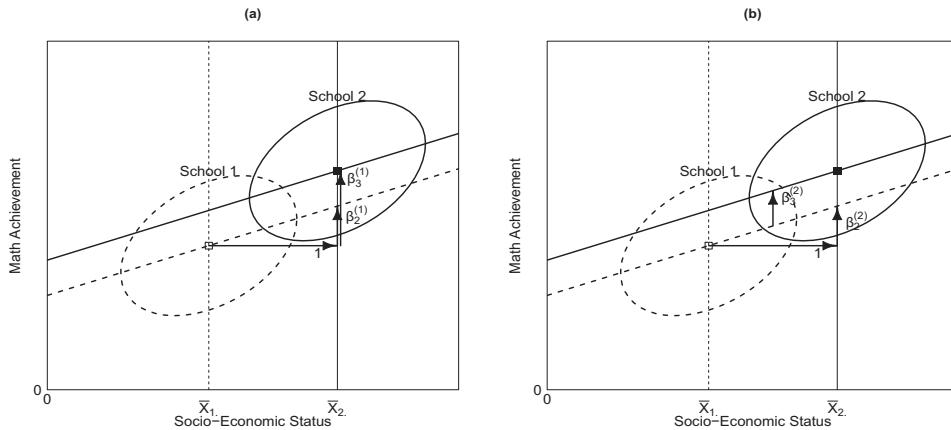


Figure 23.12 When the compositional explanatory variable $\bar{X}_{i\cdot}$ is included in the model, centering X at the group means does not fundamentally alter the model, but the meaning of the parameter β_3 associated with $\bar{X}_{i\cdot}$ changes. These graphs are for two schools that differ by 1 in their X (SES) means. The model with X centered at the school means (Model 1) is represented in panel (a), while the model with uncentered X (Model 2) is represented in panel (b).

23.7.1 Fixed Versus Random Effects

It is instructive to compare the mixed-effects models that we have just considered to analogous *fixed-effects models* in which the intercept can vary systematically across groups but in which the only random effect is the individual-level error:

1. X centered at the group means:

$$Y_{ij} = \beta_{1i}^{(1)} + \beta_2^{(1)}(X_{ij} - \bar{X}_{i\cdot}) + \varepsilon_{ij}$$

2. X uncentered:

$$Y_{ij} = \beta_{1i}^{(2)} + \beta_2^{(2)}X_{ij} + \varepsilon_{ij}$$

3. X centered at its overall mean:

$$Y_{ij} = \beta_{1i}^{(3)} + \beta_2^{(3)}(X_{ij} - \bar{X}) + \varepsilon_{ij}$$

4. X centered at the mean of the group means:

$$Y_{ij} = \beta_{1i}^{(4)} + \beta_2^{(4)}(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$

These models are equivalent to dummy-variable regression or analysis-of-covariance models treating group as a factor,⁴⁴ and the models can be fit by ordinary least-squares regression.

⁴⁴See Chapter 7 and Section 8.4.

Moreover, without random intercepts, all four models are observationally equivalent. This conclusion continues to hold when we add varying fixed-effect slopes for the groups, for example,

$$Y_{ij} = \beta_{1i}^{(1)} + \beta_{2i}^{(1)}(X_{ij} - \bar{X}_{i\cdot}) + \varepsilon_{ij} \quad (23.14)$$

The resulting model is equivalent to a dummy regression or analysis of covariance with interactions.

Perhaps even more interesting is the inability of the fixed-effects model with different group intercepts to incorporate a term for the contextual variable, $\beta_3 \bar{X}_{i\cdot}$, because $\bar{X}_{i\cdot}$ is perfectly collinear with the regressors representing the intercepts. This conclusion is general: Fixed-effects models with a different intercept for each group cannot incorporate contextual or compositional variables because these variables are invariant within groups. Indeed, this property is sometimes touted as an *advantage* of the fixed-effects model, because the model implicitly controls for all within-group-invariant explanatory variables, including those that are not available—or even known—to the researcher.

Consider, however, the following, likely surprising, fact: The least-squares estimate of the coefficient $\beta_2^{(1)}$ in fixed-effects Model 1 (or the corresponding coefficients in Model 2, 3, or 4) is the same as the estimate of $\beta_2^{(5)}$ in the following model, in which the intercept $\beta_1^{(5)}$ is constant across groups, but in which the compositional variable $\bar{X}_{i\cdot}$ is added as a regressor:⁴⁵

$$5. \quad Y_{ij} = \beta_1^{(5)} + \beta_2^{(5)}(X_{ij} - \bar{X}_{i\cdot}) + \beta_3^{(5)}\bar{X}_{i\cdot} + \varepsilon_{ij}$$

Although Model 5 is not observationally equivalent to fixed-effects Models 1 through 4—Model 5 has many fewer parameters—if our object is to estimate the individual-level coefficient β_2 of X , controlling for all group-level differences, it suffices to control for the within-group means, $\bar{X}_{i\cdot}$.

The test that $\beta_3^{(5)}$, the coefficient of the compositional variable, is 0 is related to the so-called *Hausman test* (introduced by Hausman, 1978), which is often used to decide whether to fit a fixed-effects model with differing group intercepts or a mixed-effects model with random intercepts, typically in the context of longitudinal data. The Hausman test, however, is more general: It can be applied whenever we have two estimators of a coefficient, one of which is known to be consistent and the other of which *may* be consistent—and, if so, is more efficient—if more restrictive assumptions are correct. In the current context, the coefficient of interest is β_2 ; the more general consistent estimator of β_2 is obtained from the fixed-effects model

$$Y_{ij} = \beta_{1i}^{(f)} + \beta_2^{(f)}X_{ij} + \varepsilon_{ij}$$

and the possibly consistent and more efficient estimator is obtained from the mixed-effects model

$$Y_{ij} = \beta_1^{(m)} + \delta_{1i}^{(m)} + \beta_2^{(m)}X_{ij} + \varepsilon_{ij}$$

If there are omitted group-constant explanatory variables correlated with X , the estimator of β_2 from the mixed-effects model will be inconsistent, but the estimator from the fixed-effects model will still be consistent as a consequence of fitting a different intercept for each group. If, on the other hand, there are no such omitted variables, then both models provide consistent

⁴⁵For the proof of this fact, see Exercise 23.3.

estimators of β_2 , but by virtue of its many fewer parameters, the mixed-effects estimator will be more efficient. The Hausman test statistic is

$$Z_0^2 = \frac{(\widehat{\beta}_2^{(f)} - \widehat{\beta}_2^{(m)})^2}{\widehat{V}(\widehat{\beta}_2^{(f)}) - \widehat{V}(\widehat{\beta}_2^{(m)})}$$

where $\widehat{V}(\widehat{\beta}_2^{(f)})$ is the estimated sampling variance of the fixed-effects estimator and $\widehat{V}(\widehat{\beta}_2^{(m)})$ is the estimated sampling variance of the mixed-effects estimator; under the hypothesis that the mixed-effects model is correct, Z_0^2 is distributed as χ^2_1 .⁴⁶ Inclusion of the compositional variable \bar{X}_i in the model renders the Hausman test effectively irrelevant.

The choice between random and fixed effects should reflect our view of the process that generates the data. If groups are literally sampled from a larger population of groups—as, for example, schools might be sampled from a population of schools—then it is natural to construe group effects as random. Similarly, if it is possible in principle to repeat a study with a different selection of groups, then it arguably makes sense to treat group effects as random, even if the groups are not literally sampled. This, or the previous situation, is commonly the case for longitudinal data on individuals and may be the case for hierarchical data. Finally, even when we cannot in principle repeat the study with different groups—as, for example, when the groups are the nations of the world—it may make sense to treat group effects as random, if it does not strain credulity to conceptualize the group-level data that we observe as the result of a partly random process, which may, therefore, have turned out differently. This is implicitly the point of view we take when we construct a hierarchical model for variation in individual-level coefficients across higher-level units such as nations.

In a mixed-effects model, centering an explanatory variable at the cluster means produces a different fit to the data—and a model with different interpretation—than centering the variable at a fixed value or leaving it uncentered. An exception occurs when the models include the compositional explanatory variable computed from the cluster means, in which case the models are observationally equivalent. Including the compositional variable also renders irrelevant an advantage that is often attributed to fixed-effects models: that is, that fixed-effects models control for all explanatory variables that are invariant within clusters, including unobserved cluster-invariant variables (but at the expense of being incapable of estimating the contextual effects of these variables).

⁴⁶*More generally, we may be interested in obtaining consistent estimates of several regression coefficients. Let us collect these coefficients in a parameter vector β_1 , which therefore will contain a common subset of parameters that appear in both the fixed- and mixed-effects models. Let the estimates of these parameters be respectively $\widehat{\beta}_1^{(f)}$ and $\widehat{\beta}_1^{(m)}$, with estimated covariance matrices $\widehat{V}(\widehat{\beta}_1^{(f)})$ and $\widehat{V}(\widehat{\beta}_1^{(m)})$. The Hausman test statistic is then

$$Z_0^2 = (\widehat{\beta}_1^{(m)} - \widehat{\beta}_1^{(f)})' [\widehat{V}(\widehat{\beta}_1^{(f)}) - \widehat{V}(\widehat{\beta}_1^{(m)})]^{-1} (\widehat{\beta}_1^{(m)} - \widehat{\beta}_1^{(f)})$$

which, under the hypothesis that the mixed-effects model is correct, is distributed as chi-square with degrees of freedom equal to the number of coefficients in β_1 . In the event that the difference in coefficient covariance matrices is singular, we can use a generalized inverse of $[\widehat{V}(\widehat{\beta}_1^{(f)}) - \widehat{V}(\widehat{\beta}_1^{(m)})]$, and the degrees of freedom for the test are reduced to the rank of this matrix.

23.8 BLUPs

Most social science applications of mixed-effects models focus on estimates of the fixed effects and, possibly, the variance components. “Estimating” the random effects for each higher-level unit is less frequently of direct interest, but considering how one goes about this process sheds light on mixed-effects models more generally, and consequently, I will address the topic briefly in this section. Recall that “estimating” is placed in quotation marks (with which I will dispense below) because the random effects are not parameters.

If the variance and covariance components were known, we could compute estimates of the fixed effects and random effects by generalized least-squares (GLS) regression.⁴⁷ It is instructive to examine the simplest instance of an LMM, the random effects one-way ANOVA model (introduced in Section 23.3.2):

$$Y_{ij} = \beta_1 + \delta_{1i} + \varepsilon_{ij}$$

The variance components for this model are $V(\delta_{1i}) = \psi_1^2$ and $V(\varepsilon_{ij}) = \sigma_\varepsilon^2$. Because $E(\delta_{1i}) = E(\varepsilon_{ij}) = 0$, we can take the mean in each group, \bar{Y}_i , as an independent, unbiased estimator of β_1 , with variance σ_ε^2/n_i . Furthermore, any weighted average of the group means

$$\hat{\beta}_1 = \frac{\sum_{i=1}^m w_i \bar{Y}_i}{\sum_{i=1}^m w_i}$$

for nonnegative weights w_i , not all 0, provides an unbiased estimator of β_1 . The GLS estimator of β_1 uses weights inversely proportional to the variances of the group means, most directly, $w_i = n_i$, which amounts simply to $\hat{\beta}_1 = \bar{Y}$, the unweighted average of the Y_{ij} . The GLS weights minimize the sampling variance of $\hat{\beta}_1$, making the GLS estimator the best linear unbiased estimator (BLUE) of β_1 .⁴⁸

Consider next how we could go about estimating δ_{1i} . One estimate is based on the group mean, $\hat{\delta}_{1i}^{(1)} = \bar{Y}_i - \bar{Y}$, which has variance σ_ε^2/n_i if we condition on the estimated fixed effect $\hat{\beta}_1 = \bar{Y}$. Another estimate is simply the expected value of δ_{1i} , that is $\hat{\delta}_{1i}^{(2)} = 0$, which has variance ψ_1^2 . As above, combining these two estimates optimally to produce the best linear unbiased predictor (BLUP) of δ_{1i} entails weighting the separate estimates inversely to their variances, yielding

$$\hat{\delta}_{1i} = \frac{\bar{Y}_i - \bar{Y}}{1 + \frac{\sigma_\varepsilon^2}{n_i \psi_1^2}}$$

Because the denominator of $\hat{\delta}_{1i}$ exceeds 1, the BLUP will always be closer to 0 than is the direct estimate $\bar{Y}_i - \bar{Y}$, with the difference dependent on how large σ_ε^2/n_i (i.e., the variance of $\hat{\delta}_{1i}^{(1)}$) is relative to ψ_1^2 . Substituting estimates of σ_ε^2 and ψ_1^2 for the variance components produces the empirical best linear unbiased predictor (EBLUP) of δ_{1i} .

⁴⁷See Section 23.9.1.

⁴⁸See Exercise 23.2. This is a consequence of the generalization to GLS of the Gauss-Markov theorem, presented in Section 9.3.2.

The BLUP is sometimes called a “shrinkage” estimator because it shrinks the estimated effect of membership in the i th group, $\bar{Y}_i - \bar{Y}$, toward 0. More generally, BLUPs combine information from a particular group with information from other groups, weighting the various sources of information inversely to their variances; “shrinkage” more generally is therefore to the rest of the data, not necessarily toward 0. For example, the BLUP of the mean of the i th group, $\hat{\mu}_i = \bar{Y} + \hat{\delta}_{1i}$, “shrinks” \bar{Y}_i toward the general mean \bar{Y} .

23.9 Statistical Details*

23.9.1 The Laird-Ware Model in Matrix Form

Like the linear model,⁴⁹ the Laird-Ware mixed model (Equation 23.1 on page 702) can be represented much more compactly and simply in matrix form:

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\delta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\delta}_i &\sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi}) \\ \boldsymbol{\delta}_i, \boldsymbol{\delta}_{i'} &\text{ are independent for } i \neq i' \\ \boldsymbol{\varepsilon}_i &\sim \mathbf{N}_{n_i}(\mathbf{0}, \sigma_{\varepsilon}^2 \boldsymbol{\Lambda}_i) \\ \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'} &\text{ are independent for } i \neq i' \\ \boldsymbol{\varepsilon}_i, \boldsymbol{\delta}_{i'} &\text{ are independent for all } i, i' \text{ including } i = i'\end{aligned}\tag{23.15}$$

where

- \mathbf{y}_i is the $n_i \times 1$ response vector for observations in the i th of m groups;
- \mathbf{X}_i is the $n_i \times p$ model matrix for the fixed effects of observations in group i , of full column rank p ;
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients, common to all groups;
- \mathbf{Z}_i is the $n_i \times q$ model matrix for the random effects of observations in group i , of full column rank q ;
- $\boldsymbol{\delta}_i$ is the $q \times 1$ vector of random-effect coefficients for group i ;
- $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of errors for observations in group i ;
- $\boldsymbol{\Psi}$ is the $q \times q$ covariance matrix for the random effects; and
- $\sigma_{\varepsilon}^2 \boldsymbol{\Lambda}_i$ is the $n_i \times n_i$ covariance matrix for the errors in group i and is $\sigma_{\varepsilon}^2 \mathbf{I}_{n_i}$ if the within-group errors have constant variance and are independent of each other.

Estimating Linear Mixed Models

Let $n \equiv \sum_{i=1}^m n_i$ represent the total number of observations over all m groups. It is convenient to write the model for all n observations simultaneously as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}\tag{23.16}$$

where

⁴⁹See Chapter 9.

- $\mathbf{y}_{(n \times 1)} \equiv [\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_m]'$ stacks up the response vectors \mathbf{y}_i for the m groups in one long column vector;
- similarly, $\boldsymbol{\varepsilon}_{(n \times 1)} \equiv [\varepsilon'_1, \varepsilon'_2, \dots, \varepsilon'_m]'$ is the stacked-up error vector;
- $\boldsymbol{\beta}_{(p \times 1)}$ is, as before, the fixed-effects parameter vector, common to all m groups;
- the model matrix for the fixed effects is

$$\mathbf{X}_{(n \times p)} \equiv \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix}$$

and is of rank p ;

- the model matrix for the random effects is block-diagonal,

$$\mathbf{Z}_{(n \times mq)} \equiv \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_m \end{bmatrix}$$

of rank mq , and each component $\mathbf{0}$ matrix is of order $n_i \times q$;

- $\boldsymbol{\delta}_{(mq \times 1)} \equiv [\boldsymbol{\delta}'_1, \boldsymbol{\delta}'_2, \dots, \boldsymbol{\delta}'_m]'$ is the stacked-up vector of random effects;
- and let

$$\boldsymbol{\Psi}^*_{(mq \times mq)} \equiv \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Psi} \end{bmatrix}$$

Notice the different structures of the fixed-effects and random-effects model matrices (\mathbf{X} and \mathbf{Z} , respectively): Because the fixed effects are common to all groups, it suffices simply to stack up the model matrices for the groups into \mathbf{X} . In contrast, the block-diagonal structure of \mathbf{Z} ensures that the proper random effects $\boldsymbol{\delta}_i$ enter the model for each group. As well, because the covariance matrix of the random effects $\boldsymbol{\Psi}$ is the same for all groups, the diagonal blocks of $\boldsymbol{\Psi}^*$ are identical.

For the model in Equation 23.16, we have

$$E \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (23.17)$$

and, because the random effects $\boldsymbol{\delta}$ and the errors $\boldsymbol{\varepsilon}$ are independent of each other,

$$V \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}^* & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \Lambda \end{bmatrix} \quad (23.18)$$

where

$$\sigma_{\varepsilon}^2 \boldsymbol{\Lambda}_{(n \times n)} \equiv \sigma_{\varepsilon}^2 \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Lambda}_2 \end{bmatrix}$$

is the block-diagonal covariance matrix for the errors. In hierarchical data with observations sampled independently within groups and with constant error variance, $V(\varepsilon) = \sigma_{\varepsilon}^2 \mathbf{I}_n$. From Equations 23.17 and 23.18, the covariance matrix of the response is⁵⁰

$$\boldsymbol{\Theta}_{(n \times n)} \equiv V(\mathbf{y}) = \mathbf{Z}\boldsymbol{\Psi}^*\mathbf{Z}' + \sigma_{\varepsilon}^2 \boldsymbol{\Lambda}$$

Summarizing, $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Theta})$.

Suppose unrealistically, for the purpose of argument, that all of the variance and covariance parameters in $\boldsymbol{\Psi}$ and $\sigma_{\varepsilon}^2 \boldsymbol{\Lambda}$, and thus $\boldsymbol{\Theta}$, are known. Treating the random effects $\boldsymbol{\delta}$ as if they were parameters, we could obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ by generalized least squares (GLS), solving the estimating equations⁵¹

$$\begin{bmatrix} \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \sigma_{\varepsilon}^2 \boldsymbol{\Psi}^{*-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{y} \\ \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{y} \end{bmatrix} \quad (23.19)$$

and producing

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\boldsymbol{\Theta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Theta}^{-1}\mathbf{y} \\ \hat{\boldsymbol{\delta}} &= \boldsymbol{\Psi}^*\mathbf{Z}'\boldsymbol{\Theta}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \quad (23.20)$$

The maximum-likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\delta}}$ contains the BLUPs of the random effects. Pursuing the application of GLS, the covariance matrix of the fixed-effects estimates is

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

As mentioned, \mathbf{y} is multivariate normal with expectation $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Theta}$. Writing out its density in detail,⁵²

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Theta}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \boldsymbol{\Theta}}} \exp [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Theta}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \quad (23.21)$$

Suppose now, and once again unrealistically, that the variance and covariance components are unknown but that the fixed effects $\boldsymbol{\beta}$ are known. Collect the independent variance- and covariance-component parameters in a vector $\boldsymbol{\sigma}$. For example, the original hierarchical model that I entertained for the High School and Beyond data (Equation 23.10 on page 714) has variance and covariance components $\psi_1^2 = V(\delta_{1i})$, $\psi_2^2 = V(\delta_{2i})$, $\psi_{12} = C(\delta_{1i}, \delta_{2i})$, and $\sigma_{\varepsilon}^2 = V(\varepsilon_{ij})$, and thus $\boldsymbol{\sigma} \equiv [\psi_1^2, \psi_2^2, \psi_{12}, \sigma_{\varepsilon}^2]'$. The covariance matrix $\boldsymbol{\Theta}$ of \mathbf{y} is a function of $\boldsymbol{\sigma}$ [say, $\boldsymbol{\Theta}(\boldsymbol{\sigma})$], and the log-likelihood for these parameters is⁵³

⁵⁰See Exercise 23.4.

⁵¹See Section 16.1 for generalized least-squares estimation and Exercise 23.5 for the derivation of the LMM estimating equations and their solution, along with the coefficient covariance matrix for the fixed-effects estimates. The covariance matrix for the random effects also follows from these results but is too complicated for me to write it out explicitly here.

⁵²See online Appendix D on probability and estimation for a description of the multivariate-normal distribution.

⁵³See Exercise 23.6.

$$\log_e L(\sigma|\beta, \mathbf{y}) = -\frac{n}{2} \log_e(2\pi) - \frac{1}{2} \log_e[\det \Theta(\sigma)] - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)\Theta^{-1}(\sigma)(\mathbf{y} - \mathbf{X}\beta) \quad (23.22)$$

We could estimate σ by maximizing this log-likelihood.

These observations suggest the basis of a procedure for maximum-likelihood estimation of all parameters in the LMM, along with the EBLUPs of the random effects:

1. Start with preliminary estimates of the variance- and covariance-component parameters, $\hat{\sigma}_0$.
2. Use $\hat{\Theta}_0 = \Theta(\hat{\sigma}_0)$ to obtain corresponding estimates of the fixed effects, $\hat{\beta}_0$, according to the first line of Equations 23.20.
3. Use the preliminary estimates of the fixed effects $\hat{\beta}_0$ to estimate the variance and covariance components, maximizing Equation 23.22 and obtaining $\hat{\sigma}_1$.
4. Iterate (i.e., repeat) Steps 2 and 3 until the estimates of β and σ converge.

REML estimation of the linear mixed model is similar, except that the REML log-likelihood is

$$\begin{aligned} \log_e L_{\text{REML}}(\sigma|\beta, \mathbf{y}) &= -\frac{n-p}{2} \log_e(2\pi) - \frac{1}{2} \log_e[\det \Theta(\sigma)] \\ &\quad - \frac{1}{2} \log_e \{ \det [\mathbf{X}' \Theta^{-1}(\sigma) \mathbf{X}] \} - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \Theta^{-1}(\sigma) (\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

EBLUPs of the random effects can be computed by substituting estimates of the fixed effects and the variance and covariance components (obtained either by ML or by REML) into the second line of Equations 23.20.

23.9.2 Wald Tests Revisited

A linear hypothesis for the fixed effects in an LMM takes the following form:⁵⁴

$$H_0: \underset{(r \times p)}{\mathbf{L}} \underset{(p \times 1)}{\beta} = \underset{(r \times 1)}{\mathbf{0}}$$

where the hypothesis matrix \mathbf{L} is formulated so that it is of full row rank r . In the simplest case, \mathbf{L} consists of $r = 1$ row with one nonzero unit entry, such as $L = (0, 1, 0, \dots, 0)$, corresponding to the hypothesis $H_0: \beta_2 = 0$.

The estimated covariance matrix of the fixed-effects coefficients follows from the solution of the mixed-model estimating equations (given in Equations 23.20):

$$\hat{V}(\hat{\beta}) = \left(\mathbf{X}' \hat{\Theta}^{-1} \mathbf{X} \right)^{-1}$$

where, recall, $\Theta = \mathbf{Z}\Psi^*\mathbf{Z}' + \sigma_\varepsilon^2 \Lambda$ is the covariance matrix of the response vector \mathbf{y} . The estimate $\hat{\Theta}$, therefore, is computed from $\hat{\sigma}_e^2$, $\hat{\Psi}^*$, and $\hat{\Lambda}$, which, in turn, typically depend on a small number of estimated variance and covariance parameters. Using $\hat{V}(\hat{\beta})$, a naive Wald F -statistic for the hypothesis is

⁵⁴Cf. Section 9.4.3 for linear hypotheses in linear models.

$$F_0 = \frac{1}{r} \widehat{\beta} \mathbf{L}' \left[\mathbf{L} \widehat{V}(\widehat{\beta}) \mathbf{L}' \right]^{-1} \mathbf{L} \widehat{\beta}$$

with r and $n - p$ degrees of freedom. For example, for the simple hypothesis $H_0: \beta_2 = 0$, where $L = (0, 1, 0, \dots, 0)$,

$$F_0 = \frac{\widehat{\beta}_2^2}{\widehat{V}(\widehat{\beta}_2)}$$

with 1 and $n - p$ degrees of freedom, and thus

$$t_0 = \sqrt{F_0} = \frac{\widehat{\beta}_2}{\text{SE}(\widehat{\beta}_2)}$$

with $n - p$ degrees of freedom.

As mentioned,⁵⁵ however, these F - and t -test statistics run up against two problems: (1) The estimated covariance matrix $\widehat{V}(\widehat{\beta})$ can be substantially biased, tending to produce coefficient standard errors that are too small, and (2) the degrees of freedom $n - p$ do not take account of the dependencies among the observations and consequently are too large. Taken together, these problems imply that naive Wald tests and confidence intervals tend to exaggerate the statistical significance and precision of estimation of the fixed effects.

Kenward and Roger (1997) suggest a method to correct for the downward bias in the estimated coefficient variances by inflating the coefficient covariance matrix $\widehat{V}(\widehat{\beta})$. Satterthwaite's (1946) method can then be applied to the adjusted coefficient covariance matrix to get corrected denominator degrees of freedom for Wald t - and F -tests.⁵⁶

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 23.1. Using the estimated fixed effects in the table on page 717 for the model fit to the High School and Beyond data, find the fixed-effect regression equations for typical low, medium, and high mean SES Catholic and public schools, as plotted in Figure 23.6.

Exercise 23.2. *BLUPs: As discussed in Section 23.8, show that for the random-effects one-way ANOVA model, $Y_{ij} = \beta_1 + \delta_{1i} + \varepsilon_{ij}$, the weights $w_i = n_i$ minimize the variance of the estimator

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^m w_i \bar{Y}_i}{\sum_{i=1}^m w_i}$$

and thus that this choice provides the best linear unbiased estimator (BLUE) of β_1 . Then explain why

⁵⁵In Section 23.5.

⁵⁶The details are sufficiently complex that I will not present them here, but see, for example, Stroup (2013, Section 5.4).

$$\hat{\delta}_{1i} = \frac{\bar{Y}_i - \bar{Y}}{1 + \frac{\sigma_e^2}{n_i \bar{y}_i^2}}$$

is the best linear unbiased predictor (BLUP) of δ_{1i} .

Exercise 23.3. *Prove that the least-squares estimates of the coefficient β_2 for X_{ij} is the same in the following two fixed-effects models (numbered as in Section 23.7.1):

1. $Y_{ij} = \beta_{1i}^{(1)} + \beta_2^{(1)}(X_{ij} - \bar{X}_i) + \varepsilon_{ij}$
5. $Y_{ij} = \beta_1^{(5)} + \beta_2^{(5)}(X_{ij} - \bar{X}_i) + \beta_3^{(5)}\bar{X}_i + \varepsilon_{ij}$

Recall the context: The data are divided into groups $i = 1, \dots, m$, with individuals $j = 1, \dots, n_i$ in the i th group. The first model (Model 1) fits a different intercept $\beta_{1i}^{(1)}$ in each group, along with the common slope $\beta_2^{(1)}$. The second model (Model 5) fits a common intercept $\beta_1^{(5)}$ and common slope $\beta_2^{(5)}$ but controls for the compositional variable \bar{X}_i . (*Hint:* Consider the added-variable plot that determines the coefficient $\hat{\beta}_2$.⁵⁷ In Model 1, this added-variable plot is the scatterplot for residuals from the regressions of Y_{ij} and $X_{ij} - \bar{X}_i$ on a set of m dummy regressors for groups. In Model 5, the added-variable plot is the scatterplot for residuals from the regressions of Y_{ij} and $X_{ij} - \bar{X}_i$ on the compositional variable \bar{X}_i , and the intercept, but the compositional variable is itself the projection of X_{ij} onto the space spanned by the group dummy regressors—that is, the group means \bar{X}_i are of course perfectly determined by the groups.)

Exercise 23.4. *Using

$$V \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}^* & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{\Lambda} \end{bmatrix}$$

show that the covariance matrix of the response variable in the compact form of the LMM, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$, can be written as $V(\mathbf{y}) = \mathbf{Z}\boldsymbol{\Psi}^*\mathbf{Z}' + \sigma_e^2 \boldsymbol{\Lambda}$.⁵⁸

Exercise 23.5. *Derive the generalized least-squares estimating equations for the LMM (repeating Equation 23.19 from page 736),

$$\begin{bmatrix} \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{Z} + \sigma_e^2 \boldsymbol{\Psi}^{*-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{y} \\ \mathbf{Z}'\boldsymbol{\Lambda}^{-1}\mathbf{y} \end{bmatrix}$$

and show that the solution can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\boldsymbol{\Theta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Theta}^{-1}\mathbf{y} \\ \hat{\boldsymbol{\delta}} &= \boldsymbol{\Psi}^*\mathbf{Z}'\boldsymbol{\Theta}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

where $\boldsymbol{\Theta} \equiv V(\mathbf{y}) = \mathbf{Z}\boldsymbol{\Psi}^*\mathbf{Z}' + \sigma_e^2 \boldsymbol{\Lambda}$. Then show that $V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\boldsymbol{\Theta}^{-1}\mathbf{X})^{-1}$. *An even more-challenging exercise:* Find an explicit expression for the covariance matrix of the BLUPs $\hat{\boldsymbol{\delta}}$. (*Hint:* Invert the partitioned matrix on the left-hand side of the GLS estimating equations for the LMM in Equations 23.19.)

⁵⁷See Section 11.6.1 and Exercises 11.5 and 11.6 for information on added-variable plots.

⁵⁸See Section 23.9.1

Exercise 23.6. *Show that the log-likelihood for the variance-covariance-component parameters σ given the fixed effects β can be written as (repeating Equation 23.22 from page 737)

$$\log_e L(\sigma|\beta, \mathbf{y}) = -\frac{n}{2} \log_e(2\pi) - \frac{1}{2} \log_e[\det \Theta(\sigma)] - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)\Theta^{-1}(\sigma)(\mathbf{y} - \mathbf{X}\beta)$$

Summary

- Clustered data commonly arise in two contexts: hierarchical data, in which lower-level units, such as individual students, are nested within higher-level units, such as schools, and longitudinal data, in which individuals (or other multiple units of observation) are followed over time. In both cases, observations within a cluster—lower-level units within higher-level units or different measurement occasions for the same individual—cannot reasonably be treated as statistically independent. Mixed-effect models take account of dependencies in hierarchical, longitudinal, and other dependent data.
- The linear mixed-effects model (LMM) is applicable both to hierarchical and longitudinal data; in Laird-Ware form, the model is written

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 X_{2ij} + \cdots + \beta_p X_{pj} + \delta_{1i} Z_{1ij} + \cdots + \delta_{qi} Z_{qij} + \varepsilon_{ij} \\ \delta_{ki} &\sim N(0, \psi_k^2), C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'} \\ \delta_{ki}, \delta_{k'i} &\text{ are independent for } i \neq i' \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2 \lambda_{ij}), C(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_\varepsilon^2 \lambda_{ij'} \\ \varepsilon_{ij}, \varepsilon_{i'j'} &\text{ are independent for } i \neq i' \\ \delta_{ki}, \varepsilon_{i'j} &\text{ are independent for all } i, i', k, j \text{ (including } i = i') \end{aligned}$$

Here, Y_{ij} is the value of the response variable for the j th of n_i observations in the i th of m clusters, the β s are fixed-effect coefficients, the X s are fixed-effect regressors, the δ s are random-effect coefficients, the Z s are random-effect regressors, and the ε s are errors for individuals within clusters. The ψ s and λ s, which capture the dependencies among the random effects and errors within clusters, are typically expressed in terms of a small number of fundamental variance- and covariance-component parameters.

- In modeling hierarchical data, it is often natural to formulate an individual-level model within clusters and then to treat the coefficients of that model as random effects that appear as the responses in a higher-level model. The models at the two levels can be combined as an LMM in Laird-Ware form. Contextual variables describe higher-level units; compositional variables also describe higher-level units but are derived from lower-level units (e.g., by averaging).
- The coefficients-as-outcomes model relates regression coefficients of lower-level units within clusters to characteristics of the clusters. Simpler hierarchical models include the random-effects one-way ANOVA model, in which each cluster has its own mean, treated as a random effect, and the random-coefficients regression model, in which several regression coefficients can vary randomly across clusters. In the random-effects one-way ANOVA model, the intraclass correlation measures the proportion of individual-level variation that is due to differences among clusters, $\rho = \psi_1^2 / (\psi_1^2 + \sigma_\varepsilon^2)$, where ψ_1^2 is the variance component for clusters and σ_ε^2 is the error variance.

- LMMs can be estimated by maximum likelihood (ML) or by restricted maximum likelihood (REML). When the number of clusters is small, REML tends to produce less biased estimates of variance components.
- In modeling longitudinal data, it is often sensible to allow for serial correlation in the errors. A common error-generating process for equally spaced observations is the first-order autoregressive process, $\text{AR}(1)$, $\varepsilon_{ij} = \phi\varepsilon_{i,j-1} + v_{ij}$, where $v_{ij} \sim N(0, \sigma_v^2)$, $|\phi| < 1$, and v_{ij} and $v_{ij'}$ are independent for $j \neq j'$. In the AR(1) process, the autocorrelation between two errors s time periods apart is $\rho(s) = \phi^{|s|}$. When the errors are unequally spaced, we may instead specify a continuous first-order autoregressive process, for which similarly $\text{Cor}(\varepsilon_{it}, \varepsilon_{i,t+s}) = \rho(s) = \phi^{|s|}$, but where the time interval between observations, s , need not be an integer.
- When there are relatively few Level 2 units, naively computed Wald t - and F -tests, confidence intervals, and confidence regions for fixed-effects coefficients estimated by REML can be inaccurate. Inference based on Wald statistics can be rendered more accurate by employing the Kenward-Roger adjustment to coefficient standard errors and Satterthwaite degrees of freedom.
- When the LMM is estimated by REML, it is inappropriate to use likelihood-ratio tests that compare models that differ in their fixed effects, even when the fixed effects for the two models are nested. We can, however, perform likelihood-ratio tests for variance and covariance components, as long as we are careful to take account of the fact that the null value of 0 for a variance parameter is on the boundary of the parameter space. If we delete one of q random effects from the model, that removes a variance component and $q - 1$ covariance components. The p -value for the resulting likelihood-ratio test statistic is computed as $p = [\Pr(\chi_q^2 > G_0^2) + \Pr(\chi_{q-1}^2 > G_0^2)]/2$.
- In a mixed-effects model, centering an explanatory variable at the cluster means produces a different fit to the data—and a model with different interpretation—than centering the variable at a fixed value or leaving it uncentered. An exception occurs when the models include the compositional explanatory variable computed from the cluster means, in which case the models are observationally equivalent. Including the compositional variable also renders irrelevant an advantage that is often attributed to fixed-effects models: that is, that fixed-effects models control for all explanatory variables that are invariant within clusters, including unobserved cluster-invariant variables (but at the expense of being incapable of estimating the contextual effects of these variables).

Recommended Reading

There are many books on mixed-effects models, and the subject can certainly profit from a more extensive treatment than I have been able to give it in this (and the next) chapter. I find the following sources especially useful:

- Snijders and Bosker (2012) provide a highly accessible treatment of mixed-effects models, emphasizing hierarchical data and linear mixed models. Careful attention is paid to the practice of mixed-effects modeling, and there are numerous examples from the social sciences.

- Raudenbush and Bryk (2012) also emphasize hierarchical data but provide greater formal detail. They do so, however, in a manner that builds insight and supports intuitive understanding of the statistical results. Examples are drawn from the social sciences, with a focus on educational research.
- Gelman and Hill (2007) present multilevel models in the more general context of regression analysis. Their treatment carefully balances the theoretical underpinnings of statistical models with a sensitivity to the data and stresses Bayesian methods.

24

Generalized Linear and Nonlinear Mixed-Effects Models

The range of application of linear models is greatly expanded by considering non-normal conditional distributions of the response variable, producing the class of generalized linear models developed in Part IV of the text. The same is true of linear mixed models, and the first section of the current chapter introduces generalized linear mixed models for non-normal responses. Generalized nonlinear mixed models are useful, for example, for modeling clustered categorical responses and count data. The second section of the chapter takes up fundamentally nonlinear mixed-effects models for clustered data, generalizing the treatment of the topic in Chapter 17, where I developed nonlinear models for independent observations.

24.1 Generalized Linear Mixed Models

Recall that a generalized linear model consists of three components:¹

- a *random component*, specifying the conditional distribution—or simply the conditional mean and variance—of the response variable Y_i for observation i , given the regressors, $X_{i1}, X_{i2}, \dots, X_{ik}$; traditionally, but not necessarily, the random component is a member of an *exponential family* of distributions—the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families;
- a linear predictor,

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

on which the expected value μ_i of the response variable depends; and

- a *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor; the inverse of the link function is the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

The *generalized linear mixed-effects model (GLMM)* is a straightforward extension of the generalized linear model, adding random effects to the linear predictor and expressing the expected value of the response conditional on the random effects: The link function $g(\cdot)$ is the same as in generalized linear models. In the GLMM, the conditional distribution of Y_{ij} , the response for

¹See Chapter 15.

observation j in group i , given the random effects, is (most straightforwardly) a member of an exponential family, with mean μ_{ij} , variance

$$V(Y_{ij}) = \phi v(\mu_{ij})\lambda_{ij}$$

and covariances

$$C(Y_{ij}, Y_{ij'}) = \phi \sqrt{v(\mu_{ij})} \sqrt{v(\mu_{ij'})} \lambda_{ijj'}$$

where ϕ is a dispersion parameter and the function $v(\mu_{ij})$ depends on the distributional family to which Y belongs. Recall, for example, that in the binomial and Poisson families, the dispersion is fixed to 1, and that in the Gaussian family, $v(\mu) = 1$.

We will make the same assumptions about the random effects δ_{ki} in the GLMM that we made in the LMM: that they are multivariate normal with 0 means, that they may be correlated for a particular observation and may have unequal variances, but that they are independent across observations. That is, $\delta_{ki} \sim N(0, \psi_k^2)$, $C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'}$, and $\delta_{ki}, \delta_{k'i}$ are independent for $i \neq i'$.

The generalized linear mixed model is fit by maximum-likelihood estimation when the conditional distribution of the response is a member of an exponential family or quasi-likelihood when it is not. Although it is not difficult to write down the likelihood for the model,² the likelihood is difficult to maximize because it is necessary to integrate out (i.e., “sum over”) the unobservable random effects. Consequently, statistical software for GLMMs resorts to various approximations, including (in ascending order of general accuracy) penalized quasi-likelihood, the Laplace approximation, and Gauss-Hermite quadrature. Bayesian approaches are also commonly employed but are beyond the scope of this chapter, as are the details of the differences among the various approximate ML methods.³

The generalized linear mixed-effects model (GLMM) may be written as

$$\begin{aligned}\eta_{ij} &= \beta_1 + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij} + \delta_{1i} Z_{1ij} + \cdots + \delta_{qi} Z_{qij} \\ g(\mu_{ij}) &= E(Y_{ij} | \delta_{1i}, \dots, \delta_{qi}) = \eta_{ij} \\ \delta_{ki} &\sim N(0, \psi_k^2), C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'} \\ \delta_{ki}, \delta_{k'i} &\text{ are independent for } i \neq i' \\ V(Y_{ij}) &= \phi v(\mu_{ij})\lambda_{ij} \\ C(Y_{ij}, Y_{ij'}) &= \phi \sqrt{v(\mu_{ij})} \sqrt{v(\mu_{ij'})} \lambda_{ijj'} \\ Y_{ij}, Y_{ij'} &\text{ are independent for } i \neq i'\end{aligned}$$

where η_{ij} is the linear predictor for observation j in cluster i ; the fixed-effect coefficients (β s), random-effect coefficients (δ s), fixed-effect regressors (X s), and random-effect regressors (Z s) are defined as in the LMM; and the dispersion parameter ϕ and variance function $v(\cdot)$ depend on the distributional family to which Y belongs; alternatively, for quasi-likelihood estimation, $v(\cdot)$ can be given directly. The GLMM is estimated by approximate maximum likelihood.

²See Section 24.1.2.

³But see the references given at the end of the chapter.

24.1.1 Example: Migraine Headaches

In an effort to reduce the severity and frequency of migraine headaches through the use of biofeedback training, Tammy Kostecki-Dillon, a psychologist, collected longitudinal data on migraine headache sufferers.⁴ The 133 patients who participated in the study were each given four weekly sessions of biofeedback training. The patients were asked to keep daily logs of their headaches for a period of 30 days prior to training, during training, and after training, to 100 days after training began. Compliance with these instructions was low, and there is therefore quite a bit of missing data; for example, only 55 patients kept a log prior to training. On average, subjects recorded information on 31 days, with the number of days ranging from 7 to 121. Subjects were divided into three self-selected groups: those who discontinued their migraine medication during the training and posttraining phase of the study, those who continued their medication but at a reduced dose, and those who continued their medication at the previous dose.

I will use a binomial GLMM—specifically, a binary logit model—to analyze the incidence of headaches during the period of the study. Examination of the data suggested that the incidence of headaches was invariant during the pretraining phase of the study, increased (as was expected by the investigator) at the start of training, and then declined at a decreasing rate. I decided to fit a linear trend prior to the start of training (before time 0), possibly to capture a trend that I failed to detect in my exploration of the data and to transform time at time 0 and later (which, for simplicity, I term *time posttreatment*) by taking the square root.⁵ In addition to the intercept, representing the level of headache incidence at the end of the pretraining period, I included a dummy regressor coded 1 posttreatment and 0 pretreatment to capture the anticipated increase in headache incidence at the start of training, dummy regressors for levels of medication, and interactions between medication and treatment, as well as between medication and the pre- and posttreatment time trends. Thus, the fixed-effects part of the model is

$$\begin{aligned} \text{logit}(\pi_{ij}) = & \beta_1 + \beta_2 M_{1i} + \beta_3 M_{2i} + \beta_4 P_{ij} + \beta_5 T_{0ij} + \beta_6 \sqrt{T_{1ij}} \\ & + \beta_7 M_{1i}P_{ij} + \beta_8 M_{2i}P_{ij} + \beta_9 M_{1i}T_{0ij} + \beta_{10} M_{2i}T_{0ij} \\ & + \beta_{11} M_{1i} \sqrt{T_{1ij}} + \beta_{12} M_{2i} \sqrt{T_{1ij}} \end{aligned} \quad (24.1)$$

where

- π_{ij} is the probability of a headache for individual $i = 1, \dots, 133$, on occasion $j = 1, \dots, n_i$;
- M_{1i} is a dummy regressor coded 1 if individual i continued taking migraine medication at a reduced dose posttreatment, and M_{2i} is a dummy regressor coded 1 if individual i continued taking medication at its previous dose posttreatment;
- P_{ij} is a dummy regressor coded 1 posttreatment (i.e., after time 0) and 0 pretreatment;
- T_{0ij} is time (in days) pretreatment, running from -29 through 0, and coded 0 after treatment began;
- T_{1ij} is time (in days) posttreatment, running from 1 through 99, and coded 0 pretreatment.

⁴The data are described by Kostecki-Dillon, Monette, and Wong (1999) and were generously made available to me by Georges Monette. The data were also used in a different context by Gao (2007).

⁵The original analysis of the data by Georges Monette used regression splines for time trends, with results generally similar to those reported here: See Exercise 24.1.

I included patient random effects for the intercept (i.e., the level of headache incidence pretreatment), for the posttreatment dummy regressor, and for the pre- and posttreatment time-trend regressors.

Wald tests for the fixed effects reveal that all of the interactions are nonsignificant, along with the pretreatment trend, while the medication and treatment effects, along with the posttreatment trend, are highly statistically significant:⁶

Term	Wald Chi-square	df	p
Medication (M_1, M_2)	22.07	2	< .0001
Treatment (P)	16.09	1	< .0001
Pretreatment Trend (T_0)	0.35	1	.55
Posttreatment Trend ($\sqrt{T_1}$)	37.87	1	<< .0001
Medication \times Treatment	2.50	2	.29
Medication \times Pretreatment Trend	1.85	2	.40
Medication \times Posttreatment Trend	0.07	2	.97

Even without explicit temporal autocorrelation, the random effects are relatively complex for such a small data set, and it would be desirable to be able to simplify this part of the model. To this end, I dropped each random effect in turn and performed likelihood-ratio tests for the corresponding variance and covariance components; in each case, one variance and three covariance components are removed from the model, and *p*-values are computed using the approach described in Section 23.6:

Random Effect Removed	G^2	p
Intercept	19.70	.0004
Treatment	12.08	.012
Pretreatment Trend	5.79	.17
Posttreatment Trend	16.21	.0019

On the basis of these tests for the fixed and random effects, I specified a final model for the migraine data that eliminates the fixed-effect interactions with medication and the pretreatment trend fixed and random effects, obtaining the following estimates for the fixed effects and variance components. I number the fixed-effect parameters and corresponding variance components as in the original model (Equation 24.1), show the variance components as standard deviations, and suppress the covariance components:

⁶These tests are constructed conforming to the principle of marginality. For example, the test for medication is computed assuming that the interactions are nil. See Sections 7.3.5 and 8.2.5 for further discussion, in the context of linear models, of formulating hypothesis tests when terms in the model are related by marginality.

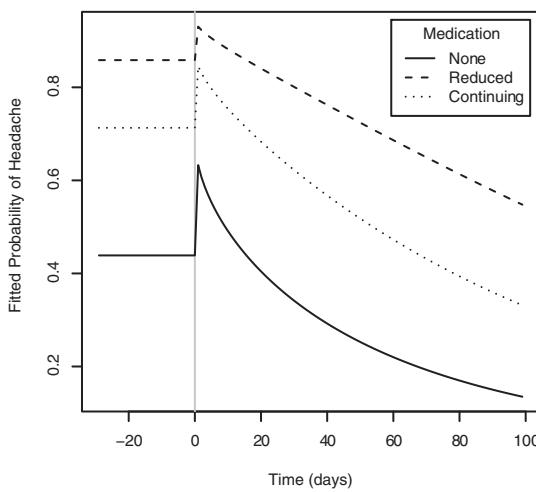


Figure 24.1 Fixed effects from a binomial GLMM fit to the migraine data. Treatment started at Time 1.

Term	Parameter	Estimate	Std. Error
Intercept	β_1	-0.246	0.344
	ψ_1	1.306	—
Medication (reduced)	β_2	2.049	0.467
(continuing)	β_3	1.156	0.383
Treatment	β_4	1.059	0.244
	ψ_2	1.313	—
Posttreatment Trend	β_6	-0.268	0.045
	ψ_4	0.239	—

Figure 24.1 shows the estimated fixed effects plotted on the probability scale; as a consequence, the posttreatment trends for the three medication conditions are not parallel, as they would be if plotted on the logit scale. It is apparent from this graph that after an initial increase at the start of treatment, the incidence of headaches declined to substantially below its pretreatment level. As well, the incidence of headaches was lowest among the patients who discontinued their medication and highest among those who reduced their medication; patients who continued their medication at pretraining levels were intermediate in headache incidence.⁷ Of course, self-selection of the medication groups renders interpretation of this pattern ambiguous.

⁷See Exercise 24.1 for the construction of this graph.

24.1.2 Statistical Details*

The Generalized Linear Mixed Model in Matrix Form

In matrix form, the GLMM is

$$\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\delta}_i \quad (24.2)$$

$$g(\boldsymbol{\mu}_i) = g[E(\mathbf{y}_i | \boldsymbol{\delta}_i)] = \boldsymbol{\eta}_i$$

$$\boldsymbol{\delta}_i \sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi})$$

$\boldsymbol{\delta}_i, \boldsymbol{\delta}_{i'}$ are independent for $i \neq i'$

$$E(\mathbf{y}_i | \boldsymbol{\delta}_i) = \boldsymbol{\mu}_i \quad (24.3)$$

$$V(\mathbf{y}_i | \boldsymbol{\delta}_i) = \phi v^{1/2}(\boldsymbol{\mu}_i) \boldsymbol{\Lambda} v^{1/2}(\boldsymbol{\mu}_i) \quad (24.4)$$

$\mathbf{y}_i, \mathbf{y}_{i'}$ are independent for $i \neq i'$

where

- \mathbf{y}_i is the $n_i \times 1$ response vector for observations in the i th of m groups;
- $\boldsymbol{\mu}_i$ is the $n_i \times 1$ expectation vector for the response, conditional on the random effects;
- $\boldsymbol{\eta}_i$ is the $n_i \times 1$ linear predictor for the elements of the response vector;
- $g(\cdot)$ is the link function, transforming the conditional expected response to the linear predictor;
- \mathbf{X}_i is the $n_i \times p$ model matrix for the fixed effects of observations in group i ;
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients;
- \mathbf{Z}_i is the $n_i \times q$ model matrix for the random effects of observations in group i ;
- $\boldsymbol{\delta}_i$ is the $q \times 1$ vector of random-effect coefficients for group i ;
- $\boldsymbol{\Psi}$ is the $q \times q$ covariance matrix of the random effects;
- $\boldsymbol{\Lambda}_i$ is $n_i \times n_i$ and expresses the dependence structure for the conditional distribution of the response within each group—for example, if the observations are sampled independently in each group, $\boldsymbol{\Lambda}_i = \mathbf{I}_{n_i}$;
- $v^{1/2}(\boldsymbol{\mu}_i) \equiv \text{diag}[\sqrt{v(\mu_{ij})}]$, with the form of the variance function $v(\cdot)$ depending on the distributional family to which \mathbf{y}_i belongs; and
- ϕ is the dispersion parameter.

Alternatively, for quasi-likelihood estimation, the variance function $v(\cdot)$ can be given directly, without assuming an exponential family for Y .⁸

Estimating Generalized Linear Mixed Models

Estimation of the GLMM is considerably more complex than estimation of the LMM, and so I will avoid the details. As in the case of the LMM, it is convenient to rewrite the GLMM (Equation 24.2) for all $n = \sum n_i$ observations simultaneously:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} \\ g(\boldsymbol{\mu}) &= g[E(\mathbf{y} | \boldsymbol{\delta})] = \boldsymbol{\eta} \end{aligned}$$

⁸As in the generalized linear model; see Section 15.3.

where the fixed-effects model matrix \mathbf{X} , the random-effects model matrix \mathbf{Z} , the response vector \mathbf{y} , the fixed-effects coefficients $\boldsymbol{\beta}$, and the random-effects coefficients $\boldsymbol{\delta}$ are similar to the analogous terms in the LMM (Equation 23.16 on page 734). The linear predictor

$$\eta_{(n \times 1)} \equiv [\eta'_1, \eta'_2, \dots, \eta'_m]'$$

is likewise the stacked-up column vector of linear predictors, and

$$\mu_{(n \times 1)} \equiv [\mu'_1, \mu'_2, \dots, \mu'_m]'$$

is the stacked-up conditional expectation vector of the response. As in the linear mixed model, the random effects are multivariately normally distributed, $\boldsymbol{\delta} \sim N_{mq}(\mathbf{0}, \boldsymbol{\Psi}^*)$, where $\boldsymbol{\Psi}^*$ is a block-diagonal matrix with $\boldsymbol{\Psi}$ on the diagonal blocks. The variance-covariance matrix of the response conditional on the random effects is

$$V(\mathbf{y}|\boldsymbol{\delta}) = \phi v^{1/2}(\boldsymbol{\mu}) \boldsymbol{\Lambda} v^{1/2}(\boldsymbol{\mu})$$

where $\boldsymbol{\Lambda}$ is an $n \times n$ block-diagonal matrix with the $\boldsymbol{\Lambda}_i$ on the diagonal blocks, and $v^{1/2}(\boldsymbol{\mu}) = \text{diag}[\sqrt{v(\mu_{ij})}]$ is an $n \times n$ diagonal matrix.

The distribution of the random effects $\boldsymbol{\delta}$ is multivariate normal:

$$p(\boldsymbol{\delta}|\boldsymbol{\Psi}) = \frac{1}{(2\pi)^{mq/2} \sqrt{\det \boldsymbol{\Psi}^*}} \exp(\boldsymbol{\delta}' \boldsymbol{\Psi}^{*-1} \boldsymbol{\delta})$$

The distribution of the response conditional on the random effects, $p(\mathbf{y}|\boldsymbol{\beta}, \phi, \boldsymbol{\delta})$, depends on the distributional family from which the response is drawn.⁹ To obtain the marginal distribution of the data, we must integrate over the random effects,

$$p(\mathbf{y}|\boldsymbol{\beta}, \phi) = \int_{\boldsymbol{\delta}} p(\mathbf{y}|\boldsymbol{\beta}, \phi, \boldsymbol{\delta}) p(\boldsymbol{\delta}|\boldsymbol{\Psi}) d\boldsymbol{\delta} \quad (24.5)$$

Then, maximizing $p(\mathbf{y}|\boldsymbol{\beta}, \phi)$ produces maximum-likelihood estimates of the fixed effects $\boldsymbol{\beta}$, along with the dispersion parameter ϕ . The integral in Equation 24.5 is difficult to evaluate, however, leading to the approximate methods mentioned in Section 24.1: penalized quasi-likelihood (PQL), the Laplace approximation, and Gauss-Hermite quadrature (in order of increasing general accuracy). Because better methods are now widely available in statistical software, it is in particular a good idea to avoid estimation by PQL. Gauss-Hermite quadrature was used to fit a binomial GLMM to the headache data in Section 24.1.1.¹⁰

24.2 Nonlinear Mixed Models*

For the i th of n independent observations, the nonlinear regression model is

$$Y_i = f(\boldsymbol{\beta}, \mathbf{x}'_i) + \varepsilon_i$$

where Y_i is the response variable, $\boldsymbol{\beta}$ is a vector of regression coefficients, \mathbf{x}'_i is a vector of explanatory variables, and ε_i is the error. We assume that $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and that ε_i and $\varepsilon_{i'}$ are independent for $i \neq i'$. The nonlinear regression function $f(\cdot)$ is specified explicitly as part of

⁹See Section 15.3.1.

¹⁰I employed the **glmer** function in the **lme4** package for R for these computations.

the model. Under these assumptions, maximum-likelihood estimates for the parameters of the model are provided by nonlinear least squares.¹¹

One extension of the nonlinear regression model to include random effects, due to Pinheiro and Bates (2000), is as follows (but with different notation than in the original source):

$$\begin{aligned}\mathbf{y}_i &= f(\boldsymbol{\theta}_i, \mathbf{X}_i) + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\theta}_i &= \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\delta}_i\end{aligned}\tag{24.6}$$

where

- \mathbf{y}_i is the $n_i \times 1$ response vector for the n_i observations in the i th of m groups.
- \mathbf{X}_i is a $n_i \times s$ matrix of explanatory variables (some of which may be categorical) for observations in group i .
- $\boldsymbol{\varepsilon}_i \sim \mathbf{N}_{n_i}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$ is a $n_i \times 1$ vector of multivariately normally distributed errors for observations in group i ; the matrix $\boldsymbol{\Lambda}_i$, which is $n_i \times n_i$, is typically parametrized in terms of a much smaller number of parameters, and $\boldsymbol{\Lambda}_i = \mathbf{I}_{n_i}$ if the observations are independently sampled within groups.
- $\boldsymbol{\theta}_i$ is a $n_i \times 1$ *composite coefficient vector* for the observations in group i , incorporating both fixed and random effects.
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect parameters.
- $\boldsymbol{\delta}_i \sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ is the $q \times 1$ vector of random-effect coefficients for group i .
- \mathbf{A}_i and \mathbf{B}_i are, respectively, $n_i \times p$ and $n_i \times q$ matrices of known constants for combining the fixed and random effects in group i . These will often be “incidence matrices” of 0s and 1s but may also include Level 1 explanatory variables, treated as conditionally fixed (as in the standard linear model).

Like fundamentally nonlinear fixed-effects regression models, *nonlinear mixed-effects models (NLMMs)* are uncommon in the social and behavioral sciences. Variable transformations, regression splines, and polynomial regressors allow us to fit a wide variety of nonlinear relationships within the ambit of the LMM. Nevertheless, as in the following example, it is occasionally more natural to specify a nonlinear mixed model, especially when the parameters of the model have compelling substantive interpretations.

The nonlinear mixed-effects model (NLMM) takes the form

$$\begin{aligned}\mathbf{y}_i &= f(\boldsymbol{\theta}_i, \mathbf{X}_i) + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\theta}_i &= \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\delta}_i\end{aligned}$$

where \mathbf{y}_i is the response vector for the i th cluster; \mathbf{X}_i is a matrix of explanatory variables, also for the i th cluster; $\boldsymbol{\theta}_i$ is the composite coefficient vector for the observations in cluster i ; $\boldsymbol{\beta}$ is the vector of fixed-effect parameters; $\boldsymbol{\delta}_i \sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ is the vector of random-effect coefficients for cluster i ; and \mathbf{A}_i and \mathbf{B}_i are matrices of known constants for combining the fixed and random effects, typically containing 0s and 1s along with Level 1 explanatory variables. Like the GLMM, the NLMM is estimated by approximate maximum likelihood.

¹¹Fundamentally nonlinear models for independent observations and nonlinear least-squares estimation are developed in Section 17.4.

24.2.1 Example: Recovery From Coma

The data and model for this example are taken from Wong, Monette, and Weiner (2001).¹² The data pertain to 200 patients who sustained traumatic brain injuries resulting in comas of varying duration. After awakening from their comas, patients were periodically administered a standard IQ test. In this section, I will examine recovery of “performance IQ” (“PIQ”) post-coma; the data set also includes a measure of verbal IQ.¹³

About half of the patients in the study (107) completed a single IQ test, but the remainder were measured on two to five irregularly timed occasions, raising the possibility of tracing the trajectory of IQ recovery postcoma. A mixed-effects model is very useful here because it allows us to pool the information in the small number of observations available per patient to estimate the typical within-subject trajectory of recovery along with variation in this trajectory.

After examining the data, Wong et al. posited the following *asymptotic growth model* for IQ recovery:

$$\begin{aligned} Y_{ij} &= \theta_{1i} + \theta_{2i} e^{-\theta_{3i} X_{1ij}} + \varepsilon_{ij} \\ \theta_{1i} &= \beta_1 + \beta_2 \sqrt{X_{2i}} + \delta_{1i} \\ \theta_{2i} &= \beta_3 + \beta_4 \sqrt{X_{2i}} + \delta_{2i} \\ \theta_{3i} &= \beta_5 \end{aligned} \tag{24.7}$$

where the variables and parameters of the model have the following interpretations (see Figure 24.2):

- Y_{ij} is the PIQ of the i th patient measured on the j th occasion, $j = 1, \dots, n_i$; as mentioned, $n_i = 1$ for about half the patients.
- X_{1ij} is the time postcoma (in days) for the i th patient at the j th occasion.
- X_{2i} is the duration of the coma (in days) for the i th patient.
- θ_{1i} is the eventual, recovered level of PIQ for patient i , specified to depend linearly on the square root of the length of the coma, with fixed-effect parameters β_1 and β_2 , as well as a random-effect component δ_{1i} . Were patients to recover PIQ fully, the average value of θ_{1i} would be 100, assuming that coma patients are representative of the general population in their precoma average level of IQ. Thus, the fixed-effect intercept β_1 is interpretable as the expected eventual level of PIQ for a patient in a coma of zero days duration.
- θ_{2i} is the negative of the amount of PIQ eventually regained by patient i , beginning at the point of recovery from coma. Like θ_{1i} , the coefficient θ_{2i} has a fixed-effect component depending linearly on length of coma, with parameters β_3 and β_4 , and a random-effect component, δ_{2i} .

¹²I am grateful to Georges Monette for making the data and associated materials available to me. The analysis reported here is very similar to that in the original source.

¹³See Exercise 24.2 for a parallel analysis of the data on verbal IQ.

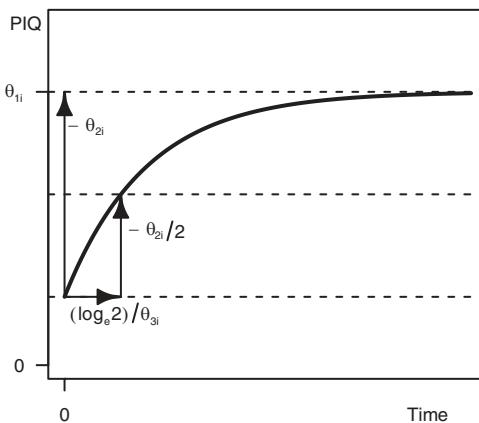


Figure 24.2 The asymptotic growth model for recovery of IQ following coma, $Y_{ij} = \theta_{1i} + \theta_{2i}e^{-\theta_{3i}X_{1ij}} + \varepsilon_{ij}$, where Y_{ij} is the PIQ and X_{1ij} is the time postcoma for subject i on occasion j . The parameter θ_{1i} represents the eventual level of PIQ for subject i , $-\theta_{2i}$ is the amount of PIQ recovered by subject i , and θ_{3i} is the rate of recovery for subject i (fixed across subjects), with $(\log_e 2)/\theta_{3i}$ representing the time to half-recovery.

- θ_{3i} is the recovery rate for patient i , treated as a fixed effect, β_5 , with $(\log_e 2)/\theta_{3i}$ representing the time required to recover half the difference between final and (expected) initial postcoma PIQ (the “half-recovery” time), that is, $-\theta_{2i}/2$.¹⁴
- ε_{ij} is the error for patient i on occasion j .

There are, therefore, four variance-covariance components in this model, $V(\varepsilon_{ij}) = \sigma_\varepsilon^2$, $V(\delta_{1i}) = \psi_1^2$, $V(\delta_{2i}) = \psi_2^2$, and $C(\delta_{1i}, \delta_{2i}) = \psi_{12}$. Although the data are longitudinal, there are too few observations per patient to entertain a model with serially correlated errors.

Before fitting this model, I will examine the data, both to determine whether the posited model seems reasonable and to provide rough guesses for the fixed-effects parameters. As in nonlinear least squares,¹⁵ initial guesses of the fixed-effects parameters provide a starting point for the iterative process of maximizing the likelihood in the NLMM.

Figure 24.3 is a scatterplot of PIQ versus number of days postcoma, with the observations for each patient connected by lines. Forty of the 331 measurements were taken after 1000 days postcoma, and these are omitted from the graph to allow us to discern more clearly the general pattern of the data. The line on the plot is drawn by local linear regression.¹⁶ Mixing together the observations from all patients makes the scatterplot difficult to interpret, but on the other hand, there are too few observations for each patient to establish clear individual trajectories.

¹⁴It makes substantive sense to treat the patients’ recovery rates as potentially variable—that is, as a random effect—but doing so introduces three additional parameters (a variance component and two covariance components) yet leaves the likelihood essentially unchanged. The very small number of observations per patient produces very little information in the data for estimating patient-specific recovery rates.

¹⁵See Section 17.4.

¹⁶See Section 18.1.2.

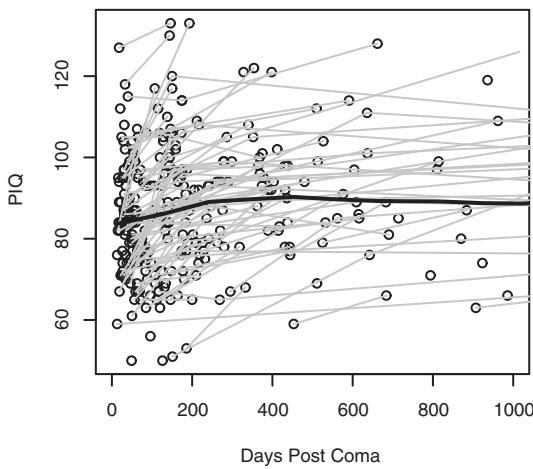


Figure 24.3 Scatterplot for PIQ versus days since awakening from coma. Observations beyond 1000 days are not shown, and the observations for each patient are connected by gray lines. The heavier black line is for a nonparametric regression smooth.

Nevertheless, the asymptotic growth model is roughly consistent with the general pattern of the data, and the patients for whom there are multiple observations do tend to improve over time.

Figure 24.4 is a scatterplot of the initial PIQ measurement for each patient against the length of the patient's coma (in days, on the square-root scale). These initial measurements were taken at varying times postcoma and therefore should not be interpreted as the PIQ at time of awakening (i.e., time 0) for each patient. The relationship of initial PIQ to square-root length of coma appears to be reasonably linear.

These two graphs also provide a basis for obtaining initial values of the fixed-effects parameters in the mixed model of Equations 24.7:

- Figure 24.3 leads me to expect that the average eventual level of recovered IQ will be less than 100, but Figure 24.4 suggests that the average eventual level for those who spent fewer days in a coma should be somewhat higher; I therefore use the start value $\beta_1^{(0)} = 100$.
- The slope of the least-squares line in Figure 24.4, relating initial PIQ to the square-root of length of coma, is -1.9 , and thus I take $\beta_2^{(0)} = -2$.
- The parameter β_3 represents the negative of the expected eventual gain in PIQ for a patient who spent 0 days in a coma. On the basis of Figure 24.3, I will guess that such patients start on average at a PIQ of 90 and eventually recover to an average of 100, suggesting the start value $\beta_3^{(0)} = -10$.
- The parameter β_4 represents the change in expected eventual PIQ gain with a 1-unit increase in the length of the coma on the square-root scale. My examination of the data does not provide a basis for guessing the value of this parameter, and so I will take $\beta_4^{(0)} = 0$.
- Recall that the time to half-recovery is $(\log_e 2)/\beta_5$. From Figure 24.3, it seems reasonable to guess that the half-recovery time is around 100 days. Thus, $\beta_5^{(0)} = (\log_e 2)/100 = 0.007$.

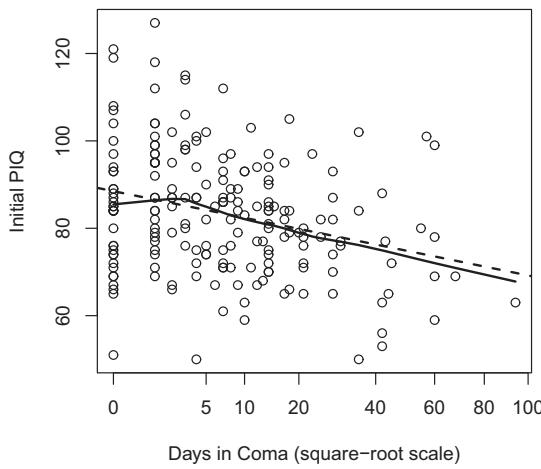


Figure 24.4 Scatterplot of the initial PIQ measurement for each patient (not necessarily taken at day 0) versus the number of days the patient spent in a coma (on the square-root scale). The broken line is a least-squares line, while the solid line is a nonparametric-regression smooth.

With these start values for the fixed effects, maximum-likelihood estimation of the model converges rapidly to the following parameter estimates:¹⁷

Parameter	ML Estimate	Std. Error
β_1	97.09	2.04
β_2	-1.245	0.480
β_3	-11.15	3.21
β_4	-3.248	1.077
β_5	0.008251	0.001651
σ_e	6.736	
ψ_1	13.77	
ψ_2	2.606	
ψ_{12}	-35.67	

All of the estimated fixed-effects parameters are considerably larger than their standard errors. The estimated correlation between the random effects δ_{1i} and δ_{2i} is very high; however, $r_{\delta_1 \delta_2} = -35.67/(13.77 \times 2.606) = -.994$. We might either simplify the model, say by eliminating random effects δ_{2i} from the equation for θ_{2i} , or by reparameterizing the model to reduce the correlation between the random effects.

The estimates of the fixed effects suggest that the average final level of recovered PIQ for individuals in a coma of 0 days duration is $\hat{\beta}_1 = 97.1$. This level declines, as anticipated, with the length of the coma, $\hat{\beta}_2 = -1.25$. On average, patients who spend 0 days in a coma recover $-\hat{\beta}_3 = 11.1$ PIQ points, and the average size of the recovery increases with the length of the

¹⁷But REML estimates do not converge without simplifying the random effects.

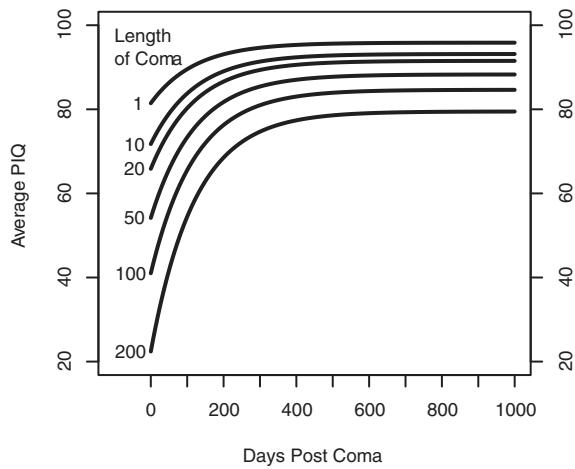


Figure 24.5 Fixed-effect plot of average PIQ by days since recovery from coma and length of coma in days, based on the NLMM fit to the coma-recovery data.

SOURCE: Adapted from Wong et al. (2001).

coma, $-\hat{\beta}_4 = 3.25$. The estimated half-recovery time is $(\log_e 2)/\hat{\beta}_5 = (\log_e 2)/0.00825 = 84$ days. The fixed-effect display in Figure 24.5, similar to one reported in the original paper by Wong et al., shows how typical PIQ recovery varies as a function of days postcoma and length of coma.

24.2.2 Estimating Nonlinear Mixed Models

As in the LMM, it is convenient to write the NLMM simultaneously for all $n = \sum n_i$ observations:

$$\begin{aligned}\mathbf{y} &= f(\boldsymbol{\theta}, \mathbf{X}) + \boldsymbol{\varepsilon} \\ \boldsymbol{\theta} &= \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\delta}\end{aligned}$$

where \mathbf{y} is the $n \times 1$ stacked-up response vector, $\boldsymbol{\varepsilon}$ is the $n \times 1$ stacked-up vector of errors, $\boldsymbol{\delta}$ is the $n \times 1$ stacked-up vector of random effects, $\boldsymbol{\theta}$ is the stacked-up $n \times 1$ composite coefficient vector, \mathbf{X} is the stacked-up $n \times s$ explanatory-variable matrix, and \mathbf{A} and \mathbf{B} are the stacked-up $n \times p$ and $n \times mq$ “incidence matrices” respectively for the fixed and random effects. The incidence matrix for the random effects has a block-diagonal structure:

$$\mathbf{B}_{(n \times mq)} \equiv \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_m \end{bmatrix}$$

The assumed distributions of the errors and the random effects are the same as in the LMM:¹⁸

¹⁸See Section 23.9.1.

$$\begin{aligned}\varepsilon &\sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{\Lambda}) \\ \boldsymbol{\delta} &\sim N_{mq}(\mathbf{0}, \boldsymbol{\Psi}^*)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\Lambda}_{(n \times n)} &\equiv \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Lambda}_2 \end{bmatrix} \\ \boldsymbol{\Psi}^*_{(mq \times mq)} &\equiv \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Psi} \end{bmatrix}\end{aligned}$$

The distribution of the random effects is therefore the same as for a linear mixed model, with multivariate-normal density

$$p(\boldsymbol{\delta}) = \frac{1}{(2\pi)^{mq/2} \sqrt{\det \boldsymbol{\Psi}^*}} \exp\left(\frac{1}{2} \boldsymbol{\delta}' \boldsymbol{\Psi}^{*-1} \boldsymbol{\delta}\right)$$

Because the individual-level errors ε have 0 expectations, the expectation and covariance matrix of \mathbf{y} , conditional on the random effects, are

$$\begin{aligned}E(\mathbf{y}|\boldsymbol{\delta}) &= \boldsymbol{\theta} \\ V(\mathbf{y}|\boldsymbol{\delta}) &= \sigma_\varepsilon^2 \boldsymbol{\Lambda}\end{aligned}$$

and the multivariate-normal conditional density of \mathbf{y} is

$$p(\mathbf{y}|\boldsymbol{\delta}) = \frac{1}{(2\pi)^{N/2} \sigma_\varepsilon \sqrt{\det \boldsymbol{\Lambda}}} \exp\left[\frac{1}{2\sigma_\varepsilon^2} (\mathbf{y} - \boldsymbol{\theta})' \boldsymbol{\Lambda}^{-1} (\mathbf{y} - \boldsymbol{\theta})\right]$$

The marginal density of \mathbf{y} , integrating over the random effects, is therefore

$$\begin{aligned}p(\mathbf{y}) &= \int_{\boldsymbol{\delta}} p(\mathbf{y}|\boldsymbol{\delta}) p(\boldsymbol{\delta}) d\boldsymbol{\delta} \\ &= \int_{\boldsymbol{\delta}} \frac{1}{(2\pi)^{N/2} \sigma_\varepsilon \sqrt{\det \boldsymbol{\Lambda}}} \exp\left[\frac{1}{2\sigma_\varepsilon^2} (\mathbf{y} - \boldsymbol{\theta})' \boldsymbol{\Lambda}^{-1} (\mathbf{y} - \boldsymbol{\theta})\right] \\ &\quad \times \frac{1}{(2\pi)^{mq/2} \sqrt{\det \boldsymbol{\Psi}^*}} \exp\left(\frac{1}{2} \boldsymbol{\delta}' \boldsymbol{\Psi}^{*-1} \boldsymbol{\delta}\right) d\boldsymbol{\delta}\end{aligned}\tag{24.8}$$

Treating this formula as a function of the parameters $\boldsymbol{\beta}$, σ_ε^2 , $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$ produces the likelihood for the NLMM. Because the integral in Equation 24.8 is not tractable analytically, numerical methods of approximation, similar to those used for GLMMs,¹⁹ are required to maximize the likelihood.

¹⁹See Section 24.1.2.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 24.1. Further on migraine headaches:

- A graph of the fixed effects for the mixed-effects logit model fit to the migraine headaches data is shown in Figure 24.1 (page 747), and the estimated parameters of the model are given on page 746. Explain how the lines on the graph, showing how the fitted probability of headache occurrence depends on medication group and time, can be computed from the estimates of the fixed effects.
- As mentioned (footnote 5 on page 745), the original analysis of the migraine headaches data used a regression spline, rather than a square-root transformation, for time posttreatment. Reanalyze the data using a regression spline for time posttreatment, and compare the results to those produced by the model employed in the text.²⁰

Exercise 24.2. Further on recovery from coma:

- The example in Section 24.2.1 on recovery from coma uses data on performance IQ. The original analysis of the data by Wong et al. (2001) also examined verbal IQ. Repeat the analysis using verbal IQ as the response variable, employing the non-linear mixed-effects model in Equations 24.7. Compare the results for postcoma recovery of performance and verbal IQ.
- Figure 24.5 (page 755) shows the trajectory of postcoma performance IQ as a function of length of coma and days postcoma, with days postcoma on the horizontal axis of the graph and lines drawn for selected values of length of coma. Using the estimates of the fixed effects (given on page 754), draw an alternative graph with length of coma on the horizontal axis and different lines for selected values of days postcoma. Then draw a 3D plot of the fitted regression surface, with average PIQ on the vertical axis and length of coma and days postcoma as the “horizontal” axes.

Summary

- The generalized linear mixed-effects model (GLMM) may be written as

$$\begin{aligned}\eta_{ij} &= \beta_1 + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij} + \delta_{1i} Z_{1ij} + \cdots + \delta_{qi} Z_{qij} \\ g(\mu_{ij}) &= E(Y_{ij} | \delta_{1i}, \dots, \delta_{qi}) = \eta_{ij} \\ \delta_{ki} &\sim N(0, \psi_k^2), C(\delta_{ki}, \delta_{k'i}) = \psi_{kk'} \\ \delta_{ki}, \delta_{k'i} &\text{ are independent for } i \neq i' \\ V(Y_{ij}) &= \phi v(\mu_{ij}) \lambda_{ij} \\ C(Y_{ij}, Y_{ij'}) &= \phi \sqrt{v(\mu_{ij})} \sqrt{v(\mu_{ij'})} \lambda_{ijj'} \\ Y_{ij}, Y_{ij'} &\text{ are independent for } i \neq i'\end{aligned}$$

²⁰See Section 17.2 for information on regression splines.

where η_{ij} is the linear predictor for observation j in cluster i ; the fixed-effect coefficients (β s), random-effect coefficients (δ s), fixed-effect regressors (X s), and random-effect regressors (Z s) are defined as in the LMM; and the dispersion parameter ϕ and variance function $\nu(\cdot)$ depend on the distributional family to which Y belongs; alternatively, for quasi-likelihood estimation, $\nu(\cdot)$ can be given directly. The GLMM is estimated by approximate maximum likelihood.

- The nonlinear mixed-effects model (NLMM) takes the form

$$\begin{aligned}\mathbf{y}_i &= f(\boldsymbol{\theta}_i, \mathbf{X}_i) + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\theta}_i &= \mathbf{A}_i \boldsymbol{\beta} + \mathbf{B}_i \boldsymbol{\delta}_i\end{aligned}$$

where \mathbf{y}_i is the response vector for the i th cluster; \mathbf{X}_i is a matrix of explanatory variables, also for the i th cluster; $\boldsymbol{\theta}_i$ is the composite coefficient vector for the observations in cluster i ; $\boldsymbol{\beta}$ is the vector of fixed-effect parameters; $\boldsymbol{\delta}_i \sim N_q(\mathbf{0}, \boldsymbol{\Psi})$ is the vector of random-effect coefficients for cluster i ; and \mathbf{A}_i and \mathbf{B}_i are matrices of known constants for combining the fixed and random effects, typically containing 0s and 1s along with Level 1 explanatory variables. Like the GLMM, the NLMM is estimated by approximate maximum likelihood.

Recommended Reading

- Of the recommended readings in the previous chapter, Raudenbush and Bryk (2002) have the most extensive coverage of generalized linear mixed models.
- Stroup (2013) strongly emphasizes the generalized linear mixed model, treating other statistical models—linear models, generalized linear models, and linear mixed-effects models—as special cases. The presentation is considerably more demanding than in the other recommended sources in this and the preceding chapter, and Stroup derives all of the basic results for linear and generalized linear mixed models. The examples in the text are not oriented toward the social sciences.

Appendix A

Notation

Specific notation is introduced at various points in the appendices and chapters. Throughout the text, I adhere to the following general conventions, with few exceptions. (Examples are shown in brackets.)

- Known scalar constants (including subscripts) are represented by lowercase italic letters [a, b, x_i, x_i^*].
- Observable scalar random variables are represented by uppercase italic letters [X, Y_i, B_0'] or if the names contain more than one character, by Roman letters, the first of which is uppercase [RegSS, RSS₀]. Where it is necessary to make the distinction, *specific values* of random variables are represented as constants [x, y_i, b_0'].
- Scalar parameters are represented by lowercase Greek letters [$\alpha, \beta, \beta_j^*, \gamma_2$]. (See the Greek alphabet in Table 1.) Their estimators are generally denoted by “corresponding” italic characters [$\hat{A}, \hat{B}, \hat{B}_j^*, \hat{C}_2$] or by Greek letters with diacritics [$\check{\alpha}, \check{\beta}$].
- Unobservable scalar random variables are also represented by lowercase Greek letters [ε_i].
- Vectors and matrices are represented by boldface characters—lowercase for vectors [$\mathbf{x}_1, \boldsymbol{\beta}$] and uppercase for matrices [$\mathbf{X}, \boldsymbol{\Sigma}_{12}$]. Roman letters are used for constants and observable random variables [$y, \mathbf{x}_1, \mathbf{X}$]. Greek letters are used for parameters and unobservable random variables [$\boldsymbol{\beta}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\varepsilon}$]. It is occasionally convenient to show the order of a vector or matrix below the matrix [$\begin{smallmatrix} \boldsymbol{\varepsilon} \\ (n \times 1) \end{smallmatrix}, \begin{smallmatrix} \mathbf{X} \\ (n \times k+1) \end{smallmatrix}$]. The order of an identity matrix is given by a subscript [\mathbf{I}_n]. A zero matrix or vector is represented by a boldface 0 [$\mathbf{0}$]; a vector of 1s is represented by a boldface 1, possibly subscripted with its number of elements [$\mathbf{1}_n$]. Vectors are column vectors, unless they are explicitly transposed [column: \mathbf{x} ; row: \mathbf{x}'].
- Diacritics and symbols such as * (asterisk) and ' (prime) are used freely as modifiers to denote alternative forms [$\mathbf{X}^*, \boldsymbol{\beta}', \tilde{\boldsymbol{\varepsilon}}$].

Table 1 The Greek Alphabet With Roman “Equivalents”

Greek Letter		Roman Equivalent	
Lowercase	Uppercase	Phonetic	Other
α	A	alpha	a
β	B	beta	b
γ	Γ	gamma	g, n
δ	Δ	delta	d
ε	\mathcal{E}	epsilon	e
ζ	Z	zeta	z
η	H	eta	e
θ	Θ	theta	th
ι	I	iota	i
κ	K	kappa	k
λ	Λ	lambda	l
μ	M	mu	m
ν	N	nu	n
ξ	Ξ	xi	x
\o	O	omicron	o
π	Π	pi	p
ρ	P	rho	r
σ	Σ	sigma	s
τ	T	tau	t
υ	Υ	upsilon	y, u
ϕ	Φ	phi	ph
χ	X	chi	ch
ψ	Ψ	psi	ps
ω	Ω	omega	o
			w

- The symbol \equiv can be read as “is defined by” or “is equal to by definition” [$\bar{X} \equiv (\sum X_i)/n$].
- The symbol \approx means “is approximately equal to” [$1/3 \approx 0.333$].
- The symbol \ll means “much less than” [$p \ll .0001$].
- The symbol \sim means “is distributed as” [$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$].
- The symbol \in denotes membership in a set [$1 \in \{1, 2, 3\}$].
- The operator $E(\)$ denotes the expectation of a scalar, vector, or matrix random variable [$E(Y_i), E(\varepsilon), E(\mathbf{X})$].
- The operator $V(\)$ denotes the variance of a scalar random variable or the variance-covariance matrix of a vector random variable [$V(\varepsilon_i), V(\mathbf{b})$].
- Estimated variances or variance-covariance matrices are indicated by a circumflex (“hat”) placed over the variance operator [$\hat{V}(\varepsilon_i), \hat{V}(\mathbf{b})$].
- The operator $C(\)$ gives the covariance of two scalar random variables or the covariance matrix of two vector random variables [$C(X, Y), C(\mathbf{x}_i, \varepsilon)$].

- The operators $\mathcal{E}(\)$ and $\mathcal{V}(\)$ denote asymptotic expectation and variance, respectively. Their usage is similar to that of $E(\)$ and $V(\)$ [$\mathcal{E}(B)$, $\mathcal{V}(\hat{\beta})$, $\hat{\mathcal{V}}(B)$].
- Probability limits are specified by plim [$\text{plim } b = \beta$].
- Standard mathematical functions are shown in lowercase [$\cos W$, trace (\mathbf{A})]. The base of the log function is always specified explicitly, unless it is irrelevant [$\log_e L$, $\log_{10} X$]. The exponential function $\exp(x)$ represents e^x .
- The summation sign \sum is used to denote continued addition [$\sum_{i=1}^n X_i \equiv X_1 + X_2 + \cdots + X_n$]. Often, the range of the index is suppressed if it is clear from the context [$\sum_i X_i$], and the index may be suppressed as well [$\sum X_i$]. The symbol \prod similarly indicates continued multiplication [$\prod_{i=1}^n p(Y_i) \equiv p(Y_1)(Y_2) \times \cdots \times p(Y_n)$]. The symbol $\#$ indicates a count [$\#\sum_{i=1}^n (T_b^* \geq T)$].
- To avoid awkward and repetitive phrasing in the statement of definitions and results, the words “if” and “when” are understood to mean “if and only if,” unless explicitly indicated to the contrary. Terms are generally set in *italics* when they are introduced. (“Two vectors are *orthogonal* if their inner product is 0.”)

References

- Abler, R., Adams, J. S., & Gould, P. (1971). *Spatial organization: The geographer's view of the world*. Englewood Cliffs, NJ: Prentice Hall.
- Achen, C. H. (1982). *Interpreting and using regression*. Beverly Hills, CA: Sage.
- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.
- Agresti, A. (1990). *Categorical data analysis*. Hoboken, NJ: John Wiley.
- Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley.
- Aitkin, M., Francis, B., & Hinde, J. (2005). *Statistical modeling in GLIM4* (2nd ed.). Oxford, UK: Clarendon.
- Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.
- Allison, P. D. (2014). *Event history and survival analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Alvarez, R. M., & Nagler, J. (1998). When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42, 55–96.
- Andersen, R. (2007). *Modern methods for robust regression*. Thousand Oaks, CA: Sage.
- Andersen, R., Heath, A., & Sinnott, R. (2002). Political knowledge and electoral choice. *British Elections and Parties Review*, 12, 11–27.
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: John Wiley.
- Andrews, D. F. (1979). The robustness of residual displays. In R. L. Launer & G. N. Wilkenson (Eds.), *Robustness in statistics* (pp. 19–32). New York: Academic Press.
- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59, 817–858.
- Andrews, F. M., Morgan, J. N., & Sonquist, J. A. (1973). *Multiple classification analysis: A report on a computer program for multiple regression using categorical predictors* (2nd ed.). Ann Arbor: Institute for Social Research, University of Michigan.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.
- Anscombe, F. J. (1960). Rejection of outliers [with commentary]. *Technometrics*, 2, 123–166.
- Anscombe, F. J. (1961). Examination of residuals. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 1–36.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27, 17–22.
- Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics*, 5, 141–160.
- Atkinson, A. C. (1985). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford, UK: Clarendon.
- Atkinson, A., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.
- Auspurg, K. & Hinz, T. (In press). *The factorial survey method*. Thousand Oaks, CA: Sage.
- Bailey, R. A., Harding, S. A., & Smith, G. L. (1989). Cross-validation. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Suppl. Vol., pp. 39–44). New York: John Wiley.
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.
- Bard, Y. (1974). *Nonlinear parameter estimation*. New York: Academic Press.

- Barnard, G. A. (1974). Discussion of Professor Stone's paper. *Journal of the Royal Statistical Society, Series B*, 36, 133–135.
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: John Wiley.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, A*, 160, 268–282.
- Basman, R. L. (1957). A generalized method of linear estimation of coefficients in a structural equation. *Econometrica*, 25, 77–83.
- Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression and its applications*. New York: John Wiley.
- Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth.
- Beckman, R. J., & Cook, R. D. (1983). Outliers. *Technometrics*, 25, 119–163.
- Beckman, R. J., & Trussell, H. J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association*, 69, 199–201.
- Belsley, D. A. (1984). Demeaning condition diagnostics through centering [with commentary]. *American Statistician*, 38, 73–93.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.
- Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.
- Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York: John Wiley.
- Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley.
- Box, G. E. P. (1979) Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York: Academic Press.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–252.
- Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4, 531–550.
- Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The Kernel approach with S-Plus illustrations*. Oxford, UK: Oxford University Press.
- Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, 1287–1294.
- Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time [with discussion]. *Journal of the Royal Statistical Society, Series B*, 37, 149–192.
- Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretical approach*. New York: Springer.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261–304.
- Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis for count data*. Cambridge, UK: Cambridge University Press.
- Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American voter*. New York: John Wiley.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96, 1022–1030.
- Chambers, J. M. (1998). *Programming with data: A guide to the S language*. New York: Springer.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.
- Chambers, J. M., & Hastie, T. J. (Eds.). (1992). *Statistical models in S*. Pacific Grove, CA: Wadsworth.
- Chatfield, C. (2003). *Analysis of time series: An introduction* (6th ed.). London: Chapman & Hall.
- Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. New York: John Wiley.
- Chatterjee, S., & Price, B. (1991). *Regression analysis by example* (2nd ed.). New York: John Wiley.
- Christensen, R. (2011). *Plane answers to complex questions: The theory of linear models* (4th ed.). New York: Springer.

- Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.
- Cleveland, W. S. (1994). *The elements of graphing data* (Rev. ed.). Summit, NJ: Hobart Press.
- Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 309–376). Pacific Grove, CA: Wadsworth.
- Collett, D. (2003). *Modelling binary data* (2nd ed.). London: Chapman & Hall.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.
- Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 351–361.
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15–18.
- Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35, 351–362.
- Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89, 177–189.
- Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics*. New York: John Wiley.
- Cook, R. D., & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22, 495–508.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.
- Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1–10.
- Cook, R. D., & Weisberg, S. (1989). Regression diagnostics with dynamic graphics [with commentary]. *Technometrics*, 31, 277–311.
- Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics*. New York: John Wiley.
- Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: John Wiley.
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice Hall.
- Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.
- Cowles, M., & Davis, C. (1987). The subject matter of psychology: Volunteers. *British Journal of Social Psychology*, 26, 97–102.
- Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, 45, 215–233.
- Davis, C. (1990). Body image and weight preoccupation: A comparison between exercising and non-exercising women. *Appetite*, 15, 13–21.
- Davis, C., Blackmore, E., Katzman, D. K., & Fox, J. (2005). Female adolescents with anorexia nervosa and their parents: A case-control study of exercise attitudes and behaviours. *Psychological Medicine*, 35, 377–386.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Dempster, A. P. (1969). *Elements of continuous multivariate analysis*. Reading, MA: Addison-Wesley.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm [with discussion]. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Department of Mines and Technical Surveys. (1962). *Major roads (map)*. Ottawa, Ontario, Canada: Author.
- Dobson, A. J. (2001). *An introduction to generalized linear models* (2nd ed.). London: Chapman & Hall.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.
- Draper, N. R., & Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimators: Review and comments. *Technometrics*, 21, 451–466.
- Duan, N., & Li, K. C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*, 19, 505–530.
- Duncan, O. D. (1961). A socioeconomic index for all occupations. In A. J. Reiss Jr. (Ed.), *Occupations and social status* (pp. 109–138). New York: Free Press.
- Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.
- Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York: Russell Sage.

- Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, 37, 409–428.
- Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, 38, 159–178.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Emerson, J. D., & Hoaglin, D. C. (1983). Analysis of two-way tables by medians. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 166–210). New York: John Wiley.
- European Values Study Group and World Values Survey Association. (2000). *World values surveys and European value surveys, 1981–1984, 1990–1993, and 1995–1997* [computer file]. Ann Arbor, MI: Institute for Social Research [producer], Inter-University Consortium for Political and Social Research [distributor].
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.
- Firth, D. (1991). Generalized linear models. In D. V. Hinkley, N. Reid, & E. J. Snell (Eds.), *Statistical theory and modelling: In honour of Sir David Cox, FRS* (pp. 55–82). London: Chapman & Hall.
- Firth, D. (2003). Overcoming the reference category problem in the presentation of statistical models. In R. M. Stolzenberg (Ed.), *Sociological methodology 2003* (pp. 1–18). Washington, DC: American Sociological Association.
- Firth, D., & De Menezes, R. X. (2004). Quasi-variances. *Biometrika*, 91, 65–80.
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507–521.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
- Foster, D. P., & Stine, R. A. (2004). Variable selection in data mining: Building a predictive model of bankruptcy. *Journal of the American Statistical Association*, 99, 303–313.
- Fournier, P., Cutler, F., Soroka, S., & Stolle, D. (2013). *Canadian Election Study, 2011: Study Documentation*. Technical report, Canadian Opinion Research Archive, Queen's University, Kingston, Ontario, Canada.
- Fox, B. (1980). *Women's domestic labour and their involvement in wage work: Twentieth-century changes in the reproduction of daily life*. Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada.
- Fox, J. (1984). *Linear statistical models and related methods: With applications to social research*. New York: John Wiley.
- Fox, J. (1987). Effect displays for generalized linear models. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 347–361). Washington, DC: American Sociological Association.
- Fox, J. (1991). *Regression diagnostics: An introduction*. Newbury Park, CA: Sage.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Fox, J. (2000a). *Multiple and generalized nonparametric regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2000b). *Nonparametric simple regression: Smoothing scatterplots*. Thousand Oaks, CA: Sage.
- Fox, J. (2000c). Statistical graphics. In E. F. Borgatta & R. J. V. Montgomery (Eds.), *Encyclopedia of sociology* (2nd ed., Vol. 5, pp. 3003–3023). New York: Macmillan.
- Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1–18.
- Fox, J. (2008). *Applied regression analysis and generalized linear models*. (2nd ed.). Thousand Oaks, CA: Sage.
- Fox, J., & Andersen, R. (2006). Effect displays for multinomial and proportional-odds logit models. In R. M. Stolzenberg (Ed.), *Sociological methodology 2006*. Washington, DC: American Sociological Association.
- Fox, J., & Hartnagel, T. F. (1979). Changing social roles and female crime in Canada: A time series analysis. *Canadian Review of Sociology and Anthropology*, 16, 96–104.
- Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87, 178–183.
- Fox, J., & Suschnigg, C. (1989). A note on gender and the prestige of occupations. *Canadian Journal of Sociology*, 14, 353–360.

- Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.
- Francis, I. (1973). Comparison of several analyses of variance programs. *Journal of the American Statistical Association*, 68, 860–865.
- Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 87, 453–476.
- Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York: Norton.
- Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, 37, 152–155.
- Freedman, D. A. (1987). As others see us: A case study in path analysis [with commentary]. *Journal of Educational Statistics*, 12, 101–223.
- Freedman, J. L. (1975). *Crowding and behavior*. New York: Viking.
- Friendly, M. (1991). *SAS system for statistical graphics*. Cary, NC: SAS Institute.
- Friendly, M. (2002). Corrrgrams: Exploratory displays for correlation matrices. *American Statistician*, 56, 316–324.
- Friendly, M., & Franklin, P. (1980). Interactive presentation in multitrial free recall. *Memory and Cognition*, 8, 265–270.
- Friendly, M., Monette, G., & Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, 28, 1–39.
- Fuller, W. A. (2009). *Sampling statistics*. Hoboken, NJ: John Wiley.
- Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, 16, 499–511.
- Gallant, A. R. (1975). Nonlinear regression. *American Statistician*, 29, 73–81.
- Gao, X. (2007). A nonparametric procedure for the two-factor mixed model with missing data. *Biometrical Journal*, 49, 774–788.
- Gass, S. I. (2003). *Linear programming: Methods and applications* (5th ed.). New York: Dover.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
- Goldberger, A. S. (1964). *Econometric theory*. New York: John Wiley.
- Goldberger, A. S. (1973). Structural equation models: An overview. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 1–18). New York: Seminar Press.
- Goodall, C. (1983). *M-estimators of location: An outline of the theory*. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 339–403). New York: John Wiley.
- Gould, S. J. (1989). *Wonderful life: The Burgess shale and the nature of history*. New York: Norton.
- Green, P. E., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.
- Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. London: Chapman & Hall.
- Greene, I., & Shaffer, P. (1992). Leave to appeal and leave to commence judicial review in Canada's refugee-determination system: Is the process fair? *International Journal of Refugee Law*, 4, 71–83.
- Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Groves, R. M., Fowler, Jr., F., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: John Wiley.
- Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures: A practical approach for behavioural scientists*. London: Chapman & Hall.
- Härdle, W. (1991). *Smoothing techniques: With implementation in S*. New York: Springer.
- Harrell, F. E., Jr. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.
- Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge: MIT Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.
- Hastie, T. J. (1992). Generalized additive models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 249–307). Pacific Grove, CA: Wadsworth.
- Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.
- Hauck, W. W., Jr., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72, 851–853.

- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1271–1271.
- Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica*, 42, 679–693.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Heckman, J. J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–107). New York: Springer.
- Hernandez, F., & Johnson, R. A. (1980). The large sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75, 855–861.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1985). *Exploring data tables, trends, and shapes*. New York: John Wiley.
- Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *American Statistician*, 32, 17–22.
- Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32, 1–49.
- Hocking, R. R. (1985). *The analysis of linear models*. Monterey, CA: Brooks/Cole.
- Hocking, R. R., & Speed, F. M. (1975). The full rank analysis of some linear model problems. *Journal of the American Statistical Association*, 70, 706–712.
- Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.
- Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12, 69–82.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial [with discussion]. *Statistical Science*, 14, 382–401.
- Holland, P. W. (1986). Statistics and causal inference [with commentary]. *Journal of the American Statistical Association*, 81, 945–970.
- Hosmer, D. W., Jr., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: John Wiley.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). Berkeley: University of California Press.
- Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.
- Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data*. Thousand Oaks, CA: Sage.
- Jacoby, W. G. (1998). *Statistical graphics for visualizing multivariate data*. Thousand Oaks, CA: Sage.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (Vol. 1, 2nd ed.). New York: John Wiley.
- Johnston, J. (1972). *Econometric methods* (2nd ed.). New York: McGraw-Hill.
- Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: John Wiley.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kenward, M. G., & Roger, J. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997.
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95, 49–69.
- Kish, L. (1987). *Statistical design for research*. New York: John Wiley.

- Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan.
- Koch, G. G., & Gillings, D. B. (1983). Inference, design based vs. model based. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 84–88). New York: John Wiley.
- Koenker, R. (2005). *Quantile regression*. Cambridge, UK: Cambridge University Press.
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.
- Kostecki-Dillon, T., Monette, G., & Wong, P. (1999). Pine trees, comas, and migraines. *York University Institute for Social Research Newsletter*, 14, 2.
- Kotz, S., Balakrishnan, N., & Johnson, N. L. (1994). *Continuous multivariate distributions: Vol. 1. Models and applications* (2nd ed.). New York: John Wiley.
- Krzanowski, W. J. (1988). *Principles of multivariate analysis: A user's perspective*. Oxford, UK: Clarendon.
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- Landwehr, J. M., Pregibon, D., & Shoemaker, A. C. (1980). Some graphical procedures for studying a logistic regression fit. In *Proceedings of the Business and Economic Statistics Section, American Statistical Association* (pp. 15–20). Alexandria, VA: American Statistical Association.
- Leinhardt, S., & Wasserman, S. S. (1979). Exploratory data analysis: An introduction to selected methods. In K. F. Schuessler (Ed.), *Sociological methodology 1979* (pp. 311–365). San Francisco: Jossey-Bass.
- Li, G. (1985). Robust regression. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 281–343). New York: John Wiley.
- Little, R. J. A., & Rubin, D. B. (1990). The analysis of social science data with missing values. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 374–409). Newbury Park CA: Sage.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224.
- Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, NJ: John Wiley.
- Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535–1546.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, 15, 661–676.
- Mallows, C. L. (1986). Augmented partial residuals. *Technometrics*, 28, 313–319.
- Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *American Statistician*, 36, 15–24.
- Manski, C. (1991). Regression. *Journal of Economic Literature*, 29, 34–50.
- McCallum, B. T. (1973). A note concerning asymptotic covariance expressions. *Econometrica*, 41, 581–583.
- McCullagh, P. (1980). Regression models for ordinal data [with commentary]. *Journal of the Royal Statistical Society, B*, 42, 109–142.
- McCullagh, P. (1991). Quasi-likelihood and estimating functions. In D. V. Hinkley, N. Reid, & E. J. Snell (Eds.), *Statistical theory and modelling: In honour of Sir David Cox, FRS* (pp. 265–286). London: Chapman & Hall.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
- Monette, G. (1990). Geometry of multiple regression and 3-D graphics. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 209–256). Newbury Park, CA: Sage.
- Moody, E. A. (1972). William of Ockham. In P. Edwards (Ed.), *The encyclopedia of philosophy* (Vol. 8, pp. 306–317). New York: Macmillan.
- Moore, D. S., Notz, W. I., & Fligner, M. A. (2013). *The basic practice of statistics* (6th ed.). New York: Freeman.
- Moore, J. C., Jr., & Krupat, E. (1971). Relationship between source status, authoritarianism, and conformity in a social setting. *Sociometry*, 34, 122–134.
- Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

- Morrison, D. F. (2005). *Multivariate statistical methods* (4th ed.). Belmont, CA: Thomson-Brooks-Cole.
- Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology: Vol. 2. Research methods* (2nd ed., pp. 80–203). Reading, MA: Addison-Wesley.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.
- Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford, UK: Oxford University Press.
- Nason, G. P., & Silverman, B. W. (2000). Wavelets for regression and other statistical problems. In M. G. Schimek (Ed.), *Smoothing and regression: Approaches, computation, and application* (pp. 159–191). New York: John Wiley.
- National Opinion Research Center (2005). *General Social Survey (GSS)*. Retrieved October 31, 2005, from www.gss.norc.org
- Nelder, J. A. (1976). Letter to the editor. *American Statistician*, 30, 103.
- Nelder, J. A. (1977). A reformulation of linear models [with commentary]. *Journal of the Royal Statistical Society, A*, 140, 48–76.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370–384.
- Nerlove, M., & Press, S. J. (1973). *Univariate and multivariate log-linear and logistic models*. Santa Monica, CA: RAND Corporation.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703–708.
- Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61, 631–653.
- Northrup, D. (2012). *The 2011 Canadian Election Survey: Technical documentation*. Technical report, Institute for Social Research, York University, Toronto, Ontario, Canada.
- Obenchain, R. L. (1977). Classical *F*-tests and confidence intervals for ridge regression. *Technometrics*, 19, 429–439.
- O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin* 97, 316–333.
- Ornstein, M. D. (1976). The boards and executives of the largest Canadian corporations: Size, composition, and interlocks. *Canadian Journal of Sociology*, 1, 411–437.
- Ornstein, M. D. (1983). *Accounting for gender differences in job income in Canada: Results from a 1981 survey*. Ottawa, Ontario, Canada: Labour Canada.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 559–572.
- Pickup, M. (2014). *Introduction to time series analysis*. Thousand Oaks, CA: Sage.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Powers, D. A., & Xie, Y. (2008). *Statistical methods for categorical data analysis* (2nd ed.). Bingley, UK: Emerald.
- Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705–724.
- Putter, J. (1967). Orthonormal bases of error spaces and their use for investigating the normality and variance of residuals. *Journal of the American Statistical Association*, 62, 1022–1036.
- R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raftery, A. E. (1995). Bayesian model selection in social research [with discussion]. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–195). Washington, DC: American Sociological Association.
- Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83, 251–266.
- Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.
- Raudenbush, S. W., & Bryk, A. S. (2012). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley.
- Rubin, D. B. (1976). Inference and missing data [with discussion]. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.
- Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29, 343–367.
- Sall, J. (1990). Leverage plots for general linear hypotheses. *American Statistician*, 44, 308–315.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Searle, S. R. (1971). *Linear models*. New York: John Wiley.
- Searle, S. R. (1987). *Linear models for unbalanced data*. New York: John Wiley.
- Searle, S. R., Speed, F. M., & Henderson, H. V. (1981). Some computational and model equivalences in analysis of variance of unequal-subclass-numbers data. *American Statistician*, 35, 16–33.
- Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *American Statistician*, 34, 216–221.
- Seber, G. A. F. (1977). *Linear regression analysis*. New York: John Wiley.
- Shryock, H. S., Siegel, J. S., & Associates. (1973). *The methods and materials of demography* (2 vols.). Washington, DC: U.S. Bureau of the Census.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.
- Silverman, B. W., & Young, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika*, 74, 469–479.
- Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238–241.
- Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys*. New York: John Wiley.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.
- Speed, F. M., & Hocking, R. R. (1976). The use of the $R(\cdot)$ -notation with unbalanced data. *American Statistician*, 30, 30–33.
- Speed, F. M., Hocking, R. R., & Hackney, O. P. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73, 105–112.
- Speed, F. M., & Monlezun, C. J. (1979). Exact F -tests for the method of unweighted means in a 2^k experiment. *American Statistician*, 33, 15–18.
- Spence, I., & Lewandowsky, S. (1990). Graphical perception. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 13–57). Newbury Park, CA: Sage.
- Statistics Canada. (1971). *Census of Canada*. Ottawa, Ontario, Canada: Author.
- Stefanski, L. A. (1991). A note on high-breakdown estimators. *Statistics and Probability Letters*, 11, 353–358.
- Steinhorst, R. K. (1982). Resolving current controversies in analysis of variance. *American Statistician*, 36, 138–139.
- Stine, R. (1990). An introduction to bootstrap methods: Examples and ideas. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 325–373). Newbury Park, CA: Sage.
- Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *American Statistician*, 49, 53–56.
- Stine, R. A. (2004). Model selection using information theory and the MDL principle. *Sociological Methods and Research*, 33, 230–260.

- Stine, R. A., & Fox, J. (Eds.). (1996). *Statistical computing environments for social research*. Thousand Oaks, CA: Sage.
- Stolley, P. D. (1991). When genius errs: R. A. Fisher and the lung cancer controversy. *American Journal of Epidemiology*, 133, 416–425.
- Stolzenberg, R. M. (1979). The measurement and decomposition of causal effects in nonlinear and nonadditive models. In K. F. Schuessler (Ed.), *Sociological methodology 1980* (pp. 459–488). San Francisco: Jossey-Bass.
- Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, 62, 494–507.
- Stone, M. (1987). *Coordinate-free multivariable statistics: An illustrated geometric progression from Halmos to Gauss and Bayes*. Oxford, UK: Clarendon.
- Street, J. O., Carroll, R. J., & Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *American Statistician*, 42, 152–154.
- Stroup, W. W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, FL: CRC Press.
- Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45(2), 1–31.
- Theil, H. (1971). *Principles of econometrics*. New York: John Wiley.
- Thompson, M. E. (1988). Superpopulation models. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 9, pp. 93–99). New York: John Wiley.
- Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics*. New York: John Wiley.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5, 232–242.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.
- Tukey, J. W. (1972). Some graphic and semigraphic displays. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor* (pp. 293–316). Ames: Iowa State University Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tukey, J. W. (1986). Discussion. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 58–62). New York: Springer.
- United Nations. (1998). *Social indicators*. Retrieved June 1, 1998, from www.un.org/Depts/unsd/social/main.htm
- U.S. Bureau of the Census. (2006). *The 2006 statistical abstract*. Washington, DC: Author.
- U.S. Bureau of the Census. (2011). *Population distribution and change: 2000 to 2010*. Technical report, United States Census Bureau, Washington, DC.
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall.
- van Buuren, S., & Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden, Germany: TNO Preventie en Gezondheid.
- Veall, M. R., & Zimmermann, K. F. (1996). Pseudo- R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, 10, 241–259.
- Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury.
- Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *American Statistician*, 35, 234–241.
- Velilla, S. (1993). A note on the multivariate Box-Cox transformation to normality. *Statistics and Probability Letters*, 17, 259–263.
- Vinod, H. D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *Review of Economics and Statistics*, 60, 121–131.
- Vinod, H. D., & Ullah, A. (1981). *Recent advances in regression methods*. New York: Dekker.
- Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13, 1378–1402.

- Wainer, H. (Ed.). (1986). *Drawing inferences from self-selected samples*. Mahwah, NJ: Lawrence Erlbaum.
- Wang, P. C. (1985). Adding a variable in generalized linear models. *Technometrics*, 27, 273–276.
- Wang, P. C. (1987). Residual plots for detecting nonlinearity in generalized linear models. *Technometrics*, 29, 435–438.
- Weakliem, D. (1999). A critique of the Bayesian information criterion for model selection [with commentary]. *Sociological Methods and Research*, 27, 359–443.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439–447.
- Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley.
- Weisberg, S. (2014). *Applied linear regression* (4th ed.). Hoboken, NJ: John Wiley.
- White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 38, 817–838.
- Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36, 181–191.
- Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, 18, 327–350.
- Wong, P. P., Monette, G., & Weiner, N. I. (2001). Mathematical models of cognitive recovery. *Brain Injury*, 15, 519–530.
- Wonnacott, R. J., & Wonnacott, T. H. (1979). *Econometrics* (2nd ed.). New York: John Wiley.
- Wood, S. N. (2006). *Generalized additive models: An introduction with R*. London: Chapman & Hall.
- Wu, L. L. (1985). Robust M -estimation of location and regression. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 316–388). San Francisco: Jossey-Bass.
- Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 29, 51–66.
- Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2(2), 181–247.
- Yeo, I., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.
- Young, G. A. (1994). Bootstrap: More than a stab in the dark? [with commentary]. *Statistical Science*, 9, 382–415.

Author Index

- Abler, A., 512, 762
Achen, C. H., 11, 638
Adams, J. S., 512, 762
Agresti, A., 392, 402–403, 417, 435, 762, 765
Aitkin, M., 472, 762
Allison, P. D., 605, 626, 628, 646, 762
Alvarez, R. M., 415, 762
Andersen, R., 146, 394, 403, 604, 762,
Anderson, D. R., 674, 677, 680, 695, 698, 763
Anderson, T. W., 244, 762
Andrews, D. F., 301, 762
Andrews, D. W. K., 488–489, 762
Andrews, F. M., 181, 762
Angrist, J. D., 232, 762
Anscombe, F. J., 28–30, 274–276, 294, 330,
602, 762
Atkinson, A. C., 54, 286, 288, 295, 300–301,
325–326, 340, 762
Ausburg, K., 227, 762
Azzalini, A., 538, 544, 585, 763
- Bailey, R. A., 698, 762
Baker, F. B., 63,
Balakrishnan, N., 630, 767, 768
Bard, Y., 519, 527, 762
Barnard, G. A., 693, 763
Barnett, V., 295, 763
Bartlett, M. S., 322, 763
Basmann, R. L., 234, 763
Bassett, G., 597–598, 768
Bates, D. M., 527, 750, 763, 769
Becker, R. A., 54, 763
Beckman, R. J., 273–274, 763
Belsley, D. A., 276–277, 279, 281–282, 291, 293,
295, 356, 364, 763
Berk, R.A., 8, 11, 763
Blackmore, E., 718, 721–722, 725, 764
Blau, P. M., 237, 763
Bollen, K., 123, 763
Bosker, R. J., 741, 770
Bowman, A. W., 538, 544, 585, 763
Box, G. E. P., 1, 55–56, 76–77, 79, 298, 316,
324–330, 337–339, 457, 526, 625, 641, 763
Breusch, T. S., 329, 763
Brown, R. L., 494, 763
Bryk, A. S., 704, 710, 727, 742, 758, 770
Burnham, K. P., 674, 677, 680, 695,
698, 763
Cameron, A. C., 464, 472, 763
Campbell, A., 11, 408, 435, 763
Campbell, D. T., 11, 763
Cantoni, E., 600, 763
Carroll, J. D., 264, 766
Carroll, R. J., 594, 771
Chambers, J. M., 54, 585, 763–764, 766
Chatfield, C., 481, 483, 485, 491, 500, 763
Chatterjee, S., 277, 279–282, 295, 357, 763
Christensen, R., 264, 763
Cleveland, W. S., 46, 51, 54, 340, 549, 585,
763–764
Collett, D., 417, 764
Collins, L. M., 625, 764
Conover, W. J., 322, 764
Converse, P. E., 408, 435, 763
Cook, R. D., 50, 54, 80, 274, 276–278, 282, 286,
291, 294, 302, 308, 314, 316–318, 329–333,
337, 340, 455–456, 460, 471, 763–764
Coombs, C. H., 63, 764
Cowles, M., 505–506, 523, 764
Cox, D. R., 55–56, 76–77, 79, 298, 316, 324–327,
329–330, 337, 339, 409, 417, 625, 641,
763–764
Cribari-Neto, F., 305, 764
Cutler, F., 461, 765
- Davis, C., 20, 23–25, 50, 83, 86, 88, 91, 96,
103, 111, 124, 267–269, 271, 274–275, 277,
279, 288, 505–506, 523, 665, 718, 721–722,
725, 764
Davison, A. C., 664, 668, 764
Dawes, R. M., 63, 764
De Menezes, R. X., 139, 468, 765
Dempster, A. P., 264, 616, 764
Diaconis, P., 32, 766
Dobson, A. J., 471, 764

- Donner, A., 382, 766
Draper, N. R., 364–365, 527, 764
Duan, N., 333, 764
Duncan, O. D., 48–49, 51, 63, 94–96, 98–100,
114, 116–117, 123–125, 155–156, 209, 216,
219–220, 237–238, 261, 271–272, 274–275,
277–278, 280, 285, 287–288, 293, 594–597,
601, 659–662, 665–666, 763–764, 766
Durbin, J., 492–494, 498, 500, 763, 765

Efron, B., 647, 652, 656, 664, 666, 668, 765
Emerson, J. D., 80, 514, 765
Ervin, L. H., 305, 768
Evans, J. M., 494, 763

Fan, J., 555, 585, 765
Fienberg, S. E., 400, 417, 435–437, 439, 765
Firth, D., 139, 443, 468, 472, 765
Fisher, R. A., 11, 153, 175, 245, 390, 443, 447,
679, 702, 765
Fligner, M. A., xxi, 4, 768
Foster, D. P., 670, 765
Fournier, P., 461, 765
Fowler, F., Jr., 472, 766
Fox, B., 346–348, 354–355, 357, 360, 367, 765
Fox, J., xx–xxi, xxiii, 20, 54, 146, 244, 357–358,
394, 403, 489–491, 493, 498–499, 505, 585,
718, 721–722, 725, 764–766, 768, 770–771
Francis, B., 472, 762
Francis, I., 176, 766
Franklin, P., 191–194, 199, 206, 766
Freedman, D. A., 4, 11, 32, 694, 766
Freedman, J. L., 9, 766
Friedman, J., 507, 510–511, 527, 671, 698, 766
Friendly, M., 54, 191–194, 199, 206, 244,
681, 766
Fuller, W. A., 472, 766
Furnival, G. M., 360, 766,

Gallant, A. R., 527, 766
Gass, S. I., 598,
Gelman, A., 621, 742, 766, 771
Gentleman, R., xx, 767
Gijbels, I., 555, 585, 765
Gillings, D. B., 11, 768
Goldberger, A. S., 117–118, 373, 766
Goodall, C., 604, 766
Gould, P., 512, 762
Gould, S. J., 11, 766
Graham, J. W., 612, 646, 770
Green, P. E., 264, 766
Green, P. J., 550, 766
Greene, I., 4–5, 691, 766
Greene, W. H., 244, 415, 417, 501, 527,
638, 643, 766

Griffiths, W. E., 479, 483, 487, 492, 494–495,
50, 767
Grosse, E., 549, 585, 764
Groves, R. M., 472, 766

Hackney, O. P., 177, 186, 770
Hadi, A. S., 277, 279–282, 295, 763
Hand, D. J., 227, 766
Harding, S. A., 698, 762
Härdle, W., 766
Harrell, F. E., 527, 670, 766
Hartnagel, T. F., 489–491, 493, 498–499, 765
Harvey, A., 479, 493, 495, 501, 766
Hastie, T. J., 54, 507, 510–511, 527, 548–549,
569, 576, 585, 671, 698, 763–764, 766
Hauck, W. W., Jr., 382, 766
Hausman, J. A., 732–732, 767
Heath, A., 394, 762
Heckman, J. J., 629, 632–638, 642–643,
645–646, 761
Henderson, H. V., 177, 770
Hernandez, F., 324, 767
Hill, E. C., 479, 484, 487, 492, 494–495,
501, 767
Hill, J., 621, 742, 766, 771
Hinde, J., 472, 762
Hinkley, D. V., 664, 668, 764–765, 768
Hinz, T., 227, 762
Hoaglin, D. C., 54, 80, 273, 514, 604,
765–768, 771
Hocking, R. R., 173, 177, 186, 244, 364,
767, 770
Hoerl, A. E., 361–364, 367, 767
Hoeting, J. A., 685, 767
Holland, P. W., 8, 11, 767
Holt, D., 472, 770
Honaker, J., 621, 646, 767
Hosmer, D. W., 605, 767
Hotelling, H., 348, 767
Huber, P. J., 303, 305, 586, 588–595, 602–603,
659–661, 665–666, 767

Ihaka, R., xx, 767
Imbens, G. W., 232, 762

Jacoby, W. G., 54, 767
Johnson, M. E., 322, 764
Johnson, M. M., 322, 764
Johnson, N. L., 630, 762, 767–768, 771
Johnson, R., 79, 772
Johnson, R. A., 324, 767
Johnston, J., 241, 767
Joseph, A., 621, 646, 767
Judge, G. G., 479, 483, 487, 492, 494–495,
501, 767

- Kaiser, M. K., 227, 769
Kam, C.-M., 625, 764
Kass, R. E., 677, 680, 685, 698, 767
Katzman, D. K., 718, 721–722, 725, 764
Kennard, R. W., 361–364, 367, 767
Kenward, M. G., 724–725, 738, 741, 767
Kim, S.-H., 63, 762
King, G., 621, 646, 767
Kish, L., 11, 767
Kleiner, B., 54, 763
Kmenta, J., 336, 768
Koch, G. G., 11, 768
Koenker, R., 597–598, 600, 604, 768
Kostecki-Dillon, T., 745, 768
Kotz, S., 630, 762, 767–768, 771
Krupat, E., 164–165, 167, 174–175, 188, 190,
 198, 200, 240, 768
Krzanowski, W. J., 244, 768
Kuh, E., 276–277, 279, 281–282, 293, 295, 356,
 364, 763
Laird, N. M., 616, 702, 704, 709–710, 712, 714,
 718, 721, 734, 740, 764, 768
Lambert, D., 433, 768
Landwehr, J. M., 454, 768
Lee, T.-C., 479, 483, 487, 492, 494–495, 501, 767
Leinhardt, S., 67, 768
Lemeshow, S., 605, 767
Lepkowski, J. M., 472, 621, 766, 769
Leroy, A. M., 596, 600, 604, 770
Lewandowsky, S., 50, 770
Lewis, T., 295, 763
Li, G., 594, 604, 768
Li, K. C., 333, 764
Little, R. J. A., 607, 610, 615–616, 620,
 646, 768
Long, J. S., 305, 406, 417, 432, 472, 768, 770
Lütkepohl, H., 479, 483, 487, 492, 494–495,
 501, 767
Madigan, D., 685, 687, 767–768
Mallows, C. L., 316–317, 337, 672, 674, 694,
 696, 768
Mandel, J., 356, 768
Manski, C., 21, 768
Mare, R. D., 646, 772
McCallum, B. T., 241, 768
McCullagh, P., 403, 432, 443, 448, 471, 768
Miller, W. E., 408, 435, 763
Milliken, G. A., 198, 770
Monette, G., 222–224, 244, 357–358, 745,
 751, 755, 757, 765–766, 768, 772
Monlezun, C. J., 177, 770
Moody, E. A., 687, 768
Moore, D. S., xxi, 4, 768
Moore, J. C., Jr., 164–165, 167, 174–175, 188,
 190, 198, 200, 240, 768
Morgan, J. N., 181, 762
Morgan, S. L., 11, 768
Morrison, D. F., 244, 349, 769
Mosteller, F., 54, 64–66, 79–80, 604, 687, 698,
 765–769
Murnane, R. J., 11, 769
Nagler, J., 415, 762
Nason, G. P., 529, 769
Nelder, J. A., 144, 173, 418, 432, 443, 447–448,
 471, 768, 769
Nerlove, M., 397, 769
Newey, W. K., 488–489, 499–500, 769
Notz, W. I., xxi, 4, 768
Obenchain, R. L., 364, 769
O'Brien, R. G., 227, 769
Ornstein, M. D., 8, 46–47, 70–71, 427–430,
 432–433, 453, 455–457, 459, 464, 602, 769
Oudshoorn, K., 621, 771
Pagan, A. R., 329, 763
Pearl, J., 11, 769
Pearson, K., 348, 769
Pickup, M., 501, 769
Pinheiro, J. C., 750, 769
Pisani, R., 4, 766
Powers, D. A., 417, 435, 769
Pregibon, D., 454, 768–769
Press, S. J., 397, 769
Price, B., 357, 763
Purves, R., 4, 766
Putter, J., 257, 769
Raftery, A. E., 361, 674, 677, 679–680, 685, 687,
 698, 767–769
Raghunathan, T. E., 621, 769
Rao, C. R., 212, 226, 244, 451, 770
Raudenbush, S. W., 704, 710, 727, 742,
 758, 770
Relles, D. A., 646, 771
Riani, M., 286, 288, 294–295, 762
Robb, R., 636, 767
Roger, J., 724–725, 738, 741, 767
Ronchetti, E., 600–601, 763
Rousseeuw, P. J., 596, 600, 604, 770
Rubin, D. B., 11, 232, 606–607, 610, 614–616,
 620–621, 623–624, 644, 646, 762, 764,
 768, 770
Ruppert, D., 594, 771
Sall, J., 291–292, 770
Satterthwaite, F. E., 725, 738, 741, 770

- Schafer, J. L., 612, 619, 621, 624–625, 628, 646, 764, 770
- Scheffé, H., 139, 222, 364, 770
- Scheve, K., 621, 646, 767
- Schwarz, G., 673, 770
- Searle, S. R., 172, 177, 186, 198, 211, 243–244, 770
- Seber, G. A. F., 213, 218, 243, 770
- Shaffer, P., 4–5, 691, 766
- Shoemaker, A. C., 454, 768
- Shryock, H. S., 515, 770
- Shyu, W. M., 549, 585, 764
- Siegel, J. S., 515, 770
- Silverman, B. W., 34–35, 529, 550, 649, 766, 769, 770
- Simonoff, J. S., 555, 585, 770
- Simpson, E. H., 129, 770
- Singer, E., 472, 766
- Sinnott, R., 394, 762
- Skinner, C. J., 472, 770
- Smith, G. L., 698, 762
- Smith, H., 364–365, 527, 764
- Smith, T. M. F., 472, 770
- Snell, E. J., 409, 417, 764, 765, 768
- Snijders, T. A. B., 741, 770
- Solenberger, P., 621, 769
- Sonquist, J. A., 181, 762
- Soroka, S., 461, 765
- Speed, F. M., 173, 177, 186, 198, 767, 770
- Spence, I., 50, 770
- Stanley, J. C., 11, 763
- Stefanski, L. A., 597, 770
- Steinhorst, R. K., 177, 770
- Stine, R. A., 54, 343, 656, 668, 670–671, 698, 765, 770–771
- Stokes, D. E., 408, 435, 763
- Stolle, D., 461, 765
- Stolley, P. D., 11, 771
- Stolzenberg, R. M., 521, 646, 765, 771
- Stone, M., 264, 771
- Street, J. O., 594, 771
- Stroup, W. W., xv, 738, 758, 771
- Su, Y.-S., 621, 771
- Suschnigg, C., 20, 765
- Taylor, C. C., 227, 766
- Theil, H., 234, 257, 356, 364, 771
- Thompson, M. E., 11, 771
- Tibshirani, R. J., 507, 510–511, 527, 548–549, 569, 576, 585, 652, 656, 664, 666, 668, 671, 698, 765–766
- Tidwell, P. W., 326–329, 338–339, 457, 526, 763
- Tierney, L., 54, 80, 771
- Tobin, J., 638–639, 648, 771
- Torgerson, W. S., 63, 771
- Tourangeau, R., 472, 766
- Trivedi, P. K., 464, 472, 763
- Trussell, H. J., 273, 763
- Tufte, E. R., 29, 54, 771
- Tukey, J. W., 30, 41, 44, 54, 57, 64, 66, 70, 74, 79–80, 275, 302, 514, 588, 604, 636, 646, 664, 687, 698, 762, 765–769, 771
- Tukey, P. A., 54, 763
- Tversky, A., 63, 764
- Ullah, A., 364, 771
- van Buuren, S., 621, 646, 771
- Van Hoewyk, J., 621, 769
- Van Nostrand, R. C., 364, 764
- Veall, M. R., 383, 771
- Velilla, S., 77, 771
- Velleman, P. E., 54, 80, 282, 293, 771
- Vinod, H. D., 364, 771
- Volinsky, C. T., 685, 767
- Wahba, G., 540, 771
- Wainer, H., 636, 646, 767, 771–772
- Wang, P. C., 454–455, 772
- Ware, J. H., 702, 704, 709–710, 712, 714, 718, 721, 734, 740, 768
- Wasserman, S. S., 67, 768
- Watson, G. S., 492–494, 498, 500, 765
- Watts, D. G., 527, 763
- Weakliem, D., 698, 772
- Wedderburn, R. W. M., 418, 447–448, 769, 772
- Weiner, N. I., 751, 755, 757, 772
- Weisberg, S., xx–xxi, 50, 54, 80, 282, 286, 290, 294, 314, 329–333, 340, 451, 764, 766, 772
- Welsch, R. E., 273, 276–277, 279, 281–282, 293, 295, 356, 369, 763, 767, 771
- West, K. D., 488–489, 499–500, 769
- White, H., 305–307, 330, 488, 643, 772
- Wilks, A. R., 54, 763
- Willett, J. B., 11, 769
- Williams, D. A., 454–455, 772
- Wilson, R. W., 360, 766
- Winship, C., 11, 646, 768, 772
- Wong, P. P., 745, 751, 755, 757, 768, 772
- Wonnacott, R. J., 212, 264, 501, 772
- Wonnacott, T. H., 212, 264, 501, 772
- Wood, S. N., 540, 550, 574–575, 585, 772
- Wu, L. L., 604, 772
- Xie, Y., 417, 435, 769
- Yajima, M., 621, 771
- Yates, F., 175–176, 702, 772
- Yeo, L., 79, 324, 772
- Young, G. A., 649, 668, 770, 772
- Zimmermann, K. F., 383, 771

Subject Index

- Adaptive-kernel density estimator, 36–37
Added-variable plots, 50, 104, 282–286
 and collinearity, 343
 for constructed variables, 324–328, 457–458
 for generalized linear models, 455
Additive regression models, 563–566
 fitting, 566–567
Additive relationships, 128–132, 135–137, 160–162
Adjusted means, 150–151, 197–198
 See also Effect displays
Akaike information criterion (AIC), 673–677, 723
 bias-corrected (AICc), 676–677
 and model averaging, 695
Analysis of covariance (ANCOVA), 187–190, 730–731
 model for, 188–189
 See also Dummy-variable regression
Analysis of deviance, 384, 386, 394, 404, 410–411, 425–426, 429, 449–450
Analysis of variance (ANOVA):
 higher-way, 180–186
 one-way, 153–159
 random-effects one-way, 710–712
 for regression, 89, 98–99, 115, 248–249, 253
 table, 115, 146, 148–149, 159–160, 172–173, 180
 three-way, 177–180
 two-way, 159–177
 use of, to test constant error variance, 322–323
 vector geometry of, 258–260
AR(1). *See* Autoregressive process, first-order
AR(p). *See* Autoregressive process, higher-order
Arcsine-square-root transformation, 74
ARMA. *See* Autoregressive-moving-average process
Assumptions of regression model. *See* Constant error variance; Independence; Linearity; Normality
Asymptotic standard errors:
 for Box-Cox transformation, 78
 for effect displays, 453
 for generalized linear model, 425, 448
 for instrumental-variables estimator, 234
 for logit model, 382, 390, 414
 for M estimator, 594
 for nonlinear functions of parameters, 452–452
 for nonlinear least squares, 519
 for polytomous logit model, 398
 in quantile regression, 598
 for weighted least squares, 304, 335
Asymptotic growth model, 751
Asymptotic variance-covariance matrix.
 See Asymptotic standard errors
 See also Standard errors; Variance-covariance matrix
Attenuation due to measurement error, 122
Autocorrelated errors. *See* Serially correlated errors
Autocorrelation, 477–479, 481–484, 496
 Of errors in linear mixed model, 722
 partial, 485
 of residuals, 487, 491
Autocovariance, 478
Autoregressive process:
 continuous first-order, 722–723
 first-order [AR(1)], 477–481, 485–487, 722, 722
 higher-order [AR(p)], 481–482, 485
Autoregressive-moving-average (ARMA) process, 482–484, 487, 496
Average squared error (ASE) of
 local regression, 539
 mean average squared error (MASE), 541
Backfitting, 566–568
Backward elimination, 359
Balanced data in ANOVA, 174–175, 197–198, 293
Bandwidth in nonparametric regression, 530, 534, 539–541
 See also Span; Window
Bartlett's test for constant error variance, 322

- Baseline category:
 in dummy regression, 131–132, 135–136,
 138–139, 144–145
 in polytomous logit model, 393
- Basis of model matrix in ANOVA, 205–208
- Bayes factor, 677–681, 686
- Bayesian information criterion (BIC), 360,
 673–675, 677–689, 723
- Best linear unbiased estimator (BLUE), 212–213,
 733, 738
See also Gauss-Markov theorem
- Best linear unbiased predictor (BLUP), 719,
 733–734, 736, 738–739
- Bias:
 bootstrap estimate of, 665–666
 of maximum-likelihood estimators of variance
 components, 711
 measurement error and, 122
 in nonparametric regression, 21–23, 536–538
 of ridge estimator, 362–363
 and specification error, 119–120, 129
- Biased estimation, 361–365
- BIC. *See* Bayesian information criterion
- Binary vs. binomial data, 412, 419
- Binomial distribution, 418, 421–422, 443–444,
 450, 466–467, 743–744
- Bins, number of, for histogram, 32
- Bisquare (biweight) objective and weight
 functions, 588–592
- Bivariate-normal distribution. *See* Multivariate-
 normal distribution
- Bonferroni outlier test, 274, 455
- Bootstrap:
 advantages of, 647
 barriers to use, 663–664
 bias estimate, 665–666
 central analogy of, 651
 confidence envelope for studentized residuals,
 300–301
 confidence intervals, 655–658
 hypothesis tests, 660–662
 for mean, 647–654
 parametric, 300–301
 procedure, 653–655
 for regression models, 658–660
 standard error, 653–655
 for survey data, 662–663
 for time-series regression, 666
- Boundary bias in nonparametric regression, 23,
 530, 536
- Bounded-influence regression, 595–597
- Box-Cox transformations. *See* Transformations,
 Box-Cox
- Boxplots, 41–44
- Box-Tidwell transformations, 326–328, 338,
 457, 526
- Breakdown point, 595–596
- “Bubble plot” of Cook’s D-statistic,
 277–278
- “Bulging rule” to select linearizing
 transformation, 64–67
- Canonical parameter, 443
- Case weights, 461–462, 662–663, 666
- Causation, 3–8, 117–120, 126, 232
- Censored normal distribution, 629–632, 642
- Censored regression, 637–639
- Censoring, 605, 629–631
- Centering, 217, 357, 503, 522, 532, 553, 706, 708,
 727–730
- CERES plots, 318, 456–457
- Clusters:
 in mixed-effects models, 700–703
 in survey sampling, 461–462, 662
- Collinearity, 94, 97, 112–113, 208–209
 and ANOVA models, 156–157, 259
 detection of, 341–358
 in dummy regression, 136
 estimation in presence of, 358–366
 in Heckman’s selection-regression
 model, 635
- Influence on, 280
- in logistic regression, 388
- in model selection, 672
- in time-series regression, 341, 495
- vector geometry of, 253, 259, 261
- Comparisons, linear. *See* Contrasts, linear
- Complementary log-log link, 419, 420
- Component-plus-residual plots, 308–312
 augmented, 316–318
 effectiveness of, 314–316, 336–337
 for generalized linear models, 456–458
 “leakage” in, 317
 for models with interactions, 313–314
- Compositional effect, 729
- Compositional variable, 708, 710, 729–730
- Condition index, 356–357
- Condition number, 356
- Conditionally undefined data, 606
- Conditioning plot (coplot), 51–53, 133–134,
 559–562
- Confidence ellipse (and ellipsoid). *See* Confidence
 regions, joint
See also Data ellipse, standard
- Confidence envelope:
 for nonparametric regression, 542–544, 555,
 559, 576
 for quantile-comparison plots, 39, 300–301

- Confidence intervals:
- bootstrap, 655–658
 - for Box-Cox transformation, 325
 - for effect displays, 453
 - generating ellipse, 221–223, 238–239
 - for generalized linear model, 426
 - jackknife, 664–665
 - for logit models, 382–383
 - and missing data, 609, 612–613
 - for multiple imputation, 622
 - for nonlinear function of parameters, 451
 - for regression coefficients, 111, 114, 216, 221–222, 238
 - and ridge regression, 364
 - and variance-inflation factor, 342–343
- Confidence regions, joint, 220–224, 279, 343, 358, 382, 390
- Consistency:
- of least-squares estimator, 230, 301
 - and missing data, 609, 614, 629, 633–634
 - of nonparametric regression, 21–22
 - of quasi-likelihood estimation, 448
- Constant error variance, assumption of, 107, 112, 156, 203
- See also* Nonconstant error variance
- Constructed variables, 324–328, 457–458
- Contextual effects, 702, 708, 710, 729
- model, 702, *See also* Linear mixed-effects model
- Contextual variable, 708, 710
- Contingency tables:
- logit models for, 408–410, 441–442
 - log-linear models for, 434–442
- Continuation ratios, 400
- See also* Dichotomies, nested
- Contour plot of regression surface, 558–559, 561
- Contrasts, linear, in ANOVA, 190–194, 198–200, 206–208, 236
- Cook's distance (Cook's D), 276–277, 282, 291, 455
- Coplot. *See* Conditioning plot
- Correlation:
- intraclass, 711
 - multiple, 99–100, 253
 - multiple, adjusted for degrees of freedom, 100, 671–672, 694–695
 - multiple, for generalized linear models, 426
 - multiple, for logit models, 383
 - partial, 104–105, 261, 485
 - simple, 88–92
 - vs. slope, 91
 - vs. standard error of regression, 88
 - vector geometry of, 248–250, 253–254
- Correlogram, 491
- See also* Autocorrelation
- Cosine of angle between vectors and correlation, 248–250
- Covariates, 187
- COVRATIO, 279–280, 282
- Cross-validation (CV):
- generalized (GCV), 540–541, 554, 574, 673
 - and model selection, 673
 - to select span in nonparametric regression, 539–542, 554
 - vs. validation, 690
- Cumulative distribution function (CDF), 37–38, 375
- Curvilinear relationship, 28–29, 64
- See also* Nonlinearity
- "Curse of dimensionality", 22, 556–557
- Cutoffs, relative vs. absolute, for diagnostics, 281
- Data craft, 13
- Data ellipse (and ellipsoid), standard, 222–224, 271, 728–730
- Degrees of freedom:
- in ANOVA, 160, 170, 173, 179–180
 - in dummy regression, 138, 148–149
 - in estimating the error variance, 257–258
 - multiple correlation corrected for, 100, 671–672, 694–695
 - in nonparametric regression, 541, 546–549, 554–556, 569
 - in regression analysis, 87, 98, 115–116, 215–216
- Satterthwaite, 725, 738
- for studentized residuals, 272–273
 - vector geometry of, 250–252, 256
- Delta method, 451–452
- Density estimation, 33–37
- Design-based vs. model-based inference, 11, 460
- Design matrix, 203
- See also* Model matrix
- Deviance:
- components of, 412–413
 - residual, 383–384, 412, 414, 425–426, 431, 449, 466–467, 574
 - scaled, 426, 449
- See also* Analysis of deviance
- Deviation regressors, 158, 171, 178, 186, 189–191, 204–206, 410
- DFBETA and DFBETAS, 276–277, 455–456
- DFFITS, 277, 282, 291
- Diagnostic methods. *See* Autocorrelation; Collinearity; Generalized linear models, diagnostic methods for; Influential

- observations; Leverage of observations;
- Nonconstant error variance; Nonlinearity;
- Non-normality of errors; Outliers
- Dichotomies, nested, 399–400, 407–408
- Dichotomous explanatory variables in dummy regression, 128–132, 140–145
- Dichotomous response variables. *See* Linear-probability model; Logit models; Probit models
- Discrete explanatory variables, 318–323
 - See also* Analysis of variance; Contingency tables; Dummy-variable regression
- Discreteness and residual plots, 457
- Dispersion parameter:
 - estimation of, 425–426, 432, 447–448, 450, 454, 749
 - in generalized linear mixed model, 744, 748
 - in generalized linear model, 421–422, 424, 426, 431, 443, 446, 449
 - in generalized nonparametric regression, 573
- Dummy-variable regression:
 - and analysis of covariance, 187–188
 - and analysis of variance, 153–154, 157, 166–168
 - collinearity in, 136
 - for dichotomous explanatory variable, 128–133, 140–145
 - interactions in, 140–149
 - and linear contrasts, 190
 - model for, 130, 142
 - model matrix for, 203–204
 - for polytomous explanatory variable, 133–138, 145
 - and semiparametric regression, 570
 - and standardized coefficients, misuse of, 149–150
 - and variable selection, caution concerning, 361
- Durbin-Watson statistic, 492–493, 498
- Effect displays:
 - in analysis of variance, 183–186
 - in dummy regression, 146–147, 149
 - following transformations, 311–314, 459, 511
 - for generalized linear models, 429–431, 453, 459
 - for logit models, 386–387, 396, 404–407, 505–506
 - for mixed-effects models, 717, 724–725, 755
 - in quantile regression, 599
- Elasticity, and log transformations, 69
- Ellipse (ellipsoid). *See* Confidence regions, joint;
 - Data ellipse (and ellipsoid), standard
- EM (expectation-maximization) algorithm, 616–618, 628, 641
- Empirical best linear unbiased predictor (EBLUP), 719, 733
- Empirical cumulative distribution function (ECDF), 37–38
- Empirical vs. structural relations, 117–120, 229–230
- Empty cells in ANOVA, 186–187
- Equal cell frequencies in ANOVA, 174–175, 197–198, 293
- Equivalent kernels in nonparametric regression, 580–581
- Error variance, estimation of:
 - in instrumental-variables estimation, 233–234, 235
 - in linear regression and linear models, 98, 111, 114–115, 117, 124, 149, 167, 214–215, 218, 256–258, 700
 - in nonlinear regression, 519
 - in nonparametric regression, 543–546, 548, 555–556, 566, 569, 571
 - in quantile regression, 598
 - in two-stage least-squares estimation, 235
- Essential nonlinearity, 515
- Estimating equations:
 - for 2SLS regression, 234
 - for additive regression model, 567–569
 - for generalized linear model, 445–446, 448
 - for logit models, 389–391, 397–398, 415
 - for mixed-effects models, 736–737, 739
 - for nonlinear regression model, 516
 - for robust estimation, 590–591, 593, 600
 - See also* Normal equations
- Expected sums of squares, 218–219
- Experimental vs. observational research, 4–8
- Exponential families, 418, 421–425, 443–445, 466–467, 743–745, 743–744
 - See also* Normal (Gaussian) distribution; Binomial distribution; Poisson distribution; Gamma distribution; Inverse-Gaussian distribution
- Extra sum of squares. *See* Incremental sum of squares
- Factors, defined, 128
- Fences, to identify outliers, 42–43
- Finite population correction, 461
- Fisher’s method of scoring, 447
- Fitted values:
 - in 2SLS, 235
 - and bootstrapping, 658–659
 - in generalized linear models, 453

- in linear regression and linear models, 83, 86–89, 92–93, 96–97, 146–147, 170, 184–185,
 in logit models, 389, 391, 404–405
 and missing data, 618–620
 and model selection, 672–673
 in multivariate linear model, 225
 in nonparametric regression, 23–24, 529–530, 532–534, 537–540, 543, 546–547, 551, 554–555, 574, 576, 580
 and regression diagnostics, 270, 289, 291, 302–303
 vector geometry of, 247–248, 252–253, 256, 259, 261
- Fitting constants, Yates's method of, for ANOVA, 176
- Five-number summary, Tukey's, 41–42
- Fixed effects, 703, 710, 755
 models, 730, 739
- Fixed explanatory variables, 108, 112, 187, 658–659, 665
- Forward search, 286–287
- Forward selection, 359
- F-tests:
 in analysis of variance, 154, 157–158, 160, 167, 173, 180, 190–191, 194–195
 for constant error variance, 322–323
 for contrasts, 194
 and Cook's D-statistic, 276, 282, 291
 in dummy regression, 132, 137–138, 146, 148–149
 for general linear hypothesis, 219, 450
 for generalized linear models, 426, 432, 449–450
 and joint confidence regions, 220–222
 for linear mixed models, 737–738
 and Mallows's C_p -statistic, 672, 694
 for multiple imputation, 624–625
 in multiple regression, 115–117, 217–219, 254–256, 545
 for nonlinearity (lack of fit), 319–322
 for nonparametric regression, 545–546, 549, 555–556, 561, 565–566, 569–571
 step-down, for polynomial regression, 322, 522
 vector geometry of, 254–256, 261
- Gamma distribution, 418, 421–422, 424, 426, 432, 444, 466–467, 743
- Gamma function, 422, 444
- Gaussian distribution. *See* Normal distribution
- Gaussian (normal) kernel function, 35–36, 529–530, 538
- Gauss–Markov theorem, 110, 212–213, 231, 238, 297, 335, 476, 496, 733
- Gauss–Newton method, for nonlinear least squares, 518, 520–521
- Generalized additive models (GAMs), 576–578
- Generalized cross validation (GCV), 540–542, 554, 574, 673, 694–695
- Generalized least squares (GLS), 475–476, 485–487, 496
 bootstrapping, 666
 empirical (EGLS), 487
 limitations of, 494–495
 and mixed-effects models, 702, 733, 736, 739
See also Weighted least squares
- Generalized linear model (GLM), 418–420, diagnostic methods for, 453–460
 robust estimation of, 600–601
 saturated, 425–426, 437–442
See also Logit models; Poisson regression; Probit models
- Generalized linear mixed-effects model (GLMM), 743–744, 748
 estimation of, 748–749
- Generalized variance, 279
- Generalized variance-inflation factor (GVIF), 357–358, 367, 459–460, 635, 647
- General linear model, 202, 212, 289, 502–503, 700
 Multivariate, 225–227, 240
 vs. generalized linear model, 418
See also Analysis of covariance; Analysis of variance; Dummy-variable regression; Multiple-regression analysis; Polynomial regression; Simple-regression analysis
- General nonlinear model, 515
- Geometric mean, 37, 77, 325
- Global (unit) nonresponse, 461, 605
- GLS. *See* Generalized least squares
- Gravity model of migration, 512–513, 523–525
- Greek alphabet, 760
- Hat-matrix, 289–290, 293, 298, 454, 547–548
- Hat-values, 270–273, 277–281, 289–290, 293, 305, 454–456, 600
- Hausman test, 731–732
- Heavy-tailed distributions, 16–17, 39, 41, 297–299, 586, 601
- Heckman's selection-regression model, 632–634
 cautions concerning, 636
- Heteroscedasticity. *See* Non-constant error variance;
See also Constant error variance; Weighted least squares;
- “White” corrected (White-Huber) standard errors

- Hierarchical data, 700–701
modeling, 704–717
- Hierarchical linear model, 702
See also Linear mixed-effects model
- Higher-way ANOVA. *See* Analysis of variance, higher-way
- Hinges (quartiles), 39, 42–44, 60–61, 70, 597
- Hinge-spread (interquartile range), 32, 36, 39, 43, 70–71, 101, 553
- Histograms, 14, 30–34
See also Density estimate; Stem-and-leaf display
- Homoscedasticity. *See* Constant error variance
See also Non-constant error variance
- Hotelling-Lawley trace test statistic, 226
- Huber objective and weight functions, 588–589, 591–592
- Hypothesis tests:
in ANCOVA, 190
in ANOVA, 154, 157–158, 160, 167, 173, 180, 190–191, 194–195
Bayesian, 677–678
bootstrap, 660–662
for Box-Cox transformation, 78, 325
for Box-Tidwell transformation, 327
for constant error variance, 322–323, 329–331
for contrasts, 190–194, 198–200, 206–208
for difference in means, 194
in dummy-variable regression, 135–136, 138, 142, 146, 148–149
for equality of regression coefficients, 124, 220, 364–365
for general linear hypothesis, 219–220, 226–227, 291–293, 390, 450, 737
for general nonlinear hypothesis, 451–452
in generalized linear models, 425–426, 437–438, 440–442, 448–452
impact of large samples on, 670
for “lack of fit”, 318–322, 545–546, 570–571
for linearity, 318–322, 545–546, 570–571
in logit models, 382–383, 390
in mixed-effects models, 713–714, 724–726, 731–732, 737–738
for multiple imputation, 621–625
in multivariate linear model, 225–227
in nonparametric regression, 545–546, 555–556, 570–571, 574
for outliers, 273–274
for overdispersion, 464
for least-squares regression coefficients, 111, 113–117, 124, 215–220, 228, 254
for serially correlated errors, 492–493
“step-down”, for polynomial terms, 322, 503, 522
See also F-tests; Likelihood-ratio test; score test; t-tests; Wald tests
- Identity link function, 419, 421, 443, 449
- Ignorable missing data, 607, 609, 616, 625, 629, 633
- Ill conditioning, 356, 362
See also Collinearity
- Incremental sum of squares:
in ANOVA, 167, 172–174, 176–177, 180, 190, 239–240
in dummy regression, 132, 136–138, 146, 148–149
for equality of regression coefficients, 124
in least-squares analysis, 116–117, 217–218
for linear hypothesis, 218
for nonlinearity, 318–320
in nonparametric regression, 545–546, 549, 555–556, 561, 569, 571
vector geometry of, 254, 261
See also F-tests
- Incremental sum-of-squares-and-products matrix, 225–226
- Independence:
assumption of, 16, 108–110, 112, 123–124, 128, 156, 203, 211–212, 214, 225, 229–230, 240, 257, 297, 304–306, 324, 326, 346, 381, 389, 397, 401, 418, 445, 460–462, 474, 477, 479, 488, 502, 586, 662–663, 666, 700–703, 718, 734, 743–744, 749–750
of nested dichotomies, 400
from irrelevant alternatives, 415
- Independent random sample, 16, 460–461, 647, 654, 662–663
- Index plots, 271–272, 276–278
- Indicator variables, 130
for polytomous logit model, 397
See also Dummy-variable regression
- Indirect effect. *See* Intervening variable
- Influence function, 587–590
- Influential observations, 29, 276–289, 290–293
- Information matrix for logit models, 389–390, 398, 414–415
- Initial estimates (start values), 391, 447, 516–517, 519–520, 591, 593, 597, 752–753
- Instrumental-variables (IV) estimation, 126, 231–234
- Intention to treat, 240
- Interaction effects:
in ANCOVA, 188–190

- in ANOVA, 161, 163–164, 166–181
 and association parameters in log-linear models, 437
 and component-plus-residual plots, 313–314
 cross-level, in linear mixed-effects model, 709
 disordinal, 163–164
 distinguished from correlation, 140–141
 in dummy regression, 140–149
 in generalized linear models, 419
 linear-by-linear, 394
 in logit models, 380, 410
 and multiple imputation, 626
 in nonparametric regression, 559, 569, 571
 in polynomial regression, 504, 506
 and structural dimension, 332
 and variable selection, 361
See also Effect displays; Marginality, principle of
- Interquartile range. *See* Hinge-spread
- Intervening variable, 7, 120
- Intraclass correlation, 711
- Invariant explanatory variables, 7, 120
- Inverse link (mean) function, 419, 573, 576, 600, 743
- Inverse Mills ratio, 630–631, 633–635
- Inverse regression, 333
- Inverse-Gaussian distribution, 418, 421, 424–426, 444, 466–467, 743
- Inverse-square link function, 419, 421
- Invertibility of MA and ARMA processes, 483
- Irrelevant regressors, 6, 119, 125, 230
- Item nonresponse, 605
- Iteratively weighted (reweighted) least squares (IWLS, IRLS), 391, 447–448, 454–455, 457, 575–576, 590–591, 593
- Jackknife, 657, 664–665
- Joint confidence regions. *See* Confidence regions, joint
- Jointly influential observations, 282–286
- Kenward-Roger standard errors, 724–725, 738
- Kernel smoothing:
 in nonparametric density estimation, 34–37
 in nonparametric regression, 528–531, 536–539, 580–581
- Kullback-Leibler information, 675–676
- Ladder of powers and roots, 56–57
See also Transformations, family of powers and roots
- Lagged variables, 495
- Least-absolute-values (LAV), 84–85, 587–588, 591–592, 597–598
- Least squares:
 criterion, 84–85
 estimators, properties of, 109–110
 nonlinear, 515–519
 objective function, 587
 vector geometry of, 246–247, 252
See also Generalized least squares; Multiple-regression analysis; Ordinary least-squares regression; Simple-regression analysis; Weighted least squares
- Least-trimmed-squares (LTS) regression, 596–597, 602
- Levene's test for constant error variance, 322–323
- Leverage of observations. *See* Hat-values
- Leverage plot, 291–293
- Likelihood-ratio tests:
 for fixed effects estimated by REML, invalidity of, 714, 724
 for generalized linear model, 426, 449
 for generalized nonparametric regression, 574, 578
 for independence, 465–466
 for linear model, 217–218
 for logit models, 382, 384, 404, 410–412
 for log-linear models, 437–438, 440–442
 and missing data, 614
 for overdispersion, 464
 of proportional-odds assumption, 405–406
 to select transformation, 77–78, 324–325
 for variance and covariance components, 713–714, 721, 726, 746
See also Analysis of deviance
- Linear estimators, 109–110, 211–213, 297
- Linear hypothesis. *See* Hypothesis tests, for general linear hypothesis
- Linear model. *See* General linear model
- Linear predictor, 375, 380, 418–419, 429, 453, 505, 743, 748–749
- Linearity:
 assumption of, 16–17, 106–107, 109–110, 112, 211, 307–308
 among explanatory variables, 316–317
See also Nonlinearity
- Linear mixed-effects model (LMM) 702–704
 estimation of, 734–737
- Laird-Ware form of, 702–704, 709–710, 712, 714, 718, 721, 734
- Linear-probability model, 372–374
 constrained, 374–375
- Link function, 419–420, 743

- canonical, 421, 443–444, 446–447, 449
 vs. linearizing transformation of response, 421
See also Complementary log-log link function;
 Identity link function;
 Inverse link function; Inverse-square link
 function; Log link function;
 Logit link function; Log-log link function; Probit
 link function;
 Square-root link function
 Local averaging, in nonparametric regression, 22–23
 Local likelihood estimation, 572–574
 Local linear regression. *See* Local-polynomial
 regression
 Local-polynomial regression, 532–534, 550–557,
 573–574, 601
 Loess. *See* Lowess smoother
 Log odds. *See* Logit
 Logarithm, as “zeroth” power, 57
 Logit (log odds), 73–75, 377
 empirical, 309
 link function, 419–421
 Logit models:
 binomial, 411–413
 for contingency tables, 408–413, 441–442
 dichotomous, 375–383
 estimation of, 381, 389–392, 397–398, 412,
 414–415
 interpretation of, 377–378, 380
 and log-linear model, 441–442
 mixed-effects model, 745
 multinomial, 393, 413, 415, 442
See also Logit models, polytomous
 for nested dichotomies, 399–400,
 407–408
 nonparametric, 572–574
 ordered (proportional-odds), 401–403,
 406–408
 polytomous, 392–393, 397–398, 407–408, 415
 problems with coefficients in, 388
 saturated, 412
 unobserved-variable formulation of, 379, 401
 Logistic distribution, 375–376
 Logistic population-growth model, 515, 519–521
 Logistic regression. *See* Logit models
 Log-linear model, 434–441
 relationship to logit model, 441–442
 Log-log link function, 419–420
 Longitudinal data, 700–701, 703, 745
 modeling, 717–724
 Lowess (loess) smoother, 23, 532
See also Local-polynomial regression
 Lurking variable, 120
 M estimator:
 of location, 586–592
 in regression, 592–595
 MA. *See* Moving-average process
 Main effects, 144, 146, 148–150, 161–164,
 166–184, 186–190
 Mallows’s C_p-statistic, 672, 694
 MAR. *See* Missing data, missing at random
 Marginal means in ANOVA, 160
 Marginal vs. partial relationship, 48, 94, 122,
 129, 308
 Marginality, principle of, 144–145, 148–149, 164,
 167–168, 172–174, 177–178, 180–181, 184,
 187, 190, 384, 404, 410, 439, 503
 Marquardt method, for nonlinear least squares, 518
 MASE. *See* Mean average squared error
 Maximum-likelihood estimation:
 of Box-Cox transformation, 76–77, 324–326,
 337–338
 of Box-Tidwell transformation, 326–328, 338
 of constrained linear-probability model, 374
 EM algorithm for, with missing data, 616–618
 of error variance, 214–215, 217, 329, 700, 711
 of general nonlinear model, 416–419
 of generalized additive models, 575–576
 and generalized least squares, 475–476
 of generalized linear mixed-effects model,
 744, 748–749
 of generalized linear model, 425, 445–448
 of Heckman’s selection-regression model, 634
 of linear mixed-effects model, 711,
 736–737, 740
 of linear regression model, 110, 113,
 123–124, 214–215, 228–229, 700
 of logit models, 381, 389–391, 397–398,
 411–412, 414–415
 of log-linear models, 438
 with missing data, 613–619
 of multivariate linear model, 225, 240
 of nonlinear mixed-effects model, 756
 with random regressors, 228–229
 restricted (REML), 711, 737
 in time-series regression, 487, 498
 of transformation parameters in regression,
 323–329
 and weighted least squares, 304, 335
 of zero-inflated negative-binomial (ZINB)
 model, 465
 of zero-inflated Poisson (ZIP) model, 433–434
 MCAR. *See* Missing data, missing completely at
 random
 Mean average squared error (MASE) in local
 regression, 541–542

- Mean function, 331–332, 419, 743
Mean-deviation form, vector geometry of, 247–250
See also Centering
“Mean-shift” outlier model, 273
Mean-squared error:
and biased estimation, 361–363
and C_p -statistic, 672
and cross-validation, 673
of least-squares estimator, 110, 212
in nonparametric regression, 537, 539, 555
and outlier rejection, 274–275
of ridge estimator, 363
Mean squares, 115
Measurement error, 120–123, 125
Median, 18, 32, 39, 42–44, 60–61, 70–71, 322, 587–590, 595, 597, 601–602
Median absolute deviation (MAD), 588–589
Method-of-moments estimator of dispersion parameter, 425, 431–432, 447–448
Missing data:
available-case analysis (pair-wise deletion) of, 610–612
complete-case analysis (list-wise, case-wise deletion) of, 610–613
conditional mean (regression) imputation of, 611
missing at random (MAR), 606–614, 617, 619, 621, 625
missing completely at random (MCAR), 606–612, 614, 616, 629
multiple imputation of, 619–626
unconditional mean imputation of, 611
univariate, 607, 611, 640
Missing information, rate of, 622
MM estimator, 597
MNAR. *See* Missing data, missing not at random
Model averaging, 685–687
based on AIC, 695
comments on, 687–688
Model matrix, 203–204, 208, 210–211, 225, 227, 232, 259, 289, 389, 397, 447, 453, 593, 734–735, 748–749
row basis of, 205–206, 208, 236, 240–241, 260
Model respecification and collinearity, 359, 365
Model selection:
avoiding, 670
and collinearity, 359, 365
comments on, 683, 685
criteria for, 671–674
and fallacy of affirming the consequent, 669
vs. model averaging, 670
and simultaneous inference, 669
See also Akaike information criterion; Bayesian information criterion;
Correlation, multiple, adjusted for degrees of freedom; Cross validation;
Mallows’s C_p -statistic; Model averaging
Model validation, 690–691, 693
Modes, multiple, in error distribution, 16, 298
Moving-average process (MA), 482–483, 485–487, 496
Multicollinearity, 344
See also Collinearity
Multinomial distribution, 413, 415, 418, 437, 621
Multinomial logit model. *See* Logit models, multinomial; Logit models, polytomous
Multiple correlation. *See* Correlation, multiple
Multiple imputation of missing data, 619–626
Multiple outliers, 282
Multiple regression analysis, 92–98, 104, 112–117, 202–203, 212, 270
and instrumental-variables estimation, 232–234
model for, 112
nonparametric, 550–571
vs. simple regression analysis, 94
vector geometry of, 252–256
Multiple-classification analysis (MCA), 181
Multiplicative errors, 512–513, 515
Multistage sampling, 461–462
Multivariate linear models, 225–227, 640–641, 702
Multivariate logistic distribution, 377, 392
Multivariate-normal distribution:
Box-Cox transformation to, 76–78
EM algorithm for, 617–618
and likelihood for linear model, 214
of errors in linear model, 203, 225
multiple imputation for, 619–621, 625–626
nonignorable, 607, 616, 629
and polytomous probit model, 392
of random effects in the nonlinear mixed-effects model, 756
of regression coefficients, 211–212, 215
of response in linear model, 203
singular, of residuals, 257, 261
Negative binomial distribution, 418, 432
Negative-binomial regression model, 432–433
zero-inflated (ZINB), 465
Nested dichotomies, 399–400
Newey-West standard errors, 488–489, 499
Newton-Raphson method, 390–391, 447
Nonconstant error variance or spread, 17
and bootstrap, 659
correction for, 305–306

- detection of, 301–304
 and dummy response variable, 373, 413
 effect on OLS estimator, 306–307,
 335–336
 in linear mixed-effects model, 703
 and quantile regression, 599
 and specification error, 303, 335
 tests for, 322–323, 329–331
 transforming, 70–72, 301–303
 and weighted least squares (WLS),
 304–305, 335
- Nonignorable missing data, 607, 616, 629
- Nonlinear least squares, 515–519, 750
- Nonlinear mixed-effects model (NLMM), 750
 estimating, 755–756
- Nonlinearity, 17
 and correlation coefficient, 89–90
 detection of, 307–318, 456–459
 and dummy response variable, 373
 essential, 515
 monotone vs. nonmonotone, 64, 66
 and multiple imputation, 625–626
 tests for, 318–320, 545–546, 570–571
 transformable, 512–514
 transformation of, 63–66, 326–327,
 456–458
- See also* Linearity, assumption of; Nonlinear least squares; Nonparametric regression
- Non-normality of errors:
 detection of, 297–301
 and dummy response variable, 373
See also Normality; Skewness
- Nonorthogonal contrasts, 236
- Nonparametric regression:
 generalized, 572–578
 by local averaging, 22–23
 naive, 18–22
 obstacles to, 556–557
See also Kernel smoothing; Local-polynomial regression; Splines, smoothing
- Normal (Gaussian) distributions:
 family of, in generalized linear mixed model,
 743–744
 family of, in generalized linear model, 418,
 421–422, 426, 433, 444, 446, 449–450,
 466–467
 as kernel function, 34–36, 529–530, 538
 of regression coefficients, 110, 113, 215
 to transform probabilities, 74, 376–377,
 379, 401
See also Censored-normal distribution;
 Multivariate-normal distribution;
- Non-normality of errors; Normality, assumption of; Quantile-comparison plots;
 Truncated-normal distribution
- Normal equations, 85, 93, 96–97, 104, 125,
 208–210, 342
- Normality, assumption of, 16–17, 107, 109, 112,
 203, 212, 214, 275, 502, 515, 570, 632–633,
 638, 647
See also Non-normality of errors
- Normalization, in principal-components analysis, 350
- Normal-probability plots. *See* Quantile-comparison plots
- Notation, 759–761
- Objective function. *See* Least absolute values;
 Least squares criterion;
 Huber objective and weight functions;
 Biweight (bisquare) objective and weight functions
- Observation space, 246, 250–251, 256–258, 260
- Observational vs. experimental research, 4–8, 10
- Occam’s window, 687
- Odds, 377–378, 380, 385, 388, 402
 posterior, 677–678, 687
- Omnibus null hypothesis, 115, 154, 158, 218–219,
 228, 238, 382, 390, 660
- Omitted-variable bias. *See* Specification error
- One-way ANOVA. *See* Analysis of variance, one-way
- Order statistics, 37, 39, 60–61, 301, 598
- Ordinal data, 400–407
- Ordinary-least-squares (OLS) regression:
 and generalized-least-squares, 476,
 486–487, 494
 and instrumental-variables estimation,
 126, 241
 for linear-probability model, 373
 and nonconstant error variance, 305–307,
 335–336
 vs. ridge estimator, 363
 in time-series regression, 480–481, 497
 and weighted least squares, 304
See also Generalized least squares; Least squares; Multiple regression analysis;
 Simple regression analysis; Weighted least squares
- Orthogonal contrasts, 208, 236, 522
- Orthogonal data in ANOVA, 174–175, 197–198
- Orthogonal (uncorrelated) regressors, 255–256
 in polynomial regression, 522
- Orthonormal basis for error subspace,
 257–258, 262

- Outliers, 19, 23, 32, 42–43, 266–270, 272–274, 288–289, 298, 454–455, 586–589, 659
Anscombe’s insurance analogy for, 274–276
multivariate, 270–271
See also Unusual data, discarding
- Overdispersion, 431–434, 464
- Overfitting, 288, 690
- Parametric equation, in ANOVA, 205–206, 236, 259–260
- Partial autocorrelation, 485
- Partial correlation. *See* Correlation, partial
- Partial regression functions, 317, 563–564, 566–569, 575–576
- Partial vs. marginal relationship, 48, 94, 122, 129, 308
- Partial-regression plots. *See* Added-variable plots; Leverage plot
- Partial-residual plots. *See* Component-plus-residual plots
- Penalized sum of squares, 549
- Perspective plot of regression surface, 557–558, 561, 564–565
- Pillai-Bartlett trace test statistic, 226
- Poisson distribution, 418, 421–423, 426–435, 444, 464, 466–467, 743–744
and multinomial distribution, 437
- Poisson regression model, 427–430
zero-inflated (ZIP), 433–434
- Polynomial regression, 28, 64, 308, 311, 317, 320–322, 357, 451–452, 503–507, 522
piece-wise, 507–512, 523
See also Local-polynomial regression
- Polytomous explanatory variables in dummy regression, 133, 135–136, 138–139, 145
- Polytomous response variables, 392–408
- Prediction in regression, 239, 361, 625, 671–673, 677, 682–683, 685, 687
- Predictive distribution of the data, 619, 621, 628, 677
- Premium-protection approach to outliers, 274–275
- Principal-components analysis, 348–354, 366
and diagnosing collinearity, 356–357
- Prior cause, common, 7, 120
- Prior information and collinearity, 364–365
- Probit:
and Heckman’s selection-regression model, 633–634
link function, 419–420
models, 376, 379–380, 392, 399, 401, 415
transformation, 74–75
- Profile log-likelihood, 325–326
- Proportional-odds model, 400–403, 407–408
- Pseudoresponse variable in logit model, 391
- Pseudo-values in jackknife, 665
- Quadratic regression. *See* Polynomial regression
- Quadratic surfaces, 503–505
- Quantile function, 38
- Quantile regression, 597–598
- Quantile-comparison plots, 37–40, 274, 298–301, 655
- Quartiles. *See* Hinges
- Quasi-binomial models, 432
- Quasi-likelihood estimation, 431–432, 448–449, 744, 748
- Quasi-Poisson regression model, 431–432
- Quasi-variances of dummy-variable coefficients, 138–140, 467–468
- Random-coefficients regression model, 702, 712–714
See also Linear mixed-effects model
- Random effects, 700–701, 703, 710, 750, 755
crossed, 701
models, 702
See also Generalized linear mixed-effects model; Linear mixed-effects model; Nonlinear mixed-effects model
- Random explanatory variables, 108, 118, 227–230, 658, 655
- Random-intercept regression model, 727
- Randomization in experimental design, 4–6, 9, 153
- Raw moments, 233
- Rectangular kernel function, 530
- Reference category. *See* Baseline category
- Regression of X on Y, 91, 103
- Regression toward the mean, 103
- Regressors, distinguished from explanatory variables, 130, 142, 502
- Repeated-measures models, 227, 702
See also Linear mixed-effects model
- Residual standard error. *See* Standard error of the regression
- Residuals, 3, 83–85, 92–93, 208, 245, 247, 252–253
augmented partial, 317
deviance, 455
distribution of, 290
in generalized linear models, 454–455, 457
partial, 308–314, 316–317, 454, 457, 564, 567–568, 570
Pearson, 454
plot of, vs. fitted values, 302
quantile-comparison plot for, 298, 300–301

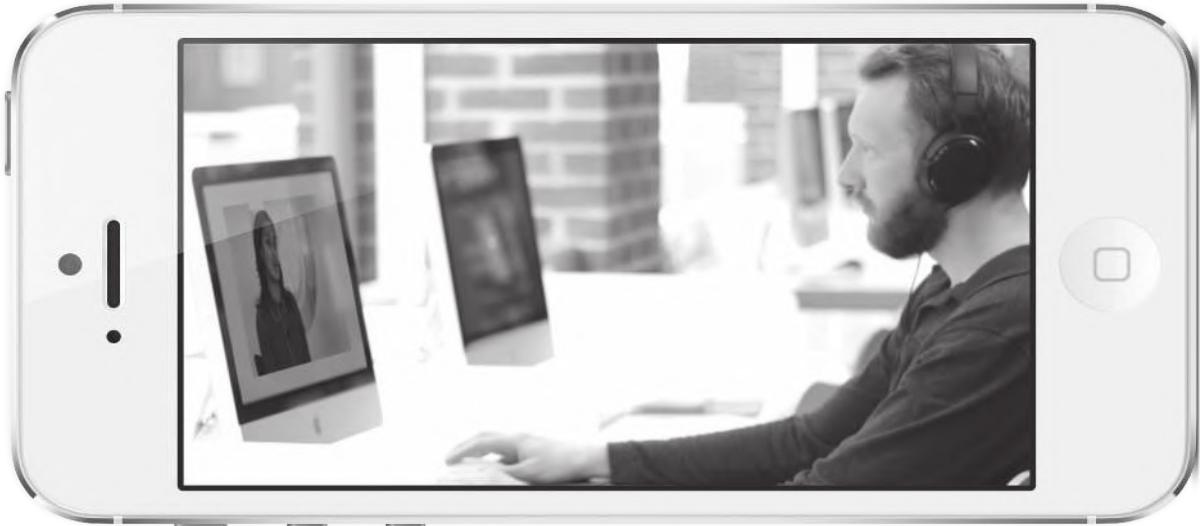
- response, 454
- standardized, 272–273, 275
- standardized deviance, 455
- standardized Pearson, 454–455
- studentized, 272–274, 280–281, 298–302, 455
- supernormality of, 301
- working, 454
- Resistance (to outliers), 85, 286, 586, 588–589, 600–601
- Restricted maximum likelihood (REML). *See* Maximum likelihood, restricted
- Restrictions (constraints) on parameters:
 - in ANCOVA, 189
 - in ANOVA, 157–158, 169, 177–178, 195, 204
 - in logit models for contingency tables, 410
 - in log-linear models, 436, 438
 - in polytomous logit model, 393
 - and ridge regression, 365
 - sigma, 157–158, 169, 178, 180, 186, 189, 195, 204–205, 240, 262, 393, 410, 436, 438, 442
- Ridge regression, 362–365, 367
- Ridge trace, 367
- Robust regression. *See* Generalized linear model, robust estimation of;
 - Least-trimmed-squares regression; M estimator; MM estimator; Quantile regression
- Robustness of efficiency and validity, 297
- Roy’s maximum root test statistic, 226
- Rug plot, 35
- Sampling fraction, 461
- Sampling variance:
 - of fitted values, 543
 - of the generalized-least-squares estimator, 476, 496
 - of least-squares estimators, 109, 113, 123, 212, 215, 237, 306, 342, 356, 363, 497
 - of the mean, 588
 - of the mean of an AR(1) process, 479
 - of the median, 588
 - of a nonlinear function of coefficients, 451
 - of nonparametric regression, 20, 537
 - of ridge-regression estimator, 363
 - of weighted-least-squares estimator, 336
- See also* Asymptotic standard errors; Standard errors; Variance-covariance matrix
- Sandwich coefficient covariance estimator, 305, 489
- Scatterplot matrices, 48–49, 333–334
- Scatterplots, 13–14, 44–45
 - coded, 50
 - jittering, 36
 - one-dimensional, 35
- smoothing, 23, 44–45, 528–550
- three-dimensional, 50–51
- vs. vector representation, 246, 260
- Scheffé intervals, 222
- Score test:
 - of constant error variance, 329–330
 - of proportional-odds assumption, 406
 - to select transformation, 324–325
- Scoring, Fisher’s method of, 447
- Seasonal effects, 479
- Semiparametric regression models, 569–571
- Separability in logit models, 388
- Serially correlated errors, 476–485
 - diagnosing, 489–493
 - effect on OLS estimation, 497
 - estimation with, 485–487
 - in mixed-effects models, 718, 746
- Sigma constraints. *See* Restrictions on parameters, sigma
- Simple random sample, 108, 460–462, 606, 647
- Simple regression analysis, 83–87, 106–112, and instrumental variable estimation, 126, 231–232
 - model for, 106–108, 245
 - vector geometry of, 245–252
- Simpson’s paradox, 129
- Skewness, 13, 16, 36, 39–41, 44, 72, 192, 297–298, 424
 - See also* Transformations, to correct skewness
- Smoother matrix, 546–548, 550, 555, 568–569
- Smoothing. *See* Density estimation; Lowess smoother; Local-polynomial regression; Scatterplots, smoothing; Splines, smoothing
- Span of smoother, 22–24, 530–532, 534–535, 538–544, 552, 554–556, 574, 579
 - See also* Bandwidth; Window
- Specification error, 118–119, 124–125, 229–230, 303, 335, 633, 670, 685
- Splines:
 - regression, 507–512, 523
 - smoothing, 549–550
- Spread-level plot, 70–71, 302–303
- Spurious association, 7, 120, 685
- Square-root link function, 419, 421
- SS notation for ANOVA, 172–175, 180, 240, 262
 - adapted to logit models, 410
- Standard error(s):
 - bootstrap, 653–655
 - of coefficients in generalized linear models, 425, 431
 - of coefficients in Heckman’s selection-regression model, 634, 643
 - of coefficients in logit models, 382, 388, 390

- of coefficients in regression, 111, 113–114, 215, 279, 284, 301
collinearity, impact of, on, 341
of differences in dummy-variable coefficients, 138–139, 467–468
of effect displays, 146, 186, 453
influence on, 277, 279
Kenward-Roger, 724–725, 738
of the mean, 648
and model selection, 670
from multiple imputations, 621–622
Newey-West, 488–489, 499
for nonlinear function of coefficients, 451–452
of order statistics, 39
of the regression, 87–88, 98, 272
of transformation-parameter estimates, 77
of two-stage least-squares estimator, 235
“White” corrected, 305
See also Asymptotic standard errors; Variance-covariance matrix
Standardized regression coefficients, 100–102, 105, 237
misuse of, 102, 149–150
“Start” for power transformation, 58–59, 79
Start values, *See* Initial estimates (start values)
Stationary time series, 476–477, 479–483, 498
Statistical models, limitations of, 1–4
Steepest descent, method of, for nonlinear least squares, 516–518
Stem-and-leaf display, 30–32
Stepwise regression, 359–360, 683
Stratified sampling, 461–462, 691
Structural dimension, 331–333, 338
Structural-equation models, 3, 123
Studentized residuals. *See* Residuals, studentized
Subset regression, 360, 367, 672
Sum of squares:
 between-group, 159
 for contrasts, 199, 208
 generalized, 475
 for orthogonal regressors, 255, 261
 penalized, 549
 prediction (PRESS), 673
 raw, 172
 regression (RegSS), 89, 98–99, 104, 113, 115–117, 141, 159–160, 172, 174–176, 193–194, 208, 218–219, 240, 248–250, 253–256, 259, 261, 292, 322
residual (RSS), 85, 89, 98–99, 115, 149, 159–160, 172–173, 198, 208, 217–218, 247–251, 253, 256, 345–346, 414, 450, 516–518, 532, 541, 543, 545, 548–549, 551, 555–556, 569, 671, 673–674, 694–695
total (TSS), 89, 98–99, 115–116, 149, 173, 180, 248–250, 253, 256, 545, 671
“Types I, II, and III”, 149, 167, 174, 384, 410
uncorrected, 250
vector geometry of, 248–249, 250–251, 253–256, 262
weighted, 304, 335, 532, 551, 593
within-group, 159
See also Incremental sum of squares; SS notation
Sum-of-squares-and-products (SSP) matrices, 225–227
Survey samples, complex, 460–464, 662–663
Tables. *See* Contingency tables
Three-way ANOVA. *See* Analysis of variance, three-way
Time-series data, 346, 474, 495
Time-series regression. *See* Generalized least squares
Tobit model, 638
Training subsample, 690
Transformable nonlinearity, 512–514
Transformations:
 arcsine-square-root, 74
 Box-Cox, 55–56, 76–77, 79, 324–325, 330, 337
 Box-Tidwell, 326–327, 338, 457, 526
 constructed variables for, 324–328, 457–458
 to correct nonconstant spread, 70–72, 303
 to correct nonlinearity, 63–69, 308–309
 to correct skewness, 59–63, 298
 family of powers and roots, 55–59
 “folded” powers and roots, 74
 and generalized least squares, 476, 486–487, 494, 497–498
 linear, effect of, on regression coefficients, 103–104, 124
 logarithms (logs), 51–52, 69
 logit, 73–74
 normalizing, *See* Transformations, Box-Cox
 of probabilities and proportions, 72–75
 probit, 74
 Yeo-Johnson, 79, 324
Trend in time series, 346, 479–481, 494–495
Tricube kernel function, 529–531, 533–534, 537–538, 543, 580–581
Truncated normal distribution, 629–630, 642
Truncation, 629–630
t-tests and confidence intervals:
 for constructed variable, 325–326, 328
 for contrasts in ANOVA, 193–194
 for difference of means, 194–195, 232, 610
 in multiple imputation, 622

- for regression coefficients, 111, 114, 117, 132, 139, 190–191, 216, 238, 450, 738
- for studentized residuals (outliers), 272–274, 298
- Tuning constant, 588–592
- Two-stage least-squares (2SLS) estimation, 234–235, 241
- Two-way ANOVA. *See* Analysis of variance, two-way
- Unbias of least-squares estimators, 109–110, 113, 118, 123, 211–213, 228, 275, 297, 301, 306, 362, 497
- Univariate missing data, 607, 611, 640–641
- Unmodeled heterogeneity, 432
- Unusual data, discarding, 288–289
 - See also* Influential observations; Leverage of observations; Outliers
- Validation. *See* Cross-validation; Model validation
- Validation subsample, 690
- Variable-selection methods in regression. *See* Model selection
- Variance components, 700, 711, 721, 727
- Variance-covariance components, 712–713, 733
- Variance-covariance matrix:
 - of errors, 188, 225, 240, 304, 335, 475, 485–486, 496, 498
 - of fitted values, 547
 - of fixed effects in the linear mixed-effects model, 737–738
 - of generalized least-squares estimator, 476
 - of generalized linear model coefficients, 448
 - of instrumental-variables estimator, 233–234, 240–241
 - of least-squares estimator, 211, 215, 305
 - of logit-model coefficients, 390–391, 398, 414
 - of M estimator coefficients, 594
 - of principal components, 351
 - of quantile-regression coefficients, 598
 - of ridge-regression estimator, 362–363, 367
 - sandwich estimator of, 305, 489
 - of two-stage least-squares estimator, 235
 - of weighted-least-squares estimator, 304, 335
- See also* Asymptotic standard errors; Standard errors
- Variance-inflation factors (VIF), 113, 342–343, 356, 459
 - generalized (GVIF), 357–358, 459–460, 635
- Vector geometry:
 - of added-variable plots, 291, 293
 - of analysis of variance, 259–260
 - of correlation, 249–250, 253–254
 - of multiple regression, 252–256, 334, 357
 - of principal components, 349–352
 - of simple regression, 245–251
- Wald tests:
 - bootstrapping, 660
 - in complex survey samples, 463
 - for generalized linear models, 425–426, 448, 450
 - for logit models, 382, 390, 400
 - with missing data, 614, 624
 - for mixed-effects models, 715, 724–725, 737–738
 - for overdispersion, 464
 - for proportional odds, 406
 - of transformation parameters, 77–78, 324
- Weighted least squares (WLS), 304–306, 335–336, 461, 475, 662, 666
 - estimation of linear probability model, 373
- See also* Iteratively weighted least squares; Local-polynomial regression; M estimator
- Weighted squares of means, method of, for ANOVA, 176
- “White” corrected (White-Huber) standard errors, 305–307, 448, 643
- White noise, 478
- Wilks’s lambda test statistic, 226
- Window:
 - in density estimation, 34–37
 - in nonparametric regression, 22–24, 529–531, 533–534, 536, 552, 573
 - Occam’s, 687
- See also* Bandwidth; Span of smoother
- Working response, 447, 575–576
- Yeo-Johnson family of transformations, 79, 324
- Yule-Walker equations, 485, 491
- Zero-inflated negative-binomial (ZINB) regression model, 465
- Zero-inflated Poisson (ZIP) regression model, 433–434, 465

Data Set Index

- Anscombe's "quartet," 28–30, 602
- B. Fox, Canadian women's labor-force time series, 346–349, 354–355, 357, 360–361, 367, 666
- Baseball salaries, 681–684, 686–689, 695
- Blau and Duncan, stratification, 237
- British Election Panel Study (BEPS), 392, 394–396
- Campbell, et al., *The American Voter*, 408–412, 435, 438, 440–442
- Canadian migration, 523–525
- Canadian occupational prestige, 20–24, 26, 32–33, 65, 67–68, 73–75, 97–102, 104, 133–134, 136–140, 145–151, 239, 333–334, 468, 530–531, 533–535, 540, 543–546, 550, 579–581
- Chilean plebiscite, 371–375, 378–379, 392, 572–574
- Cowles and Davis, volunteering, 505–506, 523
- Davis, height and weight of exercisers, 19–21, 24–26, 50, 83, 86–88, 91–92, 96, 103–104, 111–112, 124, 267–270, 274–275, 277, 279, 288, 665
- Davis et al., exercise and eating disorders, 718–725
- Duncan, U.S. occupational prestige, 48–49, 51, 94–96, 98–100, 114, 116–117, 124–125, 155–156, 209–210, 216, 219–220, 238, 261, 271–272, 274–275, 277–278, 280, 285–288, 293, 594–597, 601, 659–662, 665–666
- Fox and Hartnagel, Canadian crime-rates time series, 489–493, 498–499
- Friendly and Franklin, memory, 191–194, 199, 206–207
- Fournier et al., 2011 Canadian Election Study, 461–464
- General Social Survey, vocabulary, 45–46, 51–52, 181–186, 318–320, 323
- Greene and Shaffer, refugee appeals, 4–6, 691–693
- Kostecki-Dillon et al., migraine headaches, 745–747, 757
- Moore and Krupat, conformity, 164–167, 174–175, 188–190, 198, 200, 240
- Ornstein, Canadian interlocking directorates, 46–47, 70–72, 427–430, 432, 453, 455–460, 464, 602–603
- Raudenbush and Bryk, High School and Beyond, 704–717, 726–727, 736, 738
- Statistics Canada, Survey of Labour and Income Dynamics (SLID), 13–15, 296–300, 302–303, 305–306, 309–315, 317–318, 320–322, 325–328, 330–331, 337, 383–387, 452, 467, 511, 526, 558–566, 570–571, 577–578, 599, 635–639
- U.S. population, 519–521, 525
- United Nations, social indicators, 30–36, 39, 41–45, 60–62, 67–69, 77–78, 507–508, 580, 626–629, 641–642
- Wong et al., recovery from coma, 751–755, 757
- World Value Survey (WVS), government action on poverty, 403–407



SAGE video

We are delighted to announce the launch of a streaming video program at SAGE!

SAGE Video online collections are developed in partnership with leading academics, societies and practitioners, including many of SAGE's own authors and academic partners, to deliver cutting-edge pedagogical collections mapped to curricular needs.

Available alongside our book and reference collections on the SAGE Knowledge platform, content is delivered with critical online functionality designed to support scholarly use.

SAGE Video combines originally commissioned and produced material with licensed videos to provide a complete resource for students, faculty, and researchers.

NEW IN 2015!

- Counseling and Psychotherapy
- Education
- Media and Communication

sagepub.com/video
#sagevideo

 **SAGE** | **50** YEARS