

Linear Regression Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

Linear Regression Analysis

Second Edition

**GEORGE A. F. SEBER
ALAN J. LEE**

Department of Statistics
University of Auckland
Auckland, New Zealand



A JOHN WILEY & SONS PUBLICATION

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permreq@wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data Is Available

ISBN 0-471-41540-5

10 9 8 7

Contents

Preface	xv
1 Vectors of Random Variables	1
1.1 Notation	1
1.2 Statistical Models	2
1.3 Linear Regression Models	4
1.4 Expectation and Covariance Operators	5
Exercises 1a	8
1.5 Mean and Variance of Quadratic Forms	9
Exercises 1b	12
1.6 Moment Generating Functions and Independence	13
Exercises 1c	15
Miscellaneous Exercises 1	15
2 Multivariate Normal Distribution	17
2.1 Density Function	17
Exercises 2a	19
2.2 Moment Generating Functions	20
Exercises 2b	23
2.3 Statistical Independence	24

Exercises 2c	26
2.4 Distribution of Quadratic Forms	27
Exercises 2d	31
Miscellaneous Exercises 2	31
3 Linear Regression: Estimation and Distribution Theory	35
3.1 Least Squares Estimation	35
Exercises 3a	41
3.2 Properties of Least Squares Estimates	42
Exercises 3b	44
3.3 Unbiased Estimation of σ^2	44
Exercises 3c	47
3.4 Distribution Theory	47
Exercises 3d	49
3.5 Maximum Likelihood Estimation	49
3.6 Orthogonal Columns in the Regression Matrix	51
Exercises 3e	52
3.7 Introducing Further Explanatory Variables	54
3.7.1 General Theory	54
3.7.2 One Extra Variable	57
Exercises 3f	58
3.8 Estimation with Linear Restrictions	59
3.8.1 Method of Lagrange Multipliers	60
3.8.2 Method of Orthogonal Projections	61
Exercises 3g	62
3.9 Design Matrix of Less Than Full Rank	62
3.9.1 Least Squares Estimation	62
Exercises 3h	64
3.9.2 Estimable Functions	64
Exercises 3i	65
3.9.3 Introducing Further Explanatory Variables	65
3.9.4 Introducing Linear Restrictions	65
Exercises 3j	66
3.10 Generalized Least Squares	66
Exercises 3k	69
3.11 Centering and Scaling the Explanatory Variables	69
3.11.1 Centering	70
3.11.2 Scaling	71

Exercises 3l	72
3.12 Bayesian Estimation	73
Exercises 3m	76
3.13 Robust Regression	77
3.13.1 M-Estimates	78
3.13.2 Estimates Based on Robust Location and Scale Measures	80
3.13.3 Measuring Robustness	82
3.13.4 Other Robust Estimates	88
Exercises 3n	93
Miscellaneous Exercises 3	93
 4 Hypothesis Testing	97
4.1 Introduction	97
4.2 Likelihood Ratio Test	98
4.3 <i>F</i> -Test	99
4.3.1 Motivation	99
4.3.2 Derivation	99
Exercises 4a	102
4.3.3 Some Examples	103
4.3.4 The Straight Line	107
Exercises 4b	109
4.4 Multiple Correlation Coefficient	110
Exercises 4c	113
4.5 Canonical Form for H	113
Exercises 4d	114
4.6 Goodness-of-Fit Test	115
4.7 <i>F</i> -Test and Projection Matrices	116
Miscellaneous Exercises 4	117
 5 Confidence Intervals and Regions	119
5.1 Simultaneous Interval Estimation	119
5.1.1 Simultaneous Inferences	119
5.1.2 Comparison of Methods	124
5.1.3 Confidence Regions	125
5.1.4 Hypothesis Testing and Confidence Intervals	127
5.2 Confidence Bands for the Regression Surface	129
5.2.1 Confidence Intervals	129
5.2.2 Confidence Bands	129

5.3	Prediction Intervals and Bands for the Response	131
5.3.1	Prediction Intervals	131
5.3.2	Simultaneous Prediction Bands	133
5.4	Enlarging the Regression Matrix	135
	Miscellaneous Exercises 5	136
6	Straight-Line Regression	139
6.1	The Straight Line	139
6.1.1	Confidence Intervals for the Slope and Intercept	139
6.1.2	Confidence Interval for the x -Intercept	140
6.1.3	Prediction Intervals and Bands	141
6.1.4	Prediction Intervals for the Response	145
6.1.5	Inverse Prediction (Calibration)	145
	Exercises 6a	148
6.2	Straight Line through the Origin	149
6.3	Weighted Least Squares for the Straight Line	150
6.3.1	Known Weights	150
6.3.2	Unknown Weights	151
	Exercises 6b	153
6.4	Comparing Straight Lines	154
6.4.1	General Model	154
6.4.2	Use of Dummy Explanatory Variables	156
	Exercises 6c	157
6.5	Two-Phase Linear Regression	159
6.6	Local Linear Regression	162
	Miscellaneous Exercises 6	163
7	Polynomial Regression	165
7.1	Polynomials in One Variable	165
7.1.1	Problem of Ill-Conditioning	165
7.1.2	Using Orthogonal Polynomials	166
7.1.3	Controlled Calibration	172
7.2	Piecewise Polynomial Fitting	172
7.2.1	Unsatisfactory Fit	172
7.2.2	Spline Functions	173
7.2.3	Smoothing Splines	176
7.3	Polynomial Regression in Several Variables	180
7.3.1	Response Surfaces	180

7.3.2 Multidimensional Smoothing	184
Miscellaneous Exercises 7	185
8 Analysis of Variance	187
8.1 Introduction	187
8.2 One-Way Classification	188
8.2.1 General Theory	188
8.2.2 Confidence Intervals	192
8.2.3 Underlying Assumptions	195
Exercises 8a	196
8.3 Two-Way Classification (Unbalanced)	197
8.3.1 Representation as a Regression Model	197
8.3.2 Hypothesis Testing	197
8.3.3 Procedures for Testing the Hypotheses	201
8.3.4 Confidence Intervals	204
Exercises 8b	205
8.4 Two-Way Classification (Balanced)	206
Exercises 8c	209
8.5 Two-Way Classification (One Observation per Mean)	211
8.5.1 Underlying Assumptions	212
8.6 Higher-Way Classifications with Equal Numbers per Mean	216
8.6.1 Definition of Interactions	216
8.6.2 Hypothesis Testing	217
8.6.3 Missing Observations	220
Exercises 8d	221
8.7 Designs with Simple Block Structure	221
8.8 Analysis of Covariance	222
Exercises 8e	224
Miscellaneous Exercises 8	225
9 Departures from Underlying Assumptions	227
9.1 Introduction	227
9.2 Bias	228
9.2.1 Bias Due to Underfitting	228
9.2.2 Bias Due to Overfitting	230
Exercises 9a	231
9.3 Incorrect Variance Matrix	231
Exercises 9b	232

9.4	Effect of Outliers	233
9.5	Robustness of the <i>F</i> -Test to Nonnormality	235
9.5.1	Effect of the Regressor Variables	235
9.5.2	Quadratically Balanced <i>F</i> -Tests	236
	Exercises 9c	239
9.6	Effect of Random Explanatory Variables	240
9.6.1	Random Explanatory Variables Measured without Error	240
9.6.2	Fixed Explanatory Variables Measured with Error	241
9.6.3	Round-off Errors	245
9.6.4	Some Working Rules	245
9.6.5	Random Explanatory Variables Measured with Error	246
9.6.6	Controlled Variables Model	248
9.7	Collinearity	249
9.7.1	Effect on the Variances of the Estimated Coefficients	249
9.7.2	Variance Inflation Factors	254
9.7.3	Variances and Eigenvalues	255
9.7.4	Perturbation Theory	255
9.7.5	Collinearity and Prediction	261
	Exercises 9d	261
	Miscellaneous Exercises 9	262
10	Departures from Assumptions: Diagnosis and Remedies	265
10.1	Introduction	265
10.2	Residuals and Hat Matrix Diagonals	266
	Exercises 10a	270
10.3	Dealing with Curvature	271
10.3.1	Visualizing Regression Surfaces	271
10.3.2	Transforming to Remove Curvature	275
10.3.3	Adding and Deleting Variables	277
	Exercises 10b	279
10.4	Nonconstant Variance and Serial Correlation	281
10.4.1	Detecting Nonconstant Variance	281
10.4.2	Estimating Variance Functions	288
10.4.3	Transforming to Equalize Variances	291
10.4.4	Serial Correlation and the Durbin–Watson Test	292
	Exercises 10c	294
10.5	Departures from Normality	295
10.5.1	Normal Plotting	295

10.5.2 Transforming the Response	297
10.5.3 Transforming Both Sides	299
Exercises 10d	300
10.6 Detecting and Dealing with Outliers	301
10.6.1 Types of Outliers	301
10.6.2 Identifying High-Leverage Points	304
10.6.3 Leave-One-Out Case Diagnostics	306
10.6.4 Test for Outliers	310
10.6.5 Other Methods	311
Exercises 10e	314
10.7 Diagnosing Collinearity	315
10.7.1 Drawbacks of Centering	316
10.7.2 Detection of Points Influencing Collinearity	319
10.7.3 Remedies for Collinearity	320
Exercises 10f	326
Miscellaneous Exercises 10	327
11 Computational Algorithms for Fitting a Regression	329
11.1 Introduction	329
11.1.1 Basic Methods	329
11.2 Direct Solution of the Normal Equations	330
11.2.1 Calculation of the Matrix $\mathbf{X}'\mathbf{X}$	330
11.2.2 Solving the Normal Equations	331
Exercises 11a	337
11.3 QR Decomposition	338
11.3.1 Calculation of Regression Quantities	340
11.3.2 Algorithms for the QR and WU Decompositions	341
Exercises 11b	352
11.4 Singular Value Decomposition	353
11.4.1 Regression Calculations Using the SVD	353
11.4.2 Computing the SVD	354
11.5 Weighted Least Squares	355
11.6 Adding and Deleting Cases and Variables	356
11.6.1 Updating Formulas	356
11.6.2 Connection with the Sweep Operator	357
11.6.3 Adding and Deleting Cases and Variables Using QR	360
11.7 Centering the Data	363
11.8 Comparing Methods	365

11.8.1 Resources	365
11.8.2 Efficiency	366
11.8.3 Accuracy	369
11.8.4 Two Examples	372
11.8.5 Summary	373
Exercises 11c	374
11.9 Rank-Deficient Case	376
11.9.1 Modifying the QR Decomposition	376
11.9.2 Solving the Least Squares Problem	378
11.9.3 Calculating Rank in the Presence of Round-off Error	378
11.9.4 Using the Singular Value Decomposition	379
11.10 Computing the Hat Matrix Diagonals	379
11.10.1 Using the Cholesky Factorization	380
11.10.2 Using the Thin QR Decomposition	380
11.11 Calculating Test Statistics	380
11.12 Robust Regression Calculations	382
11.12.1 Algorithms for L_1 Regression	382
11.12.2 Algorithms for M- and GM-Estimation	384
11.12.3 Elemental Regressions	385
11.12.4 Algorithms for High-Breakdown Methods	385
Exercises 11d	388
Miscellaneous Exercises 11	389
12 Prediction and Model Selection	391
12.1 Introduction	391
12.2 Why Select?	393
Exercises 12a	399
12.3 Choosing the Best Subset	399
12.3.1 Goodness-of-Fit Criteria	400
12.3.2 Criteria Based on Prediction Error	401
12.3.3 Estimating Distributional Discrepancies	407
12.3.4 Approximating Posterior Probabilities	410
Exercises 12b	413
12.4 Stepwise Methods	413
12.4.1 Forward Selection	414
12.4.2 Backward Elimination	416
12.4.3 Stepwise Regression	418
Exercises 12c	420

12.5 Shrinkage Methods	420
12.5.1 Stein Shrinkage	420
12.5.2 Ridge Regression	423
12.5.3 Garrote and Lasso Estimates	425
Exercises 12d	427
12.6 Bayesian Methods	428
12.6.1 Predictive Densities	428
12.6.2 Bayesian Prediction	431
12.6.3 Bayesian Model Averaging	433
Exercises 12e	433
12.7 Effect of Model Selection on Inference	434
12.7.1 Conditional and Unconditional Distributions	434
12.7.2 Bias	436
12.7.3 Conditional Means and Variances	437
12.7.4 Estimating Coefficients Using Conditional Likelihood	437
12.7.5 Other Effects of Model Selection	438
Exercises 12f	438
12.8 Computational Considerations	439
12.8.1 Methods for All Possible Subsets	439
12.8.2 Generating the Best Regressions	442
12.8.3 All Possible Regressions Using QR Decompositions	446
Exercises 12g	447
12.9 Comparison of Methods	447
12.9.1 Identifying the Correct Subset	447
12.9.2 Using Prediction Error as a Criterion	448
Exercises 12h	456
Miscellaneous Exercises 12	456
 Appendix A Some Matrix Algebra	457
A.1 Trace and Eigenvalues	457
A.2 Rank	458
A.3 Positive-Semidefinite Matrices	460
A.4 Positive-Definite Matrices	461
A.5 Permutation Matrices	464
A.6 Idempotent Matrices	464
A.7 Eigenvalue Applications	465
A.8 Vector Differentiation	466
A.9 Patterned Matrices	466

A.10 Generalized Inverse	469
A.11 Some Useful Results	471
A.12 Singular Value Decomposition	471
A.13 Some Miscellaneous Statistical Results	472
A.14 Fisher Scoring	473
Appendix B Orthogonal Projections	475
B.1 Orthogonal Decomposition of Vectors	475
B.2 Orthogonal Complements	477
B.3 Projections on Subspaces	477
Appendix C Tables	479
C.1 Percentage Points of the Bonferroni t -Statistic	480
C.2 Distribution of the Largest Absolute Value of k Student t Variables	482
C.3 Working–Hotelling Confidence Bands for Finite Intervals	489
Outline Solutions to Selected Exercises	491
References	531
Index	549

Preface

Since publication of the first edition in 1977, there has been a steady flow of books on regression ranging over the pure–applied spectrum. Given the success of the first edition in both English and other languages (Russian and Chinese), we have therefore decided to maintain the same theoretical approach in this edition, so we make no apologies for a lack of data! However, since 1977 there have major advances in computing, especially in the use of powerful statistical packages, so our emphasis has changed. Although we cover much the same topics, the book has been largely rewritten to reflect current thinking. Of course, some theoretical aspects of regression, such as least squares and maximum likelihood are almost set in stone. However, topics such as analysis of covariance which, in the past, required various algebraic techniques can now be treated as a special case of multiple linear regression using an appropriate package.

We now list some of the major changes. Chapter 1 has been reorganized with more emphasis on moment generating functions. In Chapter 2 we have changed our approach to the multivariate normal distribution and the ensuing theorems about quadratics. Chapter 3 has less focus on the dichotomy of full-rank and less-than-full-rank models. Fitting models using Bayesian and robust methods are also included. Hypothesis testing again forms the focus of Chapter 4. The methods of constructing simultaneous confidence intervals have been updated in Chapter 5. In Chapter 6, on the straight line, there is more emphasis on modeling and piecewise fitting and less on algebra. New techniques of smoothing, such as splines and loess, are now considered in Chapters 6 and 7. Chapter 8, on analysis of variance and covariance, has

been updated, and the thorny problem of the two-way unbalanced model is addressed in detail. Departures from the underlying assumptions as well as the problem of collinearity are addressed in Chapter 9, and in Chapter 10 we discuss diagnostics and strategies for detecting and coping with such departures. Chapter 11 is a major update on the computational aspects, and Chapter 12 presents a comprehensive approach to the problem of model selection. There are some additions to the appendices and more exercises have been added.

One of the authors (GAFS) has been very encouraged by positive comments from many people, and he would like to thank those who have passed on errors found in the first edition. We also express our thanks to those reviewers of our proposed table of contents for their useful comments and suggestions.

GEORGE A. F. SEBER
ALAN J. LEE

*Auckland, New Zealand
November 2002*

1

Vectors of Random Variables

1.1 NOTATION

Matrices and vectors are denoted by boldface letters \mathbf{A} and \mathbf{a} , respectively, and scalars by italics. Random variables are represented by capital letters and their values by lowercase letters (e.g., Y and y , respectively). This use of capitals for random variables, which seems to be widely accepted, is particularly useful in regression when distinguishing between fixed and random regressor (independent) variables. However, it does cause problems because a vector of random variables, \mathbf{Y} , say, then looks like a matrix. Occasionally, because of a shortage of letters, a boldface lowercase letter represents a vector of random variables.

If X and Y are random variables, then the symbols $E[Y]$, $\text{var}[Y]$, $\text{cov}[X, Y]$, and $E[X|Y = y]$ (or, more briefly, $E[X|Y]$) represent expectation, variance, covariance, and conditional expectation, respectively.

The $n \times n$ matrix with diagonal elements d_1, d_2, \dots, d_n and zeros elsewhere is denoted by $\text{diag}(d_1, d_2, \dots, d_n)$, and when all the d_i 's are unity we have the identity matrix \mathbf{I}_n .

If \mathbf{a} is an $n \times 1$ column vector with elements a_1, a_2, \dots, a_n , we write $\mathbf{a} = (a_i)$, and the *length* or *norm* of \mathbf{a} is denoted by $\|\mathbf{a}\|$. Thus

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}'\mathbf{a}} = (a_1^2 + a_2^2 + \cdots + a_n^2)^{1/2}.$$

The vector with elements all equal to unity is represented by $\mathbf{1}_n$, and the set of all vectors having n elements is denoted by \mathfrak{R}_n .

If the $m \times n$ matrix \mathbf{A} has elements a_{ij} , we write $\mathbf{A} = (a_{ij})$, and the sum of the diagonal elements, called the *trace* of \mathbf{A} , is denoted by $\text{tr}(\mathbf{A})$ ($= a_{11} + a_{22} + \cdots + a_{kk}$, where k is the smaller of m and n). The *transpose*

of \mathbf{A} is represented by $\mathbf{A}' = (a'_{ij})$, where $a'_{ij} = a_{ji}$. If \mathbf{A} is square, its determinant is written $\det(\mathbf{A})$, and if \mathbf{A} is nonsingular its inverse is denoted by \mathbf{A}^{-1} . The space spanned by the columns of \mathbf{A} , called the *column space* of \mathbf{A} , is denoted by $\mathcal{C}(\mathbf{A})$. The null space or kernel of \mathbf{A} ($= \{x : \mathbf{Ax} = 0\}$) is denoted by $\mathcal{N}(\mathbf{A})$.

We say that $Y \sim N(\theta, \sigma^2)$ if Y is normally distributed with mean θ and variance σ^2 : Y has a *standard normal* distribution if $\theta = 0$ and $\sigma^2 = 1$. The t - and chi-square distributions with k degrees of freedom are denoted by t_k and χ_k^2 , respectively, and the F -distribution with m and n degrees of freedom is denoted by $F_{m,n}$.

Finally we mention the *dot* and *bar* notation, representing sum and average, respectively; for example,

$$a_{i\cdot} = \sum_{j=1}^J a_{ij} \quad \text{and} \quad \bar{a}_{i\cdot} = \frac{a_{i\cdot}}{J}.$$

In the case of a single subscript, we omit the dot.

Some knowledge of linear algebra by the reader is assumed, and for a short review course several books are available (see, e.g., Harville [1997]). However, a number of matrix results are included in Appendices A and B at the end of this book, and references to these appendices are denoted by, e.g., A.2.3.

1.2 STATISTICAL MODELS

A major activity in statistics is the building of statistical models that hopefully reflect the important aspects of the object of study with some degree of realism. In particular, the aim of *regression analysis* is to construct mathematical models which describe or explain relationships that may exist between variables. The simplest case is when there are just two variables, such as height and weight, income and intelligence quotient (IQ), ages of husband and wife at marriage, population size and time, length and breadth of leaves, temperature and pressure of a certain volume of gas, and so on. If we have n pairs of observations (x_i, y_i) ($i = 1, 2, \dots, n$), we can plot these points, giving a *scatter diagram*, and endeavor to fit a smooth curve through the points in such a way that the points are as close to the curve as possible. Clearly, we would not expect an exact fit, as at least one of the variables is subject to chance fluctuations due to factors outside our control. Even if there is an “exact” relationship between such variables as temperature and pressure, fluctuations would still show up in the scatter diagram because of errors of measurement. The simplest two-variable regression model is the straight line, and it is assumed that the reader has already come across the fitting of such a model.

Statistical models are fitted for a variety of reasons. One important reason is that of trying to uncover causes by studying relationships between vari-

ables. Usually, we are interested in just one variable, called the *response* (or *predicted* or *dependent*) *variable*, and we want to study how it depends on a set of variables called the *explanatory variables* (or *regressors* or *independent variables*). For example, our response variable might be the risk of heart attack, and the explanatory variables could include blood pressure, age, gender, cholesterol level, and so on. We know that statistical relationships do not necessarily imply causal relationships, but the presence of any statistical relationship does give us a starting point for further research. Once we are confident that a statistical relationship exists, we can then try to model this relationship mathematically and then use the model for prediction. For a given person, we can use their values of the explanatory variables to predict their risk of a heart attack. We need, however, to be careful when making predictions outside the usual ranges of the explanatory variables, as the model may not be valid there.

A second reason for fitting models, over and above prediction and explanation, is to examine and test scientific hypotheses, as in the following simple examples.

EXAMPLE 1.1 Ohm's law states that $Y = rX$, where X amperes is the current through a resistor of r ohms and Y volts is the voltage across the resistor. This give us a straight line through the origin so that a linear scatter diagram will lend support to the law. \square

EXAMPLE 1.2 The theory of gravitation states that the force of gravity F between two objects is given by $F = \alpha/d^\beta$. Here d is the distance between the objects and α is a constant related to the masses of the two objects. The famous inverse square law states that $\beta = 2$. We might want to test whether this is consistent with experimental measurements. \square

EXAMPLE 1.3 Economic theory uses a *production function*, $Q = \alpha L^\beta K^\gamma$, to relate Q (production) to L (the quantity of labor) and K (the quantity of capital). Here α , β , and γ are constants that depend on the type of goods and the market involved. We might want to estimate these parameters for a particular market and use the relationship to predict the effects of infusions of capital on the behavior of that market. \square

From these examples we see that we might use models developed from theoretical considerations to (a) check up on the validity of the theory (as in the Ohm's law example), (b) test whether a parameter has the value predicted from the theory, under the assumption that the model is true (as in the gravitational example and the inverse square law), and (c) estimate the unknown constants, under the assumption of a valid model, and then use the model for prediction purposes (as in the economic example).

1.3 LINEAR REGRESSION MODELS

If we denote the response variable by Y and the explanatory variables by X_1, X_2, \dots, X_K , then a general model relating these variables is

$$E[Y|X_1 = x_1, X_2 = x_2, \dots, X_K = x_K] = \phi(x_1, x_2, \dots, x_K),$$

although, for brevity, we will usually drop the conditioning part and write $E[Y]$. In this book we direct our attention to the important class of linear models, that is,

$$\phi(x_1, x_2, \dots, x_K) = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K,$$

which is linear in the parameters β_j . This restriction to linearity is not as restrictive as one might think. For example, many functions of several variables are approximately linear over sufficiently small regions, or they may be made linear by a suitable transformation. Using logarithms for the gravitational model, we get the straight line

$$\log F = \log \alpha - \beta \log d. \quad (1.1)$$

For the linear model, the x_i could be functions of other variables z, w , etc.; for example, $x_1 = \sin z$, $x_2 = \log w$, and $x_3 = zw$. We can also have $x_i = x^i$, which leads to a polynomial model; the linearity refers to the parameters, not the variables. Note that “categorical” models can be included under our umbrella by using *dummy (indicator)* x -variables. For example, suppose that we wish to compare the means of two populations, say, $\mu_i = E[U_i]$ ($i = 1, 2$). Then we can combine the data into the single model

$$\begin{aligned} E[Y] &= \mu_1 + (\mu_2 - \mu_1)x \\ &= \beta_0 + \beta_1 x, \end{aligned}$$

where $x = 0$ when Y is a U_1 observation and $x = 1$ when Y is a U_2 observation. Here $\mu_1 = \beta_0$ and $\mu_2 = \beta_0 + \beta_1$, the difference being β_1 . We can extend this idea to the case of comparing m means using $m - 1$ dummy variables.

In a similar fashion we can combine two straight lines,

$$U_j = \alpha_j + \gamma_j x_1 \quad (j = 1, 2),$$

using a dummy x_2 variable which takes the value 0 if the observation is from the first line, and 1 otherwise. The combined model is

$$\begin{aligned} E[Y] &= \alpha_1 + \gamma_1 x_1 + (\alpha_2 - \alpha_1)x_2 + (\gamma_2 - \gamma_1)x_1 x_2 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \end{aligned} \quad (1.2)$$

say, where $x_3 = x_1 x_2$. Here $\alpha_1 = \beta_0$, $\alpha_2 = \beta_0 + \beta_2$, $\gamma_1 = \beta_1$, and $\gamma_2 = \beta_1 + \beta_3$.

In the various models considered above, the explanatory variables may or may not be random. For example, dummy variables are nonrandom. With random X -variables, we carry out the regression conditionally on their observed values, provided that they are measured exactly (or at least with sufficient accuracy). We effectively proceed as though the X -variables were not random at all. When measurement errors cannot be ignored, the theory has to be modified, as we shall see in Chapter 9.

1.4 EXPECTATION AND COVARIANCE OPERATORS

In this book we focus on vectors and matrices, so we first need to generalize the ideas of expectation, covariance, and variance, which we do in this section.

Let Z_{ij} ($i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$) be a set of random variables with expected values $E[Z_{ij}]$. Expressing both the random variables and their expectations in matrix form, we can define the general expectation operator of the matrix $\mathbf{Z} = (Z_{ij})$ as follows:

Definition 1.1

$$E[\mathbf{Z}] = (E[Z_{ij}]).$$

THEOREM 1.1 *If $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$, and $\mathbf{C} = (c_{ij})$ are $l \times m$, $n \times p$, and $l \times p$ matrices, respectively, of constants, then*

$$E[\mathbf{A}\mathbf{Z}\mathbf{B} + \mathbf{C}] = \mathbf{A}E[\mathbf{Z}]\mathbf{B} + \mathbf{C}.$$

Proof. Let $\mathbf{W} = \mathbf{A}\mathbf{Z}\mathbf{B} + \mathbf{C}$; then $W_{ij} = \sum_{r=1}^m \sum_{s=1}^n a_{ir} Z_{rs} b_{sj} + c_{ij}$ and

$$\begin{aligned} E[\mathbf{A}\mathbf{Z}\mathbf{B} + \mathbf{C}] &= (E[W_{ij}]) = \left(\sum_r \sum_s a_{ir} E[Z_{rs}] b_{sj} + c_{ij} \right) \\ &= ((\mathbf{A}E[\mathbf{Z}]\mathbf{B})_{ij}) + (c_{ij}) \\ &= \mathbf{A}E[\mathbf{Z}]\mathbf{B} + \mathbf{C}. \end{aligned}$$

□

In this proof we note that l , m , n , and p are any positive integers, and the matrices of constants can take any values. For example, if \mathbf{X} is an $m \times 1$ vector, then $E[\mathbf{AX}] = \mathbf{AE}[\mathbf{X}]$. Using similar algebra, we can prove that if \mathbf{A} and \mathbf{B} are $m \times n$ matrices of constants, and \mathbf{X} and \mathbf{Y} are $n \times 1$ vectors of random variables, then

$$E[\mathbf{AX} + \mathbf{BY}] = \mathbf{AE}[\mathbf{X}] + \mathbf{BE}[\mathbf{Y}].$$

In a similar manner we can generalize the notions of covariance and variance for vectors. If \mathbf{X} and \mathbf{Y} are $m \times 1$ and $n \times 1$ vectors of random variables, then we define the generalized covariance operator Cov as follows:

Definition 1.2

$$\text{Cov}[\mathbf{X}, \mathbf{Y}] = (\text{cov}[X_i, Y_j]).$$

THEOREM 1.2 If $E[\mathbf{X}] = \boldsymbol{\alpha}$ and $E[\mathbf{Y}] = \boldsymbol{\beta}$, then

$$\text{Cov}[\mathbf{X}, \mathbf{Y}] = E[(\mathbf{X} - \boldsymbol{\alpha})(\mathbf{Y} - \boldsymbol{\beta})'].$$

Proof.

$$\begin{aligned}\text{Cov}[\mathbf{X}, \mathbf{Y}] &= (\text{cov}[X_i, Y_j]) \\ &= \{E[(X_i - \alpha_i)(Y_j - \beta_j)]\} \\ &= E\{[(X_i - \alpha_i)(Y_j - \beta_j)]\} \\ &= E[(\mathbf{X} - \boldsymbol{\alpha})(\mathbf{Y} - \boldsymbol{\beta})'].\end{aligned}\quad \square$$

Definition 1.3 When $\mathbf{Y} = \mathbf{X}$, $\text{Cov}[\mathbf{X}, \mathbf{X}]$, written as $\text{Var}[\mathbf{X}]$, is called the variance (variance-covariance or dispersion) matrix of \mathbf{X} . Thus

$$\begin{aligned}\text{Var}[\mathbf{X}] &= (\text{cov}[X_i, X_j]) \\ &= \begin{pmatrix} \text{var}[X_1] & \text{cov}[X_1, X_2] & \cdots & \text{cov}[X_1, X_n] \\ \text{cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{cov}[X_2, X_n] \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}[X_n, X_1] & \text{cov}[X_n, X_2] & \cdots & \text{var}[X_n] \end{pmatrix}. \quad (1.3)\end{aligned}$$

Since $\text{cov}[X_i, X_j] = \text{cov}[X_j, X_i]$, the matrix above is symmetric. We note that when $\mathbf{X} = X_1$ we write $\text{Var}[\mathbf{X}] = \text{var}[X_1]$.

From Theorem 1.2 with $\mathbf{Y} = \mathbf{X}$ we have

$$\text{Var}[\mathbf{X}] = E[(\mathbf{X} - \boldsymbol{\alpha})(\mathbf{X} - \boldsymbol{\alpha})'], \quad (1.4)$$

which, on expanding, leads to

$$\text{Var}[\mathbf{X}] = E[\mathbf{X}\mathbf{X}'] - \boldsymbol{\alpha}\boldsymbol{\alpha}'. \quad (1.5)$$

These last two equations are natural generalizations of univariate results.

EXAMPLE 1.4 If \mathbf{a} is any $n \times 1$ vector of constants, then

$$\text{Var}[\mathbf{X} - \mathbf{a}] = \text{Var}[\mathbf{X}].$$

This follows from the fact that $X_i - a_i - E[X_i - a_i] = X_i - E[X_i]$, so that

$$\text{cov}[X_i - a_i, X_j - a_j] = \text{cov}[X_i, X_j]. \quad \square$$

THEOREM 1.3 If \mathbf{X} and \mathbf{Y} are $m \times 1$ and $n \times 1$ vectors of random variables, and \mathbf{A} and \mathbf{B} are $l \times m$ and $p \times n$ matrices of constants, respectively, then

$$\text{Cov}[\mathbf{AX}, \mathbf{BY}] = \mathbf{A} \text{Cov}[\mathbf{X}, \mathbf{Y}] \mathbf{B}' . \quad (1.6)$$

Proof. Let $\mathbf{U} = \mathbf{AX}$ and $\mathbf{V} = \mathbf{BY}$. Then, by Theorems 1.2 and 1.1,

$$\begin{aligned} \text{Cov}[\mathbf{AX}, \mathbf{BY}] &= \text{Cov}[\mathbf{U}, \mathbf{V}] \\ &= E[(\mathbf{U} - E[\mathbf{U}])(\mathbf{V} - E[\mathbf{V}])'] \\ &= E[(\mathbf{AX} - \mathbf{A}\alpha)(\mathbf{BY} - \mathbf{B}\beta)'] \\ &= E[\mathbf{A}(\mathbf{X} - \alpha)(\mathbf{Y} - \beta)' \mathbf{B}'] \\ &= \mathbf{A}E[(\mathbf{X} - \alpha)(\mathbf{Y} - \beta)'] \mathbf{B}' \\ &= \mathbf{A} \text{Cov}[\mathbf{X}, \mathbf{Y}] \mathbf{B}' . \end{aligned} \quad \square$$

From the theorem above we have the special cases

$$\text{Cov}[\mathbf{AX}, \mathbf{Y}] = \mathbf{A} \text{Cov}[\mathbf{X}, \mathbf{Y}] \quad \text{and} \quad \text{Cov}[\mathbf{X}, \mathbf{BY}] = \text{Cov}[\mathbf{X}, \mathbf{Y}] \mathbf{B}' .$$

Of particular importance is the following result, obtained by setting $\mathbf{B} = \mathbf{A}$ and $\mathbf{Y} = \mathbf{X}$:

$$\text{Var}[\mathbf{AX}] = \text{Cov}[\mathbf{AX}, \mathbf{AX}] = \mathbf{A} \text{Cov}[\mathbf{X}, \mathbf{X}] \mathbf{A}' = \mathbf{A} \text{Var}[\mathbf{X}] \mathbf{A}' . \quad (1.7)$$

EXAMPLE 1.5 If \mathbf{X} , \mathbf{Y} , \mathbf{U} , and \mathbf{V} are any (not necessarily distinct) $n \times 1$ vectors of random variables, then for all real numbers a , b , c , and d (including zero),

$$\begin{aligned} &\text{Cov}[a\mathbf{X} + b\mathbf{Y}, c\mathbf{U} + d\mathbf{V}] \\ &= ac \text{Cov}[\mathbf{X}, \mathbf{U}] + ad \text{Cov}[\mathbf{X}, \mathbf{V}] + bc \text{Cov}[\mathbf{Y}, \mathbf{U}] + bd \text{Cov}[\mathbf{Y}, \mathbf{V}] . \end{aligned} \quad (1.8)$$

To prove this result, we simply multiply out

$$\begin{aligned} &E[(a\mathbf{X} + b\mathbf{Y} - aE[\mathbf{X}] - bE[\mathbf{Y}])(c\mathbf{U} + d\mathbf{V} - cE[\mathbf{U}] - dE[\mathbf{V}])'] \\ &= E[(a(\mathbf{X} - E[\mathbf{X}]) + b(\mathbf{Y} - E[\mathbf{Y}]))(c(\mathbf{U} - E[\mathbf{U}]) + d(\mathbf{V} - E[\mathbf{V}]))'] . \end{aligned}$$

If we set $\mathbf{U} = \mathbf{X}$ and $\mathbf{V} = \mathbf{Y}$, $c = a$ and $d = b$, we get

$$\begin{aligned} \text{Var}[a\mathbf{X} + b\mathbf{Y}] &= \text{Cov}[a\mathbf{X} + b\mathbf{Y}, a\mathbf{X} + b\mathbf{Y}] \\ &= a^2 \text{Var}[\mathbf{X}] + ab(\text{Cov}[\mathbf{X}, \mathbf{Y}] + \text{Cov}[\mathbf{Y}, \mathbf{X}]) \\ &\quad + b^2 \text{Var}[\mathbf{Y}] . \end{aligned} \quad (1.9)$$

\square

In Chapter 2 we make frequent use of the following theorem.

THEOREM 1.4 *If \mathbf{X} is a vector of random variables such that no element of \mathbf{X} is a linear combination of the remaining elements [i.e., there do not exist $\mathbf{a} (\neq \mathbf{0})$ and b such that $\mathbf{a}'\mathbf{X} = b$ for all values of $\mathbf{X} = \mathbf{x}$], then $\text{Var}[\mathbf{X}]$ is a positive-definite matrix (see A.4)].*

Proof. For any vector \mathbf{c} , we have

$$\begin{aligned} 0 &\leq \text{var}[\mathbf{c}'\mathbf{X}] \\ &= \mathbf{c}' \text{Var}[\mathbf{X}] \mathbf{c} \quad [\text{by equation (1.7)}]. \end{aligned}$$

Now equality holds if and only if $\mathbf{c}'\mathbf{X}$ is a constant, that is, if and only if $\mathbf{c}'\mathbf{X} = d$ ($\mathbf{c} \neq \mathbf{0}$) or $\mathbf{c} = \mathbf{0}$. Because the former possibility is ruled out, $\mathbf{c} = \mathbf{0}$ and $\text{Var}[\mathbf{X}]$ is positive-definite. \square

EXAMPLE 1.6 If \mathbf{X} and \mathbf{Y} are $m \times 1$ and $n \times 1$ vectors of random variables such that no element of \mathbf{X} is a linear combination of the remaining elements, then there exists an $n \times m$ matrix \mathbf{M} such that $\text{Cov}[\mathbf{X}, \mathbf{Y} - \mathbf{MX}] = \mathbf{0}$. To find \mathbf{M} , we use the previous results to get

$$\begin{aligned} \text{Cov}[\mathbf{X}, \mathbf{Y} - \mathbf{MX}] &= \text{Cov}[\mathbf{X}, \mathbf{Y}] - \text{Cov}[\mathbf{X}, \mathbf{MX}] \\ &= \text{Cov}[\mathbf{X}, \mathbf{Y}] - \text{Cov}[\mathbf{X}, \mathbf{X}]\mathbf{M}' \\ &= \text{Cov}[\mathbf{X}, \mathbf{Y}] - \text{Var}[\mathbf{X}]\mathbf{M}'. \end{aligned} \quad (1.10)$$

By Theorem 1.4, $\text{Var}[\mathbf{X}]$ is positive-definite and therefore nonsingular (A.4.1). Hence (1.10) is zero for

$$\mathbf{M}' = (\text{Var}[\mathbf{X}])^{-1} \text{Cov}[\mathbf{X}, \mathbf{Y}]. \quad \square$$

EXAMPLE 1.7 We now give an example of a singular variance matrix by using the two-cell multinomial distribution to represent a binomial distribution as follows:

$$\text{pr}(X_1 = x_1, X_2 = x_2) = \frac{n!}{x_1! x_2!} p_1^{x_1} p_2^{x_2}, \quad p_1 + p_2 = 1, \quad x_1 + x_2 = n.$$

If $\mathbf{X} = (X_1, X_2)'$, then

$$\text{Var}[\mathbf{X}] = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 \\ -np_1p_2 & np_2(1-p_2) \end{pmatrix},$$

which has rank 1 as $p_2 = 1 - p_1$. \square

EXERCISES 1a

1. Prove that if \mathbf{a} is a vector of constants with the same dimension as the random vector \mathbf{X} , then

$$E[(\mathbf{X} - \mathbf{a})(\mathbf{X} - \mathbf{a})'] = \text{Var}[\mathbf{X}] + (E[\mathbf{X}] - \mathbf{a})(E[\mathbf{X}] - \mathbf{a})'.$$

If $\text{Var}[\mathbf{X}] = \Sigma = (\sigma_{ij})$, deduce that

$$E[||\mathbf{X} - \mathbf{a}||^2] = \sum_i \sigma_{ii} + ||E[\mathbf{X}] - \mathbf{a}||^2.$$

2. If \mathbf{X} and \mathbf{Y} are $m \times 1$ and $n \times 1$ vectors of random variables, and \mathbf{a} and \mathbf{b} are $m \times 1$ and $n \times 1$ vectors of constants, prove that

$$\text{Cov}[\mathbf{X} - \mathbf{a}, \mathbf{Y} - \mathbf{b}] = \text{Cov}[\mathbf{X}, \mathbf{Y}].$$

3. Let $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ be a vector of random variables, and let $Y_1 = X_1$, $Y_i = X_i - X_{i-1}$ ($i = 2, 3, \dots, n$). If the Y_i are mutually independent random variables, each with unit variance, find $\text{Var}[\mathbf{X}]$.
4. If X_1, X_2, \dots, X_n are random variables satisfying $X_{i+1} = \rho X_i$ ($i = 1, 2, \dots, n-1$), where ρ is a constant, and $\text{var}[X_1] = \sigma^2$, find $\text{Var}[\mathbf{X}]$.

1.5 MEAN AND VARIANCE OF QUADRATIC FORMS

Quadratic forms play a major role in this book. In particular, we will frequently need to find the expected value of a quadratic form using the following theorem.

THEOREM 1.5 *Let $\mathbf{X} = (X_i)$ be an $n \times 1$ vector of random variables, and let \mathbf{A} be an $n \times n$ symmetric matrix. If $E[\mathbf{X}] = \mu$ and $\text{Var}[\mathbf{X}] = \Sigma = (\sigma_{ij})$, then*

$$E[\mathbf{X}' \mathbf{A} \mathbf{X}] = \text{tr}(\mathbf{A} \Sigma) + \mu' \mathbf{A} \mu.$$

Proof.

$$\begin{aligned} E[\mathbf{X}' \mathbf{A} \mathbf{X}] &= \text{tr}(E[\mathbf{X}' \mathbf{A} \mathbf{X}]) \\ &= E[\text{tr}(\mathbf{X}' \mathbf{A} \mathbf{X})] \\ &= E[\text{tr}(\mathbf{A} \mathbf{X} \mathbf{X}'')] \quad [\text{by A.1.2}] \\ &= \text{tr}(E[\mathbf{A} \mathbf{X} \mathbf{X}'']) \\ &= \text{tr}(\mathbf{A} E[\mathbf{X} \mathbf{X}'']) \\ &= \text{tr}[\mathbf{A} (\text{Var}[\mathbf{X}] + \mu \mu')] \quad [\text{by (1.5)}] \\ &= \text{tr}(\mathbf{A} \Sigma) + \text{tr}(\mathbf{A} \mu \mu') \\ &= \text{tr}(\mathbf{A} \Sigma) + \mu' \mathbf{A} \mu \quad [\text{by A.1.2}.] \end{aligned} \quad \square$$

We can deduce two special cases. First, by setting $\mathbf{Y} = \mathbf{X} - \mathbf{b}$ and noting that $\text{Var}[\mathbf{Y}] = \text{Var}[\mathbf{X}]$ (by Example 1.4), we have

$$E[(\mathbf{X} - \mathbf{b})' \mathbf{A} (\mathbf{X} - \mathbf{b})] = \text{tr}(\mathbf{A} \Sigma) + (\mu - \mathbf{b})' \mathbf{A} (\mu - \mathbf{b}). \quad (1.11)$$

Second, if $\Sigma = \sigma^2 \mathbf{I}_n$ (a common situation in this book), then $\text{tr}(\mathbf{A}\Sigma) = \sigma^2 \text{tr}(\mathbf{A})$. Thus in this case we have the simple rule

$$E[\mathbf{X}'\mathbf{AX}] = \sigma^2(\text{sum of coefficients of } X_i^2) + (\mathbf{X}'\mathbf{AX})_{\mathbf{x}=\mu}. \quad (1.12)$$

EXAMPLE 1.8 If X_1, X_2, \dots, X_n are independently and identically distributed with mean μ and variance σ^2 , then we can use equation (1.12) to find the expected value of

$$Q = (X_1 - X_2)^2 + (X_2 - X_3)^2 + \cdots + (X_{n-1} - X_n)^2.$$

To do so, we first write

$$Q = \mathbf{X}'\mathbf{AX} = 2 \sum_{i=1}^n X_i^2 - X_1^2 - X_n^2 - 2 \sum_{i=1}^{n-1} X_i X_{i+1}.$$

Then, since $\text{cov}[X_i, X_j] = 0$ ($i \neq j$), $\Sigma = \sigma^2 \mathbf{I}_n$ and from the squared terms, $\text{tr}(\mathbf{A}) = 2n - 2$. Replacing each X_i by μ in the original expression for Q , we see that the second term of $E[\mathbf{X}'\mathbf{AX}]$ is zero, so that $E[Q] = \sigma^2(2n - 2)$. \square

EXAMPLE 1.9 Suppose that the elements of $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ have a common mean μ and \mathbf{X} has variance matrix Σ with $\sigma_{ii} = \sigma^2$ and $\sigma_{ij} = \rho\sigma^2$ ($i \neq j$). Then, when $\rho = 0$, we know that $Q = \sum_i (X_i - \bar{X})^2$ has expected value $\sigma^2(n - 1)$. To find its expected value when $\rho \neq 0$, we express Q in the form $\mathbf{X}'\mathbf{AX}$, where $\mathbf{A} = [\delta_{ij} - n^{-1}]$ and

$$\begin{aligned} \mathbf{A}\Sigma &= \sigma^2 \begin{pmatrix} 1 - n^{-1} & -n^{-1} & \cdots & -n^{-1} \\ -n^{-1} & 1 - n^{-1} & \cdots & -n^{-1} \\ \cdots & \cdots & \cdots & \cdots \\ -n^{-1} & -n^{-1} & \cdots & 1 - n^{-1} \end{pmatrix} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \\ &= \sigma^2(1 - \rho)\mathbf{A}. \end{aligned}$$

Once again the second term in $E[Q]$ is zero, so that

$$E[Q] = \text{tr}(\mathbf{A}\Sigma) = \sigma^2(1 - \rho) \text{tr}(\mathbf{A}) = \sigma^2(1 - \rho)(n - 1). \quad \square$$

THEOREM 1.6 Let X_1, X_2, \dots, X_n be independent random variables with means $\theta_1, \theta_2, \dots, \theta_n$, common variance μ_2 , and common third and fourth moments about their means, μ_3 and μ_4 , respectively (i.e., $\mu_r = E[(X_i - \theta_i)^r]$). If \mathbf{A} is any $n \times n$ symmetric matrix and \mathbf{a} is a column vector of the diagonal elements of \mathbf{A} , then

$$\text{var}[\mathbf{X}'\mathbf{AX}] = (\mu_4 - 3\mu_2^2)\mathbf{a}'\mathbf{a} + 2\mu_2^2 \text{tr}(\mathbf{A}^2) + 4\mu_2\mathbf{\theta}'\mathbf{A}^2\mathbf{\theta} + 4\mu_3\mathbf{\theta}'\mathbf{A}\mathbf{a}.$$

(This result is stated without proof in Atiqullah [1962].)

Proof. We note that $E[\mathbf{X}] = \boldsymbol{\theta}$, $\text{Var}[\mathbf{X}] = \mu_2 \mathbf{I}_n$, and

$$\text{Var}[\mathbf{X}'\mathbf{AX}] = E[(\mathbf{X}'\mathbf{AX})^2] - (E[\mathbf{X}'\mathbf{AX}])^2. \quad (1.13)$$

Now

$$\mathbf{X}'\mathbf{A}\mathbf{X} = (\mathbf{X} - \boldsymbol{\theta})'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta}) + 2\boldsymbol{\theta}'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta},$$

so that squaring gives

$$\begin{aligned} (\mathbf{X}'\mathbf{A}\mathbf{X})^2 &= [(\mathbf{X} - \boldsymbol{\theta})'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta})]^2 + 4[\boldsymbol{\theta}'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta})]^2 + (\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta})^2 \\ &\quad + 2\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}[(\mathbf{X} - \boldsymbol{\theta})'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta}) + 4\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta})] \\ &\quad + 4\boldsymbol{\theta}'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta})(\mathbf{X} - \boldsymbol{\theta})'\mathbf{A}(\mathbf{X} - \boldsymbol{\theta}). \end{aligned}$$

Setting $\mathbf{Y} = \mathbf{X} - \boldsymbol{\theta}$, we have $E[\mathbf{Y}] = \mathbf{0}$ and, using Theorem 1.5,

$$\begin{aligned} E[(\mathbf{X}'\mathbf{A}\mathbf{X})^2] &= E[(\mathbf{Y}'\mathbf{A}\mathbf{Y})^2] + 4E[(\boldsymbol{\theta}'\mathbf{A}\mathbf{Y})^2] + (\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta})^2 \\ &\quad + 2\boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}\mu_2 \text{tr}(\mathbf{A}) + 4E[\boldsymbol{\theta}'\mathbf{A}\mathbf{Y}\mathbf{Y}'\mathbf{A}\mathbf{Y}]. \end{aligned}$$

As a first step in evaluating the expression above we note that

$$(\mathbf{Y}'\mathbf{A}\mathbf{Y})^2 = \sum_i \sum_j \sum_k \sum_l a_{ij} a_{kl} Y_i Y_j Y_k Y_l.$$

Since the Y_i are mutually independent with the same first four moments about the origin, we have

$$E[Y_i Y_j Y_k Y_l] = \begin{cases} \mu_4, & i = j = k = l, \\ \mu_2^2, & i = j, k = l; i = k, j = l; i = l, j = k, \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$\begin{aligned} E[(\mathbf{Y}'\mathbf{A}\mathbf{Y})^2] &= \mu_4 \sum_i a_{ii}^2 + \mu_2^2 \sum_i \left(\sum_{k \neq i} a_{ii} a_{kk} + \sum_{j \neq i} a_{ij}^2 + \sum_{j \neq i} a_{ij} a_{ji} \right) \\ &= (\mu_4 - 3\mu_2^2)\mathbf{a}'\mathbf{a} + \mu_2^2 [\text{tr}(\mathbf{A})^2 + 2\text{tr}(\mathbf{A}^2)], \end{aligned} \tag{1.14}$$

since \mathbf{A} is symmetric and $\sum_i \sum_j a_{ij}^2 = \text{tr}(\mathbf{A}^2)$. Also,

$$(\boldsymbol{\theta}'\mathbf{A}\mathbf{Y})^2 = (\mathbf{b}'\mathbf{Y})^2 = \sum_i \sum_j b_i b_j Y_i Y_j,$$

say, and

$$\boldsymbol{\theta}'\mathbf{A}\mathbf{Y}\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_i \sum_j \sum_k b_i a_{jk} Y_i Y_j Y_k,$$

so that

$$E[(\boldsymbol{\theta}'\mathbf{A}\mathbf{Y})^2] = \mu_2 \sum_i b_i^2 = \mu_2 \mathbf{b}'\mathbf{b} = \mu_2 \boldsymbol{\theta}'\mathbf{A}^2\boldsymbol{\theta}$$

and

$$E[\boldsymbol{\theta}'\mathbf{A}\mathbf{Y}\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \mu_3 \sum_i b_i a_{ii} = \mu_3 \mathbf{b}'\mathbf{a} = \mu_3 \boldsymbol{\theta}'\mathbf{A}\mathbf{a}.$$

Finally, collecting all the terms and substituting into equation (1.13) leads to the desired result. \square

EXERCISES 1b

1. Suppose that X_1, X_2 , and X_3 are random variables with common mean μ and variance matrix

$$\text{Var}[\mathbf{X}] = \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

Find $E[X_1^2 + 2X_1X_2 - 4X_2X_3 + X_3^2]$.

2. If X_1, X_2, \dots, X_n are independent random variables with common mean μ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, prove that $\sum_i (X_i - \bar{X})^2/[n(n-1)]$ is an unbiased estimate of $\text{var}[\bar{X}]$.
3. Suppose that in Exercise 2 the variances are known. Let $\bar{X}_w = \sum_i w_i X_i$ be an unbiased estimate of μ (i.e., $\sum_i w_i = 1$).
- Prove that $\text{var}[\bar{X}_w]$ is minimized when $w_i \propto 1/\sigma_i^2$. Find this minimum variance v_{\min} .
 - Let $S_w^2 = \sum_i w_i (X_i - \bar{X}_w)^2/(n-1)$. If $w_i \sigma_i^2 = a$ ($i = 1, 2, \dots, n$), prove that $E[S_w^2]$ is an unbiased estimate of v_{\min} .
4. The random variables X_1, X_2, \dots, X_n have a common nonzero mean μ , a common variance σ^2 , and the correlation between any pair of random variables is ρ .
- Find $\text{var}[\bar{X}]$ and hence prove that $-1/(n-1) \leq \rho \leq 1$.
 - If

$$Q = a \sum_{i=1}^n X_i^2 + b \left(\sum_{i=1}^n X_i \right)^2$$

is an unbiased estimate of σ^2 , find a and b . Hence show that, in this case,

$$Q = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{(1-\rho)(n-1)}.$$

5. Let X_1, X_2, \dots, X_n be independently distributed as $N(\mu, \sigma^2)$. Define

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

and

$$Q = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2.$$

- (a) Prove that $\text{var}[S^2] = 2\sigma^4/(n-1)$.
- (b) Show that Q is an unbiased estimate of σ^2 .
- (c) Find the variance of Q and hence show that as $n \rightarrow \infty$, the efficiency of Q relative to S^2 is $\frac{2}{3}$.

1.6 MOMENT GENERATING FUNCTIONS AND INDEPENDENCE

If \mathbf{X} and \mathbf{t} are $n \times 1$ vectors of random variables and constants, respectively, then the *moment generating function* (m.g.f.) of \mathbf{X} is defined to be

$$M_{\mathbf{X}}(\mathbf{t}) = E[\exp(\mathbf{t}'\mathbf{X})].$$

A key result about m.g.f.'s is that if $M_{\mathbf{X}}(\mathbf{t})$ exists for all $\|\mathbf{t}\| \leq t_0$ ($t_0 > 0$) (i.e., in an interval containing the origin), then it determines the distribution uniquely. Fortunately, most of the common distributions have m.g.f.'s, one important exception being the t -distribution (with some of its moments being infinite, including the Cauchy distribution with 1 degree of freedom). We give an example where this uniqueness is usefully exploited. It is assumed that the reader is familiar with the m.g.f. of χ_r^2 : namely, $(1-2t)^{-r/2}$.

EXAMPLE 1.10 Suppose that $Q_i \sim \chi_{r_i}^2$ for $i = 1, 2$, and $Q = Q_1 - Q_2$ is statistically independent of Q_2 . We now show that $Q \sim \chi_r^2$, where $r = r_1 - r_2$. Writing

$$\begin{aligned} (1-2t)^{-r_1/2} &= E[\exp(tQ_1)] \\ &= E[\exp(tQ + tQ_2)] \\ &= E[\exp(tQ)]E[\exp(tQ_2)] \\ &= E[\exp(tQ)](1-2t)^{-r/2}, \end{aligned}$$

we have

$$E[\exp(tQ)] = (1-2t)^{-(r_1-r_2)/2},$$

which is the m.g.f. of χ_r^2 . □

Moment generating functions also provide a convenient method for proving results about statistical independence. For example, if $M_{\mathbf{X}}(\mathbf{t})$ exists and

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}}(t_1, \dots, t_r, 0, \dots, 0)M_{\mathbf{X}}(0, \dots, 0, t_{r+1}, \dots, t_n),$$

then $\mathbf{X}_1 = (X_1, \dots, X_r)'$ and $\mathbf{X}_2 = (X_{r+1}, \dots, X_n)'$ are statistically independent. An equivalent result is that \mathbf{X}_1 and \mathbf{X}_2 are independent if and only if we have the factorization

$$M_{\mathbf{X}}(\mathbf{t}) = a(t_1, \dots, t_r)b(t_{r+1}, \dots, t_n)$$

for some functions $a(\cdot)$ and $b(\cdot)$.

EXAMPLE 1.11 Suppose that the joint distribution of the vectors of random variables \mathbf{X} and \mathbf{Y} have a joint m.g.f. which exists in an interval containing the origin. Then if \mathbf{X} and \mathbf{Y} are independent, so are any (measurable) functions of them. This follows from the fact that if $\mathbf{c}(\cdot)$ and $\mathbf{d}(\cdot)$ are suitable vector functions,

$$E[\exp\{\mathbf{s}'\mathbf{c}(\mathbf{X}) + \mathbf{s}'\mathbf{d}(\mathbf{Y})\}] = E[\exp\{\mathbf{s}'\mathbf{c}(\mathbf{X})\}]E[\exp\{\mathbf{s}'\mathbf{d}(\mathbf{Y})\}] = a(\mathbf{s})b(\mathbf{t}),$$

say. This result is, in fact, true for any \mathbf{X} and \mathbf{Y} , even if their m.g.f.'s do not exist, and can be proved using characteristic functions. \square

Another route that we shall use for proving independence is via covariance. It is well known that $\text{cov}[X, Y] = 0$ does not in general imply that X and Y are independent. However, in one important special case, the bivariate normal distribution, X and Y are independent if and only if $\text{cov}[X, Y] = 0$. A generalization of this result applied to the multivariate normal distribution is given in Chapter 2. For more than two variables we find that for multivariate normal distributions, the variables are mutually independent if and only if they are pairwise independent. However, pairwise independence does not necessarily imply mutual independence, as we see in the following example.

EXAMPLE 1.12 Suppose that X_1 , X_2 , and X_3 have joint density function

$$\begin{aligned} f(x_1, x_2, x_3) &= (2\pi)^{-3/2} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right] \\ &\quad \times \{1 + x_1 x_2 x_3 \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right]\} \\ &\quad -\infty < x_i < \infty \quad (i = 1, 2, 3). \end{aligned}$$

Then the second term in the braces above is an odd function of x_3 , so that its integral over $-\infty < x_3 < \infty$ is zero. Hence

$$\begin{aligned} f_{12}(x_1, x_2) &= (2\pi)^{-1} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2)\right] \\ &= f_1(x_1)f_2(x_2), \end{aligned}$$

and X_1 and X_2 are independent $N(0, 1)$ variables. Thus although X_1 , X_2 , and X_3 are pairwise independent, they are not mutually independent, as

$$f(x_1, x_2, x_3) \neq (2\pi)^{-3/2} \exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right] = f_1(x_1)f_2(x_2)f_3(x_3). \quad \square$$

EXERCISES 1c

1. If X and Y are random variables with the same variance, prove that $\text{cov}[X + Y, X - Y] = 0$. Give a counterexample which shows that zero covariance does not necessarily imply independence.
2. Let X and Y be discrete random variables taking values 0 or 1 only, and let $\text{pr}(X = i, Y = j) = p_{ij}$ ($i = 1, 0; j = 1, 0$). Prove that X and Y are independent if and only if $\text{cov}[X, Y] = 0$.
3. If X is a random variable with a density function symmetric about zero and having zero mean, prove that $\text{cov}[X, X^2] = 0$.
4. If X, Y and Z have joint density function

$$f(x, y, z) = \frac{1}{8}(1 + xyz) \quad (-1 \leq x, y, z \leq 1),$$

prove that they are pairwise independent but not mutually independent.

MISCELLANEOUS EXERCISES 1

1. If X and Y are random variables, prove that

$$\text{var}[X] = E_Y \{ \text{var}[X|Y] \} + \text{var}_Y \{ E[X|Y] \}.$$

Generalize this result to vectors \mathbf{X} and \mathbf{Y} of random variables.

2. Let $\mathbf{X} = (X_1, X_2, X_3)'$ with

$$\text{Var}[\mathbf{X}] = \begin{pmatrix} 5 & 2 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix}.$$

- (a) Find the variance of $X_1 - 2X_2 + X_3$.
 - (b) Find the variance matrix of $\mathbf{Y} = (Y_1, Y_2)'$, where $Y_1 = X_1 + X_2$ and $Y_2 = X_1 + X_2 + X_3$.
3. Let X_1, X_2, \dots, X_n be random variables with a common mean μ . Suppose that $\text{cov}[X_i, X_j] = 0$ for all i and j such that $j > i + 1$. If

$$Q_1 = \sum_{i=1}^n (X_i - \bar{X})^2$$

and

$$Q_2 = (X_1 - X_2)^2 + (X_2 - X_3)^2 + \cdots + (X_{n-1} - X_n)^2 + (X_n - X_1)^2,$$

prove that

$$E \left[\frac{3Q_1 - Q_2}{n(n-3)} \right] = \text{var}[\bar{X}].$$

4. Given a random sample X_1, X_2, X_3 from the distribution with density function

$$f(x) = \frac{1}{2} \quad (-1 \leq x \leq 1),$$

find the variance of $(X_1 - X_2)^2 + (X_2 - X_3)^2 + (X_3 - X_1)^2$.

5. If X_1, \dots, X_n are independently and identically distributed as $N(0, \sigma^2)$, and \mathbf{A} and \mathbf{B} are any $n \times n$ symmetric matrices, prove that

$$\text{Cov}[\mathbf{X}' \mathbf{A} \mathbf{X}, \mathbf{X}' \mathbf{B} \mathbf{X}] = 2\sigma^4 \text{tr}(\mathbf{AB}).$$

2

Multivariate Normal Distribution

2.1 DENSITY FUNCTION

Let Σ be a positive-definite $n \times n$ matrix and μ an n -vector. Consider the (positive) function

$$f(y_1, \dots, y_n) = k^{-1} \exp[-\frac{1}{2}(y - \mu)' \Sigma^{-1} (y - \mu)], \quad (2.1)$$

where k is a constant. Since Σ (and hence Σ^{-1} by A.4.3) is positive-definite, the quadratic form $(y - \mu)' \Sigma^{-1} (y - \mu)$ is nonnegative and the function f is bounded, taking its maximum value of k^{-1} at $y = \mu$.

Because Σ is positive-definite, it has a symmetric positive-definite square root $\Sigma^{1/2}$, which satisfies $(\Sigma^{1/2})^2 = \Sigma$ (by A.4.12).

Let $z = \Sigma^{-1/2}(y - \mu)$, so that $y = \Sigma^{1/2}z + \mu$. The Jacobian of this transformation is

$$J = \det \left(\frac{\partial y_i}{\partial z_j} \right) = \det(\Sigma^{1/2}) = [\det(\Sigma)]^{1/2}.$$

Changing the variables in the integral, we get

$$\begin{aligned} & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[-\frac{1}{2}(y - \mu)' \Sigma^{-1} (y - \mu)] dy_1 \cdots dy_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-\frac{1}{2}z' \Sigma^{1/2} \Sigma^{-1} \Sigma^{1/2} z) |J| dz_1 \cdots dz_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(-\frac{1}{2}z' z) |J| dz_1 \cdots dz_n \end{aligned}$$

$$\begin{aligned}
&= |J| \prod_{i=1}^n \int_{-\infty}^{\infty} \exp(-\frac{1}{2}z_i^2) dz_i \\
&= |J| \prod_{i=1}^n (2\pi)^{1/2} \\
&= (2\pi)^{n/2} \det(\Sigma)^{1/2}.
\end{aligned}$$

Since $f > 0$, it follows that if $k = (2\pi)^{n/2} \det(\Sigma)^{1/2}$, then (2.1) represents a density function.

Definition 2.1 *The distribution corresponding to the density (2.1) is called the multivariate normal distribution.*

THEOREM 2.1 *If a random vector \mathbf{Y} has density (2.1), then $E[\mathbf{Y}] = \mu$ and $\text{Var}[\mathbf{Y}] = \Sigma$.*

Proof. Let $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{Y} - \mu)$. Repeating the argument above, we see, using the change-of-variable formula, that \mathbf{Z} has density

$$\begin{aligned}
g(z_1, z_2, \dots, z_n) &= f[\mathbf{y}(\mathbf{z})] |J| \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z_i^2) \tag{2.2} \\
&= \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2}\mathbf{z}'\mathbf{z}). \tag{2.3}
\end{aligned}$$

The factorization of the joint density function in (2.2) implies that the Z_i are mutually independent normal variables and $Z_i \sim N(0, 1)$. Thus $E[\mathbf{Z}] = \mathbf{0}$ and $\text{Var}[\mathbf{Z}] = \mathbf{I}_n$, so that

$$E[\mathbf{Y}] = E[\Sigma^{1/2}\mathbf{Z} + \mu] = \Sigma^{1/2}E[\mathbf{Z}] + \mu = \mu$$

and

$$\text{Var}[\mathbf{Y}] = \text{Var}[\Sigma^{1/2}\mathbf{Z} + \mu] = \text{Var}[\Sigma^{1/2}\mathbf{Z}] = \Sigma^{1/2}\mathbf{I}_n\Sigma^{1/2} = \Sigma. \quad \square$$

We shall use the notation $\mathbf{Y} \sim N_n(\mu, \Sigma)$ to indicate that \mathbf{Y} has the density (2.1). When $n = 1$ we drop the subscript.

EXAMPLE 2.1 Let Z_1, \dots, Z_n be independent $N(0, 1)$ random variables. The density of $\mathbf{Z} = (Z_1, \dots, Z_n)'$ is the product of the univariate densities given by (2.2), so that by (2.3) the density of \mathbf{Z} is of the form (2.1) with $\mu = \mathbf{0}$ and $\Sigma = \mathbf{I}_n$ [i.e., $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$]. \square

We conclude that if $\mathbf{Y} \sim N_n(\mu, \Sigma)$ and $\mathbf{Y} = \Sigma^{1/2}\mathbf{Z} + \mu$, then $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{Y} - \mu)$ and $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. The distribution of \mathbf{Z} is the simplest and most fundamental example of the multivariate normal. Just as any univariate normal can be obtained by rescaling and translating a standard normal with

mean zero and variance 1, so can any multivariate normal be thought of as a rescaled and translated $N_n(\mathbf{0}, \mathbf{I}_n)$. Multiplying by $\Sigma^{1/2}$ is just a type of rescaling of the elements of \mathbf{Z} , and adding μ is just a translation by μ .

EXAMPLE 2.2 Consider the function

$$f(x, y) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}\sigma_x\sigma_y} \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right] \right\}$$

where $\sigma_x > 0$, $\sigma_y > 0$, and $|\rho| < 1$. Then f is of the form (2.1) with

$$\mu' = (\mu_x, \mu_y) \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

The density f above is the density of the *bivariate normal distribution*. \square

EXERCISES 2a

1. Show that

$$f(y_1, y_2) = k^{-1} \exp[-\frac{1}{2}(2y_1^2 + y_2^2 + 2y_1y_2 - 22y_1 - 14y_2 + 65)]$$

is the density of a bivariate normal random vector $\mathbf{Y} = (Y_1, Y_2)'$.

(a) Find k .

(b) Find $E[\mathbf{Y}]$ and $\text{Var}[\mathbf{Y}]$.

2. Let \mathbf{U} have density g and let $\mathbf{Y} = \mathbf{A}(\mathbf{U} + \mathbf{c})$, where \mathbf{A} is nonsingular. Show that the density f of \mathbf{Y} satisfies

$$f(\mathbf{y}) = g(\mathbf{u}) / |\det(\mathbf{A})|,$$

where $\mathbf{y} = \mathbf{A}(\mathbf{u} + \mathbf{c})$.

3. (a) Show that the 3×3 matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

is positive-definite for $\rho > -\frac{1}{2}$.

(b) Find $\Sigma^{1/2}$ when

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

2.2 MOMENT GENERATING FUNCTIONS

We can use the results of Section 2.1 to calculate the moment generating function (m.g.f.) of the multivariate normal. First, if $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, then, by the independence of the Z_i 's, the m.g.f. of \mathbf{Z} is

$$\begin{aligned}
E[\exp(t' \mathbf{Z})] &= E \left[\exp \left(\sum_{i=1}^n t_i Z_i \right) \right] \\
&= E \left[\prod_{i=1}^n \exp(t_i Z_i) \right] \\
&= \prod_{i=1}^n E[\exp(t_i Z_i)] \\
&= \prod_{i=1}^n \exp(\frac{1}{2} t_i^2) \\
&= \exp(\frac{1}{2} t' t).
\end{aligned} \tag{2.4}$$

Now if $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can write $\mathbf{Y} = \boldsymbol{\Sigma}^{1/2} \mathbf{Z} + \boldsymbol{\mu}$, where $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Hence using (2.4) and putting $\mathbf{s} = \boldsymbol{\Sigma}^{1/2} \mathbf{t}$, we get

$$\begin{aligned}
E[\exp(t' \mathbf{Y})] &= E[\exp\{\mathbf{t}' (\boldsymbol{\Sigma}^{1/2} \mathbf{Z} + \boldsymbol{\mu})\}] \\
&= E[\exp(\mathbf{s}' \mathbf{Z})] \exp(\mathbf{t}' \boldsymbol{\mu}) \\
&= \exp(\frac{1}{2} \mathbf{s}' \mathbf{s}) \exp(\mathbf{t}' \boldsymbol{\mu}) \\
&= \exp(\frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{t} + \mathbf{t}' \boldsymbol{\mu}) \\
&= \exp(\mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Sigma} \mathbf{t}).
\end{aligned} \tag{2.5}$$

Another well-known result for the univariate normal is that if $Y \sim N(\mu, \sigma^2)$, then $aY + b$ is $N(a\mu + b, a^2\sigma^2)$ provided that $a \neq 0$. A similar result is true for the multivariate normal, as we see below.

THEOREM 2.2 *Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, \mathbf{C} be an $m \times n$ matrix of rank m , and \mathbf{d} be an $m \times 1$ vector. Then $\mathbf{CY} + \mathbf{d} \sim N_m(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'')$.*

Proof. The m.g.f. of $\mathbf{CY} + \mathbf{d}$ is

$$\begin{aligned}
E\{\exp[t'(\mathbf{CY} + \mathbf{d})]\} &= E\{\exp[(\mathbf{C}' \mathbf{t})' \mathbf{Y} + \mathbf{t}' \mathbf{d}]\} \\
&= \exp[(\mathbf{C}' \mathbf{t})' \boldsymbol{\mu} + \frac{1}{2} (\mathbf{C}' \mathbf{t})' \boldsymbol{\Sigma} \mathbf{C}' \mathbf{t} + \mathbf{t}' \mathbf{d}] \\
&= \exp[t'(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}) + \frac{1}{2} \mathbf{t}' \mathbf{C} \boldsymbol{\Sigma} \mathbf{C}' \mathbf{t}].
\end{aligned}$$

Since $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$ is positive-definite, the equation above is the moment generating function of $N_m(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'')$. We stress that \mathbf{C} must be of full rank to ensure that $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$ is positive-definite (by A.4.5), since we have only defined the multivariate normal for positive-definite variance matrices. \square

COROLLARY If $\mathbf{Y} = \mathbf{AZ} + \boldsymbol{\mu}$, where \mathbf{A} is an $n \times n$ nonsingular matrix, then $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{AA}')$.

Proof. We replace \mathbf{Y} , $\boldsymbol{\mu}$, Σ and \mathbf{d} by \mathbf{Z} , $\mathbf{0}$, \mathbf{I}_n and $\boldsymbol{\mu}$, respectively, in Theorem 2.2. \square

EXAMPLE 2.3 Suppose that $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and that \mathbf{T} is an orthogonal matrix. Then, by Theorem 2.2, $\mathbf{Z} = \mathbf{T}'\mathbf{Y}$ is $N_n(\mathbf{0}, \mathbf{I}_n)$, since $\mathbf{T}'\mathbf{T} = \mathbf{I}_n$. \square

In subsequent chapters, we shall need to deal with random vectors of the form \mathbf{CY} , where \mathbf{Y} is multivariate normal but the matrix \mathbf{C} is not of full rank. For example, the vectors of fitted values and residuals in a regression are of this form. In addition, the statement and proof of many theorems become much simpler if we admit the possibility of singular variance matrices. In particular we would like the Corollary above to hold in some sense when \mathbf{C} does not have full row rank.

Let $\mathbf{Z} \sim N_m(\mathbf{0}, \mathbf{I}_m)$, and let \mathbf{A} be an $n \times m$ matrix and $\boldsymbol{\mu}$ an $n \times 1$ vector. By replacing $\Sigma^{1/2}$ by \mathbf{A} in the derivation of (2.5), we see that the m.g.f. of $\mathbf{Y} = \mathbf{AZ} + \boldsymbol{\mu}$ is $\exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$, with $\Sigma = \mathbf{AA}'$. Since distributions having the same m.g.f. are identical, the distribution of \mathbf{Y} depends on \mathbf{A} only through \mathbf{AA}' . We note that $E[\mathbf{Y}] = \mathbf{AE}[\mathbf{Z}] + \boldsymbol{\mu} = \boldsymbol{\mu}$ and $\text{Var}[\mathbf{Y}] = \mathbf{A}\text{Var}[\mathbf{Z}]\mathbf{A}' = \mathbf{AA}'$. These results motivate us to introduce the following definition.

Definition 2.2 A random $n \times 1$ vector \mathbf{Y} with mean $\boldsymbol{\mu}$ and variance matrix Σ has a multivariate normal distribution if it has the same distribution as $\mathbf{AZ} + \boldsymbol{\mu}$, where \mathbf{A} is any $n \times m$ matrix satisfying $\Sigma = \mathbf{AA}'$ and $\mathbf{Z} \sim N_m(\mathbf{0}, \mathbf{I}_m)$. We write $\mathbf{Y} \sim \mathbf{AZ} + \boldsymbol{\mu}$ to indicate that \mathbf{Y} and $\mathbf{AZ} + \boldsymbol{\mu}$ have the same distribution.

We need to prove that when Σ is positive-definite, the new definition is equivalent to the old. As noted above, the distribution is invariant to the choice of \mathbf{A} , as long as $\Sigma = \mathbf{AA}'$. If Σ is of full rank (or, equivalently, is positive-definite), then there exists a nonsingular \mathbf{A} with $\Sigma = \mathbf{AA}'$, by A.4.2. If \mathbf{Y} is multivariate normal by Definition 2.1, then Theorem 2.2 shows that $\mathbf{Z} = \mathbf{A}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$ is $N_n(\mathbf{0}, \mathbf{I}_n)$, so \mathbf{Y} is multivariate normal in the sense of Definition 2.2. Conversely, if \mathbf{Y} is multivariate normal by Definition 2.2, then its m.g.f. is given by (2.5). But this is also the m.g.f. of a random vector having density (2.1), so by the uniqueness of the m.g.f.'s, \mathbf{Y} must also have density (2.1).

If Σ is of rank $m < n$, the probability distribution of \mathbf{Y} cannot be expressed in terms of a density function. In both cases, irrespective of whether Σ is positive-definite or just positive-semidefinite, we saw above that the m.g.f. is

$$\exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}). \quad (2.6)$$

We write $\mathbf{Y} \sim N_m(\boldsymbol{\mu}, \Sigma)$ as before. When Σ has less than full rank, \mathbf{Y} is sometimes said to have a singular distribution. From now on, no assumption that Σ is positive-definite will be made unless explicitly stated.

EXAMPLE 2.4 Let $Y \sim N(\mu, \sigma^2)$ and put $\mathbf{Y}' = (Y, -Y)$. The variance-covariance matrix of \mathbf{Y} is

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Put $Z = (Y - \mu)/\sigma$. Then

$$\mathbf{Y} = \begin{pmatrix} \sigma \\ -\sigma \end{pmatrix} Z + \begin{pmatrix} \mu \\ \mu \end{pmatrix} = \mathbf{A}Z + \boldsymbol{\mu}$$

and

$$\Sigma = \mathbf{A}\mathbf{A}'.$$

Thus \mathbf{Y} has a multivariate normal distribution. \square

EXAMPLE 2.5 We can show that Theorem 2.2 remains true for random vectors having this extended definition of the multivariate normal without the restriction on the rank of \mathbf{A} . If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{Y} \sim \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$. Hence $C\mathbf{Y} \sim C\mathbf{A}\mathbf{Z} + C\boldsymbol{\mu} = \mathbf{B}\mathbf{Z} + \mathbf{b}$, say, and $C\mathbf{Y}$ is multivariate normal with $E[C\mathbf{Y}] = \mathbf{b} = C\boldsymbol{\mu}$ and $\text{Var}[C\mathbf{Y}] = \mathbf{B}\mathbf{B}' = \mathbf{C}\mathbf{A}\mathbf{A}'\mathbf{C}' = \mathbf{C}\Sigma\mathbf{C}'$. \square

EXAMPLE 2.6 Under the extended definition, a constant vector has a multivariate normal distribution. (Take \mathbf{A} to be a matrix of zeros.) In particular, if \mathbf{A} is a zero row vector, a scalar constant has a (univariate) normal distribution under this definition, so that we regard constants (with zero variance) as being normally distributed. \square

EXAMPLE 2.7 (Marginal distributions) Suppose that $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ and we partition \mathbf{Y} , $\boldsymbol{\mu}$ and Σ conformably as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then $\mathbf{Y}_1 \sim N_p(\boldsymbol{\mu}_1, \Sigma_{11})$. We see this by writing $\mathbf{Y}_1 = \mathbf{B}\mathbf{Y}$, where $\mathbf{B} = (\mathbf{I}_p, \mathbf{0})$. Then $\mathbf{B}\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{B}\Sigma\mathbf{B}' = \Sigma_{11}$, so the result follows from Theorem 2.2. Clearly, \mathbf{Y}_1 can be any subset of \mathbf{Y} . In other words, the marginal distributions of the multivariate normal are multivariate normal. \square

Our final result in this section is a characterization of the multivariate normal.

THEOREM 2.3 A random vector \mathbf{Y} with variance-covariance matrix Σ and mean vector $\boldsymbol{\mu}$ has a $N_n(\boldsymbol{\mu}, \Sigma)$ distribution if and only if $\mathbf{a}'\mathbf{Y}$ has a univariate normal distribution for every vector \mathbf{a} .

Proof. First, assume that $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$. Then $\mathbf{Y} \sim \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$, so that $\mathbf{a}'\mathbf{Y} \sim \mathbf{a}'\mathbf{A}\mathbf{Z} + \mathbf{a}'\boldsymbol{\mu} = (\mathbf{A}'\mathbf{a})'Z + \mathbf{a}'\boldsymbol{\mu}$. This has a (univariate) normal distribution in the sense of Definition 2.2.

Conversely, assume that $\mathbf{t}'\mathbf{Y}$ is a univariate normal random variable for all \mathbf{t} . Its mean is $\mathbf{t}'\boldsymbol{\mu}$ and the variance is $\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$. Then using the formula for the m.g.f. of the univariate normal, we get

$$E\{\exp[s(\mathbf{t}'\mathbf{Y})]\} = \exp[s(\mathbf{t}'\boldsymbol{\mu}) + \frac{1}{2}s^2(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})].$$

Putting $s = 1$ shows that the m.g.f. of \mathbf{Y} is given by (2.6), and thus $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. \square

We have seen in Example 2.7 that the multivariate normal has normal marginals; and in particular the univariate marginals are normal. However, the converse is not true, as the following example shows. Consider the function

$$f(y_1, y_2) = (2\pi)^{-1} \exp[-\frac{1}{2}(y_1^2 + y_2^2)] \{1 + y_1 y_2 \exp[-\frac{1}{2}(y_1^2 + y_2^2)]\},$$

which is nonnegative (since $1 + ye^{-y^2} > 0$) and integrates to 1 (since the integral $\int_{-\infty}^{+\infty} ye^{-y^2/2} dy$ has value 0). Thus f is a joint density, but it is not bivariate normal. However,

$$\begin{aligned} \int_{-\infty}^{+\infty} f(y_1, y_2) dy_2 &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y_1^2) \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-\frac{1}{2}y_2^2) dy_2 \\ &\quad + \frac{1}{\sqrt{2\pi}} y_1 \exp(-\frac{1}{2}y_1^2) \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y_2 \exp(-\frac{1}{2}y_2^2) dy_2 \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y_1^2), \end{aligned}$$

so that the marginals are $N(0, 1)$. In terms of Theorem 2.3, to prove that \mathbf{Y} is bivariate normal, we must show that $\mathbf{a}'\mathbf{Y}$ is bivariate normal for all vectors \mathbf{a} , not just for the vectors $(1, 0)$ and $(0, 1)$. Many other examples such as this are known; see, for example, Pierce and Dykstra [1969], Joshi [1970], and Kowalski [1970].

EXERCISES 2b

- Find the moment generating function of the bivariate normal distribution given in Example 2.2.
- If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, show that $Y_i \sim N(\mu_i, \sigma_{ii})$.
- Suppose that $\mathbf{Y} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Find the joint distribution of $Z_1 = Y_1 + Y_2 + Y_3$ and $Z_2 = Y_1 - Y_2$.

- Given $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{I}_n)$, find the joint density of $\mathbf{a}'\mathbf{Y}$ and $\mathbf{b}'\mathbf{Y}$, where $\mathbf{a}'\mathbf{b} = 0$, and hence show that $\mathbf{a}'\mathbf{Y}$ and $\mathbf{b}'\mathbf{Y}$ are independent.

5. Let $(X_i, Y_i), i = 1, 2, \dots, n$, be a random sample from a bivariate normal distribution. Find the joint distribution of (\bar{X}, \bar{Y}) .
6. If Y_1 and Y_2 are random variables such that $Y_1 + Y_2$ and $Y_1 - Y_2$ are independent $N(0, 1)$ random variables, show that Y_1 and Y_2 have a bivariate normal distribution. Find the mean and variance matrix of $\mathbf{Y} = (Y_1, Y_2)'$.
7. Let X_1 and X_2 have joint density

$$f(x_1, x_2) = \frac{1}{2\pi} \exp[-\frac{1}{2}(x_1^2 + x_2^2)] \left[1 - \frac{x_1 x_2}{(1+x_1^2)(1+x_2^2)} \right], \\ -\infty < x_1, x_2 < \infty.$$

Show that X_1 and X_2 have $N(0, 1)$ marginal distributions.

(Joshi [1970])

8. Suppose that Y_1, Y_2, \dots, Y_n are independently distributed as $N(0, 1)$. Calculate the m.g.f. of the random vector

$$(\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y})$$

and hence show that \bar{Y} is independent of $\sum_i (Y_i - \bar{Y})^2$.

(Hogg and Craig [1970])

9. Let X_1 , X_2 , and X_3 be i.i.d. $N(0, 1)$. Let

$$\begin{aligned} Y_1 &= (X_1 + X_2 + X_3)/\sqrt{3}, \\ Y_2 &= (X_1 - X_2)/\sqrt{2}, \\ Y_3 &= (X_1 + X_2 - 2X_3)/\sqrt{6}. \end{aligned}$$

Show that Y_1 , Y_2 and Y_3 are i.i.d. $N(0, 1)$. (The transformation above is a special case of the so-called *Helmert transformation*.)

2.3 STATISTICAL INDEPENDENCE

For any pair of random variables, independence implies that the pair are uncorrelated. For the normal distribution the converse is also true, as we now show.

THEOREM 2.4 *Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partition \mathbf{Y} , $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as in Example 2.7. Then \mathbf{Y}_1 and \mathbf{Y}_2 are independent if and only if $\boldsymbol{\Sigma}_{12} = 0$.*

Proof. The m.g.f. of \mathbf{Y} is $\exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$. Partition \mathbf{t} conformably with \mathbf{Y} . Then the exponent in the m.g.f. above is

$$\mathbf{t}'_1 \boldsymbol{\mu}_1 + \mathbf{t}'_2 \boldsymbol{\mu}_2 + \frac{1}{2} \mathbf{t}'_1 \boldsymbol{\Sigma}_{11} \mathbf{t}_1 + \frac{1}{2} \mathbf{t}'_2 \boldsymbol{\Sigma}_{22} \mathbf{t}_2 + \mathbf{t}'_1 \boldsymbol{\Sigma}_{12} \mathbf{t}_2. \quad (2.7)$$

If $\Sigma_{12} = 0$, the exponent can be written as a function of just t_1 plus a function of just t_2 , so the m.g.f. factorizes into a term in t_1 alone times a term in t_2 alone. This implies that \mathbf{Y}_1 and \mathbf{Y}_2 are independent.

Conversely, if \mathbf{Y}_1 and \mathbf{Y}_2 are independent, then

$$M(t_1, \mathbf{0})M(\mathbf{0}, t_2) = M(t_1, t_2),$$

where M is the m.g.f. of \mathbf{Y} . By (2.7) this implies that $t'_1 \Sigma_{12} t_2 = 0$ for all t_1 and t_2 , which in turn implies that $\Sigma_{12} = 0$. [This follows by setting $t_1 = (1, 0, \dots, 0)'$, etc.] \square

We use this theorem to prove our next result.

THEOREM 2.5 *Let $\mathbf{Y} \sim N_n(\mu, \Sigma)$ and define $\mathbf{U} = \mathbf{A}\mathbf{Y}$, $\mathbf{V} = \mathbf{B}\mathbf{Y}$. Then \mathbf{U} and \mathbf{V} are independent if and only if $\text{Cov}[\mathbf{U}, \mathbf{V}] = \mathbf{A}\Sigma\mathbf{B}' = \mathbf{0}$.*

Proof. Consider

$$\mathbf{W} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \mathbf{Y}.$$

Then, by Theorem 2.2, the random vector \mathbf{W} is multivariate normal with variance-covariance matrix

$$\text{Var}[\mathbf{W}] = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \text{Var}[\mathbf{Y}] (\mathbf{A}', \mathbf{B}') = \begin{pmatrix} \mathbf{A}\Sigma\mathbf{A}' & \mathbf{A}\Sigma\mathbf{B}' \\ \mathbf{B}\Sigma\mathbf{A}' & \mathbf{B}\Sigma\mathbf{B}' \end{pmatrix}.$$

Thus, by Theorem 2.4, \mathbf{U} and \mathbf{V} are independent if and only if $\mathbf{A}\Sigma\mathbf{B}' = \mathbf{0}$. \square

EXAMPLE 2.8 Let $\mathbf{Y} \sim N_n(\mu, \sigma^2 \mathbf{I}_n)$ and let $\mathbf{1}_n$ be an n -vector of 1's. Then the sample mean $\bar{\mathbf{Y}} = n^{-1} \sum_i Y_i$ is independent of the sample variance $S^2 = (n-1)^{-1} \sum_i (Y_i - \bar{Y})^2$. To see this, let $\mathbf{J}_n = \mathbf{1}_n \mathbf{1}'_n$ be the $n \times n$ matrix of 1's. Then $\bar{\mathbf{Y}} = n^{-1} \mathbf{1}'_n \mathbf{Y}$ ($= \mathbf{A}\mathbf{Y}$, say) and

$$\begin{pmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{pmatrix} = (\mathbf{I}_n - n^{-1} \mathbf{J}_n) \mathbf{Y} = \mathbf{B}\mathbf{Y},$$

say. Now

$$\mathbf{A}\Sigma\mathbf{B}' = n^{-1} \mathbf{1}'_n \sigma^2 \mathbf{I}_n (\mathbf{I}_n - n^{-1} \mathbf{J}_n) = \sigma^2 n^{-1} \mathbf{1}_n - \sigma^2 n^{-1} \mathbf{1}_n = \mathbf{0},$$

so by Theorem 2.5, $\bar{\mathbf{Y}}$ is independent of $(Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})$, and hence independent of S^2 . \square

EXAMPLE 2.9 Suppose that $\mathbf{Y} \sim N_n(\mu, \Sigma)$ with Σ positive-definite, and \mathbf{Y} is partitioned into two subvectors $\mathbf{Y}' = (\mathbf{Y}'_1, \mathbf{Y}'_2)$, where \mathbf{Y}_1 has dimension

r . Partition μ and Σ similarly. Then the conditional distribution of \mathbf{Y}_1 given $\mathbf{Y}_2 = \mathbf{y}_2$ is $N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

To derive this, put

$$\begin{aligned}\mathbf{U}_1 &= \mathbf{Y}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{Y}_2 - \mu_2), \\ \mathbf{U}_2 &= \mathbf{Y}_2 - \mu_2.\end{aligned}$$

Then

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 - \mu_1 \\ \mathbf{Y}_2 - \mu_2 \end{pmatrix} = \mathbf{A}(\mathbf{Y} - \mu),$$

so that \mathbf{U} is multivariate normal with mean $\mathbf{0}$ and variance matrix $\mathbf{A}\Sigma\mathbf{A}'$ given by

$$\begin{aligned}\begin{pmatrix} \mathbf{I}_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ -\Sigma_{22}^{-1}\Sigma_{21} & \mathbf{I}_{n-r} \end{pmatrix} \\ = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}.\end{aligned}$$

Hence, \mathbf{U}_1 and \mathbf{U}_2 are independent, with joint density of the form $g(\mathbf{u}_1, \mathbf{u}_2) = g_1(\mathbf{u}_1)g_2(\mathbf{u}_2)$.

Now consider the conditional density function of \mathbf{Y}_1 given \mathbf{Y}_2 :

$$f_{1|2}(\mathbf{y}_1 | \mathbf{y}_2) = f(\mathbf{y}_1, \mathbf{y}_2) / f_2(\mathbf{y}_2) \quad (2.8)$$

and write

$$\begin{aligned}\mathbf{u}_1 &= \mathbf{y}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2), \\ \mathbf{u}_2 &= \mathbf{y}_2 - \mu_2.\end{aligned}$$

By Exercises 2a, No. 2, $f_2(\mathbf{y}_2) = g_2(\mathbf{u}_2)$ and $f(\mathbf{y}_1, \mathbf{y}_2) = g_1(\mathbf{u}_1)g_2(\mathbf{u}_2)$, so that from (2.8), $f_{1|2}(\mathbf{y}_1 | \mathbf{y}_2) = g_1(\mathbf{u}_1) = g_1(\mathbf{y}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \mu_2))$. The result now follows from the fact that g_1 is the density of the $N_r(\mathbf{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ distribution. \square

EXERCISES 2c

- If Y_1, Y_2, \dots, Y_n have a multivariate normal distribution and are pairwise independent, are they mutually independent?
- Let $\mathbf{Y} \sim N_n(\mu \mathbf{1}_n, \Sigma)$, where $\Sigma = (1 - \rho)\mathbf{I}_n + \rho \mathbf{J}_n$ and $\rho > -1/(n - 1)$. When $\rho = 0$, \bar{Y} and $\sum_i (Y_i - \bar{Y})^2$ are independent, by Example 2.8. Are they independent when $\rho \neq 0$?

3. Given $\mathbf{Y} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix},$$

for what value(s) of ρ are $Y_1 + Y_2 + Y_3$ and $Y_1 - Y_2 - Y_3$ statistically independent?

2.4 DISTRIBUTION OF QUADRATIC FORMS

Quadratic forms in normal variables arise frequently in the theory of regression in connection with various tests of hypotheses. In this section we prove some simple results concerning the distribution of such quadratic forms.

Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive-definite. We are interested in the distribution of random variables of the form $\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} Y_i Y_j$. Note that we can always assume that the matrix \mathbf{A} is symmetric, since if not we can replace a_{ij} with $\frac{1}{2}(a_{ij} + a_{ji})$ without changing the value of the quadratic form. Since \mathbf{A} is symmetric, we can diagonalize it with an orthogonal transformation; that is, there is an orthogonal matrix \mathbf{T} and a diagonal matrix \mathbf{D} with

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{D} = \text{diag}(d_1, \dots, d_n). \quad (2.9)$$

The diagonal elements d_i are the eigenvalues of \mathbf{A} and can be any real numbers.

We begin by assuming that the random vector in the quadratic form has a $N_n(\mathbf{0}, \mathbf{I}_n)$ distribution. The general case can be reduced to this through the usual transformations. By Example 2.3, if \mathbf{T} is an orthogonal matrix and \mathbf{Y} has an $N_n(\mathbf{0}, \mathbf{I}_n)$ distribution, so does $\mathbf{Z} = \mathbf{T}'\mathbf{Y}$. Thus we can write

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{Y}'\mathbf{T}\mathbf{D}\mathbf{T}'\mathbf{Y} = \mathbf{Z}'\mathbf{D}\mathbf{Z} = \sum_{i=1}^n d_i Z_i^2, \quad (2.10)$$

so the distribution of $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ is a linear combination of independent χ_1^2 random variables. Given the values of d_i , it is possible to calculate the distribution, at least numerically. Farebrother [1990] describes algorithms for this.

There is an important special case that allows us to derive the distribution of the quadratic form exactly, without recourse to numerical methods. If r of the eigenvalues d_i are 1 and the remaining $n - r$ zero, then the distribution is the sum of r independent χ_1^2 's, which is χ_r^2 . We can recognize when the eigenvalues are zero or 1 using the following theorem.

THEOREM 2.6 *Let \mathbf{A} be a symmetric matrix. Then \mathbf{A} has r eigenvalues equal to 1 and the rest zero if and only if $\mathbf{A}^2 = \mathbf{A}$ and $\text{rank } \mathbf{A} = r$.*

Proof. See A.6.1. □

Matrices \mathbf{A} satisfying $\mathbf{A}^2 = \mathbf{A}$ are called *idempotent*. Thus, if \mathbf{A} is symmetric, idempotent, and has rank r , we have shown that the distribution of $\mathbf{Y}'\mathbf{AY}$ must be χ_r^2 . The converse is also true: If \mathbf{A} is symmetric and $\mathbf{Y}'\mathbf{AY}$ is χ_r^2 , then \mathbf{A} must be idempotent and have rank r . To prove this by Theorem 2.6, all we need to show is that r of the eigenvalues of \mathbf{A} are 1 and the rest are zero. By (2.10) and Exercises 2d, No. 1, the m.g.f. of $\mathbf{Y}'\mathbf{AY}$ is $\prod_{i=1}^n (1 - 2d_i t)^{-1/2}$. But since $\mathbf{Y}'\mathbf{AY}$ is χ_r^2 , the m.g.f. must also equal $(1 - 2t)^{-r/2}$. Thus

$$\prod_{i=1}^n (1 - 2d_i t) = (1 - 2t)^r,$$

so by the unique factorization of polynomials, r of the d_i are 1 and the rest are zero.

We summarize these results by stating them as a theorem.

THEOREM 2.7 *Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and let \mathbf{A} be a symmetric matrix. Then $\mathbf{Y}'\mathbf{AY}$ is χ_r^2 if and only if \mathbf{A} is idempotent of rank r .*

EXAMPLE 2.10 Let $\mathbf{Y} \sim N_n(\mu, \sigma^2 \mathbf{I}_n)$ and let S^2 be the sample variance as defined in Example 2.8. Then $(n - 1)S^2/\sigma^2 \sim \chi_{n-1}^2$. To see this, recall that $(n - 1)S^2/\sigma^2$ can be written as $\sigma^{-2}\mathbf{Y}'(\mathbf{I}_n - n^{-1}\mathbf{J}_n)\mathbf{Y}$. Now define $\mathbf{Z} = \sigma^{-1}(\mathbf{Y} - \mu\mathbf{1}_n)$, so that $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Then we have

$$(n - 1)S^2/\sigma^2 = \mathbf{Z}'(\mathbf{I}_n - n^{-1}\mathbf{J}_n)\mathbf{Z},$$

where the matrix $\mathbf{I}_n - n^{-1}\mathbf{J}_n$ is symmetric and idempotent, as can be verified by direct multiplication. To calculate its rank, we use the fact that for symmetric idempotent matrices, the rank and trace are the same (A.6.2). We get

$$\begin{aligned} \text{rank}(\mathbf{I}_n - n^{-1}\mathbf{J}_n) &= \text{tr}(\mathbf{I}_n - n^{-1}\mathbf{J}_n) \\ &= \text{tr}(\mathbf{I}_n) - n^{-1} \text{tr}(\mathbf{J}_n) \\ &= n - 1, \end{aligned}$$

so the result follows from Theorem 2.7. □

Our next two examples illustrate two very important additional properties of quadratic forms, which will be useful in Chapter 4.

EXAMPLE 2.11 Suppose that \mathbf{A} is symmetric and $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Then if $\mathbf{Y}'\mathbf{AY}$ is χ_r^2 , the quadratic form $\mathbf{Y}'(\mathbf{I}_n - \mathbf{A})\mathbf{Y}$ is χ_{n-r}^2 . This follows because \mathbf{A} must be idempotent, which implies that $(\mathbf{I}_n - \mathbf{A})$ is also idempotent. (Check by direct multiplication.) Furthermore,

$$\text{rank}(\mathbf{I}_n - \mathbf{A}) = \text{tr}(\mathbf{I}_n - \mathbf{A}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{A}) = n - r,$$

so that $\mathbf{Y}'(\mathbf{I}_n - \mathbf{A})\mathbf{Y}$ is χ^2_{n-r} . \square

EXAMPLE 2.12 Suppose that \mathbf{A} and \mathbf{B} are symmetric, $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, and $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are both chi-squared. Then $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independent if and only if $\mathbf{AB} = 0$.

To prove this, suppose first that $\mathbf{AB} = 0$. Since \mathbf{A} and \mathbf{B} are idempotent, we can write the quadratic forms as $\mathbf{Y}'\mathbf{AY} = \mathbf{YA}'\mathbf{AY} = \|\mathbf{AY}\|^2$ and $\mathbf{Y}'\mathbf{BY} = \|\mathbf{BY}\|^2$. By Theorem 2.5, \mathbf{AY} and \mathbf{BY} are independent, which implies that the quadratic forms are independent.

Conversely, suppose that the quadratic forms are independent. Then their sum is the sum of independent chi-squareds, which implies that $\mathbf{Y}'(\mathbf{A} + \mathbf{B})\mathbf{Y}$ is also chi-squared. Thus $\mathbf{A} + \mathbf{B}$ must be idempotent and

$$\mathbf{A} + \mathbf{B} = (\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2 = \mathbf{A} + \mathbf{AB} + \mathbf{BA} + \mathbf{B},$$

so that

$$\mathbf{AB} + \mathbf{BA} = 0.$$

Multiplying on the left by \mathbf{A} gives $\mathbf{AB} + \mathbf{ABA} = 0$, while multiplying on the right by \mathbf{A} gives $\mathbf{ABA} + \mathbf{BA} = 0$; hence $\mathbf{AB} = \mathbf{BA} = 0$. \square

EXAMPLE 2.13 (Hogg and Craig [1958, 1970]) Let $\mathbf{Y} \sim N_n(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n)$ and let $Q_i = (\mathbf{Y} - \boldsymbol{\theta})' \mathbf{P}_i (\mathbf{Y} - \boldsymbol{\theta}) / \sigma^2$ ($i = 1, 2$). We will show that if $Q_i \sim \chi^2_{r_i}$ and $Q_1 - Q_2 \geq 0$, then $Q_1 - Q_2$ and Q_2 are independently distributed as $\chi^2_{r_1 - r_2}$ and $\chi^2_{r_2}$, respectively.

We begin by noting that if $Q_i \sim \chi^2_{r_i}$, then $\mathbf{P}_i^2 = \mathbf{P}_i$ (Theorem 2.7). Also, $Q_1 - Q_2 \geq 0$ implies that $\mathbf{P}_1 - \mathbf{P}_2$ is positive-semidefinite and therefore idempotent (A.6.5). Hence, by Theorem 2.7, $Q_1 - Q_2 \sim \chi^2_r$, where

$$\begin{aligned} r &= \text{rank}(\mathbf{P}_1 - \mathbf{P}_2) \\ &= \text{tr}(\mathbf{P}_1 - \mathbf{P}_2) \\ &= \text{tr } \mathbf{P}_1 - \text{tr } \mathbf{P}_2 \\ &= \text{rank } \mathbf{P}_1 - \text{rank } \mathbf{P}_2 \\ &= r_1 - r_2. \end{aligned}$$

Also, by A.6.5, $\mathbf{P}_1 \mathbf{P}_2 = \mathbf{P}_2 \mathbf{P}_1 = \mathbf{P}_2$, and $(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{P}_2 = 0$. Therefore, since $\mathbf{Z} = (\mathbf{Y} - \boldsymbol{\theta})/\sigma^2 \sim N_n(\mathbf{0}, \mathbf{I}_n)$, we have, by Example 2.12, that $Q_1 - Q_2 [= \mathbf{Z}'(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{Z}]$ is independent of Q_2 ($= \mathbf{Z}'\mathbf{P}_2\mathbf{Z}$). \square

We can use these results to study the distribution of quadratic forms when the variance-covariance matrix Σ is any positive-semidefinite matrix. Suppose that \mathbf{Y} is now $N_n(\mathbf{0}, \Sigma)$, where Σ is of rank s ($s \leq n$). Then, by Definition 2.2 (Section 2.2), \mathbf{Y} has the same distribution as $\mathbf{R}\mathbf{Z}$, where $\Sigma = \mathbf{R}\mathbf{R}'$ and \mathbf{R} is $n \times s$ of rank s (A.3.3). Thus the distribution of $\mathbf{Y}'\mathbf{AY}$ is that of $\mathbf{Z}'\mathbf{R}'\mathbf{ARZ}$,

which, by Theorem 2.7, will be χ_r^2 if and only if $\mathbf{R}'\mathbf{A}\mathbf{R}$ is idempotent of rank r . However, this is not a very useful condition. A better one is contained in our next theorem.

THEOREM 2.8 *Suppose that $\mathbf{Y} \sim N_n(\mathbf{0}, \Sigma)$, and \mathbf{A} is symmetric. Then $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ is χ_r^2 if and only if r of the eigenvalues of $\mathbf{A}\Sigma$ are 1 and the rest are zero.*

Proof. We assume that $\mathbf{Y}'\mathbf{A}\mathbf{Y} = \mathbf{Z}'\mathbf{R}'\mathbf{A}\mathbf{R}\mathbf{Z}$ is χ_r^2 . Then $\mathbf{R}'\mathbf{A}\mathbf{R}$ is symmetric and idempotent with r unit eigenvalues and the rest zero (by A.6.1), and its rank equals its trace (A.6.2). Hence, by (A.1.2),

$$r = \text{rank}(\mathbf{R}'\mathbf{A}\mathbf{R}) = \text{tr}(\mathbf{R}'\mathbf{A}\mathbf{R}) = \text{tr}(\mathbf{A}\mathbf{R}\mathbf{R}') = \text{tr}(\mathbf{A}\Sigma).$$

Now, by (A.7.1), $\mathbf{R}'\mathbf{A}\mathbf{R}$ and $\mathbf{A}\mathbf{R}\mathbf{R}' = \mathbf{A}\Sigma$ have the same eigenvalues, with possibly different multiplicities. Hence the eigenvalues of $\mathbf{A}\Sigma$ are 1 or zero. As the trace of any square matrix equals the sum of its eigenvalues (A.1.3), r of the eigenvalues of $\mathbf{A}\Sigma$ must be 1 and the rest zero. The converse argument is just the reverse of the one above. \square

For nonsymmetric matrices, idempotence implies that the eigenvalues are zero or 1, but the converse is not true. However, when Σ (and hence \mathbf{R}) has full rank, the fact that $\mathbf{R}'\mathbf{A}\mathbf{R}$ is idempotent implies that $\mathbf{A}\Sigma$ is idempotent. This is because the equation

$$\mathbf{R}'\mathbf{A}\mathbf{R}\mathbf{R}'\mathbf{A}\mathbf{R} = \mathbf{R}'\mathbf{A}\mathbf{R}$$

can be premultiplied by $(\mathbf{R}')^{-1}$ and postmultiplied by \mathbf{R}' to give

$$\mathbf{A}\Sigma\mathbf{A}\Sigma = \mathbf{A}\Sigma.$$

Thus we have the following corollary to Theorem 2.8.

COROLLARY Let $\mathbf{Y} \sim N_n(\mathbf{0}, \Sigma)$, where Σ is positive-definite, and suppose that \mathbf{A} is symmetric. Then $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ is χ_r^2 if and only $\mathbf{A}\Sigma$ is idempotent and has rank r .

For other necessary and sufficient conditions, see Good [1969, 1970] and Khatri [1978].

Our final theorem concerns a very special quadratic form that arises frequently in statistics.

THEOREM 2.9 *Suppose that $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$, where Σ is positive-definite. Then $Q = (\mathbf{Y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi_n^2$.*

Proof. Making the transformation $\mathbf{Y} = \Sigma^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ considered in Theorem 2.1, we get

$$Q = \mathbf{Z}'\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}\mathbf{Z} = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^n Z_i^2.$$

Since the Z_i^2 's are independent χ_1^2 variables, $Q \sim \chi_n^2$. \square

EXERCISES 2d

1. Show that the m.g.f. for (2.10) is $\prod_1^n (1 - 2td_i)^{-1/2}$.
2. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and let \mathbf{A} be symmetric.
 - (a) Show that the m.g.f. of $\mathbf{Y}'\mathbf{AY}$ is $[\det(\mathbf{I}_n - 2t\mathbf{A})]^{-1/2}$.
 - (b) If \mathbf{A} is idempotent of rank r , show that the m.g.f. is $(1 - 2t)^{-r/2}$.
 - (c) Find the m.g.f. if $\mathbf{Y} \sim N_n(\mathbf{0}, \Sigma)$.
3. If $\mathbf{Y} \sim N_2(\mathbf{0}, \mathbf{I}_2)$, find values of a and b such that

$$a(Y_1 - Y_2)^2 + b(Y_1 + Y_2)^2 \sim \chi_2^2.$$

4. Suppose that $\mathbf{Y} \sim N_3(\mathbf{0}, \mathbf{I}_n)$. Show that

$$\frac{1}{3} [(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + (Y_3 - Y_1)^2]$$

has a χ_2^2 distribution. Does some multiple of

$$(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + \cdots + (Y_{n-1} - Y_n)^2 + (Y_n - Y_1)^2$$

have a chi-squared distribution for general n ?

5. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and let \mathbf{A} and \mathbf{B} be symmetric. Show that the joint m.g.f. of $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ is $[\det(\mathbf{I}_n - 2s\mathbf{A} - 2t\mathbf{B})]^{-1/2}$. Hence show that the two quadratic forms are independent if $\mathbf{AB} = \mathbf{0}$.

MISCELLANEOUS EXERCISES 2

1. Suppose that $\varepsilon \sim N_3(\mathbf{0}, \sigma^2 \mathbf{I}_3)$ and that Y_0 is $N(0, \sigma_0^2)$, independently of the ε_i 's. Define

$$Y_i = \rho Y_{i-1} + \varepsilon_i \quad (i = 1, 2, 3).$$

- (a) Find the variance-covariance matrix of $\mathbf{Y} = (Y_1, Y_2, Y_3)'$.
- (b) What is the distribution of \mathbf{Y} ?

2. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, and put $\mathbf{X} = \mathbf{AY}$, $\mathbf{U} = \mathbf{BY}$ and $\mathbf{V} = \mathbf{CY}$. Suppose that $\text{Cov}[\mathbf{X}, \mathbf{U}] = \mathbf{0}$ and $\text{Cov}[\mathbf{X}, \mathbf{V}] = \mathbf{0}$. Show that \mathbf{X} is independent of $\mathbf{U} + \mathbf{V}$.

3. If Y_1, Y_2, \dots, Y_n is a random sample from $N(\mu, \sigma^2)$, prove that \bar{Y} is independent of $\sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2$.

4. If \mathbf{X} and \mathbf{Y} are n -dimensional vectors with independent multivariate normal distributions, prove that $a\mathbf{X} + b\mathbf{Y}$ is also multivariate normal.

5. If $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and \mathbf{a} is a nonzero vector, show that the conditional distribution of $\mathbf{Y}'\mathbf{Y}$ given $\mathbf{a}'\mathbf{Y} = 0$ is χ_{n-1}^2 .
6. Let $\mathbf{Y} \sim N_n(\mu\mathbf{1}_n, \Sigma)$, where $\Sigma = (1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}'_n$ and $\rho > -1/(n-1)$. Show that $\sum_i(Y_i - \bar{Y})^2/(1-\rho)$ is χ_{n-1}^2 .
7. Let \mathbf{Y}_i , $i = 1, \dots, n$, be independent $N_p(\boldsymbol{\mu}, \Sigma)$ random vectors. Show that

$$\frac{1}{n-1} \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})'$$

is an unbiased estimate of Σ .

8. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and let \mathbf{A} and \mathbf{B} be symmetric idempotent matrices with $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$. Show that $\mathbf{Y}'\mathbf{AY}$, $\mathbf{Y}'\mathbf{BY}$ and $\mathbf{Y}'(\mathbf{I}_n - \mathbf{A} - \mathbf{B})\mathbf{Y}$ have independent chi-square distributions.
9. Let (X_i, Y_i) , $i = 1, 2, \dots, n$, be a random sample from a bivariate normal distribution, with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 , and correlation ρ , and let

$$\mathbf{W} = (\mathbf{X}', \mathbf{Y}')' = (X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_n)'.$$

- (a) Show that \mathbf{W} has a $N_{2n}(\boldsymbol{\mu}, \Sigma)$ distribution, where

$$\boldsymbol{\mu} = (\mu_1 \mathbf{1}'_n, \mu_2 \mathbf{1}'_n)' \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 \mathbf{I}_n & \rho \sigma_1 \sigma_2 \mathbf{I}_n \\ \rho \sigma_1 \sigma_2 \mathbf{I}_n & \sigma_2^2 \mathbf{I}_n \end{pmatrix}.$$

- (b) Find the conditional distribution of \mathbf{X} given \mathbf{Y} .

10. If $\mathbf{Y} \sim N_2(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})$, prove that

$$\left(\mathbf{Y}' \Sigma^{-1} \mathbf{Y} - \frac{Y_1^2}{\sigma_{11}} \right) \sim \chi_1^2.$$

11. Let a_0, a_1, \dots, a_n be independent $N(\mathbf{0}, \sigma^2)$ random variables and define

$$Y_i = a_i + \phi a_{i-1} \quad (i = 1, 2, \dots, n).$$

Show that $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ has a multivariate normal distribution and find its variance-covariance matrix. (The sequence Y_1, Y_2, \dots is called a *moving average process of order one* and is a commonly used model in time series analysis.)

12. Suppose that $\mathbf{Y} \sim N_3(\mathbf{0}, \mathbf{I}_n)$. Find the m.g.f. of $2(Y_1 Y_2 - Y_2 Y_3 - Y_3 Y_1)$. Hence show that this random variable has the same distribution as that of $2U_1 - U_2 - U_3$, where the U_i 's are independent χ_1^2 random variables.

13. Theorem 2.3 can be used as a definition of the multivariate normal distribution. If so, deduce from this definition the following results:
- If Z_1, Z_2, \dots, Z_n are i.i.d. $N(0, 1)$, then $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.
 - If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then \mathbf{Y} has m.g.f. (2.5).
 - If $\boldsymbol{\Sigma}$ is positive-definite, prove that \mathbf{Y} has density function (2.1).
14. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ be a vector of n random variables ($n > 3$) with density function

$$f(\mathbf{y}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n y_i^2\right) \left\{ 1 + \prod_{i=1}^n [y_i \exp(-\frac{1}{2} y_i^2)] \right\},$$

$$-\infty < y_i < \infty \quad (i = 1, 2, \dots, n).$$

Prove that any subset of $n - 1$ random variables are mutually independent $N(0, 1)$ variables.

(Pierce and Dykstra [1969])

15. Suppose that $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)' \sim N_4(\mathbf{0}, \mathbf{I}_4)$, and let $Q = Y_1 Y_2 - Y_3 Y_4$.
- Prove that Q does not have a chi-square distribution.
 - Find the m.g.f. of Q .
16. If $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, find the variance of
- $$(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + \cdots + (Y_{n-1} - Y_n)^2.$$

17. Given $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, prove that

$$\text{var}[\mathbf{Y}' \mathbf{A} \mathbf{Y}] = 2 \text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma}) + 4 \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu}.$$

3

Linear Regression: Estimation and Distribution Theory

3.1 LEAST SQUARES ESTIMATION

Let Y be a random variable that fluctuates about an unknown parameter η ; that is, $Y = \eta + \varepsilon$, where ε is the fluctuation or *error*. For example, ε may be a “natural” fluctuation inherent in the experiment which gives rise to η , or it may represent the error in measuring η , so that η is the true response and Y is the observed response. As noted in Chapter 1, our focus is on linear models, so we assume that η can be expressed in the form

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1},$$

where the *explanatory* variables x_1, x_2, \dots, x_{p-1} are known constants (e.g., experimental variables that are controlled by the experimenter and are measured with negligible error), and the β_j ($j = 0, 1, \dots, p - 1$) are unknown parameters to be estimated. If the x_j are varied and n values, Y_1, Y_2, \dots, Y_n , of Y are observed, then

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad (i = 1, 2, \dots, n), \quad (3.1)$$

where x_{ij} is the i th value of x_j . Writing these n equations in matrix form, we have

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.2)$$

where $x_{10} = x_{20} = \dots = x_{n0} = 1$. The $n \times p$ matrix \mathbf{X} will be called the *regression matrix*, and the x_{ij} 's are generally chosen so that the columns of \mathbf{X} are linearly independent; that is, \mathbf{X} has rank p , and we say that \mathbf{X} has *full rank*. However, in some experimental design situations, the elements of \mathbf{X} are chosen to be 0 or 1, and the columns of \mathbf{X} may be linearly dependent. In this case \mathbf{X} is commonly called the *design matrix*, and we say that \mathbf{X} has less than full rank.

It has been the custom in the past to call the x_j 's the independent variables and Y the dependent variable. However, this terminology is confusing, so we follow the more contemporary usage as in Chapter 1 and refer to x_j as a *explanatory variable* or *regressor* and Y as the *response variable*.

As we mentioned in Chapter 1, (3.1) is a very general model. For example, setting $x_{ij} = x_i^j$ and $k = p - 1$, we have the polynomial model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_k x_i^k + \varepsilon_i.$$

Again,

$$Y_i = \beta_0 + \beta_1 e^{w_{i1}} + \beta_2 w_{i1} w_{i2} + \beta_3 \sin w_{i3} + \varepsilon_i$$

is also a special case. The essential aspect of (3.1) is that it is linear in the unknown parameters β_j ; for this reason it is called a *linear model*. In contrast,

$$Y_i = \beta_0 + \beta_1 e^{-\beta_2 x_i} + \varepsilon_i$$

is a nonlinear model, being nonlinear in β_2 .

Before considering the problem of estimating β , we note that all the theory in this and subsequent chapters is developed for the model (3.2), where x_{i0} is not necessarily constrained to be unity. In the case where $x_{i0} \neq 1$, the reader may question the use of a notation in which i runs from 0 to $p - 1$ rather than 1 to p . However, since the major application of the theory is to the case $x_{i0} \equiv 1$, it is convenient to "separate" β_0 from the other β_j 's right from the outset. We shall assume the latter case until stated otherwise.

One method of obtaining an estimate of β is the method of least squares. This method consists of minimizing $\sum_i \varepsilon_i^2$ with respect to β ; that is, setting $\theta = \mathbf{X}\beta$, we minimize $\varepsilon' \varepsilon = \|\mathbf{Y} - \theta\|^2$ subject to $\theta \in \mathcal{C}(\mathbf{X}) = \Omega$, where Ω is the column space of \mathbf{X} ($= \{y : y = \mathbf{X}x \text{ for any } x\}$). If we let θ vary in Ω , $\|\mathbf{Y} - \theta\|^2$ (the square of the length of $\mathbf{Y} - \theta$) will be a minimum for $\theta = \hat{\theta}$ when $(\mathbf{Y} - \hat{\theta}) \perp \Omega$ (cf. Figure 3.1). This is obvious geometrically, and it is readily proved algebraically as follows.

We first note that $\hat{\theta}$ can be obtained via a symmetric idempotent (projection) matrix \mathbf{P} , namely $\hat{\theta} = \mathbf{PY}$, where \mathbf{P} represents the orthogonal projection onto Ω (see Appendix B). Then

$$\mathbf{Y} - \theta = (\mathbf{Y} - \hat{\theta}) + (\hat{\theta} - \theta),$$

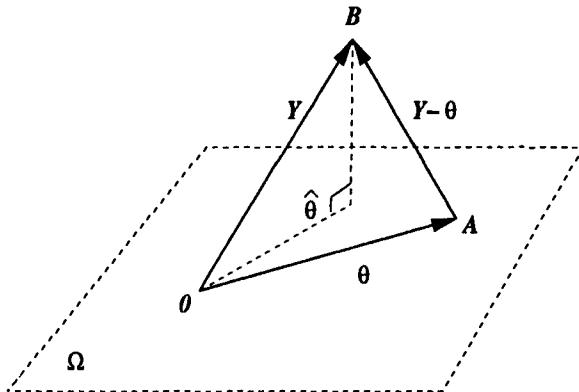


Fig. 3.1 The method of least squares consists of finding A such that AB is a minimum.

where from $P\theta = \theta$, $P' = P$ and $P^2 = P$, we have

$$\begin{aligned} (\mathbf{Y} - \hat{\theta})'(\hat{\theta} - \theta) &= (\mathbf{Y} - P\mathbf{Y})'P(\mathbf{Y} - \theta) \\ &= \mathbf{Y}'(I_n - P)P(\mathbf{Y} - \theta) \\ &= 0. \end{aligned}$$

Hence

$$\begin{aligned} \|\mathbf{Y} - \theta\|^2 &= \|\mathbf{Y} - \hat{\theta}\|^2 + \|\hat{\theta} - \theta\|^2 \\ &\geq \|\mathbf{Y} - \hat{\theta}\|^2, \end{aligned}$$

with equality if and only if $\theta = \hat{\theta}$. Since $\mathbf{Y} - \hat{\theta}$ is perpendicular to Ω ,

$$\mathbf{X}'(\mathbf{Y} - \hat{\theta}) = 0$$

or

$$\mathbf{X}'\hat{\theta} = \mathbf{X}'\mathbf{Y}. \quad (3.3)$$

Here $\hat{\theta}$ is uniquely determined, being the *unique* orthogonal projection of \mathbf{Y} onto Ω (see Appendix B).

We now assume that the columns of \mathbf{X} are linearly independent so that there exists a unique vector $\hat{\beta}$ such that $\hat{\theta} = \mathbf{X}\hat{\beta}$. Then substituting in (3.3), we have

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}, \quad (3.4)$$

the *normal equations*. As \mathbf{X} has rank p , $\mathbf{X}'\mathbf{X}$ is positive-definite (A.4.6) and therefore nonsingular. Hence (3.4) has a unique solution, namely,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (3.5)$$

Here $\hat{\beta}$ is called the (ordinary) *least squares estimate* of β , and computational methods for actually calculating the estimate are given in Chapter 11.

We note that $\hat{\beta}$ can also be obtained by writing

$$\begin{aligned}\epsilon' \epsilon &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

[using the fact that $\beta'\mathbf{X}'\mathbf{Y} = (\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$] and differentiating $\epsilon' \epsilon$ with respect to β . Thus from $\partial\epsilon' \epsilon / \partial\beta = 0$ we have (A.8)

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta = 0 \quad (3.6)$$

or

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}.$$

This solution for β gives us a stationary value of $\epsilon' \epsilon$, and a simple algebraic identity (see Exercises 3a, No. 1) confirms that $\hat{\beta}$ is a minimum.

In addition to the method of least squares, several other methods are used for estimating β . These are described in Section 3.13.

Suppose now that the columns of \mathbf{X} are not linearly independent. For a particular $\hat{\theta}$ there is no longer a unique $\hat{\beta}$ such that $\hat{\theta} = \mathbf{X}\hat{\beta}$, and (3.4) does not have a unique solution. However, a solution is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ is any generalized inverse of $(\mathbf{X}'\mathbf{X})$ (see A.10). Then

$$\hat{\theta} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y},$$

and since \mathbf{P} is unique, it follows that \mathbf{P} does not depend on which generalized inverse is used.

We denote the *fitted values* $\mathbf{X}\hat{\beta}$ by $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)'$. The elements of the vector

$$\begin{aligned}\mathbf{Y} - \hat{\mathbf{Y}} &= \mathbf{Y} - \mathbf{X}\hat{\beta} \\ &= (\mathbf{I}_n - \mathbf{P})\mathbf{Y}, \quad \text{say,}\end{aligned} \quad (3.7)$$

are called the *residuals* and are denoted by \mathbf{e} . The minimum value of $\epsilon' \epsilon$, namely

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'[\mathbf{X}'\mathbf{X}\hat{\beta} - \mathbf{X}'\mathbf{Y}] \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \quad [\text{by (3.4)}], \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta},\end{aligned} \quad (3.8)$$

is called the *residual sum of squares* (RSS). As $\hat{\theta} = \mathbf{X}\hat{\beta}$ is unique, we note that $\hat{\mathbf{Y}}$, \mathbf{e} , and RSS are unique, irrespective of the rank of \mathbf{X} .

EXAMPLE 3.1 Let Y_1 and Y_2 be independent random variables with means α and 2α , respectively. We will now find the least squares estimate of α and the residual sum of squares using both (3.5) and direct differentiation as in (3.6). Writing

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \alpha + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

we have $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, where $\mathbf{X} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\beta = \alpha$. Hence, by the theory above,

$$\begin{aligned} \hat{\alpha} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \left\{ (1, 2) \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}^{-1} (1, 2)\mathbf{Y} \\ &= \frac{1}{5}(1, 2) \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= \frac{1}{5}(Y_1 + 2Y_2) \end{aligned}$$

and

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\alpha}(Y_1 + 2Y_2) \\ &= Y_1^2 + Y_2^2 - \frac{1}{5}(Y_1 + 2Y_2)^2. \end{aligned}$$

We note that

$$\mathbf{P} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \left\{ (1, 2) \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}^{-1} (1, 2) = \frac{1}{5} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

The problem can also be solved by first principles as follows: $\mathbf{e}'\mathbf{e} = (Y_1 - \alpha)^2 + (Y_2 - 2\alpha)^2$ and $\partial\mathbf{e}'\mathbf{e}/\partial\alpha = 0$ implies that $\hat{\alpha} = \frac{1}{5}(Y_1 + 2Y_2)$. Further,

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (Y_1 - \hat{\alpha})^2 + (Y_2 - 2\hat{\alpha})^2 \\ &= Y_1^2 + Y_2^2 - \hat{\alpha}(2Y_1 + 4Y_2) + 5\hat{\alpha}^2 \\ &= Y_1^2 + Y_2^2 - \frac{1}{5}(Y_1 + 2Y_2)^2. \end{aligned}$$

In practice, both approaches are used. □

EXAMPLE 3.2 Suppose that Y_1, Y_2, \dots, Y_n all have mean β . Then the least squares estimate of β is found by minimizing $\sum_i(Y_i - \beta)^2$ with respect to β . This leads readily to $\hat{\beta} = \bar{Y}$. Alternatively, we can express the observations in terms of the regression model

$$\mathbf{Y} = \mathbf{1}_n\beta + \boldsymbol{\varepsilon},$$

where $\mathbf{1}_n$ is an n -dimensional column of 1's. Then

$$\hat{\beta} = (\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n \mathbf{Y} = \frac{1}{n} \mathbf{1}'_n \mathbf{Y} = \bar{Y}.$$

Also,

$$\mathbf{P} = \mathbf{1}_n (\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n = \frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \frac{1}{n} \mathbf{J}_n. \quad \square$$

We have emphasized that \mathbf{P} is the linear transformation representing the orthogonal projection of n -dimensional Euclidean space, \mathfrak{R}_n , onto Ω , the space spanned by the columns of \mathbf{X} . Similarly, $\mathbf{I}_n - \mathbf{P}$ represents the orthogonal projection of \mathfrak{R}_n onto the orthogonal complement, Ω^\perp , of Ω . Thus $\mathbf{Y} = \mathbf{PY} + (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$ represents a unique orthogonal decomposition of \mathbf{Y} into two components, one in Ω and the other in Ω^\perp . Some basic properties of \mathbf{P} and $(\mathbf{I}_n - \mathbf{P})$ are proved in Theorem 3.1 and its corollary, although these properties follow directly from the more general results concerning orthogonal projections stated in Appendix B. For a more abstract setting, see Seber [1980].

THEOREM 3.1 Suppose that \mathbf{X} is $n \times p$ of rank p , so that $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then the following hold.

- (i) \mathbf{P} and $\mathbf{I}_n - \mathbf{P}$ are symmetric and idempotent.
- (ii) $\text{rank}(\mathbf{I}_n - \mathbf{P}) = \text{tr}(\mathbf{I}_n - \mathbf{P}) = n - p$.
- (iii) $\mathbf{PX} = \mathbf{X}$.

Proof. (i) \mathbf{P} is obviously symmetric and $(\mathbf{I}_n - \mathbf{P})' = \mathbf{I}_n - \mathbf{P}' = \mathbf{I}_n - \mathbf{P}$. Also,

$$\begin{aligned} \mathbf{P}^2 &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{XI}_p(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}, \end{aligned}$$

and $(\mathbf{I}_n - \mathbf{P})^2 = \mathbf{I}_n - 2\mathbf{P} + \mathbf{P}^2 = \mathbf{I}_n - \mathbf{P}$.

(ii) Since $\mathbf{I}_n - \mathbf{P}$ is symmetric and idempotent, we have, by A.6.2,

$$\begin{aligned} \text{rank}(\mathbf{I}_n - \mathbf{P}) &= \text{tr}(\mathbf{I}_n - \mathbf{P}) \\ &= n - \text{tr}(\mathbf{P}), \end{aligned}$$

where

$$\begin{aligned} \text{tr}(\mathbf{P}) &= \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \\ &= \text{tr}[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \quad (\text{by A.1.2}) \\ &= \text{tr}(\mathbf{I}_p) \\ &= p. \end{aligned}$$

$$(iii) \mathbf{P}\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}. \quad \square$$

COROLLARY If \mathbf{X} has rank r ($r < p$), then Theorem 3.1 still holds, but with p replaced by r .

Proof. Let \mathbf{X}_1 be an $n \times r$ matrix with r linearly independent columns and having the same column space as \mathbf{X} [i.e., $\mathcal{C}(\mathbf{X}_1) = \Omega$]. Then $\mathbf{P} = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$, and (i) and (ii) follow immediately. We can find a matrix \mathbf{L} such that $\mathbf{X} = \mathbf{X}_1\mathbf{L}$, which implies that (cf. Exercises 3j, No. 2)

$$\mathbf{P}\mathbf{X} = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_1\mathbf{L} = \mathbf{X}_1\mathbf{L} = \mathbf{X},$$

which is (iii). \square

EXERCISES 3a

1. Show that if \mathbf{X} has full rank,

$$(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta),$$

and hence deduce that the left side is minimized uniquely when $\beta = \hat{\beta}$.

2. If \mathbf{X} has full rank, prove that $\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$. *Hint:* Consider the first column of \mathbf{X} .

3. Let

$$\begin{aligned} Y_1 &= \theta + \varepsilon_1 \\ Y_2 &= 2\theta - \phi + \varepsilon_2 \\ Y_3 &= \theta + 2\phi + \varepsilon_3, \end{aligned}$$

where $E[\varepsilon_i] = 0$ ($i = 1, 2, 3$). Find the least squares estimates of θ and ϕ .

4. Consider the regression model

$$E[Y_i] = \beta_0 + \beta_1 x_i + \beta_2 (3x_i^2 - 2) \quad (i = 1, 2, 3),$$

where $x_1 = -1$, $x_2 = 0$, and $x_3 = +1$. Find the least squares estimates of β_0 , β_1 , and β_2 . Show that the least squares estimates of β_0 and β_1 are unchanged if $\beta_2 = 0$.

5. The tension T observed in a nonextensible string required to maintain a body of unknown weight w in equilibrium on a smooth inclined plane of angle θ ($0 < \theta < \pi/2$) is a random variable with mean $E[T] = w \sin \theta$. If for $\theta = \theta_i$ ($i = 1, 2, \dots, n$) the corresponding values of T are T_i ($i = 1, 2, \dots, n$), find the least squares estimate of w .
6. If \mathbf{X} has full rank, so that $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, prove that $\mathcal{C}(\mathbf{P}) = \mathcal{C}(\mathbf{X})$.

7. For a general regression model in which \mathbf{X} may or may not have full rank, show that

$$\sum_{i=1}^n \hat{Y}_i(Y_i - \hat{Y}_i) = 0.$$

8. Suppose that we scale the explanatory variables so that $x_{ij} = k_j w_{ij}$ for all i, j . By expressing \mathbf{X} in terms of a new matrix \mathbf{W} , prove that $\hat{\mathbf{Y}}$ remains unchanged under this change of scale.

3.2 PROPERTIES OF LEAST SQUARES ESTIMATES

If we assume that the errors are unbiased (i.e., $E[\varepsilon] = \mathbf{0}$), and the columns of \mathbf{X} are linearly independent, then

$$\begin{aligned} E[\hat{\beta}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta, \end{aligned} \tag{3.10}$$

and $\hat{\beta}$ is an unbiased estimate of β . If we assume further that the ε_i are uncorrelated and have the same variance, that is, $\text{cov}[\varepsilon_i, \varepsilon_j] = \delta_{ij}\sigma^2$, then $\text{Var}[\varepsilon] = \sigma^2\mathbf{I}_n$ and

$$\text{Var}[\mathbf{Y}] = \text{Var}[\mathbf{Y} - \mathbf{X}\beta] = \text{Var}[\varepsilon].$$

Hence, by (1.7),

$$\begin{aligned} \text{Var}[\hat{\beta}] &= \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}[\mathbf{Y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{3.11}$$

The question now arises as to why we chose $\hat{\beta}$ as our estimate of β and not some other estimate. We show below that for a reasonable class of estimates, $\hat{\beta}_j$ is the estimate of β_j with the smallest variance. Here $\hat{\beta}_j$ can be extracted from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})'$ simply by premultiplying by the row vector \mathbf{c}' , which contains unity in the $(j+1)$ th position and zeros elsewhere. It transpires that this special property of $\hat{\beta}_j$ can be generalized to the case of any linear combination $\mathbf{a}'\hat{\beta}$ using the following theorem.

THEOREM 3.2 *Let $\hat{\theta}$ be the least squares estimate of $\theta = \mathbf{X}\beta$, where $\theta \in \Omega = \mathcal{C}(\mathbf{X})$ and \mathbf{X} may not have full rank. Then among the class of linear unbiased estimates of $\mathbf{c}'\theta$, $\mathbf{c}'\hat{\theta}$ is the unique estimate with minimum variance. [We say that $\mathbf{c}'\hat{\theta}$ is the best linear unbiased estimate (BLUE) of $\mathbf{c}'\theta$.]*

Proof. From Section 3.1, $\hat{\theta} = PY$, where $P\theta = PX\beta = X\beta = \theta$ (Theorem 3.1, Corollary). Hence $E[c'\hat{\theta}] = c'P\theta = c'\theta$ for all $\theta \in \Omega$, so that $c'\hat{\theta}$ [$= (Pc)'Y$] is a linear unbiased estimate of $c'\theta$. Let $d'Y$ be any other linear unbiased estimate of $c'\theta$. Then $c'\theta = E[d'Y] = d'\theta$ or $(c - d)'\theta = 0$, so that $(c - d) \perp \Omega$. Therefore, $P(c - d) = 0$ and $Pc = Pd$.

Now

$$\begin{aligned}\text{var}[c'\hat{\theta}] &= \text{var}[(Pc)'Y] \\ &= \text{var}[(Pd)'Y] \\ &= \sigma^2 d'P'Pd \\ &= \sigma^2 d'P^2d \\ &= \sigma^2 d'Pd \quad (\text{Theorem 3.1})\end{aligned}$$

so that

$$\begin{aligned}\text{var}[d'Y] - \text{var}[c'\hat{\theta}] &= \text{var}[d'Y] - \text{var}[(Pd)'Y] \\ &= \sigma^2(d'd - d'Pd) \\ &= \sigma^2 d'(I_n - P)d \\ &= \sigma^2 d'(I_n - P)'(I_n - P)d \\ &= \sigma^2 d'_1 d_1, \quad \text{say,} \\ &\geq 0,\end{aligned}$$

with equality only if $(I_n - P)d = 0$ or $d = Pd = Pc$. Hence $c'\hat{\theta}$ has minimum variance and is unique. \square

COROLLARY If X has full rank, then $a'\hat{\beta}$ is the BLUE of $a'\beta$ for every vector a .

Proof. Now $\theta = X\beta$ implies that $\beta = (X'X)^{-1}X'\theta$ and $\hat{\beta} = (X'X)^{-1}X'\hat{\theta}$. Hence setting $c' = a'(X'X)^{-1}X'$ we have that $a'\hat{\beta}$ ($= c'\hat{\theta}$) is the BLUE of $a'\beta$ ($= c'\theta$) for every vector a . \square

Thus far we have not made any assumptions about the distribution of the ε_i . However, when the ε_i are independently and identically distributed as $N(0, \sigma^2)$, that is, $\varepsilon \sim N(0, \sigma^2 I_n)$ or, equivalently, $Y \sim N_n(X\beta, \sigma^2 I_n)$, then $a'\hat{\beta}$ has minimum variance for the entire class of unbiased estimates, not just for linear estimates (cf. Rao [1973: p. 319] for a proof). In particular, $\hat{\beta}_i$, which is also the maximum likelihood estimate of β_i (Section 3.5), is the most efficient estimate of β_i .

When the common underlying distribution of the ε_i is not normal, then the least squares estimate of β_i is not the same as the asymptotically most efficient maximum likelihood estimate. The asymptotic efficiency of the least squares estimate is, for this case, derived by Cox and Hinkley [1968].

Eicker [1963] has discussed the question of the consistency and asymptotic normality of $\hat{\beta}$ as $n \rightarrow \infty$. Under weak restrictions he shows that $\hat{\beta}$ is a

consistent estimate of β if and only if the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$ tends to infinity. This condition on the smallest eigenvalue is a mild one, so that the result has wide applicability. Eicker also proves a theorem giving necessary and sufficient conditions for the asymptotic normality of each $\hat{\beta}_j$ (see Anderson [1971: pp. 23–27]).

EXERCISES 3b

- Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), where $E[\varepsilon] = \mathbf{0}$ and $\text{Var}[\varepsilon] = \sigma^2 \mathbf{I}_n$. Find the least squares estimates of β_0 and β_1 . Prove that they are uncorrelated if and only if $\bar{x} = 0$.
- In order to estimate two parameters θ and ϕ it is possible to make observations of three types: (a) the first type have expectation θ , (b) the second type have expectation $\theta + \phi$, and (c) the third type have expectation $\theta - 2\phi$. All observations are subject to uncorrelated errors of mean zero and constant variance. If m observations of type (a), m observations of (b), and n observations of type (c) are made, find the least squares estimates $\hat{\theta}$ and $\hat{\phi}$. Prove that these estimates are uncorrelated if $m = 2n$.
- Let Y_1, Y_2, \dots, Y_n be a random sample from $N(\theta, \sigma^2)$. Find the linear unbiased estimate of θ with minimum variance.
- Let

$$Y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where $\bar{x}_j = \sum_{i=1}^n x_{ij}/n$, $E[\varepsilon] = \mathbf{0}$, and $\text{Var}[\varepsilon] = \sigma^2 \mathbf{I}_n$. If $\hat{\beta}_1$ is the least squares estimate of β_1 , show that

$$\text{var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_i (x_{i1} - \bar{x}_1)^2(1 - r^2)},$$

where r is the correlation coefficient of the n pairs (x_{i1}, x_{i2}) .

3.3 UNBIASED ESTIMATION OF σ^2

We now focus our attention on σ^2 ($= \text{var}[\varepsilon_i]$). An unbiased estimate is described in the following theorem.

THEOREM 3.3 *If $E[\mathbf{Y}] = \mathbf{X}\beta$, where \mathbf{X} is an $n \times p$ matrix of rank r ($r \leq p$), and $\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$, then*

$$S^2 = \frac{(\mathbf{Y} - \hat{\theta})'(\mathbf{Y} - \hat{\theta})}{n - r} = \frac{RSS}{n - r}$$

is an unbiased estimate of σ^2 .

Proof. Consider the full-rank representation $\theta = \mathbf{X}_1\alpha$, where \mathbf{X}_1 is $n \times r$ of rank r . Then

$$\mathbf{Y} - \hat{\theta} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y},$$

where $\mathbf{P} = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$. From Theorem 3.1 we have

$$\begin{aligned}(n-r)S^2 &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})^2\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}.\end{aligned}\tag{3.12}$$

Since $\mathbf{P}\theta = \theta$, it follows from Theorems 1.5 and 3.1(iii) applied to \mathbf{X}_1 that

$$\begin{aligned}E[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] &= \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{P}) + \theta'(\mathbf{I}_n - \mathbf{P})\theta \\ &= \sigma^2(n-r),\end{aligned}$$

and hence $E[S^2] = \sigma^2$. \square

When \mathbf{X} has full rank, $S^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})/(n-p)$. In this case it transpires that S^2 , like $\hat{\beta}$, has certain minimum properties which are partly summarized in the following theorem.

THEOREM 3.4 (*Atiqullah [1962]*) *Let Y_1, Y_2, \dots, Y_n be n independent random variables with common variance σ^2 and common third and fourth moments, μ_3 and μ_4 , respectively, about their means. If $E[Y] = \mathbf{X}\beta$, where \mathbf{X} is $n \times p$ of rank p , then $(n-p)S^2$ is the unique nonnegative quadratic unbiased estimate of $(n-p)\sigma^2$ with minimum variance when $\mu_4 = 3\sigma^4$ or when the diagonal elements of \mathbf{P} are all equal.*

Proof. Since $\sigma^2 \geq 0$ it is not unreasonable to follow Rao [1952] and consider estimates that are nonnegative. Let $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ be a member of the class \mathcal{C} of nonnegative quadratic unbiased estimates of $(n-p)\sigma^2$. Then, by Theorem 1.5,

$$(n-p)\sigma^2 = E[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \sigma^2 \text{tr}(\mathbf{A}) + \beta'\mathbf{X}'\mathbf{A}\mathbf{X}\beta$$

for all β , so that $\text{tr}(\mathbf{A}) = n-p$ (setting $\beta = \mathbf{0}$) and $\beta'\mathbf{X}'\mathbf{A}\mathbf{X}\beta = 0$ for all β . Thus $\mathbf{X}'\mathbf{A}\mathbf{X} = \mathbf{0}$ (A.11.2) and, since \mathbf{A} is positive semidefinite, $\mathbf{A}\mathbf{X} = \mathbf{0}$ (A.3.5) and $\mathbf{X}'\mathbf{A} = \mathbf{0}$. Hence if \mathbf{a} is a vector of diagonal elements of \mathbf{A} , and $\gamma_2 = (\mu_4 - 3\sigma^4)/\sigma^4$, it follows from Theorem 1.6 that

$$\begin{aligned}\text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] &= \sigma^4\gamma_2\mathbf{a}'\mathbf{a} + 2\sigma^4\text{tr}(\mathbf{A}^2) + 4\sigma^2\beta'\mathbf{X}'\mathbf{A}^2\mathbf{X}\beta + 4\mu_3\beta'\mathbf{X}'\mathbf{A}\mathbf{a} \\ &= \sigma^4\gamma_2\mathbf{a}'\mathbf{a} + 2\sigma^4\text{tr}(\mathbf{A}^2).\end{aligned}\tag{3.13}$$

Now by Theorem 3.3, $(n-p)S^2 [= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{R}\mathbf{Y}$, say] is a member of the class \mathcal{C} . Also, by Theorem 3.1,

$$\text{tr}(\mathbf{R}^2) = \text{tr}(\mathbf{R}) = n-p,$$

so that if we substitute in (3.13),

$$\text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] = \sigma^4 \gamma_2 \mathbf{r}' \mathbf{r} + 2\sigma^4(n-p). \quad (3.14)$$

To find sufficient conditions for $\mathbf{Y}'\mathbf{R}\mathbf{Y}$ to have minimum variance for class \mathcal{C} , let $\mathbf{A} = \mathbf{R} + \mathbf{D}$. Then \mathbf{D} is symmetric, and $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{R}) + \text{tr}(\mathbf{D})$; thus $\text{tr}(\mathbf{D}) = 0$. Since $\mathbf{A}\mathbf{X} = \mathbf{0}$, we have $\mathbf{A}\mathbf{P} = \mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0}$, and combining this equation with $\mathbf{P}^2 = \mathbf{P}$, that is, $\mathbf{R}\mathbf{P} = \mathbf{0}$, leads to

$$\mathbf{0} = \mathbf{A}\mathbf{P} = \mathbf{R}\mathbf{P} + \mathbf{D}\mathbf{P} = \mathbf{D}\mathbf{P}$$

and

$$\mathbf{D}\mathbf{R} = \mathbf{D} \quad (= \mathbf{D}' = \mathbf{R}\mathbf{D}).$$

Hence

$$\begin{aligned} \mathbf{A}^2 &= \mathbf{R}^2 + \mathbf{D}\mathbf{R} + \mathbf{R}\mathbf{D} + \mathbf{D}^2 \\ &= \mathbf{R} + 2\mathbf{D} + \mathbf{D}^2 \end{aligned}$$

and

$$\begin{aligned} \text{tr}(\mathbf{A}^2) &= \text{tr}(\mathbf{R}) + 2\text{tr}(\mathbf{D}) + \text{tr}(\mathbf{D}^2) \\ &= (n-p) + \text{tr}(\mathbf{D}^2). \end{aligned}$$

Substituting in (3.13), setting $\mathbf{a} = \mathbf{r} + \mathbf{d}$, and using (3.14), we have

$$\begin{aligned} \text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] &= \sigma^4 \gamma_2 \mathbf{a}' \mathbf{a} + 2\sigma^4[(n-p) + \text{tr}(\mathbf{D}^2)] \\ &= \sigma^4 \gamma_2 (\mathbf{r}' \mathbf{r} + 2\mathbf{r}' \mathbf{d} + \mathbf{d}' \mathbf{d}) + 2\sigma^4[(n-p) + \text{tr}(\mathbf{D}^2)] \\ &= \sigma^4 \gamma_2 \mathbf{r}' \mathbf{r} + 2\sigma^4(n-p) + 2\sigma^4[\gamma_2(\mathbf{r}' \mathbf{d} + \frac{1}{2} \mathbf{d}' \mathbf{d}) + \text{tr}(\mathbf{D}^2)] \\ &= \text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] \\ &\quad + 2\sigma^4 \left[\gamma_2 \left(\sum_i r_{ii} d_{ii} + \frac{1}{2} \sum_i d_{ii}^2 \right) + \sum_i \sum_j d_{ij}^2 \right]. \end{aligned}$$

To find the estimate with minimum variance, we must minimize $\text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}]$ subject to $\text{tr}(\mathbf{D}) = 0$ and $\mathbf{D}\mathbf{R} = \mathbf{D}$. The minimization in general is difficult (cf. Hsu [1938]) but can be done readily in two important special cases. First, if $\gamma_2 = 0$, then

$$\text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] + 2\sigma^4 \sum_i \sum_j d_{ij}^2,$$

which is minimized when $d_{ij} = 0$ for all i, j , that is, when $\mathbf{D} = \mathbf{0}$ and $\mathbf{A} = \mathbf{R}$. Second, if the diagonal elements of \mathbf{P} are all equal, then they are equal to p/n [since, by Theorem 3.1(ii), $\text{tr}(\mathbf{P}) = p$]. Hence $r_{ii} = (n-p)/n$ for each i and

$$\begin{aligned} \text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] &= \text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] + 2\sigma^4 \left[\gamma_2 \left(0 + \frac{1}{2} \sum_i d_{ii}^2 \right) + \sum_i \sum_j d_{ij}^2 \right] \\ &= \text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] + 2\sigma^4 \left[(\frac{1}{2} \gamma_2 + 1) \sum_i d_{ii}^2 + \sum_{i \neq j} \sum_j d_{ij}^2 \right], \end{aligned}$$

as $\sum_i r_{ii} d_{ii} = [(n - p)/n] \text{tr}(\mathbf{D}) = 0$. Now $\gamma_2 > -2$ (A.13.1), so that $\text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}]$ is minimized when $d_{ij} = 0$ for all i, j . Thus in both cases we have minimum variance if and only if $\mathbf{A} = \mathbf{R}$. \square

This theorem highlights the fact that a uniformly minimum variance quadratic unbiased estimate of σ^2 exists only under certain restrictive conditions like those stated in the enunciation of the theorem. If normality can be assumed ($\gamma_2 = 0$), then it transpires that (Rao [1973: p. 319]) S^2 is the minimum variance unbiased estimate of σ^2 in the entire class of unbiased estimates (not just the class of quadratic estimates).

Rao [1970, 1972] has also introduced another criterion for choosing the estimate of σ^2 : *minimum norm quadratic unbiased estimation* (MINQUE). Irrespective of whether or not we assume normality, this criterion also leads to S^2 (cf. Rao [1970, 1974: p. 448]).

EXERCISES 3c

- Suppose that $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, where \mathbf{X} is $n \times p$ of rank p .

- (a) Find $\text{var}[S^2]$.
- (b) Evaluate $E[(\mathbf{Y}'\mathbf{A}_1\mathbf{Y} - \sigma^2)^2]$ for

$$\mathbf{A}_1 = \frac{1}{n-p+2} [\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'].$$

- (c) Prove that $\mathbf{Y}'\mathbf{A}_1\mathbf{Y}$ is an estimate of σ^2 with a smaller mean-squared error than S^2 .

(Theil and Schweitzer [1961])

- Let Y_1, Y_2, \dots, Y_n be independently and identically distributed with mean θ and variance σ^2 . Find the nonnegative quadratic unbiased estimate of σ^2 with the minimum variance.

3.4 DISTRIBUTION THEORY

Until now the only assumptions we have made about the ε_i are that $E[\varepsilon] = 0$ and $\text{Var}[\varepsilon] = \sigma^2 \mathbf{I}_n$. If we assume that the ε_i are also normally distributed, then $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and hence $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$. A number of distributional results then follow.

THEOREM 3.5 *If $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, where \mathbf{X} is $n \times p$ of rank p , then:*

- (i) $\hat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$.
- (ii) $(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) / \sigma^2 \sim \chi_p^2$.

(iii) $\hat{\beta}$ is independent of S^2 .

$$(iv) \text{RSS}/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi_{n-p}^2.$$

Proof. (i) Since $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}$, say, where \mathbf{C} is a $p \times n$ matrix such that $\text{rank } \mathbf{C} = \text{rank } \mathbf{X}' = \text{rank } \mathbf{X} = p$ (by A.2.4), $\hat{\beta}$ has a multivariate normal distribution (Theorem 2.2 in Section 2.2). In particular, from equations (3.10) and (3.11), we have $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

(ii) $(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)/\sigma^2 = (\hat{\beta} - \beta)' (\text{Var}[\hat{\beta}])^{-1} (\hat{\beta} - \beta)$, which, by (i) and Theorem 2.9, is distributed as χ_p^2 .

(iii)

$$\begin{aligned} \text{Cov}[\hat{\beta}, \mathbf{Y} - \mathbf{X}\hat{\beta}] &= \text{Cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, (\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{Cov}[\mathbf{Y}] (\mathbf{I}_n - \mathbf{P})' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{P}) \\ &= \mathbf{0} \quad [\text{by Theorem 3.1(iii)}]. \end{aligned}$$

If $\mathbf{U} = \hat{\beta}$ and $\mathbf{V} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ in Theorem 2.5 (Section 2.3), $\hat{\beta}$ is independent of $\|(\mathbf{Y} - \mathbf{X}\hat{\beta})\|^2$ and therefore of S^2 .

(iv) This result can be proved in various ways, depending on which theorems relating to quadratic forms we are prepared to invoke. It is instructive to examine two methods of proof, although the first method is the more standard one.

Method 1: Using Theorem 3.3, we have

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\beta) \quad [\text{by Theorem 3.1(iii)}] \\ &= \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}, \end{aligned} \tag{3.15}$$

where $\mathbf{I}_n - \mathbf{P}$ is symmetric and idempotent of rank $n-p$. Since $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $\text{RSS}/\sigma^2 \sim \chi_{n-p}^2$ (Theorem 2.7 in Section 2.4).

Method 2:

$$\begin{aligned} Q_1 &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}(\hat{\beta} - \beta))' (\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}(\hat{\beta} - \beta)) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &\quad + 2(\hat{\beta} - \beta)' \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \\ &= Q + Q_2, \text{ say,} \end{aligned} \tag{3.16}$$

since, from the normal equations,

$$(\hat{\beta} - \beta)' \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = (\hat{\beta} - \beta)' (\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta}) = 0. \tag{3.17}$$

Now Q_1/σ^2 ($= \sum_i \epsilon_i^2/\sigma^2$) is χ_n^2 , and $Q_2/\sigma^2 \sim \chi_p^2$ [by (ii)]. Also, Q_2 is a continuous function of $\hat{\beta}$, so that by Example 1.11 and (iii), Q is independent of Q_2 . Hence $Q/\sigma^2 \sim \chi_{n-p}^2$ (Example 1.10, Section 1.6). \square

EXERCISES 3d

1. Given Y_1, Y_2, \dots, Y_n independently distributed as $N(\theta, \sigma^2)$, use Theorem 3.5 to prove that:
 - (a) \bar{Y} is statistically independent of $Q = \sum_i (Y_i - \bar{Y})^2$.
 - (b) $Q/\sigma^2 \sim \chi_{n-1}^2$.
2. Use Theorem 2.5 to prove that for the full-rank regression model, RSS is independent of $(\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta)$.

3.5 MAXIMUM LIKELIHOOD ESTIMATION

Assuming normality, as in Section 3.4, the likelihood function, $L(\beta, \sigma^2)$ say, for the full-rank regression model is the probability density function of \mathbf{Y} , namely,

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\beta\|^2 \right\}.$$

Let $l(\beta, v) = \log L(\beta, \sigma^2)$, where $v = \sigma^2$. Then, ignoring constants, we have

$$l(\beta, v) = -\frac{n}{2} \log v - \frac{1}{2v} \|\mathbf{y} - \mathbf{X}\beta\|^2,$$

and from (3.6) it follows that

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2v} (-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta)$$

and

$$\frac{\partial l}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2} \|\mathbf{y} - \mathbf{X}\beta\|^2.$$

Setting $\partial l / \partial \beta = 0$, we get the least squares estimate of β , which clearly maximizes $l(\beta, v)$ for any $v > 0$. Hence

$$L(\beta, v) \leq L(\hat{\beta}, v) \quad \text{for all } v > 0$$

with equality if and only if $\beta = \hat{\beta}$.

We now wish to maximize $L(\hat{\beta}, v)$, or equivalently $l(\hat{\beta}, v)$, with respect to v . Setting $\partial l / \partial v = 0$, we get a stationary value of $\hat{v} = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2/n$. Then

$$\begin{aligned} l(\hat{\beta}, \hat{v}) - l(\hat{\beta}, v) &= -\frac{n}{2} \left[\log \left(\frac{\hat{v}}{v} \right) + 1 - \frac{\hat{v}}{v} \right] \\ &\geq 0, \end{aligned}$$

since $x \leq e^{x-1}$ and therefore $\log x \leq x - 1$ for $x \geq 0$ (with equality when $x = 1$). Hence

$$L(\beta, v) \leq L(\hat{\beta}, \hat{v}) \quad \text{for all } v > 0$$

with equality if and only if $\beta = \hat{\beta}$ and $v = \hat{v}$. Thus $\hat{\beta}$ and \hat{v} are the maximum likelihood estimates of β and v . Also, for future use,

$$L(\hat{\beta}, \hat{v}^2) = (2\pi\hat{v}^2)^{-n/2} e^{-n/2}. \quad (3.18)$$

In determining the efficiency of the estimates above, we derive the (expected) information matrix

$$\begin{aligned} \mathbf{I} &= -E[\partial^2 l / \partial \theta \partial \theta'] \\ &= \text{Var}[\partial l / \partial \theta], \end{aligned} \quad (3.19)$$

where $\theta = (\beta', v)'$. As a first step we find that

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta \partial \beta'} &= -\frac{1}{v^2} (\mathbf{X}' \mathbf{X}), \\ \frac{\partial^2 l}{\partial \beta \partial v} &= \frac{1}{v^2} (-2\mathbf{X}' \mathbf{y} + \mathbf{X}' \mathbf{X} \beta) \end{aligned}$$

and

$$\frac{\partial^2 l}{\partial v^2} = \frac{n}{2v^2} - \frac{1}{v^3} \|\mathbf{y} - \mathbf{X}\beta\|^2.$$

We note that $\|\mathbf{Y} - \mathbf{X}\beta\|^2/v = \epsilon' \epsilon / v \sim \chi_n^2$, so that $E[\epsilon' \epsilon] = nv$ (as $E[\chi_n^2] = n$). Replacing \mathbf{y} by \mathbf{Y} and taking expected values in the equations above gives us

$$\mathbf{I} = \begin{pmatrix} \frac{1}{v} (\mathbf{X}' \mathbf{X}) & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2v^2} \end{pmatrix}.$$

This gives us the multivariate Cramer–Rao lower bound for unbiased estimates of θ , namely,

$$\mathbf{I}^{-1} = \begin{pmatrix} v(\mathbf{X}' \mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2v^2}{n} \end{pmatrix}.$$

Since $\text{Var}[\hat{\beta}] = v(\mathbf{X}' \mathbf{X})^{-1}$, $\hat{\beta}$ is the best unbiased estimate of β in the sense that for any \mathbf{a} , $\mathbf{a}' \hat{\beta}$ is the minimum variance unbiased estimate (MINVUE) of $\mathbf{a}' \beta$.

Since $(n-p)S^2/v \sim \chi_{n-p}^2$ [by Theorem 3.5(iv)] and $\text{var}[\chi_{n-p}^2] = 2(n-p)$, it follows that $\text{var}[S^2] = 2v^2/(n-p)$, which tends to $2v^2/n$ as $n \rightarrow \infty$. This tells us that S^2 is, asymptotically, the MINVUE of v . However, the Cramer–Rao lower bound gives us just a lower bound on the minimum variance rather than the actual minimum. It transpires that S^2 is exactly MINVUE, and a different approach is needed to prove this (e.g., Rao [1973: p. 319]).

3.6 ORTHOGONAL COLUMNS IN THE REGRESSION MATRIX

Suppose that in the full-rank model $E[\mathbf{Y}] = \mathbf{X}\beta$ the matrix \mathbf{X} has a column representation

$$\mathbf{X} = (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)}),$$

where the columns are all mutually orthogonal. Then

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \left(\begin{array}{cccc} \mathbf{x}^{(0)'}\mathbf{x}^{(0)} & 0 & \cdots & 0 \\ 0 & \mathbf{x}^{(1)'}\mathbf{x}^{(1)} & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \mathbf{x}^{(p-1)'}\mathbf{x}^{(p-1)} \end{array} \right)^{-1} \left(\begin{array}{c} \mathbf{x}^{(0)'}\mathbf{Y} \\ \mathbf{x}^{(1)'}\mathbf{Y} \\ \vdots \\ \mathbf{x}^{(p-1)'}\mathbf{Y} \end{array} \right) \\ &= \left(\begin{array}{c} (\mathbf{x}^{(0)'}\mathbf{x}^{(0)})^{-1}\mathbf{x}^{(0)'}\mathbf{Y} \\ (\mathbf{x}^{(1)'}\mathbf{x}^{(1)})^{-1}\mathbf{x}^{(1)'}\mathbf{Y} \\ \vdots \\ (\mathbf{x}^{(p-1)'}\mathbf{x}^{(p-1)})^{-1}\mathbf{x}^{(p-1)'}\mathbf{Y} \end{array} \right).\end{aligned}$$

Thus $\hat{\beta}_j = \mathbf{x}^{(j)'}\mathbf{Y}/\mathbf{x}^{(j)'}\mathbf{x}^{(j)}$ turns out to be the least squares estimate of β_j for the model $E[\mathbf{Y}] = \mathbf{x}^{(j)}\beta_j$, which means that the least squares estimate of β_j is unchanged if any of the other β_l ($l \neq j$) are put equal to zero. Also, from equations (3.8) and (3.9), the residual sum of squares takes the form

$$\begin{aligned}\text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \sum_{j=0}^{p-1} \hat{\beta}_j \mathbf{x}^{(j)'}\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \sum_{j=0}^{p-1} \hat{\beta}_j^2 (\mathbf{x}^{(j)'}\mathbf{x}^{(j)}). \tag{3.20}\end{aligned}$$

If we put $\beta_j = 0$ in the model, the only change in the residual sum of squares is the addition of the term $\hat{\beta}_j \mathbf{x}^{(j)'}\mathbf{Y}$, so that we now have

$$\mathbf{Y}'\mathbf{Y} - \sum_{r=0, r \neq j}^{p-1} \hat{\beta}_r \mathbf{x}^{(r)'}\mathbf{Y}. \tag{3.21}$$

Two applications of this model are discussed in Sections 7.1.2 and 7.3.1.

EXAMPLE 3.3 Consider the full-rank model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where the ε_i are i.i.d. $N(0, \sigma^2)$ and the x_{ij} are standardized so that for $j = 1, 2, \dots, p-1$, $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = c$. We now show that

$$\frac{1}{p} \sum_{j=0}^{p-1} \text{var}[\hat{\beta}_j] \quad (3.22)$$

is minimized when the columns of \mathbf{X} are mutually orthogonal.

From

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{C} \end{pmatrix},$$

say, we have

$$\begin{aligned} \sum_{j=0}^{p-1} \text{var}[\hat{\beta}_j] &= \text{tr}(\text{Var}[\hat{\beta}]) \\ &= \sigma^2 \left[\text{tr}(\mathbf{C}^{-1}) + \frac{1}{n} \right] \\ &= \sigma^2 \sum_{j=0}^{p-1} \lambda_j^{-1}, \end{aligned} \quad (3.23)$$

where $\lambda_0 = n$ and λ_j ($j = 1, 2, \dots, p-1$) are the eigenvalues of \mathbf{C} (A.1.6). Now the minimum of (3.23) subject to the condition $\text{tr}(\mathbf{X}'\mathbf{X}) = n + c(p-1)$, or $\text{tr}(\mathbf{C}) = c(p-1)$, is given by $\lambda_j = \text{constant}$, that is, $\lambda_j = c$ ($j = 1, 2, \dots, p-1$). Hence there exists an orthogonal matrix \mathbf{T} such that $\mathbf{T}'\mathbf{C}\mathbf{T} = c\mathbf{I}_{p-1}$, or $\mathbf{C} = c\mathbf{I}_{p-1}$, so that the columns of \mathbf{X} must be mutually orthogonal. \square

This example shows that using a particular optimality criterion, the “optimum” choice of \mathbf{X} is the design matrix with mutually orthogonal columns. A related property, proved by Hotelling (see Exercises 3e, No. 3), is the following: Given any design matrix \mathbf{X} such that $\mathbf{x}^{(j)'}\mathbf{x}^{(j)} = c_j^2$, then

$$\text{var}[\hat{\beta}_j] \geq \frac{\sigma^2}{c_j^2},$$

and the minimum is attained when $\mathbf{x}^{(j)'}\mathbf{x}^{(r)} = 0$ (all $r, r \neq j$) [i.e., when $\mathbf{x}^{(j)}$ is perpendicular to the other columns].

EXERCISES 3e

1. Prove the statement above that the minimum is given by $\lambda_j = c$ ($j = 1, 2, \dots, p-1$).
2. It is required to fit a regression model of the form

$$E[Y_i] = \beta_0 + \beta_1 x_i + \beta_2 \phi(x_i) \quad (i = 1, 2, 3),$$

where $\phi(x)$ is a second-degree polynomial. If $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$, find ϕ such that the design matrix \mathbf{X} has mutually orthogonal columns.

3. Suppose that $\mathbf{X} = (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)}, \mathbf{x}^{(p)}) = (\mathbf{W}, \mathbf{x}^{(p)})$ has linearly independent columns.

- (a) Using A.9.5, prove that

$$\det(\mathbf{X}'\mathbf{X}) = \det(\mathbf{W}'\mathbf{W}) \left(\mathbf{x}^{(p)'}\mathbf{x}^{(p)} - \mathbf{x}^{(p)'}\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{x}^{(p)} \right).$$

- (b) Deduce that

$$\frac{\det(\mathbf{W}'\mathbf{W})}{\det(\mathbf{X}'\mathbf{X})} \geq \frac{1}{\mathbf{x}^{(p)'}\mathbf{x}^{(p)}},$$

and hence show that $\text{var}[\hat{\beta}_p] \geq \sigma^2(\mathbf{x}^{(p)'}\mathbf{x}^{(p)})^{-1}$ with equality if and only if $\mathbf{x}^{(p)'}\mathbf{x}^{(j)} = 0$ ($j = 0, 1, \dots, p-1$).

(Rao [1973: p. 236])

4. What modifications in the statement of Example 3.3 proved above can be made if the term β_0 is omitted?

5. Suppose that we wish to find the weights β_i ($i = 1, 2, \dots, k$) of k objects. One method is to weigh each object r times and take the average; this requires a total of kr weighings, and the variance of each average is σ^2/r (σ^2 being the variance of the weighing error). Another method is to weigh the objects in combinations; some of the objects are distributed between the two pans and weights are placed in one pan to achieve equilibrium. The regression model for such a scheme is

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where $x_i = 0, 1$, or -1 according as the i th object is not used, placed in the left pan or in the right pan, ε is the weighing error (assumed to be the same for all weighings), and Y is the weight required for equilibrium (Y is regarded as negative if placed in the left pan). After n such weighing operations we can find the least squares estimates $\hat{\beta}_i$ of the weights.

- (a) Show that the estimates of the weights have maximum precision (i.e., minimum variance) when each entry in the design matrix \mathbf{X} is ± 1 and the columns of \mathbf{X} are mutually orthogonal.
- (b) If the objects are weighed individually, show that kn weighings are required to achieve the same precision as that given by the optimal design with n weighings.

(Rao [1973: p. 309])

3.7 INTRODUCING FURTHER EXPLANATORY VARIABLES

3.7.1 General Theory

Suppose that after having fitted the regression model

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, \quad \text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n,$$

we decide to introduce additional x_j 's into the model so that the model is now enlarged to

$$\begin{aligned} G: E[\mathbf{Y}] &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma \\ &= (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix} \\ &= \mathbf{W}\boldsymbol{\delta}, \end{aligned} \tag{3.24}$$

say, where \mathbf{X} is $n \times p$ of rank p , \mathbf{Z} is $n \times t$ of rank t , and the columns of \mathbf{Z} are linearly independent of the columns of \mathbf{X} ; that is, \mathbf{W} is $n \times (p+t)$ of rank $p+t$. Then to find the least squares estimate $\hat{\boldsymbol{\delta}}_G$ of $\boldsymbol{\delta}$ there are two possible approaches. We can either compute $\hat{\boldsymbol{\delta}}_G$ and its dispersion matrix directly from

$$\hat{\boldsymbol{\delta}}_G = (\mathbf{W}\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} \quad \text{and} \quad \text{Var}[\hat{\boldsymbol{\delta}}_G] = \sigma^2(\mathbf{W}'\mathbf{W})^{-1},$$

or to reduce the amount of computation, we can utilize the calculations already carried out in fitting the original model, as in Theorem 3.6 below. A geometrical proof of this theorem, which allows \mathbf{X} to have less than full rank, is given in Section 3.9.3. But first a lemma.

LEMMA If $\mathbf{R} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\mathbf{Z}'\mathbf{R}\mathbf{Z}$ is positive-definite.

Proof. Let $\mathbf{Z}'\mathbf{R}\mathbf{Z}\mathbf{a} = \mathbf{0}$; then, by Theorem 3.1(i),

$$\mathbf{a}'\mathbf{Z}'\mathbf{R}'\mathbf{R}\mathbf{Z}\mathbf{a} = \mathbf{a}'\mathbf{Z}'\mathbf{R}\mathbf{Z}\mathbf{a} = \mathbf{0},$$

or $\mathbf{R}\mathbf{Z}\mathbf{a} = \mathbf{0}$. Hence $\mathbf{Z}\mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{a} = \mathbf{X}\mathbf{b}$, say, which implies that $\mathbf{a} = \mathbf{0}$, as the columns of \mathbf{Z} are linearly independent of the columns of \mathbf{X} . Because $\mathbf{Z}'\mathbf{R}\mathbf{Z}\mathbf{a} = \mathbf{0}$ implies that $\mathbf{a} = \mathbf{0}$, $\mathbf{Z}'\mathbf{R}\mathbf{Z}$ has linearly independent columns and is therefore nonsingular. Also, $\mathbf{a}'\mathbf{Z}'\mathbf{R}\mathbf{Z}\mathbf{a} = (\mathbf{R}\mathbf{Z}\mathbf{a})'(\mathbf{R}\mathbf{Z}\mathbf{a}) \geq 0$. \square

THEOREM 3.6 Let $\mathbf{R}_G = \mathbf{I}_n - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$, $\mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$, $\mathbf{M} = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}$, and

$$\hat{\boldsymbol{\delta}}_G = \begin{pmatrix} \hat{\boldsymbol{\beta}}_G \\ \hat{\gamma}_G \end{pmatrix}.$$

Then:

$$(i) \hat{\gamma}_G = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{Y}.$$

$$(ii) \hat{\beta}_G = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) = \hat{\beta} - \mathbf{L}\hat{\gamma}_G.$$

$$(iii) \mathbf{Y}'\mathbf{R}_G\mathbf{Y} = (\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G)' \mathbf{R}(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) = \mathbf{Y}'\mathbf{R}\mathbf{Y} - \hat{\gamma}_G'\mathbf{Z}'\mathbf{R}\mathbf{Y}.$$

(iv)

$$\text{Var}[\hat{\delta}_G] = \sigma^2 \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}' & -\mathbf{L}\mathbf{M} \\ -\mathbf{M}\mathbf{L}' & \mathbf{M} \end{pmatrix}. \quad (3.25)$$

Proof. (i) We first “orthogonalize” the model. Since $\mathcal{C}(\mathbf{PZ}) \subset \mathcal{C}(\mathbf{X})$,

$$\begin{aligned} \mathbf{X}\beta + \mathbf{Z}\gamma &= \mathbf{X}\beta + \mathbf{PZ}\gamma + (\mathbf{I}_n - \mathbf{P})\mathbf{Z}\gamma \\ &= \mathbf{X}\alpha + \mathbf{RZ}\gamma \\ &= (\mathbf{X}, \mathbf{RZ}) \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \\ &= \mathbf{V}\lambda, \end{aligned}$$

say, where $\alpha = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma = \beta + \mathbf{L}\gamma$ is unique. We note that $\mathcal{C}(\mathbf{X}) \perp \mathcal{C}(\mathbf{RZ})$. Also, by A.2.4 and the previous lemma,

$$\text{rank}(\mathbf{RZ}) = \text{rank}(\mathbf{Z}'\mathbf{R}'\mathbf{RZ}) = \text{rank}(\mathbf{Z}'\mathbf{RZ}) = t,$$

so that \mathbf{V} has full rank $p + t$. Since $\mathbf{XR} = \mathbf{0}$, the least squares estimate of λ is

$$\begin{aligned} \tilde{\lambda} &= (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{Y} \\ &= \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{RZ} \\ \mathbf{Z}'\mathbf{RX} & \mathbf{Z}'\mathbf{R}'\mathbf{RZ} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}'\mathbf{R} \end{pmatrix} \mathbf{Y} \\ &= \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'\mathbf{RZ} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}'\mathbf{R} \end{pmatrix} \mathbf{Y} \\ &= \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ (\mathbf{Z}'\mathbf{RZ})^{-1}\mathbf{Z}'\mathbf{RY} \end{pmatrix} = \begin{pmatrix} \tilde{\alpha} \\ \tilde{\gamma} \end{pmatrix}. \end{aligned}$$

Now the relationship between (β, γ) and (α, γ) is one-to-one, so that the same relationships exist between their least square estimates. Hence

$$\hat{\gamma}_G = \tilde{\gamma} = (\mathbf{Z}'\mathbf{RZ})^{-1}\mathbf{Z}'\mathbf{RY}. \quad (3.26)$$

(ii) We also have

$$\begin{aligned} \hat{\beta}_G &= \tilde{\alpha} - \mathbf{L}\tilde{\gamma} \\ &= \hat{\beta} - \mathbf{L}\hat{\gamma}_G \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G). \end{aligned} \quad (3.27)$$

(iii) Using (3.27) gives

$$\mathbf{R}_G \mathbf{Y} = \mathbf{Y} - \mathbf{X}\hat{\beta}_G - \mathbf{Z}\hat{\gamma}_G$$

$$= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) - \mathbf{Z}\hat{\gamma}_G$$

$$= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G)$$

$$= \mathbf{R}(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) \quad (3.28)$$

$$= \mathbf{R}\mathbf{Y} - \mathbf{R}\mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{Y}, \quad (3.29)$$

so that by (3.28),

$$\begin{aligned} \mathbf{Y}'\mathbf{R}_G \mathbf{Y} &= (\mathbf{Y} - \mathbf{W}\hat{\delta}_G)'(\mathbf{Y} - \mathbf{W}\hat{\delta}_G) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta}_G - \mathbf{Z}\hat{\gamma}_G)'(\mathbf{Y} - \mathbf{X}\hat{\beta}_G - \mathbf{Z}\hat{\gamma}_G) \\ &= (\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G)'\mathbf{R}'\mathbf{R}(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) \\ &= (\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G)'\mathbf{R}(\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G), \end{aligned} \quad (3.30)$$

since \mathbf{R} is symmetric and idempotent [Theorem 3.1(i)].

Expanding equation (3.30) gives us

$$\begin{aligned} \mathbf{Y}'\mathbf{R}_G \mathbf{Y} &= \mathbf{Y}'\mathbf{R}\mathbf{Y} - 2\hat{\gamma}_G' \mathbf{Z}'\mathbf{R}\mathbf{Y} + \hat{\gamma}_G' \mathbf{Z}'\mathbf{R}\mathbf{Z}\hat{\gamma}_G \\ &= \mathbf{Y}'\mathbf{R}\mathbf{Y} - \hat{\gamma}_G' \mathbf{Z}'\mathbf{R}\mathbf{Y} - \hat{\gamma}_G' (\mathbf{Z}'\mathbf{R}\mathbf{Y} - \mathbf{Z}'\mathbf{R}\mathbf{Z}\hat{\gamma}_G) \\ &= \mathbf{Y}'\mathbf{R}\mathbf{Y} - \hat{\gamma}_G' \mathbf{Z}'\mathbf{R}\mathbf{Y} \quad [\text{by (3.26)}]. \end{aligned}$$

(iv)

$$\begin{aligned} \text{Var}[\hat{\gamma}_G] &= (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{R} \text{Var}[\mathbf{Y}] \mathbf{R}\mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} \\ &= \sigma^2 (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} (\mathbf{Z}'\mathbf{R}\mathbf{Z})(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} \\ &= \sigma^2 (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} = \sigma^2 \mathbf{M}. \end{aligned}$$

Now, by Theorem 1.3,

$$\begin{aligned} \text{Cov}[\hat{\beta}, \hat{\gamma}_G] &= \text{Cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{Y}] \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{R}\mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1} \\ &= 0, \end{aligned} \quad (3.31)$$

since $\mathbf{X}'\mathbf{R} = \mathbf{0}$. Hence using (i) above, we have, from Theorem 1.3,

$$\begin{aligned} \text{Cov}[\hat{\beta}_G, \hat{\gamma}_G] &= \text{Cov}[\hat{\beta} - \mathbf{L}\hat{\gamma}_G, \hat{\gamma}_G] \\ &= \text{Cov}[\hat{\beta}, \hat{\gamma}_G] - \mathbf{L} \text{Var}[\hat{\gamma}_G] \\ &= -\sigma^2 \mathbf{L}\mathbf{M} \quad [\text{by (3.31)}] \end{aligned}$$

and

$$\begin{aligned} \text{Var}[\hat{\beta}_G] &= \text{Var}[\hat{\beta} - \mathbf{L}\hat{\gamma}_G] \\ &= \text{Var}[\hat{\beta}] - \text{Cov}[\hat{\beta}, \mathbf{L}\hat{\gamma}_G] - \text{Cov}[\mathbf{L}\hat{\gamma}_G, \hat{\beta}] + \text{Var}[\mathbf{L}\hat{\gamma}_G] \\ &= \text{Var}[\hat{\beta}] + \mathbf{L} \text{Var}[\hat{\gamma}_G] \mathbf{L}' \quad [\text{by (3.31)}] \\ &= \sigma^2 [(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}']. \end{aligned}$$

□

□

From Theorem 3.6 we see that once $\mathbf{X}'\mathbf{X}$ has been inverted, we can find $\hat{\delta}_G$ and its variance-covariance matrix simply by inverting the $t \times t$ matrix $\mathbf{Z}'\mathbf{R}\mathbf{Z}$; we need not invert the $(p+t) \times (p+t)$ matrix $\mathbf{W}'\mathbf{W}$. The case $t = 1$ is considered below.

3.7.2 One Extra Variable

Let the columns of \mathbf{X} be denoted by $\mathbf{x}^{(j)}$ ($j = 0, 1, 2, \dots, p-1$), so that

$$\begin{aligned} E[\mathbf{Y}] &= (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)})\boldsymbol{\beta} \\ &= \mathbf{x}^{(0)}\beta_0 + \mathbf{x}^{(1)}\beta_1 + \dots + \mathbf{x}^{(p-1)}\beta_{p-1}. \end{aligned}$$

Suppose now that we wish to introduce a further explanatory variable, x_p , say, into the model so that in terms of the notation above we have $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{x}^{(p)}\beta_p$. Then by Theorem 3.6, the least squares estimates for the enlarged model are readily calculated, since $\mathbf{Z}'\mathbf{R}\mathbf{Z}$ ($= \mathbf{x}^{(p)'}\mathbf{R}\mathbf{x}^{(p)}$) is only a 1×1 matrix, that is, a scalar. Hence

$$\hat{\beta}_{p,G} = \hat{\gamma}_G = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{Y} = \frac{\mathbf{x}^{(p)'}\mathbf{R}\mathbf{Y}}{\mathbf{x}^{(p)'}\mathbf{R}\mathbf{x}^{(p)}}, \quad (3.32)$$

$$\hat{\beta}_G = (\hat{\beta}_{0,G}, \dots, \hat{\beta}_{p-1,G})' = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}^{(p)}\hat{\beta}_{p,G},$$

$$\mathbf{Y}'\mathbf{R}_G\mathbf{Y} = \mathbf{Y}'\mathbf{R}\mathbf{Y} - \hat{\beta}_{p,G}\mathbf{x}^{(p)'}\mathbf{R}\mathbf{Y}, \quad (3.33)$$

and the matrix $\text{Var}[\hat{\delta}_G]$ is readily calculated from $(\mathbf{X}'\mathbf{X})^{-1}$. The ease with which “corrections” can be made to allow for a single additional x variable suggests that if more than one variable is to be added into the regression model, then the variables should be brought in one at a time. We return to this stepwise procedure in Chapter 11.

The technique above for introducing one extra variable was first discussed in detail by Cochran [1938] and generalized to the case of several variables by Quenouille [1950].

EXAMPLE 3.4 A recursive algorithm was given by Wilkinson [1970] (see also James and Wilkinson [1971], Rogers and Wilkinson [1974], and Pearce et al. [1974]) for fitting analysis-of-variance models by regression methods. This algorithm amounts to proving that the residuals for the augmented model are given by \mathbf{RSRY} , where $\mathbf{S} = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'$. We now prove this result. By (3.28) the residuals required are

$$\begin{aligned} \mathbf{R}_G\mathbf{Y} &= \mathbf{R}\mathbf{Y} - \mathbf{R}\mathbf{Z}\hat{\gamma}_G \\ &= \mathbf{R}(\mathbf{R}\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G) \\ &= \mathbf{R}[\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{R}\mathbf{Y} \\ &= \mathbf{RSRY}. \end{aligned} \quad (3.34)$$

□

The basic steps of the Wilkinson algorithm are as follows:

Algorithm 3.1

Step 1: Compute the residuals $\mathbf{R}\mathbf{Y}$.

Step 2: Use the operator \mathbf{S} , which Wilkinson calls a *sweep* (not to be confused with the sweep method of Section 11.2.2), to produce a vector of *apparent residuals* $\mathbf{R}\mathbf{Y} - \mathbf{Z}\hat{\gamma}_G$ ($= \mathbf{S}\mathbf{R}\mathbf{Y}$).

Step 3: Applying the operator \mathbf{R} once again, reanalyze the apparent residuals to produce the correct residuals $\mathbf{R}\mathbf{S}\mathbf{R}\mathbf{Y}$.

If the columns of \mathbf{Z} are perpendicular to the columns of \mathbf{X} , then $\mathbf{R}\mathbf{Z} = \mathbf{Z}$ and, by (3.34),

$$\begin{aligned}\mathbf{R}\mathbf{S}\mathbf{R} &= \mathbf{R}(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{R} \\ &= \mathbf{R} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R} \\ &= \mathbf{S}\mathbf{R},\end{aligned}$$

so that step 3 is unnecessary. We see later (Section 3.9.3) that the procedure above can still be used when the design matrix \mathbf{X} does not have full rank.

By setting \mathbf{X} equal to the first k columns of \mathbf{X} , and \mathbf{Z} equal to the $(k+1)$ th column ($k = 1, 2, \dots, p-1$), this algorithm can be used to fit the regression one column of \mathbf{X} at a time. Such a stepwise procedure is appropriate in experimental design situations because the columns of \mathbf{X} then correspond to different components of the model, such as the grand mean, main effects, block effects, and interactions, and some of the columns are usually orthogonal. Also, the elements of the design matrix \mathbf{X} are 0 or 1, so that in many standard designs the sweep operator \mathbf{S} amounts to a simple operation such as subtracting means, or a multiple of the means, from the residuals.

EXERCISES 3f

1. Prove that

$$\mathbf{Y}'\mathbf{R}\mathbf{Y} - \mathbf{Y}'\mathbf{R}_G\mathbf{Y} = \sigma^2 \hat{\gamma}'_G (\text{Var}[\hat{\gamma}_G])^{-1} \hat{\gamma}_G.$$

2. Prove that $\hat{\gamma}_G$ can be obtained by replacing \mathbf{Y} by $\mathbf{Y} - \mathbf{Z}\gamma$ in $\mathbf{Y}'\mathbf{R}\mathbf{Y}$ and minimizing with respect to γ . Show further that the minimum value thus obtained is $\mathbf{Y}'\mathbf{R}_G\mathbf{Y}$.
3. If $\hat{\beta}_G = (\hat{\beta}_{G,j})$ and $\hat{\beta} = (\hat{\beta}_j)$, use Theorem 3.6(iv) to prove that

$$\text{var}[\hat{\beta}_{G,j}] \geq \text{var}[\hat{\beta}_j].$$

4. Given that Y_1, Y_2, \dots, Y_n are independently distributed as $N(\theta, \sigma^2)$, find the least squares estimate of θ .

- (a) Use Theorem 3.6 to find the least squares estimates and the residual sum of squares for the augmented model

$$Y_i = \theta + \gamma x_i + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where the ε_i are independently distributed as $N(0, \sigma^2)$.

- (b) Verify the formulae for the least square estimates of θ and γ by differentiating the usual sum of squares.

3.8 ESTIMATION WITH LINEAR RESTRICTIONS

As a prelude to hypothesis testing in Chapter 4, we now examine what happens to least squares estimation when there are some hypothesized constraints on the model. We lead into this by way of an example.

EXAMPLE 3.5 A surveyor measures each of the angles α , β , and γ and obtains unbiased measurements Y_1 , Y_2 , and Y_3 in radians, respectively. If the angles form a triangle, then $\alpha + \beta + \gamma = \pi$. We can now find the least squares estimates of the unknown angles in two ways. The first method uses the constraint to write $\gamma = \pi - \alpha - \beta$ and reduces the number of unknown parameters from three to two, giving the model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 - \pi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}.$$

We then minimize $(Y_1 - \alpha)^2 + (Y_2 - \beta)^2 + (Y_3 - \pi + \alpha + \beta)^2$ with respect to α and β , respectively. Unfortunately, this method is somewhat ad hoc and not easy to use with more complicated models.

An alternative and more general approach is to use the model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

and minimize $(Y_1 - \alpha)^2 + (Y_2 - \beta)^2 + (Y_3 - \gamma)^2$ subject to the constraint $\alpha + \beta + \gamma = \pi$ using Lagrange multipliers. We consider this approach for a general model below. \square

3.8.1 Method of Lagrange Multipliers

Let $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where \mathbf{X} is $n \times p$ of full rank p . Suppose that we wish to find the minimum of $\epsilon'\epsilon$ subject to the linear restrictions $\mathbf{A}\beta = \mathbf{c}$, where \mathbf{A} is a known $q \times p$ matrix of rank q and \mathbf{c} is a known $q \times 1$ vector. One method of solving this problem is to use Lagrange multipliers, one for each linear constraint $\mathbf{a}'_i\beta = c_i$ ($i = 1, 2, \dots, q$), where \mathbf{a}'_i is the i th row of \mathbf{A} . As a first step we note that

$$\begin{aligned}\sum_{i=1}^q \lambda_i(\mathbf{a}'_i\beta - c_i) &= \lambda'(\mathbf{A}\beta - \mathbf{c}) \\ &= (\beta'\mathbf{A}' - \mathbf{c}')\lambda\end{aligned}$$

(since the transpose of a 1×1 matrix is itself). To apply the method of Lagrange multipliers, we consider the expression $r = \epsilon'\epsilon + (\beta'\mathbf{A}' - \mathbf{c}')\lambda$ and solve the equations

$$\mathbf{A}\beta = \mathbf{c} \quad (3.35)$$

and $\partial r / \partial \beta = 0$; that is (from A.8),

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta + \mathbf{A}'\lambda = 0. \quad (3.36)$$

For future reference we denote the solutions of these two equations by $\hat{\beta}_H$ and $\hat{\lambda}_H$. Then, from (3.36),

$$\begin{aligned}\hat{\beta}_H &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\lambda}_H \\ &= \hat{\beta} - \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\lambda}_H,\end{aligned} \quad (3.37)$$

and from (3.35),

$$\begin{aligned}\mathbf{c} &= \mathbf{A}\hat{\beta}_H \\ &= \mathbf{A}\hat{\beta} - \frac{1}{2}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\lambda}_H.\end{aligned}$$

Since $(\mathbf{X}'\mathbf{X})^{-1}$ is positive-definite, being the inverse of a positive-definite matrix, $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$ is also positive-definite (A.4.5) and therefore nonsingular. Hence

$$-\frac{1}{2}\hat{\lambda}_H = [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta})$$

and substituting in (3.37), we have

$$\hat{\beta}_H = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{c} - \mathbf{A}\hat{\beta}). \quad (3.38)$$

To prove that $\hat{\beta}_H$ actually minimizes $\epsilon'\epsilon$ subject to $\mathbf{A}\beta = \mathbf{c}$, we note that

$$\begin{aligned}\|\mathbf{X}(\hat{\beta} - \beta)\|^2 &= (\hat{\beta} - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta) \\ &= (\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \hat{\beta}_H + \hat{\beta}_H - \beta) \\ &= (\hat{\beta} - \hat{\beta}_H)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \hat{\beta}_H) + (\hat{\beta}_H - \beta)' \mathbf{X}' \mathbf{X} (\hat{\beta}_H - \beta) \quad (3.39)\end{aligned}$$

$$= \|\mathbf{X}(\hat{\beta} - \hat{\beta}_H)\|^2 + \|\mathbf{X}(\hat{\beta}_H - \beta)\|^2 \quad (3.40)$$

since from (3.37),

$$2(\hat{\beta} - \hat{\beta}_H)' \mathbf{X}' \mathbf{X}(\hat{\beta}_H - \beta) = \hat{\lambda}'_H \mathbf{A}(\hat{\beta}_H - \beta) = \hat{\lambda}'_H(\mathbf{c} - \mathbf{c}) = 0. \quad (3.41)$$

Hence from (3.16) in Section 3.4 and (3.40),

$$\begin{aligned}\epsilon' \epsilon &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\hat{\beta} - \beta)\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 + \|(\mathbf{X}(\hat{\beta} - \hat{\beta}_H))\|^2 + \|\mathbf{X}(\hat{\beta}_H - \beta)\|^2\end{aligned} \quad (3.42)$$

is a minimum when $\|\mathbf{X}(\hat{\beta}_H - \beta)\|^2 = 0$, that is, when $\mathbf{X}(\hat{\beta}_H - \beta) = \mathbf{0}$, or $\beta = \hat{\beta}_H$ (since the columns of \mathbf{X} are linearly independent).

Setting $\beta = \hat{\beta}_H$, we obtain the useful identity

$$\|\mathbf{Y} - \mathbf{X}\hat{\beta}_H\|^2 = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\hat{\beta} - \hat{\beta}_H)\|^2 \quad (3.43)$$

or, writing $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ and $\hat{\mathbf{Y}}_H = \mathbf{X}\hat{\beta}_H$,

$$\|\mathbf{Y} - \hat{\mathbf{Y}}_H\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2. \quad (3.44)$$

This identity can also be derived directly (see Exercises 3g, No. 2, at the end of Section 3.8.2).

3.8.2 Method of Orthogonal Projections

It is instructive to derive (3.38) using the theory of B.3. In order to do this, we first “shift” \mathbf{c} , in much the same way that we shifted π across into the left-hand side of Example 3.5.

Suppose that β_0 is any solution of $\mathbf{A}\beta = \mathbf{c}$. Then

$$\mathbf{Y} - \mathbf{X}\beta_0 = \mathbf{X}(\beta - \beta_0) + \epsilon \quad (3.45)$$

or $\tilde{\mathbf{Y}} = \mathbf{X}\gamma + \epsilon$, and $\mathbf{A}\gamma = \mathbf{A}\beta - \mathbf{A}\beta_0 = \mathbf{0}$. Thus we have the model $\tilde{\mathbf{Y}} = \theta + \epsilon$, where $\theta \in \Omega = \mathcal{C}(\mathbf{X})$, and since \mathbf{X} has full rank, $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\theta = \mathbf{A}\gamma = \mathbf{0}$. Setting $\mathbf{A}_1 = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\omega = \mathcal{N}(\mathbf{A}_1) \cap \Omega$, it follows from B.3.3 that $\omega^\perp \cap \Omega = \mathcal{C}(\mathbf{P}_\Omega \mathbf{A}_1')$, where

$$\mathbf{P}_\Omega \mathbf{A}_1' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$$

is $n \times q$ of rank q (by Exercises 3g, No. 5, below). Therefore, by B.3.2,

$$\begin{aligned}\mathbf{P}_\Omega - \mathbf{P}_\omega &= \mathbf{P}_{\omega^\perp \cap \Omega} \\ &= (\mathbf{P}_\Omega \mathbf{A}_1')[\mathbf{A}_1 \mathbf{P}_\Omega^2 \mathbf{A}_1']^{-1} (\mathbf{P}_\Omega \mathbf{A}_1')' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.\end{aligned}$$

Hence

$$\begin{aligned}\mathbf{X}\hat{\beta}_H - \mathbf{X}\beta_0 &= \mathbf{X}\hat{\gamma}_H = \mathbf{P}_\omega \tilde{\mathbf{Y}} = \mathbf{P}_\Omega \tilde{\mathbf{Y}} - \mathbf{P}_{\omega^\perp \cap \Omega} \tilde{\mathbf{Y}} \\ &= \mathbf{P}_\Omega \mathbf{Y} - \mathbf{X}\beta_0 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c}),\end{aligned} \quad (3.46)$$

since $\mathbf{P}_\Omega \mathbf{X} \boldsymbol{\beta}_0 = \mathbf{X} \boldsymbol{\beta}_0$ and $\mathbf{A} \boldsymbol{\beta}_0 = \mathbf{c}$. Therefore, canceling $\mathbf{X} \boldsymbol{\beta}_0$ and multiplying both sides by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ leads to $\hat{\boldsymbol{\beta}}_H$ of (3.38). Clearly, this gives a minimum as $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H\|^2 = \|\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\gamma}}_H\|^2$.

EXERCISES 3g

1. (a) Find the least squares estimates of α and β in Example 3.5 using the two approaches described there. What is the least squares estimate of γ ?
- (b) Suppose that a further constraint is introduced: namely, $\alpha = \beta$. Find the least squares estimates for this new situation using both methods.
2. By considering the identity $\mathbf{Y} - \hat{\mathbf{Y}}_H = \mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H$, prove that

$$\|\mathbf{Y} - \hat{\mathbf{Y}}_H\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2.$$

3. Prove that

$$\text{Var}[\hat{\boldsymbol{\beta}}_H] = \sigma^2 \left\{ (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1} \right\}.$$

Hence deduce that

$$\text{var}[\hat{\beta}_{Hj}] \leq \text{var}[\hat{\beta}_j],$$

where $\hat{\beta}_{Hj}$ and $\hat{\beta}_j$ are the j th elements of $\hat{\boldsymbol{\beta}}_H$ and $\hat{\boldsymbol{\beta}}$, respectively.

4. Show that

$$\|\mathbf{Y} - \hat{\mathbf{Y}}_H\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \sigma^2 \hat{\lambda}'_H \left(\text{Var}[\hat{\lambda}_H] \right)^{-1} \hat{\lambda}_H.$$

5. If \mathbf{X} is $n \times p$ of rank p and \mathbf{B} is $p \times q$ of rank q , show that $\text{rank}(\mathbf{X}\mathbf{B}) = q$.

3.9 DESIGN MATRIX OF LESS THAN FULL RANK

3.9.1 Least Squares Estimation

When the techniques of regression analysis are used for analyzing data from experimental designs, we find that the elements of \mathbf{X} are 0 or 1 (Chapter 8), and the columns of \mathbf{X} are usually linearly dependent. We now give such an example.

EXAMPLE 3.6 Consider the randomized block design with two treatments and two blocks: namely,

$$Y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij} \quad (i = 1, 2; j = 1, 2),$$

where Y_{ij} is the response from the i th treatment in the j th block. Then

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \hline Y_{21} \\ Y_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ \hline 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix} \quad (3.47)$$

or $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, where, for example, the first column of \mathbf{X} is linearly dependent on the other columns. \square

In Section 3.1 we developed a least squares theory which applies whether or not \mathbf{X} has full rank. If \mathbf{X} is $n \times p$ of rank r , where $r < p$, we saw in Section 3.1 that $\hat{\beta}$ is no longer unique. In fact, $\hat{\beta}$ should be regarded as simply a solution of the normal equations [e.g., $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$] which then enables us to find $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ and $\text{RSS} = \mathbf{e}'\mathbf{e}$, all of which are unique. We note that the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$ always have a solution for β as $\mathcal{C}(\mathbf{X}') = \mathcal{C}(\mathbf{X}'\mathbf{X})$ (by A.2.5). Our focus now is to consider methods for finding $\hat{\beta}$.

So far in this chapter our approach has been to replace \mathbf{X} by an $n \times r$ matrix \mathbf{X}_1 which has the same column space as \mathbf{X} . Very often the simplest way of doing this is to select r appropriate columns of \mathbf{X} , which amounts to setting some of the β_i in $\mathbf{X}\beta$ equal to zero. Algorithms for carrying this out are described in Section 11.9.

In the past, two other methods have been used. The first consists of imposing *identifiability constraints*, $\mathbf{H}\beta = \mathbf{0}$ say, which take up the “slack” in β so that there is now a unique β satisfying $\theta = \mathbf{X}\beta$ and $\mathbf{H}\beta = \mathbf{0}$. This approach is described by Scheffé [1959: p. 17]. The second method involves computing a generalized inverse. In Section 3.1 we saw that a $\hat{\beta}$ is given by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where $(\mathbf{X}'\mathbf{X})^{-1}$ is a suitable generalized inverse of $\mathbf{X}'\mathbf{X}$. One commonly used such inverse of a matrix \mathbf{A} is the Moore–Penrose inverse \mathbf{A}^+ , which is unique (see A.10).

EXAMPLE 3.7 In Example 3.6 we see that the first column of \mathbf{X} in (3.47) is the sum of columns 2 and 3, and the sum of columns 4 and 5. Although \mathbf{X} is 4×5 , it has only three linearly independent columns, so it is of rank 3. To reduce the model to one of full rank, we can set $\alpha_2 = 0$ and $\tau_2 = 0$, thus effectively removing the third and fifth columns. Our model is now

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \hline Y_{21} \\ 22 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \tau_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix}.$$

Alternatively, we can use two identifiability constraints, the most common being $\sum_i \alpha_i = 0$ and $\sum_j \tau_j = 0$. If we add these two constraints below \mathbf{X} , we

get

$$\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{H} \end{pmatrix} \boldsymbol{\beta} = \left(\begin{array}{ccccc} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ \hline 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right) \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \tau_1 \\ \tau_2 \end{pmatrix},$$

where the augmented matrix now has five linearly independent columns. Thus given $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ is now unique. \square

EXERCISES 3h

1. Suppose that \mathbf{X} does not have full rank, and let $\hat{\boldsymbol{\beta}}_i$ ($i = 1, 2$) be any two solutions of the normal equations. Show directly that

$$\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2.$$

2. If the columns of \mathbf{X} are linearly dependent, prove that there is no matrix \mathbf{C} such that \mathbf{CY} is an unbiased estimate of $\boldsymbol{\beta}$.

3.9.2 Estimable Functions

Since $\hat{\boldsymbol{\beta}}$ is not unique, $\boldsymbol{\beta}$ is not estimable. The question then arises: What can we estimate? Since each element θ_i of $\boldsymbol{\theta}$ ($= \mathbf{X}\boldsymbol{\beta}$) is estimated by the i th element of $\hat{\boldsymbol{\theta}} = \mathbf{PY}$, then every linear combination of the θ_i , say $\mathbf{b}'\boldsymbol{\theta}$, is also estimable. This means that the θ_i form a linear subspace of estimable functions, where $\theta_i = \mathbf{x}_i'\boldsymbol{\beta}$, \mathbf{x}_i' being the i th row of \mathbf{X} . Usually, we define estimable functions formally as follows.

Definition 3.1 *The parametric function $\mathbf{a}'\boldsymbol{\beta}$ is said to be estimable if it has a linear unbiased estimate, $\mathbf{b}'\mathbf{Y}$, say.*

We note that if $\mathbf{a}'\boldsymbol{\beta}$ is estimable, then $\mathbf{a}'\boldsymbol{\beta} = E[\mathbf{b}'\mathbf{Y}] = \mathbf{b}'\boldsymbol{\theta} = \mathbf{b}'\mathbf{X}\boldsymbol{\beta}$ identically in $\boldsymbol{\beta}$, so that $\mathbf{a}' = \mathbf{b}'\mathbf{X}$ or $\mathbf{a} = \mathbf{X}'\mathbf{b}$ (A.11.1). Hence $\mathbf{a}'\boldsymbol{\beta}$ is estimable if and only if $\mathbf{a} \in \mathcal{C}(\mathbf{X}')$.

EXAMPLE 3.8 If $\mathbf{a}'\boldsymbol{\beta}$ is estimable, and $\hat{\boldsymbol{\beta}}$ is any solution of the normal equations, then $\mathbf{a}'\hat{\boldsymbol{\beta}}$ is unique. To show this we first note that $\mathbf{a} = \mathbf{X}'\mathbf{b}$ for some \mathbf{b} , so that $\mathbf{a}'\boldsymbol{\beta} = \mathbf{b}'\mathbf{X}\boldsymbol{\beta} = \mathbf{b}'\boldsymbol{\theta}$. Similarly, $\mathbf{a}'\hat{\boldsymbol{\beta}} = \mathbf{b}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{b}'\hat{\boldsymbol{\theta}}$, which is unique. Furthermore, by Theorem 3.2, $\mathbf{b}'\hat{\boldsymbol{\theta}}$ is the BLUE of $\mathbf{b}'\boldsymbol{\theta}$, so that $\mathbf{a}'\hat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{a}'\boldsymbol{\beta}$. \square

In conclusion, the simplest approach to estimable functions is to avoid them altogether by transforming the model into a full-rank model!

EXERCISES 3i

1. Prove that $\mathbf{a}'E[\hat{\beta}]$ is an estimable function of β .
2. If $\mathbf{a}_1'\beta, \mathbf{a}_2'\beta, \dots, \mathbf{a}_k'\beta$ are estimable, prove that any linear combination of these is also estimable.
3. If $\mathbf{a}'\hat{\beta}$ is invariant with respect to $\hat{\beta}$, prove that $\mathbf{a}'\beta$ is estimable.
4. Prove that $\mathbf{a}'\beta$ is estimable if and only if

$$\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{a}'.$$

(Note that $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.)

5. If $\mathbf{a}'\beta$ is an estimable function, prove that

$$\text{Var}[\mathbf{a}'\hat{\beta}] = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a}.$$

6. Prove that all linear functions $\mathbf{a}'\beta$ are estimable if and only if the columns of \mathbf{X} are linearly independent.

3.9.3 Introducing Further Explanatory Variables

If we wish to introduce further explanatory variables into a less-than-full-rank model, we can, once again, reduce the model to one of full rank. As in Section 3.7, we see what happens when we add $\mathbf{Z}\gamma$ to our model $\mathbf{X}\beta$. It makes sense to assume that \mathbf{Z} has full column rank and that the columns of \mathbf{Z} are linearly independent of the columns of \mathbf{X} . Using the full-rank model

$$\mathbf{Y} = \mathbf{X}_1\alpha + \mathbf{Z}\gamma + \varepsilon,$$

where \mathbf{X}_1 is $n \times r$ of rank r , we find that Theorem 3.6(ii), (iii), and (iv) of Section 3.7.1 still hold. To see this, one simply works through the same steps of the theorem, but replacing \mathbf{X} by \mathbf{X}_1 , β by α , and \mathbf{R} by $\mathbf{I}_n - \mathbf{P}$, where $\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1$ is the unique projection matrix projecting onto $\mathcal{C}(\mathbf{X})$.

3.9.4 Introducing Linear Restrictions

Referring to Section 3.8, suppose that we have a set of linear restrictions $\mathbf{a}_i'\beta = 0$ ($i = 1, 2, \dots, q$), or in matrix form, $\mathbf{A}\beta = 0$. Then a realistic assumption is that these constraints are all estimable. This implies that $\mathbf{a}_i' = \mathbf{m}_i'\mathbf{X}$ for some \mathbf{m}_i , or $\mathbf{A} = \mathbf{M}\mathbf{X}$, where \mathbf{M} is $q \times n$ of rank q [as $q = \text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{M})$]

by A.2.1]. Since $\mathbf{A}\boldsymbol{\beta} = \mathbf{M}\mathbf{X}\boldsymbol{\beta} = \mathbf{M}\boldsymbol{\theta}$, we therefore find the restricted least squares estimate of $\boldsymbol{\theta}$ by minimizing $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ subject to $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X}) = \Omega$ and $\mathbf{M}\boldsymbol{\theta} = \mathbf{0}$, that is, subject to

$$\boldsymbol{\theta} \in \mathcal{N}(\mathbf{M}) \cap \Omega \quad (= \omega, \text{ say}).$$

If \mathbf{P}_Ω and \mathbf{P}_ω are the projection matrices projecting onto Ω and ω , respectively, then we want to find $\hat{\boldsymbol{\theta}}_\omega = \mathbf{P}_\omega \mathbf{Y}$. Now, from B.3.2 and B.3.3,

$$\mathbf{P}_\Omega - \mathbf{P}_\omega = \mathbf{P}_{\omega^\perp \cap \Omega},$$

where $\omega^\perp \cap \Omega = \mathcal{C}(\mathbf{B})$ and $\mathbf{B} = \mathbf{P}_\Omega \mathbf{M}'$. Thus

$$\begin{aligned}\hat{\boldsymbol{\theta}}_\omega &= \mathbf{P}_\omega \mathbf{Y} \\ &= \mathbf{P}_\Omega \mathbf{Y} - \mathbf{P}_{\omega^\perp \cap \Omega} \mathbf{Y} \\ &= \hat{\boldsymbol{\theta}}_\Omega - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{Y}.\end{aligned}$$

EXERCISES 3j

1. If \mathbf{P} projects onto $\mathcal{C}(\mathbf{X})$, show that $\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}$ is nonsingular.
2. Prove that if \mathbf{X}_1 is $n \times r$ of rank r and consists of a set of r linearly independent columns of \mathbf{X} , then $\mathbf{X} = \mathbf{X}_1 \mathbf{L}$, where \mathbf{L} is $r \times p$ of rank r .
3. Prove that \mathbf{B} has full column rank [i.e., $(\mathbf{B}'\mathbf{B})^- = (\mathbf{B}'\mathbf{B})^{-1}$].
4. If \mathbf{X} has full rank and $\hat{\boldsymbol{\theta}}_\omega = \mathbf{X}\hat{\boldsymbol{\beta}}_H$, show that

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

[This is a special case of (3.38).]

5. Show how to modify the theory above to take care of the case when the restrictions are $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ ($\mathbf{c} \neq \mathbf{0}$).

3.10 GENERALIZED LEAST SQUARES

Having developed a least squares theory for the full-rank model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$, we now consider what modifications are necessary if we allow the ε_i to be correlated. In particular, we assume that $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{V}$, where \mathbf{V} is a *known* $n \times n$ positive-definite matrix.

Since \mathbf{V} is positive-definite, there exists an $n \times n$ nonsingular matrix \mathbf{K} such that $\mathbf{V} = \mathbf{K}\mathbf{K}'$ (A.4.2). Therefore, setting $\mathbf{Z} = \mathbf{K}^{-1}\mathbf{Y}$, $\mathbf{B} = \mathbf{K}^{-1}\mathbf{X}$, and

$\eta = \mathbf{K}^{-1}\epsilon$, we have the model $\mathbf{Z} = \mathbf{B}\beta + \eta$, where \mathbf{B} is $n \times p$ of rank p (A.2.2). Also, $E[\eta] = \mathbf{0}$ and

$$\text{Var}[\eta] = \text{Var}[\mathbf{K}^{-1}\epsilon] = \mathbf{K}^{-1}\text{Var}[\epsilon]\mathbf{K}^{-1'} = \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}'\mathbf{K}'^{-1} = \sigma^2\mathbf{I}_n.$$

Minimizing $\eta'\eta$ with respect to β , and using the theory of Section 3.1, the least squares estimate of β for this transformed model is

$$\begin{aligned}\beta^* &= (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{Z} \\ &= (\mathbf{X}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},\end{aligned}$$

with expected value

$$E[\beta^*] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta = \beta,$$

dispersion matrix

$$\begin{aligned}\text{Var}[\beta^*] &= \sigma^2(\mathbf{B}'\mathbf{B})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1},\end{aligned}\tag{3.48}$$

and residual sum of squares

$$\begin{aligned}\mathbf{f}'\mathbf{f} &= (\mathbf{Z} - \mathbf{B}\beta^*)'(\mathbf{Z} - \mathbf{B}\beta^*) \\ &= (\mathbf{Y} - \mathbf{X}\beta^*)'(\mathbf{K}\mathbf{K}')^{-1}(\mathbf{Y} - \mathbf{X}\beta^*) \\ &= (\mathbf{Y} - \mathbf{X}\beta^*)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta^*).\end{aligned}$$

Alternatively, we can obtain β^* simply by differentiating

$$\begin{aligned}\eta'\eta &= \epsilon'\mathbf{V}^{-1}\epsilon \\ &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} + \beta'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta\end{aligned}$$

with respect to β . Thus, by A.8,

$$\frac{\partial\eta'\eta}{\partial\beta} = -2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} + 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta,\tag{3.49}$$

and setting this equal to zero leads once again to β^* . Using this approach instead of the general theory above, we see that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ has an inverse, as it is positive-definite (by A.4.5). We note that the coefficient of 2β in (3.49) gives us the inverse of $\text{Var}[\beta^*]/\sigma^2$.

There is some variation in terminology among books dealing with the model above: Some texts call β^* the weighted least squares estimate. However, we call β^* the *generalized least squares estimate* and reserve the expression *weighted least squares* for the case when \mathbf{V} is a diagonal matrix: The diagonal

case is discussed in various places throughout this book (see, e.g., Section 10.4).

EXAMPLE 3.9 Let $\mathbf{Y} = \mathbf{x}\beta + \boldsymbol{\epsilon}$, where $\mathbf{Y} = (Y_i)$ and $\mathbf{x} = (x_i)$ are $n \times 1$ vectors, $E[\boldsymbol{\epsilon}] = 0$ and $\text{Var}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{V}$. If $\mathbf{V} = \text{diag}(w_1^{-1}, w_2^{-1}, \dots, w_n^{-1})$ ($w_i > 0$), we now find the weighted least squares estimate of β and its variance. Here it is simpler to differentiate $\boldsymbol{\eta}'\boldsymbol{\eta}$ directly rather than use the general matrix theory. Thus, since $\mathbf{V}^{-1} = \text{diag}(w_1, w_2, \dots, w_n)$,

$$\boldsymbol{\eta}'\boldsymbol{\eta} = \sum_i (Y_i - x_i\beta)^2 w_i$$

and

$$\frac{\partial \boldsymbol{\eta}'\boldsymbol{\eta}}{\partial \beta} = -2 \sum_i x_i(Y_i - x_i\beta)w_i. \quad (3.50)$$

Setting the right side of (3.50) equal to zero leads to

$$\beta^* = \frac{\sum_i w_i Y_i x_i}{\sum_i w_i x_i^2}$$

and from the coefficient of 2β ,

$$\text{var}[\beta^*] = \sigma^2 \left(\sum_i w_i x_i^2 \right)^{-1}.$$

We can also find the variance directly from

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} = \left(\sum_i w_i x_i^2 \right)^{-1}. \quad \square$$

Since the generalized least squares estimate is simply the ordinary least squares estimate (OLSE) for a transformed model, we would expect β^* to have the same optimal properties, namely, that $\mathbf{a}'\beta^*$ is the best linear unbiased estimate (BLUE) of $\mathbf{a}'\beta$. To see this, we note that

$$\mathbf{a}'\beta^* = \mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{b}'\mathbf{Y},$$

say, is linear and unbiased. Let $\mathbf{b}'_1\mathbf{Y}$ be any other linear unbiased estimate of $\mathbf{a}'\beta$. Then, using the transformed model, $\mathbf{a}'\beta^* = \mathbf{a}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{Z}$ and $\mathbf{b}'_1\mathbf{Y} = \mathbf{b}'_1\mathbf{K}\mathbf{K}^{-1}\mathbf{Y} = (\mathbf{K}'\mathbf{b}_1)'Z$. By Theorem 3.2 (Section 3.2) and the ensuing argument,

$$\text{var}[\mathbf{a}'\beta^*] \leq \text{var}[(\mathbf{K}'\mathbf{b}_1)'Z] = \text{var}[\mathbf{b}'_1\mathbf{Y}].$$

Equality occurs if and only if $(\mathbf{K}'\mathbf{b}_1)' = \mathbf{a}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$, or

$$\mathbf{b}'_1 = \mathbf{a}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{K}^{-1} = \mathbf{a}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{b}'.$$

Thus $\mathbf{a}'\beta^*$ is the unique BLUE of $\mathbf{a}'\beta$. Note that the ordinary least squares estimate $\mathbf{a}'\hat{\beta}$ will still be an unbiased estimate of $\mathbf{a}'\beta$, but $\text{var}[\mathbf{a}'\hat{\beta}] \geq \text{var}[\mathbf{a}'\beta^*]$.

EXERCISES 3k

1. Let $Y_i = \beta x_i + \varepsilon_i$ ($i = 1, 2$), where $\varepsilon_1 \sim N(0, \sigma^2)$, $\varepsilon_2 \sim N(0, 2\sigma^2)$, and ε_1 and ε_2 are statistically independent. If $x_1 = +1$ and $x_2 = -1$, obtain the weighted least squares estimate of β and find the variance of your estimate.
 2. Let Y_i ($i = 1, 2, \dots, n$) be independent random variables with a common mean θ and variances σ^2/w_i ($i = 1, 2, \dots, n$). Find the linear unbiased estimate of θ with minimum variance, and find this minimum variance.
 3. Let Y_1, Y_2, \dots, Y_n be independent random variables, and let Y_i have a $N(i\theta, i^2\sigma^2)$ distribution for $i = 1, 2, \dots, n$. Find the weighted least squares estimate of θ and prove that its variance is σ^2/n .
 4. Let Y_1, Y_2, \dots, Y_n be random variables with common mean θ and with dispersion matrix $\sigma^2 \mathbf{V}$, where $v_{ii} = 1$ ($i = 1, 2, \dots, n$) and $v_{ij} = \rho$ ($0 < \rho < 1$; $i, j = 1, 2, \dots, n$; $i \neq j$). Find the generalized least squares estimate of θ and show that it is the same as the ordinary least squares estimate. Hint: \mathbf{V}^{-1} takes the same form as \mathbf{V} .
- (McElroy [1967])
5. Let $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{V})$, where \mathbf{X} is $n \times p$ of rank p and \mathbf{V} is a known positive-definite $n \times n$ matrix. If $\boldsymbol{\beta}^*$ is the generalized least squares estimate of $\boldsymbol{\beta}$, prove that
 - (a) $Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*)/\sigma^2 \sim \chi^2_{n-p}$.
 - (b) Q is the quadratic nonnegative unbiased estimate of $(n-p)\sigma^2$ with minimum variance.
 - (c) If $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta}^* = \mathbf{P}^*\mathbf{Y}$, then \mathbf{P}^* is idempotent but not, in general, symmetric.
 6. Suppose that $E[\mathbf{Y}] = \boldsymbol{\theta}$, $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, and $\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{V}$, where \mathbf{A} is a $q \times n$ matrix of rank q and \mathbf{V} is a known $n \times n$ positive-definite matrix. Let $\boldsymbol{\theta}^*$ be the generalized least squares estimate of $\boldsymbol{\theta}$; that is, $\boldsymbol{\theta}^*$ minimizes $(\mathbf{Y} - \boldsymbol{\theta})' \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\theta})$ subject to $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$. Show that

$$\mathbf{Y} - \boldsymbol{\theta}^* = \mathbf{V}\mathbf{A}'\boldsymbol{\gamma}^*,$$

where $\boldsymbol{\gamma}^*$ is the generalized least squares estimate of $\boldsymbol{\gamma}$ for the model $E[\mathbf{Y}] = \mathbf{V}\mathbf{A}'\boldsymbol{\gamma}$, $\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{V}$.

(Wedderburn [1974])

3.11 CENTERING AND SCALING THE EXPLANATORY VARIABLES

It is instructive to consider the effect of centering and scaling the x -variables on the regression model. We shall use this theory later in the book.

3.11.1 Centering

Up until now, we have used the model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Suppose, however, that we center the x -data and use the reparameterized model

$$Y_i = \alpha_0 + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{p-1}) + \varepsilon_i,$$

where

$$\alpha_0 = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_{p-1}\bar{x}_{p-1}$$

and $\bar{x}_j = \sum_i x_{ij}/n$. Our model is now $\mathbf{Y} = \mathbf{X}_c\alpha + \varepsilon$, where

$$\alpha' = (\alpha_0, \beta_1, \dots, \beta_{p-1}) = (\alpha_0, \beta_c'),$$

$\mathbf{X}_c = (\mathbf{1}_n, \tilde{\mathbf{X}})$, and $\tilde{\mathbf{X}}$ has typical element $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$. Because the transformation between α and β is one-to-one, the least squares estimate of β_c remains the same. Then, since $\tilde{\mathbf{X}}'\mathbf{1}_n = 0$,

$$\begin{aligned}\hat{\alpha} &= (\mathbf{X}'_c \mathbf{X}_c)^{-1} \mathbf{X}'_c \mathbf{Y} \\ &= \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}'_n \mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{Y} \end{pmatrix} \\ &= \begin{pmatrix} n^{-1} & \mathbf{0}' \\ \mathbf{0} & (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_n \mathbf{Y} \\ \tilde{\mathbf{X}}'\mathbf{Y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{Y} \\ (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y} \end{pmatrix},\end{aligned}\tag{3.51}$$

so that $\hat{\alpha}_0 = \bar{Y}$ and $\hat{\beta}_c = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}$. Now $C(\mathbf{X}_c) = C(\mathbf{X})$, which can be proved by subtracting $\bar{x}_j \times$ column (1) from column $(j+1)$ of \mathbf{X} for each $j = 1, \dots, p-1$. Hence \mathbf{X}_c and \mathbf{X} have the same projection matrices, so that

$$\begin{aligned}\mathbf{P} &= \mathbf{X}_c(\mathbf{X}'_c \mathbf{X}_c)^{-1} \mathbf{X}'_c \\ &= (\mathbf{1}_n, \tilde{\mathbf{X}}) \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{pmatrix}^{-1} (\mathbf{1}_n, \tilde{\mathbf{X}})' \\ &= \frac{1}{n} \mathbf{1}_n \mathbf{1}'_n + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'.\end{aligned}\tag{3.52}$$

Let \mathbf{x}_i now represent the i th row of \mathbf{X} , but *reduced* in the sense that the initial unit element (corresponding to α_0) is omitted. Picking out the i th diagonal element of (3.52), we get

$$\begin{aligned}p_{ii} &= n^{-1} + (n-1)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}_{zz}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= n^{-1} + (n-1)^{-1} \text{MD}_i,\end{aligned}\tag{3.53}$$

where $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i$, \mathbf{S}_{zz} is the sample covariance matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}/(n-1)$, and MD_i is the Mahalanobis distance between the i th reduced row of \mathbf{X} and the

average reduced row (cf. Seber [1984: p. 10]). Thus p_{ii} is a measure of how far away \mathbf{x}_i is from the center of the x -data.

We note that the centering and subsequent reparameterization of the model do not affect the fitted model $\hat{\mathbf{Y}}$, so that the residuals for both the centered and uncentered models are the same. Hence, from (3.52), the residual sum of squares for both models is given by

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= \mathbf{Y}' \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' - \tilde{\mathbf{X}} (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}} \right) \mathbf{Y} \\ &= \mathbf{Y}' \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) \mathbf{Y} - \mathbf{Y}' \tilde{\mathbf{P}} \mathbf{Y} \\ &= \sum_i (Y_i - \bar{Y})^2 - \mathbf{Y}' \tilde{\mathbf{P}} \mathbf{Y}, \end{aligned} \quad (3.54)$$

where $\tilde{\mathbf{P}} = \tilde{\mathbf{X}} (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}'$. We will use this result later.

3.11.2 Scaling

Suppose that we now also scale the columns of $\tilde{\mathbf{X}}$ so that they have unit length. Let $s_j^2 = \sum_i \tilde{x}_{ij}^2$ and consider the new variables $x_{ij}^* = \tilde{x}_{ij}/s_j$. Then our model becomes

$$Y_i = \alpha_0 + \gamma_1 x_{i1}^* + \cdots + \gamma_{p-1} x_{ip-1}^*,$$

where $\gamma_j = \beta_j s_j$. Because the transformation is still one-to-one, $\hat{\gamma}_j = \hat{\beta}_j s_j$ and $\hat{\alpha}_0 = \bar{Y}$. If $\mathbf{X}^* = (x_{ij}^*)$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{p-1})'$, then replacing $\tilde{\mathbf{X}}$ by \mathbf{X}^* in (3.51) gives us

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} \mathbf{X}^{*\prime} \mathbf{Y} = \mathbf{R}_{xx}^{-1} \mathbf{X}^{*\prime} \mathbf{Y}, \quad (3.55)$$

where \mathbf{R}_{xx} is now the (symmetric) correlation matrix

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{pmatrix}$$

and

$$r_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/(s_j s_k)$$

is the (sample) correlation coefficient of the j th and k th explanatory variables. If we introduce the notation $\mathbf{X}^* = (\mathbf{x}^{*(1)}, \dots, \mathbf{x}^{*(p-1)})$ for the columns of \mathbf{X}^* , we see that $r_{jk} = \mathbf{x}^{*(j)'} \mathbf{x}^{*(k)}$.

EXAMPLE 3.10 For later reference we consider the special case of $p = 3$. Then

$$\mathbf{R}_{xx} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where $r = \mathbf{x}^{*(1)'} \mathbf{x}^{*(2)}$. Also, from (3.55), we have

$$\begin{aligned} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} &= \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}^{*(1)'} \mathbf{Y} \\ \mathbf{x}^{*(2)'} \mathbf{Y} \end{pmatrix} \\ &= \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}^{*(1)'} \mathbf{Y} \\ \mathbf{x}^{*(2)'} \mathbf{Y} \end{pmatrix}, \end{aligned}$$

so that

$$\hat{\gamma}_1 = \frac{1}{1-r^2} (\mathbf{x}^{*(1)'} \mathbf{Y} - r \mathbf{x}^{*(2)'} \mathbf{Y}) \quad \text{and} \quad \hat{\beta}_1 = \hat{\gamma}_1 / s_1. \quad (3.56)$$

By interchanging the superscripts (1) and (2), we get

$$\hat{\gamma}_2 = \frac{1}{1-r^2} (\mathbf{x}^{*(2)'} \mathbf{Y} - r \mathbf{x}^{*(1)'} \mathbf{Y}) \quad \text{and} \quad \hat{\beta}_2 = \hat{\gamma}_2 / s_2.$$

Since

$$\bar{\mathbf{X}} = \mathbf{X}^* \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix} = \mathbf{X}^* \mathbf{S}_d,$$

say, it follows that

$$\begin{aligned} \mathbf{P} &= n^{-1} \mathbf{1}_n \mathbf{1}_n' + \mathbf{X}^* \mathbf{S}_d (\mathbf{S}_d \mathbf{X}^{*'} \mathbf{X}^* \mathbf{S}_d)^{-1} \mathbf{S}_d \mathbf{X}^{*'} \\ &= n^{-1} \mathbf{1}_n \mathbf{1}_n' + \mathbf{X}^* (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \\ &= n^{-1} \mathbf{1}_n \mathbf{1}_n' \\ &\quad + \frac{1}{1-r^2} (\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}) \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} (\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)})' \end{aligned} \quad (3.57)$$

$$\begin{aligned} &= n^{-1} \mathbf{1}_n \mathbf{1}_n' + \mathbf{x}^{*(2)} \mathbf{x}^{*(2)'} \\ &\quad + \frac{1}{1-r^2} (\mathbf{x}_*^{(1)} - r \mathbf{x}^{*(2)}) (\mathbf{x}^{*(1)} - r \mathbf{x}^{*(2)})'. \end{aligned} \quad (3.58)$$

□

EXERCISES 3I

- If $\tilde{Y}_i = Y_i - \bar{Y}$ and $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)'$, prove from (3.54) that $\text{RSS} = \tilde{\mathbf{Y}}' (\mathbf{I}_n - \tilde{\mathbf{P}}) \tilde{\mathbf{Y}}$.
- Suppose that we consider fitting a model in which the Y -data are centered and scaled as well as the x -data. This means that we use $Y_i^* = (Y_i - \bar{Y})/s_y$ instead of Y_i , where $s_y^2 = \sum_i (Y_i - \bar{Y})^2$. Using (3.54), obtain an expression for RSS from this model.

3.12 BAYESIAN ESTIMATION

This method of estimation utilizes any prior information that we have about the parameter vector $\theta = (\beta', \sigma')$. We begin with the probability density function of \mathbf{Y} , $f(\mathbf{y}, \theta)$, say, which we have assumed to be multivariate normal in this chapter, and we now wish to incorporate prior knowledge about θ , which is expressed in terms of some density function $f(\theta)$ of θ . Our aim is to make inferences on the basis of the density function of θ given $\mathbf{Y} = \mathbf{y}$, the *posterior density function* of θ . To do this, we use Bayes' formula,

$$\begin{aligned} f(\theta|\mathbf{y}) &= \frac{f(\theta, \mathbf{y})}{f(\mathbf{y})} \\ &= \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})} \\ &= cf(\mathbf{y}|\theta)f(\theta), \end{aligned} \quad (3.59)$$

where c does not involve θ . It is usual to assume that β and σ have independent prior distributions, so that

$$f(\theta) = f_1(\beta)f_2(\sigma).$$

Frequently, one uses the *noninformative prior* (see Box and Taio [1973: Section 1.3] for a justification) in which β and $\log \sigma$ are assumed to be locally uniform and $\sigma > 0$. This translates into $f_1(\beta) = \text{constant}$ and $f_2(\sigma) \propto 1/\sigma$. These priors are described as *improper*, as their integrals are technically infinite (although we can get around this by making the intervals of the uniform distributions sufficiently large). Using these along with the independence assumption, we obtain from (3.59)

$$\begin{aligned} f(\beta, \sigma|\mathbf{y}) &= cf(\mathbf{y}|\theta)\sigma^{-1} \\ &= c(2\pi)^{-n/2}\sigma^{-(n+1)}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2\right). \end{aligned}$$

Using the result

$$\int_0^\infty x^{-(b+1)}\exp(-a/x^2)dx = \frac{1}{2}a^{-b/2}\Gamma(b/2) \quad (3.60)$$

derived from the gamma distribution, we find that

$$\begin{aligned} f(\beta|\mathbf{y}) &\propto \int_0^\infty f(\beta, \sigma|\mathbf{y})d\sigma \\ &\propto \|\mathbf{y} - \mathbf{X}\beta\|^{-n}. \end{aligned} \quad (3.61)$$

Now, from Exercises 3a, No. 1,

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\beta\|^2 &= \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2 \\ &= (n-p)s^2 + \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2, \end{aligned} \quad (3.62)$$

so that

$$f(\beta|y) \propto \left[1 + \frac{(\beta - \hat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \hat{\beta})}{(n-p)s^2} \right]^{-n/2}.$$

This is a special case of the p -dimensional multivariate t -distribution

$$f(t) = \frac{\Gamma(\frac{1}{2}[\nu + p])}{(\pi\nu)^{p/2}\Gamma(\frac{1}{2}\nu)|\Sigma|^{1/2}} [1 + \nu^{-1}(t - \mu)' \Sigma^{-1}(t - \mu)]^{-(\nu+p)/2},$$

with $\nu = n - p$, $\Sigma = s^2(\mathbf{X}'\mathbf{X})^{-1}$, and $\mu = \hat{\beta}$.

What estimate of β do we use? If we use the mean or the mode of the posterior distribution (which are the same in this case, as the distribution is symmetric) we get $\hat{\beta}$, the least squares estimate. For interval inferences, the marginal posterior distribution of β_r is a t -distribution given by

$$\frac{\beta_r - \hat{\beta}_r}{s\sqrt{c^{r+1,r+1}}} \sim t_{n-p},$$

where $(c^{rs}) = (\mathbf{X}'\mathbf{X})^{-1}$.

If some information is available on θ , it is convenient, computationally, to use a *conjugate prior*, one that combines with $f(y|\theta)$ to give a posterior distribution which has the same form as the prior. For example, suppose that $f(\beta|\sigma^2)$ is the density function for the $N_p(\mathbf{m}, \sigma^2 \mathbf{V})$ distribution and that σ^2 has an inverted gamma distribution with density function

$$f(\sigma^2) \propto (\sigma^2)^{-(d+2)/2} \exp\left(-\frac{a}{2\sigma^2}\right). \quad (3.63)$$

Then

$$\begin{aligned} f(\beta, \sigma^2) &= f(\beta|\sigma^2)f(\sigma^2) \\ &\propto (\sigma^2)^{-(d+p+2)/2} \exp\left\{-\frac{1}{2\sigma^2}[(\beta - \mathbf{m})' \mathbf{V}^{-1}(\beta - \mathbf{m}) + a]\right\}. \end{aligned}$$

Combining this prior with the normal likelihood function, we obtain

$$\begin{aligned} f(\beta, \sigma^2|y) &\propto f(y|\beta, \sigma^2)f(\beta, \sigma^2) \\ &\propto (\sigma^2)^{-(d+p+n+2)/2} \exp[-(Q + a)/(2\sigma^2)], \end{aligned}$$

where

$$Q = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \mathbf{m})' \mathbf{V}^{-1}(\beta - \mathbf{m}).$$

We can now integrate out σ^2 to get the posterior density of β . Thus

$$\begin{aligned} f(\beta|y) &= \int_0^\infty f(\beta, \sigma^2|y) d\sigma^2 \\ &\propto \int_0^\infty (\sigma^2)^{-(d+n+p+2)/2} \exp[-(Q + a)/(2\sigma^2)] d\sigma^2. \end{aligned}$$

Using the standard integral formula

$$\int_0^\infty x^{-(\nu+1)} \exp(-k/x) dx = \Gamma(\nu) k^{-\nu},$$

we see that the posterior density is proportional to

$$(Q + a)^{-(d+n+p)/2}. \quad (3.64)$$

To make further progress, we need the following result.

THEOREM 3.7 Define $\mathbf{V}_* = (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}$ and let \mathbf{m}_* be given by $\mathbf{m}_* = \mathbf{V}_*(\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m})$. Then

$$Q = (\beta - \mathbf{m}_*)'\mathbf{V}_*^{-1}(\beta - \mathbf{m}_*) + (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}). \quad (3.65)$$

Proof.

$$\begin{aligned} Q &= \beta'(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})\beta - 2\beta'(\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}) + \mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} \\ &= \beta'\mathbf{V}_*^{-1}\beta - 2\beta'\mathbf{V}_*^{-1}\mathbf{m}_* + \mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} \\ &= (\beta - \mathbf{m}_*)'\mathbf{V}_*^{-1}(\beta - \mathbf{m}_*) + \mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}_*\mathbf{V}_*^{-1}\mathbf{m}_*. \end{aligned}$$

Thus, it is enough to show that

$$\mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}_*\mathbf{V}_*^{-1}\mathbf{m}_* = (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}). \quad (3.66)$$

Consider

$$\begin{aligned} \mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} - \mathbf{m}_*\mathbf{V}_*^{-1}\mathbf{m}_* &= \mathbf{y}'\mathbf{y} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} - (\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m})'\mathbf{V}_*(\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}) \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{V}_*\mathbf{V}^{-1}\mathbf{m} \\ &\quad + \mathbf{m}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{V}_*\mathbf{V}^{-1})\mathbf{m}. \end{aligned} \quad (3.67)$$

By the definition of \mathbf{V}_* , we have $\mathbf{V}_*(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1}) = \mathbf{I}_p$, so that

$$\mathbf{V}_*\mathbf{V}^{-1} = \mathbf{I}_p - \mathbf{V}_*\mathbf{X}'\mathbf{X}$$

and

$$\begin{aligned} \mathbf{X}\mathbf{V}_*\mathbf{V}^{-1} &= \mathbf{X} - \mathbf{X}\mathbf{V}_*\mathbf{X}'\mathbf{X} \\ &= (\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{X}. \end{aligned} \quad (3.68)$$

Also, by A.9.3,

$$\begin{aligned} \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{V}_*\mathbf{V}^{-1} &= \mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}\mathbf{V}^{-1} \\ &= (\mathbf{V} + (\mathbf{X}'\mathbf{X})^{-1})^{-1} \\ &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{X}'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{X}. \end{aligned} \quad (3.69)$$

Substituting (3.68) and (3.69) into (3.67), we get

$$\begin{aligned} \mathbf{y}'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{y} - 2\mathbf{y}'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{m} + \mathbf{m}'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')\mathbf{m} \\ = (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{m}). \end{aligned} \quad (3.70)$$

Finally, again using A.9.3,

$$\begin{aligned} (\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1} &= \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}\mathbf{X}' \\ &= \mathbf{I}_n - \mathbf{X}\mathbf{V}_*\mathbf{X}', \end{aligned}$$

proving the theorem. \square

Using Theorem 3.7, we see from (3.64) that the posterior density of β is proportional to

$$[a_* + (\beta - \mathbf{m}_*)'\mathbf{V}_*^{-1}(\beta - \mathbf{m}_*)]^{-(n+d+p)/2},$$

where

$$a_* = a + (\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}).$$

This is proportional to

$$[1 + (n + d)^{-1}(\beta - \mathbf{m}_*)'\mathbf{W}_*^{-1}(\beta - \mathbf{m}_*)]^{-(d+n+p)/2},$$

where $\mathbf{W}_* = a_*\mathbf{V}_*/(n + d)$, so from A.13.5, the posterior distribution of β is a multivariate $t_p(n + d, \mathbf{m}_*, \mathbf{W}_*)$. In particular, the posterior mean (and mode) is \mathbf{m}_* , which we can take as our Bayes' estimate of β .

These arguments give the flavor of the algebra involved in Bayesian regression. Further related distributions are derived by O'Hagen [1994: Chapter 9] and in Section 12.6.2. Clearly, the choice of prior is critical and a necessary requirement in the conjugate prior approach is the choice of the values of \mathbf{m} and \mathbf{V} . These might come from a previous experiment, for example. Distributions other than the normal can also be used for the likelihood, and numerical methods are available for computing posterior likelihoods when analytical solutions are not possible. Numerical methods are surveyed in Evans and Swartz [1995]. For further practical details, the reader is referred to Gelman et al. [1995], for example.

EXERCISES 3m

1. Derive equations (3.60) and (3.61).
2. Using the noninformative prior for θ , show that the conditional posterior density $f(\beta|\mathbf{y}, \sigma)$ is multivariate normal. Hence deduce that the posterior mean of β is $\hat{\beta}$.
3. Suppose that we use the noninformative prior for θ .
 - (a) If $v = \sigma^2$, show that $f(v) \propto 1/v$.

- (b) Obtain an expression proportional to $f(\beta, v|y)$.
 (c) Using (3.62), integrate out β to obtain

$$f(v|y) \propto v^{-(\nu/2+1)} \exp\left(-\frac{a}{v}\right),$$

where $\nu = n - p$ and $a = ||y - X\hat{\beta}||^2/2$.

- (d) Find the posterior mean of v .

3.13 ROBUST REGRESSION

Least squares estimates are the most efficient unbiased estimates of the regression coefficients when the errors are normally distributed. However, they are not very efficient when the distribution of the errors is long-tailed. Under these circumstances, we can expect outliers in the data: namely, observations whose errors ϵ_i are extreme. We will see in Section 9.5 that least squares fits are unsatisfactory when outliers are present in the data, and for this reason alternative methods of fitting have been developed that are not as sensitive to outliers.

When fitting a regression, we minimize some average measure of the size of the residuals. We can think of least squares as “least mean of squares” which fits a regression by minimizing the mean of the squared residuals (or, equivalently, the sum of the squared residuals). Thus, least squares solves the minimization problem

$$\min_{\mathbf{b}} \frac{1}{n} \sum_{i=1}^n e_i^2(\mathbf{b}),$$

where $e_i(\mathbf{b}) = Y_i - \mathbf{x}'_i \mathbf{b}$. Here, *average* is interpreted as the mean and *size* as the square. The sensitivity of least squares to outliers is due to two factors. First, if we measure size using the squared residual, any residual with a large magnitude will have a very large size relative to the others. Second, by using a measure of location such as a mean that is not robust, any large square will have a very strong impact on the criterion, resulting in the extreme data point having a disproportionate influence on the fit.

Two remedies for this problem have become popular. First, we can measure size in some other way, by replacing the square e^2 by some other function $\rho(e)$ which reflects the size of the residual in a less extreme way. To be a sensible measure of size, the function ρ should be symmetric [i.e., $\rho(e) = \rho(-e)$], positive [$\rho(e) \geq 0$] and monotone [$\rho(|e_1|) \geq \rho(|e_2|)$ if $|e_1| \geq |e_2|$]. This idea leads to the notion of M-estimation, discussed by, for example, Huber [1981: Chapter 7], Hampel et al. [1986: Chapter 6], and Birkes and Dodge [1993: Chapter 5].

Second, we can replace the sum (or, equivalently, the mean) by a more robust measure of location such as the median or a trimmed mean. Regression

methods based on this idea include least median of squares and least trimmed squares, described in Rousseeuw [1984] and Rousseeuw and Leroy [1987]. A related idea is to minimize some robust measure of the scale of the residuals (Rousseeuw and Yohai [1984]).

3.13.1 M-Estimates

Suppose that the observed responses Y_i are independent and have density functions

$$f_i(y_i; \beta, \sigma) = \frac{1}{\sigma} f\left(\frac{y_i - \mathbf{x}'_i \beta}{\sigma}\right), \quad (3.71)$$

where σ is a scale parameter. For example, if f is the standard normal density, then the model described by (3.71) is just the standard regression model and σ is the standard deviation of the responses.

The log likelihood corresponding to this density function is

$$l(\beta, \sigma) = -n \log \sigma + \sum_{i=1}^n \log f[(Y_i - \mathbf{x}'_i \beta)/\sigma],$$

which, putting $\rho = -\log f$, we can write as

$$-\left\{ n \log \sigma + \sum_{i=1}^n \rho[(Y_i - \mathbf{x}'_i \beta)/\sigma] \right\}.$$

Thus, to estimate β and σ using maximum likelihood, we must minimize

$$n \log s + \sum_{i=1}^n \rho[e_i(\mathbf{b})/s] \quad (3.72)$$

as a function of \mathbf{b} and s . Differentiating leads to the estimating equations

$$\sum_{i=1}^n \psi[e_i(\mathbf{b})/s] \mathbf{x}_i = 0, \quad (3.73)$$

$$\sum_{i=1}^n \psi[e_i(\mathbf{b})/s] e_i(\mathbf{b}) = ns, \quad (3.74)$$

where $\psi = \rho'$.

EXAMPLE 3.11 Let $\rho(x) = \frac{1}{2}x^2$ so that $\psi(x) = x$. Then (3.73) reduces to the normal equations (3.4) with solution the least squares estimate (LSE) $\hat{\beta}$, and (3.74) gives the standard maximum likelihood estimate

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i(\hat{\beta})^2. \quad \square$$

EXAMPLE 3.12 Let $\rho(x) = |x|$. The corresponding estimates are values of s and \mathbf{b} that minimize

$$n \log s + \frac{1}{s} \sum_{i=1}^n |e_i(\mathbf{b})|. \quad (3.75)$$

Clearly, a value of \mathbf{b} minimizing (3.75) is also a value that minimizes

$$\sum_{i=1}^n |e_i(\mathbf{b})|$$

and is called the L_1 estimate. Note that there may be more than one value of \mathbf{b} that minimizes (3.75). There is a large literature devoted to L_1 estimation; see, for example, Bloomfield and Steiger [1983] and Dodge [1987]. Note that the L_1 estimate is the maximum likelihood estimator if f in (3.71) is the double exponential density proportional to $\exp(-|y|)$. An alternative term for the L_1 estimate is the LAD (Least Absolute Deviations) estimate. \square

If we have no particular density function f in mind, we can choose ρ to make the estimate robust by choosing a ρ for which $\psi = \rho'$ is bounded. We can generalize (3.73) and (3.74) to the estimating equations

$$\sum_{i=1}^n \psi[e_i(\mathbf{b})/s] \mathbf{x}_i = 0, \quad (3.76)$$

$$\sum_{i=1}^n \chi[e_i(\mathbf{b})/s] = 0, \quad (3.77)$$

where χ is also chosen to make the scale estimate robust. The resulting estimates are called *M-estimates*, since their definition is motivated by the maximum likelihood estimating equations (3.73) and (3.74). However, there is no requirement that ψ and χ be related to the density function f in (3.71).

EXAMPLE 3.13 (Huber “Proposal 2,” Huber [1981: p. 137]) Let

$$\psi(x) = \begin{cases} -k, & x < -k, \\ x, & -k \leq x \leq k, \\ k, & x > k, \end{cases} \quad (3.78)$$

where k is a constant to be chosen. The function (3.78) was derived by Huber using minimax asymptotic variance arguments and truncates the large residuals. The value of k is usually chosen to be 1.5, which gives a reasonable compromise between least squares (which is the choice giving greatest efficiency at the normal model) and L_1 estimation, which will give more protection from outliers. \square

An estimate $\hat{\theta}$ of a parameter θ is *consistent* if $\hat{\theta} \rightarrow \theta$ as the sample size increases. (Roughly speaking, consistency means that θ is the parameter

actually being estimated by $\hat{\theta}$.) It can be shown that a necessary condition for consistency when the parameters are estimated using (3.76) and (3.77) is

$$E[\psi(Z)] = 0, \quad (3.79)$$

and

$$E[\chi(Z)] = 0, \quad (3.80)$$

where Z has density function f . Equation (3.79) will be satisfied if f is symmetric about zero and if ψ is antisymmetric [i.e., $\psi(-z) = -\psi(z)$]. This will be the case if ρ is symmetric about zero. We note that the conditions (3.79) and (3.80) are only necessary conditions, so the estimates may be biased even if they are satisfied. However, Huber [1981: p. 171] observes that in practice the bias will be small, even if the conditions are not satisfied.

EXAMPLE 3.14 In Huber's Proposal 2, the function ψ is asymmetric, so condition (3.79) is satisfied. The scale parameter is estimated by taking $\chi(x) = \psi^2(x) - c$ for some constant c , which is chosen to make the estimate consistent when f is the normal density function. From (3.80), we require that $c = E[\psi(Z)^2]$, where Z is standard normal. \square

EXAMPLE 3.15 Another popular choice is to use $\chi(z) = \text{sign}(|z| - 1/c)$ for some constant c . Then (3.77) becomes

$$\sum_{i=1}^n \text{sign}(|e_i(\mathbf{b})| - s/c) = 0,$$

which has solution (see Exercises 3n, No. 1, at the end of this chapter)

$$s = c \text{median}_i |e_i(\mathbf{b})|.$$

This estimate is called the *median absolute deviation* (MAD); to make it consistent for the normal distribution, we require that $c^{-1} = \Phi^{-1}(3/4) = 0.6749$ (i.e., $c = 1.4326$). \square

Regression coefficients estimated using M-estimators are almost as efficient as least squares if the errors are normal, but are much more robust if the error distribution is long-tailed. Unfortunately, as we will see in Example 3.23 below, M-estimates of regression coefficients are just as vulnerable as least squares estimates to outliers in the explanatory variables.

3.13.2 Estimates Based on Robust Location and Scale Measures

As an alternative to M-estimation, we can replace the mean by a robust measure of location but retain the squared residual as a measure of size. This leads to the *least median of squares estimate* (LMS estimate), which minimizes

$$\text{median}_i e_i(\mathbf{b})^2.$$

The LMS estimator was popularized by Rousseeuw [1984] and is also discussed by Rousseeuw and Leroy [1987]. An alternative is to use the trimmed mean rather than the median, which results in the *least trimmed squares estimate* (LTS estimate), which minimizes

$$\sum_{i=1}^h e_{(i)}(\mathbf{b})^2, \quad (3.81)$$

where h is chosen to achieve a robust estimator and $e_{(1)}(\mathbf{b})^2 \leq \dots \leq e_{(n)}(\mathbf{b})^2$ are the ordered squared residuals. The amount of trimming has to be quite severe to make the estimate robust. The choice $h = [n/2] + 1$ (where $[x]$ is the greatest integer $\leq x$) is a popular choice, which amounts to trimming 50% of the residuals. The choice of h is discussed further in Section 3.13.3.

These estimates are very robust to outliers in both the errors and the explanatory variables but can be unstable in a different way. In certain circumstances, small changes in nonextreme points can make a very large change in the fitted regression. In Figure 3.2(a), the eight points lie on one of two lines, with the point marked A lying on both. If a line is fitted through the five collinear points, all five residuals corresponding to those points are zero. Since a majority of the residuals are zero, the median squared residual is also zero, so a line through these points minimizes the LMS criterion.

Now move the point marked B to be collinear with the remaining three points, resulting in Figure 3.2(b). This results in a new set of five collinear points. Using the same argument, this small change has resulted in the fitted LMS line now passing through the new set of collinear points. A small change in point B has resulted in a big change in the fit.

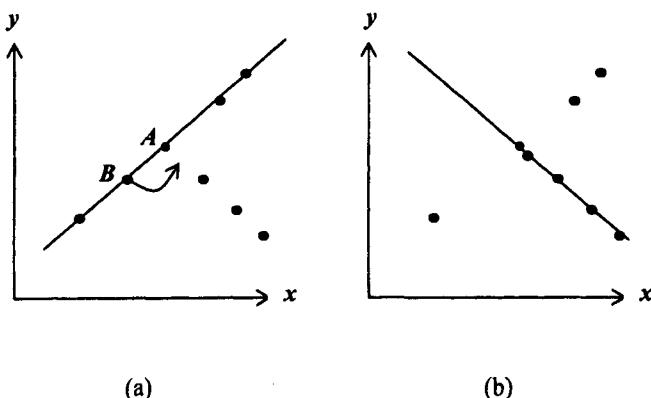


Fig. 3.2 Instability of the LMS estimator.

In addition, these estimates are very inefficient compared to least squares if the data are actually normally distributed. In this case, the asymptotic relative efficiency of LMS relative to the LSE is zero. (That is, the ratio of the variance of the LSE to that of the LMS estimate approaches zero as the sample size increases.) The equivalent for the LTS is 8% (Stromberg et al. [2000]). These poor efficiencies have motivated a search for methods that are at the same time robust and efficient. Before describing these, we need to discuss ways of quantifying robustness more precisely.

3.13.3 Measuring Robustness

We will discuss two measures of robustness. The first is the notion of breakdown point, which measures how well an estimate can resist gross corruption of a fraction of the data. The second is the influence curve, which gives information on how a single outlier affects the estimate.

Breakdown Point of an Estimate

Suppose that we select a fraction of the data. Can we cause an arbitrarily large change in the estimate by making a suitably large change in the selected data points?

Clearly, for some estimates the answer is yes; in the case of the sample mean we can make an arbitrarily large change in the mean by making a sufficiently large change in a single data point. On the other hand, for the sample median we can make large changes to almost 50% of the data without changing the median to the same extent.

Definition 3.2 *The breakdown point of an estimate is the smallest fraction of the data that can be changed by an arbitrarily large amount and still cause an arbitrarily large change in the estimate.*

Thus, the sample mean has a breakdown point of $1/n$ and the sample median a breakdown point of almost $1/2$. We note that a breakdown point of $1/2$ is the best possible, for if more than 50% of the sample is contaminated, it is impossible to distinguish between the “good” and “bad” observations, since the outliers are now typical of the sample.

Since the least squares estimate of a regression coefficient is a linear combination of the responses, it follows that an arbitrarily large change in a single response will cause an arbitrarily large change in at least one regression coefficient. Thus, the breakdown point of the least squares estimate is $1/n$.

Since the median has a very high breakdown point, and the median of the data Y_1, \dots, Y_n minimizes the least absolute deviation $\sum_i |Y_i - \theta|$ as a function of θ , it might be thought that the L_1 estimator of the regression coefficients would also have a high breakdown point. Unfortunately, this is not the case; in fact, the breakdown point of L_1 is the same as that of least squares. It can be shown, for example in Bloomfield and Steiger [1983: p. 7], that when the regression matrix X is of full rank, there is a value minimizing

$\sum_{i=1}^n |e_i(\mathbf{b})|$ for which at least p residuals are zero. Further, Bloomfield and Steiger [1983: p. 55] also prove that if one data point is arbitrarily far from the others, this data point must have a zero residual. It follows that by moving the data point an arbitrary amount, we must also be moving the fitted plane by an arbitrary amount, since the fitted plane passes through the extreme data point. Thus replacing a single point can cause an an arbitrarily large change in the regression plane, and the breakdown point of the L_1 estimate is $1/n$. The same is true of M-estimates (Rousseeuw and Leroy [1987: p. 149]).

We saw above that the LMS and LTS estimates were inefficient compared to M-estimates. They compensate for this by having breakdown points of almost $1/2$, the best possible. If we make a small change in the definition of the LMS, its breakdown point can be slightly improved. Let

$$h = [n/2] + [(p+1)/2], \quad (3.82)$$

where $[x]$ denotes the largest integer $\leq x$. If we redefine the LMS estimator as the value of \mathbf{b} that minimizes $e_{(h)}(\mathbf{b})^2$, rather than the median squared residual, the LMS breakdown point becomes $([(n-p)/2] + 1)/n$. If h is given by (3.82), then the LTS estimate which minimizes

$$\sum_{i=1}^h e_{(i)}(\mathbf{b})^2$$

also has breakdown point $([(n-p)/2] + 1)/n$, slightly higher than with the choice $h = [n/2] + 1$. These results are discussed in Rousseeuw and Leroy [1987: pp. 124, 132].

Influence Curves

Suppose that F is a k -dimensional distribution function (d.f.), and $\boldsymbol{\theta}$ is a population parameter that depends on F , so that we may write $\boldsymbol{\theta} = T(F)$. We call T a *statistical functional*, since it is a function of a function.

EXAMPLE 3.16 Perhaps the simplest example of a statistical functional is the mean $E_F[X]$ of a random variable X , where the subscript F denotes expectation with respect to the d.f. F . In terms of integrals,

$$\begin{aligned} T(F) &= E_F[X] \\ &= \int x dF(x). \end{aligned} \quad (3.83) \quad \square$$

EXAMPLE 3.17 If \mathbf{Z} is a random k -vector with distribution function F , then the matrix $E_F[\mathbf{Z}\mathbf{Z}']$ is a statistical functional, also given by the k -dimensional integral

$$T(F) = \int \mathbf{z}\mathbf{z}' dF(\mathbf{z}). \quad (3.84) \quad \square$$

Definition 3.3 If $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ are independent and identically distributed random vectors each with distribution function F , the empirical distribution function (e.d.f.) \hat{F}_n is the d.f. which places mass n^{-1} at each of the n points \mathbf{Z}_i , $i = 1, \dots, n$.

Integration with respect to the e.d.f. is just averaging; if h is a function, then

$$\int h(\mathbf{z}) d\hat{F}_n(\mathbf{z}) = n^{-1} \sum_{i=1}^n h(\mathbf{Z}_i).$$

Many statistics used to estimate parameters $T(F)$ are *plug-in estimates* of the form $T(\hat{F}_n)$, where \hat{F}_n is the e.d.f. based on a random sample from F .

EXAMPLE 3.18 (Vector sample mean) Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be a random sample from some multivariate distribution having d.f. F . The plug-in estimate of

$$T(F) = \int \mathbf{z} dF(\mathbf{z})$$

is

$$\begin{aligned} T(\hat{F}_n) &= \int \mathbf{z} d\hat{F}_n(\mathbf{z}) \\ &= n^{-1} \sum_{i=1}^n \mathbf{Z}_i, \end{aligned}$$

the sample mean. □

EXAMPLE 3.19 The plug-in estimate of (3.84) is

$$\begin{aligned} T(\hat{F}_n) &= \int \mathbf{z} \mathbf{z}' d\hat{F}_n(\mathbf{z}) \\ &= n^{-1} \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i'. \end{aligned} \tag{3.85} \quad \square$$

Consider a regression with a response variable Y and explanatory variables x_1, \dots, x_{p-1} . When studying the statistical functionals that arise in regression, it is usual to assume that the explanatory variables are random. We regard the regression data $(\mathbf{x}_i, Y_i), i = 1, \dots, n$, as n identically and independently distributed random $(p+1)$ -vectors, distributed as (\mathbf{x}, Y) , having a joint distribution function F , say. Thus, in contrast with earlier sections, we think of the vectors \mathbf{x}_i as being random and having initial element 1 if the regression contains a constant term. As before, we write \mathbf{X} for the (random) matrix with i th row \mathbf{x}_i' .

We shall assume that the conditional distribution of Y given \mathbf{x} has density function $g[(y - \beta' \mathbf{x})/\sigma]$, where g is a known density, for example the standard

normal. For simplicity, we will sometimes assume that the scale parameter σ is known. In this case, we can absorb σ into g and write the conditional density as $g(y - \beta' \mathbf{x})$.

EXAMPLE 3.20 (Least squares) Consider the functional

$$\mathbf{T}(F) = \{E_F[\mathbf{x}\mathbf{x}']\}^{-1} \dot{E}_F[\mathbf{x}Y].$$

The plug-in estimator of \mathbf{T} is

$$\begin{aligned} \mathbf{T}(\hat{F}_n) &= \left(n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(n^{-1} \sum_{i=1}^n \mathbf{x}_i Y_i \right) \\ &= (\mathbf{X}\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}. \end{aligned} \quad \square \quad (3.86)$$

To assess the robustness of a plug-in estimator $T(\hat{F}_n)$, we could study how it responds to a small change in a single data point. An alternative, which we adopt below, is to examine the population version: We look at the effect of small changes in F on the functional $T(F)$. This allows us to examine the sensitivity of T more generally, without reference to a particular set of data.

Suppose that F is a distribution function. We can model a small change in F at a fixed (i.e., nonrandom) value $\mathbf{z}_0 = (\mathbf{x}'_0, y_0)'$ by considering the mixture of distributions $F_t = (1-t)F + t\delta_{\mathbf{z}_0}$, where $\delta_{\mathbf{z}_0}$ is the distribution function of the constant \mathbf{z}_0 , and $t > 0$ is close to zero. The sensitivity of T can be measured by the rate at which $T(F_t)$ changes for small values of t .

Definition 3.4 *The influence curve (IC) of a statistical functional T is the derivative with respect to t of $T(F_t)$ evaluated at $t = 0$, and is a measure of the rate at which T responds to a small amount of contamination at \mathbf{z}_0 .*

We note that the influence curve depends on both F and \mathbf{z}_0 , and we use the notation

$$\text{IC}(F, \mathbf{z}_0) = \frac{dT(F_t)}{dt} \Big|_{t=0}$$

to emphasize this. Cook and Weisberg [1982: Chapter 3], Hampel et al. [1986] and Davison and Hinkley [1997] all have more information on influence curves.

EXAMPLE 3.21 (IC of the mean) Let T be the mean functional defined in Example 3.16. Then

$$\begin{aligned} T(F_t) &= \int x dF_t(x) \\ &= (1-t) \int x dF(x) + t \int x d\delta_{x_0}(x) \\ &= (1-t)T(F) + tx_0, \end{aligned}$$

so that

$$\frac{T(F_t) - T(F)}{t} = x_0 - T(F)$$

and

$$\text{IC}(x_0, F) = x_0 - T(F).$$

This is unbounded in x_0 , suggesting that a small amount of contamination can cause an arbitrarily large change. In other words, the mean is highly nonrobust. \square

EXAMPLE 3.22 (IC for the LSE) Let \mathbf{T} be the LSE functional defined in Example 3.20. Write $\Sigma_F = E_F[\mathbf{x}\mathbf{x}']$ and $\gamma_F = E_F[\mathbf{x}Y]$. Then

$$\mathbf{T}(F_t) = \{\Sigma_{F_t}\}^{-1} \gamma_{F_t}. \quad (3.87)$$

We have

$$\begin{aligned} \Sigma_{F_t} &= E_{F_t}[\mathbf{x}\mathbf{x}'] \\ &= (1-t)E_F[\mathbf{x}\mathbf{x}'] + t\mathbf{x}_0\mathbf{x}'_0 \\ &= (1-t)\Sigma_F + t\mathbf{x}_0\mathbf{x}'_0 \\ &= (1-t)\{\Sigma_F + t'\mathbf{x}_0\mathbf{x}'_0\}, \end{aligned}$$

where $t' = t/(1-t)$. By A.9.1 we get

$$\Sigma_{F_t}^{-1} = (1-t)^{-1} \left[\Sigma_F^{-1} - t' \frac{\Sigma_F^{-1} \mathbf{x}_0 \mathbf{x}'_0 \Sigma_F^{-1}}{1+o(t)} \right]. \quad (3.88)$$

Similarly,

$$\gamma_{F_t} = (1-t)\gamma_F + t\mathbf{x}_0 y_0. \quad (3.89)$$

Substituting (3.88) and (3.89) in (3.87) yields

$$\mathbf{T}(F_t) = \mathbf{T}(F) + t' \Sigma_F^{-1} \mathbf{x}_0 y_0 - t' \Sigma_F^{-1} \mathbf{x}_0 \mathbf{x}'_0 \Sigma_F^{-1} + o(t),$$

so that

$$\frac{\mathbf{T}(F_t) - \mathbf{T}(F)}{t} = \Sigma_F^{-1} \mathbf{x}_0 y_0 - \Sigma_F^{-1} \mathbf{x}_0 \mathbf{x}'_0 \mathbf{T}(F) + o(1).$$

Letting $t \rightarrow 0$, we get

$$\text{IC}(\mathbf{z}_0, F) = \Sigma_F^{-1} \mathbf{x}_0 [y_0 - \mathbf{x}'_0 \mathbf{T}(F)].$$

We see that this is unbounded in both \mathbf{x}_0 and y_0 , indicating that the LSE is not robust. \square

The situation is somewhat better for M-estimates.

EXAMPLE 3.23 (IC for M-estimates) For simplicity we will assume that the scale parameter σ is known. Consider the functional T defined implicitly by the equation

$$E_F (\psi\{[Y - \mathbf{x}'\mathbf{T}(F)]/\sigma\}\mathbf{x}) = 0. \quad (3.90)$$

The plug-in version is $\mathbf{T}(\hat{F}_n)$ is the solution of

$$\frac{1}{n} \sum_{i=1}^n \psi\{[Y_i - \mathbf{x}'_i \mathbf{T}(\hat{F}_n)]/\sigma\} \mathbf{x}_i = 0,$$

which is of the form (3.76). Thus, the functional \mathbf{T} defined by (3.90) is the M-estimation functional.

To derive its influence curve, we substitute F_t for F in (3.90). This yields

$$(1-t)E_F[\psi\{[Y - \mathbf{x}'\mathbf{T}(F_t)]/\sigma\}\mathbf{x}] + t\psi\{[y_0 - \mathbf{x}'_0 \mathbf{T}(F_t)]/\sigma\} \mathbf{x}_0 = \mathbf{0}. \quad (3.91)$$

Let $\dot{\mathbf{T}}_t = d\mathbf{T}(F_t)/dt$, $\varepsilon_t = [Y - \mathbf{x}'\mathbf{T}(F_t)]/\sigma$ and $\eta_t = [y_0 - \mathbf{x}'_0 \mathbf{T}(F_t)]/\sigma$, and note the derivatives

$$\frac{d\psi(\eta_t)}{dt} = -\psi'(\eta_t)\mathbf{x}'_0 \dot{\mathbf{T}}_t / \sigma$$

and

$$\begin{aligned} \frac{dE_F[\psi(\varepsilon_t)\mathbf{x}]}{dt} &= E_F \left[\frac{d\psi(\varepsilon_t)}{dt} \mathbf{x} \right] \\ &= -E_F [\psi'(\varepsilon_t)\mathbf{x}\mathbf{x}'] \dot{\mathbf{T}}_t / \sigma. \end{aligned}$$

Now differentiate both sides of (3.91). We obtain

$$(1-t) \frac{dE_F[d\psi(\varepsilon_t)\mathbf{x}]}{dt} - E_F[\psi(\varepsilon_t)\mathbf{x}] + t \frac{d\psi(\eta_t)}{dt} \mathbf{x}_0 + \psi(\eta_t)\mathbf{x}_0 = \mathbf{0},$$

which, using the derivatives above, gives

$$-(1-t)E_F[\psi'(\varepsilon_t)\mathbf{x}\mathbf{x}'] \dot{\mathbf{T}}_t / \sigma - E_F[\psi(\varepsilon_t)\mathbf{x}] - t\psi'(\eta_t)\mathbf{x}_0 \mathbf{x}_0' \dot{\mathbf{T}}_t / \sigma + \psi(\eta_t)\mathbf{x}_0 = \mathbf{0}.$$

Now set $t = 0$. Noting that $F_t = F$ when $t = 0$, and using (3.90), we get

$$\begin{aligned} E_F[\psi(\varepsilon_0)] &= E_F[\psi\{[Y - \mathbf{x}'\mathbf{T}(F)]/\sigma\}] \\ &= 0, \end{aligned}$$

and from the definition of the IC, $\dot{\mathbf{T}}_0 = \text{IC}(\mathbf{z}_0, F)$. Thus,

$$-E_F[\psi'\{[Y - \mathbf{x}'\mathbf{T}(F)]/\sigma\}\mathbf{x}\mathbf{x}'] \text{IC}(\mathbf{z}_0, F) / \sigma + \psi\{[y_0 - \mathbf{x}'_0 \mathbf{T}(F)]/\sigma\} \mathbf{x}_0 = \mathbf{0},$$

so finally,

$$\text{IC}(\mathbf{z}_0, F) = \sigma \psi\{[y_0 - \mathbf{x}'_0 \mathbf{T}(F)]/\sigma\} \mathbf{M}^{-1} \mathbf{x}_0, \quad (3.92)$$

where $\mathbf{M} = E_F[\psi\{(Y - \mathbf{x}'\mathbf{T}(F))/\sigma\}\mathbf{x}\mathbf{x}']$. Thus, assuming that ψ is bounded, the influence curve is bounded in y_0 , suggesting that M-estimates are robust

with respect to outliers in the errors. However, the IC is not bounded in \mathbf{x}_0 , so M-estimates are not robust with respect to high-leverage points (i.e., points with outliers in the explanatory variables; see Section 9.4). \square

The robust estimates discussed so far are not entirely satisfactory, since the high breakdown estimators LMS and LTS have poor efficiency, and the efficient M-estimators are not robust against outliers in the explanatory variables and have breakdown points of zero. Next, we describe some other robust estimates that have high breakdown points but much greater efficiency than LMS or LTS.

3.13.4 Other Robust Estimates

Bounded Influence Estimators

As we saw above, M-estimators have influence curves that are unbounded in \mathbf{x}_0 and so are not robust with respect to high-leverage points. However, it is possible to modify the estimating equation (3.73) so that the resulting IC is bounded in \mathbf{x}_0 . Consider an estimating equation of the form

$$\sum_{i=1}^n w(\mathbf{x}_i) \psi\{e_i(\mathbf{b})/[\sigma w(\mathbf{x}_i)]\} \mathbf{x}_i = 0, \quad (3.93)$$

where, for simplicity, we will assume that the scale parameter σ is known. This modified estimating equation was first suggested by Handschin et al. [1975], and the weights are known as *Schweppe weights*. It can be shown (see Hampel et al. [1986: p. 316]) that the IC for this estimate is

$$\text{IC}(\mathbf{z}_0, F) = \sigma w(\mathbf{x}_0) \psi\{(y_0 - \mathbf{x}'_0 \mathbf{T}(F))/(\sigma w(\mathbf{x}_0))\} \mathbf{M}^{-1} \mathbf{x}_0,$$

where \mathbf{M} is a matrix [different from the \mathbf{M} appearing in (3.92)] not depending on \mathbf{z}_0 . The weight function w is chosen to make the IC bounded, and the resulting estimates are called *bounded influence estimates* or *generalized M-estimates* (GM-estimates).

To make the IC bounded, the weights are chosen to downweight cases that are high-leverage points. However, including a high-leverage point that is not an outlier (in the sense of not having an extreme error) increases the efficiency of the estimate. This is the reason for including the weight function w in the denominator in the expression $e_i(\mathbf{b})/[\sigma w(\mathbf{x}_i)]$, so that the effect of a small residual at a high-leverage point will be magnified. An earlier version of (3.93), due to Mallows [1975], does not include the weight $w(\mathbf{x}_i)$ in the denominator and seems to be less efficient (Hill [1977]).

The weights can be chosen to minimize the asymptotic variance of the estimates, subject to the influence curve being bounded by some fixed amount. This leads to weights of the form $w(\mathbf{x}) = \|\mathbf{Ax}\|^{-1}$ for some matrix \mathbf{A} . More details may be found in Ronchetti [1987] and Hampel et al. [1986: p. 316].

Krasker and Welsch [1982] give additional references and discuss some other proposals for choosing the weights.

The breakdown point of these estimators is better than for M-estimators, but cannot exceed $1/p$ (Hampel et al. [1986: p. 328]). This can be low for problems with more than a few explanatory variables. To improve the breakdown point of GM-estimators, we could combine them in some way with high breakdown estimators, in the hope that the combined estimate will inherit the desirable properties of both.

The estimating equation (3.93) that defines the GM-estimate is usually solved iteratively by either the Newton–Raphson method or Fisher scoring (A.14), using some other estimate as a starting value. (This procedure is discussed in more detail in Section 11.12.2.)

A simple way of combining a high breakdown estimate with a GM-estimate is to use the high breakdown estimate as a starting value and then perform a single Newton–Raphson or Fisher scoring iteration using the GM iteration scheme discussed in Section 11.12.2; the resulting estimate is called a *one-step GM-estimate*. This idea has been suggested informally by several authors: for example, Hampel et al. [1986: p. 328] and Ronchetti [1987].

Simpson et al. [1992] have carried out a formal investigation of the properties of the one-step GM-estimate. They used the Mallows form of the estimating equation (3.93), with weights $w(\mathbf{x}_i)$ based on a robust Mahalanobis distance. The Mallows weights are given by

$$w(\mathbf{x}_i) = \min \left[1, \left\{ \frac{b}{(\mathbf{x}_i - \mathbf{m})' \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{m})} \right\}^{\alpha/2} \right], \quad (3.94)$$

where b and α are tuning constants, \mathbf{m} and \mathbf{C} are robust measures of the location and dispersion of the explanatory variables, and the \mathbf{x}_i 's are to be interpreted in the “reduced” sense, without the initial 1. Thus, the denominator in the weight function is a robust Mahalanobis distance, measuring the distance of \mathbf{x}_i from a typical \mathbf{x} . Suitable estimates \mathbf{m} and \mathbf{C} are furnished by the minimum volume ellipsoid described in Section 10.6.2 and in Rousseeuw and Leroy [1987: p. 258].

If the robust distance used to define the weights and the initial estimate of the regression coefficients both have a breakdown point of almost 50%, then the one-step estimator will also inherit this breakdown point. Thus, if LMS is used as the initial estimator, and the minimum volume ellipsoid (see Section 10.6.2) is used to calculate the weights, the breakdown point of the one-step estimator will be almost 50%. The one-step estimator also inherits the bounded-influence property of the GM-estimator. Coakley and Hettmansperger [1993] suggest that efficiency can be improved by using the Schweppe form of the estimating equation and starting with the LTS estimate rather than the LMS.

S-Estimators

We can think of the “average size” of the residuals as a measure of their dispersion, so we can consider more general regression estimators based on some dispersion or scale estimator $s(e_1, \dots, e_n)$. This leads to minimizing

$$D(\mathbf{b}) = s[e_1(\mathbf{b}), \dots, e_n(\mathbf{b})], \quad (3.95)$$

where s is a estimator of scale. The scale parameter σ is estimated by the minimum value of (3.95).

EXAMPLE 3.24 If we use the standard deviation as an estimate of scale, (3.95) reduces to

$$\sum_{i=1}^n [e_i(\mathbf{b}) - \overline{e_i(\mathbf{b})}]^2 = \sum_{i=1}^n [Y_i - \bar{Y} - b_1(x_{i1} - \bar{x}_1) - \dots - b_{p-1}(x_{ip-1} - \bar{x}_{p-1})]^2,$$

which is the residual sum of squares. The estimates minimizing this are the least squares estimates. Thus, in the case of a regression with a constant term, taking the scale estimate s to be the standard deviation is equivalent to estimating the regression coefficients by least squares. \square

EXAMPLE 3.25 Using the MAD as an estimate of scale leads to minimizing $\text{median}_i |e_i(\mathbf{b})|$, which is equivalent to minimizing $\text{median}_i |e_i(\mathbf{b})|^2$. Thus, using the estimate based on the MAD is equivalent to LMS. \square

Rousseeuw and Yohai [1984] considered using robust scale estimators $s = s(e_1, \dots, e_n)$ defined by the equation

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{e_i}{s}\right) = K,$$

where $K = E[\rho(Z)]$ for a standard normal Z , and the function ρ is symmetric and positive. They also assume that ρ is strictly increasing on $[0, c]$ for some value c and is constant on (c, ∞) . Estimators defined in this way are called *S-estimators*.

Rousseeuw and Yohai show that the breakdown point of such an estimator can be made close to 50% by a suitable choice of the function ρ . The *biweight function*, defined by

$$\rho(x) = \begin{cases} x^2/2 - x^4/(2c^2) + x^6/(6c^4), & |x| \leq c, \\ c^2/6, & |x| > c, \end{cases}$$

is a popular choice. If the constant c satisfies $\rho(c) = 2E[\rho(Z)]$, where Z is standard normal, then Rousseeuw and Yohai prove that the breakdown point of the estimator is $([n/2] - p + 2)/n$, or close to 50%. For the biweight estimator, this implies that $c = 1.547$. The efficiency at the normal distribution is about 29%.

R-Estimators

Another class of estimators based on a measure of dispersion are the *R-estimators*, where the dispersion measure is defined using ranks. Let $a_n(i)$, $i = 1, \dots, n$, be a set of scores, given by

$$a_n(i) = h[i/(n+1)], \quad (3.96)$$

where h is a function defined on $[0, 1]$. Examples from nonparametric statistics include the Wilcoxon scores, $[h(u) = u - 0.5]$, the van der Waerden scores $[h(u) = \Phi^{-1}(u)]$ and median scores $[h(u) = \text{sign}(u - 0.5)]$. All these scores satisfy $\sum_{i=1}^n a_n(i) = 0$.

Jaeckel [1972] defined a dispersion measure by

$$s(e_1, \dots, e_n) = \sum_{i=1}^n a_n(R_i) e_i, \quad (3.97)$$

where R_i is the rank of e_i (i.e., its position in the sequence $\{e_1, \dots, e_n\}$). Since the scores sum to zero, the dispersion measure will be close to zero if the e_i 's are similar. For a fixed vector \mathbf{b} , let R_i be the rank of $e_i(\mathbf{b})$. Jaeckel proposed as a robust estimator the vector that minimizes $s[e_1(\mathbf{b}), \dots, e_n(\mathbf{b})]$.

Note that since the scores satisfy $\sum_{i=1}^n a_n(i) = 0$, the measure s has the property

$$s(e_1 + c, \dots, e_n + c) = s(e_1, \dots, e_n).$$

Thus, for any vector \mathbf{b} , if the regression contains a constant term, the quantity $s[e_1(\mathbf{b}), \dots, e_n(\mathbf{b})]$ does not depend on the initial element b_0 of \mathbf{b} . If we write $\mathbf{b} = (b_0, \mathbf{b}'_1)'$, then $s[e_1(\mathbf{b}), \dots, e_n(\mathbf{b})]$ is a function of \mathbf{b}'_1 alone, which we can denote by $D(\mathbf{b}'_1)$. It follows that we cannot obtain an estimate of β_0 by minimizing $D(\mathbf{b}'_1)$; this must be obtained separately, by using a robust location measure such as the median applied to the residuals $e_i(\mathbf{b})$, where $\mathbf{b} = (0, \tilde{\mathbf{b}}'_1)'$, $\tilde{\mathbf{b}}'_1$ being the minimizer of $D(\mathbf{b}'_1)$.

The estimate defined in this way has properties similar to those of an M-estimator: For the Wilcoxon scores it has an influence function that is bounded in y_0 but not in x_0 , has a breakdown point of $1/n$, and has high efficiency at the normal distribution. These facts are proved in Jaeckel [1972], Jureckova [1971], and Naranjo and Hettmansperger [1994].

The estimate can be modified to have a better breakdown point by modifying the scores and basing the ranks on the absolute values of the residuals, or equivalently, on the ordered absolute residuals, which satisfy

$$|e_{(1)}(\mathbf{b})| \leq \dots \leq |e_{(n)}(\mathbf{b})|.$$

Consider an estimate based on minimizing

$$D(\mathbf{b}) = \sum_{i=1}^n a_n(i) |e_{(i)}(\mathbf{b})|, \quad (3.98)$$

where the scores are now of the form $a_n(i) = h^+(i/(n+1))$, where h^+ is a nonnegative function defined on $[0, 1]$ and is zero on $[\alpha, 1]$ for $0 < \alpha \leq 1$. Then Hössjer [1994] shows that for suitable choice of h^+ , the breakdown point of the estimate approaches $\min(\alpha, 1 - \alpha)$ as the sample size increases. The efficiency decreases as the breakdown point increases. For a breakdown point of almost 50%, the efficiency is about 7% at the normal model, similar to LTS.

The efficiency can be improved while retaining the high breakdown property by considering estimates based on differences of residuals. Sievers [1983], Naranjo and Hettmansperger [1994], and Chang et al. [1999] considered estimates of the regression coefficients (excluding the constant term) based on minimizing the criterion

$$D(\mathbf{b}_1) = \sum_{1 \leq i < j \leq n} w_{ij} |e_i(\mathbf{b}) - e_j(\mathbf{b})|, \quad (3.99)$$

which, like Jaeckel's estimator, does not depend on b_0 . The weights w_{ij} can be chosen to achieve a high breakdown point, bounded influence, and high efficiency. Suppose that \mathbf{b} and $\tilde{\sigma}$ are preliminary 50% breakdown estimates of β and σ . For example, we could use LMS to estimate β , and estimate σ using the MAD. Chang et al. [1999] show that if the weights are defined by

$$w_{ij} = \max \left\{ 1, \left| \frac{cw(\mathbf{x}_i)w(\mathbf{x}_j)}{[e_i(\tilde{\mathbf{b}})/\tilde{\sigma}][e_j(\tilde{\mathbf{b}})/\tilde{\sigma}]} \right| \right\}, \quad (3.100)$$

then the efficiency can be raised to about 67% while retaining a 50% breakdown point. In (3.100), the weights $w(\mathbf{x}_i)$ are the Mallows weights defined in (3.94), and c is a tuning constant. If $w_{ij} = 1$ for all $i \leq j$, then the estimate reduces to Jaeckel's estimate with Wilcoxon scores (see Exercises 3n, No. 2).

Similar efficiencies can be achieved using a modified form of S-estimate which is also based on differences of residuals. Croux et al. [1994] define a scale estimate $s = s(e_1, \dots, e_n)$ as the solution to the equation

$$\sum_{1 \leq i < j \leq n} \rho \left(\frac{e_i(\mathbf{b}) - e_j(\mathbf{b})}{s} \right) = \binom{n}{2} - \binom{h}{2} + 1, \quad (3.101)$$

where $h = [(n+p+1)/2]$. Then the estimate based on minimizing $s(\mathbf{b}_1) = s[e_1(\mathbf{b}), \dots, e_n(\mathbf{b})]$ is called a *generalized S-estimate*. Note that again this criterion does not depend on b_0 .

Defining

$$\rho(x) = \begin{cases} 1, & |x| \geq 1, \\ 0, & |x| < 1, \end{cases} \quad (3.102)$$

gives an estimate called the *least quartile difference estimate* (LQD estimate), since (see Exercises 3n, No. 3) the resulting s is approximately the lower quartile of all the $\binom{n}{2}$ differences $|e_i(\mathbf{b}) - e_j(\mathbf{b})|$. Croux et al. [1994] show

that the LQD estimator has a breakdown point of almost 50% and roughly 67% efficiency. It does not have a bounded influence function.

A similar estimate, based on a trimmed mean of squared differences, is the *least trimmed difference estimate* (LTD estimate), which minimises the sum of the first $\binom{h}{2}$ ordered squared differences. This estimate, introduced in Stromberg et al. [2000], has properties similar to those of the LQD.

EXERCISES 3n

- Let $\chi(z) = \text{sign}(|z| - 1/c)$ for some constant c . Show that the solution of (3.77) is the MAD estimate

$$s = c \text{ median}_i |e_i(\mathbf{b})|.$$

- Show that if we put $w_{ij} = 1$ in (3.99), we get Jaeckel's estimator defined by (3.97) with Wilcoxon weights.
- Show that if s is the solution of (3.101) with ρ given by (3.102), then the resulting s is approximately the lower quartile of the $\binom{n}{2}$ differences $|e_i(\mathbf{b}) - e_j(\mathbf{b})|$.

MISCELLANEOUS EXERCISES 3

- Let $Y_i = a_i\beta_1 + b_i\beta_2 + \varepsilon_i$ ($i = 1, 2, \dots, n$), where the a_i, b_i are known and the ε_i are independently and identically distributed as $N(0, \sigma^2)$. Find a necessary and sufficient condition for the least squares estimates of β_1 and β_2 to be independent.
- Let $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $E[\boldsymbol{\varepsilon}] = \mathbf{0}$. Prove that the value of $\boldsymbol{\theta}$ that minimizes $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ subject to $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, where \mathbf{A} is a known $q \times n$ matrix of rank q , is

$$\hat{\boldsymbol{\theta}} = (\mathbf{I}_n - \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A})\mathbf{Y}.$$

- Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$, and \mathbf{X} is $n \times p$ of rank p . If \mathbf{X} and $\boldsymbol{\beta}$ are partitioned in the form

$$\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}_1, \mathbf{X}_2) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

prove that the least squares estimate $\hat{\boldsymbol{\beta}}_2$ of $\boldsymbol{\beta}_2$ is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= [\mathbf{X}'_2 \mathbf{X}_2 - \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2]^{-1} \\ &\quad \times [\mathbf{X}'_2 \mathbf{Y} - \mathbf{X}'_2 \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{Y}]. \end{aligned}$$

Find $\text{Var}[\hat{\beta}_2]$.

4. Suppose that $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$. Prove that $\mathbf{a}'\mathbf{Y}$ is the linear unbiased estimate of $E[\mathbf{a}'\mathbf{Y}]$ with minimum variance if and only if $\text{cov}[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}] = 0$ for all \mathbf{b} such that $E[\mathbf{b}'\mathbf{Y}] = 0$ (i.e., $\mathbf{b}'\mathbf{X} = 0'$).
 (Rao [1973])

5. If \mathbf{X} has full rank and $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, prove that

$$\sum_{i=1}^n \text{var}[\hat{Y}_i] = \sigma^2 p.$$

6. Estimate the weights β_i ($i = 1, 2, 3, 4$) of four objects from the following weighing data (see Exercises 3e, No. 5, at the end of Section 3.6 for notation):

x_1	x_2	x_3	x_4	Weight (Y)
1	1	1	1	20.2
1	-1	1	-1	8.0
1	1	-1	-1	9.7
1	-1	-1	1	1.9

7. Three parcels are weighed at a post office singly, in pairs, and all together, giving weights Y_{ijk} ($i, j, k = 0, 1$), the suffix 1 denoting the presence of a particular parcel and the suffix 0 denoting its absence. Find the least squares estimates of the weights.

(Rahman [1967])

8. An experimenter wishes to estimate the density d of a liquid by weighing known volumes of the liquid. Let Y_i be the weight for volume x_i ($i = 1, 2, \dots, n$) and let $E[Y_i] = dx_i$ and $\text{var}[Y_i] = \sigma^2 f(x_i)$. Find the least squares estimate of d for the following cases:

$$(a) f(x_i) \equiv 1. \quad (b) f(x_i) = x_i. \quad (c) f(x_i) = x_i^2.$$

9. Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, 3$), where $E[\varepsilon] = 0$, $\text{Var}[\varepsilon] = \sigma^2 V$ with

$$V = \begin{pmatrix} 1 & \rho a & \rho \\ \rho a & a^2 & \rho a \\ \rho & \rho a & 1 \end{pmatrix}, \quad (a, \rho \text{ unknown}) \quad 0 < \rho < 1$$

and $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$. Show that the generalized least squares estimates of β_0 and β_1 are

$$\begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} = \begin{pmatrix} r^{-1} \{(a^2 - a\rho)Y_1 + (1 - 2a\rho + \rho)Y_2 + (a^2 - a\rho)Y_3\} \\ -\frac{1}{2}Y_1 + \frac{1}{2}Y_3 \end{pmatrix}$$

where $r = 1 + \rho + 2a^2 - 4a\rho$. Also prove the following:

- (a) If $a = 1$, then the fitted regression $Y_i^* = \beta_0^* + \beta_1^* x_i$ cannot lie wholly above or below the values of Y_i (i.e., the $Y_i - Y_i^*$ cannot all have the same sign).
- (b) If $0 < a < \rho < 1$, then the fitted regression line can lie wholly above or below the observations.

(Canner [1969])

10. If \mathbf{X} is not of full rank, show that any solution $\boldsymbol{\beta}$ of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ minimizes $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

11. Let

$$\begin{aligned} Y_1 &= \theta_1 + \theta_2 + \varepsilon_1, \\ Y_2 &= \theta_1 - 2\theta_2 + \varepsilon_2, \end{aligned}$$

and

$$Y_3 = 2\theta_1 - \theta_2 + \varepsilon_3,$$

where $E[\varepsilon_i] = 0$ ($i = 1, 2, 3$). Find the least squares estimates of θ_1 and θ_2 . If the equations above are augmented to

$$\begin{aligned} Y_1 &= \theta_1 + \theta_2 + \theta_3 + \varepsilon_1, \\ Y_2 &= \theta_1 - 2\theta_2 + \theta_3 + \varepsilon_2, \\ Y_3 &= 2\theta_1 - \theta_2 + \theta_3 + \varepsilon_3, \end{aligned}$$

find the least squares estimate of θ_3 .

12. Given the usual full-rank regression model, prove that the random variables \bar{Y} and $\sum_i(Y_i - \hat{Y}_i)^2$ are statistically independent.
13. Let $Y_i = \beta x_i + u_i$, $x_i > 0$ ($i = 1, 2, \dots, n$), where $u_i = \rho u_{i-1} + \varepsilon_i$ and the ε_i are independently distributed as $N(0, \sigma^2)$. If $\hat{\beta}$ is the ordinary least squares estimate of β , prove that $\text{var}[\hat{\beta}]$ is inflated when $\rho > 0$.
14. Suppose that $E[Y_t] = \beta_0 + \beta_1 \cos(2\pi k_1 t/n) + \beta_2 \sin(2\pi k_2 t/n)$, where $t = 1, 2, \dots, n$, and k_1 and k_2 are positive integers. Find the least squares estimates of β_0 , β_1 , and β_2 .
15. Suppose that $E[Y_i] = \alpha_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)$, $i = 1, 2, \dots, n$. Show that the least squares estimates of α_0 , β_1 , and β_2 can be obtained by the following two-stage procedure:
- (i) Fit the model $E[Y_i] = \alpha_0 + \beta_1(x_{i1} - \bar{x}_1)$.
 - (ii) Regress the residuals from (i) on $(x_{i2} - \bar{x}_2)$.

4

Hypothesis Testing

4.1 INTRODUCTION

In this chapter we develop a procedure for testing a linear hypothesis for a linear regression model. To motivate the general theory given below, we consider several examples.

EXAMPLE 4.1 From (1.1) we have the model

$$\log F = \log c - \beta \log d,$$

representing the force of gravity between two bodies distance d apart. Setting $Y = \log F$ and $x = -\log d$, we have the usual linear model $Y = \beta_0 + \beta_1 x + \varepsilon$, where an error term ε has been added to allow for uncontrolled fluctuations in the experiment. The inverse square law states that $\beta = 2$, and we can test this by taking n pairs of observations (x_i, y_i) and seeing if the least squares line has a slope close enough to 2, given the variability in the data. \square

Testing whether a particular β in a regression model takes a value other than zero is not common and generally arises in models constructed from some underlying theory rather than from empirical considerations.

EXAMPLE 4.2 From (1.2) we have the following model for comparing two straight lines:

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where $\beta_0 = \alpha_1$, $\beta_1 = \gamma_1$, $\beta_2 = \alpha_2 - \alpha_1$, and $\beta_3 = \gamma_2 - \gamma_1$. To test whether the two lines have the same slope, we test $\beta_3 = 0$; while to test whether the two lines are identical, we test $\beta_2 = \beta_3 = 0$. Here we are interested in testing whether certain prespecified β_i are zero. \square

EXAMPLE 4.3 Suppose that we have the general linear model

$$G : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i,$$

or $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. When p is large we will usually be interested in considering whether we can set some of the β_i equal to zero. This is the problem of model selection discussed in Chapter 12. If we test the hypothesis $\beta_r = \beta_{r+1} = \cdots = \beta_{p-1} = 0$, then our model becomes

$$H : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{r-1} x_{ir-1} + \varepsilon_i,$$

or $\mathbf{Y} = \mathbf{X}_r \boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here \mathbf{X}_r consists of the first r columns of \mathbf{X} . \square

Examples 4.1 and 4.2 are special cases of Example 4.3 whereby we wish to test a submodel H versus the full model G . The same computer package used to fit G and obtain RSS can also be used to fit H and obtain $RSS_H = \|\mathbf{Y} - \mathbf{X}_r \hat{\boldsymbol{\beta}}_H\|^2$. We can also express the hypothesis constraints in the matrix form

$$\mathbf{0} = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} = \mathbf{A}\boldsymbol{\beta},$$

where the rows of \mathbf{A} are linearly independent.

Combining the three examples above, a general hypothesis can be expressed in the form $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. In the next section we develop a likelihood ratio test for testing H .

4.2 LIKELIHOOD RATIO TEST

Given the linear model $G : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{X} is $n \times p$ of rank p and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, we wish to test the hypothesis $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where \mathbf{A} is $q \times p$ of rank q . The likelihood function for G is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right].$$

In Section 3.5 we showed that the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, the least squares estimate, and $\hat{\sigma}^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/n$. The maximum value of the likelihood is given by [see equation (3.18)]

$$L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2} e^{-n/2}.$$

The next step is to find the maximum likelihood estimates subject to the constraints H . This requires use of the Lagrange multiplier approach of Section 3.8, where we now consider

$$\begin{aligned} r &= \log L(\boldsymbol{\beta}, \sigma^2) + (\boldsymbol{\beta}'\mathbf{A}' - \mathbf{c}')\lambda \\ &= \text{constant} - \frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + (\boldsymbol{\beta}'\mathbf{A}' - \mathbf{c}')\lambda. \end{aligned}$$

Using algebra almost identical to that which led to $\hat{\beta}_H$ of (3.38), we find that the maximum likelihood estimates are $\hat{\beta}_H$ and $\hat{\sigma}_H^2 = \|\mathbf{Y} - \mathbf{X}\hat{\beta}_H\|^2/n$ with a maximum of

$$L(\hat{\beta}_H, \hat{\sigma}_H^2) = (2\pi\hat{\sigma}_H^2)^{-n/2} e^{-n/2}. \quad (4.1)$$

The likelihood ratio test of H is given by

$$\Lambda = \frac{L(\hat{\beta}_H, \hat{\sigma}_H^2)}{L(\hat{\beta}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_H^2} \right)^{n/2}, \quad (4.2)$$

and according to the likelihood principle, we reject H if Λ is too small. Unfortunately, Λ is not a convenient test statistic and we show in the next section that

$$F = \frac{n-p}{q} (\Lambda^{-2/n} - 1)$$

has an $F_{q,n-p}$ distribution when H is true. We then reject H when F is too large.

4.3 F-TEST

4.3.1 Motivation

Since we want to test $H : \mathbf{A}\beta = \mathbf{c}$, a natural statistic for testing this is $\mathbf{A}\hat{\beta} - \mathbf{c}$; H will be rejected if $\mathbf{A}\hat{\beta}$ is sufficiently different from \mathbf{c} . However, not every element in $\mathbf{A}\hat{\beta}$ should be treated the same, as they have different precisions. One way of incorporating the precision of each $\hat{\beta}_i$ into a suitable distance measure is to use the quadratic $(\mathbf{A}\hat{\beta} - \mathbf{c})' \left(\text{Var}[\mathbf{A}\hat{\beta}] \right)^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c})$, where $\text{Var}[\mathbf{A}\hat{\beta}] = \sigma^2 \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}'$. If we estimate σ^2 by its unbiased estimate $S^2 = \text{RSS}/(n-p)$, we arrive at $(\mathbf{A}\hat{\beta} - \mathbf{c})' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}']^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c})/S^2$. We will now derive a test statistic which is a constant times this quadratic measure.

4.3.2 Derivation

Before we derive our main theorem, we recall some notation. We have

$$\text{RSS} = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \quad [= (n-p)S^2]$$

and

$$\text{RSS}_H = \|\mathbf{Y} - \mathbf{X}\hat{\beta}_H\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}_H\|^2,$$

where, from (3.38),

$$\hat{\beta}_H = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}']^{-1} (\mathbf{c} - \mathbf{A}\hat{\beta}). \quad (4.3)$$

Here RSS_H is the minimum value of $\epsilon'\epsilon$ subject to $\mathbf{A}\beta = \mathbf{c}$. An F -statistic for testing H is described in the following theorem.

THEOREM 4.1

$$(i) \text{ RSS}_H - \text{RSS} = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2 = (\mathbf{A}\hat{\beta} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{c}).$$

(ii)

$$\begin{aligned} E[\text{RSS}_H - \text{RSS}] &= \sigma^2 q + (\mathbf{A}\beta - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - \mathbf{c}) \\ &= \sigma^2 q + (\text{RSS}_H - \text{RSS})_{\mathbf{Y}=\mathbf{E}[\mathbf{Y}]} \end{aligned}$$

(iii) When H is true,

$$F = \frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-p)} = \frac{(\mathbf{A}\hat{\beta} - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{c})}{qS^2}$$

is distributed as $F_{q,n-p}$ (the F -distribution with q and $n-p$ degrees of freedom, respectively).

(iv) When $\mathbf{c} = \mathbf{0}$, F can be expressed in the form

$$F = \frac{n-p}{q} \frac{\mathbf{Y}'(\mathbf{P} - \mathbf{P}_H)\mathbf{Y}}{\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}},$$

where \mathbf{P}_H is symmetric and idempotent, and $\mathbf{P}_H\mathbf{P} = \mathbf{P}\mathbf{P}_H = \mathbf{P}_H$.

Proof. (i) From (3.43) and (3.44) in Section 3.8.1 we have

$$\begin{aligned} \text{RSS}_H - \text{RSS} &= \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2 \\ &= (\hat{\beta} - \hat{\beta}_H)' \mathbf{X}'\mathbf{X} (\hat{\beta} - \hat{\beta}_H), \end{aligned}$$

and substituting for $\hat{\beta} - \hat{\beta}_H$ using equation (4.3) leads to the required result.

(ii) The rows of \mathbf{A} are linearly independent and $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, so from Theorem 2.2 in Section 2.2, we get $\mathbf{A}\hat{\beta} \sim N_q(\mathbf{A}\beta, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$. Let $\mathbf{Z} = \mathbf{A}\hat{\beta} - \mathbf{c}$ and $\mathbf{B} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$; then $E[\mathbf{Z}] = \mathbf{A}\beta - \mathbf{c}$ and

$$\text{Var}[\mathbf{Z}] = \text{Var}[\mathbf{A}\hat{\beta}] = \sigma^2 \mathbf{B}.$$

Hence, using Theorem 1.5 in Section 1.5,

$$\begin{aligned} E[\text{RSS}_H - \text{RSS}] &= E[\mathbf{Z}'\mathbf{B}^{-1}\mathbf{Z}] \quad [\text{by (i)}] \\ &= \text{tr}(\sigma^2 \mathbf{B}^{-1} \mathbf{B}) + (\mathbf{A}\beta - \mathbf{c})' \mathbf{B}^{-1} (\mathbf{A}\beta - \mathbf{c}) \\ &= \text{tr}(\sigma^2 \mathbf{I}_q) + (\mathbf{A}\beta - \mathbf{c})' \mathbf{B}^{-1} (\mathbf{A}\beta - \mathbf{c}) \\ &= \sigma^2 q + (\mathbf{A}\beta - \mathbf{c})' \mathbf{B}^{-1} (\mathbf{A}\beta - \mathbf{c}). \end{aligned} \tag{4.4}$$

(iii) From (i), $\text{RSS}_H - \text{RSS}$ is a continuous function of $\hat{\beta}$ and is therefore independent of RSS [by Theorem 3.5(iii) in Section 3.4 and Example 1.11 in Section 1.5]. Also, when H is true, $\mathbf{A}\hat{\beta} \sim N_q(\mathbf{c}, \sigma^2 \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$, so that by Theorem 2.9,

$$\frac{\text{RSS}_H - \text{RSS}}{\sigma^2} = (\mathbf{A}\hat{\beta} - \mathbf{c})'(\text{Var}[\mathbf{A}\hat{\beta}])^{-1}(\mathbf{A}\hat{\beta} - \mathbf{c})$$

is χ^2_q . Finally, since $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$ [Theorem 3.5(iv)], we have that

$$F = \frac{(\text{RSS}_H - \text{RSS})/\sigma^2 q}{\text{RSS}/\sigma^2(n-p)}$$

is of the form $[\chi^2_q/q]/[\chi^2_{n-p}/(n-p)]$ when H is true. Hence $F \sim F_{q,n-p}$ when H is true.

(iv) Using equation (4.3) with $\mathbf{c} = \mathbf{0}$, we have

$$\begin{aligned} \hat{\mathbf{Y}}_H &= \mathbf{X}\hat{\beta}_H \\ &= \{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} \mathbf{Y} \end{aligned} \quad (4.5)$$

$$\begin{aligned} &= (\mathbf{P} - \mathbf{P}_1)\mathbf{Y} \\ &= \mathbf{P}_H\mathbf{Y}, \end{aligned} \quad (4.6)$$

say, where \mathbf{P}_H is symmetric. Multiplying the matrices together and canceling matrices with their inverses where possible, we find that \mathbf{P}_1 is symmetric and idempotent and $\mathbf{P}_1\mathbf{P} = \mathbf{P}\mathbf{P}_1 = \mathbf{P}_1$. Hence

$$\begin{aligned} \mathbf{P}_H^2 &= \mathbf{P}^2 - \mathbf{P}_1\mathbf{P} - \mathbf{P}\mathbf{P}_1 + \mathbf{P}_1^2 \\ &= \mathbf{P} - 2\mathbf{P}_1 + \mathbf{P}_1 \\ &= \mathbf{P} - \mathbf{P}_1 \\ &= \mathbf{P}_H, \end{aligned} \quad (4.7)$$

$$\mathbf{P}_H\mathbf{P} = (\mathbf{P} - \mathbf{P}_1)\mathbf{P} = \mathbf{P} - \mathbf{P}_1 = \mathbf{P}_H \quad (4.8)$$

and taking transposes, $\mathbf{P}\mathbf{P}_H = \mathbf{P}_H$. To complete the proof, we recall that $\text{RSS} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}$ and, in a similar fashion, obtain

$$\begin{aligned} \text{RSS}_H &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}_H\|^2 \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_H)^2\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_H)\mathbf{Y}. \end{aligned} \quad (4.9)$$

Thus $\text{RSS}_H - \text{RSS} = \mathbf{Y}'(\mathbf{P} - \mathbf{P}_H)\mathbf{Y}$. \square

We note that if $S_H^2 = (\text{RSS}_H - \text{RSS})/q$, then from Theorem 4.1(ii),

$$\begin{aligned} E[S_H^2] &= \sigma^2 + \frac{(\mathbf{A}\beta - \mathbf{c})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - \mathbf{c})}{q} \\ &= \sigma^2 + \delta, \quad \text{say}, \end{aligned}$$

where $\delta \geq 0$ [since $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' = \text{Var}[\mathbf{A}\hat{\beta}]/\sigma^2$ is positive-definite]. Also (Theorem 3.3, Section 3.3),

$$E[S^2] = \sigma^2.$$

When H is true, $\delta = 0$ and S_H^2 and S^2 are both unbiased estimates of σ^2 ; that is, $F = S_H^2/S^2 \approx 1$. When H is false, $\delta > 0$ and $E[S_H^2] > E[S^2]$, so that

$$E[F] = E[S_H^2]E\left[\frac{1}{S^2}\right] > E[S_H^2]/E[S^2] > 1$$

(by the independence of S_H^2 and S^2 , and A.13.3). Thus F gives some indication as to the “true state of affairs”; H is rejected if F is significantly large.

When $q > 2$ it is usually more convenient to obtain RSS and RSS_H by finding the unrestricted and restricted minimum values of $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ directly. However, if $q \leq 2$, F can usually be found most readily by applying the general matrix theory above; the matrix $[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']$ to be inverted is only of order one or two. It can also be found directly using the fact that $[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} = \text{Var}[\mathbf{A}\hat{\beta}]/\sigma^2$. Examples are given in Section 4.3.3. It should be noted that since RSS_H is unique, it does not matter what method we use for obtaining it. We could, for example, use the constraints $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ to eliminate some of the β_j and then minimize $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ with respect to the remaining β_j 's.

Part (iv) of Theorem 4.1 highlights the geometry underlying the F -test. This geometry can be used to extend the theory to the less-than-full-rank case (cf. Theorem 4.3 in Section 4.7).

From $\hat{\sigma}_H^2 = \text{RSS}_H/n$ and $\hat{\sigma}^2 = \text{RSS}/n$ we see that

$$\begin{aligned} F &= \frac{n-p}{q} \cdot \frac{\hat{\sigma}_H^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \\ &= \frac{n-p}{q} \left(\frac{\hat{\sigma}_H^2}{\hat{\sigma}^2} - 1 \right) \\ &= \frac{n-p}{q} (\Lambda^{-2/n} - 1), \end{aligned}$$

where Λ is the likelihood ratio test statistic (4.2).

EXERCISES 4a

1. Prove that $\text{RSS}_H - \text{RSS} \geq 0$.
2. If $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ is true, show that F can be expressed in the form

$$\frac{n-p}{q} \cdot \frac{\boldsymbol{\varepsilon}'(\mathbf{P} - \mathbf{P}_H)\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}}.$$

3. If $\hat{\lambda}_H$ is the least squares estimate of the Lagrange multiplier associated with the constraints $\mathbf{A}\beta = \mathbf{c}$ (cf. Section 3.8), show that

$$\text{RSS}_H - \text{RSS} = \sigma^2 \hat{\lambda}'_H (\text{Var}[\hat{\lambda}_H])^{-1} \hat{\lambda}_H.$$

(This idea is used to construct Lagrange multiplier tests.)

4. Suppose that we want to test $\mathbf{A}\beta = \mathbf{0}$, where \mathbf{A} is $q \times p$ of rank q . Assume that the last q columns of \mathbf{A} , \mathbf{A}_2 say, are linearly independent, so that $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$, where \mathbf{A}_2 is a nonsingular matrix. By expressing β_2 in terms of β_1 , find a matrix \mathbf{X}_A so that under H the linear model becomes $E[\mathbf{Y}] = \mathbf{X}_A\gamma$. Prove that \mathbf{X}_A has full rank.
5. Consider the full-rank model with $\mathbf{X}\beta = (\mathbf{X}_1, \mathbf{X}_2)(\beta'_1, \beta'_2)'$, where \mathbf{X}_2 is $n \times q$.
- (a) Obtain a test statistic for testing $H : \beta_2 = \mathbf{0}$ in the form of the right-hand side of Theorem 4.1(i). *Hint:* Use A.9.1.
- (b) Find $E[\text{RSS}_H - \text{RSS}]$.

4.3.3 Some Examples

EXAMPLE 4.4 Let

$$\begin{aligned} Y_1 &= \alpha_1 + \varepsilon_1, \\ Y_2 &= 2\alpha_1 - \alpha_2 + \varepsilon_2, \\ Y_3 &= \alpha_1 + 2\alpha_2 + \varepsilon_3, \end{aligned}$$

where $\varepsilon \sim N_3(\mathbf{0}, \sigma^2 \mathbf{I}_3)$. We now derive the F -statistic for testing $H : \alpha_1 = \alpha_2$.

We note first that

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix},$$

or $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where \mathbf{X} is 3×2 of rank 2. Also, H is equivalent to

$$(1, -1) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = 0,$$

or $\mathbf{A}\beta = \mathbf{0}$, where \mathbf{A} is 1×2 of rank 1. Hence the theory above applies with $n = 3$, $p = 2$, and $q = 1$.

The next step is to find

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 5 \end{pmatrix}.$$

Then

$$\begin{aligned}\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} &= \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} Y_1 + 2Y_2 + Y_3 \\ -Y_2 + Y_3 \end{pmatrix}, \\ \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} &= \begin{pmatrix} \frac{1}{6}(Y_1 + 2Y_2 + Y_3) \\ \frac{1}{5}(-Y_2 + Y_3) \end{pmatrix}\end{aligned}$$

and from equation (3.9),

$$\begin{aligned}\text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= Y_1^2 + Y_2^2 + Y_3^2 - 6\hat{\alpha}_1^2 - 5\hat{\alpha}_2^2.\end{aligned}$$

We have at least two methods of finding the F -statistic.

Method 1

$$\begin{aligned}\mathbf{A}\hat{\beta} &= \hat{\alpha}_1 - \hat{\alpha}_2, \\ \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' &= (1, -1) \begin{pmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{6} + \frac{1}{5} = \frac{11}{30},\end{aligned}$$

and

$$\begin{aligned}F &= \frac{(\mathbf{A}\hat{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\hat{\beta}}{qS^2} \\ &= \frac{(\hat{\alpha}_1 - \hat{\alpha}_2)^2}{\frac{11}{30}S^2},\end{aligned}$$

where $S^2 = \text{RSS}/(n - p) = \text{RSS}$. When H is true, $F \sim F_{q, n-p} = F_{1,1}$.

Method 2

Let $\alpha_1 = \alpha_2 = \alpha$. When H is true, we have

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (Y_1 - \alpha)^2 + (Y_2 - \alpha)^2 + (Y_3 - 3\alpha)^2$$

and $\partial\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/\partial\alpha = 0$ implies that $\hat{\alpha}_H = \frac{1}{11}(Y_1 + Y_2 + 3Y_3)$. Hence

$$\text{RSS}_H = (Y_1 - \hat{\alpha}_H)^2 + (Y_2 - \hat{\alpha}_H)^2 + (Y_3 - 3\hat{\alpha}_H)^2 \quad (4.10)$$

and

$$F = \frac{\text{RSS}_H - \text{RSS}}{\text{RSS}}. \quad \square$$

EXAMPLE 4.5 Let U_1, \dots, U_{n_1} be sampled independently from $N(\mu_1, \sigma^2)$, and let V_1, \dots, V_{n_2} be sampled independently from $N(\mu_2, \sigma^2)$. We now derive a test statistic for $H : \mu_1 = \mu_2$.

Writing

$$U_i = \mu_1 + \varepsilon_i \quad (i = 1, 2, \dots, n_1)$$

and

$$V_j = \mu_2 + \varepsilon_{n_1+j} \quad (j = 1, 2, \dots, n_2),$$

we have the matrix representation

$$\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{n_1} \\ V_1 \\ V_2 \\ \vdots \\ V_{n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad (4.11)$$

where $n = n_1 + n_2$. Thus our model is of the form $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where \mathbf{X} is $n \times 2$ of rank 2 and $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Also, as in Example 4.4, H takes the form $\mathbf{A}\beta = \mathbf{0}$, so that our general regression theory applies with $p = 2$ and $q = 1$. Now

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix},$$

so that

$$\hat{\beta} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} \sum U_i \\ \sum V_j \end{pmatrix} = \begin{pmatrix} \bar{U} \\ \bar{V} \end{pmatrix},$$

$$\mathbf{A}\hat{\beta} = \hat{\mu}_2 - \hat{\mu}_1 = \bar{U} - \bar{V},$$

and

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \sum_i U_i^2 + \sum_j V_j^2 - n_1 \bar{U}^2 - n_2 \bar{V}^2 \\ &= \sum_i (U_i - \bar{U})^2 + \sum_j (V_j - \bar{V})^2. \end{aligned}$$

Also,

$$\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' = \frac{1}{n_1} + \frac{1}{n_2},$$

so that the F -statistic for H is

$$\begin{aligned} F &= \frac{(\mathbf{A}\hat{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\hat{\beta}}{qS^2} \\ &= \frac{(\bar{U} - \bar{V})^2}{S^2(1/n_1 + 1/n_2)}, \end{aligned} \quad (4.12)$$

where $S^2 = \text{RSS}/(n - p) = \text{RSS}/(n_1 + n_2 - 2)$. When H is true, $F \sim F_{1,n_1+n_2-2}$.

Since, distribution-wise, we have the identity $F_{1,k} \equiv t_k^2$, the F -statistic above is the square of the usual t -statistic for testing the difference of two normal means (assuming equal variances). \square

EXAMPLE 4.6 Given the general linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

we can obtain a test statistic for $H : \beta_j = c$, where $j > 0$.

We first need the following partition:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} l & \mathbf{m}' \\ \mathbf{m} & \mathbf{D} \end{pmatrix},$$

where l is 1×1 . Now H is of the form $\mathbf{a}'\boldsymbol{\beta} = \mathbf{c}$, where \mathbf{a}' is the row vector with unity in the $(j+1)$ th position and zeros elsewhere. Therefore, using the general matrix theory, $\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} = d_{jj}$ (the j th diagonal element of \mathbf{D}), $\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{c} = \hat{\beta}_j - c$, and the F -statistic is

$$F = \frac{(\hat{\beta}_j - c)^2}{S^2 d_{jj}}, \quad (4.13)$$

which has the $F_{1,n-p}$ distribution when H is true. As in Example 4.5, F is again the square of the usual t -statistic.

The matrix \mathbf{D} can be identified using the method of A.9 for inverting a partitioned symmetric matrix. Let $\mathbf{1}_n$ be an $n \times 1$ column vector of 1's and let $\bar{\mathbf{x}}' = (\bar{x}_{\cdot 1}, \bar{x}_{\cdot 2}, \dots, \bar{x}_{\cdot p-1})$. Then $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_1)$,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{\mathbf{x}}' \\ n\bar{\mathbf{x}} & \mathbf{X}_1'\mathbf{X}_1 \end{pmatrix},$$

and by A.9.1,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \bar{\mathbf{x}}'\mathbf{V}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{V}^{-1} \\ -\mathbf{V}^{-1}\bar{\mathbf{x}} & \mathbf{V}^{-1} \end{pmatrix}, \quad (4.14)$$

where $\mathbf{V} = (v_{jk}) = \mathbf{X}_1'\mathbf{X}_1 - n\bar{\mathbf{x}}\bar{\mathbf{x}}'$ and

$$\begin{aligned} v_{jk} &= \sum_i x_{ij}x_{ik} - n\bar{x}_{\cdot j}\bar{x}_{\cdot k} \\ &= \sum_i (x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}). \end{aligned} \quad (4.15)$$

Thus \mathbf{D} is the inverse of \mathbf{V} , where \mathbf{V} is the matrix of *corrected sums of squares and products* of the x 's. In the notation of Section 3.11, $\mathbf{V} = \tilde{\mathbf{X}}'\tilde{\mathbf{X}}$. Similar examples are considered in Section 9.7. \square

EXAMPLE 4.7 Suppose that in Example 4.6 we want to test $H : \mathbf{a}'\boldsymbol{\beta} = c$. Then $q = 1$,

$$\text{var}[\mathbf{a}'\hat{\boldsymbol{\beta}}] = \mathbf{a}'\text{Var}[\hat{\boldsymbol{\beta}}]\mathbf{a} = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}$$

and

$$F = \frac{(\mathbf{a}'\hat{\boldsymbol{\beta}} - c)^2}{s^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}},$$

which is distributed as $F_{1,n-p}$ when H is true. Again this is the square of usual t -statistic, which we can also derive directly as follows.

By Theorem 2.2, $\mathbf{a}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{a}'\boldsymbol{\beta}, \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a})$, so that

$$U_i = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}}{\sigma \{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\}^{1/2}} \sim N(0, 1).$$

Also, by Theorem 3.5 (Section 3.4), $V = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$, and since S^2 is statistically independent of $\hat{\boldsymbol{\beta}}$, V is independent of U . Hence

$$\begin{aligned} T &= \frac{U}{\sqrt{V/(n-p)}} \\ &= \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}}{S\{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\}^{1/2}} \end{aligned} \quad (4.16)$$

has the t_{n-p} distribution. To test $H : \mathbf{a}'\boldsymbol{\beta} = c$ we set $\mathbf{a}'\boldsymbol{\beta}$ equal to c in T and reject H at the α level of significance if $|T| \geq t_{n-p}^{(1/2)\alpha}$; here $t_{n-p}^{(1/2)\alpha}$ is the upper $\alpha/2$ point of the t_{n-p} distribution; that is, $\text{pr}(T > t_{n-p}^{(1/2)\alpha}) = \alpha/2$. Alternatively, we can construct a $100(1 - \alpha)\%$ confidence interval for $\mathbf{a}'\boldsymbol{\beta}$, namely,

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{n-p}^{(1/2)\alpha} S\{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\}^{1/2}, \quad (4.17)$$

or since $S^2\{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\}$ is an unbiased estimate of $\sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}$ (the variance of $\mathbf{a}'\hat{\boldsymbol{\beta}}$),

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{n-p}^{(1/2)\alpha} \hat{\sigma}_{\mathbf{a}'\hat{\boldsymbol{\beta}}}, \quad \text{say,} \quad (4.18)$$

and see if the interval above contains c . \square

4.3.4 The Straight Line

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), and suppose that we wish to test $H : \beta_1 = c$. Then $\mathbf{X} = (\mathbf{1}_n, \mathbf{x})$,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n, & n\bar{x} \\ n\bar{x}, & \sum x_i^2 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n} \sum x_i^2, & -\bar{x} \\ -\bar{x}, & 1 \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Also, from $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ we have, after some simplification,

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

and

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{Y} + \hat{\beta}_1(x_i - \bar{x}).\end{aligned}$$

(Actually, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained more readily by differentiating $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ with respect to β_0 and β_1 .) Finally, from Example 4.6 with $p = 2$, the F -statistic for testing H is given by

$$F = \frac{(\hat{\beta}_1 - c)^2}{S^2 d_{11}} = \frac{(\hat{\beta}_1 - c)^2}{S^2 / \sum(x_i - \bar{x})^2}, \quad (4.19)$$

where

$$\begin{aligned}(n - 2)S^2 &= \sum(Y_i - \hat{Y}_i)^2 \\ &= \sum[Y_i - \bar{Y} - \hat{\beta}_1(x_i - \bar{x})]^2 \\ &= \sum(Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 \\ &= \sum(Y_i - \bar{Y})^2 - \sum(\hat{Y}_i - \bar{Y})^2.\end{aligned} \quad (4.20)$$

We note from (4.21) that

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \quad (4.22)$$

$$= \sum(Y_i - \hat{Y}_i)^2 + r^2 \sum(Y_i - \bar{Y})^2, \quad (4.23)$$

where

$$\begin{aligned}r^2 &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= \frac{\hat{\beta}_1^2 \sum(x_i - \bar{x})^2}{\sum(Y_i - \bar{Y})^2} \\ &= \frac{[\sum(Y_i - \bar{Y})(x_i - \bar{x})]^2}{\sum(Y_i - \bar{Y})^2 \sum(x_i - \bar{x})^2}\end{aligned} \quad (4.24)$$

is the square of the sample correlation between Y and x . Also, r is a measure of the degree of linearity between Y and x since, from (4.23),

$$\begin{aligned}\text{RSS} &= \sum(Y_i - \hat{Y}_i)^2 \\ &= (1 - r^2) \sum(Y_i - \bar{Y})^2,\end{aligned} \quad (4.25)$$

so that the larger the value of r^2 , the smaller RSS and the better the fit of the estimated regression line to the observations.

Although $1 - r^2$ is a useful measure of fit, the correlation r itself is of doubtful use in making inferences. Tukey [1954] makes the provocative but not unreasonable statement that “correlation coefficients are justified in two and only two circumstances, when they are regression coefficients, or when the measurement of one or both variables on a determinate scale is hopeless.” The first part of his statement refers to the situation where X and Y have a bivariate normal distribution; we have (Example 2.9)

$$\begin{aligned} E[Y|X = x] &= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \\ &= \beta_0 + \beta_1 x, \end{aligned}$$

and when $\sigma_X^2 = \sigma_Y^2$, $\beta_1 = \rho$. One area where correlation coefficients are widely used, and determinate scales seem hopeless, is in the social sciences. Here the measuring scales are often completely arbitrary, so that observations are essentially only ranks. A helpful discussion on the question of correlation versus regression is given by Warren [1971].

We note that when $c = 0$, the F -statistic (4.19) can also be expressed in terms of r^2 . From equation (4.25) we have

$$(n - 2)S^2 = (1 - r^2) \sum (Y_i - \bar{Y})^2,$$

so that

$$\begin{aligned} F &= \frac{\hat{\beta}_1^2 \sum (x_i - \bar{x})^2 (n - 2)}{(1 - r^2) \sum (Y_i - \bar{Y})^2} \\ &= \frac{r^2(n - 2)}{1 - r^2}. \end{aligned}$$

The usual t -statistic for testing $\beta_1 = 0$ can also be expressed in the same form, namely,

$$T = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}. \quad (4.26)$$

EXERCISES 4b

- Let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$, $i = 1, 2, \dots, n$, where the ε_i are independent $N(0, \sigma^2)$. Prove that the F -statistic for testing the hypothesis $H : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$ ($0 < q \leq p - 1$) is unchanged if a constant, c , say, is subtracted from each Y_i .
- Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $(i = 1, 2, \dots, n)$, where the ε_i are independent $N(0, \sigma^2)$.
 - Show that the correlation coefficient of $\hat{\beta}_0$ and $\hat{\beta}_1$ is $-n\bar{x}/(n\sqrt{\sum x_i^2})$.

- (b) Derive an F -statistic for testing $H : \beta_0 = 0$.
3. Given that $\bar{x} = 0$, derive an F -statistic for testing the hypothesis $H : \beta_0 = \beta_1$ in Exercise No. 2 above. Show that it is equivalent to a certain t -test.
4. Let
- $$\begin{aligned} Y_1 &= \theta_1 + \theta_2 + \varepsilon_1, \\ Y_2 &= 2\theta_2 + \varepsilon_2, \end{aligned}$$
- and
- $$Y_3 = -\theta_1 + \theta_2 + \varepsilon_3,$$
- where the ε_i ($i = 1, 2, 3$) are independent $N(0, \sigma^2)$. Derive an F -statistic for testing the hypothesis $H : \theta_1 = 2\theta_2$.
5. Given $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_4(\mathbf{0}, \sigma^2 \mathbf{I}_4)$ and $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 0$, show that the F -statistic for testing $H : \theta_1 = \theta_3$ is

$$\frac{2(Y_1 - Y_3)^2}{(Y_1 + Y_2 + Y_3 + Y_4)^2}.$$

4.4 MULTIPLE CORRELATION COEFFICIENT

For a straight line, from equation (4.25) we have

$$\text{RSS} = (1 - r^2) \sum (Y_i - \bar{Y})^2.$$

Thus, r^2 is a measure of how well the least squares line fits the data. Noting that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1(x_i - \bar{x})$, we have

$$\begin{aligned} r &= \frac{\hat{\beta}_1 \sum (Y_i - \bar{Y})(x_i - \bar{x})}{\hat{\beta}_1 [(\sum (Y_i - \bar{Y})^2)(\sum (x_i - \bar{x})^2)]^{1/2}} \\ &= \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{[\sum (Y_i - \bar{Y})^2](\sum (\hat{Y}_i - \bar{Y})^2)^{1/2}}, \end{aligned}$$

which is the correlation coefficient of the pairs (Y_i, \hat{Y}_i) . To demonstrate this, we note that $\sum (Y_i - \hat{Y}_i) = \sum [Y_i - \bar{Y}_i - \hat{\beta}_1(x_i - \bar{x})] = 0$, so that the mean of the \hat{Y}_i , $\bar{\hat{Y}}$ say, is the same as \bar{Y} .

This reformulation of r suggests how we might generalize this measure from a straight line to a general linear model. We can now define the *sample*

multiple correlation coefficient R as the correlation coefficient of the pairs (Y_i, \hat{Y}_i) , namely,

$$R = \frac{\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\left\{ \sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{\hat{Y}})^2 \right\}^{1/2}}. \quad (4.27)$$

The quantity R^2 is commonly called the *coefficient of determination*. We now prove a useful theorem that generalizes equations (4.22) and (4.24).

THEOREM 4.2

(i)

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

(ii)

$$\begin{aligned} R^2 &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{RSS}}{\sum(Y_i - \bar{Y})^2}. \end{aligned}$$

Proof. (i) $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$, so that

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P}^2\mathbf{Y} = \mathbf{Y}'\mathbf{P}\mathbf{Y} = \mathbf{Y}'\hat{\mathbf{Y}}. \quad (4.28)$$

Also, by differentiating $\sum_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{p-1} x_{ip-1})^2$ with respect to $\hat{\beta}_0$, we have one of the normal equations for $\hat{\beta}$, namely,

$$\sum_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{p-1} x_{ip-1}) = 0$$

or

$$\sum_i(Y_i - \hat{Y}_i) = 0. \quad (4.29)$$

Hence

$$\begin{aligned} \sum(Y_i - \bar{Y})^2 &= \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2, \end{aligned}$$

since

$$\begin{aligned} \sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum(Y_i - \hat{Y}_i)\hat{Y}_i \quad [\text{by equation (4.29)}] \\ &= (\mathbf{Y} - \hat{\mathbf{Y}})' \hat{\mathbf{Y}} \\ &= 0 \quad [\text{by equation (4.28)}]. \end{aligned}$$

(ii) From equation (4.29), we get $\bar{Y} = \bar{Y}$, so that

$$\begin{aligned}\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= \sum(\hat{Y}_i - \bar{Y})^2,\end{aligned}$$

and the required expression for R^2 follows immediately from (4.27). The second expression for R^2 follows from (i). \square

From the theorem above, we have a generalization of (4.25), namely,

$$\text{RSS} = (1 - R^2) \sum(Y_i - \bar{Y})^2, \quad (4.30)$$

and the greater the value of R^2 , the closer the fit of the estimated surface to the observed data; if $Y_i = \hat{Y}_i$, we have a perfect fit and $R^2 = 1$, otherwise $R^2 < 1$. When there is just a single x -regressor then $R^2 = r^2$. By writing $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, where $(\mathbf{X}'\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$, we find that the theorem above still holds even when \mathbf{X} is not of full rank. Alternatively, we can write $\mathbf{P} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, where \mathbf{X}_1 is the matrix of linearly independent columns of \mathbf{X} .

EXAMPLE 4.8 Given the linear model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ ($i = 1, 2, \dots, n$), suppose that we wish to test whether or not the regression on the regressor variables is significant; that is, test $H : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. Then H takes the form $\mathbf{A}\beta = \mathbf{0}$, where $\mathbf{A} = (\mathbf{0}, \mathbf{I}_{p-1})$ is a $(p-1) \times p$ matrix of rank $p-1$, so that the general regression theory applies with $q = p-1$. We therefore find that

$$\begin{aligned}\text{RSS}_H &= \underset{\beta_0}{\text{minimum}} \sum_i (Y_i - \beta_0)^2 \\ &= \sum(Y_i - \bar{Y})^2,\end{aligned}$$

and by Theorem 4.2 and (4.30),

$$\begin{aligned}F &= \frac{(\text{RSS}_H - \text{RSS})/(p-1)}{\text{RSS}/(n-p)} \\ &= \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\text{RSS}} \frac{n-p}{p-1} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{(1-R^2)\sum(Y_i - \bar{Y})^2} \frac{n-p}{p-1} \\ &= \frac{R^2}{1-R^2} \frac{n-p}{p-1},\end{aligned} \quad (4.31)$$

where $F \sim F_{p-1, n-p}$ when H is true.

The statistic F provides a test for “overall” regression, and we reject H if $F > F_{p-1,n-p}^\alpha$, $F_{p-1,n-p}^\alpha$ being the upper α point for the $F_{p-1,n-p}$ distribution. If we reject H , we say that there is a significant regression and the x_{ij} values cannot be totally ignored. However, the rejection of H does not mean that the fitted equation $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ is necessarily adequate, particularly for predictive purposes. Since a large R^2 leads to a large F statistic, a working rule suggested by Draper and Smith [1998: p. 247] for model adequacy is that the observed F -ratio must be at least four or five times $F_{p-1,n-p}^\alpha$. \square

EXERCISES 4c

- Suppose that $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. Find the distribution of R^2 and hence prove that

$$E[R^2] = \frac{p-1}{n-1}.$$

- For the general linear full-rank regression model, prove that R^2 and the F -statistic for testing $H : \beta_j = 0$ ($j \neq 0$) are independent of the units in which the Y_i and the x_{ij} are measured.
- Given the full-rank model, suppose that we wish to test $H : \beta_j = 0$, $j \neq 0$. Let R_H^2 be the coefficient of determination for the model with $\beta_j = 0$.
 - Prove that the F -statistic for testing H is given by

$$F = \frac{R^2 - R_H^2}{1 - R^2} \cdot \frac{n-p}{1}.$$

(This result shows that F is a test for a significant reduction in R^2 .)

- Deduce that R^2 can never increase when a β coefficient is set equal to zero.

4.5 CANONICAL FORM FOR H

Suppose that we wish to test $H : \mathbf{A}\beta = \mathbf{0}$, where \mathbf{A} is $q \times p$ of rank q , for the full-rank model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Since \mathbf{A} has q linearly independent columns, we can assume without loss of generality (by relabeling the β_j if necessary) that these are the last q columns; thus $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$, where \mathbf{A}_2 is a $q \times q$ nonsingular matrix. Partitioning β in the same way, we have

$$\mathbf{0} = \mathbf{A}\beta = \mathbf{A}_1\beta_1 + \mathbf{A}_2\beta_2,$$

and multiplying through by \mathbf{A}_2^{-1} leads to

$$\beta_2 = -\mathbf{A}_2^{-1}\mathbf{A}_1\beta_1. \quad (4.32)$$

This means that under the hypothesis H , the regression model takes the "canonical" form

$$\begin{aligned} \mathbf{X}\beta &= (\mathbf{X}_1, \mathbf{X}_2)\beta \\ &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 \\ &= (\mathbf{X}_1 - \mathbf{X}_2\mathbf{A}_2^{-1}\mathbf{A}_1)\beta_1 \\ &= \mathbf{X}_H\gamma, \end{aligned} \quad (4.33)$$

say, where \mathbf{X}_H is $n \times (p - q)$ of rank $p - q$ and $\gamma = \beta_1$. The matrix \mathbf{X}_H has linearly independent columns since

$$\mathbf{X}_H\beta_1 = \mathbf{0} \Leftrightarrow \mathbf{X}\beta = \mathbf{0} \Leftrightarrow \beta = \mathbf{0} \Leftrightarrow \beta_1 = \mathbf{0}.$$

By expressing the hypothesized model $H : E[\mathbf{Y}] = \mathbf{X}_H\gamma$ in the same form as the original model $E[\mathbf{Y}] = \mathbf{X}\beta$, we see that the same computer package can be used for calculating both RSS and RSS_H , provided, of course, that \mathbf{X}_H can be found easily and accurately. If \mathbf{X}_H is not readily found, then the numerator of the F -statistic for testing H can be computed directly using the method of Section 11.11. We note that $q = \text{rank}(\mathbf{X}) - \text{rank}(\mathbf{X}_H)$.

One very simple application of the theory above is to test $H : \beta_2 = \mathbf{0}$; \mathbf{X}_H is simply the first $p - q$ columns of \mathbf{X} . Further applications are given in Section 6.4, Chapter 8, and in Section 4.6.

EXERCISES 4d

1. Express the hypotheses in Examples 4.4 and 4.5 in canonical form.
2. Suppose that we have n_1 observations on w_1, w_2, \dots, w_{p-1} and U , giving the model

$$U_i = \gamma_0^{(1)} + \gamma_1^{(1)}w_{i1} + \cdots + \gamma_{p-1}^{(1)}w_{i,p-1} + \eta_i \quad (i = 1, 2, \dots, n_1).$$

We are now given n_2 ($> p$) additional observations which can be expressed in the same way, namely,

$$\begin{aligned} U_i &= \gamma_0^{(2)} + \gamma_1^{(2)}w_{i1} + \cdots + \gamma_{p-1}^{(2)}w_{i,p-1} + \eta_i \\ &\quad (i = n_1 + 1, n_2 + 2, \dots, n_1 + n_2). \end{aligned}$$

Derive an F -statistic for testing the hypothesis H that the additional observations come from the same model.

4.6 GOODNESS-OF-FIT TEST

Suppose that for each set of values taken by the regressors in the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad (4.34)$$

we have repeated observations on Y , namely,

$$Y_{ir} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_{ir}, \quad (4.35)$$

where $E[\varepsilon_{ir}] = 0$, $\text{var}[\varepsilon_{ir}] = \sigma^2$, $r = 1, 2, \dots, R_i$, and $i = 1, 2, \dots, n$. We assume that the R_i repetitions Y_{ir} for a particular set $(x_{i1}, \dots, x_{i,p-1})$ are genuine replications and not just repetitions of the same reading for Y_i in a given experiment. For example, if $p = 2$, Y is yield and x_1 is temperature, then the replicated observations Y_{ir} ($r = 1, 2, \dots, R_i$) are obtained by having R_i experiments with $x_1 = x_{i1}$ in each experiment, not by having a single experiment with $x_1 = x_{i1}$ and measuring the yield R_i times. Clearly, the latter method would supply only information on the variance of the device for measuring yield, which is just part of the variance σ^2 ; our definition of σ^2 also includes the variation in yield between experiments at the same temperature. However, given genuine replications, it is possible to test whether the model (4.34) is appropriate using the F -statistic derived below.

Let $Y_{ir} = \phi_i + \varepsilon_{ir}$, say. Then writing

$$\mathbf{Y}' = (Y_{11}, Y_{12}, \dots, Y_{1R_1}, \dots, Y_{n1}, Y_{n2}, \dots, Y_{nR_n}), \text{ etc.,}$$

we have $\mathbf{Y} = \mathbf{W}\phi + \varepsilon$, where

$$\mathbf{W}\phi = \begin{pmatrix} \mathbf{1}_{R_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{R_2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{R_n} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix}. \quad (4.36)$$

Defining $N = \sum_i R_i$, then \mathbf{W} is an $N \times n$ matrix of rank n ; we also assume that $\varepsilon \sim N_N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$. Now testing the adequacy of (4.34) is equivalent to testing the hypothesis

$$H : \phi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} \quad (i = 1, 2, \dots, n)$$

or $H : \phi = \mathbf{X}\beta$, where \mathbf{X} is $n \times p$ of rank p . We thus have the canonical form (cf. Section 4.5) $E[\mathbf{Y}] = \mathbf{WX}\beta$. We note in passing that H can be converted into the more familiar constraint equation form using the following lemma.

LEMMA $\phi \in \mathcal{C}(\mathbf{X})$ if and only if $\mathbf{A}\phi = \mathbf{0}$ for some $(n-p) \times n$ matrix \mathbf{A} of rank $n-p$.

Proof. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. If $\phi \in \mathcal{C}(\mathbf{X})$, that is, $\phi = \mathbf{X}\beta$ for some β , then $(\mathbf{I}_n - \mathbf{P})\phi = (\mathbf{I}_n - \mathbf{P})\mathbf{X}\beta = \mathbf{0}$ [by Theorem 3.1(iii)]. Conversely, if $(\mathbf{I}_n - \mathbf{P})\phi = \mathbf{0}$, then $\phi = \mathbf{P}\phi = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\phi = \mathbf{X}\gamma \in \mathcal{C}(\mathbf{X})$. Hence

$\phi \in \mathcal{C}(\mathbf{X})$ if and only if $(\mathbf{I}_n - \mathbf{P})\phi = \mathbf{0}$. By Theorem 3.1(ii) the $n \times n$ matrix $\mathbf{I}_n - \mathbf{P}$ has rank $n - p$ and therefore has $n - p$ linearly independent rows which we can take as our required matrix \mathbf{A} . \square

Using the Lemma above or the canonical form, we see that the general regression theory applies to H , but with n , p , and q replaced by N , n , and $n - p$, respectively; hence

$$F = \frac{(RSS_H - RSS)/(n - p)}{RSS/(N - n)}.$$

Here RSS is found directly by minimizing $\sum_i \sum_r (Y_{ir} - \phi)^2$. Thus differentiating partially with respect to ϕ , we have

$$\hat{\phi}_i = \frac{\sum_r Y_{ir}}{R_i} = \bar{Y}_{i.} \quad \text{and} \quad RSS = \sum \sum (Y_{ir} - \bar{Y}_{i.})^2.$$

To find RSS_H we minimize $\sum_i \sum_r (Y_{ir} - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1})^2$ ($= d$, say). Therefore, setting $\partial d / \partial \beta_0 = 0$ and $\partial d / \partial \beta_j = 0$ ($j \neq 0$), we have

$$\sum_i R_i (\bar{Y}_{i.} - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1}) = 0 \quad (4.37)$$

and

$$\sum_i \sum_r x_{ij} (Y_{ir} - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1}) = 0 \quad (j = 1, 2, \dots, p-1),$$

that is,

$$\sum_i R_i x_{ij} (\bar{Y}_{i.} - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1}) = 0. \quad (4.38)$$

Since equations (4.37) and (4.38) are identical to the usual normal equations, except that Y_i is replaced by $Z_i = \bar{Y}_{i.}$, we have

$$\hat{\beta}_H = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$$

and

$$RSS_H = \sum_i \sum_r (Y_{ir} - \hat{\beta}_{0H} - \hat{\beta}_{1H} x_{i1} - \cdots - \hat{\beta}_{p-1,H} x_{ip-1})^2.$$

4.7 F-TEST AND PROJECTION MATRICES

The theory of Theorem 4.1 can be generalized to the case when \mathbf{X} has less than full rank and the rows of \mathbf{A} in testing $H : \mathbf{A}\beta = \mathbf{0}$ are linearly dependent, so that some of the hypothesis constraints are redundant. However, the algebra involves the use of generalized inverses, and the resulting formulation is not the one used to actually carry out the computations. Theorem 4.1(iv) suggests

that a more elegant approach is to use projection matrices. To set the scene, suppose that we have the model $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\theta} \in \Omega$ (an r -dimensional subspace of \mathfrak{R}_n), and we wish to test $H : \boldsymbol{\theta} \in \omega$, where ω is an $(r - q)$ -dimensional subspace of Ω . Then we have the following theorem.

THEOREM 4.3 *When H is true and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$,*

$$F = \frac{(\text{RSS}_H - \text{RSS})/q}{\text{RSS}/(n-r)} = \frac{\boldsymbol{\varepsilon}'(\mathbf{P}_\Omega - \mathbf{P}_\omega)\boldsymbol{\varepsilon}/q}{\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{\varepsilon}/(n-r)} \sim F_{q,n-r},$$

where \mathbf{P}_Ω and \mathbf{P}_ω are the symmetric idempotent matrices projecting \mathfrak{R}_n onto Ω and ω , respectively (Appendix B).

Proof. $\hat{\boldsymbol{\theta}} = \mathbf{P}_\Omega \mathbf{Y}$ and $\hat{\boldsymbol{\theta}}_H = \mathbf{P}_\omega \mathbf{Y}$ are the respective least squares estimates of $\boldsymbol{\theta}$, so that

$$\text{RSS} = \|\mathbf{Y} - \hat{\boldsymbol{\theta}}\|^2 = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\Omega)\mathbf{Y}$$

and

$$\text{RSS}_H = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\omega)\mathbf{Y}.$$

Also, $(\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{\theta} = \mathbf{0}$ (since $\boldsymbol{\theta} \in \Omega$), which implies that

$$\text{RSS} = (\mathbf{Y} - \boldsymbol{\theta})'(\mathbf{I}_n - \mathbf{P}_\Omega)(\mathbf{Y} - \boldsymbol{\theta}) = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{\varepsilon}.$$

Similarly, when H is true, $\boldsymbol{\theta} \in \omega$ and

$$\text{RSS}_H = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_\omega)\boldsymbol{\varepsilon}.$$

Now $(\mathbf{I}_n - \mathbf{P}_\Omega)$ and $(\mathbf{P}_\Omega - \mathbf{P}_\omega)$ project onto Ω^\perp and $\omega^\perp \cap \Omega$ (by B.1.6 and B.3.2), so that these matrices are symmetric and idempotent (B.1.4) and have ranks $n - r$ and $r - (r - q) = q$ by B.1.5. Since $\mathbf{P}_\Omega \mathbf{P}_\omega = \mathbf{P}_\omega$ we have $(\mathbf{I}_n - \mathbf{P}_\Omega)(\mathbf{P}_\Omega - \mathbf{P}_\omega) = \mathbf{0}$. Hence by Theorem 2.7 and Example 2.12 in Section 2.4, $\boldsymbol{\varepsilon}'(\mathbf{P}_\Omega - \mathbf{P}_\omega)\boldsymbol{\varepsilon}/\sigma^2$ and $\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{\varepsilon}/\sigma^2$ are independently distributed as χ_q^2 and χ_{n-r}^2 , respectively. Thus $F \sim F_{q,n-r}$. \square

It is readily seen that Theorem 4.1(iv) is a special case of the above; there $\Omega = \mathcal{C}(\mathbf{X})$, and when $\mathbf{c} = \mathbf{0}$, $\omega = \mathcal{N}(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \cap \Omega$.

MISCELLANEOUS EXERCISES 4

1. Aerial observations Y_1, Y_2, Y_3 , and Y_4 are made of angles $\theta_1, \theta_2, \theta_3$, and θ_4 , respectively, of a quadrilateral on the ground. If the observations are subject to independent normal errors with zero means and common variance σ^2 , derive a test statistic for the hypothesis that the quadrilateral is a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$.

(Adapted from Silvey [1970].)

2. Given the two regression lines

$$Y_{ki} = \beta_k x_i + \varepsilon_{ki} \quad (k = 1, 2; i = 1, 2, \dots, n),$$

show that the F -statistic for testing $H : \beta_1 = \beta_2$ can be put in the form

$$F = \frac{(\hat{\beta}_1 - \hat{\beta}_2)^2}{2S^2 (\sum_i x_i^2)^{-1}}.$$

Obtain RSS and RSS_H and verify that

$$\text{RSS}_H - \text{RSS} = \frac{\sum_i x_i^2 (\hat{\beta}_1 - \hat{\beta}_2)^2}{2}.$$

3. Show that the usual full-rank regression model and hypothesis $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ can be transformed to the model $\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\eta}$, where $\mu_{p+1} = \mu_{p+2} = \dots = \mu_n = 0$ and $\boldsymbol{\eta} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, and the hypothesis $H : \mu_1 = \mu_2 = \dots = \mu_q = 0$. *Hint:* Choose an orthonormal basis of $p - q$ vectors $\{\alpha_{q+1}, \alpha_{q+2}, \dots, \alpha_p\}$ for $C(\mathbf{X}_A)$, where \mathbf{X}_A is defined in Exercises 4a, No. 4; extend this to an orthonormal basis $\{\alpha_1, \alpha_2, \dots, \alpha_p\}$ for $C(\mathbf{X})$; and then extend once more to an orthonormal basis $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ for \mathfrak{R}_n . Consider the transformation $\mathbf{Z} = \mathbf{T}'\mathbf{Y}$, where $\mathbf{T} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is orthogonal.
4. A series of $n + 1$ observations Y_i ($i = 1, 2, \dots, n + 1$) are taken from a normal distribution with unknown variance σ^2 . After the first n observations it is suspected that there is a sudden change in the mean of the distribution. Derive a test statistic for testing the hypothesis that the $(n + 1)$ th observation has the same population mean as the previous observations.

5

Confidence Intervals and Regions

5.1 SIMULTANEOUS INTERVAL ESTIMATION

5.1.1 Simultaneous Inferences

We begin with the full-rank model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where \mathbf{X} is $n \times p$ of rank p . A common statistical problem for such a model is that of finding two-sided confidence intervals for k linear combinations $\mathbf{a}'_j \beta$ ($j = 1, 2, \dots, k$). One solution would simply be to write down k t -intervals of the form given in (4.18) of Section 4.3.3, namely,

$$\mathbf{a}'_j \hat{\beta} \pm t_{n-p}^{(1/2)\alpha} \hat{\sigma}_{\mathbf{a}'_j \hat{\beta}}. \quad (5.1)$$

A typical application of this would be to write $\mathbf{a}'_1 = (1, 0, \dots, 0)$, $\mathbf{a}'_2 = (0, 1, \dots, 0)$, etc., and $k = p$, so that we are interested in confidence intervals for all the β_j ($j = 0, 1, \dots, p - 1$). The intervals above would then become

$$\hat{\beta}_j \pm t_{n-p}^{(1/2)\alpha} S d_{jj}^{1/2}, \quad (5.2)$$

where d_{jj} is the $(j + 1)$ th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ (see Example 4.6). For $j = 1, \dots, p - 1$ we note that d_{jj} is also the j th diagonal element of \mathbf{V}^{-1} , where \mathbf{V} is given by equation (4.15).

If we attach a probability of $1 - \alpha$ to each separate interval, as we have done above, the overall probability that the confidence statements are true *simultaneously* is, unfortunately, not $1 - \alpha$. To see this, suppose that E_j ($j = 1, 2, \dots, k$) is the event that the j th statement is correct, and let $\text{pr}[E_j] =$

$1 - \alpha_j$. If \bar{E}_j denotes the complementary event of E_j , then

$$\begin{aligned} 1 - \delta &= \text{pr} \left[\bigcap_{j=1}^k E_j \right] = 1 - \text{pr} \left[\overline{\bigcap_j E_j} \right] = 1 - \text{pr} \left[\bigcup_j \bar{E}_j \right] \\ &\geq 1 - \sum_{j=1}^k \text{pr}[\bar{E}_j] = 1 - \sum_{j=1}^k \alpha_j. \end{aligned} \quad (5.3)$$

Here δ , the probability of getting at least one statement wrong, is referred to variously as the *probability of a nonzero family error rate*, the *abbreviated probability error rate*, (Miller [1981: p. 6]), the *familywise error rate* (FWE, Hochberg and Tamhane [1987: p. 7]), or the *experimentwise error rate* (Tukey [1953]). For the case $\alpha_j = \alpha$ ($j = 1, 2, \dots, k$),

$$\text{pr} \left[\bigcap_{j=1}^k E_j \right] \geq 1 - k\alpha, \quad (5.4)$$

so that the probability of all the statements being correct is not $1 - \alpha$ but something greater than $1 - k\alpha$. For example, if $\alpha = 0.05$ and $k = 10$, then $1 - k\alpha = 0.5$. Furthermore, as pointed out by Miller [1977: p. 779; 1981: p. 8], the inequality (5.4) is surprisingly sharp: It is not as crude as one might expect, provided that k is not too large (say, $k \leq 5$) and α is small, say, 0.01.

It is also worth noting that

$$\begin{aligned} \text{pr} \left[\bigcap_j E_j \right] &= \text{pr}[E_1] \text{pr}[E_2 | E_1] \cdots \text{pr}[E_k | E_1, \dots, E_{k-1}] \\ &\simeq \text{pr}[E_1] \text{pr}[E_2] \cdots \text{pr}[E_k] \\ &= (1 - \alpha_1)(1 - \alpha_2) \cdots (1 - \alpha_k) \end{aligned} \quad (5.5)$$

if the dependence between the events E_i is small. As we shall see below, (5.5) can sometimes provide a lower bound for $\text{pr}[\bigcap_j E_j]$ (Miller [1977: p. 780]). Other related probability inequalities are given by Hochberg and Tamhane [1987: Appendix 2]; for a general review, see Tong [1980].

There is one other problem associated with the E_j . If $\alpha_j = 0.05$ ($j = 1, 2, \dots, k$), there is 1 chance in 20 of making an incorrect statement about $\alpha'_j \beta$, so that for every 20 statements made we can expect 1 to be incorrect. In other words, 5% of our k confidence intervals can be expected to be unreliable; there is an expected error rate of 1 in 20.

A number of authors (cf. Hochberg and Tamhane [1987: pp. 9–11]) recommend that δ should be the quantity to control in any given multiple-comparison situation. When k is finite, Spjøtvoll [1972] suggested that $\gamma = \sum_j \alpha_j$ should be controlled, where γ is the expected number of incorrect statements (see Miscellaneous Exercises 5, No. 1). Hochberg and Tamhane

[1987: pp. 9–12] discuss the relative merits of controlling δ , γ , or γ/k ; it depends on whether k is infinite or not and whether the focus is exploratory or explanatory. It turns out that $\gamma/k \leq \delta \leq \gamma$ [cf. (5.3)].

We now consider several ways of avoiding some of the problems mentioned above.

Bonferroni t -Intervals

If we use an individual significance level of α/k instead of α (i.e., use $t_{n-p}^{\alpha/(2k)}$) for each of the k confidence intervals, then, from (5.4),

$$\text{pr} \left[\bigcap_{j=1}^k E_j \right] \geq 1 - k \left(\frac{\alpha}{k} \right) = 1 - \alpha, \quad (5.6)$$

so that the overall probability is at least $1 - \alpha$. However, a word of caution. When k is large, this method could lead to confidence intervals that are so wide as to be of little practical use. This means that a reasonable compromise may be to increase α : for example, use $\alpha = 0.10$.

To use the method above we frequently require significance levels for the t -distribution which are not listed in the common t -tables. The following approximation due to Scott and Smith [1970] may therefore be useful:

$$t_\nu^\alpha \approx z_\alpha \left(1 - \frac{z_\alpha^2 + 1}{4\nu} \right)^{-1},$$

where z_α denotes the upper α point of the $N(0, 1)$ distribution. Statistical packages (e.g., S-PLUS and R) generally provide t_ν^α for any α . Hochberg and Tamhane [1987: Appendix 3, Table 1] give an extensive table for small values of α together with rules for interpolation. A table of $t_\nu^{\alpha/(2k)}$ for $\alpha = 0.05, 0.01$; $k = 2(1)10(5)50, 100, 250$; $\nu = 5, 7, 10, 12, 15, 20, 24, 30, 40, 60, 120, \infty$ is given in Appendix C.1.

The intervals described above based on replacing α by α/k are called *Bonferroni t -intervals*, as (5.3) is a Bonferroni inequality (Feller [1968: p. 110]). The corresponding tests are called *Bonferroni tests* and a number of modifications of such tests have been proposed (see Rencher [1998: Section 3.4.5] for a brief summary).

Maximum Modulus t -Intervals

Let $u_{k,\nu,\rho}^\alpha$ be the upper-tail α significance point of the distribution of the maximum absolute value of k Student t -variables, each based on ν degrees of freedom and having a common pairwise correlation ρ (these variables have a joint multivariate t -distribution—see A.13.5); when $\rho = 0$, we simply denote this point by $u_{k,\nu}^\alpha$. Now if the $\mathbf{a}_j' \hat{\boldsymbol{\beta}}$ ($j = 1, 2, \dots, k$) are mutually independent ($k \leq p$) and also independent of S^2 (as $\hat{\boldsymbol{\beta}}$ is independent of S^2 by Theorem

3.5(iii)), the pairwise covariances of the t -variables

$$\begin{aligned} T_j &= \frac{\mathbf{a}'_j \hat{\beta} - \mathbf{a}'_j \beta}{\hat{\sigma}_{\mathbf{a}'_j \hat{\beta}}} \\ &= \frac{\mathbf{a}'_j \hat{\beta} - \mathbf{a}'_j \beta}{S \{ \mathbf{a}'_j (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_j \}^{1/2}}, \end{aligned}$$

conditional on S^2 , are zero. Since $E[T_j | S^2 = 0]$, it follows from the Miscellaneous Exercises 5, No. 3 (with $X = T_i$, $Y = T_j$, and $Z = S^2$) that the unconditional covariances (and correlations) are also zero. Hence

$$\begin{aligned} 1 - \alpha &= \text{pr} \left[\max_{1 \leq j \leq k} |T_j| \leq u_{k,n-p}^\alpha \right] \\ &= \text{pr} [|T_j| \leq u_{k,n-p}^\alpha, \text{ all } j], \end{aligned}$$

and the set of k intervals

$$\mathbf{a}'_j \hat{\beta} \pm u_{k,n-p}^\alpha \hat{\sigma}_{\mathbf{a}'_j \hat{\beta}} \quad (5.7)$$

will have an overall confidence probability of *exactly* $1 - \alpha$. Thus $\delta = \alpha$. However, if the $\mathbf{a}'_j \hat{\beta}$ are not independent, which is the more usual situation, then the intervals (5.7) can still be used, but they will be conservative; the overall probability will be at least $1 - \alpha$. (This result follows from a theorem by Sidak [1968]; see Hahn and Hendrickson [1971] and Hahn [1972].) We note in passing that if Bonferroni t -intervals are used with α/k instead of α [as in (5.6)], then (5.5) becomes a lower bound (Miller [1977: p. 780]), so that

$$\text{pr}[\cap_i E_i] \geq \left(1 - \frac{\alpha}{k}\right)^k > (1 - \alpha),$$

which is a slight improvement on the Bonferroni inequality. However, these Bonferroni intervals won't be as narrow as those given by (5.7).

Hahn [1972] showed that when $k = 2$, the intervals

$$\mathbf{a}'_i \hat{\beta} \pm u_{k,n-p,\rho}^\alpha \hat{\sigma}_{\mathbf{a}'_i \hat{\beta}} \quad (i = 1, 2),$$

where ρ , the correlation between $\mathbf{a}'_1 \hat{\beta}$ and $\mathbf{a}'_2 \hat{\beta}$, is given by

$$\rho = \frac{\mathbf{a}'_1 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_2}{\{ \mathbf{a}'_1 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_1 \mathbf{a}'_2 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_2 \}^{1/2}}, \quad (5.8)$$

have an exact overall probability of $1 - \alpha$. This result can be used in straight-line regression (see Chapter 6).

A table of $u_{k,\nu,\rho}^\alpha$ for $\alpha = 0.05, 0.01$; $k = 1(1)6, 8, 10, 12, 15, 20$; $\nu = 3(1)12, 15, 20, 25, 30, 40, 60$; and $\rho = 0.0, 0.2, 0.4$, and 0.5 is given in Appendix C.2. Hochberg and Tamhane [1987: Appendix 3, Tables 4 and 7] give a slightly

more extensive table for $\rho = 0.1, 0.3, 0.5$ and 0.7 , and an extensive table for $\rho = 0$.

Scheffé's S-Method

We may assume without loss of generality that the first d vectors of the set $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k\}$ are linearly independent, and the remaining vectors (if any) are linearly dependent on the first d vectors; thus $d \leq \min(k, p)$. Consider the $d \times p$ matrix \mathbf{A} , where $\mathbf{A}' = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d)$, and let $\phi = \mathbf{A}\beta$. Now \mathbf{A} is a $d \times p$ matrix of rank d so that using the same argument as that given in proving Theorem 4.1(iii) and setting $\hat{\phi} = \mathbf{A}\hat{\beta}$, we have

$$\frac{(\hat{\phi} - \phi)' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\hat{\phi} - \phi)}{dS^2} \sim F_{d,n-p}. \quad (5.9)$$

Setting $\mathbf{L} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$, it now follows that

$$\begin{aligned} 1 - \alpha &= \text{pr}[F_{d,n-p} \leq F_{d,n-p}^\alpha] \\ &= \text{pr}\left[(\hat{\phi} - \phi)' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\hat{\phi} - \phi) \leq dS^2 F_{d,n-p}^\alpha\right] \\ &= \text{pr}\left[(\hat{\phi} - \phi)' \mathbf{L}^{-1} (\hat{\phi} - \phi) \leq m\right], \quad \text{say,} \end{aligned} \quad (5.10)$$

$$\begin{aligned} &= \text{pr}[\mathbf{b}' \mathbf{L}^{-1} \mathbf{b} \leq m] \\ &= \text{pr}\left[\sup_{\mathbf{h} \neq 0} \left\{ \frac{(\mathbf{h}' \mathbf{b})^2}{\mathbf{h}' \mathbf{L} \mathbf{h}} \right\} \leq m\right] \quad (\text{by A.4.11}) \\ &= \text{pr}\left[\frac{(\mathbf{h}' \mathbf{b})^2}{\mathbf{h}' \mathbf{L} \mathbf{h}} \leq m, \text{ all } \mathbf{h} (\neq 0)\right] \\ &= \text{pr}\left[\frac{|\mathbf{h}' \hat{\phi} - \mathbf{h}' \phi|}{S(\mathbf{h}' \mathbf{L} \mathbf{h})^{1/2}} \leq (dF_{d,n-p}^\alpha)^{1/2}, \text{ all } \mathbf{h} (\mathbf{h} \neq 0)\right]. \end{aligned} \quad (5.11)$$

We can therefore construct a confidence interval for *any* linear function $\mathbf{h}'\phi$, namely,

$$\mathbf{h}'\hat{\phi} \pm (dF_{d,n-p}^\alpha)^{1/2} S(\mathbf{h}' \mathbf{L} \mathbf{h})^{1/2}, \quad (5.12)$$

and the overall probability for the entire class of such intervals is exactly $1 - \alpha$. The term $S^2 \mathbf{h}' \mathbf{L} \mathbf{h}$ involved in the calculation of (5.12) is simply an unbiased estimate of $\text{var}[\mathbf{h}'\hat{\phi}]$; frequently, the latter expression can be found directly without the need for matrix inversion (e.g., see Section 8.2.2). The interval (5.12) can therefore be written in the more compact form

$$\mathbf{h}'\hat{\phi} \pm (dF_{d,n-p}^\alpha)^{1/2} \hat{\sigma}_{\mathbf{h}'\hat{\phi}}. \quad (5.13)$$

Since $\mathbf{h}'\phi = \phi_j$ for certain \mathbf{h} , we see that a confidence interval every $\mathbf{a}_j'\beta = \phi_j$ ($j = 1, 2, \dots, d$) is included in the set of intervals (5.13). In addition, an interval for every ϕ_j ($j = d+1, d+2, \dots, k$) is also included in this set, owing to the linear dependence of the \mathbf{a}_j ($j = d+1, \dots, k$) on the other \mathbf{a}_j 's. For

example, if $\mathbf{a}_{d+1} = h_1 \mathbf{a}_1 + \cdots + h_d \mathbf{a}_d$, then $\phi_{d+1} = \mathbf{a}'_{d+1} \beta = \sum_{j=1}^d h_j \phi_j = \mathbf{h}'\phi$. Therefore, if E_j is the event that $\mathbf{a}'_j \beta$ lies in the interval

$$\mathbf{a}'_j \hat{\beta} \pm (dF_{d,n-p}^\alpha)^{1/2} \hat{\sigma}_{\mathbf{a}'_j \hat{\beta}}, \quad (5.14)$$

then since the complete set of intervals (5.13) is more than the intervals for ϕ_j ($j = 1, 2, \dots, k$) that we asked for,

$$\text{pr} \left[\bigcap_{j=1}^k E_j \right] \geq 1 - \alpha.$$

We note that the class of parametric functions $\mathbf{h}'\phi$ form a linear space, \mathcal{L} say, with basis $\phi_1, \phi_2, \dots, \phi_d$. In fact, \mathcal{L} is the smallest linear space containing the k functions ϕ_j ($j = 1, 2, \dots, k$). The method above is due to Scheffé [1953] and is called the *S-method of multiple comparisons* in his book (Scheffé [1959: p. 68]). Other methods for constructing simultaneous confidence intervals for special subsets of \mathcal{L} are discussed in Section 8.2.2. For general references on the subject of multiple comparisons, the reader is referred to Miller [1977, 1981], Hochberg and Tamahane [1987], and Hsu [1996]. The class of linear functions \mathcal{L} of the form $\mathbf{h}'\phi$ ($= \mathbf{h}'\mathbf{A}\beta$) is only a subclass of all possible linear functions $\mathbf{a}'\beta$, where \mathbf{a} is now any $p \times 1$ vector. However, setting $d = k = p$ and $\mathbf{A} = \mathbf{I}_p$, we have $\phi = \beta$, and the corresponding confidence intervals for the class of all functions $\mathbf{h}'\beta$ take the form [cf. (5.13)]

$$\mathbf{h}'\hat{\beta} \pm (pF_{p,n-p}^\alpha)^{1/2} \hat{\sigma}_{\mathbf{h}'\hat{\beta}}. \quad (5.15)$$

5.1.2 Comparison of Methods

For k confidence intervals, the Bonferroni t -intervals, the maximum modulus t -intervals (5.7), and Scheffé's S -intervals (5.14) all give a lower bound of $1 - \alpha$ for $\text{pr}[\bigcap_j E_j]$. By comparing Tables 5.1 and 5.2 we see that for $\alpha = 0.05$, $d \leq k$, and k not much greater than d ,

$$t_\nu^{\alpha/(2k)} < (dF_{d,\nu}^\alpha)^{1/2}. \quad (5.16)$$

When k is much greater than d , the reverse inequality holds. Also, it can be shown theoretically (compare Table 5.1 and Appendix C.2) that

$$u_{k,\nu}^\alpha < t_\nu^{\alpha/(2k)}, \quad (5.17)$$

so that for the common situation of $d = k$ (i.e., no "redundant" confidence intervals), the maximum modulus intervals are the shortest and the F -intervals are the widest. For example, when $\alpha = 0.05$, $d = k = 5$, $p = 6$, and $n = 26$, we have

$$\nu = 20, \quad (kF_{k,\nu}^\alpha)^{1/2} = 3.68, \quad t_\nu^{\alpha/(2k)} = 2.85, \quad \text{and} \quad u_{k,\nu}^\alpha = 2.82.$$

If we were interested in just a single t -interval, we would use $t_\nu^{(1/2)\alpha} = 2.09$, which is much smaller than the previous three numbers.

Table 5.1 Values of $t_{\nu}^{\alpha/(2k)}$ for $\alpha = 0.05$

$\nu \setminus k$	1	2	3	4	5	6	7	8	9	10	15	20	50
5	2.57	3.16	3.54	3.81	4.04	4.22	4.38	4.53	4.66	4.78	5.25	5.60	6.87
10	2.23	2.64	2.87	3.04	3.17	3.28	3.37	3.45	3.52	3.58	3.83	4.01	4.59
15	2.13	2.49	2.70	2.84	2.95	3.04	3.11	3.18	3.24	3.29	3.48	3.62	4.08
20	2.09	2.42	2.61	2.75	2.85	2.93	3.00	3.06	3.11	3.16	3.33	3.46	3.85
24	2.07	2.39	2.58	2.70	2.80	2.88	2.94	3.00	3.05	3.09	3.26	3.38	3.75
30	2.04	2.36	2.54	2.66	2.75	2.83	2.89	2.94	2.99	3.03	3.19	3.30	3.65
40	2.02	2.33	2.50	2.62	2.70	2.78	2.84	2.89	2.93	2.97	3.12	3.23	3.55
60	2.00	2.30	2.47	2.58	2.66	2.73	2.79	2.84	2.88	2.92	3.06	3.16	3.46
120	1.98	2.27	2.43	2.54	2.62	2.68	2.74	2.79	2.83	2.86	3.00	3.09	3.38
∞	1.96	2.24	2.40	2.50	2.58	2.64	2.69	2.74	2.78	2.81	2.94	3.03	3.29

SOURCE : Dunn [1959]. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright (1959) by the American Statistical Association. All rights reserved.

Table 5.2 Values of $(dF_{d,\nu}^{\alpha})^{1/2}$ for $\alpha = 0.05$

$\nu \setminus d$	1	2	3	4	5	6	7	8
5	2.57	3.40	4.03	4.56	5.02	5.45	5.84	6.21
10	2.23	2.86	3.34	3.73	4.08	4.40	4.69	4.96
15	2.13	2.71	3.14	3.50	3.81	4.09	4.36	4.60
20	2.09	2.64	3.05	3.39	3.68	3.95	4.19	4.43
24	2.06	2.61	3.00	3.34	3.62	3.88	4.12	4.34
30	2.04	2.58	2.96	3.28	3.56	3.81	4.04	4.26
40	2.02	2.54	2.92	3.23	3.50	3.75	3.97	4.18
60	2.00	2.51	2.88	3.18	3.44	3.67	3.90	4.10
120	1.98	2.48	2.84	3.13	3.38	3.62	3.83	4.02
∞	1.96	2.45	2.79	3.08	3.32	3.55	3.75	3.94

SOURCE : Dunn [1959]. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright (1959) by the American Statistical Association. All rights reserved.

5.1.3 Confidence Regions

Suppose that $d = k$. Then from (5.10) we have

$$1 - \alpha = \text{pr} \left[(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathbf{L}^{-1} (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) \leq m \right],$$

where the region $(\phi - \hat{\phi})' \mathbf{L}^{-1} (\phi - \hat{\phi}) \leq m$ is a solid ellipsoid with center $\hat{\phi}$, since $\mathbf{L} [= \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{A}']$, and therefore \mathbf{L}^{-1} , is positive definite. This ellipsoid gives us a $100(1 - \alpha)\%$ confidence *region* for ϕ . However, unless k is small, say, 2 or 3, such a region will not be computed readily, nor interpreted easily. In this respect, suitable contour lines or surfaces may be sufficient to give a reasonable description of the region. For example, if $k = 3$, the region may be pictured in two dimensions by means of a contour map as in Figure 5.1; here we have a plot of ϕ_1 versus ϕ_2 for three values of ϕ_3 . For $k > 3$ it is still possible to convey the general shape of the confidence region by using a set of contour maps. However, generally speaking, the contour region approach is of limited value.

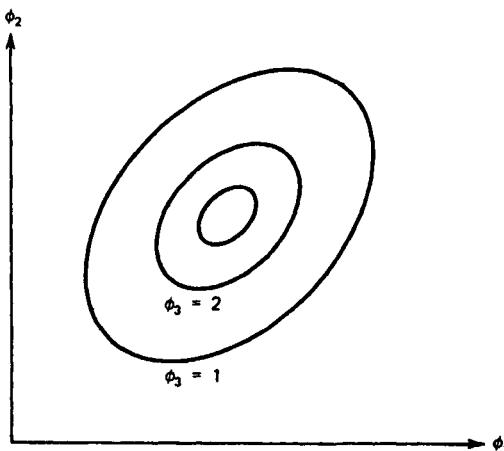


Fig. 5.1 Contour map of the confidence region for $\phi' = (\phi_1, \phi_2, \phi_3)$.

If our focus of interest is just β , which will usually be the case, we can set $\mathbf{A} = \mathbf{I}_p$ and $\phi = \beta$ in the preceding equation to get

$$1 - \alpha = \text{pr} \left[(\beta - \hat{\beta})' \mathbf{X}' \mathbf{X} (\beta - \hat{\beta}) \leq p S^2 F_{p,n-p}^\alpha \right], \quad (5.18)$$

a confidence ellipsoid for β .

EXAMPLE 5.1 All the conservative confidence intervals for the β_j [cf. (5.2)] take the form $\hat{\beta}_j \pm t S d_{jj}^{1/2}$, where t is one of $t_\nu^{\alpha/2p}$, $(p F_{p,\nu}^\alpha)^{1/2}$, or $u_{p,\nu}^\alpha$. We shall derive a formula for comparing the rectangular volume R contained within the joint confidence intervals for all of the β_j , derived from one of the conservative methods, with the volume E of the confidence ellipsoid for β .

First,

$$R = 2^p S^p t^p \left(\prod_{j=0}^{p-1} d_{jj} \right)^{1/2},$$

where d_{jj} is the $(j+1)$ th diagonal element of $\mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1}$. Second, the p -dimensional ellipsoid

$$(\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) = pS^2 F_{p,n-p}^\alpha = c,$$

say, has volume (cf. Seber and Wild [1989: p. 679])

$$E = \frac{\pi^{p/2}}{\Gamma(p/2 + 1)} \prod_{j=0}^{p-1} a_j,$$

where the a_j are the lengths of the semimajor axes. For $p = 2$, we can compare $\lambda_1 x_1^2 + \lambda_2 x_2^2 = c$ with the standardized form $(x_1^2/a_1^2) + (x_2^2/a_2^2) = 1$ and we see that $a_j = (c/\lambda_j)^{1/2}$, where λ_j is an eigenvalue of $\mathbf{X}'\mathbf{X}$. Thus, for general p ,

$$E = \frac{\pi^{p/2}}{\Gamma(p/2 + 1)} c^{p/2} |\mathbf{D}|^{1/2},$$

since $|\mathbf{X}'\mathbf{X}| = \prod_j \lambda_j$ (by A.1.3).

Comparing E and R , we have

$$\frac{E}{R} = \frac{\pi^{p/2} p^{p/2}}{2^p \Gamma(p/2 + 1)} \frac{(F_{p,n-p}^\alpha)^{p/2}}{t^p} \frac{|\mathbf{D}|^{1/2}}{(\Pi_j d_{jj})^{1/2}}. \quad (5.19)$$

Draper and Smith [1998: p. 143] express $|\mathbf{D}|^{1/2}(\Pi_j d_{jj})^{-1/2}$ in the form $|\mathbf{W}|^{1/2}$, where $w_{ij} = d_{ij}/(d_{ii}d_{jj})^{1/2}$. \square

5.1.4 Hypothesis Testing and Confidence Intervals

An interesting relationship exists between the set of confidence intervals (5.12) and the F -statistic for testing the hypothesis $H: \phi = \mathbf{c}$. From (5.9) we see that the F -statistic is not significant at the α level of significance if and only if

$$F = \frac{(\hat{\phi} - \mathbf{c})' \mathbf{L}^{-1} (\hat{\phi} - \mathbf{c})}{dS^2} \leq F_{d,n-p}^\alpha,$$

which is true if and only if $\phi = \mathbf{c}$ is contained in the region $(\phi - \hat{\phi})' \mathbf{L}^{-1} (\phi - \hat{\phi}) \leq m$ [by (5.10)], that is, if and only if $\mathbf{h}'\mathbf{c}$ is contained in (5.12) for every \mathbf{h} . Therefore, F is significant if one or more of the intervals (5.12) does not contain $\mathbf{h}'\mathbf{c}$, and the situation can arise where each interval for ϕ_i contains c_i ($i = 1, 2, \dots, k$) but H is rejected. For example, when $k = 2$ the separate intervals for ϕ_1 and ϕ_2 form the rectangle given in Figure 5.2, and the ellipse is the region $(\phi - \hat{\phi})' \mathbf{L}^{-1} (\phi - \hat{\phi}) \leq m$; a point \mathbf{c} that lies within the rectangle does not necessarily lie within the ellipse.

Usually, interval estimation is preceded by an F -test of some hypothesis $H: \mathbf{A}\beta = \mathbf{c}$. However, when a preliminary test is carried out, the appropriate probability to be considered is now the *conditional* probability $\text{pr}[\cap_i E_i | F]$

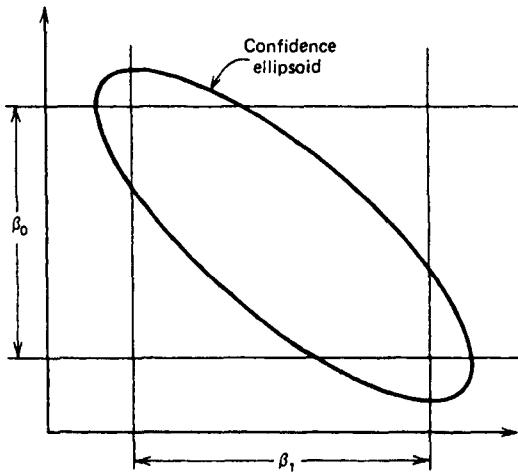


Fig. 5.2 Separate confidence intervals for β_0 and β_1 compared with a joint confidence region.

significant], which may be greater or less than the unconditional probability $\text{pr}[\cap_i E_i]$ (Olshen [1973]).

A null hypothesis is inevitably false, as, for example, two unknown parameters are never exactly equal. The question is whether there are sufficient data to detect the differences. Confidence intervals should therefore always be constructed.

EXAMPLE 5.2 Suppose that we test $H : \beta_1 = \beta_2 = \dots = \beta_d = 0$ ($d < p - 1$). We then should examine each β_j ($j = 1, 2, \dots, d$) separately using the confidence intervals $\hat{\beta}_j \pm t\hat{\sigma}_{\hat{\beta}_j}$ provided by any one of the three methods given above. However, the maximum modulus intervals would normally be preferred if they are the shortest. We hope that those intervals that do not contain zero will indicate which of the β_j are significantly different from zero, and by how much. We can also obtain intervals for all linear combinations $\sum_{i=1}^d a_i \beta_i$ using Scheffé's method. \square

EXAMPLE 5.3 Suppose that we wish to test $H : \beta_1 = \beta_2 = \dots = \beta_{d+1}$. Subsequent to the test, we will be interested in all $k [= d(d+1)/2]$ pairs $\beta_i - \beta_j$ ($i < j$). For example, if $d = 4$, $n - p = \nu = 20$, and $\alpha = 0.05$, then $k = 10$, $(dF_{d,\nu}^\alpha)^{1/2} = 3.39$, $t_\nu^{\alpha/(2k)} = 3.16$, and $u_{k,\nu}^\alpha = 3.114$, so that the maximum modulus intervals are still the shortest. Now H can also be written in the form $\phi_i = \beta_i - \beta_{d+1} = 0$ ($i = 1, 2, \dots, d$), so that Scheffé's method will provide confidence intervals for all linear combinations

$$\sum_{i=1}^d h_i \phi_i = \sum_{i=1}^d h_i \beta_i - \left(\sum_{i=1}^d h_i \right) \beta_{d+1} = \sum_{i=1}^{d+1} c_i \beta_i, \quad (5.20)$$

where $\sum_{i=1}^{d+1} c_i = 0$; thus every linear combination of the ϕ_i is a contrast in the β_i . By reversing the argument above we see that every contrast in the β_i is a linear combination of the ϕ_i . Hence Scheffé's method provides a set of multiple confidence intervals for all contrasts in the β_i ($i = 1, 2, \dots, d + 1$). \square

5.2 CONFIDENCE BANDS FOR THE REGRESSION SURFACE

5.2.1 Confidence Intervals

Once we have estimated β from n observations \mathbf{Y} , we can use the predictor

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{p-1} x_{p-1} \quad (= \mathbf{x}' \hat{\beta}, \text{ say})$$

for studying the shape of the regression surface

$$f(x_1, x_2, \dots, x_{p-1}) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1} = \mathbf{x}' \beta$$

over a range of values of the regressors x_j . In particular, we can construct a two-sided $100(1 - \alpha)\%$ confidence interval for the value of f at a particular value of \mathbf{x} , say, $\mathbf{x}_0 = (1, x_{01}, x_{02}, \dots, x_{0,p-1})'$, using $\hat{Y}_0 = \mathbf{x}'_0 \hat{\beta}$. Thus from (4.17), we have the interval

$$\hat{Y}_0 \pm t_{n-p}^{(1/2)\alpha} S \sqrt{v_0}, \quad (5.21)$$

where $v_0 = \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0$.

If we are interested in k particular values of \mathbf{x} , say, $\mathbf{x} = \mathbf{a}_j$ ($j = 1, 2, \dots, k$), then we can use any of the three methods discussed in Section 5.1 to obtain k two-sided confidence intervals for the $\mathbf{a}'_j \beta$ with a joint confidence probability of at least $1 - \alpha$. (Application of the Bonferroni and the Scheffé intervals to this problem seems to be due to Lieberman [1961].)

5.2.2 Confidence Bands

If we are interested in *all* values of \mathbf{x} , then using Scheffé's method we have from (5.15) that $\mathbf{x}' \beta$ lies in

$$\mathbf{x}' \hat{\beta} \pm (p F_{p,n-p}^\alpha)^{1/2} S \{ \mathbf{x}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x} \}^{1/2} \quad (5.22)$$

for all $\mathbf{x} = (1, x_1, x_2, \dots, x_{p-1})'$, with an exact overall probability of $1 - \alpha$. (Although the first element of \mathbf{x} is constrained to be unity, this does not mean that the appropriate constant in (5.22) should now be $[(p-1) F_{p-1,n-p}^\alpha]^{1/2}$; the interval is invariant under a scale change of one element of \mathbf{x} ; cf. Miller [1981: pp. 110–114]). The expression above gives two surfaces defined by the functions f^0 and f_0 , where

$$\begin{aligned} \text{pr}[f^0(x_1, x_2, \dots, x_{p-1}) &\geq f(x_1, x_2, \dots, x_{p-1}) \\ &\geq f_0(x_1, x_2, \dots, x_{p-1}), \text{ all } x_1, x_2, \dots, x_{p-1}] \\ &= 1 - \alpha. \end{aligned}$$

The region between f^0 and f_0 is commonly called a *confidence band*. As pointed out by Miller [1981], the band over that part of the regression surface that is not of interest, or is physically meaningless, is ignored. This means that the probability associated with the regression band over a limited region exceeds $1 - \alpha$, and the intervals given by (5.22) will be somewhat conservative. The question of constructing a confidence band over a limited region, with an *exact* probability $1 - \alpha$, is discussed by Wynn and Bloomfield [1971]. A solution is given by Halperin and Gurian [1968] for the case of an ellipsoidal region centered on the vector of means $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{p-1})$. Various solutions for the straight-line case are considered in detail in Section 6.1.3.

Scheffé's method described by (5.22) is a special case of a more general procedure developed by Bowden [1970]. Let

$$\begin{aligned}\|\mathbf{a}\|_m &= \left(\sum_{i=1}^p |a_i|^m \right)^{1/m} \quad 1 \leq m < \infty \\ &= \max_i |a_i| \quad m = \infty;\end{aligned}$$

then, as Bowden proves,

$$\Pr \left[|\mathbf{x}'\hat{\beta} - \mathbf{x}'\beta| \leq S\|\mathbf{x}\|_m z_m^\alpha, \text{ all } \mathbf{x} \right] = 1 - \alpha, \quad (5.23)$$

where z_m^α is the upper α significant point of the distribution of $\|(\hat{\beta} - \beta)/S\|_m$. By taking $m = 1, 2$, or ∞ and varying the value of \mathbf{x} , different types of regression bands can be obtained; when $p = 2$ (the straight-line case), the band has uniform or trapezoidal width ($m = 1$), or is hyperbolic ($m = 2$), or is bounded by straight-line segments ($m = \infty$). However, it transpires that for $p > 2$, Scheffé's method ($m = 2$) and its corresponding one-sided analog have certain optimal properties (Bohrer [1973]). When k is large it is natural to ask whether the maximum modulus t -intervals of (5.7) are still shorter than the intervals given by the confidence band of (5.22), particularly when k is much greater than p . Hahn [1972] has calculated

$$r = \frac{u_{k,n-p}^\alpha}{(pF_{p,n-p}^\alpha)^{1/2}}, \quad (5.24)$$

the ratio of the interval widths, for $\alpha = 0.1, 0.05, 0.01$, and for different values of k , p , and $n - p$. Table 5.3 gives the maximum value of k (for $\alpha = 0.05$, $p = 2, 3, 5$ and $n - p = 5, 10, 20, 40, 60$) for which $r < 1$. Hahn also found that, for these values of α , r increased slightly as α decreased.

Sometimes a model consists of several regression models combined together using dummy explanatory variables. For example, suppose that we have J straight lines

$$E[Y_j] = \alpha_j + \gamma_j \mathbf{x}, \quad (j = 1, 2, \dots, J),$$

and for all $\mathbf{x} [= (1, \mathbf{x})']$ and all j ($j = 1, \dots, J$), we want to construct simultaneous confidence intervals for $\mathbf{x}'\beta_j$ [$\beta_j = (\alpha_j, \gamma_j)'$]. An obvious method

Table 5.3 Maximum Value of k for which $r < 1$ [r is defined by equation (5.24)], $\alpha = 0.05$.

$n - p$	$p = 2$	$p = 3$	$p = 5$
5	3	6	20+
10	3	8	20+
20	3	8	20+
40	3	9	20+
60	3	9	20+

SOURCE : Hahn [1972: Table 3]. Reprinted with permission from *Technometrics*. Copyright (1972) by the American Statistical Association. All rights reserved.

would be to allocate α/J to each model, as with the Bonferroni method, and then apply Scheffé's method to each of the J individual models. We can write each individual model as a regression model with regression matrix \mathbf{X}_0 and p_0 ($= 2$) parameters and then combine them using dummy explanatory variables, as demonstrated by equation (1.2), to obtain a regression model with p ($= 2J$) parameters. Then, for the combined model, S^2 is independent of each $\hat{\beta}_j$, so that we can use S^2 to obtain a confidence ellipsoid for β_j . Thus from (5.22) it follows that $\mathbf{x}'\beta_j$ lies in

$$\mathbf{x}'\hat{\beta}_j \pm S\{p_0 \mathbf{x}'(\mathbf{X}'_0\mathbf{X}_0)^{-1}\mathbf{x} F_{p_0, n-p}^{\alpha/J}\}^{1/2} \text{ for all } \mathbf{x} \in \mathfrak{R}_{p_0}, x_0 \equiv 1, \quad (5.25)$$

with probability at least $1 - \alpha/J$. Using (5.3) we can combine all J models to obtain a probability of at least $1 - \sum \alpha/J$ ($= 1 - \alpha$) that the statement above is true for all $j = 1, 2, \dots, J$ with probability at least $(1 - \alpha)$. This method is essentially that proposed by Lane and DuMouchel [1994]. They compare the intervals given above with those obtained by the less efficient procedure of using the combined model and constructing Scheffé intervals for all $\mathbf{x}'\beta$, where now $\mathbf{x} \in \mathfrak{R}_p$.

5.3 PREDICTION INTERVALS AND BANDS FOR THE RESPONSE

5.3.1 Prediction Intervals

In the preceding section we discussed the problem of predicting the value of a regression surface $\mathbf{x}'\beta$ at a given value of $\mathbf{x} = \mathbf{x}_0$, say. However, in practice, we are generally more interested in predicting the value, Y_0 , say, of the random variable Y , where

$$Y_0 = \mathbf{x}'_0\beta + \varepsilon_0.$$

If we assume that $\varepsilon_0 \sim N(0, \sigma^2)$ and that ε_0 is independent of Y , then

$$\begin{aligned} E[\hat{Y}_0 - Y_0] &= \mathbf{x}'_0 \boldsymbol{\beta} - \mathbf{x}'_0 \boldsymbol{\beta} = 0, \\ \text{var}[\hat{Y}_0 - Y_0] &= \text{var}[\hat{Y}_0] + \text{var}[Y_0] \\ &= \sigma^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0 + \sigma^2 \\ &= \sigma^2 (v_0 + 1), \end{aligned} \quad (5.26)$$

say, and $(\hat{Y}_0 - Y_0) \sim N(0, \sigma^2(v_0 + 1))$. We can therefore construct a t -statistic and obtain the $100(1 - \alpha)\%$ confidence interval for Y_0 given by

$$\hat{Y}_0 \pm t_{n-p}^{(1/2)\alpha} S(v_0 + 1)^{1/2}, \quad (5.27)$$

which may be compared with the interval (5.21).

If we are interested in predicting Y for k different values of \mathbf{x} , say, $\mathbf{x} = \mathbf{a}_j$ ($j = 1, 2, \dots, k$), then we can use any of the three methods discussed in Section 5.1.1 to obtain k confidence intervals with an overall confidence probability of at least $1 - \alpha$. For $k = 2$, Hahn [1972] shows that the intervals

$$\hat{Y}_0^{(j)} \pm u_{2,n-p,\rho}^\alpha S(v_0^{(j)} + 1)^{1/2} \quad (j = 1, 2),$$

where $\hat{Y}_0^{(j)} = \mathbf{a}'_j \hat{\boldsymbol{\beta}}$, $v_0^{(j)} = \mathbf{a}'_j (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_j$, and

$$\rho = \frac{\mathbf{a}'_1 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{a}_2}{\{(v_0^{(1)} + 1)(v_0^{(2)} + 1)\}^{1/2}}$$

have an *exact* overall probability of $1 - \alpha$.

Fitted regressions are used for two types of prediction (Box [1966]). First, we may wish to predict Y in the future from passive observation of the x_j 's. We assume that the system is not interfered with, so that the regression model proposed is still appropriate in the future. Second, we want to discover how deliberate changes in the x_j 's will affect Y , with the intention of actually modifying the system to get a better value of Y . The need to distinguish between these two situations is borne out by the following example adapted from Box [1966]. In a chemical process it is found that undesirable frothing can be reduced by increasing the pressure (x_1); it is also known that the yield (Y) is unaffected directly by a change in pressure. The standard operating procedure then consists of increasing the pressure whenever frothing occurs. Suppose, however, that the frothing is actually caused by the presence of an unsuspected impurity (x_2), and that unknown to the experimenter, an increase in concentration of impurity causes an increase in frothing and a decrease in Y . If x_1 and x_2 are positively correlated because an increase in pressure causes an increase in impurity, then although Y is unaffected directly by changes in x_1 , there is a spurious negative correlation between Y and x_1 as Y and x_1 are both affected by x_2 , but in opposite directions. This means that there will be a significant regression of Y on x_1 , and the fitted regression

can be used for adequately predicting Y , provided that the system continues to run in the same fashion as when the data were recorded. However, this regression does not indicate the true causal situation. We are mistaken if we think that we can increase Y by decreasing x_1 .

5.3.2 Simultaneous Prediction Bands

Carlstein [1986] gives two examples where the k predictions mentioned in the preceding section are at *unknown* values of $\mathbf{x} = \mathbf{a}_j$ ($j = 1, 2, \dots, k$). In this case what is needed is a prediction band for each of k future values of Y . Writing $Y^{(j)} = Y^{(j)}(\mathbf{x})$, Carlstein [1986] proves the following theorem.

THEOREM 5.1 *The event that $Y^{(j)}$ lies in the interval*

$$\mathbf{x}'\hat{\beta} \pm S\{(p+k)(1+\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})F_{p+k,n-p}^{\alpha}\}^{1/2}$$

for all $\mathbf{x} \in \mathfrak{R}_p$, with $x_0 \equiv 1$, and all $j = 1, 2, \dots, k$ has probability at least $(1 - \alpha)$.

Proof. Let $Y^{(j)}(\mathbf{x}) = \mathbf{x}'\beta + \varepsilon_{n+j}$ ($j = 1, \dots, k$), where the ε_{n+j} are independent of the n initial observations \mathbf{Y} . Let $\varepsilon'_0 = (\varepsilon_{n+1}, \dots, \varepsilon_{n+k})$, $\mathbf{b}' = (\hat{\beta}' - \beta', \varepsilon'_0)$, and

$$\mathbf{W} = \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}.$$

We recall that $S^2 = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2/(n-p)$ is independent of $\hat{\beta}$, and therefore of \mathbf{b} , where $\mathbf{b} \sim N_{p+k}(\mathbf{0}, \sigma^2 \mathbf{W})$. Also, $\mathbf{b}'\mathbf{W}^{-1}\mathbf{b}/\sigma^2 \sim \chi^2_{p+k}$, so that

$$\mathbf{b}'\mathbf{W}^{-1}\mathbf{b}/(p+k)S^2 \sim F_{p+k,n-p}.$$

Then arguing as in the theory leading to equation (5.11) yields

$$\begin{aligned} 1 - \alpha &= \text{pr}\left\{ \mathbf{b}'\mathbf{W}^{-1}\mathbf{b} \leq S[(p+k)F_{p+k,n-p}^{\alpha}]^{1/2} \right\} \\ &= \text{pr}\left\{ \frac{|\mathbf{h}'\mathbf{b}|}{(\mathbf{h}'\mathbf{W}\mathbf{h})^{1/2}} \leq S[(p+k)F_{p+k,n-p}^{\alpha}]^{1/2} \text{ for all } \mathbf{h} \in \mathfrak{R}_{p+k}, \mathbf{h} \neq \mathbf{0} \right\}. \end{aligned}$$

Now consider only those \mathbf{h} such that $\mathbf{h}' = (\mathbf{x}', \delta_1, \delta_2, \dots, \delta_k)$, where $\mathbf{x} \in \mathfrak{R}_p$ is arbitrary and the δ_j 's are all zero except for a single $\delta_j = -1$ ($j = 1, 2, \dots, k$). Then

$$\mathbf{h}'\mathbf{b} = \mathbf{x}'\hat{\beta} - Y^{(j)}(\mathbf{x}) \quad \text{and} \quad \mathbf{h}'\mathbf{W}\mathbf{h} = 1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}.$$

The result follows by noting that we are only looking at a subset of the possible vectors \mathbf{h} , so that $(1 - \alpha)$ is now a lower bound. \square

Carlstein [1986] also gives an alternative method by obtaining separate confidence intervals for the two components $\mathbf{x}'\beta$ and ε_{n+j} of $Y^{(j)}$, as in the following theorem.

THEOREM 5.2 Let $0 < \tilde{\alpha} < \alpha$. The event that $Y^{(j)}$ lies in the interval

$$\mathbf{x}'\hat{\beta} \pm S\{[p\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}]F_{p,n-p}^{\tilde{\alpha}}\}^{1/2} + [kF_{k,n-p}^{\alpha-\tilde{\alpha}}]^{1/2}\}$$

for all $\mathbf{x} \in \mathfrak{R}_p$, with $x_0 \equiv 1$, and all $j = 1, 2, \dots, k$ has probability at least $(1 - \alpha)$.

Proof. We use a Bonferroni argument and allocate $\tilde{\alpha}$ to a confidence band for $\mathbf{x}'\beta$, and $\alpha - \tilde{\alpha}$ to k simultaneous intervals for the elements $\varepsilon_{n+1}, \dots, \varepsilon_{n+k}$ of ε_0 .

Now $\mathbf{x}'\beta$ lies in

$$\mathbf{x}'\beta \pm S\{p\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}F_{p,n-p}^{\tilde{\alpha}}\}^{1/2} \text{ for all } \mathbf{x} \in \mathfrak{R}_p, x_0 \equiv 1$$

with probability $(1 - \tilde{\alpha})$. Also, since $\varepsilon_0'\varepsilon_0/\sigma^2 \sim \chi_k^2$ and is independent of S^2 , we have $\varepsilon_0'\varepsilon_0/kS^2 \sim F_{k,n-p}$ and, from (5.11),

$$\begin{aligned} 1 - \alpha + \tilde{\alpha} &= \text{pr}\left(\frac{|\mathbf{a}'\varepsilon_0|}{\mathbf{a}'\mathbf{a}} \leq S\{kF_{k,n-p}^{\alpha-\tilde{\alpha}}\}^{1/2} \text{ for all } \mathbf{a} \in \mathfrak{R}_k\right) \\ &\geq \text{pr}(|\varepsilon_{n+j}| \leq S\{kF_{k,n-p}^{\alpha-\tilde{\alpha}}\}^{1/2} \text{ for all } j = 1, 2, \dots, k). \end{aligned}$$

The last equation follows by setting $\mathbf{a} = (1, 0, \dots, 0)$, etc.

The probability statements above are then combined by using $\text{pr}(A \cup B) \leq 1$, which leads to

$$\begin{aligned} \text{pr}(A \cap B) &\geq \text{pr}(A) + \text{pr}(B) - 1 \\ &= 1 - \tilde{\alpha} + 1 - (\alpha - \tilde{\alpha}) - 1 \\ &= 1 - \alpha. \end{aligned}$$

[This result is a special case of (5.3) with $k = 2$.] □

We note that $\tilde{\alpha}$ can be chosen to give the shortest intervals. Carlstein [1986] gives an example which demonstrates that neither of the two methods above is uniformly better than the other.

By noting that

$$\begin{aligned} 1 - \alpha + \tilde{\alpha} &= \text{pr}\left[\max_{j=1, \dots, k} \frac{|\varepsilon_{n+j}|}{S} \leq u_{k,n-p}^{\alpha-\tilde{\alpha}}\right] \\ &= \text{pr}[|\varepsilon_{n+j}| \leq Su_{k,n-p}^{\alpha-\tilde{\alpha}} \text{ for all } j = 1, 2, \dots, k], \end{aligned}$$

where $u_{k,n-p}^\alpha$ is defined prior to equation (5.7), Zimmerman [1987] obtained shorter intervals by replacing $(kF_{k,n-p}^{\alpha-\tilde{\alpha}})^{1/2}$ by the smaller value $u_{k,n-p}^{\alpha-\tilde{\alpha}}$ in the statement of Theorem 5.2.

A third method, proposed by Lane and DuMouchel [1994], is based on the fact that with $x_0 \equiv 1$,

$$\begin{aligned} \mathbf{x}'\hat{\beta} - Y^{(j)} &= \mathbf{x}'\hat{\beta} - \mathbf{x}'\beta - \varepsilon_{n+j} \\ &= \mathbf{x}'\mathbf{b}_j, \end{aligned}$$

where $\mathbf{b}_j = (\hat{\beta}_0 - \beta_0 - \varepsilon_{n+j}, \hat{\beta}_1 - \beta_1, \dots, \hat{\beta}_{p-1} - \beta_{p-1})'$. Replacing \mathbf{W} in Theorem 5.1 by $\text{Var}[\mathbf{b}_j]$, and using (5.25), it can be shown that $Y^{(j)}$ lies in

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm S\{p(1 + \mathbf{x}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})F_{p,n-p}^{\alpha/2}\}^{1/2} \text{ for all } \mathbf{x} \in \mathfrak{R}_p, x_0 \equiv 1, \\ \text{and all } j = 1, 2, \dots, k \quad (5.28)$$

with probability at least $1 - \alpha$.

Lane and DuMouchel [1994] give some examples where this method is better than the previous ones. Clearly, intervals could be computed using all the methods, and the shortest selected. If k , the number of predictions is large or unknown, another method is to use simultaneous tolerance intervals (cf. Limam and Thomas [1988] for several methods).

5.4 ENLARGING THE REGRESSION MATRIX

Suppose that our original regression model is enlarged by the addition of an extra regressor x_p , say, so that our model is now

$$G: Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (i = 1, 2, \dots, n).$$

What effect will this have on the width of the confidence intervals given in Sections 5.2 and 5.3? Surprisingly, the answer is that the intervals will be at least as wide and, in fact, almost invariably wider! To see this, we use the general theory of Section 3.7 to show that $\sigma^2 v$, the variance of the predictor \hat{Y} , cannot decrease when another regressor is added to the model. Setting

$$\beta_p = \gamma, \quad (x_{ip}) = \mathbf{x}_p = \mathbf{z}, \quad \text{and} \quad \mathbf{W} = (\mathbf{X}, \mathbf{z}),$$

we can write the model G in the form

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \varepsilon \\ &= \mathbf{W}\boldsymbol{\delta} + \varepsilon, \end{aligned}$$

and the least squares estimate of $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}}_G = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}.$$

For G , the new predictor at (\mathbf{x}'_0, x_{0p}) is

$$\hat{Y}_{0G} = (\mathbf{x}'_0, x_{0p})\hat{\boldsymbol{\delta}}_G,$$

and from Theorem 3.6(iv) in Section 3.7.1,

$$\begin{aligned} \text{var}[\hat{Y}_{0G}] &= (\mathbf{x}'_0, x_{0p}) \text{Var}[\hat{\boldsymbol{\delta}}_G] (\mathbf{x}'_0, x_{0p})' \\ &= \sigma^2 (\mathbf{x}'_0, x_{0p})(\mathbf{W}'\mathbf{W})^{-1} (\mathbf{x}'_0, x_{0p})' \\ &= \sigma^2 (\mathbf{x}'_0, x_{0p}) \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} + m\mathbf{k}\mathbf{k}', & -m\mathbf{k} \\ -m\mathbf{k}', & m \end{pmatrix} \begin{pmatrix} \mathbf{x}_0 \\ x_{0p} \end{pmatrix}, \end{aligned}$$

where $m = (\mathbf{z}'\mathbf{R}\mathbf{z})^{-1}$ and $\mathbf{k} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$. Multiplying out the matrix expression above, and completing the square on $\mathbf{k}'\mathbf{x}_0$, we have

$$\begin{aligned}\text{var}[\hat{Y}_{0G}] &= \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + m\sigma^2 (\mathbf{k}'\mathbf{x}_0 - x_{0p})^2 \\ &\geq \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 (= \sigma^2 v_0) \\ &= \text{var}[\hat{Y}_0],\end{aligned}\quad (5.29)$$

with equality if and only if $x_{0p} = \mathbf{k}'\mathbf{x}_0 = \mathbf{z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$. Since variances and covariances are independent of a change of origin, we see that the result above holds even if $E[\mathbf{Y}]$ is not equal to either $\mathbf{X}\beta$ or $\mathbf{W}\delta$; in this case, both predictors \hat{Y}_{0G} and \hat{Y}_0 are biased estimates of $E[\mathbf{Y}]$. We conclude, therefore, that although we may sometimes reduce the bias and improve the fit by enlarging the regression model, the variance of the predictor is not reduced. Walls and Weeks [1969] give an example in which the variance of prediction at a particular point is increased tenfold when the model is enlarged from a straight line to a quadratic. If we use the mean-squared error (MSE) of prediction as our criterion, then the MSE may increase or decrease when extra regressors are added to the model. Mallows' C_p statistic (see Section 12.3.2), which is used for comparing different regression models, is based on an "average" MSE criterion. By setting $x_{0p} = 0$ and setting \mathbf{x}_0 equal to the column vector with unity in the $(j+1)$ th position and zeros elsewhere in the theory above, we have $\text{var}[\hat{\beta}_{jG}] \geq \text{var}[\hat{\beta}_j]$ with equality if and only if $\mathbf{z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0$. Equality holds if \mathbf{z} is orthogonal to the columns of \mathbf{X} . However, in general, the variance of the least squares estimate of β_j increases when the model is enlarged. The lesson to be learned from this discussion is that we should avoid "overfitting" regression models. See Section 12.2 for further discussion of this point.

MISCELLANEOUS EXERCISES 5

- Referring to Section 5.1.1, prove that $\gamma = \sum_j \alpha_j$ is the expected number of incorrect statements. *Hint:* Let $I_j = 1$ if E_j is incorrect, and 0 otherwise.
- Prove that $(1 - \alpha/k)^k > 1 - \alpha$ ($k > 1$).
- Suppose X , Y , and Z are random variables and $a(\cdot)$ and $b(\cdot)$ are functions. Define

$$\text{cov}_Z[a(Z), b(Z)] = E_Z[(a(Z) - E\{a(Z)\})(b(Z) - E\{b(Z)\})].$$

Prove that

$$\text{cov}[X, Y] = E_Z[\text{cov}(X, Y|Z)] + \text{cov}_Z[E(X|Z), E(Y|Z)].$$

- Given the predictor $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1}$, show that \hat{Y} has a minimum variance of σ^2/n at the x point $x_j = \bar{x}_{.j}$ ($j = 1, 2, \dots, p-1$). *Hint:* Consider the model

$$Y_i = \alpha_0 + \beta_1(x_{i1} - \bar{x}_{.1}) + \cdots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{.p-1}) + \varepsilon_i.$$

(Kupper [1972: p. 52])

5. Generalize the argument given in Section 5.4; namely, show that the addition of several regressors to a regression model cannot decrease the variance of the prediction \hat{Y} . [Such a proof is, of course, not necessary, as we can add in the regressors just one at a time and evoke (5.29).]
6. Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), where the ε_i are independently distributed as $N(0, \sigma^2)$. Obtain a set of multiple confidence intervals for all linear combinations $a_0\beta_0 + a_1\beta_1$ (a_0, a_1 not both zero) such that the overall confidence for the set is $100(1 - \alpha)\%$.
7. In constructing simultaneous confidence intervals for all $\mathbf{x}'\boldsymbol{\beta}$, explain why setting $x_0 \equiv 1$ does not affect the theory. What modifications to the theory are needed if $\beta_0 = 0$?

6

Straight-Line Regression

6.1 THE STRAIGHT LINE

The simplest regression model is that of a straight line, namely,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where the ε_i are independently and identically distributed as $N(0, \sigma^2)$. The least squares theory was derived in Section 4.3.4 and we recall the following results:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}, \quad (6.1)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{\sum Y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2},$$

and

$$S^2 = \frac{1}{n-2} \left\{ \sum(Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 \right\}.$$

We now use the theory of Chapter 5 to construct various confidence intervals and bands.

6.1.1 Confidence Intervals for the Slope and Intercept

Using the maximum modulus method of Section 5.1.1 [equation (5.8)] with $\mathbf{a}'_1 = (1, 0)$ and $\mathbf{a}'_2 = (0, 1)$, we have an *exact* overall confidence probability of

$1 - \alpha$ for the following confidence intervals for β_0 and β_1 :

$$\hat{\beta}_0 \pm u_{2,n-2,\rho}^\alpha S \left\{ \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right\}^2$$

and

$$\hat{\beta}_1 \pm u_{2,n-2,\rho}^\alpha S \left\{ \frac{1}{\sum (x_i - \bar{x})^2} \right\}^2,$$

where from Exercises 4b, No. 2(a),

$$\rho = \frac{-n\bar{x}}{(n \sum x_i^2)^{1/2}}.$$

Conservative intervals are obtained by setting $\rho = 0$ or using the Bonferroni method with multiplier $t_{n-2}^{\alpha/4}$.

The two intervals can also be used for *jointly* testing a hypothesis about β_0 and a hypothesis about β_1 . However, if we are interested in just a single hypothesis, say, $H: \beta_1 = c$, then we can use the usual t -statistic,

$$T = \frac{\hat{\beta}_1 - c}{S / \{\sum (x_i - \bar{x})^2\}^{1/2}}, \quad (6.2)$$

and reject H at the α level of significance if $|T| > t_{n-2}^{(1/2)\alpha}$. This statistic can be derived directly from the fact that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum (x_i - \bar{x})^2)$ and S^2 is independent of $\hat{\beta}_1$. The F -statistic T^2 is given by (4.19).

6.1.2 Confidence Interval for the x -Intercept

When $E[Y] = 0$, $0 = \beta_0 + \beta_1 x$ and the x -intercept is $\phi = -\beta_0/\beta_1$. We now derive a confidence interval for ϕ using a technique due to Fieller [1940].

Let

$$\begin{aligned} \delta &= \frac{E[\bar{Y}]}{E[\hat{\beta}_1]} \\ &= \frac{\beta_0 + \beta_1 \bar{x}}{\beta_1} \\ &= -\phi + \bar{x}; \end{aligned} \quad (6.3)$$

then $E[\bar{Y} - \delta \hat{\beta}_1] = 0$. Also,

$$\begin{aligned} \text{cov}[\bar{Y}, \hat{\beta}_1] &= \text{cov}[\mathbf{a}' \mathbf{Y}, \mathbf{b}' \mathbf{Y}] \\ &= \mathbf{a}' \text{Var}[\mathbf{Y}] \mathbf{b} \\ &= \sigma^2 \mathbf{a}' \mathbf{b} \\ &= \sigma^2 \frac{\sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2 n} \\ &= 0, \end{aligned}$$

so that

$$\begin{aligned}\text{var}[(\bar{Y} - \delta\hat{\beta}_1)] &= \text{var}[\bar{Y}] + \delta^2 \text{var}[\hat{\beta}_1] \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{\delta^2}{\sum(x_i - \bar{x})^2} \right\} \\ &= \sigma^2 w,\end{aligned}$$

say. Now $\bar{Y} - \delta\hat{\beta}_1$ is of the form $\mathbf{c}'\mathbf{Y}$, so that it is univariate normal, namely, $N(0, \sigma^2 w)$. Also, S^2 is independent of $(\hat{\beta}_0, \hat{\beta}_1)$ [Theorem 3.5(iii), Section 3.4] and therefore of $\bar{Y} - \delta\hat{\beta}_1 [= \hat{\beta}_0 + \hat{\beta}_1(\bar{x} - \delta)]$. Hence, by the usual argument for constructing t -variables [see equation (4.16)],

$$T = \frac{\bar{Y} - \delta\hat{\beta}_1}{S\sqrt{w}} \sim t_{n-2},$$

and a $100(1 - \alpha)\%$ confidence set for δ is given by

$$T^2 \leq (t_{n-2}^{(1/2)\alpha})^2 = F_{1,n-2}^\alpha.$$

It transpires that this set reduces to the simple interval $d_1 \leq \delta \leq d_2$, where d_1 and d_2 are the roots of the quadratic

$$d^2 \left\{ \hat{\beta}_1^2 - \frac{S^2 F_{1,n-2}^\alpha}{\sum(x_i - \bar{x})^2} \right\} - 2d\bar{Y}\hat{\beta}_1 + \left(\bar{Y}^2 - \frac{1}{n} S^2 F_{1,n-2}^\alpha \right) = 0 \quad (6.4)$$

if and only if the coefficient of d^2 in equation (6.4) is positive (i.e., the line is not too flat). In this case, from equation (6.3), the corresponding interval for ϕ is $[\bar{x} - d_2, \bar{x} - d_1]$, and $\hat{\phi} = -\hat{\beta}_0/\hat{\beta}_1$ lies in this interval.

We note that $\hat{\phi}$ is the ratio of two correlated normal random variables; the exact distribution of such a ratio is given by Hinkley [1969a].

EXAMPLE 6.1 A model that often arises in animal population studies (cf. Seber [1982: p. 298]) is the following:

$$\begin{aligned}E[Y] &= \gamma(N - x) \\ &= \gamma N - \gamma x \\ (\quad &= \beta_0 + \beta_1 x, \text{ say}).\end{aligned}$$

In such applications we are interested in finding a confidence interval for the population size $N = -\beta_0/\beta_1$. \square

6.1.3 Prediction Intervals and Bands

The fitted regression line is

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \bar{Y} + \hat{\beta}_1(x - \bar{x}),\end{aligned}$$

which passes through the point (\bar{x}, \bar{Y}) . From the general theory of Section 5.2 we see that we can use the prediction $\hat{Y}_0 = \mathbf{x}'_0 \hat{\beta} = (1, x_0) \hat{\beta}$ to obtain a $100(1 - \alpha)\%$ confidence interval for $E[Y_0] = (1, x_0) \beta$, the expected value of Y at $x = x_0$. This interval is

$$\hat{Y}_0 \pm t_{n-2}^{(1/2)\alpha} S \sqrt{v_0}, \quad (6.5)$$

where, from equation (6.1),

$$\begin{aligned} v_0 &= \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0 \\ &= \frac{\sum x_i^2 - 2x_0 n \bar{x} + nx_0^2}{n \sum (x_i - \bar{x})^2} \\ &= \frac{\sum x_i^2 - n \bar{x}^2 + n(x_0 - \bar{x})^2}{n \sum (x_i - \bar{x})^2} \\ &= \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}. \end{aligned} \quad (6.6)$$

(Here v_0 can also be obtained directly—see Exercises 6a, No. 1.) We note that v_0 is a minimum when $x_0 = \bar{x}$; the farther we are from \bar{x} , the wider our confidence interval.

If we require k prediction intervals, then our critical constant $t_{n-2}^{\alpha/2}$ in (6.5) is replaced by $t_{n-2}^{\alpha/(2k)}$, $(2F_{2,n-2}^\alpha)^{1/2}$, and $u_{k,n-2}^\alpha$ for the Bonferroni, Scheffé, and maximum modulus methods, respectively. However, if k is unknown or is so large that the intervals are too wide, we can construct a confidence band for the entire regression line and thus obtain an unlimited number of confidence intervals with an overall confidence probability of at least $1 - \alpha$. From equation (5.22) this infinite band is the region between the two curves (Figure 6.1)

$$y = \bar{Y} + \hat{\beta}_1(x - \bar{x}) \pm \lambda S \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}^{1/2}, \quad (6.7)$$

where $\lambda = (2F_{2,n-2}^\alpha)^{1/2}$. This band, commonly called the *Working-Hotelling confidence band* (Working and Hotelling [1929]), is of variable vertical width d , d being a minimum at the point (\bar{x}, \bar{Y}) . The intervals obtained from this band are simply the Scheffé F -intervals.

An alternative confidence band with straight sides (Figure 6.2) has been proposed by Graybill and Bowden [1967], namely,

$$y = \bar{Y} + \hat{\beta}_1(x - \bar{x}) \pm u_{2,n-2}^\alpha S \frac{1}{\sqrt{n}} \left(1 + \frac{|x - \bar{x}|}{s_x} \right), \quad (6.8)$$

where $s_x^2 = \sum (x_i - \bar{x})^2 / n$. This band has two advantages over (6.7): (1) it is easier to graph, and (2) it has a smaller average width, although this is misleading since the average is taken over the entire band, including extreme

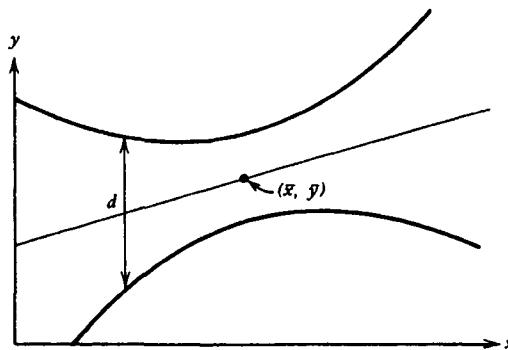


Fig. 6.1 Working-Hotelling confidence band.

values of x . However, Dunn [1968] and Halperin and Gurian [1968: p. 1027] show that for $\alpha = 0.05$, (6.7) provides narrower intervals than (6.8) when x satisfies (approximately)

$$0.1 \leq \frac{|x - \bar{x}|}{s_x} \leq 9.$$

Since, in practice, one would not expect the experimental range of $|x - \bar{x}|$ to exceed $5s_x$, the Working-Hotelling band is preferred. A similar conclusion holds for 90% confidence levels ($\alpha = 0.1$). Both bands can be derived as special cases of a general procedure given by Bowden [1970] [cf. equation (5.23) and the following discussion].

The problem of obtaining an exact confidence band for the regression line when x_0 is restricted to the finite interval $[a, b]$ was first solved by Gafarian [1964]. He showed how to construct a band of uniform width 2δ and provided appropriate tables for the case $\bar{x} = \frac{1}{2}(a + b)$ and even n . Miller [1981: p. 121] gave a useful discussion of this method and pointed out that the two conditions necessary for the use of the tables are not very restrictive: The interval $[a, b]$

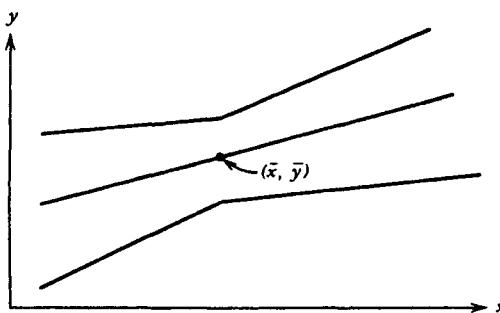


Fig. 6.2 Graybill-Bowden confidence band.

is usually sufficiently ill-defined to permit adjustment so that \bar{x} is the middle point, and interpolation in the tables gives approximate results for odd values of n . However, Bowden and Graybill [1966] later provided tables for *any* finite interval $[a, b]$ and even n . Their tables can also be used for computing exact trapezoidal confidence bands, which may be more appropriate than uniform width bands when \bar{x} lies outside $[a, b]$.

Dunn [1968] provided a truncated modification of (6.8) which gave a conservative confidence band. Halperin et al. [1967] and Halperin and Gurian [1968] gave an exact confidence band of the form (6.7) but with a different value of λ , and truncated at $x = a$ and $x = b$. However, their tables can only be used for the case $\bar{x} = \frac{1}{2}(a + b)$; λ is tabulated in Halperin et al. [1967] for different values of Q^{-1} , where

$$Q = 1 + \frac{(b - a)^2}{4s_x^2}.$$

Wynn and Bloomfield [1971], however, tackled the problem from a different viewpoint and provided tables (reproduced in Appendix C.3) for any interval $[a, b]$. One simply calculates a “standardized” version of the interval width, namely,

$$c = \frac{(b - a)s_x}{[\{s_x^2 + (a - \bar{x})^2\}\{s_x^2 + (b - \bar{x})^2\}]^{1/2} + s_x^2 + (a - \bar{x})(b - \bar{x})} \quad (6.9)$$

and looks up the corresponding value of λ in Appendix C.3. When $\bar{x} = \frac{1}{2}(a + b)$ we note that $c = (b - a)/2s_x$ and $Q = 1 + c^2$, thus linking the tables in Halperin et al. [1967] with Appendix C.3. Letting $a \rightarrow -\infty, b \rightarrow \infty$, we have $c = \infty$ and $\lambda = (2F_{2,n-2}^\alpha)^{1/2}$, as expected. Calculations given by Halperin and Gurian [1968] suggest that this modification of the Working-Hotelling band generally provides narrower confidence intervals than either the uniform or trapezoidal bands mentioned above. In conclusion, therefore, we recommend the general use of (6.7) but with λ obtained from Appendix C.3 in the case $x \in [a, b]$.

Finally, we mention one-sided confidence intervals. Bohrer and Francis [1972] give an (upper) one-sided analog of (6.7), namely (modifying their model slightly so that $x \in [a, b]$ instead of $x - \bar{x} \in [a, b]$),

$$1 - \alpha = \text{pr} \left\{ \beta_0 + \beta_1 x \leq \bar{y} + \hat{\beta}_1(x - \bar{x}) + \lambda S \left[\frac{1}{n} + \frac{x - \bar{x}}{\sum(x_i - \bar{x})^2} \right]^{1/2}, \right. \\ \left. \text{all } x \in [a, b] \right\}, \quad (6.10)$$

where λ ($= c\#$ in their notation) is tabulated for different n , ϕ^* ($= \arctan[(b - \bar{x})/s_x] - \arctan[(a - \bar{x})/s_x]$), and α ($= 1 - \alpha$ in their notation). Lower one-sided intervals are obtained by reversing the inequality and replacing λ by $-\lambda$.

6.1.4 Prediction Intervals for the Response

From the general theory of Section 5.3, we can use the predictor \hat{Y}_0 to obtain a $100(1 - \alpha)\%$ confidence interval for the random variable Y_0 , namely,

$$\hat{Y}_0 \pm t_{n-2}^{(1/2)\alpha} S(1 + v_0)^{1/2},$$

where v_0 is given by equation (6.6). If k intervals are required at $x = x_0^{(j)}$ ($j = 1, 2, \dots, k$), then we can use

$$\hat{Y}_0^{(i)} \pm \lambda S(1 + v_0^{(j)})^{1/2} \quad (j = 1, 2, \dots, k),$$

where λ is $t_{n-2}^{\alpha/(2k)}$, $(kF_{k,n-2}^\alpha)^{1/2}$, and $u_{k,n-2}^\alpha$ for the Bonferroni, Scheffé, and maximum modulus methods, respectively. However, if k is so large that the intervals are hopelessly wide, or k is unknown, then we can use simultaneous tolerance intervals (see Limam and Thomas [1988] for several methods).

6.1.5 Inverse Prediction (Calibration)

Single Observation

Suppose that we wish to calibrate an instrument, say, a pressure gauge, and we know that the gauge reading is a linear function of the pressure, namely,

$$\text{"gauge reading"} = \beta_0 + \beta_1 \text{"pressure"} + \text{"error"}$$

or

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

In order to calibrate the gauge, we subject it to two or more (say, n) *controlled pressures* x_i ($i = 1, 2, \dots, n$) and note the gauge readings Y_i . Using these data we obtain the fitted equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$, which can be used for estimating (predicting) the unknown pressure x_0 for a given gauge reading Y_0 . This is the inverse problem to the one considered in Section 6.1.4 of predicting Y_0 for a given $x = x_0$, and it is commonly referred to as the *controlled calibration problem*. The case when x is fixed rather than random, which we consider here, is also referred to as the *absolute calibration problem*.

A natural estimate of x_0 (which is also the maximum likelihood estimate) is found by solving the fitted equation $Y_0 = \hat{\beta}_0 + \hat{\beta}_1 x$, namely,

$$\hat{x}_0 = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1} = \bar{x} + \frac{Y_0 - \bar{Y}}{\hat{\beta}_1}. \quad (6.11)$$

This ratio-type estimate is biased because, in general,

$$E[\hat{x}_0] \neq \frac{E[Y_0 - \hat{\beta}_0]}{E[\hat{\beta}_1]} = x_0.$$

However, a confidence interval for x_0 can be constructed using the method of Section 6.1.2. From equation (5.26), we get

$$Y_0 - \hat{Y}_0 = Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0 \sim N(0, \sigma^2(1 + v_0)),$$

so that

$$T = \frac{Y_0 - \hat{Y}_0}{S\sqrt{1+v_0}} = \frac{Y_0 - \bar{Y} - \hat{\beta}_1(x_0 - \bar{x})}{S\sqrt{1+v_0}} \sim t_{n-2},$$

where v_0 is given by (6.6). Since

$$\begin{aligned} 1 - \alpha &= \text{pr} [|T| \leq t_{n-2}^{(1/2)\alpha}] \\ &= \text{pr} [T^2 \leq (t_{n-2}^{(1/2)\alpha})^2], \end{aligned}$$

the set of all values of x satisfying the inequality

$$\left\{ Y_0 - \bar{Y} - \hat{\beta}_1(x - \bar{x}) \right\}^2 \leq \lambda^2 S^2 \left\{ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right\}, \quad (6.12)$$

with $\lambda = t_{n-2}^{(1/2)\alpha}$ (and $\lambda^2 = F_{1,n-2}^\alpha$), will provide a $100(1 - \alpha)\%$ confidence region for the unknown x_0 . This set of points, commonly called the *discrimination interval*, may give a finite interval, two semi-infinite lines, or the entire real line (see Miller [1981: pp. 118–119; Figures 2, 3, and 4] and Hoadley [1970]). One obtains a finite interval if and only if $\hat{\beta}_1^2 > \lambda^2 S^2 / \sum(x_i - \bar{x})^2$; that is, the F -test for $\beta_1 = 0$ is significant, which we would expect for any sensible calibration curve. In this case the interval contains the estimate \hat{x}_0 and is given by $[d_1 + \bar{x}, d_2 + \bar{x}]$, where d_1 and d_2 are the (real unequal) roots of

$$\begin{aligned} d^2 \left\{ \hat{\beta}_1^2 - \frac{\lambda^2 S^2}{\sum(x_i - \bar{x})^2} \right\} - 2d\hat{\beta}_1(Y_0 - \bar{Y}) \\ + \left\{ (Y_0 - \bar{Y})^2 - \lambda^2 S^2 \left(1 + \frac{1}{n} \right) \right\} = 0. \end{aligned} \quad (6.13)$$

[This equation follows from (6.12) by setting $d = x - \bar{x}$.] If \hat{x}_0 does not lie in $[d_1 + \bar{x}, d_2 + \bar{x}]$, then the confidence region for x_0 is the union of two semi-infinite lines. However, if (6.13) has no real roots, then the region is the entire real line. The confidence region defined by (6.13) can also be derived by inverting a test of the hypothesis $x = x_0$ (Cox and Hinkley [1974: p. 268]). A bootstrap approach to the problem is given by Jones and Rocke [1999].

The theory above is readily extended in two directions. If k values of Y_0 are observed at *different* values of x_0 , then one simply substitutes $Y_0^{(j)}$ ($j = 1, \dots, k$) in (6.13) and sets λ equal to $t_{n-2}^{\alpha/(2k)}$ and $u_{k,n-2}^\alpha$ for the Bonferroni and maximum modulus intervals, respectively. Unfortunately, this method cannot be used when k is unknown. Such will be the case in calibration

problems where the estimated calibration line is used to "correct" an unlimited number of future readings taken with the instrument; for example, in bioassay a standard curve is constructed for making future assays (discriminations). If k is large, λ may be so large as to render the discrimination intervals useless. However, when k is large or unknown, several simultaneous confidence intervals can be constructed (see Mee and Eberhardt [1996]).

Krutchoff [1967, 1969] resurrected an alternative estimate \tilde{x}_0 , called the *inverse estimate*, obtained by regressing x on Y (even when x is not a random variable) and then predicting x_0 from Y_0 in the usual manner. There has been an extensive debate on the relative merits of \hat{x}_0 and \tilde{x}_0 in the literature, and details are given by Osborne [1991]. It is clear that \hat{x}_0 is satisfactory provided that $\hat{\beta}_1$ is not too small, as already mentioned above, and the properties of \tilde{x}_0 should be derived conditional on this requirement. Hoadley [1970] showed that \tilde{x}_0 is a Bayes solution with respect to a particular prior distribution on x_0 . He also gave a confidence interval based on \tilde{x}_0 when the particular prior could be justified. For further comments on this issue, see Brown [1993: pp. 31–33].

In practice there is not a great deal of difference between \hat{x}_0 and \tilde{x}_0 when the data are close to a straight line [i.e., when r^2 is large; see Exercises 6a, No. 4].

Replicated Observations

Suppose that we have m replications Y_{0j} ($j = 1, 2, \dots, m$; $m > 1$), with sample mean \bar{Y}_0 , at the unknown value $x = x_0$. In this situation we have two estimates of σ^2 , namely, S^2 and $\sum_j (Y_{0j} - \bar{Y}_0)^2 / (m - 1)$, which can be combined to give a confidence interval for x_0 as follows. Following Graybill [1961: pp. 125–127], let $U = \bar{Y}_0 - \bar{Y} - \hat{\beta}_1(x_0 - \bar{x})$. Then $E[U] = 0$,

$$\text{var}[U] = \sigma^2 \left\{ \frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right\} = \sigma_U^2,$$

say, and $U/\sigma_U \sim N(0, 1)$. If

$$V_1 = \sum_{i=1}^n \left[Y_i - \bar{Y} - \hat{\beta}_1(x_i - \bar{x}) \right]^2 = \text{RSS}$$

and

$$V_2 = \sum_{j=1}^m (Y_{0j} - \bar{Y}_0)^2,$$

then U , V_1 , and V_2 are mutually independent and

$$\frac{(n+m-3)\hat{\sigma}^2}{\sigma^2} = \frac{V_1 + V_2}{\sigma^2} \sim \chi^2_{n-2+m-1}. \quad (6.14)$$

Therefore,

$$\begin{aligned} T &= \frac{U/\sigma_U}{\hat{\sigma}/\sigma} \\ &= \frac{U}{\hat{\sigma} \left\{ \frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right\}^{1/2}} \sim t_{n+m-3} \end{aligned} \quad (6.15)$$

and (6.13) now becomes

$$\begin{aligned} d^2 \left\{ \hat{\beta}_1^2 - \frac{\mu^2 \hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \right\} - 2d\hat{\beta}_1(\bar{Y}_0 - \bar{Y}) \\ + \left\{ (\bar{Y}_0 - \bar{Y})^2 - \mu^2 \hat{\sigma}^2 \left(\frac{1}{m} + \frac{1}{n} \right) \right\} = 0, \end{aligned} \quad (6.16)$$

where $\mu^2 = (t_{n+m-3}^{(1/2)\alpha})^2 = F_{1,n+m-3}^\alpha$. We note that $\hat{\sigma}^2$, based on $n+m-3$ degrees of freedom, has a smaller sampling variance than S^2 with $n-2$ degrees of freedom; also, $\mu^2 < \lambda^2$. These two facts imply that (Cox [1971]) (1) the intervals given by (6.16) will, on the average, be narrower than those given by (6.13), and (2) the coefficient of d^2 in (6.16) is generally larger than that in (6.13), so that the probability of obtaining a *finite* confidence interval for x_0 is greater when there are replications of Y_0 .

Using a profile likelihood approach, Brown [1993: pp. 26–30] showed that the profile likelihood is a monotonic function of (6.15).

EXERCISES 6a

1. In fitting the straight line $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), prove that \bar{Y} and $\hat{\beta}_1$ are uncorrelated. If $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, deduce that

$$\text{var}[\hat{Y}_0] = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right\}.$$

2. Using the notation of Section 6.1.2, prove that $\hat{\phi} = -\hat{\beta}_0/\hat{\beta}_1$ is the maximum likelihood estimate of ϕ .
3. Given a general linear regression model, show how to find a confidence interval for the ratio $\mathbf{a}'_1 \boldsymbol{\beta} / \mathbf{a}'_2 \boldsymbol{\beta}$ of two linear parametric functions.
4. Using the notation of Section 6.1.5, show that when $\bar{x} = 0$,

$$\frac{\hat{x}_0 - \tilde{x}_0}{\hat{x}_0} = 1 - r^2,$$

where r is the correlation coefficient of the pairs (x_i, Y_i) .

6.2 STRAIGHT LINE THROUGH THE ORIGIN

In many situations it is known that $E[Y] = 0$ when $x = 0$, so that the appropriate regression line is $Y_i = \beta_1 x_i + \varepsilon_i$. The least squares estimate of β_1 is now

$$\tilde{\beta}_1 = \frac{\sum_i Y_i x_i}{\sum_i x_i^2},$$

and the unbiased estimate of σ^2 is

$$S^2 = \frac{1}{n-1} \left(\sum Y_i^2 - \tilde{\beta}_1^2 \sum x_i^2 \right). \quad (6.17)$$

Because $\tilde{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum x_i^2)$, a t -confidence interval for β_1 is

$$\tilde{\beta}_1 \pm t_{n-1}^{(1/2)\alpha} S \left(\sum_i x_i^2 \right)^{-1/2}. \quad (6.18)$$

We can use the predictor $\tilde{Y}_0 = x_0 \tilde{\beta}_1$ to obtain a confidence interval for $E[Y_0] = x_0 \beta_1$ at $x = x_0$, namely,

$$\tilde{Y}_0 \pm t_{n-1}^{(1/2)\alpha} S \sqrt{v_0}, \quad (6.19)$$

where $v_0 = x_0^2 / \sum x_i^2$; this interval gets wider as we move away from the origin. Since β_1 lies in the interval (6.18) if and only if $x_0 \beta_1$ lies in (6.19) for every x_0 , a $100(1 - \alpha)\%$ confidence band for the entire regression line is the region between the two lines,

$$y = \tilde{\beta}_1 x \pm t_{n-1}^{(1/2)\alpha} S |x| \left(\sum x_i^2 \right)^{-1/2}.$$

Prediction intervals for Y_0 , or k values of Y_0 , are obtained as in Section 6.1.4; however, v_0 is defined as above and the appropriate degrees of freedom are now $n-1$ instead of $n-2$. Inverse prediction is also straightforward. Following the method of Section 6.1.5, we find that x_0 is estimated by $\tilde{x}_0 = Y_0 / \tilde{\beta}_1$, and the corresponding confidence interval for x_0 is given by the roots of

$$x^2 \left(\tilde{\beta}_1^2 - \frac{\lambda^2 S^2}{\sum x_i^2} \right) - 2x\tilde{\beta}_1 Y_0 + Y_0^2 - \lambda^2 S^2 = 0, \quad (6.20)$$

where $\lambda = t_{n-1}^{(1/2)\alpha}$ and S^2 is given by (6.17). For m replications Y_{0j} at $x = x_0$, the corresponding quadratic is (cf. Cox [1971])

$$x^2 \left(\tilde{\beta}_1^2 - \frac{\mu^2 \hat{\sigma}^2}{\sum x_i^2} \right) - 2x\tilde{\beta}_1 \bar{Y}_0 + \bar{Y}_0^2 - \frac{\mu^2 \hat{\sigma}^2}{m} = 0,$$

where $\mu = t_{n+m-2}^{(1/2)\alpha}$, and

$$\hat{\sigma}^2 = \frac{1}{n+m-2} \left\{ \sum_{i=1}^n (Y_i - \tilde{\beta}_1 x_i)^2 + \sum_{j=1}^m (Y_{0j} - \bar{Y}_0)^2 \right\}.$$

6.3 WEIGHTED LEAST SQUARES FOR THE STRAIGHT LINE

6.3.1 Known Weights

Let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), where the ε_i are independently distributed as $N(0, \sigma^2 w_i^{-1})$, and the w_i are known positive numbers. Then, from Section 3.10, the weighted least squares estimates β_0^* and β_1^* of β_0 and β_1 , respectively, are obtained by minimizing $\sum w_i(Y_i - \beta_0 - \beta_1 x_i)^2$. Therefore, differentiating this expression partially with respect to β_0 and β_1 , we have

$$\beta_0^* \sum w_i + \beta_1^* \sum w_i x_i = \sum w_i Y_i \quad (6.21)$$

and

$$\beta_0^* \sum w_i x_i + \beta_1^* \sum w_i x_i^2 = \sum w_i Y_i x_i. \quad (6.22)$$

Dividing (6.21) by $\sum w_i$ and defining the weighted means $\bar{Y}_w = \sum w_i Y_i / \sum w_i$, etc., we have

$$\beta_0^* = \bar{Y}_w - \beta_1^* \bar{x}_w. \quad (6.23)$$

Substituting (6.23) in (6.22) leads to

$$\begin{aligned} \beta_1^* &= \frac{\sum w_i Y_i x_i - \sum w_i x_i \bar{Y}_w}{\sum w_i x_i^2 - \sum w_i x_i \bar{x}_w} \\ &= \frac{\sum w_i (Y_i - \bar{Y}_w)(x_i - \bar{x}_w)}{\sum w_i (x_i - \bar{x}_w)^2}. \end{aligned}$$

From the alternative expression

$$\beta_1^* = \frac{\sum w_i Y_i (x_i - \bar{x}_w)}{\sum w_i (x_i - \bar{x}_w)^2} \quad (6.24)$$

it readily follows that

$$\text{var}[\beta_1^*] = \frac{\sigma^2}{\sum w_i (x_i - \bar{x}_w)^2}.$$

Using the general theory of Section 3.10, we can show that

$$S_w^2 = \frac{1}{n-2} \left\{ \sum w_i [Y_i - \bar{Y}_w - \beta_1^*(x_i - \bar{x}_w)]^2 \right\} \quad (6.25)$$

$$= \frac{1}{n-2} \left\{ \sum w_i (Y_i - \bar{Y}_w)^2 - (\beta_1^*)^2 \sum w_i (x_i - \bar{x}_w)^2 \right\} \quad (6.26)$$

is an unbiased estimate of σ^2 , and a $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\beta_1^* \pm t_{n-2}^{(1/2)\alpha} \left[\frac{S_w^2}{\sum w_i (x_i - \bar{x}_w)^2} \right]^{1/2}. \quad (6.27)$$

When $\beta_0 = 0$ and $\beta_1 = \beta$ we have, from Example 3.9 in Section 3.10,

$$\beta^* = \frac{\sum w_i Y_i x_i}{\sum w_i x_i^2}, \quad (6.28)$$

and the appropriate confidence interval for β is now

$$\beta^* \pm t_{n-1}^{(1/2)\alpha} \left(\frac{S_w^2}{\sum w_i x_i^2} \right)^{1/2},$$

where

$$S_w^2 = \frac{1}{n-1} \left\{ \sum w_i Y_i^2 - (\beta^*)^2 \sum w_i x_i^2 \right\}. \quad (6.29)$$

[We note that these formulas follow from those given by equations (6.24) to (6.27) by setting $\bar{Y}_w = \bar{x}_w = 0$ and replacing $n - 2$ by $n - 1$.] Under the normality assumptions, β^* is the maximum likelihood estimate of β . However, Turner [1960] has shown that for certain w_i , β^* can still be the maximum likelihood estimate when Y is not normally distributed (cf. Exercises 6b, No. 1). Inverse prediction (discrimination) for this model is discussed by Cox [1971].

6.3.2 Unknown Weights

Let

$$\begin{aligned} Y_i &= \theta_i + \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n), \end{aligned}$$

where the ε_i are independently distributed as $N(0, vg(\theta_i))$; here $v = \sigma^2$, g is a known positive function, and the weights $w_i = 1/g(\theta_i)$ are now unknown. Two methods are available for estimating β_0 and β_1 .

Maximum Likelihood Method

If $g_i = g(\theta_i)$, then L , the logarithm of the likelihood function, is given by

$$L = -\frac{1}{2}n \log 2\pi - \frac{1}{2} \sum_i \log(vg_i) - \frac{1}{2} \sum_i \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{vg_i}.$$

Now

$$\frac{\partial \log g}{\partial \theta} = \frac{1}{g} \cdot \frac{\partial g}{\partial \theta} = h,$$

say, so that

$$\frac{\partial g}{\partial \beta_0} = \frac{\partial g}{\partial \theta} \cdot \frac{\partial \theta}{\partial \beta_0} = gh$$

and

$$\frac{\partial g}{\partial \beta_1} = \frac{\partial g}{\partial \theta} \cdot \frac{\partial \theta}{\partial \beta_1} = ghx.$$

The maximum likelihood estimates $\tilde{\beta}_0$, $\tilde{\beta}_1$, and \tilde{v} are obtained by solving $\partial L/\partial\beta_0 = \partial L/\partial\beta_1 = \partial L/\partial v = 0$, namely,

$$\begin{aligned} -\frac{1}{2}\sum_i \tilde{h}_i + \frac{1}{2}\sum_i \left\{ \frac{\tilde{h}_i(Y_i - \tilde{\theta}_i)^2}{\tilde{v}\tilde{g}_i} \right\} + \sum_i \left\{ \frac{(Y_i - \tilde{\theta}_i)}{\tilde{v}\tilde{g}_i} \right\} &= 0, \\ -\frac{1}{2}\sum_i \tilde{h}_i x_i + \frac{1}{2}\sum_i \left\{ \frac{\tilde{h}_i x_i(Y_i - \tilde{\theta}_i)^2}{\tilde{v}\tilde{g}_i} \right\} + \sum_i \left\{ \frac{x_i(Y_i - \tilde{\theta}_i)}{\tilde{v}\tilde{g}_i} \right\} &= 0, \end{aligned}$$

and

$$-\frac{1}{2}\frac{n}{\tilde{v}} + \frac{1}{2}\sum \frac{(Y_i - \tilde{\theta}_i)^2}{\tilde{v}^2\tilde{g}_i} = 0,$$

where \tilde{h}_i , \tilde{g}_i , and $\tilde{\theta}_i$ are functions of $\tilde{\beta}_0$ and $\tilde{\beta}_1$. Multiplying through by \tilde{v} , setting $\tilde{Y}_i = \tilde{\theta}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$, and $\tilde{w}_i = 1/\tilde{g}_i$, we can reduce the equations above to

$$\tilde{\beta}_0 \sum \tilde{w}_i + \tilde{\beta}_1 \sum \tilde{w}_i x_i = \sum \tilde{w}_i Y_i + \frac{1}{2} \sum \tilde{h}_i [\tilde{w}_i(Y_i - \tilde{Y}_i)^2 - \tilde{v}], \quad (6.30)$$

$$\tilde{\beta}_0 \sum \tilde{w}_i x_i + \tilde{\beta}_1 \sum \tilde{w}_i x_i^2 = \sum \tilde{w}_i x_i Y_i + \frac{1}{2} \sum \tilde{h}_i x_i [\tilde{w}_i(Y_i - \tilde{Y}_i)^2 - \tilde{v}] \quad (6.31)$$

and

$$\tilde{v} = \frac{1}{n} \sum \tilde{w}_i (Y_i - \tilde{Y}_i)^2. \quad (6.32)$$

Equations (6.30) and (6.31) may be compared with (6.21) and (6.22). Therefore, given initial approximations to $\tilde{\beta}_0$ and $\tilde{\beta}_1$ (say, unweighted least squares estimates), we can evaluate the corresponding values of w_i , h_i , and v , solve (6.30) and (6.31), and obtain new approximations for $\tilde{\beta}_0$ and $\tilde{\beta}_1$. This process is then repeated. When n is large, the variance-covariance matrix of the maximum likelihood estimates is approximately

$$\begin{aligned} &\left(\begin{array}{ccc} -E\left[\frac{\partial^2 L}{\partial \beta_0^2}\right] & -E\left[\frac{\partial^2 L}{\partial \beta_0 \partial \beta_1}\right] & -E\left[\frac{\partial^2 L}{\partial \beta_0 \partial v}\right] \\ -E\left[\frac{\partial^2 L}{\partial \beta_0 \partial \beta_1}\right] & -E\left[\frac{\partial^2 L}{\partial \beta_1^2}\right] & -E\left[\frac{\partial^2 L}{\partial \beta_1 \partial v}\right] \\ -E\left[\frac{\partial^2 L}{\partial \beta_0 \partial v}\right] & -E\left[\frac{\partial^2 L}{\partial \beta_1 \partial v}\right] & -E\left[\frac{\partial^2 L}{\partial v^2}\right] \end{array} \right)^{-1} \\ &\approx v \left(\begin{array}{ccc} \sum_i a_i & \sum_i x_i a_i & \frac{1}{2} \sum h_i \\ \sum_i x_i a_i & \sum_i x_i^2 a_i & \frac{1}{2} \sum_i h_i x_i \\ \frac{1}{2} \sum_i h_i & \frac{1}{2} \sum_i h_i x_i & \frac{1}{2} \left(\frac{n}{v}\right) \end{array} \right)^{-1}, \end{aligned}$$

where

$$\begin{aligned} a_i &= g_i^{-1} + \frac{1}{2} h_i^2 v \\ &= \frac{1}{g_i} \left\{ 1 + \frac{v}{2g_i} \left(\frac{\partial g_i}{\partial \theta_i} \right)^2 \right\}. \end{aligned}$$

Frequently, the second term of the preceding equation is small. For example, if $g_i = \theta_i^2$, then the second term is $2v$, which can be neglected if v is much smaller than $\frac{1}{2}$. In this case the variance-covariance matrix for $\tilde{\beta}_0$ and $\tilde{\beta}_1$ is approximately

$$v \left(\begin{array}{cc} \sum_i w_i & \sum_i x_i w_i \\ \sum_i x_i w_i & \sum_i x_i^2 w_i \end{array} \right)^{-1}, \quad (6.33)$$

which is the variance-covariance matrix of β_0^* and β_1^* in Section 6.3.1.

The treatment above is based on Williams [1959: pp. 67–70], although with the following differences: We have worked with σ^2 instead of σ , and Williams's g_i^2 is our g_i (his g_i may be negative).

Least Squares Method

This technique consists of estimating the weights w_i [$= 1/g(\beta_0 + \beta_1 x_i)$] from trial estimates of β_0 and β_1 , say, the unweighted least squares estimates (which are unbiased), and then solving equations (6.21) and (6.22) for new estimates of β_0 and β_1 . These new values may be used for recalculating the w_i , and the process can be repeated. Williams [1959] suggests that only two cycles of iteration are generally required, as great accuracy in the weights is not necessary for giving accurate estimates of β_0 and β_1 . Ignoring the fact that the estimated w_i are strictly random variables, the variance-covariance matrix of the least squares estimates is given approximately by (6.33). By the same argument, approximate tests and confidence intervals can be obtained using the theory of Section 6.3.1, but with the w_i estimated. For this reason, and for computational simplicity, the least squares method is often preferred to the maximum likelihood approach.

EXERCISES 6b

1. Let Y_1, Y_2, \dots, Y_n be independent random variables such that for $i = 1, 2, \dots, n$,

$$E[Y_i | X = x_i] = \beta_1 x_i$$

and

$$\text{var}[Y_i | X = x_i] = \sigma^2 w_i^{-1} \quad (w_i > 0).$$

- (a) If the conditional distribution of Y given x is the Type III (scaled gamma) distribution,

$$f(y | x) = \frac{1}{a_x^p \Gamma(p)} y^{p-1} \exp(-y/a_x) \quad 0 \leq y < \infty, \quad p > 0,$$

where a_x is a function of x , and $w_i^{-1} = x_i^2$, prove that the maximum likelihood estimate of β_1 is also the weighted least squares estimate.

- (b) If the conditional distribution of Y given x is Poisson, and $w_i^{-1} = x_i$, show that the maximum likelihood estimate is just the weighted least squares estimate.

(Turner [1960])

2. Given the model $Y_i = \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \dots, n$), where the ε_i are independently distributed as $N(0, \sigma^2 w_i^{-1})$, $w_i > 0$, show how to predict x_0 for a given value Y_0 of Y . Describe briefly a method for constructing a confidence interval for x_0 .

3. Given the regression line

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where the ε_i are independent with $E[\varepsilon_i] = 0$ and $\text{var}[\varepsilon_i] = \sigma^2 x_i^2$, show that weighted least squares estimation is equivalent to ordinary least squares estimation for the model

$$\frac{Y_i}{x_i} = \beta_1 + \frac{\beta_0}{x_i} + \delta_i.$$

6.4 COMPARING STRAIGHT LINES

6.4.1 General Model

Suppose that we wish to compare K regression lines

$$Y = \alpha_k + \beta_k x + \varepsilon \quad (k = 1, 2, \dots, K),$$

where $E[\varepsilon] = 0$, and $\text{var}[\varepsilon]$ ($= \sigma^2$, say) is the same for each line. If we are given n_k pairs of observations (x_{ki}, Y_{ki}) ($i = 1, 2, \dots, n_k$) on the k th line, then we have the model

$$Y_{ki} = \alpha_k + \beta_k x_{ki} + \varepsilon_{ki} \quad (i = 1, 2, \dots, n_k), \tag{6.34}$$

where the ε_{ki} are independently and identically distributed as $N(0, \sigma^2)$. Writing

$$\mathbf{Y}' = (Y_{11}, Y_{12}, \dots, Y_{1n_1}, \dots, Y_{K1}, Y_{K2}, \dots, Y_{Kn_K}),$$

etc., we have $\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where

$$\mathbf{X}\boldsymbol{\gamma} = \left(\begin{array}{ccccccccc} 1 & 0 & \cdots & 0 & x_{11} & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & x_{12} & 0 & \cdots & 0 \\ \cdots & \cdots \\ 1 & 0 & \cdots & 0 & x_{1n_1} & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 & 0 & x_{21} & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & x_{22} & \cdots & 0 \\ \cdots & \cdots \\ 0 & 1 & \cdots & 0 & 0 & x_{2n_2} & \cdots & 0 \\ \hline \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{K1} \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{K2} \\ \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{Kn_K} \end{array} \right) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}. \quad (6.35)$$

Here \mathbf{X} is an $N \times 2K$ matrix of rank $2K$, where $N = \sum_{k=1}^K n_k$, so that we can test any hypothesis of the form $H: \mathbf{A}\boldsymbol{\gamma} = \mathbf{c}$ using the general regression theory of Chapter 4; examples of three such hypotheses are considered below. With a regression software package we do not need to derive any algebraic formulas. All we need to do is identify the \mathbf{X} matrix corresponding to each hypothesis and then compute RSS for that model. However, it is informative to derive least squares estimates and residual sums of squares for each hypothesis, and such derivations are relegated to Exercises 6c. The following examples can also be handled using the analysis of covariance method (Section 8.8).

EXAMPLE 6.2 (Test for parallelism) Suppose that we wish to test whether the K lines are parallel; then our hypothesis is $H_1: \beta_1 = \beta_2 = \cdots = \beta_K (= \beta$, say) or $\beta_1 - \beta_K = \beta_2 - \beta_K = \cdots = \beta_{K-1} - \beta_K = 0$; in matrix form this is

$$\left(\begin{array}{c|cccccc} & 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & 0 & \cdots & 0 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{array} \right) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0$$

or $\mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$, where \mathbf{A} is $(K-1) \times 2K$ of rank $K-1$. Applying the general regression theory with $q = K-1$, $n = N$, and $p = 2K$, the test statistic for H_1 is

$$F = \frac{(RSS_{H_1} - RSS/(K-1))}{RSS/(N-2K)}.$$

To obtain RSS_{H_1} we see that when H_1 is true, the design matrix, \mathbf{X}_1 say, is obtained by simply adding together the last K columns of \mathbf{X} in (6.35). \square

EXAMPLE 6.3 (Test for coincidence) To test whether the K lines are coincident, we consider $H_2: \alpha_1 = \alpha_2 = \cdots = \alpha_K (= \alpha$, say) and $\beta_1 = \beta_2 =$

$\dots = \beta_K$ ($= \beta$, say). Arguing as in Example 6.2, we see that H_2 is of the form $\mathbf{A}\gamma = \mathbf{0}$, where \mathbf{A} is now $(2K - 2) \times 2K$ of rank $2K - 2$. The F -statistic for testing H_2 is

$$F = \frac{\text{RSS}_{H_2} - \text{RSS}}{\text{RSS}/(N - 2K)}. \quad (6.36)$$

To compute RSS_{H_2} , we note that the design matrix for H_2 , \mathbf{X}_2 say, is obtained by adding together the first K columns of \mathbf{X} and then the last K columns.

In practice we would probably test for parallelism first and then, if H_1 is not rejected, test for H_2 (given that H_1 is true) using

$$F = \frac{\text{RSS}_{H_2} - \text{RSS}_{H_1}}{\text{RSS}_{H_1}/(N - K - 1)}.$$

If this also is not significant, then we can check this nested procedure using (6.36) as a final test statistic. \square

EXAMPLE 6.4 (Test for concurrence with x -coordinate known) Suppose that we wish to test the hypothesis H_3 that all the lines meet at a point on the y -axis ($x = 0$), that is, $H_3: \alpha_1 = \alpha_2 = \dots = \alpha_K$ ($= \alpha$, say). The F -statistic for testing H_3 is then

$$F = \frac{\text{RSS}_{H_3} - \text{RSS}}{\text{RSS}/(N - 2K)}.$$

The design matrix for H_3 is obtained by adding together the first K columns of \mathbf{X} . An estimate of α , the y coordinate of the point of concurrence is obtained automatically when the model for H_3 is fitted.

If we wish to test whether the lines meet on the line $x = c$, we simply replace x_{ki} by $x_{ki} - c$ in the theory above; we shift the origin from $(0, 0)$ to $(c, 0)$. In this case the y -coordinate of the concurrence point is still given by the estimate of α . \square

EXAMPLE 6.5 (Test for concurrence with x -coordinate unknown) The hypothesis that the lines meet at $x = \phi$, where ϕ is now unknown, takes the form $H: \alpha_k + \beta_k\phi = \text{constant}$ for $k = 1, 2, \dots, K$, or, eliminating ϕ ,

$$H: \frac{\alpha_1 - \bar{\alpha}}{\beta_1 - \bar{\beta}} = \dots = \frac{\alpha_K - \bar{\alpha}}{\beta_K - \bar{\beta}}.$$

Since H is no longer a linear hypothesis, we cannot use the general regression theory to derive a test statistic. However, an approximate test is provided by Saw [1966]. \square

6.4.2 Use of Dummy Explanatory Variables

Suppose that we wish to compare just two regression lines,

$$Y_{ki} = \alpha_k + \beta_k x_{ki} + \varepsilon_{ki} \quad (k = 1, 2; i = 1, 2, \dots, n_k).$$

By introducing the dummy variable d , where

$$d = \begin{cases} 1, & \text{if the observation comes from the second line,} \\ 0, & \text{otherwise,} \end{cases}$$

we can combine these two lines into a single model, namely,

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 x_i + (\alpha_2 - \alpha_1)d_i + (\beta_2 - \beta_1)(dx)_i + \varepsilon_i \\ &= \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \varepsilon_i, \end{aligned} \quad (6.37)$$

where

$$(x_i, Y_i) = \begin{cases} (x_{1i}, Y_{1i}), & i = 1, 2, \dots, n_1, \\ (x_{2i}, Y_{2i}), & i = n_1 + 1, \dots, n_1 + n_2 \end{cases}$$

and

$$d_i = \begin{cases} 0, & i = 1, 2, \dots, n_1, \\ 1, & i = n_1 + 1, \dots, n_1 + n_2. \end{cases}$$

We note that the model (6.37) is simply a reparameterization of (6.34) (with $K = 2$); the parameters α_1 , α_2 , β_1 , and β_2 are now replaced by $\gamma_0 = \alpha_1$, $\gamma_1 = \beta_1$, $\gamma_2 = \alpha_2 - \alpha_1$, and $\gamma_3 = \beta_2 - \beta_1$. For this new model, the various tests discussed above reduce to the following: $\gamma_3 = 0$ (parallelism), $\gamma_2 = 0$ (common intercept on the y -axis), and $\gamma_2 = \gamma_3 = 0$ (coincidence). In the case of three straight lines, we introduce two dummy variables:

$$\begin{aligned} d_1 &= \begin{cases} 1, & \text{if the observation comes from the second line,} \\ 0, & \text{otherwise;} \end{cases} \\ d_2 &= \begin{cases} 1, & \text{if the observation comes from the third line,} \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

and obtain

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 x_i + (\alpha_2 - \alpha_1)d_{i1} + (\alpha_3 - \alpha_1)d_{i2} + (\beta_2 - \beta_1)(d_1 x)_i \\ &\quad + (\beta_3 - \beta_1)(d_2 x)_i + \varepsilon_i \\ &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \varepsilon_i, \end{aligned}$$

say. Further generalizations are straightforward (see, e.g., Gujarati [1970]).

EXERCISES 6c

- Using the notation of Example 6.2, prove that the least squares estimates of α_k and the common slope β (under the null hypothesis of parallelism) are given by

$$\tilde{\alpha}_k = \bar{Y}_{k\cdot} - \tilde{\beta} \bar{x}_{k\cdot}$$

and

$$\tilde{\beta} = \frac{\sum_k \sum_i (Y_{ki} - \bar{Y}_{k\cdot})(x_{ki} - \bar{x}_{k\cdot})}{\sum_k \sum_i (x_{ki} - \bar{x}_{k\cdot})^2}.$$

Prove that

$$\text{RSS}_{H_1} = \sum_k \sum_i (Y_{ki} - \bar{Y}_{k\cdot})^2 - \tilde{\beta}^2 \sum_k \sum_i (x_{ki} - \bar{x}_{k\cdot})^2.$$

If $\hat{\beta}_k$ is the least squares estimate of β_k under the general model, prove that

$$\text{RSS}_{H_1} - \text{RSS} = \sum_k \hat{\beta}_k^2 \sum_i (x_{ki} - \bar{x}_{k\cdot})^2 - \tilde{\beta}^2 \sum_k \sum_i (x_{ki} - \bar{x}_{k\cdot})^2.$$

2. In Example 6.3 find the least squares estimates of α and β when H_2 is true.
 3. In Example 6.4 derive the following results.
- (a) Show that the least squares estimates of α and β_k , under H_3 , are given by

$$\begin{aligned} & \left(N - \frac{x_{1\cdot}^2}{\sum_i x_{1i}^2} - \dots - \frac{x_{K\cdot}^2}{\sum_i x_{Ki}^2} \right) \alpha' \\ &= \left(Y_{..} - \frac{x_{1\cdot} \sum_i Y_{1i} x_{1i}}{\sum_i x_{1i}^2} - \dots - \frac{x_{K\cdot} \sum_i Y_{Ki} x_{Ki}}{\sum_i x_{Ki}^2} \right) \end{aligned}$$

and

$$\beta'_k = \frac{\sum_i (Y_{ki} - \alpha') x_{ki}}{\sum_i x_{ki}^2} \quad (k = 1, 2, \dots, K).$$

- (b) When the values of x are the same for each line so that $n_k = n$ and $x_{ki} = x_i$ ($k = 1, 2, \dots, K$), prove that

$$\alpha' = \bar{Y}_{..} - \frac{\bar{x} \sum \sum Y_{ki} (x_i - \bar{x})}{K \sum (x_i - \bar{x})^2} = \bar{Y}_{..} - \bar{x} \frac{\sum_k \hat{\beta}_k}{K},$$

$$\text{var}[\alpha'] = \frac{\sigma^2 \sum x_i^2}{n K \sum (x_i - \bar{x})^2},$$

and

$$\text{RSS}_{H_3} = \sum_k \sum_i Y_{ki}^2 - \frac{(\sum \sum Y_{ki} x_i)^2}{\sum x_i^2} - (\alpha')^2 K n \frac{\sum (x_i - \bar{x})^2}{\sum x_i^2}.$$

Hint: $\bar{Y}_{..}$ and $\hat{\beta}_k$ are uncorrelated.

4. The examples in Section 6.4 are all special cases of the following problem. Consider the model

$$G : E[Y] = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 = \mathbf{X} \boldsymbol{\beta},$$

where \mathbf{X}_i is $n \times p_i$ of rank p_i ($i = 1, 2$). We want to test the hypothesis H that all the elements of β_2 are equal (to β , i.e., $\beta_2 = \mathbf{1}_{p_2}\beta$).

(a) If $\tilde{\beta}$ is the least squares estimate of β under H , show that

$$\tilde{\beta} = \frac{\mathbf{1}'_{p_2} \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_1) \mathbf{Y}}{\mathbf{1}'_{p_2} \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_1) \mathbf{X}_2 \mathbf{1}_{p_2}},$$

where $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$.

(b) If $\hat{\beta}_2$ is the least squares estimate of β_2 under G , prove that

$$\text{RSS}_H - \text{RSS}_G = (\hat{\beta}_2 - \tilde{\beta} \mathbf{1}_{p_2})' \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_1) \mathbf{Y}.$$

(c) If $\hat{\mathbf{Y}}_G$ is the fitted regression for G , prove that

$$\hat{\mathbf{Y}}_G = \mathbf{P}_1 \mathbf{Y} + (\mathbf{I}_n - \mathbf{P}_1) \mathbf{X}_2 \hat{\beta}_2$$

and

$$\text{RSS}_H - \text{RSS}_G = (\mathbf{1}_{p_2} \tilde{\beta} - \hat{\beta}_2)' \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{P}_1) \mathbf{X}_2 (\mathbf{1}_{p_2} \tilde{\beta} - \hat{\beta}_2).$$

Hint: For (a) and (b) apply Theorem 3.6 with \mathbf{Z} equal to \mathbf{X}_2 and $\mathbf{1}_{p_2}$, respectively. For part (c) note that $\text{RSS}_G - \text{RSS}_H = ||\hat{\mathbf{Y}}_G - \hat{\mathbf{Y}}_H||^2$.

6.5 TWO-PHASE LINEAR REGRESSION

Multiphase regression models are models that undergo one of more changes in their structure. Such models, including nonlinear models, are described in detail by Seber and Wild [1989: Chapter 9], and in this section we consider just a linear two-phase model. Looking at this simple model will give a good idea as to the kinds of problems one might encounter in this topic. Here the underlying model is a straight line, but it undergoes a change in slope at some point $x = \gamma$, where the change may be continuous, smooth, or abrupt. Also, γ may be (1) known, (2) unknown but known to lie between two observed values of x , or (3) completely unknown. Assuming that the change is continuous, we have the model

$$E[Y] = \begin{cases} \alpha_1 + \beta_1 x, & x \leq \gamma, \\ \alpha_2 + \beta_2 x, & x \geq \gamma, \end{cases}$$

where continuity requires that

$$\alpha_1 + \beta_1 \gamma = \alpha_2 + \beta_2 \gamma \quad (= \theta). \quad (6.38)$$

For example, x may be an increasing function of time, and at time t_c a treatment is applied that may possibly affect the slope of the regression line either

immediately or after a time lag. Following Sprent [1961], we call $x = \gamma$ the *changeover point* and θ the *changeover value*.

Known Changeover Point

Given n_1 observations on the first line and n_2 on the second, we have, using (6.38),

$$\begin{aligned} Y_{1i} &= \alpha_1 + \beta_1 x_{1i} + \varepsilon_{1i} \quad (i = 1, 2, \dots, n_1), \\ Y_{2i} &= \alpha_1 + \beta_1 \gamma + \beta_2 (x_{2i} - \gamma) + \varepsilon_{2i} \quad (i = 1, 2, \dots, n_2), \end{aligned}$$

where

$$x_{11} < x_{12} < \dots < x_{1n_1} < \gamma < x_{21} < x_{22} < \dots < x_{2n_2},$$

and γ is known. Writing $\mathbf{Y}_j = (Y_{j1}, Y_{j2}, \dots, Y_{jn_j})'$, $\mathbf{x}_j = (x_{j1}, \dots, x_{jn_j})'$, and $\boldsymbol{\varepsilon}_j = (\varepsilon_{j1}, \dots, \varepsilon_{jn_j})'$, for $j = 1, 2$, we have

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{x}_1 & \mathbf{0} \\ \mathbf{1}_{n_2} & \gamma \mathbf{1}_{n_2} & \mathbf{x}_2 - \gamma \mathbf{1}_{n_2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where \mathbf{X} is $(n_1 + n_2) \times 3$ of rank 3. Given a value of γ , we can use a regression software package to find the least squares estimate $\hat{\boldsymbol{\beta}} = (\hat{\alpha}_1, \hat{\beta}_1, \hat{\beta}_2)'$ and $(\mathbf{X}'\mathbf{X})^{-1}$. Assuming that $\boldsymbol{\varepsilon} \sim N_{n_1+n_2}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, we can apply several inferential procedures. For example, to estimate θ we can use

$$\hat{\theta} = \hat{\alpha}_1 + \gamma \hat{\beta}_1 = (1, \gamma, 0) \hat{\boldsymbol{\beta}} \quad (= \mathbf{a}' \hat{\boldsymbol{\beta}}, \text{ say}),$$

and we can construct a t -confidence interval for $\mathbf{a}'\boldsymbol{\beta}$ in the usual manner [cf. (4.17)].

If we want to test $H : \gamma = c$, where c lies between a pair of x values, say $x_{1n_1} < c \leq x_{21}$, then testing H is equivalent to testing whether two lines concur at $x = c$. This can be done using Example 6.4.

Unknown Changeover Point

In practice the changeover point γ will be unknown. Suppose, however, it is known that $x_{1n_1} < \gamma < x_{21}$; then γ can be estimated by [cf. equation (6.38)]

$$\hat{\gamma} = -\frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\beta}_1 - \hat{\beta}_2},$$

where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the usual least squares estimates for the k th line ($k = 1, 2$). Since $\hat{\gamma}$ is the ratio of two correlated normal variables, we can use Fieller's method for finding a confidence interval for γ as follows. Consider

$U = (\hat{\alpha}_1 - \hat{\alpha}_2) + \gamma(\hat{\beta}_1 - \hat{\beta}_2)$. Then $E[U] = 0$ and from (6.6) in Section 6.1.3 with $x_0 = \gamma$, we have

$$\begin{aligned}\text{var}[U] &= \sigma^2 \left\{ \frac{1}{n_1} + \frac{(\bar{x}_{1\cdot} - \gamma)^2}{\sum(x_{1i} - \bar{x}_{1\cdot})^2} + \frac{1}{n_2} + \frac{(\bar{x}_{2\cdot} - \gamma)^2}{\sum(x_{2i} - \bar{x}_{2\cdot})^2} \right\} \\ &= \sigma^2 w,\end{aligned}$$

say.

Arguing as in Section 6.1.2, a $100(1 - \alpha)\%$ confidence interval for γ is given by the roots of

$$[\hat{\alpha}_1 - \hat{\alpha}_2 + \gamma(\hat{\beta}_1 - \hat{\beta}_2)]^2 - F_{1,n-4}^\alpha S^2 w = 0,$$

that is, of

$$\begin{aligned}\gamma^2 &\left[(\hat{\beta}_1 - \hat{\beta}_2)^2 - F_{1,n-4}^\alpha S^2 \left\{ \sum_{k=1}^2 \frac{1}{\sum_i (x_{ki} - \bar{x}_{k\cdot})^2} \right\} \right] \\ &+ 2\gamma \left[(\hat{\alpha}_1 - \hat{\alpha}_2)(\hat{\beta}_1 - \hat{\beta}_2) + F_{1,n-4}^\alpha S^2 \left\{ \sum_{k=1}^2 \frac{\bar{x}_{k\cdot}}{\sum_i (x_{ki} - \bar{x}_{k\cdot})^2} \right\} \right] \\ &- F_{1,n-4}^n S^2 \left\{ \sum_{k=1}^2 \left[\frac{\bar{x}_{k\cdot}^2}{\sum_i (x_{ki} - \bar{x}_{k\cdot})^2} + \frac{1}{n_k} \right] \right\} + (\hat{\alpha}_1 - \hat{\alpha}_2)^2 = 0,\end{aligned}$$

where $S^2 = \text{RSS}/(n - 4)$ and $n = n_1 + n_2$.

If $\hat{\gamma}$ does not lie in the interval (x_{1n_1}, x_{21}) , then the experimenter must decide whether to attribute this to sampling errors (and the confidence interval for γ will shed some light on this) or to an incorrect assumption about the position of γ . When the position of γ is unknown, the problem becomes much more difficult, as it is now nonlinear. In this case the two-phase model can be written in the form (Hinkley [1971])

$$Y_i = \begin{cases} \theta + \beta_1(x_i - \gamma) + \varepsilon_i, & (i = 1, 2, \dots, \kappa), \\ \theta + \beta_2(x_i - \gamma) + \varepsilon_i, & (i = \kappa + 1, \dots, n), \end{cases}$$

where $x_1 < \dots < x_\kappa \leq \gamma < x_{\kappa+1} < \dots < x_n$, θ is the changeover value, and κ is now unknown and has to be estimated. Hinkley summarizes the maximum likelihood estimation procedure for estimating γ , θ , β_1 , β_2 , and κ : This is described in detail in Hudson [1966] and Hinkley [1969b]. Hinkley also provides approximate large sample confidence intervals for the parameters and gives large sample tests for the hypotheses $\beta_1 = \beta_2$ (no change in slope) and $\beta_2 = 0$. Another approach to testing $\beta_1 = \beta_2$ is given by Farley and Hinich [1970]. We note that Hudson's technique was generalized by Williams [1970] to the case of three-phase linear regression. For a general discussion of the problem, see Seber and Wild [1989: Section 9.3]. Instead of changing slope abruptly from one straight line to the next, the transition can be modeled smoothly. Methods for doing this are described by Seber and Wild [1989: Section 9.4].

6.6 LOCAL LINEAR REGRESSION

A wide range of methods are available for fitting a nonlinear curve to a scatter plot. One of the conceptually simplest methods is to fit a series of straight-line segments, thus giving a piecewise linear regression reminiscent of multiphase linear regression. In fitting a straight-line segment at a target point, x_0 , say, it is clear that data points close to x_0 should carry more weight than those farther away. One method of doing this is to use *lowess*, developed by Cleveland [1979] and implemented in S-PLUS as the function `lowess`. The name *lowess*, which stands for “locally weighted scatterplot smoother”, is essentially a robust version of a locally weighted regression in which a local regression model such as a polynomial is fitted at each point. If the underlying model is $Y = f(x) + \epsilon$, the linear version of lowess (the default for `lowess`) consists of carrying out a weighted least squares by minimizing $\sum_1^n w(x_0, x_i)(Y_i - \beta_0 - \beta_1 x_i)^2$ with respect to β_0 and β_1 . Although we use all the data to fit the line, we use only the fitted line to evaluate the fit at the single point x_0 , namely, $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (1, x_0)\hat{\beta}$. If \mathbf{W} is the diagonal matrix with i th diagonal element $w(x_0, x_i)$ and $\mathbf{X} = (\mathbf{1}_n, \mathbf{x})$, where $\mathbf{x} = (x_1, \dots, x_n)'$, we have from Section 3.10 that

$$\begin{aligned}\hat{f}(x_0) &= (1, x_0)'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \\ &= \sum_{i=1}^n l_i(x_0)Y_i,\end{aligned}$$

say. Then, using a Taylor expansion for $f(x_i)$ (Hastie et al. [2001: p. 170]), we get

$$\begin{aligned}E[\hat{f}(x_0)] &= \sum_{i=1}^n l_i(x_0)f(x_i) \\ &= f(x_0) \sum_{i=1}^n l_i(x_0) + f'(x_0) \sum_{i=1}^n (x_i - x_0)l_i(x_0) + R \\ &= f(x_0) + R,\end{aligned}$$

by the following lemma. Here the remainder term R involves second and higher-order derivatives of f and is typically small under suitable smoothness assumptions, so that $E[\hat{f}(x_0)] \approx f(x_0)$.

LEMMA

$$\sum_{i=1}^n l_i(x_0) = 1 \text{ and } \sum_{i=1}^n l_i(x_0)(x_i - x_0) = 0.$$

Proof. Taking expected values of the two expressions for $\hat{f}(x_0)$ yields

$$\begin{aligned}\sum_{i=1}^n l_i(x_0)(1, x_i)\beta &= (1, x_0)(\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{X} \beta \\ &= (1, x_0)\beta\end{aligned}$$

for all β and all x_0 . Setting $\beta_0 = 1$ and $\beta_1 = 0$ in the equation above we get $\sum_1^n l_i(x_0) = 1$. Also, setting $\beta_0 = 0$, $\beta_1 = 1$ and replacing x by $x - x_0$, we have

$$\sum_{i=1}^n l_i(x_0)(x_i - x_0) = (1, x_0 - x_0)(0, 1)' = 0. \quad \square$$

In applying this theory, we need to choose an appropriate weight function. Typically, we set $w(X_0, x_i) = K_\lambda(x_0 - x_i)$, where K is a *kernel function* and λ is an appropriate scale constant. A more general version of lowess (Cleveland and Devlin [1988]) is *loess*, which is also implemented in S-PLUS and can incorporate several regressors. For further details, see, for example, Hastie and Loader [1993] and Hastie et al. [2001].

MISCELLANEOUS EXERCISES 6

- Let $F = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 / S^2$, the F -statistic for testing $H : \beta_1 = 0$ for a straight line. Using the notation of Section 6.1.5, prove that

$$\tilde{x}_0 - \bar{x} = \frac{F}{F + (n - 2)}(\hat{x}_0 - \bar{x}).$$

(Hoadley [1970])

- Derive an F -statistic for testing the hypothesis that two straight lines intersect at the point (a, b) .
- Obtain an estimate and a confidence interval for the horizontal distance between two parallel lines.
- Show how to transform the following equation into a straight line so that α and β can be estimated by least squares:

$$y = \frac{\alpha\beta}{\alpha \sin^2 \theta + \beta \cos^2 \theta}.$$

(Williams [1959]: p. 19))

7

Polynomial Regression

7.1 POLYNOMIALS IN ONE VARIABLE

7.1.1 Problem of Ill-Conditioning

When faced with a well-behaved curved trend in a scatter plot a statistician would be tempted to try and fit a low-degree polynomial. Technical support for this decision comes from the Weierstrass approximation theorem (see Davis [1975: Chapter VI]), which implies that any continuous function on a finite interval can be approximated arbitrarily closely by a polynomial. This amounts to lumping any remainder terms from a Taylor series expansion of the unknown model function into the error term. Although the approximation can be improved by increasing the order of the polynomial, the cost is an increase in the number of unknown parameters and some oscillation between data points. However, another problem arises when fitting a high-degree polynomial, which we now discuss.

If we set $x_{ij} = x_i^j$ and $k = p - 1$ ($\leq n - 1$) in the general multiple linear regression model, we have the k th-degree [($k + 1$)th-order] polynomial model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i \quad (i = 1, 2, \dots, n). \quad (7.1)$$

Although it is theoretically possible to fit a polynomial of degree up to $n - 1$, a number of practical difficulties arise when k is large. First, for k greater than about 6, we find that the regression matrix \mathbf{X} associated with (7.1) becomes ill-conditioned (Section 11.4). For example, assuming that x_i is distributed

approximately uniformly on $[0, 1]$, then for large n we have (Forsythe [1957])

$$\begin{aligned} (\mathbf{X}'\mathbf{X})_{rs} &= n \sum_{i=1}^n x_i^r x_i^s \cdot \frac{1}{n} \\ &\approx n \int_0^1 x^r x^s dx \\ &= n \int_0^1 x^{r+s} dx \\ &= \frac{n}{r+s+1}. \end{aligned} \quad (7.2)$$

Hence $\mathbf{X}'\mathbf{X}$ is something like n times the matrix $[1/(r+s+1)]$, ($r, s = 0, 1, \dots, k$), which is the $(k+1) \times (k+1)$ principal minor of the *Hilbert matrix*

$$\mathbf{H} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

It is well known that \mathbf{H} is very ill-conditioned (Todd [1954, 1961]); for example, when $k = 9$, the inverse of \mathbf{H}_{10} , the 10×10 principal minor of \mathbf{H} , has elements of magnitude 3×10^{10} (Savage and Lukacs [1954]). Thus a small error of 10^{-10} in one element of $\mathbf{X}'\mathbf{Y}$ will lead to an error of about 3 in an element of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Two things can be done to help overcome this ill-conditioning and the instability in the computations. The first is to “normalize” the x_i so that they run from -1 to $+1$. The normalized x is given by

$$x' = \frac{2x - \max(x_i) - \min(x_i)}{\max(x_i) - \min(x_i)}.$$

The second is to use orthogonal polynomials, which we now discuss.

7.1.2 Using Orthogonal Polynomials

General Statistical Properties

Consider the model

$$Y_i = \gamma_0 \phi_0(x_i) + \gamma_1 \phi_1(x_i) + \cdots + \gamma_k \phi_k(x_i) + \varepsilon_i,$$

where $\phi_r(x_i)$ is an r th-degree polynomial in x_i ($r = 0, 1, \dots, k$), and the polynomials are orthogonal over the x -set: namely,

$$\sum_{i=1}^n \phi_r(x_i) \phi_s(x_i) = 0 \quad (\text{all } r, s, r \neq s). \quad (7.3)$$

Then $\mathbf{Y} = \mathbf{X}\gamma + \epsilon$, where

$$\mathbf{X} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_k(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_k(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_k(x_n) \end{pmatrix}$$

has mutually orthogonal columns, and

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_i \phi_0^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum_i \phi_1^2(x_i) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sum_i \phi_k^2(x_i) \end{pmatrix}.$$

Hence, from $\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we have

$$\hat{\gamma}_r = \frac{\sum_i \phi_r(x_i) Y_i}{\sum_i \phi_r^2(x_i)} \quad (r = 0, 1, \dots, k), \quad (7.4)$$

which holds for all k . The orthogonal structure of \mathbf{X} implies that the least squares estimate of γ_r ($r \leq k$) is independent of the degree k of the polynomial (cf. Section 3.6)—a very desirable property. Since $\phi_0(x_i)$ is a polynomial of degree zero, we can set $\phi_0(x) \equiv 1$ and obtain

$$\hat{\gamma}_0 = \frac{\sum_i 1 \cdot Y_i}{\sum_i 1} = \bar{Y}.$$

The residual sum of squares is then

$$\begin{aligned} \text{RSS}_{k+1} &= (\mathbf{Y} - \mathbf{X}\hat{\gamma})'(\mathbf{Y} - \mathbf{X}\hat{\gamma}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\gamma}'\mathbf{X}'\mathbf{X}\hat{\gamma} \\ &= \sum_i Y_i^2 - \sum_{r=0}^k \left[\sum_i \phi_r^2(x_i) \right] \hat{\gamma}_r^2 \\ &= \sum_i (Y_i - \bar{Y})^2 - \sum_{r=1}^k \left[\sum_i \phi_r^2(x_i) \right] \hat{\gamma}_r^2. \end{aligned} \quad (7.5)$$

If we wish to test $H : \gamma_k = 0$ [which is equivalent to testing $\beta_k = 0$ in equation (7.1)], then the residual sum of squares for the model H is

$$\begin{aligned} \text{RSS}_k &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{r=1}^{k-1} \left[\sum_i \phi_r^2(x_i) \right] \hat{\gamma}_r^2 \\ &= \text{RSS}_{k+1} + \left[\sum_i \phi_k^2(x_i) \right] \hat{\gamma}_k^2, \end{aligned} \quad (7.6)$$

and the appropriate F -statistic is

$$\begin{aligned} F &= \frac{\text{RSS}_k - \text{RSS}_{k+1}}{\text{RSS}_{k+1}/(n - k - 1)} \\ &= \frac{\sum_i \phi_k^2(x_i) \hat{\gamma}_k^2}{\text{RSS}_{k+1}/(n - k - 1)}. \end{aligned}$$

The question arises as to how we choose the degree, K say, of our polynomial. We note from (7.6) that RSS_{k+1} , the residual sum of squares for a polynomial of degree k , decreases as k increases. Ideally, RSS_{k+1} decreases consistently at first and then levels off to a fairly constant value, at which stage it is usually clear when to stop (see, e.g., Hayes [1970: Section 8, Example A]). In cases of doubt, we can test for significance the coefficient of the last polynomial added to the model; this is the *forward selection procedure* procedure with a predetermined order for the regressors (although it is used only at an appropriate stage of the fitting, not necessarily right from the beginning). However, this test procedure should be used cautiously because it may lead to stopping prematurely. For example, we have the possibility that RSS_{k+1} may level off for awhile before decreasing again. To safeguard against these contingencies, it is preferable to go several steps beyond the first insignificant term, and then look carefully at RSS.

Another test procedure that can be used is the *backward elimination procedure*. In this case the maximum degree that will be fitted is determined in advance and then the highest-degree terms are eliminated one at a time using the F -test; the process stops when there is a significant F -statistic. The procedure is more efficient than forward selection, and it is suggested that the best significance level to use at each step is $\alpha \approx 0.10$ (Kennedy and Bancroft [1971: p. 1281]). However, there remains the problem of deciding the maximum degree to be fitted. Unfortunately, the forward and backward procedures do not necessarily lead to the same answer. These procedures are discussed in more detail in Section 12.4.

Generating Orthogonal Polynomials

Orthogonal polynomials can be obtained in a number of ways. Following Forsythe [1957], a pioneer in the field, Hayes [1974] suggested using the three-term recurrence relationship

$$\phi_{r+1}(x) = 2(x - a_{r+1})\phi_r(x) - b_r\phi_{r-1}(x) \quad (7.7)$$

beginning with initial polynomials

$$\phi_0(x) = 1 \quad \text{and} \quad \phi_1(x) = 2(x - a_1).$$

Here x is normalized so that $-1 \leq x \leq +1$, and the a_{r+1} and b_r are chosen to make the orthogonal relations (7.3) hold, namely,

$$a_{r+1} = \frac{\sum_{i=1}^n x_i \phi_r^2(x_i)}{\sum_{i=1}^n \phi_r^2(x_i)} \quad (7.8)$$

and

$$b_r = \frac{\sum_{i=1}^n \phi_r^2(x_i)}{\sum_{i=1}^n \phi_{r-1}^2(x_i)}, \quad (7.9)$$

where $r = 0, 1, 2, \dots, k-1$, $b_0 = 0$, and $a_1 = \bar{x}$. (Forsythe used the range -2 to $+2$ and the factor unity instead of the factor 2 given in equation (7.7). These two differences in detail are essentially compensatory, for there is an arbitrary constant factor associated with each orthogonal polynomial; cf. Hayes [1969].) We note that the method of generating the ϕ_r is similar to Gram-Schmidt orthogonalization, with the difference that only the preceding two polynomials are involved at each stage. A computer program based on Forsythe's method is given by Cooper [1968, 1971a,b]. Each $\phi_r(x)$ can be represented in the computer by its values at the (normalized) points x_i or by its a 's and b 's. However, Clenshaw [1960] has given a useful modification of the method above in which each $\phi_r(x)$ is represented by the coefficients $\{c_j^{(r)}\}$ in its Chebyshev series form, namely,

$$\phi_r(x) = \frac{1}{2} c_0^{(r)} T_0(x) + c_1^{(r)} T_1(x) + \cdots + c_r^{(r)} T_r(x), \quad (7.10)$$

where

$$T_{r+1}(x) = 2xT_r(x) - T_{r-1}(x) \quad (r = 1, 2, \dots),$$

starting with $T_0(x) = 1$ and $T_1(x) = x$. The recurrence (7.7) is now carried out in terms of the coefficients $\{c_j^{(r)}\}$, and the fitted polynomial can be expressed in terms of Chebyshev polynomials, namely,

$$\begin{aligned} \hat{Y} &= \hat{f}_k(x) \\ &= \frac{1}{2} d_0^{(k)} T_0(x) + d_1^{(k)} T_1(x) + \cdots + d_k^{(k)} T_k(x), \end{aligned} \quad (7.11)$$

say. The appropriate recurrence relationships for carrying out these computations are [by substituting (7.10) in (7.7)]

$$c_j^{(r+1)} = c_{j+1}^{(r)} + c_{|j-1|}^{(r)} - 2a_{r+1}c_j^{(r)} - b_rc_j^{(r-1)}, \quad (7.12)$$

and by substituting (7.10) and (7.11) in the equation

$$\hat{f}_{k+1}(x) = \hat{f}_k(x) + \hat{\gamma}_{k+1}\phi_{k+1}(x),$$

we get

$$d_j^{(r)} = d_j^{(r-1)} + \hat{\gamma}_r c_j^{(r)}, \quad (7.13)$$

where $j = 0, 1, \dots, r+1$ and $c_j^{(r)} = d_j^{(r)} = 0$ for $j > r$.

Although the modification above takes about two to three times as long as Forsythe's method, the computing time is generally small in either case. Therefore, because time is not generally the decisive factor, the modification

is recommended by Clenshaw and Hayes [1965: p. 168] as it presents a convenient output in concise form. The $c_j^{(r)}$, for example, carry more information than the a_r and b_r . Hayes [1969] also shows that the recurrence relation (7.7) can operate entirely in terms of the coefficients $c_j^{(r)}$ and certain of the quantities $\sum_i \phi_r(x_i) T_s(x_i)$. If these numbers are stored, then we need not store either the x_i or the $\phi_r(x_i)$. Another useful feature of Clenshaw's modification, pointed out by Hayes [1970: p. 52], is that the coefficients $c_j^{(k)}$ (for increasing j and fixed k) behave in a very similar manner to RSS_k (for increasing k); they decrease steadily, except possibly at the start and then settle down to constant values. This feature, illustrated by Examples A and B in Section 8 of Hayes [1970], provides additional "evidence" for determining the degree of the polynomial fit.

When the coefficients $d_j^{(k)}$ in (7.11) have been computed, \hat{f} can be evaluated at any desired value of x by a procedure given by Clenshaw [1955]. In this we first compute the auxiliary numbers g_k, g_{k-1}, \dots, g_0 from the recurrence relation

$$g_i = 2xg_{i+1} - g_{i+2} + d_i^{(k)}$$

starting with $g_{k+1} = g_{k+2} = 0$. The required value of \hat{f} is then given by

$$\hat{f}_k(x) = \frac{1}{2}(g_0 - g_2). \quad (7.14)$$

An error analysis of Clenshaw's modification is given by Clenshaw and Hayes [1965: p. 169]. In particular, they give a method for estimating the numerical error in each $\hat{\gamma}_j$; this error can then be used to deduce an error estimate for $d_j^{(r)}$ in equation (7.11), using (7.13) and the computed values of $c_j^{(r)}$.

Equally Spaced x -Values

Suppose that the x -values are equally spaced so that they can be transformed to

$$x_i = i - \frac{1}{2}(n + 1) \quad (i = 1, 2, \dots, n). \quad (7.15)$$

Then we have the following system of orthogonal polynomials (generally ascribed to Chebyshev):

$$\begin{aligned} \phi_0(x) &= 1 \\ \phi_1(x) &= \lambda_1 x \\ \phi_2(x) &= \lambda_2 (x^2 - \frac{1}{12}(n^2 - 1)) \\ \phi_3(x) &= \lambda_3 (x^3 - \frac{1}{20}(3n^2 - 7)x) \\ \phi_4(x) &= \lambda_4 (x^4 - \frac{1}{14}(3n^2 - 13)x + \frac{3}{560}(n^2 - 1)(n^2 - 9)), \text{ etc.,} \end{aligned}$$

where the λ_r are chosen so that the values $\phi_r(x_i)$ are all positive and negative integers. These polynomials are tabulated extensively in Pearson and Hartley

[1970] for $n = 1(1)52$ and $r = 1(1)6$ ($r \leq n - 1$); a section of their table is given in Table 7.1.

To illustrate the use of this table, suppose that $n = 3$. Then $x_i = -1, 0, 1$, $\phi_0(x) = 1$, $\phi_1(x) = \lambda_1 x = x$, $\phi_2(x) = \lambda_2(x - \frac{2}{3}) = 3x^2 - 2$ and the fitted polynomial is

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 (3x^2 - 2),$$

where

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y}, \\ \hat{\beta}_1 &= \frac{\sum_i \phi_1(x_i) Y_i}{\sum_i \phi_1^2(x_i)} = \frac{1}{2} \{(-1)Y_1 + (0)Y_2 + (1)Y_3\} = \frac{1}{2}(Y_3 - Y_1),\end{aligned}$$

and

$$\hat{\beta}_2 = \frac{1}{6}(Y_1 - 2Y_2 + Y_3).$$

Also, the residual sum of squares is given by [equation (7.5)]

$$\begin{aligned}\text{RSS}_3 &= \sum_{i=1}^3 (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum_i \phi_1^2(x_i) - \hat{\beta}_2^2 \sum_i \phi_2^2(x_i) \\ &= \sum (Y_i - \bar{Y})^2 - 2\hat{\beta}_1^2 - 6\hat{\beta}_2^2.\end{aligned}$$

The theory of this section and the tables can be used for fitting polynomials up to degree 6. However, its main application is in the theory of experimental design, where various sums of squares are sometimes split into linear, quadratic, etc. components. A simple method for generating the orthogonal polynomials iteratively when $x = 0, 1, \dots, n - 1$, due to Fisher and Yates [1957], is described by Jennrich and Sampson [1971].

Table 7.1 Values of the orthogonal polynomials, $\phi_r(x)$, for the equally spaced x -data of equation (7.15)

$n = 3$		$n = 4$			$n = 5$			
ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_3	ϕ_1	ϕ_2	ϕ_3	ϕ_4
-1	1	-3	1	-1	-2	2	-1	1
0	-2	-1	-1	3	-1	-1	2	-4
1	1	1	-1	-3	0	-2	0	6
		3	1	1	1	-1	-2	-4
					2	2	1	1
$\sum_i \phi_r^2(x_i)$	2	6	20	4	20	10	14	10
λ_r	1	3	2	1	$\frac{10}{3}$	1	1	$\frac{5}{6}$
								$\frac{35}{12}$

Application of Constraints

A possible requirement in curve fitting is for the fitting function $f(x)$, and possibly its derivatives also, to take specified values at certain values of x . For example, the function may be required to pass through the origin or to join smoothly onto a straight line at some point, or we may wish to fit the data in two adjoining ranges separately, forcing continuity up to some order of derivative at the joint. For examples, see Clenshaw and Hayes [1965], Payne [1970] and Hayes [1974]. If the polynomial is constrained to be nonnegative, nondecreasing, or convex, then the quadratic programming type method of Hudson [1969] can be used for fitting the polynomial.

7.1.3 Controlled Calibration

Suppose that we have fitted a polynomial calibration curve (see Section 6.1.5)

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_k x^k \\ &= \mathbf{x}' \hat{\boldsymbol{\beta}},\end{aligned}$$

say. We observe a value of Y , Y_* say, and we want to predict the corresponding value of x , say ξ . If $\mathbf{x}'_\xi = (1, \xi, \xi^2, \dots, \xi^k)$, then an estimate of ξ is found by solving $Y_* = \mathbf{x}'_\xi \hat{\boldsymbol{\beta}}$ for ξ . The solution will be unique if the polynomial is monotonic in the region of interest. To construct a confidence interval we can proceed as in Section 5.3.1 and consider the distribution of $Y_* - \hat{Y}_\xi$, where $\hat{Y}_\xi = \mathbf{x}'_\xi \hat{\boldsymbol{\beta}}$. From (5.26) this is $N(0, \sigma^2[1 + v_\xi])$, where $v_\xi = \mathbf{x}'_\xi (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_\xi$. Then

$$T = \frac{Y_* - \hat{Y}_\xi}{S \sqrt{1 + v_\xi}} \sim t_{n-k-1}$$

and a $100(1 - \alpha)\%$ confidence interval for ξ is the set of all ξ satisfying

$$|T| \leq t_{n-k-1}^{(1/2)\alpha}.$$

Brown [1993: pp. 47–88] shows that this interval is essentially a profile likelihood interval and generalizes the result to replicated data. He also extends the theory to orthogonal polynomials.

7.2 PIECEWISE POLYNOMIAL FITTING

7.2.1 Unsatisfactory Fit

Sometimes a polynomial fit is unsatisfactory even when orthogonal polynomials up to, say, degree 20 are fitted. This lack of fit is usually revealed in several ways. One symptom is the failure of RSS_k to settle down to a constant value as k increases; the residual sum of squares may, for example, just continue to

decrease slowly. Another symptom is the behavior of the residuals: a residual plot (see Chapter 10) of $r_i = Y_i - \hat{Y}_i$ versus x_i will continue to exhibit a systematic pattern instead of a random one (see, e.g., Hayes [1970: Section 8, Example E]). In the worst cases there will be waves in the fitted curve which eventually become oscillations between adjacent data points, usually near the ends of the range. These difficulties most frequently arise when the behavior of the underlying function is very different in one part of the range from another. It may, for example, be varying rapidly in one region and varying slowly in another.

An alternative approach to the problem is to divide up the range of x into segments and fit a low-degree polynomial in each segment (e.g., Seber and Wild [1989: Section 9.3.3]. There are several ways of doing this, and the most useful method employs the theory of *spline functions* pioneered by Schoenberg [1946].

7.2.2 Spline Functions

The method of splines consists of dividing up the range of x into segments with join points called *knots*. A polynomial of a fixed degree is then fitted to each segment with constraints applied to ensure appropriate continuity at the knots. Questions arise as to the number and placement of the knots, the degree of the polynomial, and the appropriate continuity constraints. The term *spline* is borrowed from a mechanical device that was used to draw cross sections of ships' hulls. The mechanical spline was a flexible piece of wood which was forced to pass through certain fixed points and otherwise allowed to find its natural position.

More formally, we define the *spline function* $s(x)$ of order M (degree $M-1$), with knots $\xi_1, \xi_2, \dots, \xi_K$ (where $\xi_1 < \xi_2 < \dots < \xi_K$) and having domain $[a, b]$ ($-\infty \leq a < \xi_1, \xi_K < b \leq \infty$), to be a function with the following properties:

1. In each of the intervals

$$a < x \leq \xi_1, \quad \xi_{j-1} \leq x \leq \xi_j \quad (j = 2, 3, \dots, K), \quad \text{and} \quad \xi_K \leq x < b,$$

$s(x)$ is a polynomial of degree $M-1$ at most.

2. $s(x)$ and its derivatives up to order $(M-2)$ are continuous. (When a and b are finite, which is the usual case in practice, some authors call $\xi_0 = a$ and $\xi_{K+1} = b$ knots also, a convention that we shall adopt.)

We usually refer to the splines described above as *regression splines*.

The cubic spline ($M=4$) is a satisfactory function for fitting data in most situations, and second-derivative continuity is usually adequate for most practical problems. Apparently, cubic splines are claimed to be the lowest-order spline for which knot discontinuity, (in this case third order discontinuity), is

not visible to the human eye (Hastie et al. [2001: p. 120]). In cubic splines there are four parameters for each cubic and three constraints at each knot, thus giving $4(K + 1) - 3K = K + 4$ free parameters to be estimated.

Unfortunately, a restricted least squares approach using the constraints at the knots is cumbersome, and more parameters need to be estimated than the “minimum” $K + 4$. However, any cubic spline with knots ξ_j has a unique representation in the form

$$s(x) = \sum_{h=0}^3 \alpha_h x^h + \sum_{j=1}^K \beta_j (x - \xi_j)_+^3, \quad (7.16)$$

where

$$u_+ = \max(0, u) = \begin{cases} u, & u \geq 0, \\ 0, & u \leq 0. \end{cases}$$

This representation contains $K + 4$ basis functions (four power terms and K one-sided cubics), the smallest number by which the general cubic spline with K knots can be represented. We can reduce the number of parameters from $K + 4$ to K by constraining second- and third-order derivatives to be zero at ξ_0 and ξ_{K+1} , thus forcing the spline to be linear on $[\xi_0, \xi_1]$ and $[\xi_K, \xi_{K+1}]$. This allows four more knots to be used. The modified spline is called a *natural cubic spline* and the additional constraints are called the *natural boundary conditions*.

The truncated power series approach of (7.16) has a certain algebraic simplicity, but computationally it has some problems. For example, each cubic term is evaluated at all points to the right of its knot, and the buildup of powers of large numbers leads to the ill-conditioning alluded to at the beginning of this chapter. Equation (7.16) is therefore not recommended for computational use.

Instead of using the truncated power series basis, a better approach computationally is to use the *B-spline basis*. This is defined for any order, and the reader is referred to the texts of de Boor [1978], Schumaker [1981] and Diercx [1993] for the underlying theory. However, we shall follow Eilers and Marx [1996: pp. 90–91] and provide a gentle approach to the topic.

An M th-order basis spline is an $(M - 1)$ th-degree piecewise polynomial which is positive in the interior of a domain of M intervals spanned by $M + 1$ consecutive knots, and zero elsewhere. It is made up of M pieces, one spanning each of the M intervals. To generate the basis, we start the first spline $M - 1$ (artificially created) intervals to the left of the lower boundary point ξ_0 so that the “important” part of the function, that is, the part actually between the boundary points, is positive on (ξ_0, ξ_1) . The second basis spline is then defined on a similar range but shifted one interval to the right. We keep shifting one interval to the right until we get the last spline, which is positive on (ξ_K, ξ_{K+1}) and extends for $M - 1$ artificially created intervals to the right of the upper boundary point ξ_{K+1} . To define the complete basis, we therefore

need to introduce $M - 1$ additional knots $\xi_{-(M-1)}, \dots, \xi_{i-2}, \xi_{-1}$ at the lower end and $M - 1$ knots $\xi_{K+2}, \xi_{K+3}, \dots, \xi_{K+M}$ at the upper end. These must satisfy

$$\xi_{-(M-1)} \leq \dots \leq \xi_{-1} \leq \xi_0 \quad \text{and} \quad \xi_{K+1} \leq \xi_{K+2} \leq \dots \leq \xi_{K+M}.$$

Notationally, it is convenient to relabel the knots as $\tau_{j+M} = \xi_j$ for $j = -(M-1), \dots, 0, \dots, K+M$, so that we have knots $\tau_1, \dots, \tau_{K+2M}$. For $m \leq M$ we can define the family of B -splines (sometimes called *fundamental splines*) as

$$B_{j,m}(x) = (\tau_{j+m} - \tau_j) \sum_{h=j}^{j+m} \frac{(x - \tau_h)_+^{m-1}}{\prod_{s=j, s \neq h}^{j+m} (\tau_h - \tau_s)} \quad (j = 1, 2, \dots, K+M),$$

where $B_{j,m}(x)$, termed the *jth basis function of order m*, is positive in the interval (τ_j, τ_{j+m}) and has a single local maximum. Although the divided difference formula above is complicated, there is a recursive relationship which is convenient for computation, namely (de Boor [1978]),

$$B_{j,1}(x) = \begin{cases} 1, & \tau_j \leq x \leq \tau_{j+1}, \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, 2, \dots, K+2M-1$ and

$$B_{j,m}(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1}(x) + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1}(x) \quad (7.17)$$

for $j = 1, 2, \dots, K+2M-m$. For a given m , it can be shown that the basis above spans the space of piecewise polynomials of order m . Thus if $s(x)$ is a piecewise polynomial of order m , we can write

$$s(x) = \sum_{j=1}^{K+m} \gamma_j B_{j,m}(x) \quad (7.18)$$

for some γ_j .

When $M = 4$, the functions $B_{j,4}(x)$ ($j = 1, 2, \dots, K+4$) are the $K+4$ cubic B -spline basis functions for the knot sequence $\xi = (\xi_1, \dots, \xi_K)'$. We can use the recursive formulae to generate the B -spline basis for any order spline.

One unanswered question is the choice of the extra knots. We recall that the M th-order spline has $(M-2)$ th-order continuity at the knots ξ . For example, a cubic spline basis has continuous first- and second-order derivatives but a discontinuous third-order derivative at an interior knot. If we duplicate an interior knot, it transpires that the resulting basis still spans the space of piecewise polynomials but with one less continuous derivative at the knot. If we repeat that knot three times, then we have two fewer continuous derivatives

there. For example, if the spline is cubic, repeating ξ_j three times means that the first derivative is discontinuous at ξ_j ; repeating it four times means that the spline itself is discontinuous at ξ_j . Thus for an M th-order spline, repeating an interior knot M times means that there is a discontinuity at the knot. We can now answer our question about the extra knots. By making the $M - 1$ knots outside each boundary point the same as the boundary point, we make the spline discontinuous at the boundary points and thus undefined beyond the boundaries. We note that some care is needed in the interpretation of the recursive formulas above when knots are duplicated. Any term with a zero denominator is set equal to zero. Some nice figures graphing the spline basis functions for $M = 1, 2, 3, 4$ are given by Hastie et al. [2001: p. 162].

We shall now briefly discuss the process of fitting the spline to the data (x_i, Y_i) , $i = 1, 2, \dots, n$. The model (7.18) is linear in the unknown parameters γ_i , so we can fit the usual linear model (without an intercept) using the values of $B_{j,m}(x_i)$ and obtain the squares estimates of these parameters. The fact that each spline has local support, being only nonzero over M intervals, means that there are a lot of zeros in the design matrix \mathbf{X} , which has a band-like structure.

The question of the number and placing of the knots needs to be considered, and Wold [1974] makes the following useful recommendations. Knots should be located at data points and should be a few as possible with at least four or five data points between each knot. No more than one extremum and one inflection should fall between knots (as a cubic cannot approximate more variations), with extrema centered in the intervals and the inflections located near the knots. Eubank [1984] gives a number of test procedures and diagnostics for assessing the appropriateness of the knot selection. The S-PLUS function `bs(x, degree=m-1, knots=c(0.1, 0.2, ...))`, with K specified interior knots 0.1, 0.2, etc. computes the values of the $K + m$ B-spline basis functions of degree $m - 1$, and returns the $n \times (K + m)$ design matrix; if the degree is not mentioned, the default is $m - 1 = 3$. Alternatively, we can specify `df` instead of `knots`, which places $df - m$ knots uniformly along the range of x . The design matrix is then $n \times df$. In both situations one can also choose whether or not to include an intercept. The function `n()` does a similar thing with natural splines.

7.2.3 Smoothing Splines

We now discuss a spline basis method that avoids the problem of knot selection completely by using a “maximal” set of knots. If we wanted to fit a smooth function f to the data (x_i, Y_i) ($i = 1, \dots, n$), we could use the least squares criterion of minimizing $\text{RSS}(f) = \sum_{i=1}^n [Y_i - f(x_i)]^2$ to get the best-fitting function. However, if we choose our function as one that passes through each data point [i.e., $f(x_i) = Y_i$ for all i], then $\text{RSS}(f)$ will be zero. Such a function with this property is called an *interpolating function*, as it interpolates the points (x_i, Y_i) . The simplest such f could be obtained simply by joining up

consecutive points with straight lines. Unfortunately, this piecewise linear graph would not be smooth, as its derivatives do not exist at each point. Alternatively, we could use a piecewise polynomial and impose the condition of two continuous derivatives at each point; this would give a smoother-looking graph with curves between consecutive points. Unfortunately f could end up looking quite “wiggly” or *rough* between consecutive points. What we would like to do is impose a penalty function that measures the degree of roughness. Since we would not want our measure to be affected by the addition of a constant or a linear function, we could utilize the second derivative f'' . Although various measures of the magnitude of f'' could be considered, a very useful one is the global measure $\int_a^b \{f''(x)\}^2 dx$. Combining the two ideas of least squares and roughness leads to the criterion of finding f which minimizes

$$\text{RSS}(f, \lambda) = \sum_{i=1}^n [Y_i - f(x_i)]^2 + \lambda \int_a^b [f''(t)]^2 dt, \quad (7.19)$$

the *penalized residual sum of squares*. The first term, which some authors (e.g., Eubank [1999]) divide by n , measures closeness of fit, while the second term penalizes curvature, with a fixed *smoothing parameter* λ providing a trade-off between the two criteria. If $\lambda = 0$, then f is any interpolating function, while if $\lambda = \infty$, there is no second derivative and we get the least squares fit of a straight line. As λ ranges between 0 and ∞ , f can vary from very rough to very smooth.

We note that (7.19) can also be motivated by approximating the remainder from a Taylor series expansion of $f(x)$ (Eubank [1999: p. 228–229]) or using a Bayes regression approach (cf. Eubank [1999: Section 5.6]).

Suppose that f is any curve with two continuous derivatives, and let g be any natural cubic spline that interpolates the n points $(x_i, f(x_i))$. Then $g(x_i) = f(x_i)$ for all i , so that the first term of (7.19) is the same for both functions. However, it can be shown that (Green and Silverman [1994: Chapter 2])

$$\int_a^b [g(x)]^2 dx \leq \int_a^b [f(x)]^2 dx$$

with strict inequality if f is not a natural cubic spline. Thus if f is any twice-differentiable function we can always find a natural cubic spline which has a smaller $\text{RSS}(f, \lambda)$. This means that when minimizing $\text{RSS}(f, \lambda)$, we only need to consider natural cubic splines with knots at each of the x_i . Although we seem to have too many knots, particularly when n is large and the points are close together, the penalty on the spline coefficients has a “linearizing” effect on the spline, so that some of the powers of the polynomials are reduced. Since $f(x)$ is a natural spline, we find that we can write

$$f(x) = \sum_{k=1}^n \theta_k N_k(x),$$

where the $N_k(x)$ are n basis functions for the family of natural splines. Setting $\{\mathbf{N}\}_{ik} = N_k(x_i)$ and $\{\mathbf{V}_N\}_{jk} = \int_a^b N_j''(x) N_k''(x) dx$, we see that (7.19) reduces to

$$\text{RSS}(f, \lambda) = (\mathbf{y} - \mathbf{N}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}' \mathbf{V}_N \boldsymbol{\theta}. \quad (7.20)$$

Differentiating (7.20) with respect to $\boldsymbol{\theta}$ and using A.8, we obtain

$$-2\mathbf{N}'\mathbf{y} + 2\mathbf{N}'\mathbf{N}\boldsymbol{\theta} + 2\lambda \mathbf{V}_N \boldsymbol{\theta} = 0,$$

which has solution

$$\hat{\boldsymbol{\theta}} = (\mathbf{N}'\mathbf{N} + \lambda \mathbf{V}_N)^{-1} \mathbf{N}'\mathbf{y}, \quad (7.21)$$

which is a form of ridge regression (see Section 10.7.3).

Following the discussion in Hastie et al. [2001: p. 163], we note that in practice, it is computationally more convenient not to use natural splines but rather to use the larger space of unconstrained B -splines. Writing

$$f(x) = \sum_{j=1}^n \gamma_j B_j(x),$$

the solution looks like (7.21), namely,

$$\hat{\boldsymbol{\gamma}} = (\mathbf{B}'\mathbf{B} + \lambda \mathbf{V}_B)^{-1} \mathbf{B}'\mathbf{y},$$

except that the $n \times n$ matrix \mathbf{N} is replaced by the $(n+4) \times n$ matrix \mathbf{B} , and the $(n+4) \times (n+4)$ penalty matrix \mathbf{V}_B replaces the $n \times n$ matrix \mathbf{V}_N . It turns out that, rather conveniently, the penalty term automatically imposes boundary derivative constraints. When n is large, it is not necessary to use all n interior knots and any suitable thinning strategy will work just as well. For example, the S-PLUS function `smooth.spline` uses an approximately logarithmic strategy, so that if $n < 50$, all the knots are used, while if $n = 5000$, only 204 knots are used.

Determining the Smoothing Parameter

We shall now consider methods for finding λ , using either a subjective estimate or an estimate computed automatically. It transpires that λ controls the trade-off between the bias and the variance of the fitted function \hat{f}_λ : the larger the value of λ , the smoother the curve (with a resulting smaller variance) and the larger the bias. If $\hat{\mathbf{Y}} = (\hat{f}_\lambda(x_1), \dots, \hat{f}_\lambda(x_n))'$, then from (7.21) we have

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda \mathbf{V}_N)^{-1} \mathbf{N}'\mathbf{Y} \\ &= \mathbf{S}_\lambda \mathbf{Y}, \end{aligned} \quad (7.22)$$

where the $n \times n$ positive-definite matrix \mathbf{S}_λ is known as the *smoother matrix*. This matrix is analogous to the projection (hat) matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$

arising from $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$. As \mathbf{S}_λ connects \mathbf{Y} to $\hat{\mathbf{Y}}$, it is also referred to as a *hat matrix*. Since the trace of \mathbf{P} equals p , being both the number of parameters in the linear model $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ and essentially the “degrees of freedom,” we can define in an analogous fashion,

$$df_\lambda = \text{tr}(\mathbf{S}_\lambda),$$

as the *effective degrees of freedom* of the smoothing spline. A related quantity is $EDF = n - df_\lambda$ which has been described as the *equivalent degrees of freedom for noise*.

Several other criteria have been proposed for df_λ , such as $\text{tr}(\mathbf{S}_\lambda \mathbf{S}'_\lambda)$ and $\text{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda \mathbf{S}'_\lambda)$. However, as noted by Hastie et al. [2001: p. 130], $\text{tr}(\mathbf{S}_\lambda)$ has a number of conceptual advantages as well as being simple and easy to compute. By trying several values of df_λ and using various diagnostics, we can arrive at a particular value and then back-solve numerically to get the corresponding value of λ . A value of df_λ equal to 4 or 5 is often used as the default in software. The appropriate function for fitting a smoothing spline in S-PLUS is `smooth.spline`.

To compute λ automatically, several criteria are available. For example, we have the popular cross-validation method, which finds λ_{CV} to minimize the cross-validation sum of squares

$$CV_\lambda = \frac{1}{n} \sum_{i=1}^n \{Y_i - \hat{f}_\lambda^{(-i)}(x_i)\}^2,$$

where $\hat{f}_\lambda^{(-i)}$ is the fitted value at x_i computed by leaving out the i th data point. If $s_{ii,\lambda}$ is the i th diagonal element of \mathbf{S}_λ , then Craven and Wahba [1979] (see also Green and Silverman [1994: p. 32]) proved that

$$Y_i - \hat{f}_\lambda^{(-i)}(x_i) = \frac{Y_i - \hat{f}_\lambda(x_i)}{1 - s_{ii,\lambda}},$$

so that

$$CV_\lambda = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{Y_i - \hat{f}_\lambda(x_i)}{1 - s_{ii,\lambda}} \right\}^2.$$

Previously, a variant of CV_λ was used called the *generalized cross-validation* (GCV_λ), in which $s_{ii,\lambda}$ is replaced by the average value $\text{tr}(\mathbf{S}_\lambda)/n$. Generally, GCV_λ is closely related to CV_λ (Eubank [1999: p. 43]), although GCV_λ tends to lead to undersmoothing with small sample sizes.

Estimating σ

There is one parameter we have not considered thus far, namely, σ . Given a value of λ , a natural estimate of σ^2 is

$$\hat{\sigma}_\lambda^2 = \frac{\sum_{i=1}^n \{Y_i - \hat{f}_\lambda(x_i)\}^2}{n - df_\lambda}.$$

Green and Silverman [1994: Section 3.4] review the topic of estimating σ^2 and compare $\hat{\sigma}_\lambda^2$ (using $df_\lambda = \text{tr}[\mathbf{S}_\lambda]$) with two other estimates based on first and second differences of the data. They note that the former has a substantially smaller asymptotic mean-squared error.

Use of Weights

The theory above can be generalized to allow for weighted smoothing; we simply use $\sum_i w_i [Y_i - f(x_i)]^2$ in (7.19). For example, if we have repeated data with x_i repeated n_i times, we can use $w_i = n_i$. Green and Silverman [1994: Section 3.5] show that the theory goes through in a similar fashion using the weighted sum of squares.

In concluding this section we note that there are many other kinds of splines such as ν -splines, P-splines, Q-splines, exponential splines, subsplines, additive splines, ANOVA splines, hybrid and partial splines, tensor product splines, and thin plate splines! In addition to smoothing using splines, there are other families of smoothing methods such as kernel, near neighbor and wavelet smoothers. All these topics might be described under a general heading of nonparametric regression, for which there is a very extensive literature: for example, general methods (Simonoff [1995], Eubank [1999], Schimek [2000], and some chapters in Hastie et al. [2001]), local polynomial smoothing (Fan and Gijbels [1996]), kernel smoothing (Härdle [1990], Wand and Jones [1995]), and spline smoothing (Wahba [1990], Green and Silverman [1994]).

7.3 POLYNOMIAL REGRESSION IN SEVERAL VARIABLES

7.3.1 Response Surfaces

An important application of polynomial regression in several variables is in the study of response surfaces. We illustrate some of the basic features of response surface methodology by considering the simple case of just two regressors.

Suppose that the *response* (yield) η from a given experiment is an unknown function of two variables, x_1 (temperature) and x_2 (concentration), namely, $\eta = g(x_1, x_2)$. It is assumed that this three-dimensional surface is well-behaved, in particular is smooth with a single well-defined peak. The response η is measured with error so that we actually observe $Y = \eta + \varepsilon$, where $E[\varepsilon] = 0$ and $\text{var}[\varepsilon] = \sigma^2$. One basic problem of response theory then is to estimate the coordinates, (x_{01}, x_{02}, η_0) say, of the summit.

One method of doing this is to use a sequence of experiments and a steepest ascent technique to “climb” up the surface. Typically, experimental data points are expensive to obtain, so we need to choose a design with a small number of data points and locate them in an optimal fashion so as to maximise the efficiency of estimation. For points away from the summit the surface is relatively linear in a small region, so that it can be represented locally by a

plane, namely,

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \quad (7.23)$$

To estimate the coefficients β_i , we can, for example, use a very simple design such as the 2^2 design; we shall see later that such a design has certain optimal properties. In this design we observe Y at the four vertices of a small rectangle, with center P_1 , in the (x_1, x_2) plane (Figure 7.1). Suppose that Y_{rs} is the Y observed at (x_{r1}, x_{s2}) , where x_{r1} ($r = 1, 2$) are the two chosen values of x_1 and x_{s2} ($s = 1, 2$) are the two chosen values of x_2 . Then we can fit the model

$$Y_{rs} = \beta_0 + \beta_1 x_{r1} + \beta_2 x_{s2} + \varepsilon_{rs}, \quad (7.24)$$

where $r = 1, 2$ and $s = 1, 2$, and obtain the fitted plane

$$\hat{Y} = \phi(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \quad (7.25)$$

If Q_1 is the point on this plane vertically above P_1 , we can use the fitted plane, which approximates the surface in the neighborhood of Q_1 , to help us climb up the surface to a higher point Q_2 and thus obtain a higher yield Y . For example, if $\hat{\beta}_1$ and $\hat{\beta}_2$ are both positive in (7.25), we would increase x_1 and x_2 . However, the most efficient way of climbing up the surface is to choose the direction of steepest slope. To find this path of steepest ascent, we now consider the following problem. Suppose that we wish to maximize $\phi(d_1, d_2) - \phi(0, 0)$ subject to $d_1^2 + d_2^2 = r^2$. Using a Lagrange multiplier λ , we have

$$\frac{\partial \phi}{\partial d_i} + 2\lambda d_i = 0 \quad (i = 1, 2),$$

or, setting ϕ equal to the right side of (7.25), $d_i \propto \hat{\beta}_i$ for a maximum. Therefore, regarding Q_1 as the origin, the (x_1, x_2) coordinates of the next experimental observation should be $(k\hat{\beta}_1, k\hat{\beta}_2)$ for some $k > 0$. By steadily increasing

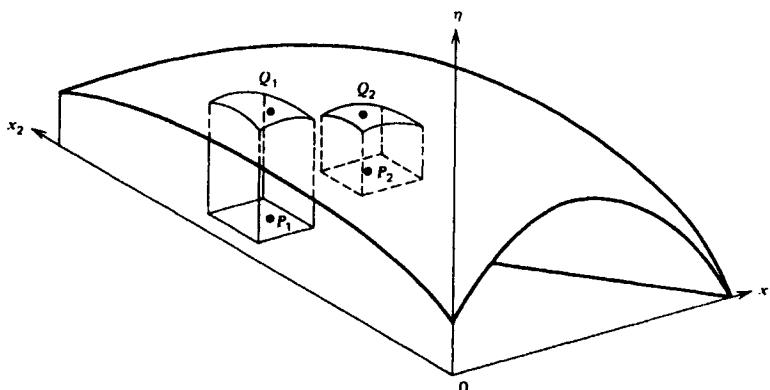


Fig. 7.1 A response surface.

k , we can go on measuring Y until we reach a point P_2 in the (x_1, x_2) plane at which the increase in Y due to a change in k becomes very small, or possibly negative. A new 2^2 experiment is then carried out on another small rectangle using P_2 as center, and another plane (7.25) is fitted. The path of steepest ascent is redetermined and, once again, we proceed in this direction until there is little change in Y . In this way we climb the surface toward the summit.

As we approach the summit, $\hat{\beta}_1$ and $\hat{\beta}_2$ get smaller, so that progress by the method of steepest ascent becomes more difficult; the curvature of the surface also begins to have a significant effect on the yield. When we are in the region of the summit, we can then fit a general quadratic form using, say, a 3^2 design; that is, we use three appropriate values for each of x_1 and x_2 and observe Y at the nine design points. Alternatively, we can use the popular 12-point central composite design, which consists of a 2^2 design with an additional four replicates at the center of the rectangle and four "coaxial" points (cf. Myers and Montgomery [1995: p. 298] for a description of its properties). By shifting the origin and rotating the axes, the fitted surface

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{22} x_2^2 \quad (7.26)$$

can be expressed in the canonical form

$$y - c_3 = \lambda_1(x_1 - c_1)^2 + \lambda_2(x_2 - c_2)^2 \quad (\lambda_1, \lambda_2 > 0), \quad (7.27)$$

where (c_1, c_2, c_3) is an estimate of the summit point (x_{01}, x_{02}, η_0) . The triple (c_1, c_2, c_3) can be found by differentiating (7.26) partially with respect to x_1 and x_2 , and solving the resulting pair of equation for x_1 and x_2 ; c_3 is the value of y in (7.26) at the solution (c_1, c_2) .

This rather sketchy description of response surface methodology leaves a number of questions unanswered, such as the following:

1. In our discussion, we used a 2^2 design to carry out a planar fit (called a *first order design*) and two possible designs for the quadratic fit (called a *second order design*). This raises the question: What is the best design to use in each case?
2. How do we know when to change from a first-order to a second-order design?
3. How do we select the values of k in $(k\hat{\beta}_1, k\hat{\beta}_2)$?
4. What happens if in our climb, we run into a stationary point which is not the maximum, or a slowly rising ridge? [Such a situation is indicated when one or other of the λ_i in equation (7.27) is negative.]

We don't have the space to consider these and other important practical questions, and we refer the readers to the comprehensive text by Myers and Montgomery [1995].

EXAMPLE 7.1 It was shown in Section 3.6 that an optimal design is one in which the design matrix has orthogonal columns. We now show that the 2^2 design lies in this category if we scale the values of x_1 and x_2 so that they take the values ± 1 . To examine this orthogonal structure, it is convenient to represent the two levels of x_1 symbolically by 1 and a , and the two levels of x_2 by 1 and b , so that the four possible combinations $(1, 1), (1, a), (1, b), (a, b)$ can be represented symbolically (by multiplying the levels together in each pair) as $1, a, b$, and ab , with observed Y values Y_1, Y_a, Y_b , and Y_{ab} , respectively. Thus setting $x_1 = -1$ and $x_2 = -1$ in (7.24), we see that $Y_1 = \beta_0 - \beta_1 - \beta_2 + \varepsilon_1$. In a similar fashion, we get

$$\begin{bmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_a \\ \varepsilon_b \\ \varepsilon_{ab} \end{bmatrix} \quad (7.28)$$

or $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the columns of \mathbf{X} are mutually orthogonal and satisfy the conditions of Example 3.3 in Section 3.6. Then $\mathbf{X}'\mathbf{X} = 4\mathbf{I}_4$ and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{bmatrix};$$

hence

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y}, \\ \hat{\beta}_1 &= \frac{1}{4}(-Y_1 + Y_a - Y_b + Y_{ab}) \\ &= \frac{1}{2}[\frac{1}{2}(Y_a + Y_{ab}) - \frac{1}{2}(Y_1 + Y_b)] \\ &= \frac{1}{2}(\text{average effect of first factor at upper level} \\ &\quad - \text{average effect of first factor at lower level}), \end{aligned}$$

$$\begin{aligned} \hat{\beta}_2 &= \frac{1}{4}(-Y_1 - Y_a + Y_b + Y_{ab}) \\ &= \frac{1}{2}[\frac{1}{2}(Y_b + Y_{ab}) - \frac{1}{2}(Y_1 + Y_a)], \end{aligned}$$

and

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \sum_c Y_c^2 - 4\bar{Y}^2 - 4\hat{\beta}_1^2 - 4\hat{\beta}_2^2. \end{aligned}$$

We note that $\text{var}[\hat{\beta}_j] = \sigma^2/4$ ($j = 0, 1, 2$), and this is the smallest variance that can be attained with a design of just four points. Finally, we note that if we use factor terminology and call x_1 and x_2 factors A and B , then $\hat{\beta}_1$ and $\hat{\beta}_2$ can be identified as estimates of what we might call the *main effects* of A and B , respectively. \square

7.3.2 Multidimensional Smoothing

Local Regression

We can generalize the method of Section 6.6 to accommodate k regressors and fit local hyperplanes $\beta' \mathbf{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ instead of lines. If \mathbf{x}_i is the value of \mathbf{x} at the i th data point, then at each \mathbf{x}_0 we minimize

$$\sum_{i=1}^n K_\lambda(\mathbf{x}_i, \mathbf{x}_0)(Y_i - \mathbf{x}'_i \beta)^2$$

with respect to β to fit $\hat{f}(\mathbf{x}_0) = \mathbf{x}'_0 \hat{\beta}$, where $\hat{\beta}$ will be a function of x_0 . Typically, $K_\lambda(\mathbf{x}_i, \mathbf{x}_0)$ will be an appropriate kernel function of the form $K(||\mathbf{x}_i - \mathbf{x}_0||/\lambda)$. As noted in Section 6.6, `loess()` in S-PLUS incorporates multidimensional local regression. It should be noted that there are some problems as k increases, and local regression tends to become less useful for $k > 3$ (cf. Hastie et al. [2001]: p. 174). Local quadratics can also be used.

Multidimensional Splines

The theory of Section 7.2.2 can be generalized using basis functions that are tensor products of (one-dimensional) B -splines. For example, if $k = 2$, we can establish a basis of functions $B_{1j}(x_1)$ ($j = 1, 2, \dots, M_1$) for x_1 and a basis $B_{2k}(x_2)$ ($k = 1, 2, \dots, M_2$) for x_2 . Then the $(M_1 \times M_2)$ -dimensional *tensor product basis*, defined by

$$B_{jk}(\mathbf{x}) = B_{1j}(x_1)B_{2k}(x_2) \quad (j = 1, \dots, M_1; \quad k = 1, \dots, M_2),$$

can be used to represent a two dimensional function, namely,

$$f(\mathbf{x}) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \theta_{jk} B_{jk}(\mathbf{x}).$$

The parameters θ_{jk} can then be estimated by least squares.

Smoothing splines can also be used. Assuming once again that $k = 2$, we now have the two-dimensional problem of minimizing

$$\text{RSS}(\lambda, f) = \sum_{i=1}^n \{Y_i - f(\mathbf{x}_i)\}^2 + \lambda J(f),$$

where

$$J(f) = \int_c^d \int_a^b \left\{ \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right\} dx_1 dx_2.$$

This leads to a two-dimensional cubic natural spline called a *thin plate spline* or *Laplacian smoothing spline*. However, the solution is computationally expensive. Software for multidimensional smoothing splines is available in S-PLUS, and on StatLib and NetLib. For further details, see Green and Silverman [1994: Chapters 7 and 8].

MISCELLANEOUS EXERCISES 7

1. Using the method of orthogonal polynomials described in Section 7.1.2, fit a third-degree equation to the following data:

y (index):	9.8	11.0	13.2	15.1	16.0
x (year):	1950	1951	1952	1953	1954

Test the hypothesis that a second-degree equation is adequate.

2. Show that the least squares estimates of β_1 and β_2 for the model (7.28) are still unbiased even when the true model includes an interaction term β_{12} , that is,

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2.$$

Find the least squares estimate of β_{12} .

3. Suppose that the regression curve

$$E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$$

has a local maximum at $x = x_m$ where x_m is near the origin. If Y is observed at n points x_i ($i = 1, 2, \dots, n$) in $[-a, a]$, $\bar{x} = 0$, and the usual normality assumptions hold, outline a method for finding a confidence interval for x_m . *Hint:* Use the method of Section 6.1.2.

(Williams [1959: p. 110])

8

Analysis of Variance

8.1 INTRODUCTION

In this chapter we look at certain special cases of the multiple regression model. When the regressors are qualitative so that indicator variables are involved (taking values 0 or 1), then we refer to the model as an *analysis-of-variance model* (ANOVA model). In this case the regression matrix \mathbf{X} is usually referred to as the *design matrix*. However, if there are both qualitative and quantitative x -variables, then we refer to the model as an *analysis-of-covariance model* (ANCOVA model).

In ANOVA models we begin with the relationship $E[\mathbf{Y}] = \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is known to belong to some space of values. We then find a matrix \mathbf{X} such that this space is $C(\mathbf{X})$. Clearly, such a representation is not unique, as

$$\begin{aligned}\boldsymbol{\theta} &= \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\mathbf{B}^{-1}\mathbf{B}\boldsymbol{\beta} \\ &= \mathbf{X}_B\boldsymbol{\gamma},\end{aligned}\tag{8.1}$$

say, where \mathbf{B} is a $p \times p$ nonsingular matrix. The choice of \mathbf{X}_B will depend on the reparameterizing transformation $\boldsymbol{\gamma} = \mathbf{B}\boldsymbol{\beta}$, which in turn depends on which linear combinations of the $\boldsymbol{\beta}$'s we are interested in. Typically, we are interested in just the individual β_j or in *contrasts* $\sum_j c_j \beta_j$, where $\sum_j c_j = 0$, for example, $\beta_j - \beta_{j-1}$. Under the broad umbrella of ANOVA models, there is a large family of designs. We shall consider some of these in the following sections.

8.2 ONE-WAY CLASSIFICATION

8.2.1 General Theory

EXAMPLE 8.1 A health researcher wishes to compare the effects of four anti-inflammatory drugs on arthritis patients. She takes a random sample of patients and divides them randomly into four groups of equal size, each of which receives one of the drugs. In the course of the study several patients became seriously ill and had to withdraw, leaving four unequal-sized groups. We thus have four independent samples, each receiving a different treatment.

□

In the example above, the type of drug is usually referred to as a *factor* or *treatment*, and the four kinds of drug are often called *levels* of the factor. The entire experiment is variously called a *single-factor experiment*, a *one-way layout*, or a *one-way classification*. Instead of four different kinds of drug, we could have used a single drug but at four different dosages.

In general, we could have I levels of the factor, giving I independent random samples with J_i observations in the i th sample ($i = 1, 2, \dots, I$). We can regard each sample as coming from its own underlying (hypothetical) population, which we assume to be normal with a common variance σ^2 . Our first task is to compare the means of these populations.

Let Y_{ij} be the j th observation ($j = 1, 2, \dots, J_i$) on the i th normal population $N(\mu_i, \sigma^2)$ ($i = 1, 2, \dots, I$), so that we have the following array of data:

		Sample mean
Population 1:	$Y_{11}, Y_{12}, \dots, Y_{1J_1}$	\bar{Y}_1 .
Population 2:	$Y_{21}, Y_{22}, \dots, Y_{2J_2}$	\bar{Y}_2 .
.....
Population I :	$Y_{I1}, Y_{I2}, \dots, Y_{IJ_I}$	\bar{Y}_I .

We can combine the information above into the single model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad (i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J_i),$$

where the ε_{ij} are i.i.d. (independently and identically distributed) as $N(0, \sigma^2)$. In Example 4.5 (Section 4.3.3) we considered the special case of comparing just two normal populations, and the theory there provides a background for what follows.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ_i})'$ represent the sample from the i th population. Then following Example 4.5, we “stack” these vectors to give

$$\mathbf{Y} = (\mathbf{Y}'_1, \dots, \mathbf{Y}'_I)',$$

a vector with $n = \sum_{i=1}^I J_i$ elements. This vector has mean

$$\boldsymbol{\theta} = (\mu_1 \mathbf{1}'_{J_1}, \mu_2 \mathbf{1}'_{J_2}, \dots, \mu_I \mathbf{1}'_{J_I})',$$

so that $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. We now wish to specify a matrix whose column space contains the set of all such vectors $\boldsymbol{\theta}$. We can find the columns of one such matrix by setting all the μ 's except μ_i equal to zero for each i . This gives us

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{J_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{J_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{J_I} \end{pmatrix}, \quad (8.2)$$

an $n \times I$ matrix of rank I (since the columns, being mutually orthogonal, are linearly independent), and $\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_I)'$. (We find it notationally more convenient in what follows to use $\boldsymbol{\mu}$ instead of $\boldsymbol{\beta}$.) This is our usual regression matrix, but without an initial column of unit elements (corresponding to β_0), as \mathbf{X} essentially describes qualitative information.

If we wish to test $H : \mu_1 = \mu_2 = \cdots = \mu_I$ ($= \mu$, say), then $\boldsymbol{\theta}$ reduces to $\boldsymbol{\theta} = \mathbf{1}_n \boldsymbol{\mu} = \mathbf{X}_H \boldsymbol{\mu}$, say, where \mathbf{X}_H is $\mathbf{1}_n$, and is obtained by adding the columns of \mathbf{X} together. This is the “canonical” form for H given in Section 4.5, so we can apply the general theory of Chapter 4 to get an F -test for H as follows.

For the original model, the least squares estimates are most readily found by direct minimization; that is, we differentiate $\sum_{i=1}^I \sum_{j=1}^{J_i} (Y_{ij} - \mu_i)^2$ with respect to μ_i , giving us

$$\hat{\mu}_i = \sum_{j=1}^{J_i} \frac{Y_{ij}}{J_i} = \frac{Y_{i\cdot}}{J_i} = \bar{Y}_{i\cdot}$$

and

$$\text{RSS} = \sum_i \sum_j (Y_{ij} - \hat{\mu}_i)^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2.$$

Similarly, under H , we minimize $\sum \sum (Y_{ij} - \mu)^2$ with respect to μ to get

$$\hat{\mu}_H = \frac{\sum_i \sum_j Y_{ij}}{\sum_i J_i} = \frac{Y_{\cdot\cdot}}{n} = \bar{Y}_{\cdot\cdot}$$

and

$$\text{RSS}_H = \sum_i \sum_j (Y_{ij} - \hat{\mu}_H)^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_i J_i (Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

Since \mathbf{X} is $n \times I$ of rank I , and \mathbf{X}_H is $n \times 1$ of rank 1, we have from Theorem 4.1 in Section 4.3.2 with $p = I$ and $q = I - 1$ (the difference in ranks between \mathbf{X} and \mathbf{X}_H)

$$F = \frac{(\text{RSS}_H - \text{RSS})/(I - 1)}{\text{RSS}/(n - I)} \sim F_{I-1, n-I},$$

when H is true. We can also express F in the form S_H^2/S^2 , a ratio of *mean sum of squares*, where $S_H^2 = (\text{RSS}_H - \text{RSS})/(I - 1)$, etc.

Setting $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{Y}}_H = \mathbf{X}_H\hat{\boldsymbol{\mu}}_H$, we have, by Theorem 4.1(i),

$$\begin{aligned}
 \text{RSS}_H - \text{RSS} &= \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2 \\
 &= \sum_i \sum_j (\hat{Y}_{ij} - \hat{Y}_{Hi})^2 \\
 &= \sum_i \sum_j (\hat{\mu}_i - \hat{\mu}_H)^2 \\
 &= \sum_i \sum_j (\bar{Y}_{i..} - \bar{Y}_{..})^2 \\
 &= \sum_i J_i (\bar{Y}_{i..} - \bar{Y}_{..})^2. \tag{8.3}
 \end{aligned}$$

Thus

$$F = \frac{\sum J_i (\bar{Y}_{i..} - \bar{Y}_{..})^2 / (I - 1)}{\sum \sum Y_{ij} - \bar{Y}_{i..})^2 / (n - I)} = \frac{S_H^2}{S^2}. \tag{8.4}$$

In preparation for thinking about two-way classification models, we shall consider some other parameterizations.

Alternative Parameterizations

The hypothesis H can also be expressed in the form $\mu_1 - \mu_I = \mu_2 - \mu_I = \cdots = \mu_{I-1} - \mu_I = 0$, so that we have $I - 1$ contrasts $\alpha_i = \mu_i - \mu_I$ ($i = 1, \dots, I - 1$). In order to transfer from the μ_i to the α_i [as demonstrated by (8.1)] we need one more parameter, for example μ_I . Putting $\mu_I = \mu$, we have $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ or $\mathbf{Y} = \mathbf{X}_1\gamma + \boldsymbol{\varepsilon}$, where $\gamma = (\mu, \alpha_1, \dots, \alpha_{I-1})'$ and \mathbf{X}_1 is simply \mathbf{X} with its last column deleted and having a new first column $\mathbf{1}_n$. Writing $\gamma = \mathbf{B}\mu$, it is readily seen that \mathbf{B} is nonsingular. We also note that \mathbf{X}_1 is $n \times I$ of rank I , and H amounts to testing $\alpha_1 = \cdots = \alpha_{I-1} = 0$. What we have done is to effectively set $\alpha_I = 0$ in $\mu + \alpha_I$. We could have also set $\alpha_1 = 0$ by writing $\mu_1 = \mu$ instead and then defining $\alpha_i = \mu_i - \mu_1$ ($i = 2, \dots, I$).

Instead of singling out one of the α 's to be zero, we can use a more symmetric approach and define the contrasts $\alpha_i = \mu_i - \mu$ ($i = 1, \dots, I$), where we now define $\mu = \sum_{i=1}^I \mu_i/I$ to be the average of the μ_i 's. However, the α_i are mathematically dependent, as $\alpha_{..} = \sum_I \alpha_i = 0$ (sometimes referred to as an *identifiability constraint*), and the corresponding model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ has a design matrix \mathbf{X}_2 , where \mathbf{X}_2 is \mathbf{X} plus a first column $\mathbf{1}_n$. This is an $n \times (I + 1)$ matrix and does not have full rank, as its first column is the sum of the other columns. We can turn this matrix into one of full rank by setting $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$. One conceptual advantage in using \mathbf{X}_2 instead of \mathbf{X} is that \mathbf{X}_2 contains \mathbf{X}_H as its first column.

Clearly, a wide range of reparameterizations are possible. Another common one is to define $\mu = \sum_i J_i \mu_i/n$ and $\alpha_i = \mu_i - \mu$. The identifiability constraint is then $\sum_i J_i \alpha_i = 0$. Although introducing all these parameterizations is not

particularly helpful in the context of the one-way classification, the idea is important for more complex models such as the two-way classification discussed later.

Finally, we note that the model with the design matrix (8.2) can also be expressed in the form

$$Z_r = \mu_1 d_{r1} + \mu_2 d_{r2} + \cdots + \mu_I d_{rI} + \varepsilon_r \quad (r = 1, 2, \dots, n),$$

where $\mathbf{Z} = \mathbf{Y}$ and d_{ri} is the r th observation on the i th dummy variable d_i ($i = 1, 2, \dots, I$); that is, $d_i = +1$ when Z_r is an observation from the i th population, and zero otherwise.

EXAMPLE 8.2 In this example we illustrate how orthogonal contrasts can be utilized in some situations. Suppose that $I = 3$, $J_i = J$ for $i = 1, 2, 3$, with the first treatment being a placebo (dummy drug) and the other two being different doses of the same drug. We are interested in testing for no differences in the three treatments (i.e. $H : \mu_1 = \mu_2 = \mu_3$). However, if we suspect that the hypothesis will be rejected, we will be interested in contrasts such as $\frac{1}{2}(\mu_2 + \mu_3) - \mu_1$ (comparing the drug with the placebo) and $\mu_2 - \mu_3$ (comparing the two dosages). Let

$$\gamma_1 = \frac{1}{\sqrt{3}}(\mu_1 + \mu_2 + \mu_3), \quad \gamma_2 = \frac{1}{\sqrt{2}}(\mu_2 - \mu_3) \quad \text{and} \quad \gamma_3 = \frac{1}{\sqrt{6}}(\mu_2 + \mu_3 - 2\mu_1).$$

Then

$$\gamma = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix} \boldsymbol{\mu} = \mathbf{B}\boldsymbol{\mu},$$

where \mathbf{B} is an orthogonal matrix (known as *Helmert's transformation*). Now, $\hat{\gamma} = \mathbf{B}\hat{\boldsymbol{\mu}} = (\bar{Y}_{1.}, \bar{Y}_{2.}, \bar{Y}_{3.})'$, so that

$$\sum_i \hat{\gamma}_i^2 = \hat{\gamma}'\hat{\gamma} = \hat{\boldsymbol{\mu}}' \mathbf{B}' \mathbf{B} \hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\mu}} = \sum_i \hat{\mu}_i^2 = \bar{Y}_{1.}^2 + \bar{Y}_{2.}^2 + \bar{Y}_{3.}^2.$$

Since $\hat{\gamma}_1^2 = \frac{1}{3}(\bar{Y}_{1.} + \bar{Y}_{2.} + \bar{Y}_{3.})^2 = 3\bar{Y}_{..}^2$, we have

$$J\hat{\gamma}_2^2 + J\hat{\gamma}_3^2 = J(\sum_i \bar{Y}_{i.}^2 - 3\bar{Y}_{..}^2) = \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{..})^2 = \text{RSS}_H - \text{RSS}.$$

Also, from

$$\text{Var}[\hat{\gamma}] = \mathbf{B} \text{Var}[\hat{\boldsymbol{\mu}}] \mathbf{B}' = \mathbf{B} \sigma^2 \left(\frac{\mathbf{I}_3}{J} \right) \mathbf{B}' = \frac{\sigma^2}{J} \mathbf{I}_3,$$

it follows that $\text{cov}[\hat{\gamma}_2, \hat{\gamma}_3] = 0$, so that $\hat{\gamma}_2$ and $\hat{\gamma}_3$ are independent (Theorem 2.5) under normality assumptions. We thus have a decomposition of $\text{RSS}_H - \text{RSS}$ into two independent components, $J\hat{\gamma}_2^2$ and $J\hat{\gamma}_3^2$, which can be tested independently. This decomposition into orthogonal components is an option in S-PLUS. \square

Computations

With a less-than-full-rank design such as X_2 , a variety of computational procedures are possible. Software packages such as S-PLUS effectively operate on the linearly independent columns of the design matrix.

It is standard practice for software packages to set out the various sums of squares in the form of a table (Table 8.1). The terminology used for the column labeled "source" varies in the literature. Instead of "between populations", one encounters "between groups" or "treatments." The "error" sum of squares, variously called "within groups," "within populations," or "residual" sum of squares, provides a pooled estimate of σ^2 .

Table 8.1 Analysis-of-variance table for a one-way classification

Source	Sum of squares (SS)	Degrees of freedom (df)	$\frac{SS}{df}$
Between populations	$J \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	$I - 1$	S_H^2
Error	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$	$IJ - I$	S^2
Corrected total	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$IJ - 1$	

8.2.2 Confidence Intervals

The literature on confidence intervals for a one-way classification (cf. Hsu [1996] and Hochberg and Tamhane [1987] for references) is very extensive and rather daunting. Some statistics packages (e.g., SAS) present a bewildering array of options. We shall therefore confine ourselves to a few basic methods that have good properties and which illustrate the general methodology.

Along with the F -test of (8.4), we are interested in seeing how the μ_i differ from one another; this leads to looking at contrasts. There may be one contrast $\theta = \sum_i c_i \mu_i$ of particular interest, and we can estimate it by $\hat{\theta} = \sum_i c_i \bar{Y}_{i\cdot}$ ($= \sum_i c_i \hat{\mu}_i = c' \hat{\mu}$). This has variance $\sigma^2 \sum_i c_i^2 / J_i$, where σ^2 is estimated by $S^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 / (n - I)$. Then, from Example 4.7, a $100(1 - \alpha)\%$ confidence interval for θ is given by

$$\sum_i c_i \bar{Y}_{i\cdot} \pm t_{n-I}^{(1/2)\alpha} S \left(\sum_i c_i^2 / J_i \right)^{1/2}.$$

If we are interested in several contrasts (or the individual μ_i), chosen prior to seeing the data, then we can use one of the three conservative methods described in Chapter 5: namely, the Bonferroni, Studentized maximum-modulus (SMM), and Scheffé methods. If we are interested in just the μ_i , then SSM should be used. However, if the upper quantiles of the SMM distribution are not available, then a conservative approximation (Hsu [1996: pp. 10–11]) is given by the t -distribution with upper quantile $t_{n-I}^{\alpha^*/2}$, where

$$\alpha^* = 1 - (1 - \alpha)^{1/k}.$$

The Bonferroni intervals using $t_{n-I}^{\alpha/(2k)}$ are wider, as $(1 - \alpha)^{1/k} < 1 - \alpha/k$ implies that $1 - \alpha^* < 1 - \alpha/k$, but the difference is small. There is therefore little advantage in using the approximation.

If we are interested in all possible contrasts, then Scheffé's method is appropriate. Since H can be expressed in the form $\phi_i = \mu_i - \mu_I = 0$ ($i = 1, \dots, I-1$), we know from Example 5.3 in Section 5.1.4 that the set of all possible linear combinations of the ϕ_i is the same as the set of all possible contrasts in the μ_i . Hence for every c such that $c'1_I = 0$, it follows from (5.13) that $c'\mu$ lies in the interval

$$\sum_i c_i \bar{Y}_i \pm [(I-1)F_{I-1,n-I}^{\alpha}]^{1/2} S [\sum_i (c_i^2/J_i)]^{1/2} \quad (8.5)$$

with an overall probability of $1 - \alpha$. From Section 5.1.4 we see that H will be rejected if and only if at least one of these confidence intervals does not contain zero.

As already noted in Chapter 5, some of the conservative intervals can be quite wide. However, if we are interested in just the set of pairwise contrasts $\mu_r - \mu_s$ (for all r and s with $r \neq s$), then we can do better. Tukey [1953] and Kramer [1956] independently proposed the intervals

$$\bar{Y}_r - \bar{Y}_s \pm q_{I,n-I}^{(\alpha)} S \sqrt{\frac{1}{2} \left(\frac{1}{J_r} + \frac{1}{J_s} \right)}, \quad (8.6)$$

where $q_{k,\nu}^{(\alpha)}$ is the upper $100\alpha\%$ point of the Studentized range distribution with parameters k and ν [i.e., the distribution of the range of k independent $N(0, 1)$ random variables divided by $(V/\nu)^{1/2}$, where V is independently distributed as χ^2_ν]. Tukey conjectured that the overall confidence probability for these intervals is at least $1 - \alpha$ with equality when the J_i are equal. This conjecture was proved by Hayter [1984], who showed that equality occurs if and only if the J_i are all equal. He also gave similar conservative intervals for all contrasts. Hsu [1996: Section 5.2.1] gives a helpful discussion on graphical ways of displaying the Tukey–Kramer (TK) intervals.

There are number of other procedures available. However, it appears that the TK intervals are to be preferred, as they are less conservative (i.e., not

as wide) (Hochberg and Tamhane [1987: pp. 93–96]; Hsu [1996: pp. 146, 158–160]).

Balanced Design

If the sample sizes are all equal (i.e., $J_i = J$ for all i), then other sets of confidence intervals are available. For example, Tukey [1953] proposed the following intervals for all contrasts $\sum_{i=1}^I c_i \mu_i$, namely,

$$\sum_i c_i \bar{Y}_{i\cdot} \pm q_{I,IJ-I}^{(\alpha)} \frac{S}{\sqrt{J}} \sum_{i=1}^I \frac{|c_i|}{2}, \quad (8.7)$$

which have an exact overall probability of $(1 - \alpha)$. If we are interested in just the pairwise contrasts $\mu_r - \mu_s$, we can use [cf. (8.6) with $J_i = J$]

$$\bar{Y}_{r\cdot} - \bar{Y}_{s\cdot} \pm \frac{q_{I,IJ-I}^{(\alpha)} S}{\sqrt{J}} \quad (8.8)$$

or

$$\bar{Y}_{r\cdot} - \bar{Y}_{s\cdot} \pm d,$$

say. Since any difference $\bar{Y}_{r\cdot} - \bar{Y}_{s\cdot}$ greater than d suggests that $\mu_r - \mu_s \neq 0$, we can sort the μ_i into groups which contain means that are not significantly different from one another. For example, if $d = 10.4$ and the ranked sample means are given as follows:

$i:$	5	1	3	2	4
$\bar{Y}_{i\cdot}:$	25.4	32.6	39.2	40.8	52.1

then $\mu_5 < \mu_3, \mu_2, \mu_4$; $\mu_1 < \mu_4$; $\mu_3 < \mu_4$; $\mu_2 < \mu_4$; and the appropriate groups of means are (μ_5, μ_1) , (μ_1, μ_3, μ_2) , and μ_4 . It is common practice to underline the groups as we have done above. However, when the design is unbalanced, Hsu [1996: p. 147] notes two situations where underlining is not helpful. For example, suppose that $\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3$, but $J_1 < J_3 < J_2$, so that the estimated standard deviation of $\bar{Y}_3 - \bar{Y}_2$ is greater than the estimated standard deviation of $\bar{Y}_3 - \bar{Y}_1$. Then the Tukey–Kramer method may find μ_2 and μ_3 significantly different, but not μ_1 and μ_3 ; this cannot be described by underlining. Furthermore, underlining will not distinguish between a confidence interval for $\mu_1 - \mu_2$ that is tight around zero, thus indicating practical equivalence of μ_1 and μ_2 , and a wide interval around zero, which is somewhat inconclusive.

When we use a pairwise comparison, we find that $\sum_i |c_i|/2 = 1$ and (8.7) is the same as (8.8). The reason for this is that the pairwise contrast vectors form a basis for the set of all contrast vectors.

8.2.3 Underlying Assumptions

In Section 9.5 we will find that *quadratically balanced F-tests* are robust with regard to departures from the assumption of normality. Now it is easy to check for quadratic balance once we have derived the *F*-statistic; we simply check both the numerator and the denominator to see if, in each case, the coefficient of Y_{rs}^2 is the same for all r, s . From (8.4) we see that this will be the case when all the samples are the same size [i.e., $J_i = J$, ($i = 1, \dots, I$)], and we say then that the model is *balanced*. Clearly, the experimenter should generally aim for balance as closely as possible; although balance is not always achievable, as we saw in Example 8.1.

Another assumption is that σ^2 is the same for all the populations. Scheffé [1959: Chapter 10] concluded that any heterogeneity of variance does not affect the *F*-test if the design is approximately balanced. Then a usually conservative rule of thumb is that heterogeneity is unlikely to do much harm if the ratio of the largest to the smallest sample standard deviation is no more than 2. However, the confidence intervals are not so robust against variance differences. If the variance differences are substantial, then a number of alternative procedures are available. For two samples there is the well known Welch procedure, which does not assume equal variances; this is usually available from statistics packages. Welch and others have generalized this method to more than two samples (see Roth [1988] for details).

There are a number of standard tests for the equality of population (group) variances based on sample standard deviations. Unfortunately, these tests are very sensitive to any nonnormality of the data and are therefore suspect as a preliminary test (cf. Markowski and Markowski [1990]). There are, however, several robust tests that are based on calculating a dispersion variable Z_{ij} and then performing a one-way analysis on the Z -data. The statistics package SAS, for example, provides a number of homogeneity of variance tests using the HOVTEST option in its GLM procedure. The best-known of these is Levene's test (Levene [1960]), which uses the absolute deviations and involves calculating for the i th group (sample)

$$Z_{ij} = |Y_{ij} - \bar{Y}_{\cdot i}| \quad (j = 1, 2, \dots, J_i).$$

One can also use Z_{ij}^2 instead of Z_{ij} . A robust version of this test, due to Brown and Forsythe [1974], uses sample medians instead of sample means when calculating the Z_{ij} data (see also Glaser [1983]). Simulation studies (Conover et al. [1981]; Olejnik and Algina [1987]) indicate that this last test seems to be the most powerful at detecting variance differences, while protecting the type I error probability. Although the test is not very powerful, we can rely on the fact that unless the group variances are very different or the number of groups (I) is large, the test is reasonably robust to variance heterogeneity when the J_i are approximately equal. If the group variances are very different, we should use a Welch-type procedure.

If the assumptions of normality and equal variances are seriously violated, then one can perhaps transform the data or use a nonparametric test such as the Kruskal-Wallis test. Such robust procedures are generally available in statistics packages.

With regard to confidence intervals, if the population variances are unequal, a number of methods are available. For example, if we are interested in just a single confidence interval for, say, $\sum a_i \mu_i$, then we can use the approximate $N(0, 1)$ statistic (Scott and Smith [1971])

$$Z = \frac{\sum a_i \bar{Y}_{i\cdot} - \sum a_i \mu_i}{(\sum a_i^2 \tilde{S}_i^2 / J_i)^{1/2}} \quad (\text{each } J_i > 3),$$

where $\tilde{S}_i^2 = \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 / (J_i - 3)$. A number of methods for computing simultaneous intervals are described by Hochberg and Tamhane [1987: Chapter 7]. Unfortunately, the upper quantiles of the various distributions used are difficult to obtain and one must resort to approximations.

We note that the F -statistic is not robust with respect to the presence of intraclass correlation, that is, when $\text{cov}[Y_{is}, Y_{ir}] = \sigma^2 \rho$ ($r \neq s, \rho \neq 0$).

In conclusion, it should be emphasized that the diagnostic methods of Chapter 10 using residuals can be applied to ANOVA models.

EXERCISES 8a

1. Express $H : \mu_1 = \dots = \mu_J$ in the form $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. What is the rank of \mathbf{A} ? As this matrix is not unique, find another matrix, \mathbf{A}_1 say, also satisfying $H : \mathbf{A}_1 \boldsymbol{\mu} = \mathbf{0}$. What would be the relationship between \mathbf{A} and \mathbf{A}_1 ?

2. Obtain the identity indicated by (8.3), namely,

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 = \sum_i \sum_i (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2.$$

3. Obtain the least squares estimates of the μ_i using $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ directly, where \mathbf{X} is given by (8.2).

4. Prove that

$$\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \sum_i \frac{Y_{i\cdot}^2}{J_i} - \frac{Y_{..}^2}{n}.$$

Obtain a similar expression for $\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$.

5. Find (a) $E[\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2]$ and (b) $E[\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2]$. Hint: Use Theorem 4.1(ii) for the first expression.

8.3 TWO-WAY CLASSIFICATION (UNBALANCED)

8.3.1 Representation as a Regression Model

Consider an experiment in which two factors A and B are allowed to vary: for example, type of drug and dosage. Suppose that there are I levels of A and J levels of B so that there are IJ different combinations of the levels. Let Y_{ijk} be the k th experimental observation ($k = 1, 2, \dots, K_{ij} : K_{ij} > 1$) on the combination of the i th level of A with the j th level of B so that there are $n = \sum_{i=1}^I \sum_{j=1}^J K_{ij} = K..$ observations altogether. The data can be regarded as providing IJ independent samples, each from a different population. We shall assume that the Y_{ijk} are independently distributed as $N(\mu_{ij}, \sigma^2)$, so that

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad (i = 1, \dots, I; j = 1, \dots, J; k = 1, \dots, K_{ij}), \quad (8.9)$$

where the ε_{ijk} are i.i.d. $N(0, \sigma^2)$.

Let $\mathbf{Y}_{ij} = (Y_{ij1}, \dots, Y_{ijK_{ij}})'$ be the ij th sample vector, and suppose that we stack these vectors to get

$$\mathbf{Y} = (\mathbf{Y}'_{11}, \mathbf{Y}'_{12}, \dots, \mathbf{Y}'_{1J_1}, \dots, \mathbf{Y}'_{I1}, \mathbf{Y}'_{I2}, \dots, \mathbf{Y}'_{IJ_I})'$$

with mean

$$\boldsymbol{\theta} = (\mu_{11} \mathbf{1}'_{K_{11}}, \mu_{12} \mathbf{1}'_{K_{12}}, \dots, \mu_{IJ} \mathbf{1}'_{K_{IJ}})'$$

If we stack the ε_{ijk} in the same way as the Y_{ijk} to get $\boldsymbol{\varepsilon}$, then $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

As in the one-way classification, we can find an $n \times IJ$ matrix \mathbf{X} of rank $p = IJ$ just like (8.2) such that $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \dots, \mu_{IJ})'$. The least squares estimates of the μ_{ij} can then be found by minimizing $\sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2$ with respect to μ_{ij} . This leads to $\hat{\mu}_{ij} = \bar{Y}_{ij..}$ and $\text{RSS} = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij..})^2$ with $n - p = n - IJ$ degrees of freedom.

8.3.2 Hypothesis Testing

The next question to consider is: What hypotheses are of interest? The first hypothesis to test would be to see if the factors make any difference at all (i.e., $H : \mu_{ij} = \mu$ for all i, j). The test statistic for this hypothesis is straightforward, as it follows immediately from the one-way classification with IJ means instead of I means. Thus from (8.4) we have

$$F = \frac{\sum_i \sum_j K_{ij} (\bar{Y}_{ij..} - \bar{Y}_{...})^2 / (IJ - 1)}{\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij..})^2 / (n - IJ)},$$

which has an $F_{IJ-1, n-IJ}$ distribution when H is true. We would generally expect the test to be significant if the experiment is a meaningful one.

Our next question is: Do the two factors interact in any way; that is, does the effect of factor A at level i , say, depend on the level of factor B ? If not,

then we could ignore the presence of factor B and treat the experiment as a one-way classification for A with $K_{i\cdot} = \sum_{j=1}^J K_{ij}$ observations on the i th level of A . By the same token, we could ignore the presence of factor A and treat the experiment as a one-way classification for B with $K_{\cdot j} = \sum_{i=1}^I K_{ij}$ observations on the j th level of B . With no interactions we essentially get two one-way classifications as a bonus. We now wish to pin down this concept of interaction mathematically.

If there were no interaction between the factors, we would expect the difference in means $\mu_{i_1 j_1} - \mu_{i_2 j_1}$ to depend only on i_1 and i_2 and not on the level j_1 of B . Mathematically, this means that

$$\mu_{i_1 j_1} - \mu_{i_2 j_1} = \mu_{i_1 j_2} - \mu_{i_2 j_2}, \quad (8.10)$$

and the hypothesis of no interactions is then

$$H_{AB} : \mu_{i_1 j_1} - \mu_{i_2 j_1} - \mu_{i_1 j_2} + \mu_{i_2 j_2} = 0 \quad \text{for all } i_1, i_2, j_1, j_2.$$

Another way of expressing this hypothesis is to use μ_{IJ} as a baseline (i.e., set $i_2 = I$ and $j_2 = J$) and write

$$H_{AB1} : \mu_{ij} - \mu_{Ij} - \mu_{iJ} + \mu_{IJ} = 0 \quad \text{all } i, j.$$

If we use μ_{11} as our baseline instead of μ_{IJ} , we get

$$H_{AB2} : \mu_{ij} - \mu_{1j} - \mu_{i1} + \mu_{11} = 0 \quad \text{all } i, j.$$

There is yet one more common method of specifying the no-interaction hypothesis. From (8.10)

$$\begin{aligned} \mu_{i_1 j} - \mu_{i_2 j} &= \phi(i_1, i_2), \quad \text{say} \\ &= \sum_{j=1}^J \frac{\phi(i_1, i_2)}{J} \\ &= \bar{\mu}_{i_1 \cdot} - \bar{\mu}_{i_2 \cdot} \quad \text{for all } i_1, i_2 \end{aligned}$$

or

$$\mu_{i_1 j} - \bar{\mu}_{i_1 \cdot} = \mu_{i_2 j} - \bar{\mu}_{i_2 \cdot}.$$

This equation implies that $\mu_{ij} - \bar{\mu}_{i\cdot}$ does not depend on i , so that

$$\begin{aligned} \mu_{ij} - \bar{\mu}_{i\cdot} &= \Phi(j), \quad \text{say} \\ &= \sum_{i=1}^I \frac{\Phi(j)}{I} \\ &= \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot \cdot}, \end{aligned}$$

or

$$H_{AB3} : \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot \cdot} = 0 \quad \text{for all } i, j. \quad (8.11)$$

We note that this expression is symmetric in i and j , so that we would arrive at the same result if we assumed the difference $\mu_{ij_1} - \mu_{ij_2}$ to depend only on j_1 and j_2 and not on i .

All the sets of hypothesis contrasts on the μ_{ij} above are equivalent, as they all lead to the same vector subspace defining the linear model corresponding to H_{AB} . This is effectively the *means model approach*. However, we might ask what this model looks like? Instead of using a constraint equation specification where the μ_{ij} 's are constrained, we can also use a "freedom equation" specification, in which we express the μ_{ij} in terms of other parameters. For example, all the forms of H_{AB} are equivalent to $\mu_{ij} = \mu + \alpha_i + \beta_j$ (Exercises 8b, No. 1), and this will specify μ as belonging to the column space of a particular matrix. To incorporate this matrix into the general model with column space $C(I_{IJ})$ (as $\mu = I_{IJ}\mu$), we simply write

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \quad (8.12)$$

where the $(\alpha\beta)_{ij}$ are usually referred to as the *interaction parameters*. In vector form this model is given by

$$\mu = L(\mu, \alpha', \beta', (\alpha\beta)')' = L\gamma,$$

say, where

$$\alpha = (\alpha_1, \dots, \alpha_I)', \quad \beta = (\beta_1, \dots, \beta_J)', \text{ and } (\alpha\beta) = ((\alpha\beta)_{11}, \dots, (\alpha\beta)_{IJ})'.$$

To get some idea of what L looks like, suppose that $I = 2$ and $J = 2$. Then

$$\begin{aligned} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{pmatrix} &= \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{pmatrix} \\ &= (1_4, A_1, A_2, A_{12})\gamma, \end{aligned} \quad (8.13)$$

say. We note that L does not have full rank; in fact, it has four linearly independent columns (as the rows are linearly independent). However, $C(L) = C(I_4)$, so that $C(L)$ also represents the general model. Since H_{AB} is obtained by setting $(\alpha\beta) = 0$, the model corresponding to H_{AB} is obtained from L simply by leaving out A_{12} . Before discussing the fitting of such less-than-full-rank models, we now consider two further hypotheses of interest.

If there are no interactions, then, as we noted above, we can treat the two factors separately and test, for example, the hypothesis that all the levels of A have the same effect. There are several other ways of expressing this

hypothesis. For example, if μ_{ij} does not vary with i , then $\mu_{ij} = \mu_{Ij}$, which combined with H_{AB1} leads to

$$H_{A1} : \mu_{ij} - \mu_{IJ} = 0, \quad i = 1, \dots, I - 1.$$

Alternatively, combining $\mu_{ij} = \mu_{1j}$ with H_{AB2} leads to

$$H_{A2} : \mu_{i1} - \mu_{11} = 0, \quad i = 2, \dots, I.$$

Also, if $\mu_{ij} = \lambda(j) = \bar{\mu}_{\cdot j}$, then combining this with H_{AB3} gives us

$$H_{A3} : \bar{\mu}_{\cdot i} - \bar{\mu}_{\cdot \cdot} = 0, \quad i = 1, \dots, I.$$

This hypothesis can also be expressed in the form

$$H_A : \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \dots = \bar{\mu}_{I\cdot} \quad \text{or} \quad \bar{\mu}_i = \bar{\mu}_{I\cdot} \quad (i = 1, \dots, I - 1).$$

In a similar fashion, we have

$$H_B : \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \dots = \bar{\mu}_{\cdot J},$$

with its alternative forms $H_{B1} : \mu_{Ij} - \mu_{IJ} = 0$, $H_{B2} : \mu_{1j} - \mu_{11} = 0$, and $H_{B3} : \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot \cdot} = 0$.

We note that if we combine with H_{AB} the hypothesis that the μ_{ij} do not vary with i , we get the model $\mu_{ij} = \mu + \beta_j$ ($i = 1, \dots, I$). To obtain this model, we simply omit \mathbf{A}_{12} and \mathbf{A}_1 from \mathbf{L} .

If we use the model (8.12), then since \mathbf{L} does not have full rank, the individual parameters are not identifiable and are therefore not estimable. Only certain functions of them, which are also linear functions of the μ_{ij} , are estimable. Estimability is determined by the model, not the method of specifying the parameters. It is common practice in textbooks (including the first edition of this book) to introduce constraints on the parameters so as to make them all identifiable and therefore estimable. However, as Nelder [1994] cogently argues, this is not necessary. All estimable functions have the same estimates, irrespective of the constraints imposed, but nonestimable functions will have different estimates under different constraints.

One reason for the introduction of constraints in statistics packages is to avoid the use of generalized inverses in fitting less-than-full-rank models. We briefly consider some of these constraints by referring to \mathbf{L} above.

Suppose that we consider the group of hypotheses H_{AB3} , H_{A3} , and H_{B3} , and define $(\alpha\beta)_{ij}^* = \mu_{ij} - \bar{\mu}_{\cdot i} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot \cdot}$, $\alpha_i^* = \bar{\mu}_{\cdot i} - \bar{\mu}_{\cdot \cdot}$, and $\beta_j^* = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot \cdot}$. These parameters then satisfy (8.12) and also satisfy the *symmetric constraints* $\alpha_i^* = 0$, $\beta_j^* = 0$, $(\alpha\beta)_{i\cdot}^* = 0$ ($i = 1, \dots, I - 1$), $(\alpha\beta)_{\cdot j}^* = 0$ ($j = 1, \dots, J - 1$), and $(\alpha\beta)_{\cdot \cdot}^* = 0$. This means that there are only

$$IJ - (I - 1) - (J - 1) - 1 = (I - 1)(J - 1)$$

mathematically independent $(\alpha\beta)_{ij}^*$, $I - 1$ independent α_i^* , and $J - 1$ independent β_j^* . Thus on the right-hand side of (8.12), we have $1 + (I - 1) +$

$(J - 1) + (I - 1)(J - 1) = IJ$ independent parameters, the same number as the μ_{ij} , so that the transformation is one-to-one. Therefore, if we proceed in a reverse direction and start with the symmetric constraints, the parameters are uniquely defined and have the definitions above (cf. Exercises 8b, No. 2). The matrix L can then be reduced to one of full rank by using the constraints to eliminate some of the parameters (e.g., $\alpha_i^* = -\sum_{i=1}^{I-1} \alpha_i^*$, etc). However, computationally, such an approach is not so helpful.

If we use the group of hypotheses H_{AB2} , etc. and set $\gamma_{ij} = \mu_{ij} - \mu_{1j} - \mu_{i1} + \mu_{11}$, $\alpha_i = \mu_{i1} - \mu_{11}$, and $\beta_j = \mu_{1j} - \mu_{11}$, we end up with a different set of constraints, namely, $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{i1} = 0$ ($i = 2, \dots, I$), and $(\alpha\beta)_{1j} = 0$ ($j = 1, \dots, J$). These constraints are the ones used by GLIM, S-PLUS, and R, for example. However, if we use the hypotheses H_{AB1} , etc. and set $\gamma_{ij} = \mu_{ij} - \mu_{1j} - \mu_{iJ} + \mu_{1J}$, $\alpha_i = \mu_{iJ} - \mu_{1J}$, and $\beta_j = \mu_{1j} - \mu_{1J}$, we get the constraints $\alpha_{IJ} = 0$, $\beta_{IJ} = 0$, $(\alpha\beta)_{iJ} = 0$ ($i = 1, \dots, I$), and $(\alpha\beta)_{1j} = 0$ ($j = 1, \dots, J - 1$). These constraints are used by the package SAS, for example, and we can see what is happening if we consider L given by (8.13). Since we have $\alpha_2 = \beta_2 = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$, we are effectively omitting columns 2, 4, 6, 7, and 8 from L . This corresponds to examining the columns of L one at a time from left to right and omitting any linearly dependent columns.

Clearly, it is preferable to work with the unambiguous means model and the hypotheses H_{AB} , H_A , and H_B . However, the parameters μ , α , etc., are introduced for computational convenience but at the expense of possible ambiguity, as we shall see below. The first hypothesis that should be tested is H_{AB} (or its equivalents), and if this is rejected, then H_A and H_B become problematical. For example, if constraints on the parameters are used, the hypotheses H_A , H_{A1} , H_{A2} , and H_{A3} (without H_{AB} being true as well) are different from one another and will produce different residual sums of squares. This means that different statistics packages can produce different answers. Clearly, it is preferable not to impose any constraints on the parameters when using (8.12), which is equivalent to working with the underlying vector spaces specified by the hypotheses. This will be our approach.

In what follows we shall use the model (8.12), but without imposing any constraints on the parameters μ_{ij} . In this case, H_A , for example, is no longer equivalent to $\mu_{ij} - \bar{\mu}_{.j} = 0$ (all i, j), the hypothesis originally intended. There is not much point in testing whether the average effects of A are zero when some of the individual effects $\mu_{ij} - \bar{\mu}_{.j}$ are not zero.

8.3.3 Procedures for Testing the Hypotheses

We begin with the general model $E[Y] = X\mu = XL\gamma$, which we shall call G . The models of interest to us are

$$\begin{aligned} H_{123} : \mu_{ij} &= \mu \\ H_{13} : \mu_{ij} &= \mu + \beta_j \end{aligned}$$

$$\begin{aligned} H_{12} : \mu_{ij} &= \mu + \alpha_i \\ H_1 : \mu_{ij} &= \mu + \alpha_i + \beta_j, \end{aligned}$$

with no constraints on the parameters. The ranks of the underlying regression matrices are the same as the number of free parameters when the symmetric constraints are applied, namely, 1, $J-1$, $I-1$, and $1+I-1+J-1 = I+J-1$, respectively. These models are all subsets of the columns of L ; here H_1 is the same as H_{AB} (we use a different notation later to illustrate the nesting effect of the models above). The residual sum of squares for each model can be computed by imposing constraints on the parameters to reduce each regression matrix to one of full rank.

The model corresponding to $\alpha = 0$ without H_{AB} being true is

$$\mu_{ij} = \mu + \beta_j + (\alpha\beta)_{ij},$$

which, looking at L , for example, is still G , our unconstrained general model. Therefore, if we tested $\alpha = 0$ against G , we would get a zero hypothesis sum of squares in the numerator of the F -test. However, if constraints are applied to all the parameters, then the hypothesis sum of squares will be nonzero.

Several procedures for testing the hypothesis sequence have been proposed, and they all begin with testing H_1 versus G . Let RSS_{123} , RSS_{13} , etc. be the corresponding residual sums of squares obtained by fitting the models H_{123} , H_{13} , and so on; RSS is the usual residual sum of squares from fitting G . The degrees of freedom associated with each residual sum of squares is n minus the rank of the regression matrix. Before proceeding further, we introduce an alternative *R-notation* for $RSS_H - RSS$ which is often used in the literature. Suppose that $\gamma' = (\gamma'_1, \gamma'_2)$ and we wish to test $H : \gamma_1 = 0$ given $\gamma_2 \neq 0$. Let $RSS_H - RSS$ for this hypothesis be denoted by $R(\gamma_1 | \gamma_2)$. Then the numerator sum of squares of the F -statistic for testing H_1 is $RSS_1 - RSS = R((\alpha\beta) | \mu, \alpha, \beta)$. The degrees of freedom associated with the difference of two residual sums of squares is simply the difference in the two degrees of freedom. For example, $RSS_1 - RSS$ has degrees of freedom

$$n - (I + J - 1) - (n - IJ) = IJ - I - J + 1 = (I - 1)(J - 1).$$

We now give three different procedures that can be used after testing H_1 versus G , and we shall use the nomenclature given in the statistics package SAS.

Type 1 Procedure

After testing H_1 , this procedure consists of the following sequence: (1) Test $\beta = 0$ by testing H_{12} versus H_1 using $RSS_{12} - RSS_1 = R(\beta | \mu, \alpha)$; then (2) test $\alpha = 0$ by testing H_{123} versus H_{12} using $RSS_{123} - RSS_{12} = R(\alpha | \mu)$. We now show that these sums of squares, along with $RSS_1 - RSS = R((\alpha\beta) | \mu, \alpha, \beta)$, are mutually independent.

Let G , H_1 , H_{12} , and H_{123} specify that $\theta = E[\mathbf{Y}]$ belongs to Ω , ω_1 , ω_{12} , and ω_{123} , respectively. If \mathbf{P}_ω represents the orthogonal projection onto ω , then consider the decomposition

$$\mathbf{I}_n = (\mathbf{I}_n - \mathbf{P}_\Omega) + (\mathbf{P}_\Omega - \mathbf{P}_1) + (\mathbf{P}_1 - \mathbf{P}_{12}) + (\mathbf{P}_{12} - \mathbf{P}_{123}) + (\mathbf{P}_{123}). \quad (8.14)$$

Since $\omega_{123} \subset \omega_{12} \subset \omega_1 \subset \Omega$, it follows from B.3.1 and B.3.2 that all the matrices in parentheses in (8.14) are symmetric and idempotent, and mutually orthogonal. For example,

$$\begin{aligned} (\mathbf{P}_\Omega - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_{12}) &= \mathbf{P}_\Omega \mathbf{P}_1 - \mathbf{P}_\Omega \mathbf{P}_{12} - \mathbf{P}_1^2 + \mathbf{P}_1 \mathbf{P}_{12} \\ &= \mathbf{P}_1 - \mathbf{P}_{12} - \mathbf{P}_1 + \mathbf{P}_{12} \\ &= \mathbf{0}. \end{aligned}$$

The matrices in parentheses therefore represent orthogonal projections onto the mutually orthogonal subspaces Ω^\perp , $\omega_1^\perp \cap \Omega$, $\omega_{12}^\perp \cap \omega_1$, $\omega_{123}^\perp \cap \omega_{12}$, and ω_{123} respectively (by B.3.2). Therefore, if we multiply (8.14) on the right by \mathbf{Y} , we will obtain an orthogonal decomposition of \mathbf{Y} on the right-hand side, namely,

$$\mathbf{Y} = (\mathbf{I}_n - \mathbf{P}_\Omega)\mathbf{Y} + (\mathbf{P}_\Omega - \mathbf{P}_1)\mathbf{Y} + (\mathbf{P}_1 - \mathbf{P}_{12})\mathbf{Y} + (\mathbf{P}_{12} - \mathbf{P}_{123})\mathbf{Y} + (\mathbf{P}_{123})\mathbf{Y}. \quad (8.15)$$

We also have from (8.15) that

$$\begin{aligned} \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{123})\mathbf{Y} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\Omega)\mathbf{Y} + \mathbf{Y}'(\mathbf{P}_\Omega - \mathbf{P}_1)\mathbf{Y} + \mathbf{Y}'(\mathbf{P}_1 - \mathbf{P}_{12})\mathbf{Y} \\ &\quad + \mathbf{Y}'(\mathbf{P}_{12} - \mathbf{P}_{123})\mathbf{Y} \end{aligned}$$

or

$$\begin{aligned} \text{RSS}_{123} &= \text{RSS} + (\text{RSS}_1 - \text{RSS}) + (\text{RSS}_{12} - \text{RSS}_1) \\ &\quad + (\text{RSS}_{123} - \text{RSS}_{12}) \\ &= \text{RSS} + R((\alpha\beta)|\mu, \alpha, \beta) + R(\beta|\mu, \alpha) + R(\alpha|\mu). \quad (8.16) \end{aligned}$$

Now, from Theorem 1.3 and $\mathbf{P}_{12}\mathbf{P}_1 = \mathbf{P}_{12}$,

$$\text{Cov}[(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y}, (\mathbf{P}_1 - \mathbf{P}_{12})\mathbf{Y}] = \sigma^2(\mathbf{I}_n - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_{12})' = \mathbf{0}.$$

In a similar fashion we can show that the pairwise covariances of all the vectors on the right-hand side of (8.15) are zero (a property of the orthogonal structure). Furthermore, from results such as $\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y}$, it follows from Theorem 2.5 that the sums of squares on the right-hand side of (8.16) are mutually independent. Hence the sums of squares used for testing our nested sequence of hypotheses are independent of each other and of RSS. It is usual to use RSS in the denominator of each test. We note that since $\theta \in \omega_{123}$ if and only if $\theta = \mu\mathbf{1}_n$, $\mathbf{P}_{123} = \mathbf{1}_n(\mathbf{1}'_n \mathbf{1}_n)^{-1} \mathbf{1}'_n = n^{-1} \mathbf{1}_n \mathbf{1}'_n$, and $\text{RSS}_{123} = \sum_i (Y_i - \bar{Y}_.)^2$. The four sums of squares then add up to the (corrected) total sum of squares RSS_{123} , which are expressed this way in computer printouts.

We note that two orderings are possible, as we can interchange the roles of the two factors; this amounts to using H_{13} instead of H_{12} , and interchanging the subscripts 2 and 3 in the theory above. These two orders will lead to different breakdowns of $\sum_i(Y_i - \bar{Y}_.)^2$, which may lead to conflicting models and the need for a model selection method. The type 1 method is also used by S-PLUS and GLIM.

Type 2 Procedure

In addition to testing H_1 , this procedure consists of the following sequence: (1) Test $\beta = 0$ by testing H_{12} versus H_1 using $RSS_{12} - RSS_1 = R(\beta|\mu, \alpha)$; then (2) test $\alpha = 0$ by testing H_{13} versus H_1 using $RSS_{13} - RSS_1 = R(\alpha|\mu, \beta)$. This method treats the two factors symmetrically and effectively combines the two orderings used in the type 1 method. Clearly, the various sums of squares no longer add to $\sum_i(Y_i - \bar{Y}_.)^2$ in the unbalanced case. This method does not have the nice independence properties enjoyed by the type 1 procedure. Also, one questions the appropriateness of moving outside a nested framework; a point considered by Nelder [1994] under the title of *marginality*.

Type 3 Procedure

In addition to testing H_1 (H_{AB}), this procedure consists of testing two *marginal means* hypotheses, namely, $H_A : \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_{I.}$, and $H_B : \bar{\mu}_{.1} = \bar{\mu}_{.2} = \dots = \bar{\mu}_{.J}$ versus G . Using the model (8.12) (without restrictions), we see that H_A , for example, is equivalent to testing that $\alpha_i + (\alpha\beta)_{i.}$ is constant for all i , which is not a very meaningful hypothesis when the $(\alpha\beta)_{ij}$ are not zero. This hypothesis can be tested formally by imposing, for computational purposes, the symmetric constraints on the parameters and then testing $H_{A3} : \alpha_i^* = \bar{\mu}_{i.} - \bar{\mu}_{..} = 0$. Then $RSS_{H_{A3}} - RSS$ can be expressed as $R(\alpha^*|\mu^*, \beta^*, (\alpha\beta)^*)$, which is the starred notation used by Speed and Hocking [1976] and Speed et al. [1978]. Using * indicates that the value of R depends on the constraints used; alternative constraints will lead to a different value. Without the constraints we saw that $R(\alpha|\mu, \beta, (\alpha\beta)) = 0$, which highlights a problem with the R -notation; computer packages don't always clarify which one is being used.

When one or more cell frequencies K_{ij} are zero, the procedures above can be adapted accordingly in statistical packages; in SAS there is then a further procedure referred to as a type 4 procedure. When there are no missing cells, types 3 and 4 sums of squares are identical.

In conclusion, we note that there is considerable controversy and confusion over which is the appropriate procedure (cf. Hocking [1996: Section 13.2]). However, if H_1 is rejected, there are problems with all three methods.

8.3.4 Confidence Intervals

As we can regard a two-way classification as a one-way classification with IJ populations (groups), we can use the methods of Section 8.2.3 to construct confidence intervals for one or several contrasts in the $\{\mu_{ij}\}$. If we use the

symmetric constraints, we can also construct simultaneous confidence intervals for contrasts in $\{\alpha_i^*\}$, $\{\beta_j^*\}$, and the $\{(\alpha\beta)_{ij}^*\}$. For example, expressing H_A in the form $\phi = \mathbf{A}\mu = 0$ [$\phi' = (\alpha_1^*, \alpha_2^*, \dots, \alpha_{I-1}^*)$], where $\alpha_i^* = \bar{\mu}_{i\cdot} - \bar{\mu}_{..}$, we can use Scheffé's method [cf. (5.13)] as follows.

We note that

$$\begin{aligned} \sum_{i=1}^{I-1} h_i \alpha_i^* &= \sum_{i=1}^{I-1} h_i (\bar{\mu}_{i\cdot} - \bar{\mu}_{..}) \\ &= \sum_{i=1}^{I-1} \left(h_i - \frac{1}{I} \sum_{i=1}^{I-1} h_i \right) \bar{\mu}_{i\cdot} + \left(-\frac{1}{I} \sum_{i=1}^{I-1} h_i \right) \bar{\mu}_I \\ &= \sum_{i=1}^I c_i \bar{\mu}_{i\cdot} \\ &= \sum_{i=1}^I c_i \alpha_i^*, \end{aligned} \quad (8.17)$$

where $\sum c_i = 0$. Conversely, by writing $\alpha_I^* = -\sum_{i=1}^{I-1} \alpha_i^*$, we see that (8.17) is expressible in the form $\sum_{i=1}^{I-1} h_i \alpha_i^*$. These two statements imply that the set of all linear combinations $\mathbf{h}'\phi$ is the set of all contrasts in $\alpha_1^*, \alpha_2^*, \dots, \alpha_I^*$. Now

$$\sum_i c_i \hat{\alpha}_i^* = \sum_i c_i (\bar{Y}_{i\cdot} - \bar{Y}_{..}) = \sum_i c_i \bar{Y}_{i\cdot},$$

so that

$$\text{var}[\sum_i c_i \hat{\alpha}_i^*] = \sigma^2 \sum_i \frac{c_i^2}{K_{i\cdot}},$$

where $K_{i\cdot} = \sum_j K_{ij}$. Hence

$$\begin{aligned} 1 - \alpha &= \text{pr} \left[\sum_i c_i \bar{\mu}_{i\cdot} \in \sum_i c_i \bar{Y}_{i\cdot} \pm [(I-1)F_{I-1, n-IJ}^\alpha]^{1/2} S \left(\sum_i \frac{c_i^2}{K_{i\cdot}} \right)^{1/2}, \right. \\ &\quad \left. \text{for all contrasts} \right], \end{aligned}$$

where

$$S^2 = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij\cdot})^2 / n \quad \text{and} \quad n = \sum_i \sum_j K_{ij} = K_{..}$$

Similar methods can be applied using H_B and H_{AB} .

EXERCISES 8b

1. Prove that H_{AB} , H_{AB1} , H_{AB2} , and H_{AB3} are all equivalent to $\mu_{ij} = \mu + \alpha_i + \beta_j$.

2. Using the symmetric constraints, express the parameters μ , α_i , etc. in terms of the μ_{ij} .
3. Obtain an F -test statistic for testing $H : \mu_{ij} = \mu$ in a two-way classification. Find the expected value of the numerator sum of squares of your test statistics.

8.4 TWO-WAY CLASSIFICATION (BALANCED)

When $K_{ij} = K$ for all i and j , the model is said to be *balanced*. In this case we find that the various sums of squares have some nice properties. To begin with, all three procedures are identical, as we shall show below that

$$R(\alpha^*|\mu^*) = R(\alpha^*|\mu^*, \beta^*) = R(\alpha^*|, \mu^*, \beta^*, (\alpha\beta)^*), \quad (8.18)$$

and

$$R(\alpha^*|\mu^*) = R(\alpha|\mu), \text{ and } R(\alpha^*|\mu^*, \beta^*) = R(\alpha|\mu, \beta). \quad (8.19)$$

We recall, however, that $R(\alpha^*|, \mu^*, \beta^*, (\alpha\beta)^*) \neq R(\alpha|, \mu, \beta, (\alpha\beta)) (= 0)$. Similar results can be obtained by interchanging the roles of α and β .

To find RSS_H for H_{AB} , H_A , and H_B , we can use the symmetric constraints which are based on the decomposition

$$\begin{aligned} \mu_{ij} &= \bar{\mu}_{..} + (\bar{\mu}_{i..} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + (\mu_{ij} - \bar{\mu}_{i..} - \bar{\mu}_{.j} + \bar{\mu}_{..}) \\ &= \mu^* + \alpha_i^* + \beta_j^* + (\alpha\beta)_{ij}^* \end{aligned} \quad (8.20)$$

with a corresponding decomposition of ε_{ijk} , namely,

$$\begin{aligned} \varepsilon_{ijk} &= \bar{\varepsilon}_{...} + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{...}) \\ &\quad + (\bar{\varepsilon}_{ij.} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j} + \bar{\varepsilon}_{...}) + (\varepsilon_{ijk} - \bar{\varepsilon}_{ij.}). \end{aligned} \quad (8.21)$$

Squaring, and summing on i , j , and k , we find that the cross-product terms vanish (because the elements come from mutually orthogonal vectors) and

$$\begin{aligned} \sum \sum \sum \varepsilon_{ijk}^2 &= \sum \sum \sum \bar{\varepsilon}^2_{...} + \sum \sum \sum (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...})^2 + \dots \\ &\quad + \sum \sum \sum (\varepsilon_{ijk} - \bar{\varepsilon}_{ij.})^2. \end{aligned} \quad (8.22)$$

Setting $\varepsilon_{ijk} = Y_{ijk} - \mu^* - \alpha_i^* - \beta_j^* - (\alpha\beta)_{ij}^*$, and using $\alpha_i^* = 0$, etc., we obtain

$$\begin{aligned} \sum \sum \sum (Y_{ijk} - \mu^* - \alpha_i^* - \beta_j^* - (\alpha\beta)_{ij}^*)^2 &= \sum \sum \sum (\bar{Y}_{...} - \mu)^2 + \sum \sum \sum (\bar{Y}_{i..} - \bar{Y}_{...} - \alpha_i^*)^2 \\ &\quad + \sum \sum \sum (\bar{Y}_{.j} - \bar{Y}_{...} - \beta_j^*)^2 \\ &\quad + \sum \sum \sum (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...} - (\alpha\beta)_{ij}^*)^2 \\ &\quad + \sum \sum \sum (Y_{ijk} - \bar{Y}_{ij.})^2. \end{aligned} \quad (8.23)$$

By inspection, the right side of (8.23) is minimized (subject to $\alpha_i^* = 0$, etc.) when the unknown parameters take the values

$$\hat{\mu}^* = \bar{Y}_{...}, \quad \hat{\alpha}_i^* = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j^* = \bar{Y}_{.j.} - \bar{Y}_{...},$$

and

$$(\widehat{\alpha\beta})_{ij}^* = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}.$$

Because of our previous discussion, we won't refer to these quantities as estimates because they depend on which constraints are used. Rather, they are intermediate quantities which assist us in finding each residual sum of squares. By substitution we get $\hat{\mu}_{ij} = \hat{\mu}^* + \hat{\alpha}_i^* + \hat{\beta}_j^* + (\widehat{\alpha\beta})_{ij}^* = \bar{Y}_{ij.}$ and

$$\text{RSS} = \sum \sum \sum (Y_{ijk} - \bar{Y}_{ij.})^2,$$

as before. This estimate of μ_{ij} and the residual sum of squares will be the same irrespective of the method of reparameterization.

To find $\text{RSS}_{H_{AB}}$, we must minimize (8.23) subject to $(\widehat{\alpha\beta})_{ij}^* = 0$ for all i, j . By inspection this minimum occurs at $\hat{\mu}^*$, $\hat{\alpha}_i^*$, and $\hat{\beta}_j^*$, so that

$$\text{RSS}_{H_{AB}} = \sum \sum \sum (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 + \sum \sum \sum (Y_{ijk} - \bar{Y}_{ij.})^2$$

and

$$\begin{aligned} \text{RSS}_{H_{AB}} - \text{RSS} &= \sum \sum \sum (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \\ &= K \sum_i \sum_j (\widehat{\alpha\beta})_{ij}^{*2}. \end{aligned}$$

The F -statistic for testing H_{AB} is therefore

$$F = \frac{K \sum_i \sum_j (\widehat{\alpha\beta})_{ij}^{*2} / (I-1)(J-1)}{\text{RSS}/(IJK - IJ)} = \frac{S_{AB}^2}{S^2}, \quad (8.24)$$

say, which has an F -distribution with $(I-1)(J-1)$ and $IJK - IJ$ degrees of freedom, respectively, when H_{AB} is true.

Test statistics for H_A and H_B are obtained in a similar fashion. Setting $\alpha_i^* = 0$, for example, in (8.23), we find that the minimum of (8.23) now occurs at $\hat{\mu}^*$, $\hat{\beta}_j^*$, and $(\widehat{\alpha\beta})_{ij}^*$. Thus

$$\text{RSS}_{H_A} = \sum \sum \sum (\bar{Y}_{i..} - \bar{Y}_{...})^2 + \sum \sum \sum (\bar{Y}_{ijk} - \bar{Y}_{i..})^2,$$

$$\text{RSS}_{H_A} - \text{RSS} = \sum \sum \sum (\bar{Y}_{i..} - \bar{Y}_{...})^2 = JK \sum_i \hat{\alpha}_i^{*2},$$

and

$$F = \frac{JK \sum_i \hat{\alpha}_i^{*2} / (I-1)}{\text{RSS}/(IJK - IJ)} = \frac{S_A^2}{S^2}, \quad (8.25)$$

say, is the F -statistic for testing H_A . The corresponding statistic for H_B is

$$F = \frac{IK \sum_j \hat{\beta}_j^{*2} / (J - 1)}{\text{RSS}/(IJK - IJ)} = \frac{S_B^2}{S^2}. \quad (8.26)$$

In all the derivations above we see that the minimizing values of the parameters are always the same, irrespective of which ones are set equal to zero. Using the fact that $\text{RSS}_H - \text{RSS} = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H\|^2$ for any pair of hypotheses with one nested within the other, we have

$$\begin{aligned} R(\alpha^* | \mu^*) &= \sum_{ijk} [\hat{\mu}^* + \hat{\alpha}_i^* - (\hat{\mu}^*)]^2 \\ &= \sum_{ijk} [\hat{\mu}^* + \hat{\alpha}_i^* + \hat{\beta}_j^* - (\hat{\mu}^* + \hat{\beta}_j^*)]^2 \\ &= \sum_{ijk} [\hat{\mu}^* + \hat{\alpha}_i^* + \hat{\beta}_j^* + (\widehat{\alpha\beta})_{ij}^* - (\hat{\mu}^* + \hat{\beta}_j^* + (\widehat{\alpha\beta})_{ij}^*)]^2 \\ &= \sum_{ijk} \hat{\alpha}_i^{*2} \quad \left(= JK \sum_i \hat{\alpha}_i^{*2} \right), \end{aligned}$$

thus proving (8.18). Equation (8.19) follows from the fact that when we have the model $\mu_{ij} = \mu + \alpha_i + \beta_j$ and any submodels of it, the symmetric constraints do not change the subspaces represented by the design matrices; the only change is that the matrices are changed to full rank, and this does not affect the residual sum of squares. Although the number of degrees of freedom for each RSS_H is given above, they can also be obtained from the coefficient of σ^2 in $E[\text{RSS}_H]$ (these expected values are given in Exercises 8c, No. 1).

Analysis-of-Variance Table

As in the one-way classification, the various sums of squares are normally set out in the form of a table (Table 8.2). The various sums of squares in Table 8.2 add up to $\sum \sum \sum (Y_{ijk} - \bar{Y}_{ij.})^2$; this follows either from (8.16), which also shows that they are independent, or from the fact that (8.22) is an *identity* in the ϵ_{ijk} and therefore also holds for the Y_{ijk} . Since

$$\sum \sum \sum (\bar{Y}_{i..} - \bar{Y}_{...})^2 = JK \sum_i \hat{\alpha}_i^{*2},$$

this sum of squares is usually called the *sum of squares due to the A main effects*, although some textbooks use the term *row sum of squares*. Similar designations apply to the next two sums of squares in Table 8.2, as they are $IK \sum_j \hat{\beta}_j^{*2}$ and $K \sum_i \sum_{ij} (\widehat{\alpha\beta})_{ij}^{*2}$, respectively. The sum of squares labeled “error” gives a pooled estimate of σ^2 based on all IJ normal populations; this term is also called the *within populations sum of squares* or the *residual sum of squares*. As in the unbalanced case, we need to look first at H_{AB} .

Table 8.2 Analysis-of-variance table for a two-way classification with $K(K > 1)$ observations per population mean

Source	Sum of squares (SS)	Degrees of freedom (df)	SS df
A main effects	$JK \sum_i \hat{\alpha}_i^{*2}$	$I - 1$	S_A^2
B main effects	$IK \sum_j \hat{\beta}_j^{*2}$	$J - 1$	S_B^2
AB interactions	$K \sum_i \sum_j (\widehat{\alpha\beta})_{ij}^{*2}$	$(I - 1)(J - 1)$	S_{AB}^2
Error	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$	$IJ(K - 1)$	S^2
Corrected total	$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2$	$IJK - 1$	

EXERCISES 8c

1. (a) Prove that

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (\bar{Y}_{i..} - \bar{Y}_{...})^2 = \sum_{i=1}^I \frac{Y_{i..}^2}{JK} - \frac{Y_{...}^2}{IJK}.$$

Obtain a similar expression for

$$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2.$$

- (b) Using Table 8.2, prove the following:

$$\begin{aligned} E[(I - 1)S_A^2] &= \sigma^2(I - 1) + JK \sum_i \alpha_i^{*2} \\ E[(J - 1)S_B^2] &= \sigma^2(J - 1) + IK \sum_j \beta_j^{*2} \\ E[(I - 1)S_{AB}^2] &= \sigma^2(I - 1)(J - 1) + K \sum_i \sum_j (\alpha\beta)_{ij}^{*2}. \end{aligned}$$

2. Given the population means μ_{ij} ($i = 1, 2, \dots, I; j = 1, 2, \dots, J$), let

$$\begin{aligned} A_i &= \sum_j v_j \mu_{ij} \quad \left(\sum_j v_j = 1 \right), \\ B_i &= \sum_i u_i \mu_{ij} \quad \left(\sum_i u_i = 1 \right), \end{aligned}$$

and

$$\mu = \sum_i u_i A_i = \sum_j v_j B_j = \sum_i \sum_j u_i v_j \mu_{ij}.$$

Define $\alpha_i = A_i - \mu$, $\beta_j = B_j - \mu$, and

$$(\alpha\beta)_{ij} = \mu_{ij} - A_i - B_j + \mu.$$

(a) Show that $\sum_i u_i \alpha_i = \sum_j v_j \beta_j = 0$, $\sum_i u_i (\alpha\beta)_{ij} = 0$ (all j), and $\sum_j v_j (\alpha\beta)_{ij} = 0$ (all i).

(b) Conversely, given

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

show that the parameters in the equation above are uniquely determined by the constraints in (a).

(c) Prove that if the interactions $\{(\alpha\beta)_{ij}\}$ are all zero for some system of weights $\{u_i\}$ and $\{v_i\}$, then they are zero for every system of weights. In that case show that every contrast in the $\{\alpha_i\}$, or $\{\beta_i\}$, has a value that does not depend on the system of weights.

(Scheffé [1959: Section 4.1])

3. Let $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, where $i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K_{ij}$; and the ε_{ijk} are independently distributed as $N(0, \sigma^2)$. Given $K_{ij} = K_i K_{.j} / K_{..}$ for all i, j , find a test statistic for testing the hypothesis $H : (\alpha\beta)_{ij} = 0$ (all i, j). Hint: By Exercise 2, the validity of H does not depend on the weights used in the identifiability constraints $\sum_i u_i \alpha_i = \sum_j v_j \beta_j = 0$. We can therefore use $u_i = K_{i.} / K_{..}$ and $v_j = K_{.j} / K_{..}$ and find the least squares estimates of α_i and β_j when H is true.

(Scheffé [1959: p. 119])

8.5 TWO-WAY CLASSIFICATION (ONE OBSERVATION PER MEAN)

Suppose that in a two-way classification there is only one observation per mean, so that the model becomes

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} \quad (i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J), \quad (8.27)$$

where the ε_{ij} are i.i.d. $N(0, \sigma^2)$. We now have IJ observations but $IJ + 1$ unknown parameters $\{\mu_{ij}\}$ and σ^2 , so that we cannot estimate all the parameters without imposing at least one constraint to reduce the number of “free” parameters. However, typically, such data come from a randomized block design with I treatments and J blocks. Here each block has I sampling units and the I treatments are applied to the units in random order so that we would expect the interaction between treatment and block number to be small. A reasonable assumption would then be that our model for this experiment is the additive model

$$G : \mu_{ij} = \mu + \alpha_i + \beta_j,$$

where there are no constraints on the parameters. In Section 8.4 we saw that to compute residual sums of squares, we can reduce the model to one of full rank using, for example, the symmetric constraints (but we now drop the * label, for convenience). Since we have $\alpha_i = \bar{\mu}_{i\cdot} - \bar{\mu}_{..}$, and similarly for β_j , we find that the additive model is equivalent to

$$\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{..} = 0 \quad \text{for all } i, j. \quad (8.28)$$

We can therefore express G in the form

$$\mu_{ij} = \bar{\mu}_{..} + (\bar{\mu}_{i\cdot} - \bar{\mu}_{..}) + (\bar{\mu}_{\cdot j} - \bar{\mu}_{..}) + (\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{..}),$$

which suggests, as in Section 8.4, the corresponding decomposition

$$\varepsilon_{ij} = \bar{\varepsilon}_{..} + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{..}) + (\bar{\varepsilon}_{\cdot j} - \bar{\varepsilon}_{..}) + (\bar{\varepsilon}_{ij} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot j} + \bar{\varepsilon}_{..}).$$

Algebra similar to that used in Section 8.4 leads to

$$\begin{aligned} \sum \sum \varepsilon_{ij}^2 &= \sum \sum \varepsilon_{..}^2 + \sum \sum (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{..})^2 \\ &\quad + \sum \sum (\bar{\varepsilon}_{\cdot j} - \bar{\varepsilon}_{..})^2 + \sum \sum (\bar{\varepsilon}_{ij} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot j} + \bar{\varepsilon}_{..})^2. \end{aligned}$$

Setting $\varepsilon_{ij} = Y_{ij} - \mu - \alpha_i - \beta_j$ and applying the constraints α_i and β_j to the model, we have

$$\begin{aligned} \sum \sum (Y_{ij} - \mu - \alpha_i - \beta_j)^2 &= \sum \sum (\bar{Y}_{..} - \mu)^2 + \sum \sum (\bar{Y}_{i\cdot} - \bar{Y}_{..} - \alpha_i)^2 \\ &\quad + \sum \sum (\bar{Y}_{\cdot j} - \bar{Y}_{..} - \beta_j)^2 \\ &\quad + \sum \sum (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2. \quad (8.29) \end{aligned}$$

The left-hand side of (8.29) is minimized when $\mu = \bar{Y}_{..}$ ($= \hat{\mu}$), $\alpha_i = \bar{Y}_{i\cdot} - \bar{Y}_{..}$ ($= \hat{\alpha}_i$), and $\beta_j = \bar{Y}_{\cdot j} - \bar{Y}_{..}$ ($= \hat{\beta}_j$), so that $\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{..}$ and

$$\begin{aligned}\text{RSS} &= \sum \sum (Y_{ij} - \hat{\mu}_{ij})^2 \\ &= \sum \sum (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2 \\ &= \sum \sum (\hat{\alpha}\hat{\beta})_{ij}^2,\end{aligned}$$

say, with degrees of freedom $IJ - (I + J - 1) = (I - 1)(J - 1)$. Since the interactions are zero, we see from the equation above that the “interaction” sum of squares takes over the role of the error sum of squares, and an unbiased estimate of σ^2 is $\text{RSS}/(I - 1)(J - 1)$. However, it should be pointed out that an estimate of σ^2 can be found by making much weaker assumptions; not all the interactions need to be zero (Johnson and Graybill [1972a,b]).

Setting $\alpha_i = 0$, we see, by inspection, that the left side of (8.29) is minimized when $\mu = \hat{\mu}$ and $\beta_j = \hat{\beta}_j$ so that

$$\text{RSS}_{H_A} = \sum \sum (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + \sum \sum (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2$$

and

$$\text{RSS}_{H_A} - \text{RSS} = \sum \sum (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2.$$

Hence the F -statistic for testing H_A is

$$\begin{aligned}F &= \frac{\sum \sum (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 / (I - 1)}{\sum \sum (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2 / (I - 1)(J - 1)} \\ &= \frac{J \sum_i \hat{\alpha}_i^2 / (I - 1)}{\sum \sum (\hat{\alpha}\hat{\beta})_{ij}^2 / (I - 1)(J - 1)}. \quad (8.30)\end{aligned}$$

The test statistic for H_B follows by interchanging i and j , namely,

$$F = \frac{I \sum_j \hat{\beta}_j^2 / (J - 1)}{\sum \sum (\hat{\alpha}\hat{\beta})_{ij}^2 / (I - 1)(J - 1)}. \quad (8.31)$$

The entire procedure can be summarized as in Table 8.3. We note that with the interactions zero, H_A is equivalent to $\mu_{ij} - \bar{\mu}_{\cdot j} = 0$; that is, μ_{ij} does not depend on i for every j . In a similar fashion we see that H_B is equivalent to μ_{ij} not depending on j for every i .

8.5.1 Underlying Assumptions

A key assumption in the analysis above is (8.28). We shall call the quantities $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{..}$, the *interactions*. Because we cannot estimate all the parameters in (8.27), we cannot test that $(\alpha\beta)_{ij} = 0$ for all i, j against

Table 8.3 Analysis-of-variance table for a two-way classification with one observation per population mean

Source	Sum of squares (SS)	Degrees of freedom (df)	SS \bar{df}
A main effects (treatments)	$J \sum_i \hat{\alpha}_i^2$	$I - 1$	S_A^2
B main effects (blocks)	$I \sum_j \hat{\beta}_j^2$	$J - 1$	S_B^2
Error	$\sum_i \sum_j (\hat{\alpha}\hat{\beta})_{ij}^2$	$(I - 1)(J - 1)$	S^2
Corrected total	$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$IJ - 1$	

a general class of alternatives $(\alpha\beta)_{ij} \neq 0$ [for at least one pair (i, j)]. We therefore have to resort to carrying out our test against a suitably restricted class of alternatives; several such classes have been considered. For example, if we assume that $(\alpha\beta)_{ij} = \gamma\alpha_i\beta_j$, then Tukey's [1949] well-known test for additivity is equivalent to testing the null hypothesis $H_\gamma : \gamma = 0$ against the alternative $\gamma \neq 0$ (Scheffé [1959: pp. 129–137]). Tukey's test statistic is

$$F = \frac{SS_\gamma}{(\text{RSS} - SS_\gamma)/[(I - 1)(J - 1) - 1]}, \quad (8.32)$$

where

$$SS_\gamma = \frac{\left(\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j Y_{ij} \right)^2}{\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2}$$

and

$$\text{RSS} = \sum_i \sum_j (\hat{\alpha}\hat{\beta})_{ij}^2.$$

Then F has an F -distribution with 1 and $IJ - I - J$ degrees of freedom, respectively, when the underlying model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}.$$

It is instructive to derive (8.32) from the following lemma due to Scheffé [1959: p. 144, Example 4.19].

LEMMA Suppose that $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, where \mathbf{X} is $n \times p$ of rank p , and define $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$. Let $\mathbf{Z} = \mathbf{f}(\hat{\boldsymbol{\theta}})$

be a continuous function of $\hat{\theta}$ (chosen before the outcome of \mathbf{Y} is inspected), and let $\hat{\phi}$ be the same linear function of \mathbf{Z} that $\hat{\theta}$ is of \mathbf{Y} . Define $R = \|\mathbf{Y} - \hat{\theta}\|^2$ and

$$R_1 = \frac{\mathbf{Z}'(\mathbf{Y} - \hat{\theta})}{\{(\mathbf{Z} - \hat{\phi})'(\mathbf{Z} - \hat{\phi})\}^{1/2}}.$$

Then

$$F_0 = \frac{R_1^2}{(R - R_1^2)/(n - p - 1)} \sim F_{1,n-p-1}.$$

Proof. $\hat{\theta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y}$, so that $\hat{\phi} = \mathbf{P}\mathbf{Z}$ and

$$R_1 = \frac{\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}}{\{\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}\}^{1/2}} \quad [= \mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}/c_Z, \text{ say}].$$

Consider the distributions of R and R_1 conditional on $\mathbf{Z} = \mathbf{z}$. Since R is independent of $\hat{\beta}$ [Theorem 3.5(iii), Section 3.4], and therefore of \mathbf{Z} , the conditional distribution of R/σ^2 is the same as the unconditional distribution, namely, χ_{n-p}^2 [Theorem 3.5(iv)]. Also, from $(\mathbf{I}_n - \mathbf{P})\mathbf{X} = \mathbf{0}$, $R_1 = \mathbf{Z}'(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\hat{\beta})/c_z$, where $\mathbf{Y} - \mathbf{X}\hat{\beta}$ is independent of $\hat{\theta}$ and therefore of \mathbf{Z} . Hence

$$E[R_1 | \mathbf{Z} = \mathbf{z}] = \mathbf{z}'(\mathbf{I}_n - \mathbf{P})E[\mathbf{Y} - \mathbf{X}\hat{\beta}]/c_z = 0$$

and

$$\begin{aligned} \text{var}[R_1 | \mathbf{Z} = \mathbf{z}] &= \frac{\mathbf{z}'(\mathbf{I}_n - \mathbf{P}) \text{Var}[\mathbf{Y} - \mathbf{X}\hat{\beta}](\mathbf{I}_n - \mathbf{P})'\mathbf{z}}{c_z^2} \\ &= \sigma^2 \frac{\mathbf{z}'(\mathbf{I}_n - \mathbf{P})(\mathbf{I}_n - \mathbf{P})(\mathbf{I}_n - \mathbf{P})\mathbf{z}}{c_z^2} = \sigma^2 \end{aligned}$$

imply that R_1 is conditionally $N(0, \sigma^2)$. As this does not involve \mathbf{z} , it is also the unconditional distribution. Now setting $\mathbf{u} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$ and $\mathbf{v} = (\mathbf{I}_n - \mathbf{P})\mathbf{z}$ and invoking the Cauchy–Schwartz inequality (A.4.11), we have

$$\begin{aligned} R - R_1^2 &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} - \frac{\{\mathbf{z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}\}^2}{\mathbf{z}'(\mathbf{I}_n - \mathbf{P})\mathbf{z}} \\ &= \mathbf{u}'\mathbf{u} - \frac{(\mathbf{u}'\mathbf{v})^2}{\mathbf{v}'\mathbf{v}} \\ &= \frac{(\mathbf{u}'\mathbf{u})(\mathbf{v}'\mathbf{v}) - (\mathbf{u}'\mathbf{v})^2}{\mathbf{v}'\mathbf{v}} \\ &\geq 0. \end{aligned}$$

Then, since $R/\sigma^2 \sim \chi_{n-p}^2$ and $R_1^2/\sigma^2 \sim \chi_1^2$, we have, by Example 2.13 in Section 2.4, that $(R - R_1^2)/\sigma^2$ and R_1^2/σ^2 are independently distributed as χ_{n-p-1}^2 and χ_1^2 , respectively. Thus $F_0 \sim F_{1,n-p-1}$, and because the F -distribution does not depend on \mathbf{z} , it is also the unconditional distribution of F_0 . \square

To apply this Lemma to (8.32), we simply define $\mathbf{Z} = \mathbf{f}(\hat{\boldsymbol{\theta}})$ by $Z_{ij} = \hat{\theta}_{ij}^2$, where $\hat{\theta}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$. Then

$$\|\mathbf{Y} - \hat{\boldsymbol{\theta}}\|^2 = \text{RSS} = \sum \sum (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..})^2$$

and

$$\begin{aligned} R_1^2 &= \frac{\{[(\mathbf{I}_n - \mathbf{P})\mathbf{Z}]'\mathbf{Y}\}^2}{\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}} \\ &= \frac{\{\sum \sum (Z_{ij} - \bar{Z}_{i\cdot} - \bar{Z}_{\cdot j} + \bar{Z}_{..})Y_{ij}\}^2}{\sum \sum (Z_{ij} - \bar{Z}_{i\cdot} - \bar{Z}_{\cdot j} + \bar{Z}_{..})^2}. \end{aligned}$$

Using $\hat{\alpha}_{\cdot} = \hat{\beta}_{\cdot} = 0$, we have after some algebra that

$$Z_{ij} - \bar{Z}_{i\cdot} - \bar{Z}_{\cdot j} + \bar{Z}_{..} = 2\hat{\alpha}_i\hat{\beta}_j$$

so that

$$R_1^2 = \frac{\{\sum \sum \hat{\alpha}_i\hat{\beta}_j Y_{ij}\}^2}{\sum \sum \hat{\alpha}_i^2\hat{\beta}_j^2}, \quad (8.33)$$

and we have derived (8.32). A similar method can be used for deriving a test for interaction for other experimental designs which assume additivity, for example, the Latin square.

Tukey's test, originally proposed without specifying any particular form for $(\alpha\beta)_{ij}$, seems to have reasonably good power for the alternatives $\gamma \neq 0$ (Ghosh and Sharma [1963]), and the effect of nonnormality on the test is examined empirically by Yates [1972]. Several generalizations of the procedure have been proposed (cf. Johnson and Graybill [1972a,b] for references), and all these tests would appear to have reasonably good power when $(\alpha\beta)_{ij}$ is a function of the α_i or β_j . Johnson and Graybill [1972b] also proposed a test for interaction that would have a reasonable power when the underlying model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \lambda\gamma_i\delta_j + \varepsilon_{ij}, \quad (8.34)$$

where $\alpha_{\cdot} = \beta_{\cdot} = \gamma_{\cdot} = \delta_{\cdot} = 0$ and $\sum_i \gamma_i^2 = \sum_j \delta_j^2 = 1$.

Residual plots based on the residuals $(\widehat{\alpha\beta})_{ij}$ must be interpreted with caution. Any irregularities could be due to either departures from the usual normality assumptions or to the presence of nonzero interactions (i.e., $E[(\widehat{\alpha\beta})_{ij}] \neq 0$ for some i, j). Because the F -tests (8.30) and (8.31) are quadratically balanced (cf. Section 9.5.2), we would expect the test statistics to be robust with regard to nonnormality. As a bonus, the effect of randomizing the treatments in the blocks induces a *randomization* distribution for the data which is approximately normal. Any heterogeneity of variance or error correlations within blocks could be tested using, for example, the methods of Han [1969].

8.6 HIGHER-WAY CLASSIFICATIONS WITH EQUAL NUMBERS PER MEAN

8.6.1 Definition of Interactions

The extension of the theory in Section 8.4 to higher-way classifications with equal numbers of observations per mean is fairly straightforward and is demonstrated briefly by considering the three-way classification:

$$Y_{ijkm} = \mu_{ijk} + \varepsilon_{ijkm}, \quad (8.35)$$

where $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; $k = 1, 2, \dots, K$; $m = 1, 2, \dots, M$, and the ε_{ijkm} are i.i.d. $N(0, \sigma^2)$. Here we have three factors: A at I levels, B at J levels, C at K levels; and there are M ($M > 1$) observations per population mean for each of the IJK means. In addition to the (first-order) interactions between A and B , B and C , and A and C , we now have the possibility of a (second-order) interaction between all three factors. If, however, the factors interact only in pairs, so that, for example, the AB interactions are not affected by C , then an AB interaction would be the same for all levels of C . Mathematically, this means that

$$\begin{aligned} \mu_{ijk} - \bar{\mu}_{i..} - \bar{\mu}_{.jk} + \bar{\mu}_{...k} &= \psi(i, j) \\ &= \sum_{k=1}^K \psi(i, j)/K \\ &= \bar{\mu}_{ij.} - \bar{\mu}_{i..} - \bar{\mu}_{.j.} + \bar{\mu}_{...} \end{aligned}$$

or

$$\begin{aligned} (\alpha\beta\gamma)_{ijk} &= \mu_{ijk} - \bar{\mu}_{ij.} - \bar{\mu}_{.jk} - \bar{\mu}_{i..} + \bar{\mu}_{i..} + \bar{\mu}_{.j.} + \bar{\mu}_{..k} - \bar{\mu}_{...} \\ &= 0. \end{aligned}$$

(A numerical example demonstrating this is given in Exercises 8d, No. 2, at the end of Section 8.6.3.) Since $(\alpha\beta\gamma)_{ijk}$ is symmetric in i, j, k , we see that we would have arrived at the same result if we considered the BC interaction at different levels of A , or the AC interaction at different levels of B . It therefore seems appropriate to define $(\alpha\beta\gamma)_{ijk}$ as the second-order interaction between the i th level of A , the j th level of B , and the k th level of C . We refer to these interactions simply as *ABC interactions*.

Our two-factor concepts of Section 8.4 can be carried over to this situation by considering a two-way table for each level of C . For example, the interaction of the i th level of A with the j th level of B , given that C is at level k , is

$$\mu_{ijk} - \bar{\mu}_{i..} - \bar{\mu}_{.jk} + \bar{\mu}_{...k}. \quad (8.36)$$

The average of these over the levels of C , namely,

$$(\alpha\beta)_{ij} = \bar{\mu}_{ij.} - \bar{\mu}_{i..} - \bar{\mu}_{.j.} + \bar{\mu}_{...},$$

we call the interaction of the i th level of A with the j th level of B . We similarly define the BC and AC interactions to be

$$(\beta\gamma)_{jk} = \bar{\mu}_{.jk} - \bar{\mu}_{.j.} - \bar{\mu}_{..k} + \bar{\mu}_{...}$$

and

$$(\alpha\gamma)_{ik} = \bar{\mu}_{ik} - \bar{\mu}_{i..} - \bar{\mu}_{..k} + \bar{\mu}_{...}.$$

By analogy with Section 8.4, we also define the following parameters, which in the context of balanced designs, are usually referred to as *main effects*. Thus

$$\begin{aligned} A \text{ main effects: } \alpha_i &= \bar{\mu}_{i..} - \bar{\mu}_{...}, \\ B \text{ main effects: } \beta_j &= \bar{\mu}_{.j.} - \bar{\mu}_{...}, \quad \text{and} \\ C \text{ main effects: } \gamma_k &= \bar{\mu}_{..k} - \bar{\mu}_{...}. \end{aligned}$$

We stated in Section 8.3 that in the unbalanced design it is not appropriate to put constraints on the parameters. Also, there were then three methods of testing the hypothesis sequence. However, when the design is balanced, we saw in Section 8.4 that there is an orthogonal structure so that the three methods are the same and the various sums of squares can be obtained by applying the symmetric constraints. The same thing applies to all higher-way balanced designs, as we shall see in the next section.

8.6.2 Hypothesis Testing

With the definitions above, and defining $\mu = \bar{\mu}_{...}$, we have the reparametrization

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk}, \quad (8.37)$$

where

$$\begin{aligned} \alpha_+ = \beta_+ &= \gamma_+ = 0, \\ (\alpha\beta)_i = (\alpha\beta)_{.j} &= (\beta\gamma)_{.k} = (\alpha\gamma)_{i..} = (\alpha\gamma)_{..k} = 0, \end{aligned}$$

and

$$(\alpha\beta\gamma)_{ij.} = (\alpha\beta\gamma)_{.jk} = (\alpha\beta\gamma)_{i..k} = 0, \quad (8.38)$$

these conditions holding for all values of the subscripts i, j , and k .

The appropriate order for hypothesis testing is as follows: second-order interactions zero [$H_{ABC} : (\alpha\beta\gamma)_{ijk} = 0$, all i, j, k]; first-order interactions zero [$H_{AB} : (\alpha\beta)_{ij} = 0$, all i, j ; $H_{BC} : (\beta\gamma)_{jk} = 0$, all j, k ; $H_{AC} : (\alpha\gamma)_{ik} = 0$, all i, k]; and main effects zero ($H_A : \alpha_i = 0$, all i ; $H_B : \beta_j = 0$, all j ; $H_C : \gamma_k = 0$, all k). When H_{ABC} is true, the three-factor experiment becomes equivalent to three independent two-factor experiments, one for each pair of factors, and the first-order interactions are readily interpreted. For example, (8.36) is now the same for all k , so it is equal to the average over k [which

is $(\alpha\beta)_{ij}$. Similarly, when H_{AB} is also true, the three-factor experiment becomes equivalent to two independent one-factor experiments for A and B , respectively, and the main effects α_i and β_j have a simple interpretation (e.g., $\alpha_i = \bar{\mu}_{i..} - \bar{\mu}_{...} = \bar{\mu}_{ij.} - \bar{\mu}_{.j.} = \mu_{ijk} - \bar{\mu}_{.jk}$). As in the two-way classification, the general regression theory can be applied here. For example, writing

$$\mathbf{Y}' = (Y_{1111}, Y_{1112}, \dots, Y_{IJKM}),$$

(8.35) can be expressed in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where \mathbf{X} is $n \times p$ of rank p , $n = IJKM$, and $p = IJK$. Minimizing $\sum_{ijkm} (Y_{ijkm} - \mu_{ijk})^2$ with respect to μ_{ijk} , we obtain $\hat{\mu}_{ijk} = \bar{Y}_{...}$ and

$$\text{RSS} = \sum_{ijkm} (Y_{ikjm} - \bar{Y}_{ijk.})^2, \quad (8.39)$$

with $(n - p)$ degrees of freedom. To find RSS_H for each hypothesis, we split up ε_{ijkm} in a manner suggested by (8.37), namely,

$$\begin{aligned} \varepsilon_{ijkm} &= \bar{\varepsilon}_{...} + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{.j..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{..k..} - \bar{\varepsilon}_{...}) \\ &\quad + (\bar{\varepsilon}_{ij..} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j..} + \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{i.k..} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{..k..} + \bar{\varepsilon}_{...}) \\ &\quad + (\bar{\varepsilon}_{.jk..} - \bar{\varepsilon}_{.j..} - \bar{\varepsilon}_{..k..} + \bar{\varepsilon}_{...}) \\ &\quad + (\bar{\varepsilon}_{ijk.} - \bar{\varepsilon}_{ij..} - \bar{\varepsilon}_{.jk..} - \bar{\varepsilon}_{i.k..} + \bar{\varepsilon}_{i..} + \bar{\varepsilon}_{.j..} + \bar{\varepsilon}_{..k..} - \bar{\varepsilon}_{...}) \\ &\quad + (\varepsilon_{ijkm} - \bar{\varepsilon}_{ijk.}). \end{aligned}$$

Squaring and summing on i, j, k , and m , we find that the cross-product terms vanish, so that

$$\sum_{ijkm} \varepsilon_{ijkm}^2 = \sum_{ijkm} \bar{\varepsilon}_{...}^2 + \dots + \sum_{ijkm} (\varepsilon_{ijkm} - \bar{\varepsilon}_{ijk.})^2.$$

Setting $\varepsilon_{ijkm} = Y_{ijkm} - \mu_{ijk}$, and using equations (8.37) and (8.38), we find that

$$\begin{aligned} \sum_{ijkm} (Y_{ijkm} - \mu - \alpha_i - \dots - (\alpha\beta\gamma)_{ijk})^2 \\ &= \sum_{ijkm} (\bar{Y}_{...} - \mu)^2 + \sum_{ijkm} (\bar{Y}_{i..} - \bar{Y}_{...} - \alpha_i)^2 \\ &\quad + \dots + \sum_{ijkm} (Y_{ijkm} - \bar{Y}_{ijk.})^2. \end{aligned} \quad (8.40)$$

By inspection, the left side of (8.40) is minimized when the unknown parameters take the values

$$\begin{aligned} \hat{\mu} &= \bar{Y}_{...}, \\ \hat{\alpha}_i &= \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j = \bar{Y}_{.j..} - \bar{Y}_{...}, \quad \hat{\gamma}_k = \bar{Y}_{..k..} - \bar{Y}_{...}, \\ \widehat{(\alpha\beta)}_{ij} &= \bar{Y}_{ij..} - \bar{Y}_{i..} - \bar{Y}_{.j..} + \bar{Y}_{...}, \text{etc.,} \\ \widehat{(\alpha\beta\gamma)}_{ijk} &= \bar{Y}_{ijk.} - \bar{Y}_{ij..} - \bar{Y}_{.jk..} - \bar{Y}_{i.k..} + \bar{Y}_{i..} + \bar{Y}_{.j..} + \bar{Y}_{..k..} - \bar{Y}_{...}, \end{aligned}$$

and the minimum value is, of course, RSS of (8.39). Testing any particular hypothesis is now very straightforward. For example, if we wish to test H_A , we set $\alpha_i = 0$ in (8.40) and minimize with respect to the other parameters. We see, by inspection, that the minimum occurs at the same values of the remaining parameters, so that

$$\text{RSS}_{H_A} = \sum_{ijkm} (\bar{Y}_{i...} - \bar{Y}_{....})^2 + \sum_{ijkm} (\bar{Y}_{ijkm} - \bar{Y}_{ijk.})^2.$$

Hence

$$\begin{aligned} \text{RSS}_{H_A} - \text{RSS} &= \sum_{ijkm} (\bar{Y}_{i...} - \bar{Y}_{....})^2 \\ &= JKM \sum_i \hat{\alpha}_i^2, \end{aligned}$$

with $(I - 1)$ degrees of freedom, and the appropriate F -statistic is

$$F = \frac{JKM \sum_i \hat{\alpha}_i^2 / (I - 1)}{\text{RSS}/[IJK(M - 1)]} = \frac{S_A^2}{S^2},$$

say. This statistic has an F -distribution with $I - 1$ and $IJK(M - 1)$ degrees of freedom when H_A is true. The various quadratic forms, together with their degrees of freedom, are listed in Table 8.4. The numbers of degrees of freedom can be obtained from the ranks of the underlying design (regression) matrices, from the numbers of free parameters in each case, or from the trace of the appropriate quadratic form. For example,

$$\begin{aligned} &\sum_{ijkm} (\bar{Y}_{ij..} - \bar{Y}_{i...} - \bar{Y}_{.j..} + \bar{Y}_{....})^2 \\ &= \sum_i \sum_j \frac{(Y_{ij..})^2}{KM} - \sum_i \frac{(Y_{i...})^2}{JKM} - \sum_j \frac{(Y_{.j..})^2}{IKM} + \frac{(Y_{....})^2}{IJKM}, \end{aligned}$$

and the trace of the symmetric matrix underlying this quadratic is the sum of the coefficients of the terms Y_{ijkm}^2 , namely,

$$\begin{aligned} &\frac{1}{KM} \sum_i \sum_j KM - \frac{1}{JKM} \sum_i JKM - \frac{1}{IKM} \sum_j IKM \\ &\quad + \frac{1}{IJKM} \cdot IJKM \\ &= IJ - I - J + 1 \\ &= (I - 1)(J - 1). \end{aligned}$$

It should be noted that if the test of a particular higher-order interaction is significant, then we need to include lower-order interactions and related main effects in the model, thus maintaining a hierarchical structure.

Table 8.4 Analysis-of-variance table for a three-way classification with M observations per population mean

Source	Sum of squares (SS)	Degrees of freedom (df)	$\frac{SS}{df}$
A main effects	$JKM \sum_i \hat{\alpha}_i^2$	$I - 1$	S_A^2
B main effects	$IKM \sum_j \hat{\beta}_j^2$	$J - 1$	S_B^2
C main effects	$IJM \sum_k \hat{\gamma}_k^2$	$K - 1$	S_C^2
AB interactions	$KM \sum_i \sum_j (\widehat{\alpha\beta})_{ij}^2$	$(I - 1)(J - 1)$	S_{AB}^2
BC interactions	$IM \sum_j \sum_k (\widehat{\beta\gamma})_{jk}^2$	$(J - 1)(K - 1)$	S_{BC}^2
AC interactions	$JM \sum_i \sum_k (\widehat{\alpha\gamma})_{ik}^2$	$(I - 1)(K - 1)$	S_{AC}^2
ABC interactions	$M \sum_i \sum_j \sum_k (\widehat{\alpha\beta\gamma})_{ijk}^2$	$(I - 1)(J - 1)(K - 1)$	S_{ABC}^2
Error	$\sum_i \sum_j \sum_k \sum_m (Y_{ijkl} - \bar{Y}_{ijkl.})^2$	$IJKM - IJK$	S^2
Corrected total	$\sum_i \sum_j \sum_k \sum_m (Y_{ijkl} - \bar{Y}_{....})^2$	$IJKM - 1$	

8.6.3 Missing Observations

We saw in Example 8.1 that an experimenter may set out to use a balanced design, but one or more of the observations may get destroyed. Since unbalanced designs are problematical in their analysis, a sensible approach might be to use a balanced design with some missing observations. The idea would be to put in suitable estimates of the missing values, which have the property that if a balanced analysis is carried out using the completed data, the least squares estimates of the unknown parameters and the residual sum of squares are correct for the original unbalanced design. This will mean that any residuals corresponding to the estimated values are zero (see Exercises 8d, No. 3). For example, if there was just one missing value, we could guess it, find its residual, and then adjust the value to make the residual zero, giving us a second iteration. We note that care is needed with standard errors, as

those derived from the completed data will not be correct. For references and further details, see Jarrett [1978] and Hunt and Triggs [1989].

EXERCISES 8d

- Verify (8.28).
- A three-factor experiment has population means μ_{ijk} ($i = 1, 2, 3; j = 1, 2, 3; k = 1, 2$), given by the following tables:

C_1	B_1	B_2	B_3	Mean	C_2	B_1	B_2	B_3	Mean
A_1	5	6	10	7	A_1	9	7	14	10
A_2	7	7	1	5	A_2	9	6	3	6
A_3	6	5	7	6	A_3	9	5	10	8
Mean	6	6	6	6	Mean	9	6	9	8

Show that the ABC interactions are zero.

- Let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where \mathbf{X} has full rank. Suppose that the observations \mathbf{Y}_2 are missing. For the data observed, the least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{Y}_1$. Let $\hat{\mathbf{Y}}_2 = \mathbf{X}_2 \hat{\boldsymbol{\beta}}$. Show that $\hat{\boldsymbol{\beta}}$ can be obtained by applying least squares to the observed data augmented by $\hat{\mathbf{Y}}_2$ and the regression matrix \mathbf{X} .

- Suppose that we have a one-way classification

$$E[Y_{ij}] = \mu_i \quad (i = 1, 2, \dots, I; j = 1, 2, \dots, J),$$

and suppose that Y_{IJ} is missing. Prove that the appropriate estimate of Y_{IJ} is the mean of the remaining observations on the mean μ_I .

8.7 DESIGNS WITH SIMPLE BLOCK STRUCTURE

In addition to the cross-classification designs considered above, there are also the *hierarchical* or *nested designs*. For example, suppose that we have I cities, J factories within each city, and a sample of size K is taken from each factory, giving the model $Y_{ijk} = \theta_{ijk} + \varepsilon_{ijk}$ ($i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K$). Then the appropriate reparametrization of the model is

$$\theta_{ijk} = \bar{\theta}_{...} + (\bar{\theta}_{i..} - \bar{\theta}_{...}) + (\bar{\theta}_{ij.} - \bar{\theta}_{i..}) + (\bar{\theta}_{ijk} - \bar{\theta}_{ij.}) \quad (8.41)$$

or, since $\theta_{ijk} = \mu_{ij}$ ($k = 1, 2, \dots, K$),

$$\mu_{ij} = \mu + \alpha_i + \beta_{ij} \quad (8.42)$$

with identifiability constraints $\alpha_i = 0$ and $\beta_{i..} = 0$ (all i). The hypotheses of interest are $H_1 : \beta_{ij} = 0$ (no variation within each city) and $H_2 : \alpha_i = 0$ (no variation between cities), and the appropriate decomposition of ε_{ijk} is

$$\varepsilon_{ijk} = \bar{\varepsilon}_{...} + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{ij.} - \bar{\varepsilon}_{i..}) + (\bar{\varepsilon}_{ijk} - \bar{\varepsilon}_{ij.}). \quad (8.43)$$

Once again we have an orthogonal decomposition of ε , and F -statistics are readily obtained for testing H_1 and H_2 ; the details are left as an exercise (Miscellaneous Exercises 8, No. 3, at the end of this chapter).

Many of the designs currently used are a mixture of both crossing and nesting. When every nesting classification used has equal numbers of subunits nested in each unit, then the experimental units are said to have a *simple block structure* and there is an elegant theory for handling such designs (due to Nelder [1965a,b]).

8.8 ANALYSIS OF COVARIANCE

In Section 8.1 we referred briefly to analysis-of-covariance (ANCOVA) models and we now wish to focus on these models, which combine both qualitative and quantitative variables in one regression matrix. Such a situation can arise in an experiment where a particular “factor” may be involved either quantitatively or qualitatively.

EXAMPLE 8.3 Suppose that we wish to compare the effects of three different drugs on people by measuring some response Y . If Y_{ij} is the response from the j th patient taking the i th drug, then a one-way analysis of variance (one factor at three levels) can be carried out using the model $E[Y_{ij}] = \mu_i$ ($i = 1, 2, 3; j = 1, 2, \dots, J$), or $E[Y] = \mathbf{X}\boldsymbol{\beta}$. However, it transpires that the effect of a drug may depend on the age of the patient, so that one model might be

$$E[Y_{ij}] = \mu_i + \gamma_{i1} z_{ij} + \gamma_{i2} z_{ij}^2,$$

where z_{ij} is the age of the j th patient taking drug i . This model can be expressed in the form

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},$$

where

$$\mathbf{Z}\boldsymbol{\gamma} = \begin{pmatrix} z_{11} & 0 & 0 & z_{11}^2 & 0 & 0 \\ z_{12} & 0 & 0 & z_{12}^2 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_{1J} & 0 & 0 & z_{1J}^2 & 0 & 0 \\ 0 & z_{21} & 0 & 0 & z_{21}^2 & 0 \\ 0 & z_{22} & 0 & 0 & z_{22}^2 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & z_{2J} & 0 & 0 & z_{2J}^2 & 0 \\ 0 & 0 & z_{31} & 0 & 0 & z_{31}^2 \\ 0 & 0 & z_{32} & 0 & 0 & z_{32}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & z_{3J} & 0 & 0 & z_{3J}^2 \end{pmatrix} \begin{pmatrix} \gamma_{11} \\ \gamma_{21} \\ \gamma_{31} \\ \gamma_{12} \\ \gamma_{22} \\ \gamma_{32} \end{pmatrix}.$$

If there is no interaction between age and type of drug, that is, the effect of age is the same for each drug, then the model can be simplified to

$$E[Y_{ij}] = \mu_i + \gamma_1 z_{ij} + \gamma_2 z_{ij}^2$$

or

$$\mathbf{Z}\boldsymbol{\gamma} = \begin{pmatrix} z_{11} & z_{11}^2 \\ z_{12} & z_{12}^2 \\ \dots & \dots \\ z_{3J} & z_{3J}^2 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

In addition to age, there may be a body weight effect which also does not interact with drug type. A suitable model might now be

$$E[Y_{ij}] = \mu_i + \gamma_1 z_{ij} + \gamma_2 z_{ij}^2 + \gamma_3 w_{ij},$$

where w_{ij} is the weight of the j th patient taking the i th drug; if the drugs change the weight, then w_{ij} could refer to the initial body weight. The three quantities *age*, $(age)^2$, and *weight* are commonly called *concomitant variables*, and frequently they are random variables rather than variables controlled by the experimenter. Random explanatory variables are discussed in Chapter 9. However, if the variables are measured accurately it transpires that we can treat the variables as (conditionally) fixed. \square

If, in Example 8.3, age and weight are likely to have a considerable effect on the drug action and we are particularly interested in this effect, then it might be more appropriate to design a three-way layout with three factors treated qualitatively: drug, age, and weight. This model would be more robust than the ANCOVA model.

A general analysis-of-covariance model takes the form

$$G : E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} = (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} = \mathbf{W}\boldsymbol{\delta}, \quad (8.44)$$

say, where \mathbf{X} is $n \times p$, \mathbf{Z} is $n \times t$ of rank t and the columns of \mathbf{Z} are linearly independent of the columns of \mathbf{X} . Thus G can be analyzed as one large regression model and hypothesis tests carried out using the general theory of Chapter 4. We emphasize the important assumption that the variables in \mathbf{Z} should not be affected by the “treatments” in \mathbf{X} . For example, as we pointed out in Example 8.3, if a particular drug causes a weight change, then w_{ij} should refer to the *initial* weight, which is, of course, unaffected by the drug.

EXAMPLE 8.4 We have already considered an ANCOVA model in Section 6.4. It is instructive to look again at Example 6.2 using a different notation.

Let

$$G : Y_{ij} = \mu_i + \gamma_i z_{ij} + \varepsilon_{ij} \quad (i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J),$$

where the ε_{ij} are independently and identically distributed as $N(0, \sigma^2)$, be I regression lines with J observations per line. Suppose we wish to test that the lines are parallel, namely, $H : \gamma_1 = \gamma_2 = \dots = \gamma_I$ ($= \gamma$, say). Then we can find RSS for the model G and RSS_H for model H , where

$$H : Y_{ij} = \mu_i + \gamma z_{ij}$$

and construct an F -test accordingly. \square

EXERCISES 8e

1. Exercises 3f, No. 2, in Section 3.7 gave the following two-stage method for finding least squares estimates for a general ANCOVA model G , say. First, we find the least squares estimate $\hat{\beta}$ of β for the model $E[\mathbf{Y}] = \mathbf{X}\beta$ and the residual sum of squares $\mathbf{Y}'\mathbf{R}\mathbf{Y}$. Second, $\hat{\gamma}_G$ is found by replacing \mathbf{Y} by $\mathbf{Y} - \mathbf{Z}\gamma$ in $\mathbf{Y}'\mathbf{R}\mathbf{Y}$ and minimizing with respect to γ . This minimum value is the correct residual sum of squares for the model G . Third, referring to Theorem 3.6, we see that $\hat{\beta}_G$ is obtained from $\hat{\beta}$ by replacing \mathbf{Y} by $\mathbf{Y} - \mathbf{X}\hat{\gamma}$. This technique can be applied to the model for H as well as G . Now apply this method to the following problem. Consider the model $Y_{ij} = \mu_i + \gamma_i x_j + \varepsilon_{ij}$, where $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; and the ε_{ij} are independently distributed as $N(0, \sigma^2)$. Derive an F -statistic for testing the hypothesis that $\gamma_1 = \gamma_2$ and show that this statistic is the square of the usual t -statistic for testing whether two lines are parallel.
2. Let $Y_{ij} = \mu_i + \gamma_1 z_{ij} + \gamma_2 w_{ij} + \varepsilon_{ij}$, where $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; and the ε_{ij} are independently distributed as $N(0, \sigma^2)$.
 - (a) Derive the least squares estimate of γ_1 and show that it is an unbiased estimate of γ_1 .
 - (b) Find the variance matrix of the least squares estimates $\hat{\gamma}_i$ of γ_i ($i = 1, 2$).

- (c) Under what conditions are $\hat{\gamma}_1$ and $\hat{\gamma}_2$ statistically independent?
3. Let $Y_{ijk} = \mu_{ij} + \gamma_{ij}z_{ijk} + \varepsilon_{ijk}$, where $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; $k = 1, 2, \dots, K$; and the ε_{ijk} are independently distributed as $N(0, \sigma^2)$. Obtain a test statistic for testing the hypothesis

$$H : \gamma_{ij} = \gamma \quad (\text{all } i, j).$$

MISCELLANEOUS EXERCISES 8

1. If the ε_{ij} ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$) are independently distributed as $N(0, \sigma^2)$, prove that

$$\sum_i \sum_j (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{..})^2 \quad \text{and} \quad \sum_i \sum_j (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot j} + \bar{\varepsilon}_{..})^2$$

are statistically independent.

2. Let $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$), where $\sum_i d_i \alpha_i = 0$ ($\sum_i d_i \neq 0$) and $E[\varepsilon_{ij}] = 0$ for all i, j . Using the method of Lagrange multipliers, find the least squares estimates of μ and α_i . Hint: Show that the Lagrange multiplier is zero.
3. Let $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$, where

$$\begin{aligned} \mu_{ij} &= \bar{\mu}_{..} + (\bar{\mu}_{i\cdot} - \bar{\mu}_{..}) + (\bar{\mu}_{\cdot j} - \bar{\mu}_{..}) \\ &= \mu + \alpha_i + \beta_{ij}, \end{aligned}$$

say, $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; $k = 1, 2, \dots, K$, and the ε_{ijk} are independently distributed as $N(0, \sigma^2)$.

- (a) Find the least squares estimates of μ , α_i , and β_{ij} , and show that they are statistically independent.
- (b) Obtain test statistics for testing the hypotheses $H_1 : \beta_{ij} = 0$ (all i, j) and $H_2 : \alpha_i = 0$ (all i).
4. Let $Y_{ij} = \mu_i + \varepsilon_{ij}$ ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$), where the ε_{ij} are independently distributed as $N(0, \sigma^2)$.
- (a) When $I = 4$, obtain an F -statistic for testing the hypothesis that $\mu_1 = 2\mu_2 = 3\mu_3$.
- (b) When $I = 2$, show that the F -statistic for testing $\mu_1 = \mu_2$ is the square of the usual t -test for testing the hypothesis that the means of two normally distributed populations are equal, given that their variances are equal.

5. Suppose that we have the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk},$$

where $i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K$, $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$, and the ε_{ijk} are independently distributed as $N(0, \sigma^2)$.

- (a) Express μ, α_i, β_j , and γ_k in terms of the parameters $\mu_{ijk} = E[Y_{ijk}]$.
- (b) Obtain a test statistic for testing the hypothesis $H : \alpha_i = 0$ (all i).
- (c) Prove that

$$\sum_i \sum_j \sum_k (\bar{\varepsilon}_{ij.} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j.} + \bar{\varepsilon}_{...})^2 / \sigma^2 \sim \chi^2_{(I-1)(J-1)}.$$

Hint: Split up $\sum_i \sum_j \sum_k (\bar{\varepsilon}_{ij.} - \bar{\varepsilon}_{i..})^2$ into two sums of squares.

6. Consider the linear model $Y_{ijk} = \mu_{ijk} + \varepsilon_{ijk}$, where $i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K$; and the ε_{ijk} are independently distributed as $N(0, \sigma^2)$. Let

$$\begin{aligned} \mu_{ijk} &= \bar{\mu}_{...} + (\bar{\mu}_{i..} - \bar{\mu}_{...}) + (\bar{\mu}_{.j.} - \bar{\mu}_{i..}) + (\bar{\mu}_{..k} - \bar{\mu}_{...}) + \Delta_{ijk} \\ &= \mu + \alpha_i + \beta_{ij} + \gamma_k, \end{aligned}$$

say, where $\Delta_{ijk} = 0$ (all i, j, k).

- (a) Find the least squares estimates of $\mu, \alpha_i, \beta_{ij}$, and γ_k .
- (b) Obtain an F -statistic for testing $H : \alpha_i = 0$ (all i).

9

Departures from Underlying Assumptions

9.1 INTRODUCTION

The basic multiple regression model that we have been studying thus far is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where \mathbf{X} is $n \times p$ of rank p . We assume that the elements of $\boldsymbol{\varepsilon}$

- (1) are unbiased;
- (2) have constant variance;
- (3) are uncorrelated, and
- (4) are normally distributed.

Assumption (1) implies that $E[\boldsymbol{\varepsilon}] = \mathbf{0}$, which implies that \mathbf{X} is the correct design matrix (i.e., $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$). Assumptions (2) and (3) imply that $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$, and (3) and (4) imply the independence of the ε_i . It is also assumed, implicitly, that the regressor variables x_j are not random variables but are predetermined constants. If the explanatory variables are random and are measured without error, then the regression can be regarded as being conditional on the observed values of the explanatory variables, a problem discussed in Section 9.6.1. In this chapter we examine each of the foregoing assumptions in detail.

It should be noted that there is a tendency for errors that occur in many real situations to be normally distributed owing to the central limit theorem. If $\boldsymbol{\varepsilon}$ is a sum of n errors from different sources, then, as n increases, $\boldsymbol{\varepsilon}$ tends to normality irrespective of the probability distributions of the individual errors.

This argument applies to small errors δ_i , say, in a nonlinear system since

$$\varepsilon = f(\delta_1 + a_1, \dots, \delta_n + a_n) - f(a_1, \dots, a_n) \approx \delta_1 \frac{\partial f}{\partial a_1} + \dots + \delta_n \frac{\partial f}{\partial a_n},$$

and ε is once again a (weighted) sum of errors.

In the next section we examine the effect of misspecifying the design matrix. If \mathbf{X} is “underfitted”, we will find that $\hat{\beta}$ is biased, S^2 is an overestimate of σ^2 , but $\text{Var}[\hat{\beta}]$ is correct. The residual vector is also biased, but its variance-covariance matrix is not. If \mathbf{X} is “overfitted”, then $\hat{\beta}$ is essentially unbiased, S^2 is unbiased, but $\text{Var}[\hat{\beta}]$ is inflated. Also, the residual vector is unbiased, but its variance-covariance matrix is inflated. In both cases there are similar problems with fitted values and predictions.

9.2 BIAS

9.2.1 Bias Due to Underfitting

If $E[\varepsilon] = 0$, then $E[\mathbf{Y}] = \mathbf{X}\beta$ and the least squares estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is an unbiased estimate of β . However, if the model is underfitted, so that the true model is actually

$$E[\mathbf{Y}] = \mathbf{X}\beta + \mathbf{Z}\gamma, \quad (9.1)$$

where the columns of the $n \times t$ matrix \mathbf{Z} are linearly independent of the columns of \mathbf{X} , then ε is biased and

$$\begin{aligned} E[\hat{\beta}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{Z}\gamma) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma \\ &= \beta + \mathbf{L}\gamma, \end{aligned} \quad (9.2)$$

say. Thus $\hat{\beta}$ is now a biased estimate of β with bias $\mathbf{L}\gamma$. This bias term depends on both the postulated and the true models, and \mathbf{L} can be interpreted as the matrix of regression coefficients of the omitted variables regressed on the x -variables actually included in the model. A good choice of design may keep the bias to a minimum even if the wrong model has been postulated and fitted. For example, if the columns of \mathbf{Z} are orthogonal to the columns of \mathbf{X} , then $\mathbf{X}'\mathbf{Z} = \mathbf{0}$, $\mathbf{L} = \mathbf{0}$, and $\hat{\beta}$ is unbiased. Under some circumstances the orthogonality of the columns of \mathbf{X} and $\mathbf{0}$ may be described as zero correlation between a pair of x and z explanatory variables (Malinvaud [1970]). In this case, inadvertently omitting an uncorrelated regressor may not be serious.

EXAMPLE 9.1 Suppose that we postulate the model $E[Y] = \beta_0 + \beta_1 x$ when the true model is $E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$. If we use observations of Y at $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$ to estimate β_0 and β_1 in the model postulated, then we can find the biases as follows.

The true model is

$$\begin{aligned} E \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} &= \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \beta_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\beta_2. \end{aligned}$$

Now

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \\ \mathbf{X}'\mathbf{Z} &= \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \end{aligned}$$

and from equation (9.2), the bias of $\hat{\boldsymbol{\beta}}$ is

$$\mathbf{L}\boldsymbol{\gamma} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix} \beta_2 = \begin{pmatrix} \frac{2}{3} \\ 0 \end{pmatrix} \beta_2.$$

Thus $\hat{\beta}_0$ has bias $\frac{2}{3}\beta_2$, and $\hat{\beta}_1$ is unbiased. \square

If equation (9.1) represents the true model, then provided that $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$, we still have $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. However, given $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with trace p , and $S^2 = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}/(n-p)$, then (Theorem 1.5 in Section 1.5)

$$E[S^2] = \sigma^2 + \frac{\boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}\boldsymbol{\gamma}}{n-p} > \sigma^2.$$

This follows from the fact that $(\mathbf{I}_n - \mathbf{P})$ is idempotent and therefore positive-semidefinite, and when $\boldsymbol{\gamma} \neq 0$, we have $(\mathbf{I}_n - \mathbf{P})\mathbf{Z}\boldsymbol{\gamma} \neq 0$ [as $\mathbf{Z}\boldsymbol{\gamma} \notin \mathcal{C}(\mathbf{X})$ implies that its projection perpendicular to $\mathcal{C}(\mathbf{X})$ is not zero]. Hence S^2 is an overestimate of σ^2 .

To examine the effect of underfitting on the fitted model, we note that $\hat{\mathbf{Y}} = \mathbf{PY}$ and

$$\begin{aligned} E[\hat{\mathbf{Y}}] &= \mathbf{PE}[\mathbf{Y}] \\ &= \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{PZ}\boldsymbol{\gamma}. \end{aligned} \tag{9.3}$$

The effect, therefore, of ignoring $\mathbf{Z}\boldsymbol{\gamma}$ in the regression amounts to using \mathbf{PZ} instead of \mathbf{Z} .

As far as the residuals are concerned, we see that

$$\begin{aligned} E[\mathbf{e}] &= E[\mathbf{Y}] - E[\hat{\mathbf{Y}}] \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{XL}\boldsymbol{\gamma}) \quad [\text{by (9.2)}] \\ &= (\mathbf{I}_n - \mathbf{P})\mathbf{Z}\boldsymbol{\gamma} \end{aligned} \tag{9.4}$$

and

$$\text{Var}[\mathbf{e}] = \text{Var}[(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] = \sigma^2(\mathbf{I}_n - \mathbf{P})^2 = \sigma^2(\mathbf{I}_n - \mathbf{P}).$$

The effect of misspecification is to bias \mathbf{e} but leave $\text{Var}[\mathbf{e}]$ unchanged. Ramsey [1969] makes use of this fact to provide several tests for this kind of misspecification.

We now examine the effect of the underfitting on prediction. Suppose that we wish to predict the value of Y at $\mathbf{w}_0 = (\mathbf{x}'_0, \mathbf{z}'_0)'$, although only \mathbf{x}_0 is observed. Then the prediction using \mathbf{x}_0 only is $\hat{Y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$, whereas the correct prediction is $\hat{Y}_{0G} = \mathbf{w}'_0 \hat{\boldsymbol{\delta}}$, where (from Section 3.7)

$$\hat{\boldsymbol{\delta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}$$

is the least squares estimate of $\boldsymbol{\delta}$ for the model $\mathbf{W}\boldsymbol{\delta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$. Now, from (9.2),

$$E[\hat{Y}_0] = \mathbf{x}'_0 \boldsymbol{\beta} + \mathbf{x}'_0 \mathbf{L}\boldsymbol{\gamma},$$

which may be compared to the “true” expectation

$$E[\hat{Y}_{0G}] = \mathbf{w}'_0 \hat{\boldsymbol{\delta}} = \mathbf{x}'_0 \boldsymbol{\beta} + \mathbf{z}'_0 \boldsymbol{\gamma}.$$

Also, from equation (3.25),

$$\begin{aligned} \text{var}[\hat{Y}_{0G}] &= (\mathbf{x}'_0, \mathbf{z}'_0) \text{Var}[\hat{\boldsymbol{\delta}}_{\mathbf{G}}](\mathbf{x}'_0, \mathbf{z}'_0)' \\ &= \sigma^2(\mathbf{x}'_0, \mathbf{z}'_0)(\mathbf{W}'\mathbf{W})^{-1}(\mathbf{x}'_0, \mathbf{z}'_0)' \\ &= \sigma^2(\mathbf{x}'_0, \mathbf{z}'_0) \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}' & -\mathbf{L}\mathbf{M} \\ -\mathbf{M}\mathbf{L}' & \mathbf{M} \end{pmatrix} (\mathbf{x}'_0, \mathbf{z}'_0)' \\ &= \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + \sigma^2 (\mathbf{L}'\mathbf{x}_0 - \mathbf{z}_0)' \mathbf{M} (\mathbf{L}'\mathbf{x}_0 - \mathbf{z}_0) \\ &\geq \text{var}[\hat{Y}_0], \end{aligned} \tag{9.5}$$

since $\mathbf{M} = [\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}]^{-1}$ is positive-definite because, by (3.25), it is the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}_G$. Hence the “apparent” prediction variance $\text{var}[\hat{Y}_0]$ will tend to be smaller than the “true” variance.

9.2.2 Bias Due to Overfitting

Suppose that the true model is $E[\mathbf{Y}] = \mathbf{X}_1 \boldsymbol{\beta}_1$, where \mathbf{X}_1 consists of the first k columns of \mathbf{X} ; thus $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, say. Then

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X}_1 \boldsymbol{\beta}_1 \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}, \end{aligned} \tag{9.6}$$

and $\hat{\beta}_1$, consisting of the first k elements of $\hat{\beta}$, is an unbiased estimate of β_1 . Also,

$$E[\hat{Y}] = E[X\hat{\beta}] = X \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} = X_1\beta_1, \quad (9.7)$$

so that the fitted model is an unbiased estimate of the true model. However, as we shall now show, the familiar formula $\sigma^2(X'X)^{-1}$ leads to inflated expressions for the variances of the elements of $\hat{\beta}_1$. From equation (3.25) in Section 3.7, with X and Z set equal to X_1 and X_2 , respectively, we have

$$(X'X)^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} + LML', & -LML \\ -ML', & M \end{pmatrix},$$

where M , and therefore LML' , is positive-definite (A.4.5). Hence, by A.4.8,

$$\begin{aligned} \text{"apparent" } \text{var}[\hat{\beta}_i] &= \text{"true" } \text{var}[\hat{\beta}_i] + (LML')_{ii} \\ &> \text{"true" } \text{var}[\hat{\beta}_i], \end{aligned}$$

where $\hat{\beta}_i$ is an element of $\hat{\beta}_1$. Since $(I_n - P)(X_1, X_2) = 0$ [Theorem 3.1(iii)], we note that

$$E[Y'(I_n - P)Y] = (n - p)\sigma^2 + \beta_1'X_1'(I_n - P)X_1\beta_1 = (n - p)\sigma^2,$$

and S^2 is still an unbiased estimate of σ^2 .

Additional effects are explored in the following exercises.

EXERCISES 9a

- Suppose in Example 9.1 that the roles of the postulated model and the true model are exchanged. Find the biases of the least squares estimates.
- In Section 9.2.2 the observed residual is $Y - X\hat{\beta}$. Find its true mean and variance-covariance matrix.
- In Section 9.2.2 suppose that $x'_0 = (x'_{10}, x'_{20})$, where x_0 is $p \times 1$ and x_{10} is $k \times 1$. The prediction at x_0 is $\hat{Y}_0 = x'_0\hat{\beta}$, whereas the correct prediction is $\hat{Y}_{10} = x'_{10}\hat{\beta}_1$. Compare \hat{Y}_0 and \hat{Y}_{10} with respect to their means and variances.

9.3 INCORRECT VARIANCE MATRIX

If we assume that $\text{Var}[\epsilon] = \sigma^2 I_n$ when in fact $\text{Var}[\epsilon] = \sigma^2 V$, then $\hat{\beta}$ is still an unbiased estimate of β . However,

$$\text{Var}[\hat{\beta}] = \text{Var}[(X'X)^{-1}X'Y] = \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}$$

is, in general, not equal to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Then (cf. Theorem 1.5 in Section 1.5)

$$\begin{aligned} E[S^2] &= \frac{\sigma^2}{n-p} E[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \\ &= \frac{\sigma^2}{n-p} \text{tr}[\mathbf{V}(\mathbf{I}_n - \mathbf{P})], \end{aligned} \quad (9.8)$$

and S^2 is generally a biased estimate of σ^2 . It follows that

$$\hat{v} = S^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{a} \quad (9.9)$$

will normally be a biased estimate of

$$\text{var}[\mathbf{a}'\hat{\beta}] = \sigma^2 \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{a}. \quad (9.10)$$

In fact, Swindel [1968] has shown that if

$$E[\hat{v}] = \text{var}[\mathbf{a}'\hat{\beta}] + b,$$

then

$$\begin{aligned} &\frac{\{\text{mean of } (n-p) \text{ least eigenvalues of } \mathbf{V}\} - (\text{greatest eigenvalue of } \mathbf{V})}{(\text{greatest eigenvalue of } \mathbf{V})} \\ &\leq \frac{b}{\text{var}[\mathbf{a}'\hat{\beta}]} \\ &\leq \frac{\{\text{mean of } (n-p) \text{ greatest eigenvalues of } \mathbf{V}\} - (\text{least eigenvalue of } \mathbf{V})}{(\text{least eigenvalue of } \mathbf{V})} \end{aligned}$$

and the bounds are attainable.

EXERCISES 9b

1. If the first column of \mathbf{X} is $\mathbf{1}_n$ and

$$\mathbf{V} = (1 - \rho)\mathbf{I}_n + \rho \mathbf{1}_n \mathbf{1}_n' \quad (0 \leq \rho < 1),$$

use (9.8) to show that

$$E[S^2] = \sigma^2(1 - \rho).$$

Hint: Consider $\mathbf{P}\mathbf{1}_n$.

2. When $\text{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{V}$, the appropriate estimate of $\boldsymbol{\beta}$ is the generalized least squares estimate $\boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. If $\mathcal{C}(\mathbf{V}^{-1}\mathbf{X}) = \mathcal{C}(\mathbf{X})$, show that $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$ are identical. *Hint:* $\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}\mathbf{W}$, where \mathbf{W} is nonsingular.

(McElroy [1967])

9.4 EFFECT OF OUTLIERS

In fitting the regression model $Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$ ($i = 1, \dots, n$), where \mathbf{x}_i is the i th row of the design matrix \mathbf{X} , we can think of the data point $(\mathbf{x}'_i, Y_i)'$ as a point in p -dimensional space. Some of these points may arouse suspicions, as they are “discordant” with the other points. Such points are usually referred to vaguely as *outliers*, and they may or may not have an effect on estimation and inference using the prescribed regression model. It is generally accepted that several percent of even (supposedly) high-quality data can be erroneous through such things as wrong measurements, wrong decimal points, and wrong copying, for example. Some fraction of the erroneous data may be sufficiently different from the other data to warrant the label *outlier*. Also, if ε_i comes from a long-tailed distribution rather than our postulated normal distribution, then a Y_i that is more extreme than usual can arise and become a candidate for outlier status.

We shall be interested in two kinds of points: those points whose error ε_i is large, and those points whose \mathbf{x}_i value is far from the bulk of the data. There is no standard terminology for these points; the former is variously labeled *outlier*, *error outlier*, *outlier in the y -direction*, and *regression outlier*, and the latter variously called an *extreme point*, *outlier in the x -direction*, *x -outlier*, *leverage point* and *high-leverage point*. We shall use the terms *outlier* and *high-leverage point*, respectively. To demonstrate these terms, suppose that our (true) underlying model is a straight line, as in Figure 9.1, with data points scattered around it.

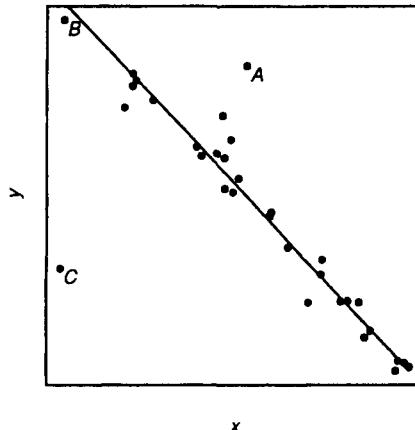


Fig. 9.1 Outliers and high-leverage points.

Here the points A , B , and C , are of particular interest, with A and C being outliers (i.e., a big vertical displacement from the true line), and B and C being high-leverage points (since their x -values are distant from the average x -value). We now consider what would happen if just one of these points were present and the other three absent. If just A were present, we would expect A to have a modest effect on the least squares fit of the line to the data, moving it up. The point B would have a negligible effect on the fit, but C , which is both an outlier and a high-leverage point, will have a considerable effect, much greater than that of A . (See Section 10.6 for more discussion.) Points like C are often referred to as *influential points*, since they exert a big influence on the position of the fitted line.

We note that, in general, the fitted model takes the form $\hat{\mathbf{Y}} = \mathbf{PY}$, or

$$\hat{Y}_i = \sum_j p_{ij} Y_j = p_{ii} Y_i + \sum_{j \neq i} p_{ij} Y_j, \quad (9.11)$$

where $\mathbf{P} = (p_{ij})$. Since $\mathbf{P} = \mathbf{P}^2$, we have

$$p_{ii} = \sum_{j=1}^n p_{ij}^2 = p_{ii}^2 + \sum_{j \neq i} p_{ij}^2, \quad (9.12)$$

which implies that $p_{ii} \geq p_{ii}^2$, or $p_{ii} \leq 1$. Furthermore, from (3.53), the non-negative distance MD_i implies that $p_{ii} \geq 1/n$, so that

$$\frac{1}{n} \leq p_{ii} \leq 1. \quad (9.13)$$

We note from (9.12) that when p_{ii} is close to 1, p_{ij} ($j \neq i$) will be close to zero, and from (9.11) this means that \hat{Y}_i will be determined largely by the value of Y_i . Thus if Y_i is both an outlier and a high-leverage point, MD_i will be large, p_{ii} will be close to 1, and \hat{Y}_i will be affected by Y_i . This supports our contention that a point like C above can have a serious affect on the fitted line.

Suppose that the i th point is an outlier, so that $Y_i = \mathbf{x}'_i \beta + \Delta_i + \varepsilon_i$, where Δ_i is a positive constant. Let Δ be the vector with i th element Δ_i and the remaining elements be zero. Then

$$\mathbf{Y} = \mathbf{X}\beta + \Delta + \varepsilon,$$

and since $\mathbf{PX} = \mathbf{P}$,

$$\begin{aligned} E[\hat{\mathbf{Y}}] &= \mathbf{PE}[\mathbf{Y}] \\ &= \mathbf{P}(\mathbf{X}\beta + \Delta) \\ &= \mathbf{X}\beta + \mathbf{P}\Delta, \end{aligned}$$

which leads to

$$E[\hat{Y}_i] = \mathbf{x}'_i \beta + p_{ii} \Delta_i. \quad (9.14)$$

Once again this emphasizes that the effect of Δ_i on the fitted surface depends on the magnitude of p_{ii} .

9.5 ROBUSTNESS OF THE F -TEST TO NONNORMALITY

9.5.1 Effect of the Regressor Variables

Box and Watson [1962] showed that the sensitivity of the F -test to normality depends very much on the numerical values of the regression variables. In terms of the experimental design situation in which the elements of the design matrix \mathbf{X} are 0 or 1, this means that some designs will have more robust tests associated with them. For example, Box and Watson show, by an appropriate choice of \mathbf{X} , that almost the same regression model can be made to reproduce, on the one hand, a test to compare means which is little affected by nonnormality, and on the other, a comparison of variances test which is notoriously sensitive to nonnormality. Let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{i,p-1} + \varepsilon_i$, and consider $H: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. When H is true and the regression assumptions are valid, then

$$F = \frac{n-p}{p-1} \cdot \frac{\text{RSS}_H - \text{RSS}}{\text{RSS}} \sim F_{k,n-p}.$$

However, if we now relax the distributional assumptions and assume that the ε_i are independently distributed with some common—not necessarily normal—distribution, then Box and Watson [1962: p. 101] show that when H is true, F is approximately distributed as F_{ν_1, ν_2} , with $\nu_1 = \delta(p-1)$, $\nu_2 = \delta(n-p)$, and

$$\delta^{-1} = 1 + \frac{(n+1)\alpha_2}{n-1-2\alpha_2},$$

where

$$\alpha_2 = \frac{n-3}{2n(n-1)} \cdot C_X \Gamma_\gamma;$$

or (to order n^{-1})

$$\delta^{-1} = 1 + \frac{C_X \Gamma_\gamma}{2n}. \quad (9.15)$$

Here $\Gamma_\gamma = E[k_4/k_2^2]$, where k_2 and k_4 are the sample cumulants for the n values of Y , and C_X is a multivariate analog of k_4/k_2^2 for the x variables. When ε , and therefore Y , has a normal distribution, then $\Gamma_\gamma = 0$, $\delta = 1$, and $F_{\nu_1, \nu_2} = F_{k,n-p}$. We see that the effect of any nonnormality in Y depends on C_X in the term δ . Box and Watson show that

$$-2 \leq \frac{n-3}{n-1} C_X \leq n-1, \quad (9.16)$$

where the lower bound is obtainable but the upper bound, although approached, cannot be attained in finite samples. When the explanatory variables can be regarded as being approximately “normal,” then $C_X \approx 0$ and the F -test is insensitive to nonnormality. Thus we may sum up by saying

that it is the extent of nonnormality in the explanatory variables which determines the sensitivity of F to nonnormality in the Y observations. Let $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$ ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, p - 1$) and let $\tilde{\mathbf{X}} = (\tilde{x}_{ij})$. If $\mathbf{M} = (m_{rs}) = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'$ and $m = \sum_{r=1}^n m_{rr}^2$, Box and Watson show that

$$C_X = \frac{n(n^2 - 1)}{(p-1)(n-p)(n-3)} \left\{ m - \frac{(p-1)^2}{n} - \frac{2(p-1)(n-p)}{n(n+1)} \right\}. \quad (9.17)$$

Now, applying Theorem 3.1(ii) to the $n \times n$ matrix \mathbf{M} yields $\text{tr}(\mathbf{M}) = p - 1$. If the diagonal elements of \mathbf{M} are all equal, we have $m_{rr} = (p-1)/n$ ($r = 1, 2, \dots, n$), $m = (p-1)^2/n$, and

$$\begin{aligned} C_X &= \frac{n(n^2 - 1)}{(p-1)(n-p)(n-3)} \left\{ -\frac{2(p-1)(n-p)}{n(n+1)} \right\} \\ &= -\frac{2(n-1)}{n-3}. \end{aligned}$$

Hence, in this case, the lower bound of (9.16) is attained, $\delta^{-1} = 1 - (\Gamma_\gamma/n) \approx 1$, and for large n the F -test is insensitive to nonnormality. From symmetry conditions it is not hard to show that many analysis-of-variance models, such as any cross-classification with equal cell frequencies in every cell or any hierarchical classification with equal cell frequencies at each stage of the hierarchy, have equal elements m_{rr} .

This theory refers only to the case $H: \beta_1 = \dots = \beta_{p-1} = 0$. However, an alternative approach, which allows a more general hypothesis, has been given by Atiqullah [1962]. We now consider his method in detail.

9.5.2 Quadratically Balanced F -Tests

Let Y_1, Y_2, \dots, Y_n be independent random variables with means $\theta_1, \theta_2, \dots, \theta_n$, respectively, common variance σ^2 , and common third and fourth moments about their means; let $\gamma_2 = (\mu_4 - 3\sigma^4)/\sigma^4$ be their common kurtosis. Then from Atiqullah [1962] we have the following theorems.

THEOREM 9.1 *Let \mathbf{P}_i ($i = 1, 2$) be a symmetric idempotent matrix of rank f_i such that $E[\mathbf{Y}'\mathbf{P}_i\mathbf{Y}] = \sigma^2 f_i$, and let $\mathbf{P}_1\mathbf{P}_2 = 0$. If \mathbf{p}_i is the column vector of the diagonal elements of \mathbf{P}_i then:*

$$(i) \quad \text{var}[\mathbf{Y}'\mathbf{P}_i\mathbf{Y}] = 2\sigma^4(f_i + \frac{1}{2}\gamma_2\mathbf{p}'_i\mathbf{p}_i).$$

$$(ii) \quad \text{cov}[\mathbf{Y}'\mathbf{P}_1\mathbf{Y}, \mathbf{Y}'\mathbf{P}_2\mathbf{Y}] = \sigma^4\gamma_2\mathbf{p}'_1\mathbf{p}_2.$$

Proof. (i) Since \mathbf{P}_i is symmetric and idempotent, $\text{tr}(\mathbf{P}_i) = \text{rank}(\mathbf{P}_i) = f_i$ (A.6.2). Also, $E[\mathbf{Y}'\mathbf{P}_i\mathbf{Y}] = \sigma^2 \text{tr}(\mathbf{P}_i) + \boldsymbol{\theta}'\mathbf{P}_i\boldsymbol{\theta} = \sigma^2 f_i$ (Theorem 1.5, Section 1.5), so that $\boldsymbol{\theta}'\mathbf{P}_i^2\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{P}_i\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta}$; that is, $\mathbf{P}_i\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta}$. Therefore,

substituting $\mathbf{A} = \mathbf{P}_i$ in Theorem 1.6, we have

$$\begin{aligned}\text{var}[\mathbf{Y}'\mathbf{P}_i\mathbf{Y}] &= 2\sigma^4 \text{tr}(\mathbf{P}_i^2) + (\mu_4 - 3\sigma^4)\mathbf{P}_i'\mathbf{P}_i \\ &= 2\sigma^4[\text{tr}(\mathbf{P}_i) + \frac{1}{2}\gamma_2\mathbf{P}_i'\mathbf{P}_i] \\ &= 2\sigma^4(f_i + \frac{1}{2}\gamma_2\mathbf{P}_i'\mathbf{P}_i).\end{aligned}$$

(ii) Given $\mathbf{P}_1\mathbf{P}_2 = \mathbf{0}$, we have

$$\begin{aligned}(\mathbf{P}_1 + \mathbf{P}_2)^2 &= \mathbf{P}_1^2 + \mathbf{P}_1\mathbf{P}_2 + \mathbf{P}_2\mathbf{P}_1 + \mathbf{P}_2^2 \\ &= \mathbf{P}_1 + \mathbf{P}_1\mathbf{P}_2 + (\mathbf{P}_1\mathbf{P}_2)' + \mathbf{P}_2 \\ &= \mathbf{P}_1 + \mathbf{P}_2.\end{aligned}$$

Therefore, $\mathbf{P}_1 + \mathbf{P}_2$ is idempotent and, by (i),

$$\begin{aligned}\text{var}[\mathbf{Y}'\mathbf{P}_1\mathbf{Y} + \mathbf{Y}'\mathbf{P}_2\mathbf{Y}] &= \text{var}[\mathbf{Y}'(\mathbf{P}_1 + \mathbf{P}_2)\mathbf{Y}] \\ &= 2\sigma^4[\text{tr}(\mathbf{P}_1 + \mathbf{P}_2) + \frac{1}{2}\gamma_2(\mathbf{P}_1 + \mathbf{P}_2)'(\mathbf{P}_1 + \mathbf{P}_2)] \\ &= 2\sigma^4[f_1 + f_2 + \frac{1}{2}\gamma_2(\mathbf{P}_1'\mathbf{P}_1 + 2\mathbf{P}_1'\mathbf{P}_2 + \mathbf{P}_2'\mathbf{P}_2)] \\ &= \text{var}[\mathbf{Y}'\mathbf{P}_1\mathbf{Y}] + \text{var}[\mathbf{Y}'\mathbf{P}_2\mathbf{Y}] + 2\sigma^4\gamma_2\mathbf{P}_1'\mathbf{P}_2.\end{aligned}$$

Hence $\text{cov}[\mathbf{Y}'\mathbf{P}_1\mathbf{Y}, \mathbf{Y}'\mathbf{P}_2\mathbf{Y}] = \sigma^4\gamma_2\mathbf{P}_1'\mathbf{P}_2$. \square

THEOREM 9.2 Suppose that \mathbf{P}_1 and \mathbf{P}_2 satisfy the conditions of Theorem 9.1, and let $Z = \frac{1}{2}\log F$, where

$$F = \frac{\mathbf{Y}'\mathbf{P}_1\mathbf{Y}/f_1}{\mathbf{Y}'\mathbf{P}_2\mathbf{Y}/f_2} \quad \left(= \frac{S_1^2}{S_2^2}, \text{ say} \right).$$

Then for large f_1 and f_2 we have, asymptotically,

$$\begin{aligned}E[Z] &\sim \frac{1}{2}(f_2^{-1} - f_1^{-1}) \\ &\times [1 + \frac{1}{2}\gamma_2(f_1\mathbf{P}_2 - f_2\mathbf{P}_1)'(f_1\mathbf{P}_2 + f_2\mathbf{P}_1)\{f_1f_2(f_1 + f_2)\}^{-1}] \quad (9.18)\end{aligned}$$

and

$$\text{var}[Z] \sim \frac{1}{2}(f_1^{-1} + f_2^{-1})[1 + \frac{1}{2}\gamma_2(f_1\mathbf{P}_2 - f_2\mathbf{P}_1)'(f_1\mathbf{P}_2 - f_2\mathbf{P}_1)\{f_1f_2(f_1 + f_2)\}^{-1}]. \quad (9.19)$$

Proof. Using a Taylor expansion of $\log S_i^2$ about $\log \sigma^2$, we have

$$\log S_i^2 \sim \log \sigma^2 + \frac{S_i^2 - \sigma^2}{\sigma^2} - \frac{(S_i^2 - \sigma^2)^2}{2\sigma^4}. \quad (9.20)$$

Taking expected values, and using $E[S_i^2] = \sigma^2$, we have

$$E[\log S_i^2] \sim \log \sigma^2 - \frac{1}{2\sigma^4} \text{var}[S_i^2],$$

where, from Theorem 9.1,

$$\text{var}[S_i^2] = \frac{\text{var}[\mathbf{Y}'\mathbf{P}_i\mathbf{Y}]}{f_i^2} = 2\sigma^4(f_i^{-1} + \frac{1}{2}\gamma_2 f_i^{-2}\mathbf{p}_i'\mathbf{p}_i).$$

Substituting in

$$E[Z] = \frac{1}{2} \{ E[\log S_i^2] - E[\log S_2^2] \}$$

leads to equation (9.18).

To find an asymptotic expression for $\text{var}[Z]$, we note first that

$$\text{var}[Z] = \frac{1}{4} \{ \text{var}[\log S_1^2] + \text{var}[\log S_2^2] - 2\text{cov}[\log S_1^2, \log S_2^2] \}. \quad (9.21)$$

Then, ignoring the third term in (9.20), we have $E[\log S_i^2] \sim \log \sigma^2$ and

$$\begin{aligned} \text{var}[\log S_i^2] &\sim E[(\log S_i^2 - \log \sigma^2)^2] \\ &\sim \frac{E[(S_i^2 - \sigma^2)^2]}{\sigma^4} \\ &= \frac{\text{var}[S_i^2]}{\sigma^4}. \end{aligned}$$

Similarly,

$$\begin{aligned} \text{cov}[\log S_1^2, \log S_2^2] &\sim E[(\log S_1^2 - \log \sigma^2)(\log S_2^2 - \log \sigma^2)] \\ &\sim \frac{E[(S_1^2 - \sigma^2)(S_2^2 - \sigma^2)]}{\sigma^4} \\ &= \frac{\text{cov}[S_1^2, S_2^2]}{\sigma^4}. \end{aligned}$$

Finally, substituting in

$$\text{var}[Z] \sim \frac{1}{4\sigma^4} \{ \text{var}[S_1^2] + \text{var}[S_2^2] - 2\text{cov}[S_1^2, S_2^2] \}$$

and using Theorem 9.1 leads to equation (9.19). \square

We can now apply the theory above to the usual F -statistic for testing $H: \mathbf{A}\beta = \mathbf{0}$. From Theorem 4.1(iv) in Section 4.3, we have

$$\begin{aligned} F &= \frac{\mathbf{Y}'(\mathbf{P} - \mathbf{P}_H)\mathbf{Y}/q}{\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}/(n-p)} \\ &= \frac{\mathbf{Y}'\mathbf{P}_1\mathbf{Y}/q}{\mathbf{Y}'\mathbf{P}_2\mathbf{Y}/(n-p)} \\ &= \frac{S_1^2}{S_2^2}, \end{aligned} \quad (9.22)$$

say, where $\mathbf{P}_1\mathbf{P}_2 = (\mathbf{P} - \mathbf{P}_H)(\mathbf{I}_n - \mathbf{P}) = \mathbf{P}_H - \mathbf{P}_H\mathbf{P} = \mathbf{0}$. Suppose that we now relax the distributional assumptions underlying F and assume only that

the ε_i are independently and identically distributed; in particular, $E[\varepsilon] = \mathbf{0}$ and $\text{Var}[\varepsilon] = \sigma^2 \mathbf{I}_n$. Then $E[S_1^2] = E[S^2] = \sigma^2$ (Theorem 3.3 in Section 3.3) and, when H is true, $E[S_1^2] = \sigma^2$ [by Theorem 4.1(ii) with $\mathbf{A}\beta = \mathbf{c} = \mathbf{0}$; the assumption of normality is not used in the proof]. Also, the Y_i satisfy the conditions stated at the beginning of this section [with $(\theta_i) = \boldsymbol{\theta} = \mathbf{X}\beta$], so that when H is true, Theorem 9.2 can be applied directly to the F -statistic (9.22) with $f_1 = q$ and $f_2 = n - p$. When the ε_i , and therefore the Y_i , are normally distributed, it is known that for large f_1 and f_2 , $Z = \frac{1}{2} \log F$ is approximately normally distributed with mean and variance given by setting $\gamma_2 = 0$ in equations (9.18) and (9.19) when H is true. As this approximation is evidently quite good even when f_1 and f_2 are as small as four, it is not unreasonable to accept Atiqullah's proposition that for a moderate amount of nonnormality, Z is still approximately normal with mean and variance given by (9.18) and (9.19). On this assumption Z , and therefore F , will be approximately independent of γ_2 if the coefficient of γ_2 in (9.18) and (9.19) is zero; that is, if

$$f_1 \mathbf{p}_2 = f_2 \mathbf{p}_1. \quad (9.23)$$

Now, using Atiqullah's terminology, we say that F is quadratically balanced if the diagonal elements of P_i ($i = 1, 2$) are equal; most of the usual F -tests for balanced experimental designs belong to this category. In this case, since $\text{tr}(\mathbf{P}_i) = f_i$, we have

$$\mathbf{p}_i = \frac{f_i}{n} \mathbf{1}_n \quad \text{and} \quad f_1 \mathbf{p}_2 = \frac{f_1 f_2}{n} \mathbf{1}_n = f_2 \mathbf{p}_1.$$

Thus a sufficient condition for (9.23) to hold is that F is quadratically balanced.

Atiqullah [1962: p. 88] also states that even if γ_2 varies among the Y_i , quadratic balance is still sufficient for $E[Z]$ and $\text{var}[Z]$ to be independent of kurtosis effects, to the order involved in Theorem 9.2. Finally, we note that if γ_2 can be estimated, equations (9.18) and (9.19) can be used to modify the degrees of freedom and improve the correspondence between the distribution of the F -ratio and an F -distribution (Prentice [1974]).

EXERCISES 9c

1. Fill in the details of the proof of Theorem 9.2.
2. Show that the theory of Section 9.5.2 can also be applied to the case $H: \mathbf{A}\beta = \mathbf{c}$, where $\mathbf{c} \neq \mathbf{0}$ [cf. (9.22)].
3. Consider the full-rank regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where the ε_i are independently and identically distributed as $N(0, \sigma^2)$, and suppose that we wish to test $H: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$.

Assuming that H is true, find an approximate expression for $E[Z]$, where $Z = \frac{1}{2} \log F$, in terms of the diagonal elements of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

9.6 EFFECT OF RANDOM EXPLANATORY VARIABLES

We shall consider four different scenarios: random explanatory variables measured without error, fixed explanatory variables measured with error, random explanatory variables measured with error, and controlled explanatory variables (commonly called *Berkson's model*).

9.6.1 Random Explanatory Variables Measured without Error

Suppose that we have a lawlike relationship

$$V = \beta_0 + \beta_1 U_1 + \cdots + \beta_{p-1} U_{p-1} \quad (9.24)$$

between the random variables V and $\{U_j\}$. This relationship is typically called a *structural relationship* with the U_j being observed exactly but V unknown (due, for example, to experimental error), so that $Y (= V + \varepsilon)$ is actually observed. The appropriate model is now

$$Y = \beta_0 + \beta_1 U_1 + \cdots + \beta_{p-1} U_{p-1} + \varepsilon \quad (9.25)$$

or

$$E[Y | \{U_j\}] = \beta_0 + \beta_1 U_1 + \cdots + \beta_{p-1} U_{p-1} \quad (= \mu, \text{say}). \quad (9.26)$$

The simplest and most popular method of fitting this model is to carry out a standard regression analysis *conditionally* on the values observed for the explanatory variables; we simply proceed as though the explanatory variables were fixed. Such an approach now requires the usual assumptions of normality, constant variance, and independence to hold conditionally on the U_j 's. The problems of bias, etc. due to model misspecification, as discussed above, will be the same as for fixed explanatory variables.

A different approach to the problem is the following. Suppose that the true model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 U_1 + \cdots + \beta_s U_s \\ &= \beta_0 + \beta_1 U_1 + \cdots + \beta_r U_r + \delta, \quad (r < s). \end{aligned}$$

where the U_j ($j = 1, 2, \dots, s$) are random variables with $E[U_j] = \theta_j$. Here the "error" δ is assumed to be due to further "hidden" variables U_{r+1}, \dots, U_s . Now

$$\begin{aligned} Y &= (\beta_0 + \beta_{r+1} \theta_{r+1} + \cdots + \beta_s \theta_s) + \beta_1 U_1 + \cdots + \beta_r U_r + \sum_{j=r+1}^s \beta_j (U_j - \theta_j) \\ &= \alpha_0 + \beta_1 U_1 + \cdots + \beta_r U_r + \varepsilon, \end{aligned}$$

where $E[\varepsilon] = 0$. This is the same model as (9.25). However, since r is arbitrary, we will always have $E[\varepsilon] = 0$ irrespective of the number of regressor variables we include in the model. In this case there is no question of finding the “true” model. What we are looking for is an adequate model, that is, one which reduces ε to a reasonable level. For this type of model it can be argued, therefore, that the question of bias due to overfitting or underfitting does not arise.

9.6.2 Fixed Explanatory Variables Measured with Error

Least Squares Estimation

Suppose that the relationship (9.24) is now between the expected values rather than the random variables, that is,

$$\begin{aligned} v &= E[Y] = \beta_0 + \beta_1 E[X_1] + \cdots + \beta_{p-1} E[X_{p-1}] \\ &= \beta_0 + \beta_1 u_1 + \cdots + \beta_{p-1} u_{p-1}. \end{aligned} \quad (9.27)$$

This lawlike relationship between the expected values is usually called a *functional relationship*. Fuller [1987: p. 2] gives the helpful mnemonic “F” for fixed and functional and “S” for stochastic and structural. Here v and the u_j ’s are unknown, and are all measured with error. Sometimes the relationship comes from a physical law (perhaps suitably transformed to achieve linearity), with the randomness in the model arising from experimental errors in measuring the mathematical variables v and u_j . For this reason the model is sometimes called the *errors-in-variables* model or, in the straight-line case, the model for *regression with both variables subject to error*. Here the appropriate model is now

$$Y = \beta_0 + \beta_1 E[X_1] + \cdots + \beta_{p-1} E[X_{p-1}] + \varepsilon, \quad (9.28)$$

where ε is assumed to be independent of the $\{X_j\}$. If we have n measurements on the model above, then

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 u_{i1} + \cdots + \beta_{p-1} u_{ip-1} + \varepsilon_i \\ &= \mathbf{u}'_i \boldsymbol{\beta} + \varepsilon_i, \end{aligned}$$

say, or

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\beta} + \varepsilon, \quad (9.29)$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)'$. Suppose that the data point \mathbf{u}_i is measured with an unbiased error of δ_i so that we actually observe $\mathbf{x}_i = \mathbf{u}_i + \delta_i$, that is, observe $\mathbf{X} = \mathbf{U} + \Delta$, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, $\Delta = (\delta_1, \delta_2, \dots, \delta_n)'$, and $E[\Delta] = 0$. It is assumed that the δ_i are uncorrelated and have the same variance matrix; thus

$$E[\delta_i \delta'_j] = \begin{cases} \mathbf{D}, & i = j, \\ 0, & i \neq j. \end{cases}$$

Since the first element of each \mathbf{u}_i and \mathbf{X}_i is unity, the first element of $\boldsymbol{\delta}_i$ is zero, and the first row and column of \mathbf{D} consists of zeros. We also assume that Δ is independent of ϵ . The usual least squares estimate of β is now

$$\hat{\beta}_\Delta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

instead of $\hat{\beta} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y}$, so that $\hat{\beta}_\Delta$ is no longer unbiased. The properties of $\hat{\beta}_\Delta$ were discussed in detail by Hodges and Moore [1972] for the common special case of $\mathbf{D} = \text{diag}(0, \sigma_1^2, \sigma_2^2, \dots, \sigma_{p-1}^2)$. However, using a more rigorous approximation theory, Davies and Hutton [1975] extended this work to the case of a general matrix \mathbf{D} (in their notation, $\mathbf{U} \rightarrow \mathbf{X}'$, $\mathbf{X} \rightarrow \mathbf{W}'$, $\mathbf{D} \rightarrow \mathbf{S}$, and $\Delta \rightarrow \Delta'$). We now consider their results below.

Bias

Since Δ is independent of ϵ (and \mathbf{Y}),

$$\begin{aligned} E[\hat{\beta}_\Delta] &= E_\Delta E[\hat{\beta}_\Delta | \Delta] \\ &= E_\Delta[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\beta] \quad [\text{from (9.29)}] \\ &= E_\Delta[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X} - \Delta)\beta] \\ &= \beta - E_\Delta[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Delta\beta] \\ &= \beta - \mathbf{b}, \end{aligned} \tag{9.30}$$

say. When n is large, Davies and Hutton [1975: Theorem 4.1] show that

$$\begin{aligned} \mathbf{b} &\approx \left(\frac{1}{n} \mathbf{U}'\mathbf{U} + \mathbf{D} \right)^{-1} \mathbf{D}\beta \\ &= n(\mathbf{U}'\mathbf{U} + n\mathbf{D})^{-1}\mathbf{D}\beta. \end{aligned} \tag{9.31}$$

[In fact, if $\lim_{n \rightarrow \infty} \{(1/n)\mathbf{X}'\mathbf{X}\} = \mathbf{A}$, say, then $\hat{\beta}_\Delta$ is a consistent estimate of $(\mathbf{A} + \mathbf{D})^{-1}\mathbf{A}\beta = \beta - (\mathbf{A} + \mathbf{D})^{-1}\mathbf{D}\beta$.] Since

$$\begin{aligned} E[\mathbf{X}'\mathbf{X}] &= E[\mathbf{U}'\mathbf{U} + \Delta'\mathbf{U} + \mathbf{U}'\Delta + \Delta'\Delta] \\ &= \mathbf{U}'\mathbf{U} + E[\Delta'\Delta] \\ &= \mathbf{U}'\mathbf{U} + E \left[\sum_{i=1}^n \delta_i \delta_i' \right] \\ &= \mathbf{U}'\mathbf{U} + n\mathbf{D}, \end{aligned} \tag{9.32}$$

an obvious estimate of the bias \mathbf{b} is

$$\hat{\mathbf{b}} = n(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{D}}\hat{\beta},$$

where $\hat{\mathbf{D}}$ is a rough estimate of \mathbf{D} available, we hope, from other experiments. When $\mathbf{D} = \text{diag}(0, \sigma_1^2, \dots, \sigma_{p-1}^2)$, the approximations used by Hodges and Moore lead to a similar estimate of \mathbf{b} (with $n - p - 1$ instead of n). Davies

and Hutton show that the magnitude of \mathbf{b} is related to how close $\mathbf{X}'\mathbf{X}$ is to being singular. If the errors are such that they may bring $\mathbf{X}'\mathbf{X}$ close to being singular, then the bias could be large. Using the central limit theorem they also show that $\sqrt{n}\hat{\beta}_\Delta$ is asymptotically normal.

EXAMPLE 9.2 We now apply the theory above to the straight-line model. Here we have

$$Y_i = \beta_0 + \beta_1 u_i + \varepsilon_i \quad \text{and} \quad X_i = u_i + \delta_i.$$

The pairs $(\delta_i, \varepsilon_i)$ are generally assumed to be a random sample from a bivariate normal distribution. If we assume further that δ_i and ε_i are independently distributed with zero means and respective unknown variances σ_δ^2 and σ_ε^2 , we have n pairs (X_i, Y_i) of data but $n+4$ unknowns $\beta_0, \beta_1, \sigma_\delta^2, \sigma_\varepsilon^2$, and u_1, \dots, u_n . Applying (9.31) with $\mathbf{D} = \text{diag}(0, \sigma_\delta^2)$ yields

$$E[\hat{\beta}_{0\Delta}] \approx \beta_0 + \frac{\bar{u}\beta_1 n\sigma_\delta^2}{\sum_i(u_i - \bar{u})^2 + n\sigma_\delta^2}$$

and

$$E[\hat{\beta}_{1\Delta}] \approx \beta_1 - \frac{\beta_1 n\sigma_\delta^2}{\sum_i(u_i - \bar{u})^2 + n\sigma_\delta^2}.$$

We note that

$$\begin{aligned} \sum_{i=1}^n(X_i - \bar{X})^2 &= \sum_i[u_i + \delta_i - (\bar{u} + \bar{\delta})]^2 \\ &= \sum_i(u_i - \bar{u})^2 - 2\sum_i(u_i - \bar{u})(\delta_i - \bar{\delta}) + \sum_i(\delta_i - \bar{\delta})^2, \end{aligned}$$

so that

$$E\left[\sum_i(X_i - \bar{X})^2\right] = \sum_i(u_i - \bar{u})^2 + (n-1)\sigma_\delta^2.$$

We then see that the relative bias in $\hat{\beta}_{1\Delta}$ is approximately

$$-\frac{n\sigma_\delta^2}{E\left[\sum_i(X_i - \bar{X})^2\right]},$$

which will generally be small if $\sum_i(X_i - \bar{X})^2/n >> \sigma_\delta^2$. This will be the case if the variation among the X_i is much greater than the error in a single X_i ; a not unexpected result. When this is not the case, we find that certain maximum likelihood estimates are no longer consistent (Sprent [1969: Chapter 3]; see also Moran [1970, 1971]). The inconsistency is related to the fact that the number of unknowns u_i increases with n . \square

Expressions for the exact values of $E[\hat{\beta}_{1\Delta}]$ and $E[(\hat{\beta}_{1\Delta} - \beta_1)^2]$, along with more accurate large sample approximations, are given by Richardson and Wu

[1970]. Their results are generalized by Halperin and Gurian [1971] to the case where δ_i and ε_i are correlated. Alternative methods of estimation are available under certain restrictions (e.g., when the ratio $\sigma_e^2/\sigma_\delta^2$ is known; see Fuller [1987] for details).

Standard Errors

Davies and Hutton [1975: equation 4.2] show that when \mathbf{D} is close to the zero matrix,

$$\text{Var}[\hat{\beta}_\Delta] \approx \frac{1}{n} \left\{ \left(\frac{1}{n} \mathbf{U}' \mathbf{U} + \mathbf{D} \right)^{-1} (\sigma^2 + \beta' \Delta \beta) + O(\mathbf{D}^2) \right\}.$$

The usual estimate of this variance-covariance matrix is $\hat{\mathbf{V}} = S^2(\mathbf{X}'\mathbf{X})^{-1}$, where

$$\begin{aligned} (n-p)S^2 &= (\mathbf{Y} - \mathbf{X}\hat{\beta}_\Delta)'(\mathbf{Y} - \mathbf{X}\hat{\beta}_\Delta) \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_X)\mathbf{Y}, \end{aligned}$$

say. The question now is: Does $\hat{\mathbf{V}}$ still provide an unbiased estimate of $\text{Var}[\hat{\beta}_\Delta]$? Since Δ is independent of ε , $\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_X) = \mathbf{0}$, and $\text{tr}(\mathbf{I}_n - \mathbf{P}_X) = n - p$, we have (Theorem 1.5)

$$\begin{aligned} E[(n-p)S^2 | \Delta] &= E[(n-p)\sigma^2 + \beta' \mathbf{U}'(\mathbf{I}_n - \mathbf{P}_X)\mathbf{U}\beta | \Delta] \\ &= E[(n-p)\sigma^2 + \beta'(\mathbf{X}' - \Delta')(\mathbf{I}_n - \mathbf{P}_X)(\mathbf{X} - \Delta)\mathbf{B} | \Delta] \\ &= E[(n-p)\sigma^2 + \beta'\Delta'(\mathbf{I}_n - \mathbf{P}_X)\Delta\beta | \Delta]. \quad (9.33) \end{aligned}$$

Now for any matrix \mathbf{C} ,

$$\begin{aligned} E_\Delta[\Delta' \mathbf{C} \Delta] &= \sum_i \sum_j c_{ij} E[\delta_i \delta'_j] \\ &= \sum_i c_{ii} \mathbf{D} \\ &= \mathbf{D} \text{tr}(\mathbf{C}), \end{aligned}$$

so that from (9.33) it transpires that

$$\begin{aligned} E[\hat{\mathbf{V}}] &= E_\Delta E[S^2(\mathbf{X}'\mathbf{X})^{-1} | \Delta] \\ &= E_\Delta \left[\left\{ \sigma^2 + \frac{1}{n-p} \beta' \Delta' (\mathbf{I}_n - \mathbf{P}_X) \Delta \beta \right\} \{(\mathbf{X}'\mathbf{X})^{-1}\} \right] \\ &\approx \left\{ \sigma^2 + \frac{1}{n-p} \beta' E[\Delta'(\mathbf{I}_n - \mathbf{P}_X)\Delta]\beta \right\} \{E[\mathbf{X}'\mathbf{X}]\}^{-1} \\ &\approx (\sigma^2 + \beta' \mathbf{D} \beta)(\mathbf{U}' \mathbf{U} + n \mathbf{D})^{-1} \\ &\approx \text{Var}[\hat{\beta}_\Delta]. \quad (9.34) \end{aligned}$$

Hence, for large n and small \mathbf{D} , $\hat{\mathbf{V}}$ is still approximately unbiased.

9.6.3 Round-off Errors

Using the notation above, we suppose, once again, that \mathbf{U} is the correct data matrix (i.e., $E[\mathbf{Y}] = \mathbf{U}\beta$), and we observe $\mathbf{X} = \mathbf{U} + \Delta$. However, following Swindel and Bower [1972], we now assume that the measurements are *accurate* but they are rounded off according to some consistent rule to give $x_{ij} = u_{ij} + \Delta_{ij}$. In this case the rounding error Δ_{ij} can be regarded as an (unknown) constant, *not* a random variable; Δ_{ij} is determined solely by the u_{ij} and the rounding rule. The matrix \mathbf{X} is now a matrix of constants rather than a random matrix as in the preceding section. The bias of $\hat{\beta}_\Delta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is

$$\begin{aligned} E[\hat{\beta}_\Delta - \beta] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U} - \mathbf{X})\beta \\ &= -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Delta\beta. \end{aligned} \quad (9.35)$$

By writing $\Delta\beta = \sum_{j=0}^{p-1} \Delta_j\beta_j$, where the Δ_j are columns of Δ , we see that the bias does not depend on β_j if $\Delta_j = 0$; the bias depends only on the explanatory variables containing rounding errors. We see from

$$\begin{aligned} E[S^2] &= \sigma^2 + \frac{\beta' \mathbf{U}'(\mathbf{I}_n - \mathbf{P}_X)\mathbf{U}\beta}{n-p} \\ &= \sigma^2 + \frac{\beta'(\mathbf{X} - \Delta)'(\mathbf{I}_n - \mathbf{P}_X)(\mathbf{X} - \Delta)\beta}{n-p} \\ &= \sigma^2 + \frac{\beta' \Delta'(\mathbf{I}_n - \mathbf{P}_X)\Delta\beta}{n-p} \\ &\geq \sigma^2 \end{aligned} \quad (9.36)$$

(since $\mathbf{I}_n - \mathbf{P}_X$ is positive-semidefinite), that S^2 will tend to overestimate σ^2 . However, $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is the correct dispersion matrix of $\hat{\beta}_\Delta$. Using eigenvalues, Swindel and Bower [1972] prove that for any \mathbf{a} the estimate $\mathbf{a}'\hat{\beta}_\Delta$ of $\mathbf{a}'\beta$ has the property that

$$0 \leq RB(\mathbf{a}'\hat{\beta}_\Delta) \leq \frac{1}{\sigma}(\beta' \Delta' \Delta \beta)^{1/2},$$

where RB is the relative bias, that is, $|\text{bias}|/\text{standard deviation}$.

9.6.4 Some Working Rules

Davies and Hutton [1975: p. 390] consider both random and round-off error in their analysis and give the following working rules. For the round-off situation, define r_j to be the square root of the j th diagonal element of $\Delta'\Delta/n$ and suppose that m_1 of the r_j 's are nonzero. We first compute

$$\rho_1 = \left\{ \sum_j r_j^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{jj} \right\}^{-1/2}. \quad (9.37)$$

If ρ_1 is not somewhat larger than $(m_1 n)^{1/2}$, or possibly $n^{1/2}$ if n is large, then at least some of the elements of $\hat{\beta}_\Delta$ are likely to have little meaning. (In practice, r_j will not be known exactly and will be replaced by, say, an upper bound.) If this test is passed, then (Davies and Hutton [1975], correction)

$$\frac{n^{1/2} \sum_j r_j |\hat{\beta}_{j,\Delta}|}{S}$$

should be evaluated. If this quantity is markedly less than 1, then the errors in \mathbf{U} can be ignored. However, if this test is failed and the situation of Section 9.6.1 prevails (with random Δ_{ij}), then the next step is to compute

$$\frac{n \left(\sum_j r_j^2 \hat{\beta}_{j,\Delta}^2 \right)^{1/2}}{(\rho_1 S)},$$

where in the formula above and the definition of ρ_1 in (9.37), r_j is now the square root of the j th diagonal element of \mathbf{D} . If the quantity above is markedly less than 1, then the effects of the errors are probably negligible, particularly if n is large. On the other hand, if this term is larger than 1, then the bias is likely to constitute a major part in the error of at least some of the estimates. The authors also suggest that the diagonal elements of \mathbf{P}_X be calculated in order to check whether any single regressor observation has an undue effect on the estimates. In particular, if any diagonal element is greater than about 0.2, it is possible for a moderate error in the corresponding regressor to affect the estimates significantly and yet go undetected when the residuals are checked (Chapter 10).

9.6.5 Random Explanatory Variables Measured with Error

Suppose that we have the structural model (9.25), but the random explanatory variables U_j are now measured with (unbiased) errors so that X_j is observed instead of U_j . Then $X_j = U_j + \gamma_j$, where $E[\gamma_j | U_j] = 0$, and our model becomes

$$\begin{aligned} E[Y|\{U_j\}] &= \beta_0 + \beta_1 U_1 + \cdots + \beta_{p-1} U_{p-1} \\ &= \beta_0 + \beta_1 E[X_1 | U_1] + \cdots + \beta_{p-1} E[X_{p-1} | U_{p-1}]. \end{aligned} \quad (9.38)$$

By treating the $\{U_j\}$ as though they are (conditionally) constant, we see that the model above is analogous to (9.28), so that the discussion in the preceding section can be applied here, conditional on the $\{U_j\}$. Since $E[X_j] = E[U_j]$, we note that the structural model (9.38) can also be written in the form

$$\begin{aligned} Y &= \beta_0 + \beta_1 E[U_1] + \cdots + \beta_{p-1} E[U_{p-1}] + \varepsilon + \sum_{j=1}^{p-1} \beta_j (U_j - E[U_j]) \\ &= \beta_0 + \beta_1 E[U_1] + \cdots + \beta_{p-1} E[U_{p-1}] + \varepsilon', \end{aligned}$$

where $E[\epsilon'] = 0$. This looks like (9.27), but there is a difference: ϵ' is not independent of the $\{U_i\}$.

EXAMPLE 9.3 As in the functional model, we shall demonstrate some of the estimation problems that arise by considering the special case of a straight line. Here $V_i = \beta_0 + \beta_1 U_i$ ($i = 1, \dots, n$) and we observe $Y_i = V_i + \epsilon_i$ and $X_i = U_i + \delta_i$. We now assume that the vectors $(U_i, \epsilon_i, \delta_i)$ are independently and identically distributed as $N_3((\mu_U, 0, 0)', \text{diag}(\sigma_U^2, \sigma_\epsilon^2, \sigma_\delta^2))$. If we use the usual least squares estimate $\hat{\beta}_1 = \sum_i Y_i(X_i - \bar{X}) / \sum(X_i - \bar{X})^2$, which ignores the fact that we should use the true values $\{U_i\}$ instead of the observed values $\{X_i\}$, we have from Richardson and Wu [1970: p. 732]

$$E[\hat{\beta}_1] = \beta_1 \frac{(\sigma_U^2 / \sigma_\delta^2)}{1 + (\sigma_U^2 / \sigma_\delta^2)}$$

and

$$\text{var}[\hat{\beta}_1] = \frac{1}{n-2} \left[\frac{\sigma_\epsilon^2}{\sigma_U^2 + \sigma_\delta^2} + \beta_1^2 \frac{\sigma_U^2 \sigma_\delta^2}{(\sigma_U^2 + \sigma_\delta^2)^2} \right].$$

We see that $\hat{\beta}_1$ is biased toward zero, and one way of describing this is to say that the regression coefficient has been *attenuated* by the measurement error. If $\text{var}[X] = \sigma_X^2 = \sigma_U^2 + \sigma_\delta^2$, then the coefficient of β_1 in the first equation is σ_U^2 / σ_X^2 . This ratio measures, in some sense, the reliability of X_i .

We note from Miscellaneous Exercises 9, No. 4, at the end of the chapter, that $(X_i, Y_i)'$ has a bivariate normal distribution with mean $(\mu_U, \beta_0 + \beta_1 \mu_U)'$ and variance-covariance matrix

$$\begin{pmatrix} \sigma_U^2 + \sigma_\delta^2 & \beta_1 \sigma_U^2 \\ \beta_1 \sigma_U^2 & \beta_1^2 \sigma_U^2 + \sigma_\epsilon^2 \end{pmatrix}.$$

There are six unknown parameters, μ_U , σ_U^2 , β_0 , β_1 , σ_δ^2 , and σ_ϵ , but they cannot all be estimated, as the bivariate normal above has only five parameters. In fact, only μ_U can be estimated; the remaining parameters are not identifiable and the structural relation $V = \beta_0 + \beta_1 U$ cannot be estimated. If the distribution of V is other than normal, it may be possible to devise methods that will identify all the parameters. However, in practice, we never know the distribution of U , and the nearer the distribution is to normality, the worse the estimates. \square

We see from this simple example that a key issue in structural models is the identifiability of the parameters. One obvious approach to the straight-line problem is to impose some constraint on the parameters, thus effectively reducing the number of unknown parameters by one. Three types of constraint have been studied:

1. σ_δ^2 or σ_ϵ^2 is known: All the parameters are now identifiable and can be estimated.

2. $\sigma_\delta/\sigma_\epsilon$ ($= k$, say) is known: All the parameters can be estimated consistently. This is perhaps the most common constraint.
3. σ_U^2/σ_X^2 is known.
4. Both σ_δ and σ_ϵ are known: This leads to “overidentification” of the model so that this simplification is of limited practical use.

In practice, “known” usually means estimated accurately using an independent method. Fuller [1987] provides the theory of the straight line for cases (1)–(3) as well as the theory for several explanatory variables, nonlinear models, and multivariate models. Another way around the identification problem is to use replication (see, e.g., Seber and Wild [1989: Chapter 10] for the more general nonlinear case).

9.6.6 Controlled Variables Model

In this model, usually called *Berkson's model*, the explanatory variables are random but their observed values are controlled, a common situation when investigating lawlike relationships in the physical sciences. We demonstrate the idea with the following simple example.

EXAMPLE 9.4 Suppose that we wish to study Ohm's law

$$v = \beta u,$$

where v is the voltage in volts, u is the current in amperes, and β is the resistance in ohms. Then, for a given resistance, a natural experimental procedure would be to adjust the current through the circuit so that the ammeter reads a certain prescribed or “target” value x_i , for example, $x_i = 1$ A, and then measure the voltage Y_i with a voltmeter. The ammeter will have a random error so that the current actually flowing through the circuit is an unknown random variable U_i , say. Similarly, the true voltage will also be an unknown random variable V_i so that our model for this experiment is now

$$Y_i = V_i + \varepsilon_i = \beta U_i + \varepsilon_i,$$

which is of the form (9.25). However, the model above reduces to a “standard” least squares model

$$\begin{aligned} Y_i &= \beta x_i + \varepsilon_i + \beta(U_i - x_i) \\ &= \beta x_i + \varepsilon'_i, \end{aligned}$$

where the error or fluctuation term is now ε'_i instead of ε_i . What the discussion above implies is that in the controlled explanatory variables situation, the model may be analyzed as though the explanatory variables are nonrandom and error free. \square

9.7 COLLINEARITY

One important assumption that we have not yet mentioned is that the regression matrix \mathbf{X} is assumed to be of full rank. In practice, the columns of \mathbf{X} could be almost linearly dependent or *collinear*, which leads to $\mathbf{X}'\mathbf{X}$ being close to singular. Since

$$\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \quad (9.39)$$

near collinearity will have a considerable effect on the precision with which β can be estimated. When the estimated regression coefficients have large variances, tests will have low power, and confidence intervals will be very wide. It will be difficult to decide if a variable makes a significant contribution to the regression.

In this section we examine in more detail the effect that almost collinear columns have on the variances of the estimated coefficients and show how the resulting fitted regressions can be unstable. In Section 10.7, we discuss how we can detect the presence of almost collinear columns and what action we can take to improve the precision of our estimates.

9.7.1 Effect on the Variances of the Estimated Coefficients

In this section we develop more detailed expressions for the variances of the estimated regression coefficients and identify patterns in \mathbf{X} that lead to large variances. We begin by considering the case of the straight line.

Straight-Line Regression

The straight-line regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, \dots, n), \quad (9.40)$$

was considered in Section 6.1. The variances of the regression coefficients are

$$\text{var}[\hat{\beta}_0] = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{ns_{xx}} \quad (9.41)$$

and

$$\text{var}[\hat{\beta}_1] = \frac{\sigma^2}{s_{xx}}, \quad (9.42)$$

where $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$. Define the *coefficient of variation* of the x 's by

$$\text{CV}_x = \frac{(s_{xx}/n)^{1/2}}{|\bar{x}|}.$$

The quantity CV_x measures the variability of the x 's relative to their average size, and is independent of the units used to measure x . Using the identity

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2,$$

we can write (9.41) as

$$\begin{aligned}\text{var}[\hat{\beta}_0] &= \frac{\sigma^2 \{s_{xx} + n\bar{x}^2\}}{ns_{xx}} \\ &= \frac{\sigma^2}{n} (1 + 1/\text{CV}_x^2).\end{aligned}$$

If CV_x is small, then $\text{var}[\hat{\beta}_0]$ will be large.

In contrast, $\text{var}[\hat{\beta}_1]$ depends only on s_{xx} (i.e., on the absolute rather than relative variability of the x 's). We note that unlike $\text{var}[\hat{\beta}_1]$, $\text{var}[\hat{\beta}_0]$ does not depend on the scale with which the x 's are measured.

We can avoid estimating β_0 altogether by considering the centered model

$$Y_i = \alpha_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i \quad (i = 1, \dots, n). \quad (9.43)$$

The parameter α_0 represents the height of the true regression line at $x = \bar{x}$ rather than at $x = 0$. The true regression line is the same in both models, but it now has a different mathematical description. Comparing (9.40) with (9.43), we see that $\alpha_0 = \beta_0 - \beta_1\bar{x}$. From Section 3.11.1, the estimate of α_0 is $\hat{\alpha}_0 = \bar{Y}$ with variance σ^2/n . The estimate of β_1 is unchanged.

The magnitude of β_1 depends on the units in which the x 's are measured. If the x 's are multiplied by a constant factor, for example, if different units are used, then β_1 is reduced by the same factor, as are its estimate and standard error. Thus the absolute size of the estimate and its variance have meaning only when referred to a particular set of units.

It is common practice to *center and scale* the explanatory variable by transforming the x 's to the quantities

$$x_i^* = (x_i - \bar{x})/s_{xx}^{1/2},$$

so that $\sum_i x_i^* = 0$ and $\sum_i x_i^{*2} = 1$. This produces a scale-invariant explanatory variable, which is dimensionless (i.e., has no units). The model becomes (see Section 3.11.2)

$$Y_i = \alpha_0 + \gamma x_i^* + \varepsilon_i \quad (i = 1, \dots, n). \quad (9.44)$$

The regression coefficients α_0 and γ now are measured in the same units (that of the response variable Y) and will typically have the same order of magnitude. The estimate of α_0 is the same as before, but the estimate of γ is

$$\begin{aligned}\hat{\gamma} &= \sum_{i=1}^n x_i^*(Y_i - \bar{Y}) \\ &= \sum_{i=1}^n x_i^* Y_i\end{aligned}$$

with variance

$$\begin{aligned}\text{var}[\hat{\gamma}] &= \text{var} \left[\sum_{i=1}^n x_i^* Y_i \right] \\ &= \sum_{i=1}^n x_i^{*2} \text{var}[Y_i] \\ &= \sigma^2.\end{aligned}$$

In terms of the original data, we have

$$\hat{\gamma} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{s_{xx}^{1/2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{s_{xx}^{1/2}}. \quad (9.45)$$

The estimates are different because the parameters describing the true regression line are different. However, the fitted values \hat{Y}_i remain the same.

Two Explanatory Variables

We now consider the case of two explanatory variables, with model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \quad (i = 1, \dots, n), \quad (9.46)$$

or, in matrix notation, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and $\mathbf{X} = (\mathbf{1}, \mathbf{x}, \mathbf{z})$. Now let $\mathbf{w} = (\bar{x}, \bar{z})'$, and let $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$ be the centered version of (\mathbf{x}, \mathbf{z}) , as in Section 3.11.1. Then

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \begin{pmatrix} n & n\bar{x} & n\bar{z} \\ n\bar{x} & \sum_i x_i^2 & \sum_i x_i z_i \\ n\bar{z} & \sum_i x_i z_i & \sum_i z_i^2 \end{pmatrix} \\ &= n \begin{pmatrix} 1 & \mathbf{w}' \\ \mathbf{w} & \mathbf{S} + \mathbf{w}\mathbf{w}' \end{pmatrix},\end{aligned}$$

where $\mathbf{S} = n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$. Using A.9.1, we see that

$$(\mathbf{X}'\mathbf{X})^{-1} = n^{-1} \begin{pmatrix} 1 + \mathbf{w}'\mathbf{S}^{-1}\mathbf{w} & -\mathbf{w}'\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{w} & \mathbf{S}^{-1} \end{pmatrix},$$

so that

$$\text{var}[\hat{\beta}_0] = \sigma^2(1 + \mathbf{w}'\mathbf{S}^{-1}\mathbf{w})/n,$$

$$\text{var}[\hat{\beta}_1] = \sigma^2/\{s_{xx}(1 - r^2)\},$$

and

$$\text{var}[\hat{\beta}_2] = \sigma^2/\{s_{zz}(1 - r^2)\},$$

where $s_{zz} = \sum_{i=1}^n (z_i - \bar{z})^2$ and r is the correlation between \mathbf{x} and \mathbf{z} .

Also, after some algebra (see Exercises 9d, No. 1), we get

$$\text{var}[\hat{\beta}_0] = \frac{\sigma^2}{n} \left[1 + \frac{1}{1 - r^2} \left(\frac{1}{\text{CV}_z^2} - \frac{2r}{\text{CV}_x \text{CV}_z} + \frac{1}{\text{CV}_z^2} \right) \right],$$

so that the variance of $\hat{\beta}_0$ does not depend on the scale used to measure x and z .

As with straight-line regression, we can center and scale the x 's and the z 's, and use the model

$$Y_i = \alpha_0 + \gamma_1 x_i^* + \gamma_2 z_i^* + \varepsilon_i \quad (i = 1, \dots, n), \quad (9.47)$$

where $x_i^* = (x_i - \bar{x})/s_{xx}^{1/2}$ and $z_i^* = (z_i - \bar{z})/s_{zz}^{1/2}$. Setting $\mathbf{X}_s = (\mathbf{1}, \mathbf{x}^*, \mathbf{z}^*)$, we get

$$\mathbf{X}'_s \mathbf{X}_s = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{R}_{xx} \end{pmatrix},$$

where

$$\mathbf{R}_{xx} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

is the correlation matrix of \mathbf{x} and \mathbf{z} . Inverting, we get

$$\text{var}[\hat{\alpha}_0] = \frac{\sigma^2}{n},$$

$$\text{var}[\hat{\gamma}_1] = \frac{\sigma^2}{1 - r^2},$$

and

$$\text{var}[\hat{\gamma}_2] = \frac{\sigma^2}{1 - r^2}.$$

Thus, given the model (9.47), the accuracy of the scaled regression coefficients depends only on the error variance σ^2 and the correlation between x and z . In particular, the scaled coefficients cannot be estimated accurately if the correlation is close to 1, or, alternatively, when the explanatory variables cluster about a straight line in the (x, z) plane. Intuitively, when the data are well spread over the (x, z) plane, the fitted regression plane is well supported by the data. When the correlation is high, and \mathbf{x} and \mathbf{z} are almost linearly dependent, the regression plane is supported by a narrow ridge of points, and is consequently unstable, with a small change in the data resulting in a big change in the fitted plane. Hocking and Pendleton [1983] and Hocking [1996: p. 262] discuss this “picket fence” analogy in more detail.

General Case

We now consider the general regression model

$$Y_i = \beta_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i \quad (i = 1, \dots, n), \quad (9.48)$$

with $p - 1$ explanatory variables and a constant term. In centered and scaled form, the model becomes (see Section 3.11.2)

$$Y_i = \alpha_0 + \gamma_1 x_{i1}^* + \gamma_2 x_{i2}^* + \cdots + \gamma_{p-1} x_{ip-1}^* + \varepsilon_i \quad (i = 1, \dots, n), \quad (9.49)$$

or in matrix terms

$$\mathbf{Y} = \mathbf{X}_s \begin{pmatrix} \alpha_0 \\ \gamma \end{pmatrix} + \boldsymbol{\varepsilon}, \quad (9.50)$$

where $\mathbf{X}_s = (\mathbf{1}, \mathbf{X}^*)$. Thus

$$\mathbf{X}'_s \mathbf{X}_s = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{R}_{xx} \end{pmatrix},$$

where \mathbf{R}_{xx} is the correlation matrix of the explanatory variables x_1, \dots, x_{p-1} . We partition \mathbf{R}_{xx} as

$$\mathbf{R}_{xx} = \begin{pmatrix} 1 & \mathbf{r}' \\ \mathbf{r} & \mathbf{R}_{22} \end{pmatrix},$$

where $\mathbf{r} = (r_{12}, r_{13}, \dots, r_{1,p-1})'$ and \mathbf{R}_{22} is the matrix formed by deleting the first row and column from \mathbf{R}_{xx} . Here $r_{1j} = \mathbf{x}^{*(1)'} \mathbf{x}^{*(j)}$ ($j = 2, \dots, p-1$) is the correlation between the variables x_1 and x_j . Then, by A.9.1, the 1,1 element of \mathbf{R}_{xx}^{-1} is given by $(1 - \mathbf{r}' \mathbf{R}_{22}^{-1} \mathbf{r})^{-1}$, so that

$$\text{var}[\hat{\gamma}_1] = \sigma^2 (1 - \mathbf{r}' \mathbf{R}_{22}^{-1} \mathbf{r})^{-1}. \quad (9.51)$$

For an alternative interpretation of (9.51), let $\mathbf{x}^{*(j)}$ be the j th column of \mathbf{X}^* , and let $\mathbf{X}^{*(j)}$ be \mathbf{X}^* with the j th column removed. Consider the formal regression of the vector $\mathbf{x}^{*(j)}$ on the columns of $\mathbf{X}^{*(j)}$, including a constant term. We will now evaluate the coefficient of determination R_j^2 for this regression. Substituting x_{ij}^* for Y_i , and using the equations $\sum_i x_{ij}^* = 0$ and $\sum_i x_{ij}^{*2} = 1$, we get

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n x_{ij}^{*2} \\ &= 1. \end{aligned}$$

Hence, from Theorem 4.2(ii), it follows that

$$\begin{aligned} R_j^2 &= 1 - \frac{\text{RSS}_j}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \text{RSS}_j, \end{aligned} \quad (9.52)$$

where RSS_j is the residual sum of squares for the formal regression. By (3.52), the residual sum of squares for the regression of \mathbf{Y} on \mathbf{X} is

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 - \mathbf{Y}' \tilde{\mathbf{X}} (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \mathbf{Y}.$$

Making the substitutions of $\mathbf{x}^{*(j)}$ for \mathbf{Y} and $\mathbf{X}^{*(j)}$ for \mathbf{X} , and using the facts that for this substitution $\sum_i (Y_i - \bar{Y})^2 = 1$ and that centering $\mathbf{X}^{*(j)}$ has no effect, we get

$$\begin{aligned} \text{RSS}_j &= 1 - \mathbf{x}^{*(j)'} \mathbf{X}^{*(j)} (\mathbf{X}^{*(j)'} \mathbf{X}^{*(j)})^{-1} \mathbf{X}^{*(j)'} \mathbf{x}^{*(j)} \\ &= 1 - \mathbf{x}^{*(j)'} \mathbf{P}_j \mathbf{x}^{*(j)}, \end{aligned} \quad (9.53)$$

where $\mathbf{P}_j = \mathbf{X}^{*(j)'}(\mathbf{X}^{*(j)'}\mathbf{X}^{*(j)})^{-1}\mathbf{X}^{*(j)'}^T$ is the projection onto $C(\mathbf{X}^{*(j)})$. When $j = 1$, $\mathbf{X}^{*(1)'}\mathbf{x}^{*(1)} = \mathbf{r}$ and $\mathbf{X}^{*(1)'}\mathbf{X}^{*(1)} = \mathbf{R}_{22}$, so that

$$\text{RSS}_1 = \mathbf{1}'\mathbf{R}_{22}^{-1}\mathbf{r}.$$

Hence, from (9.52), $R_1^2 = \mathbf{r}'\mathbf{R}_{22}^{-1}\mathbf{r}$ and

$$\text{var}[\hat{\gamma}_1] = \sigma^2 / (1 - R_1^2).$$

Similar formulas hold for the other coefficients, namely,

$$\text{var}[\hat{\gamma}_j] = \sigma^2 / (1 - R_j^2).$$

Since $\mathbf{x}^{*(j)'}\mathbf{x}^{*(j)} = 1$ and $(\mathbf{I}_n - \mathbf{P}_j)$ is symmetric and idempotent, it follows from (9.52) and (9.53) that

$$1 - R_j^2 = \mathbf{x}^{*(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{*(j)} = \|(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{*(j)}\|^2. \quad (9.54)$$

Therefore, geometrically, $1 - R_j^2$ measures how close $\mathbf{x}^{*(j)}$ is to the subspace $C(\mathbf{X}^{*(j)})$, since it is the squared length of the residual vector when $\mathbf{x}^{*(j)}$ is projected onto $C(\mathbf{X}^{*(j)})$. In other words, $1 - R_j^2$ measures how near $\mathbf{x}^{*(j)}$ is to being a linear combination of the other explanatory variables. Thus, when the columns of the centered and scaled regression matrix are “almost collinear” in this sense, we can expect at least some of the regression coefficients to have large variances.

9.7.2 Variance Inflation Factors

The formula

$$\text{var}[\hat{\gamma}_j] = \sigma^2 / (1 - R_j^2)$$

given above expresses the variance of the scaled regression coefficient in terms of a coefficient of determination. Since R_j^2 is a squared correlation, we must have $0 \leq R_j^2 \leq 1$, so it follows that

$$\text{Var}[\hat{\gamma}_j] \geq \sigma^2$$

with equality if and only if $R_j^2 = 0$. By Exercises 9d, No. 2, this occurs when $\mathbf{x}^{*(j)}$ is orthogonal to the other columns of \mathbf{X}^* .

The term $(1 - R_j^2)^{-1}$ is called the j th *variance inflation factor* or VIF_j . From the discussion above, we have

$$\text{VIF}_j = \text{Var}[\hat{\gamma}_j]/\sigma^2$$

and

$$\text{VIF}_j = \|(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{*(j)}\|^{-2}.$$

In terms of the original variables \mathbf{X} , and using the relationship $\hat{\gamma}_j = \hat{\beta}_j s_j$ (cf. Section 3.11.2), we have

$$\begin{aligned}\text{VIF}_j &= \text{Var}[\hat{\gamma}_j]/\sigma^2 \\ &= \text{Var}[\hat{\beta}_j s_j]/\sigma^2 \\ &= s_j^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,j+1}.\end{aligned}$$

9.7.3 Variances and Eigenvalues

Another expression for the variance of the estimated regression coefficient $\hat{\gamma}_j$ can be derived from the spectral representation A.1.4 of the correlation matrix \mathbf{R}_{xx} . Since \mathbf{R}_{xx} is positive-definite (see Exercises 9d, No. 3), we can write its spectral representation as

$$\mathbf{R}_{xx} = \mathbf{T}\Lambda\mathbf{T}', \quad (9.55)$$

where $\mathbf{T} = (t_{ij})$ is orthogonal, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{p-1})$ and the λ 's are the eigenvalues of \mathbf{R}_{xx} , which are positive by A.4.1. From (9.55), we get $\mathbf{R}_{xx}^{-1} = \mathbf{T}\Lambda^{-1}\mathbf{T}'$, so that

$$\begin{aligned}\text{var}[\hat{\gamma}_j] &= \sigma^2 (\mathbf{R}_{xx}^{-1})_{jj} \\ &= \sigma^2 \sum_{l=1}^{p-1} t_{jl}^2 \lambda_l^{-1}.\end{aligned} \quad (9.56)$$

Now, since the rows of \mathbf{T} are orthonormal, we must have $\sum_{l=1}^{p-1} t_{jl}^2 = 1$, so that $|t_{jl}| \leq 1$ for all j and l . Thus, if any eigenvalue λ_l is close to zero and the element t_{jl} is *not* close to zero, then $\text{var}[\hat{\gamma}_j]$ must be large.

Thus, we have two ways of recognising when the variances of the regression coefficients in the centered and scaled model are large: (1) when one or more columns are “almost collinear” with the others, as measured by a large VIF, or, equivalently, when the projection of one column on the space spanned by the others has a small residual, and (2) one or more eigenvalues of the correlation matrix are small.

9.7.4 Perturbation Theory

Ideally, a small change in the regression data should cause only a small change in the regression coefficients. Statistically, the variance of $\hat{\beta}$ measures the change expected in $\hat{\beta}$ when the responses Y_i are subjected to changes, whose magnitudes are described by the error variance σ^2 .

A more direct method is to examine the relative change in the estimated regression coefficients when the data are subjected to small changes or perturbations. Below, we derive some bounds on these changes. First, we review the concepts of *matrix norm* and the *condition number* of a matrix.

Definition 9.1 If \mathbf{X} is an $n \times p$ matrix, then the 2-norm of \mathbf{X} is defined by

$$\|\mathbf{X}\|_2 = \max_{\mathbf{a}} \frac{\|\mathbf{X}\mathbf{a}\|}{\|\mathbf{a}\|} \quad (\|\mathbf{a}\| \neq 0), \quad (9.57)$$

where the maximum is taken over all nonzero p -vectors \mathbf{a} , and $\|\mathbf{a}\| = (\mathbf{a}'\mathbf{a})^{1/2}$ is the norm of \mathbf{a} .

Note that many definitions of matrix norm are possible; see, for example Higham [1996: Chapter 6], Björck [1996: p. 24], Golub and van Loan [1996: p. 52], and Trefethen and Bau [1997: p. 17]. It can be shown that a consequence of (9.57) is that

$$\|\mathbf{X}\|_2 = \sigma_{\text{MAX}}, \quad (9.58)$$

where σ_{MAX} is the largest singular value of \mathbf{X} , the square root of the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ (see A.12). Also, note the inequality

$$\|\mathbf{X}\mathbf{a}\| \leq \|\mathbf{X}\|_2 \|\mathbf{a}\|, \quad (9.59)$$

which follows directly from (9.57).

EXAMPLE 9.5 Let \mathbf{P} be a projection matrix. Then, since \mathbf{P} is symmetric and $\mathbf{P}\mathbf{P} = \mathbf{P}$, the eigenvalues of \mathbf{P} are zero or 1 (A.6.1). The largest eigenvalue of $\mathbf{P}'\mathbf{P} = \mathbf{P}$ is therefore 1, so $\|\mathbf{P}\|_2 = 1$. Note also from (9.59) that $\|\mathbf{P}\mathbf{a}\| \leq \|\mathbf{a}\|$. \square

Definition 9.2 The condition number of the matrix \mathbf{X} is the ratio of the largest and smallest singular values of \mathbf{X} , and is written $\kappa(\mathbf{X})$.

For matrices that are almost rank-deficient, the condition number is large and is theoretically infinite for matrices of less than full rank. The minimum value of the condition number is unity; this occurs when the matrix has orthonormal columns. If \mathbf{X} has full rank, then the condition number can also be written as

$$\kappa(\mathbf{X}) = (\lambda_{\text{MAX}}/\lambda_{\text{MIN}})^{1/2}, \quad (9.60)$$

where λ_{MAX} and λ_{MIN} are the largest and smallest eigenvalues of $\mathbf{X}'\mathbf{X}$.

EXAMPLE 9.6 (Straight-line regression) Let

$$\mathbf{X}_c = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}.$$

Then

$$\mathbf{X}'_c \mathbf{X}_c = \begin{pmatrix} n & 0 \\ 0 & s_{xx} \end{pmatrix},$$

with eigenvalues n and $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$. If $\hat{\sigma}^2 = (1/n)s_{xx}$, then assuming that $\hat{\sigma} < 1$ (or $s_{xx} < n$), we have $\|\mathbf{X}_c\|_2 = n^{1/2}$ and $\kappa(\mathbf{X}_c) = 1/\hat{\sigma}$. \square

Let us now derive a bound for changes in the regression coefficient $\hat{\gamma}$ in the centered and scaled straight-line regression model (9.44). Suppose that the original noncentered and unscaled x_i 's are changed by small amounts δ_i , where $|\delta_i| < \epsilon$, and assume that $0 < \epsilon < \hat{\sigma} < 1$. Then $\sum_i \delta_i^2 < n\epsilon^2$, so that $\|\delta\| < n^{1/2}\epsilon$. Also consider the projection $\mathbf{P} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$, so that

$$\begin{aligned}\mathbf{Px} &= (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})', \\ \|\mathbf{Px}\|^2 &= s_{xx} = n\hat{\sigma}^2,\end{aligned}\tag{9.61}$$

and

$$\mathbf{x}^* = \mathbf{Px}/\|\mathbf{Px}\|.$$

Using (9.45), we can now write $\hat{\gamma}$ in terms of \mathbf{P} as

$$\hat{\gamma} = \frac{\mathbf{Y}'\mathbf{Px}}{\|\mathbf{Px}\|}.$$

Now suppose that the x_i 's are perturbed as described above, and let $\delta = (\delta_1, \dots, \delta_n)'$. The new estimate of γ is

$$\begin{aligned}\hat{\gamma}_\epsilon &= \frac{\mathbf{Y}'\mathbf{P}(\mathbf{x} + \delta)}{\|\mathbf{P}(\mathbf{x} + \delta)\|} \\ &= \frac{\hat{\gamma}\|\mathbf{Px}\| + \mathbf{Y}'\mathbf{P}\delta}{\|\mathbf{P}(\mathbf{x} + \delta)\|}.\end{aligned}$$

Thus the relative change in $\hat{\gamma}$ is

$$\frac{\hat{\gamma} - \hat{\gamma}_\epsilon}{\hat{\gamma}} = \frac{\|\mathbf{P}(\mathbf{x} + \delta)\| - \|\mathbf{Px}\| - \mathbf{Y}'\mathbf{P}\delta/\hat{\gamma}}{\|\mathbf{P}(\mathbf{x} + \delta)\|},\tag{9.62}$$

so that

$$\left| \frac{\hat{\gamma} - \hat{\gamma}_\epsilon}{\hat{\gamma}} \right| \leq \frac{\|\mathbf{P}(\mathbf{x} + \delta)\| - \|\mathbf{Px}\| + |\mathbf{Y}'\mathbf{P}\delta/\hat{\gamma}|}{\|\mathbf{P}(\mathbf{x} + \delta)\|}.\tag{9.63}$$

Now consider the inequality (cf. A.11.4)

$$|\|\mathbf{a}\| - \|\mathbf{b}\|| \leq \|\mathbf{a} - \mathbf{b}\|,\tag{9.64}$$

which is valid for all vectors \mathbf{a} and \mathbf{b} . Putting $\mathbf{a} = \mathbf{P}(\mathbf{x} + \delta)$ and $\mathbf{b} = \mathbf{Px}$, we get

$$\begin{aligned}|\|\mathbf{P}(\mathbf{x} + \delta)\| - \|\mathbf{Px}\|| &\leq \|\mathbf{P}\delta\| \\ &\leq \|\delta\|, \\ &< n^{1/2}\epsilon,\end{aligned}\tag{9.65}$$

using Example 9.5. Also, since $\|\mathbf{P}\delta\| < n^{1/2}\epsilon$, we have from (9.61) that

$$\begin{aligned} \|\mathbf{Px}\| - \|\mathbf{P}\delta\| &> \|\mathbf{Px}\| - n^{1/2}\epsilon \\ &= n^{1/2}(\hat{\sigma} - \epsilon) \\ &> 0. \end{aligned}$$

Putting $\mathbf{a} = \mathbf{Px}$ and $\mathbf{b} = -\mathbf{P}\delta$ in (9.64), we get from the equation above

$$\begin{aligned} \|\mathbf{P}(\mathbf{x} + \delta)\| &\geq \|\mathbf{Px}\| - \|\mathbf{P}\delta\| \\ &= \|\mathbf{Px}\| - \|\mathbf{P}\delta\| \\ &> n^{1/2}(\hat{\sigma} - \epsilon). \end{aligned} \tag{9.66}$$

Now the estimate of α_0 is \bar{Y} , so that

$$Y_i - \bar{Y} = \hat{\gamma}x_i^* + e_i,$$

and thus

$$\mathbf{PY} = \hat{\gamma}\mathbf{Px}/\|\mathbf{Px}\| + \mathbf{e}.$$

Using this and the Cauchy–Schwartz inequality (A.4.11), we can write

$$\begin{aligned} |\delta' \mathbf{PY}| / |\hat{\gamma}| &\leq |\delta' \mathbf{Px}| / \|\mathbf{Px}\| + |\delta' \mathbf{e}| / |\hat{\gamma}| \\ &\leq \|\delta\|(1 + \|\mathbf{e}\| / |\hat{\gamma}|). \end{aligned} \tag{9.67}$$

Combining (9.63), (9.65), (9.66), and (9.67), we have

$$\begin{aligned} \left| \frac{\hat{\gamma} - \hat{\gamma}_\epsilon}{\hat{\gamma}} \right| &< \frac{n^{1/2}\epsilon + |\mathbf{Y}'\mathbf{P}\delta/\hat{\gamma}|}{n^{1/2}(\hat{\sigma} - \epsilon)} \\ &< \frac{n^{1/2}\epsilon(2 + \|\mathbf{e}\| / |\hat{\gamma}|)}{n^{1/2}(\hat{\sigma} - \epsilon)}. \end{aligned}$$

Using the formula $\kappa(\mathbf{X}_c) = \hat{\sigma}^{-1}$ from Example 9.6 and the assumption that $\hat{\sigma} < 1$, we get

$$\left| \frac{\hat{\gamma} - \hat{\gamma}_\epsilon}{\hat{\gamma}} \right| \leq \frac{\kappa(\mathbf{X}_c)\epsilon}{1 - \kappa(\mathbf{X}_c)\epsilon} \left(2 + \frac{\|\mathbf{e}\|}{|\hat{\gamma}|} \right).$$

This shows that provided the condition number is not too large and $|\hat{\gamma}|$ is not too small, a small change in the x_i 's will not cause too great a change in the estimate.

What if the condition number is large? Then, under some circumstances, a small change in the x_i 's can cause a large relative change in the estimate.

EXAMPLE 9.7 Suppose that the sample correlation of δ with both \mathbf{x} and \mathbf{y} is zero and that δ has its elements summing to zero. Since

$$\sum_I (Y_i - \bar{Y})(\delta_i - \bar{\delta}) = \sum_i Y_i(\delta_i - \bar{\delta}) = \sum_i Y_i\delta_i, \tag{9.68}$$

we have $\mathbf{Y}'\delta = 0$ and $\mathbf{x}'\delta = 0$. Then, from $\mathbf{1}_n'\delta = 0$, we have $\mathbf{P}\delta = \delta$, $\mathbf{Y}'\mathbf{P}\delta = \mathbf{Y}'\delta = 0$, and $(\mathbf{P}\delta)'\mathbf{P}\mathbf{x} = \delta'\mathbf{P}\mathbf{x} = \delta'\mathbf{x} = 0$. Thus, $\mathbf{P}\mathbf{x}$ and $\mathbf{P}\delta$ are orthogonal, so that

$$\begin{aligned}\|\mathbf{P}(\mathbf{x} + \delta)\|^2 &= \|\mathbf{Px}\|^2 + \|\mathbf{P}\delta\|^2 \\ &= \|\mathbf{Px}\|^2 + \|\delta\|^2.\end{aligned}$$

Substituting into (9.62), and using $\mathbf{Y}'\mathbf{P}\delta = 0$, the relative change is

$$1 - \sqrt{\frac{\|\mathbf{Px}\|^2}{\|\mathbf{Px}\|^2 + \|\delta\|^2}} = 1 - (1 + \kappa(\mathbf{X}_c)^2 \|\delta\|^2/n)^{-1/2},$$

from (9.61), since $\kappa(\mathbf{X}_c)^2 = 1/\hat{\sigma}^2 = n/\|\mathbf{Px}\|^2$. Thus, for a fixed but arbitrarily small change in the x 's, the relative change in $\hat{\gamma}$ can be as much as 100% if the condition number is large enough. \square

The arguments above show that the condition number $\kappa(\mathbf{X}_c)$ is a good diagnostic for instability. On the other hand, the variance of $\hat{\gamma}$ is always σ^2 , no matter what the x_i 's, so can give us no information about any possible instability of the estimate when there are small changes in the x_i 's.

We can also look at the relative change in $\hat{\gamma}$ compared to the relative change in a single x_i . Suppose that x_i changes to $x_i + \Delta x_i$ where Δx_i is small, resulting in a change $\Delta\hat{\gamma}$ in $\hat{\gamma}$. The ratio of relative changes is

$$\left| \frac{\Delta\hat{\gamma}}{\hat{\gamma}} \right| / \left| \frac{\Delta x_i}{x_i} \right| = \left| \frac{\Delta\hat{\gamma}}{\Delta x_i} \right| \cdot \left| \frac{x_i}{\hat{\gamma}} \right|.$$

Letting $\Delta \rightarrow 0$, the right-hand side of this expression approaches

$$E_i = \left| \frac{\partial\hat{\gamma}}{\partial x_i} \right| \cdot \left| \frac{x_i}{\hat{\gamma}} \right|. \quad (9.69)$$

The quantity (9.69), called the *i*th *elasticity*, measures the sensitivity of the estimate to small relative changes in the original data. To evaluate the E_i 's, we make use of the results (see Exercises 9d, No. 5)

$$\frac{\partial}{\partial x_i} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = Y_i - \bar{Y}$$

and

$$\frac{\partial}{\partial x_i} \sum_{i=1}^n (x_i - \bar{x})^2 = 2(x_i - \bar{x}).$$

Using these results and formula (9.45), we get

$$\begin{aligned}\frac{\partial\hat{\gamma}}{\partial x_i} &= \frac{Y_i - \bar{Y} - \hat{\gamma}x_i^*}{s_{xx}^{1/2}} \\ &= \frac{e_i}{s_{xx}^{1/2}},\end{aligned}$$

where e_i is the i th residual. Thus

$$E_i = \left| \frac{e_i}{\hat{\gamma}} \right| \cdot \frac{|x_i|}{s_{xx}^{1/2}}. \quad (9.70)$$

Since $|\mathbf{e}'\mathbf{x}| \leq \|\mathbf{e}\| \|\mathbf{x}\|$ (A.4.11), the sum of the elasticities is bounded by

$$\frac{\|\mathbf{e}\| \|\mathbf{x}\|}{|\hat{\gamma}| s_{xx}^{1/2}}.$$

Using the relationship $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$, we have

$$\|\mathbf{x}\|/s_{xx}^{1/2} = (1 + CV_x^{-1})^{1/2}. \quad (9.71)$$

Thus, if the CV of the x 's is small, all the elasticities will also be small, provided that $|\hat{\gamma}|$ is not too close to zero.

In summary, if the condition number of the centered matrix is not too large, the regression will be stable with respect to changes in the original x -data. If CV_x is not too large, the elasticities (where the relative change in the coefficient is compared to the relative change in the x_i 's) will not be too large. See Belsley [1991] for more information on elasticities.

General Case

A more general perturbation result can also be proved, which allows for changes in both the explanatory variables and the response. Consider the regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

which we can take to be either centered or uncentered. Suppose that the data is subject to small changes, resulting in new data $\mathbf{X} + \delta\mathbf{X}$ and $\mathbf{Y} + \delta\mathbf{Y}$, where $\|\delta\mathbf{X}\|_2 < \epsilon \|\mathbf{X}\|_2$, $\|\delta\mathbf{Y}\| < \epsilon \|\mathbf{Y}\|$, and $\kappa(\mathbf{X})\epsilon < 1$. Let $\hat{\boldsymbol{\beta}}_\epsilon$ be the new least squares estimate after these changes. Then (see e.g., Higham [1996: p. 392])

$$\frac{\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_\epsilon\|}{\|\hat{\boldsymbol{\beta}}\|} \leq \frac{\kappa(\mathbf{X})\epsilon}{1 - \kappa(\mathbf{X})\epsilon} \left(2 + (1 + \kappa(\mathbf{X})) \frac{\|\boldsymbol{\epsilon}\|}{\|\mathbf{X}\|_2 \|\hat{\boldsymbol{\beta}}\|} \right). \quad (9.72)$$

Thus, as in the case of simple linear regression, if the condition number of the regression matrix is not too large, the regression coefficients will be stable with respect to small relative changes in the regression matrix.

The idea of stability of regression coefficients is another way of approaching the concept of multicollinearity. Note that if the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$ is not too small, then the condition number of \mathbf{X} cannot be too large, at least relative to the size of the elements of \mathbf{X} , so the regression will be stable. On the other hand, if the smallest eigenvalue is not too small, then the variances of the regression coefficients cannot be too large; so the ideas of stability and small variance are connected.

9.7.5 Collinearity and Prediction

Collinearity affects the estimation of regression coefficients much more than prediction. To demonstrate this, suppose that we want to predict the response at $\mathbf{x}_0 = (x_{01}, \dots, x_{0,p-1})'$, using the predictor $\hat{Y} = \hat{\beta}_0 + \hat{\beta}'_c \mathbf{x}_0$, where $\hat{\beta}_c = (\hat{\beta}_1, \dots, \hat{\beta}_{p-1})$. Then from Section 3.11, we have

$$\hat{Y} = \bar{Y} + \hat{\beta}'_c(\mathbf{x}_0 - \bar{\mathbf{x}}),$$

where $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_{p-1})'$. Also, we recall from Section 3.11 the estimate $\hat{\beta}_c = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}$, where $\tilde{\mathbf{X}}$ is the centered version of \mathbf{X} (without $\mathbf{1}_n$).

From Theorem 1.3 and Section 3.11, we get

$$\begin{aligned}\text{Cov}[\bar{Y}, \hat{\beta}_c] &= \text{Cov}[n^{-1}\mathbf{1}'\mathbf{Y}, (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y}] \\ &= n^{-1}\mathbf{1}'\text{Var}[\mathbf{Y}]\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \\ &= \sigma^2 n^{-1}\mathbf{1}'\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \\ &= \mathbf{0},\end{aligned}$$

since $\mathbf{1}'\tilde{\mathbf{X}} = \mathbf{0}$. Thus

$$\begin{aligned}\text{var}[\hat{Y}] &= \text{var}[\bar{Y} + \hat{\beta}'_c(\mathbf{x}_0 - \bar{\mathbf{x}})] \\ &= \text{var}[\bar{Y}] + \text{var}[\hat{\beta}'_c(\mathbf{x}_0 - \bar{\mathbf{x}})] \\ &= \sigma^2 n^{-1} + (\mathbf{x}_0 - \bar{\mathbf{x}})' \text{Var}[\hat{\beta}_c](\mathbf{x}_0 - \bar{\mathbf{x}}) \\ &= \sigma^2 \left\{ n^{-1} + (\mathbf{x}_0 - \bar{\mathbf{x}})' (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}) \right\}. \quad (9.73)\end{aligned}$$

Up to the factor σ^2 , the second term in (9.73) is just the Mahalanobis distance between \mathbf{x}_0 and $\bar{\mathbf{x}}$. Thus, if we are predicting the response at \mathbf{x}_0 , the variance of the predictor depends on how outlying \mathbf{x}_0 is: Predictions at points close to $\bar{\mathbf{x}}$ have a small error.

Conversely, predictions made at outlying points have large errors. This is not of much consequence, as it is unwise to make a prediction at an outlying point. We cannot be sure that the model holds at points remote from the observed data.

If the data are collinear, then points close in the sense of Mahalanobis distance to \mathbf{x}_0 will lie almost in a subspace of lower dimension. For example, in the case $p = 2$, points close to $\bar{\mathbf{x}}$ will cluster about a line. Despite the collinearity, predictions made at points close to this line will have small errors, provided that they are not too far from \mathbf{x}_0 .

EXERCISES 9d

- Prove that in a regression with two explanatory variables x and z ,

$$\text{var}[\hat{\beta}_0] = \frac{\sigma^2}{n} \left[1 + \frac{1}{1-r^2} \left(\frac{1}{\text{CV}_x^2} - \frac{2r}{\text{CV}_x \text{CV}_z} + \frac{1}{\text{CV}_z^2} \right) \right].$$

2. Prove that $R_j^2 = 0$ if and only if $\mathbf{x}^{*(j)}$ is orthogonal to the columns of $\mathbf{X}^{*(j)}$.
3. Show that \mathbf{R}_{xx} is positive-definite provided that \mathbf{X} has full rank and contains a column of 1's.
4. Show that the eigenvalues λ_j of \mathbf{R}_{xx} satisfy

$$1 \leq \lambda_j \leq p - 1.$$

Hence prove that $VIF_j \leq \kappa^2$, where κ is the condition number of \mathbf{R}_{xx} .

5. Prove that

$$\frac{\partial}{\partial x_i} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = Y_i - \bar{Y}$$

and

$$\frac{\partial}{\partial x_i} \sum_{i=1}^n (x_i - \bar{x})^2 = 2(x_i - \bar{x}).$$

MISCELLANEOUS EXERCISES 9

1. Suppose that the regression model postulated is

$$E[Y] = \beta_0 + \beta_1 x$$

when, in fact, the true model is

$$E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

If we use observations of Y at $x = -3, -2, -1, 0, 1, 2, 3$ to estimate β_0 and β_1 in the postulated model, what bias will be introduced in these estimates?

2. Suppose that the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

is fitted when the true model is actually

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

If e_i is the residual from the fitted model, prove that

$$E[e_i] = \beta_2(x_{i2} + gx_{i1} + h),$$

where g and h are functions of the x_{ij} .

3. Suppose we wish to test the hypothesis H that the means of two populations are equal, given n_i observations from the i th population ($i = 1, 2$). Assuming that the populations have the same variance and kurtosis (γ_2), find approximate expressions for $E[Z]$ and $\text{var}[Z]$ on the assumption that H is true (cf. Theorem 9.2). Show that to the order of approximation used, these expressions are independent of γ_2 if $n_1 = n_2$.
4. Prove that (X_i, Y_i) in Example 9.3 of Section 9.6.5 has the bivariate normal distribution stated.

10

Departures from Assumptions: Diagnosis and Remedies

10.1 INTRODUCTION

In Chapter 9, we described what happens when the standard regression assumptions are not met. In this chapter we consider how departures from these assumptions may be detected and how their effects may be overcome by suitable transformations of the variables and weighting of the cases.

Most diagnostic techniques make use of various kinds of residuals as well as measures such as *hat matrix diagonals* that measure how “outlying” are the rows of the regression matrix. These measures are discussed in Section 10.2.

The most serious form of model misspecification occurs when we use a model that is linear in the explanatory variables when in fact the conditional mean of the responses is a *nonlinear* function of the x -variables. More precisely, suppose that we have a set of explanatory variables $\mathbf{x} = (x_0, x_1, \dots, x_{p-1})'$ (with $x_0 = 1$, say) and we attempt to model the response Y as

$$Y = \mathbf{x}'\beta + \varepsilon.$$

Then the least squares estimate of β will not be estimating anything very meaningful if the true model has $E[Y|\mathbf{x}] = \mu(\mathbf{x})$, where μ is a nonlinear function of \mathbf{x} . We need ways of visualizing the true nature of μ , and of deciding if a linear form $\mu(\mathbf{x}) = \mathbf{x}'\beta$ is at least approximately satisfied. If it is not, we will want to transform the x 's in order to achieve a better fit. In Section 10.3 we discuss ways of visualizing the form of μ , deciding if the linear assumption is adequate, and choosing a transformation if it is not.

The standard regression model assumes that the variance function $\text{var}[Y|\mathbf{x}]$ does not depend functionally on the explanatory variables. If, in fact, we have $\text{var}[Y|\mathbf{x}] = w(\mathbf{x})$, where $w(\mathbf{x})$ is not constant, then (see Section 9.3)

the least squares estimates will still be unbiased estimates of the regression coefficients but will not be efficient. If $w(\mathbf{x})$ is known, we can use weighted least squares as described in Section 3.10. If $w(\mathbf{x})$ is not known, we must decide if it can be assumed constant. If we cannot assume this, we must estimate $w(\mathbf{x})$ and incorporate the estimate into a new estimation procedure for the regression coefficients. Alternatively, we can transform the response in the hope of making the errors more homoscedastic. The details are given in Sections 10.4.1 to 10.4.3.

Even if the variance function is constant, the errors ϵ may fail to be independent. For example, if the data have been collected sequentially, successive observations may be serially correlated. This can be detected by the Durbin–Watson test, which is discussed in Section 10.4.4.

If the errors are not normally distributed, then not too much goes wrong provided that the joint distribution of the explanatory variables is approximately normal. If this is not the case, we can often use transformation methods to improve the situation, either transforming the response alone or using the *transform both sides* technique. Such transformations to normality are discussed in Section 10.5.

Outliers in the data can “attract” the fitted line or plane, resulting in a poor fit to the remaining observations. This can be particularly pronounced if a case has extreme values of the explanatory variables. We have two options here; either identify the outliers and downweight or delete them before fitting the model, or use a robust fitting method that is resistant to the outliers. We discuss both these approaches in Section 10.6.

Finally, in Section 10.7 we discuss how to detect collinear columns in the regression matrix and suggest some possible remedies for collinearity.

10.2 RESIDUALS AND HAT MATRIX DIAGONALS

We begin with the usual model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where \mathbf{X} is $n \times p$ of rank p . The major tools for diagnosing model faults are the *residuals*, which were introduced in Section 3.1. In terms of the projection matrix \mathbf{P} , which projects onto $C(\mathbf{X})$, they are given by

$$\begin{aligned}\mathbf{e} &= (\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= (\mathbf{I}_n - \mathbf{P})\epsilon,\end{aligned}$$

since $(\mathbf{I}_n - \mathbf{P})\mathbf{X} = \mathbf{0}$ [by Theorem 3.1(iii)].

The elements of the fitted regression $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ are called the *fitted values* and satisfy $\hat{\mathbf{Y}} = \mathbf{PY}$. In much of the literature on regression diagnostics, the projection matrix \mathbf{P} is called the *hat matrix*, since it transforms the responses (the Y_i 's) into the fitted values (the \hat{Y}_i 's). For this reason, it is often denoted by \mathbf{H} rather than \mathbf{P} . We use \mathbf{H} in this chapter.

Using the theorems of Chapter 1 and the idempotence of \mathbf{H} , we get

$$E[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H})E[\mathbf{Y}] = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\beta = \mathbf{0},$$

$$\begin{aligned} \text{Var}[\mathbf{e}] &= \text{Var}[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}] \\ &= (\mathbf{I}_n - \mathbf{H})\text{Var}[\mathbf{Y}](\mathbf{I}_n - \mathbf{H})' \\ &= (\mathbf{I}_n - \mathbf{H})\sigma^2\mathbf{I}_n(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H}), \end{aligned} \quad (10.1)$$

$$E[\hat{\mathbf{Y}}] = \mathbf{H}E[\mathbf{Y}] = \mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta,$$

and

$$\text{Var}[\hat{\mathbf{Y}}] = \mathbf{H}\text{Var}[\mathbf{Y}]\mathbf{H}' = \sigma^2\mathbf{H}. \quad (10.2)$$

Moreover,

$$\text{Cov}[\mathbf{e}, \hat{\mathbf{Y}}] = \text{Cov}[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \mathbf{H}\mathbf{Y}] = \sigma^2(\mathbf{I}_n - \mathbf{H})\mathbf{H} = \mathbf{0},$$

which implies the independence of \mathbf{e} and $\hat{\mathbf{Y}}$ under normality assumptions (by Theorem 2.5). If $\mathbf{H} = (h_{ij})$, the diagonal elements h_{ii} of \mathbf{H} are called the *hat matrix diagonals* and following general practice, we denote them by h_i rather than h_{ii} . We note from (10.1) that $\text{var}[e_i] = \sigma^2(1 - h_i)$.

The results above show that, when the model is correct, the variances of the residuals depend on the hat matrix diagonals. For this reason the residuals are sometimes scaled to have approximately unit variance; this leads to the *internally Studentized residual*

$$r_i = \frac{e_i}{S(1 - h_i)^{1/2}}, \quad (10.3)$$

where $S^2 = \mathbf{e}'\mathbf{e}/(n - p)$ is the usual estimate of σ^2 . It can be shown that (see Cook and Weisberg [1982: p. 18] and the references cited there; also Exercises 10a, No. 3) that $r_i^2/(n - p)$ has a beta $[\frac{1}{2}, \frac{1}{2}(n - p - 1)]$ distribution (A.13.6), so that the internally Studentized residuals are identically distributed.

Since the residuals (and hence the estimate S^2 of σ^2) can be affected by outliers, some writers favor using the *externally Studentized residual*

$$t_i = \frac{e_i}{S(i)(1 - h_i)^{1/2}}. \quad (10.4)$$

The estimate S is replaced by the estimate $S(i)$ which is calculated in the usual way from the $n - 1$ data points that remain after deleting the i th case. This results in an estimate of σ that will not be affected if the i th case is an outlier.

To derive the distribution of t_i , we first prove a theorem that we need to establish the relationship between $S(i)^2$ and S^2 .

THEOREM 10.1 Let $\hat{\beta}$ and $\hat{\beta}(i)$ denote the least squares estimate of β with and without the i th case included in the data. Then

$$\hat{\beta} - \hat{\beta}(i) = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_i}. \quad (10.5)$$

Proof. Let $\mathbf{X}(i)$ denote the regression matrix with the i th row deleted. Since $\mathbf{X}(i)'\mathbf{X}(i) = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'$, we have from A.9.4 that

$$\begin{aligned} (\mathbf{X}(i)'\mathbf{X}(i))^{-1} &= (\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} \\ &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_i}. \end{aligned} \quad (10.6)$$

Hence

$$\begin{aligned} \hat{\beta}(i) &= [\mathbf{X}(i)'\mathbf{X}(i)]^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{x}_i Y_i) \\ &= \left[\mathbf{X}'\mathbf{X}^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_i} \right] (\mathbf{X}'\mathbf{Y} - \mathbf{x}_i Y_i) \\ &= \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1 - h_i} [Y_i(1 - h_i) - \mathbf{x}_i' \hat{\beta} + h_i Y_i] \\ &= \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_i}. \end{aligned} \quad (10.7)$$

□

Using Theorem 10.1, we have

$$\begin{aligned} (n - p - 1)S(i)^2 &= \sum_{l \neq i} [Y_l - \mathbf{x}_l' \hat{\beta}(i)]^2 \\ &= \sum_{l \neq i} \left[Y_l - \mathbf{x}_l' \left(\hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_i} \right) \right]^2 \\ &= \sum_{l \neq i} \left(e_l + \frac{h_{li} e_i}{1 - h_i} \right)^2 \\ &= \sum_{l=1}^n \left(e_l + \frac{h_{li} e_i}{1 - h_i} \right)^2 - \frac{e_i^2}{(1 - h_i)^2}. \end{aligned} \quad (10.8)$$

Since \mathbf{H} is symmetric and satisfies $\mathbf{He} = \mathbf{0}$ and $\mathbf{H}^2 = \mathbf{H}$, it follows that $\sum_l h_{li} e_l = 0$ and $\sum_l h_{il}^2 = h_i$. Using these expressions in the right-hand side of (10.8) leads to

$$(n - p - 1)S(i)^2 = (n - p)S^2 - \frac{e_i^2}{1 - h_i}. \quad (10.9)$$

Hence, from $e_i^2/(1 - h_i) = r_i^2 S^2$,

$$\begin{aligned}
 t_i^2 &= \frac{e_i^2(n-p-1)}{(n-p-1)S(i)^2(1-h_i)} \\
 &= \frac{e_i^2(n-p-1)}{[(n-p)S^2 - e_i^2/(1-h_i)](1-h_i)} \\
 &= \frac{e_i^2}{S^2(1-h_i)} \frac{n-p-1}{n-p-r_i^2} \\
 &= \frac{r_i^2(n-p-1)}{n-p-r_i^2} \\
 &= \frac{B}{1-B}(n-p-1),
 \end{aligned} \tag{10.10}$$

say, where, by Exercises 10a, No. 3, the random variable $B = r_i^2(n-p)^{-1}$ has a beta $\left[\frac{1}{2}, \frac{1}{2}(n-p-1)\right]$ distribution (see A.13.6). We now use the fact (see Exercises 10a, No. 2) that if B has a beta $(\frac{1}{2}\alpha, \frac{1}{2}\beta)$ distribution, then $\beta B \{\alpha(1-B)\}^{-1}$ has an $F_{\alpha, \beta}$ distribution. Setting $\alpha = 1$ and $\beta = n-p-1$, we see that t_i^2 has an $F_{1, (n-p-1)}$ distribution, or equivalently, that t_i has a t_{n-p-1} distribution.

The hat matrix diagonals can be interpreted as a measure of distance in $(p-1)$ -dimensional space. To see this, we recall from (3.53) that

$$h_i = n^{-1} + (n-1)^{-1} \text{MD}_i, \tag{10.11}$$

where MD_i is the Mahalanobis distance

$$\text{MD}_i = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

between \mathbf{x}_i (now interpreted as the “reduced” i th row of \mathbf{X} without the initial element of 1) and the average reduced row. Thus, the hat matrix diagonal is a measure of how “outlying” the i th data point is, at least as far as the explanatory variables are concerned. This measure, however, is based on the sample covariance matrix \mathbf{S} and the mean vector $\bar{\mathbf{x}}$, which are not resistant to outliers. More robust measures are considered in Section 10.6.2.

Assuming that the regression model has a constant term, so that \mathbf{X} contains a column $\mathbf{1}_n$, it follows from (9.13) that

$$n^{-1} \leq h_i \leq 1. \tag{10.12}$$

The upper bound is attained in the limit as \mathbf{x}_i moves farther and farther away from $\bar{\mathbf{x}}$. Finally, from the proof of Theorem 3.1(ii),

$$\sum_i h_i = \text{tr}(\mathbf{H}) = p, \tag{10.13}$$

so that the average hat matrix diagonal is p/n .

EXAMPLE 10.1 Consider the *regression through the origin* model

$$Y_i = \beta x_i + \varepsilon_i.$$

Since $\mathbf{X} = (x_1, x_2, \dots, x_n)' = \mathbf{x}$, the hat matrix is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{x}\mathbf{x}'/\|\mathbf{x}\|^2$$

and $h_i = x_i^2 / \sum_k x_k^2$. □

EXAMPLE 10.2 In the simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

the hat matrix diagonals are, from Section 4.3.4,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}. \quad (10.14)$$

□

EXERCISES 10a

1. Prove (10.14).
2. Show that if the random variable B has a beta($\frac{1}{2}\alpha, \frac{1}{2}\beta$) distribution, then $F = \beta B\{\alpha(1 - B)\}^{-1}$ has an $F_{\alpha, \beta}$ distribution. Express B in terms of F .
3. (a) Express e_i in the form $e_i = \mathbf{c}_i'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}$ for a suitable vector \mathbf{c}_i .
 (b) Show that $(n - p)^{-1}r_i^2$ can be written as

$$(n - p)^{-1}r_i^2 = \frac{\mathbf{Z}'\mathbf{Q}\mathbf{Z}}{\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}},$$

where $\mathbf{Q} = (1 - h_i)^{-1}(\mathbf{I}_n - \mathbf{H})\mathbf{c}_i\mathbf{c}_i'(\mathbf{I}_n - \mathbf{H})$ and $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.

- (c) Prove that \mathbf{Q} is a projection matrix (i.e., show that $\mathbf{Q}^2 = \mathbf{Q}$ and $\mathbf{Q}' = \mathbf{Q}$).
- (d) Show that $(\mathbf{I}_n - \mathbf{H}) - \mathbf{Q}$ is a projection matrix and prove that $\mathbf{Z}'\mathbf{Q}\mathbf{Z}$ and $\mathbf{Z}'(\mathbf{I}_n - \mathbf{H}) - \mathbf{Q})\mathbf{Z}$ are independent. Hence prove that $(n - p)^{-1}r_i^2$ has a beta($\frac{1}{2}, \frac{1}{2}(n - p - 1)$) distribution.
4. Show that $(1 - h_i)^2 + \sum_{j \neq i} h_{ij}^2 = (1 - h_i)$.

10.3 DEALING WITH CURVATURE

Let $E[Y|\mathbf{x}]$ denote the conditional mean of the response Y given the explanatory variables \mathbf{x} . In order to assess the suitability of the linear model $E[Y|\mathbf{x}] = \beta' \mathbf{x}$, we need to be able to visualize the true regression surface $E[Y|\mathbf{x}] = \mu(\mathbf{x})$ and decide if it can be adequately represented by a linear function of \mathbf{x} .

10.3.1 Visualizing Regression Surfaces

In the case of a single explanatory variable x , a simple plot of Y versus x will reveal the relationship between the variables. The relationship can be made more apparent by smoothing the plot using a readily available smoother such as loess (end of Section 6.6) and smoothing splines (Section 7.2.3). Figure 10.1 illustrates the results of smoothing.

If the relationship appears linear, we can proceed to fit a linear model. If not, we can transform the response variable using a power transformation as described in Section 10.3.2, and replot. Alternatively, we can fit a polynomial model as described in Chapter 7.

In the case of two explanatory variables x_1 and x_2 , we can plot the response Y versus $\mathbf{x} = (x_1, x_2)'$ using a three-dimensional plot. Such plots are enhanced by dynamic rotation (spinning) and are available in standard computer packages. For more detail on the construction of such plots, see, for

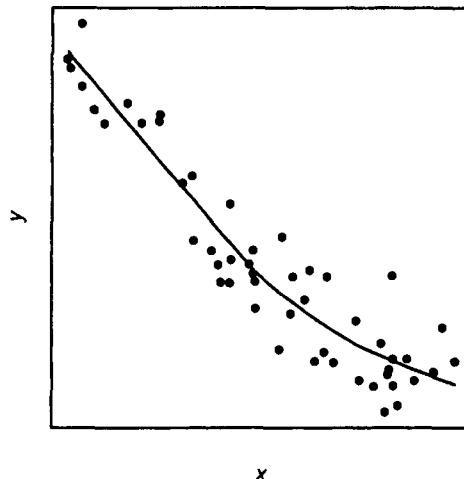


Fig. 10.1 Smoothing a plot of Y versus x .

example, Bekker et al. [1988], Cook and Weisberg [1994: p. 57; 1999: p. 185]. Interactive inspection of the plot will reveal if the regression surface is planar or curved. Once again, if the surface is not planar, we can either transform or fit a polynomial model.

When we have more than two explanatory variables, we can no longer visualize the surface directly, and our task is much harder. A useful plot is to plot residuals versus fitted values, which, as we saw in Section 10.2, are independent under normality assumptions. Thus, if the linear model is correct, the plot should display no pattern but should appear as a horizontal band of points, as shown in Figure 10.2.

A curved regression surface reveals itself as a curved plot; once again, interpretation of the plot is enhanced by smoothing. However, a disadvantage of the residuals versus fitted value plot is that the *nature* of the curvature is not revealed. Even more seriously, it is possible for the regression surface to be curved without the residuals versus fitted value plot revealing the curvature. Cook [1994: Example 7.1] gives an example of a regression where the regression surface is highly nonplanar, but the plot of residuals versus fitted values reveals no problems.

To reveal the nature of the curvature, we can use various forms of *partial residual plots*. These are plots of suitably modified residuals versus the explanatory variables. Suppose that the true regression surface is of the form

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 g(x_1) + \beta_2' \mathbf{x}_2, \quad (10.15)$$

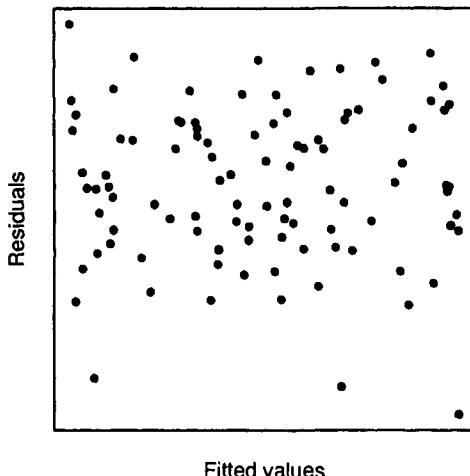


Fig. 10.2 Plotting residuals versus fitted values.

where $\mathbf{x} = (x_1, x_2')'$ and g is an unknown function. If we could discover the nature of g , we could replace x_1 with the transformed variable $x_1^* = g(x_1)$ and fit the model

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1^* + \beta_2' \mathbf{x}_2,$$

which would now be correct. Partial residual plots are designed to reveal the form of g . This type of plot was first considered by Ezekiel [1924] and subsequently by Ezekiel and Fox [1959] and Larsen and McCleary [1972], and is based on the following heuristic justification.

Suppose that we fit the linear model $E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2' \mathbf{x}_2$ when in fact (10.15) is the true model. Writing

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2' \mathbf{x}_2 + [\beta_1 g(x_1) - \beta_1 x_1 + \varepsilon] \\ &\approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2' \mathbf{x}_2 + e, \end{aligned}$$

we might expect that the residual e from the linear fit would approximate $\beta_1[g(x_1) - x_1] \approx \hat{\beta}_1[g(x_1) - x_1]$. This suggests that a plot of $e_i^\dagger = e_i + \hat{\beta}_1 x_{i1}$ [$\approx \hat{\beta}_1 g(x_{i1})$] versus x_{i1} might reveal the shape of g . The modified residuals e_i^\dagger are called *partial residuals*.

This plot has several interesting features. To begin with, a least squares line fitted through the origin of the plot will have slope $\hat{\beta}_1$, since the slope is

$$\frac{\sum_i x_{i1}(\hat{\beta}_1 x_{i1} + e_i)}{\sum_i x_{i1}^2} = \hat{\beta}_1 + \frac{\sum_i x_{i1} e_i}{\sum_i x_{i1}^2} = \hat{\beta}_1, \quad (10.16)$$

from $\mathbf{Xe} = \mathbf{X}(\mathbf{I}_p - \mathbf{H})\mathbf{Y} = \mathbf{0}$. Also, the residuals from this fit are just the ordinary residuals, since the former are just $e_i^* - \hat{\beta}_1 x_{i1} = e_i + \hat{\beta}_1 x_{i1} - \hat{\beta}_1 x_{i1} = e_i$.

Unfortunately, partial residual plots suffer from two drawbacks. First, the appearance of the plot may overemphasize the importance of the explanatory variable x_1 in the fit in the sense that the points in the plot are “overly close” to the fitted line. This point is developed further in Section 10.3.3, and an alternative plot is described. Second, and more important, if g is highly nonlinear, then the least squares estimates of the coefficients β_2 are not very good estimates and the heuristic argument sketched above breaks down.

We now examine this effect by using the formulas given in Example 3.10 in Section 3.11.2. For simplicity we shall assume that $p = 3$ and that the x -variables have been standardized to x^* -variables with zero means and unit lengths; for the general case, see Mansfield and Conerly [1987]. To reflect this change we use γ instead of β , and we apply the theory from Example 3.10. Thus

$$\hat{\gamma}_1 = \frac{1}{1 - r^2} (\mathbf{x}^{*(1)'} \mathbf{Y} - r \mathbf{x}^{*(2)'} \mathbf{Y}),$$

where $\mathbf{x}^{*(1)} = (x_{11}^*, \dots, x_{n1}^*)'$, $\mathbf{x}^{*(2)} = (x_{12}^*, \dots, x_{n2}^*)'$ and $r = \mathbf{x}^{*(1)'} \mathbf{x}^{*(2)}$ is the correlation between the columns. Using (3.52), after some algebra we get

$$\mathbf{H} = n^{-1} \mathbf{1}_n \mathbf{1}'_n + \mathbf{x}^{*(2)'} \mathbf{x}^{*(2)'} + \frac{1}{1 - r^2} (\mathbf{x}^{*(1)} - r \mathbf{x}^{*(2)}) (\mathbf{x}^{*(1)} - r \mathbf{x}^{*(2)}'). \quad (10.17)$$

Now suppose that

$$\mathbf{Y} = \gamma_0 \mathbf{1}_n + \gamma_1 \mathbf{g} + \gamma_2 \mathbf{x}^{*(2)} + \boldsymbol{\varepsilon},$$

where $\mathbf{g} = (g(x_{i1}^*), \dots, g(x_{in}^*))'$. Then, since \mathbf{H} is the projection onto $\mathbf{X}_s = (\mathbf{1}_n, \mathbf{x}^{*(1)}, \mathbf{x}^{*(2)})'$, we have $(\mathbf{I}_n - \mathbf{H})\mathbf{X}_s = \mathbf{0}$,

$$\begin{aligned} E[\mathbf{e}] &= E[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}] \\ &= (\mathbf{I}_n - \mathbf{H})(\gamma_0 \mathbf{1}_n + \gamma_1 \mathbf{g} + \gamma_2 \mathbf{x}^{*(2)}) \\ &= \gamma_1 (\mathbf{I}_n - \mathbf{H})\mathbf{g} \end{aligned}$$

and since $\mathbf{x}^{*(1)'} \mathbf{x}^{*(2)} = r$ and $\mathbf{x}^{*(j)'} \mathbf{1}_n = 0$ ($j = 1, 2$),

$$\begin{aligned} E[\hat{\gamma}_1] &= \frac{1}{1-r^2} (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' E[\mathbf{Y}] \\ &= \frac{1}{1-r^2} (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' (\gamma_0 \mathbf{1}_n + \gamma_1 \mathbf{g} + \gamma_2 \mathbf{x}^{*(2)}) \\ &= \frac{1}{1-r^2} \gamma_1 (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' \mathbf{g}. \end{aligned}$$

It follows from (10.17) that

$$\begin{aligned} E[\mathbf{e}^\dagger] &= E[\mathbf{e}] + \mathbf{x}^{*(1)} E[\hat{\gamma}_1] \\ &= \gamma_1 (\mathbf{I}_n - \mathbf{H})\mathbf{g} + \frac{1}{1-r^2} \gamma_1 \mathbf{x}^{*(1)} (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' \mathbf{g} \\ &= \gamma_1 (\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n') \mathbf{g} - \gamma_1 \left[\mathbf{x}^{*(2)'} \mathbf{x}^{*(2)}' \right. \\ &\quad \left. + \frac{1}{1-r^2} (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)}) (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' \right. \\ &\quad \left. - \frac{1}{1-r^2} \mathbf{x}^{*(1)} (\mathbf{x}^{*(1)} - r\mathbf{x}^{*(2)})' \right] \mathbf{g} \\ &= \gamma_1 (\mathbf{g} - \bar{g} \mathbf{1}_n) - \frac{1}{1-r^2} \gamma_1 \mathbf{x}^{*(2)} (\mathbf{x}^{*(2)} - r\mathbf{x}^{*(1)})' \mathbf{g}. \end{aligned}$$

This indicates that the plot does not reveal the shape of g , but rather the shape contaminated by additional terms. In particular, the plot will suffer from a large amount of contamination if the correlation between the columns is large (see also Berk and Booth [1995] and Mansfield and Conerly [1987]). On the other hand, if the correlation is small, then (apart from a constant, which is not important), the contamination is approximately $(\mathbf{x}^{*(2)'} \mathbf{g}) \mathbf{x}^{*(2)}$. This will be small if the correlation between \mathbf{g} and $\mathbf{x}^{*(2)}$ is small, as will most likely be the case, as the correlation between $\mathbf{x}^{*(1)}$ and $\mathbf{x}^{*(2)}$ is small. This analysis shows that the plot will do a good job of revealing the shape of g if the columns of the regression matrix are uncorrelated.

In an interesting paper, Cook [1993] describes other circumstances when partial residual plots will be effective. He assumes that the data (\mathbf{x}_i, Y_i) , $i =$

$1, \dots, n$ are a random sample from a p -variate distribution and that the conditional distribution of Y given \mathbf{x} is of the form

$$Y = \alpha_0 + \alpha_1 g(x_1) + \alpha'_2 \mathbf{x}_2 + \varepsilon, \quad (10.18)$$

where $E[\varepsilon|x_1, \mathbf{x}_2] = 0$.

Suppose that we fit a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta'_2 \mathbf{x}_2 + \varepsilon$$

to the data. What now does the least squares estimate $\hat{\beta}$ of $\beta = (\beta_0, \beta_1, \beta'_2)'$ actually estimate? It can be shown that $\hat{\beta}$ converges to

$$\delta = \{E[\mathbf{x}\mathbf{x}']\}^{-1} E[\mathbf{x}Y].$$

Since the partial residual plot for variable x_1 is a plot of e_i^\dagger versus $x_{1,i}$, and

$$e_i^\dagger = Y_i - \hat{\beta}_0 - \hat{\beta}'_2 \mathbf{x}_{i2} = (\alpha_0 - \hat{\beta}_0) + \alpha_1 g(x_{i1}) + (\alpha_2 - \hat{\beta}_2)' \mathbf{x}_{i2} + \varepsilon_i,$$

we cannot expect the plot to reveal the shape of g unless $\hat{\beta}_2$ is a good estimate of α_2 . Cook [1993] proves that this will be the case if $E[\mathbf{x}_2|x_1]$ is a linear function of x_1 , which can be checked by plotting the other explanatory variables against x_1 . However, if $E[\mathbf{x}_2|x_1] = m(x_1)$, say, where m is not linear, then the partial residual plot must be modified. Suppose now that we transform x_1 using m , and fit a linear model using the transformed x_1 and \mathbf{x}_2 , obtaining residuals e_i . If we form the modified residuals $e_i^\dagger = \hat{\beta}_1 m(x_{i1}) + e_i$, then the partial residual plot will reveal the shape of g . Plots modified in this way are called CERES plots (*combine conditional expectations and residuals*). The function m is not usually known but can be estimated by smoothing. For further details, see Cook [1993] and Cook and Weisberg [1994, 1999]. An earlier version of this modification is due to Mallows [1986], who assumed that the conditional expectation was a quadratic function.

In the case of just two explanatory variables, Berk and Booth [1995] show that the unmodified partial residual plot is just a two-dimensional view of the three-dimensional plot: if we rotate a three-dimensional plot of the data so that we are looking parallel to the fitted least squares plane and also parallel to the x_2, y plane, the resulting view is exactly the partial residual plot, apart from a shift in the e^\dagger direction. This characterisation helps us to understand how partial plots can fail to reveal the true curvature and why CERES plots are necessary – see the interesting examples in the Berk and Booth article.

10.3.2 Transforming to Remove Curvature

If the plots described in the preceding section suggest that the regression surface is curved, we can modify the model by considering a surface of the form

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 g_1(\mathbf{x}) + \cdots + \beta_r g_r(\mathbf{x}),$$

where g_1, \dots, g_r are functions chosen to linearize the regression surface. Very often, we will have $r = p - 1$ and $g_j(\mathbf{x}) = g_j(x_j)$, so that the original explanatory variables are each subjected to a transformation before refitting.

In the simple case $p = 2$, with just one explanatory variable, we fit the model

$$E[Y|x] = \beta_0 + \beta_1 g_1(x),$$

where g is a suitable transformation. Possible transformations include powers, polynomials, or logarithms. A flexible family of transformations, suitable when the values of the explanatory variable are all positive, is the family (Box and Cox [1964])

$$g(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log x, & \lambda = 0, \end{cases} \quad (10.19)$$

which is discussed further in Section 10.5.2. If this family is used, we can try different powers and repeatedly plot y versus $g(x, \lambda)$ until the plot is linear. Suitable interactive software, as described by Cook and Weisberg [1999], makes this easy. Alternatively, we can fit a polynomial in x , as described in Chapter 7.

For $p > 2$, we can use the CERES plots described in Section 10.3.1 to guide us in the choice of the functions g_j . For a more automatic guide, we can fit a *generalized additive model*.

Generalized Additive Models

These are discussed by Tibshirani and Hastie [1990] and allow us to fit the model

$$E[Y|x] = \beta_0 + g_1(x_1) + \dots + g_{p-1}(x_{p-1}). \quad (10.20)$$

The functions g_j are required to be smooth but are otherwise unspecified. Tibshirani and Hastie [1990] describe a method known as *backfitting* to obtain estimates of the g_j . The algorithm is described below.

Algorithm 10.1

Step 1: Fit a linear model by least squares, and estimate β_0 by its least squares estimate and $g_j(x_j)$ by $\hat{g}_j(x_j) = \hat{\beta}_j x_j$.

Step 2: For $j = 1, 2, \dots, p - 1$, compute the residuals

$$e_{ij} = Y_i - \hat{\beta}_0 - \sum_{l \neq j} \hat{g}_l(x_{il})$$

and then smooth the plot of e_{ij} versus x_{ij} using one of the smoothers described in Section 7.2.3. The resulting smooth function is taken to be \hat{g}_j .

Step 3: Repeat step 2 until there is no further change.

By plotting the estimate of $g_j(x_{ij})$ versus x_{ij} , we can get an idea of the form of g_j . This will help us select an appropriate power or polynomial transformation.

10.3.3 Adding and Deleting Variables

In Section 10.3.2 we mentioned that partial residual plots can “oversell” the importance of an explanatory variable. In this section we explain how this can happen and describe an alternative type of plot, the *added variable plot*, which gives a better indication of the contribution that each explanatory variable makes to the fit. We assume, once again, that a constant term has been included in the model.

Suppose that there is a reasonable straight-line relationship in the partial residual plot for x_j . The strength of the relationship is measured by the correlation \tilde{r}_j between x_{ij} and the partial residuals $e_{ij}^\dagger = e_i + \hat{\beta}_j x_{ij}$. By Exercises 10b, No. 1,

$$\tilde{r}_j^2 = \frac{\hat{\beta}_j^2 \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{\text{RSS} + \hat{\beta}_j^2 \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}. \quad (10.21)$$

Now consider the variance inflation factors

$$\text{VIF}_j = \sigma^{-2} \text{Var}[\hat{\beta}_j] \times \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2,$$

which were introduced in Section 9.7.2. We saw there that VIF_j will be large if x_j can be predicted accurately from the other explanatory variables. If F_j is the F -statistic [cf. (4.13)] for testing the hypothesis that $\beta_j = 0$, then

$$\begin{aligned} F_j &= \frac{\hat{\beta}_j^2}{S^2 (\mathbf{X}' \mathbf{X})_{j+1,j+1}^{-1}} \\ &= \frac{\hat{\beta}_j^2 \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{S^2 \times \text{VIF}_j}. \end{aligned}$$

Hence from (10.21), with $\text{RSS} = (n-p)S^2$,

$$\tilde{r}_j^2 = \frac{F_j \times \text{VIF}_j}{(n-p) + F_j \times \text{VIF}_j}. \quad (10.22)$$

This shows that even if the variable x_j makes very little contribution to the fit (in the sense of having a small value of F_j), the correlation can be close to 1 if the VIF is sufficiently large. Put another way, if x_j can be well predicted from the other explanatory variables, then the partial residual plot may show a strong linear relationship indicated by a large \tilde{r}_j^2 , and mislead us as to the importance of including the variable, even if the contribution of x_j to the fit is negligible. This is illustrated numerically in Exercises 10b, No. 3.

We can get a better picture of the value of adding the variable x_j to the regression by examining the relationship between x_j , on the one hand, and the residuals from a fit excluding the variable x_j , on the other. If $\mathbf{X} = (\mathbf{1}_n, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)})$, and $\mathbf{X}^{(j)}$ is \mathbf{X} with $\mathbf{x}^{(j)}$ omitted, then these residuals, $\mathbf{e}^{(j)}$ say, are given by

$$\mathbf{e}^{(j)} = (\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y},$$

where $\mathbf{P}_j = \mathbf{X}^{(j)}(\mathbf{X}^{(j)'}\mathbf{X}^{(j)})^{-1}\mathbf{X}^{(j)'}$. When considering whether to add x_j to the regression, we must assess how well we can predict the residuals $\mathbf{e}^{(j)}$ by using x_j . However, since x_j can be predicted partly by the other explanatory variables, we are really interested in how well the residuals $\mathbf{e}^{(j)}$ can be predicted by the part of x_j that is not predicted by the other explanatory variables. This suggests looking at the relationship between $\mathbf{e}^{(j)}$ and the “ x_j residual” $(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}$. The relationship is assessed by plotting $\mathbf{e}^{(j)} = (\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}$ versus $(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}$, with a strong linear relationship indicating that the variable should be included in the regression. This plot, called an *added variable plot*, has been discussed by many authors: see Cook [1998: p. 136] for some history and references.

These plots share some of the properties of the partial residual plot. First, assuming that a constant term is fitted in the original model, the least squares line through the origin and fitted to the plot has slope $\hat{\beta}_j$. To see this, we first note that by Section 6.2, the least squares estimate of the slope of a line through the origin for data points (x_i, y_i) is $\mathbf{y}'\mathbf{x}/\mathbf{x}'\mathbf{x}$. Setting $\mathbf{y} = \mathbf{e}^{(j)}$ and $\mathbf{x} = (\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}$, the slope of the least squares line through the plot can be written as

$$\begin{aligned} \frac{\mathbf{e}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}}{[(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]'\mathbf{x}^{(j)}} &= \frac{\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}}{\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}} \\ &= \hat{\beta}_j, \end{aligned} \quad (10.23)$$

by (3.32). Also, from (3.28) with $(\mathbf{X}, \mathbf{Z}) \rightarrow \mathbf{X}, \mathbf{Z} \rightarrow \mathbf{X}^{(j)}$ and $\hat{\gamma}_G \rightarrow \hat{\beta}_j$, we see that the vector of residuals \mathbf{e} from the full model is given by

$$\begin{aligned} \mathbf{e} &= (\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= (\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} - (\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}\hat{\beta}_j \\ &= \mathbf{e}^{(j)} - (\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}\hat{\beta}_j \end{aligned} \quad (10.24)$$

(or in terms of \mathbf{y} and \mathbf{x} , $\mathbf{e} = \mathbf{y} - \mathbf{x}\hat{\beta}_j$) and the residuals from the full fit are exactly the residuals from fitting a least squares line through the origin.

Now let us calculate the squared correlation, ρ_j^2 say, between the quantities in the added variable plot. We first note that when fitting a model with a constant term, the sum (and therefore the mean) of the residuals is zero [cf. (4.29)]. Applying this to the residual vectors \mathbf{x} and \mathbf{y} , we have $\bar{x} = \bar{y} = 0$,

and using the fact that $\mathbf{I}_n - \mathbf{P}_j$ is symmetric and idempotent, we have

$$\begin{aligned}\rho_j^2 &= \frac{(\mathbf{y}'\mathbf{x})^2}{(\mathbf{y}'\mathbf{y})(\mathbf{x}'\mathbf{x})} \\ &= \frac{[\mathbf{e}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]^2}{[\mathbf{e}^{(j)'}\mathbf{e}^{(j)}][\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]} \\ &= \frac{[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]^2}{[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}][\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]} \\ &= \frac{\hat{\beta}_j^2[\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}]}{\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}},\end{aligned}$$

by (10.23). Premultiplying (10.24) by \mathbf{Y}' , we can write the denominator as

$$\begin{aligned}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} + \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}\hat{\beta}_j \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} + \hat{\beta}_j^2[\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{x}^{(j)}].\end{aligned}\quad (10.25)$$

Also, by A.9.2,

$$(\mathbf{X}'\mathbf{X})_{j+1,j+1}^{-1} = [\mathbf{x}^{(j)'}(\mathbf{I}_p - \mathbf{P}_j)\mathbf{x}^{(j)}]^{-1}$$

(with $j+1$ instead of j because of β_0), so that the F -statistic for testing $\beta_j = 0$ can be written

$$F_j = \frac{\hat{\beta}_j^2\mathbf{x}^{(j)'}(\mathbf{I}_p - \mathbf{P}_j)\mathbf{x}^{(j)}}{S^2}.$$

Using this and (10.25), we find that

$$\rho_j^2 = \frac{F_j}{n-p+F_j}. \quad (10.26)$$

Thus, the linear relationship displayed in an added variable plot will be strong only if the F -statistic is large (i.e., only if the variable x_j makes a significant contribution to the regression). This is in marked contrast to the partial residual plot, as may be seen by comparing (10.22) and (10.26).

Variables such as time order that are implicitly rather than explicitly recorded in the data often have a strong effect on regression analyses. Joiner [1981] gives some sobering examples of this. The effect of such “lurking variables” (variables that are not included in the fit) should always be examined, and added variable plots are a good way to do this.

EXERCISES 10b

1. (a) Show that provided that the regression matrix \mathbf{X} contains a column of 1's, $\sum_{i=1}^n (x_{ij} - \bar{x}_j)e_i = 0$ for each j .
- (b) Use part (a) to prove (10.21).

The following exercises require familiarity with a computing package such as S-PLUS, R, or Matlab. We provide some R code to assist readers.

2. (a) Generate a set of regression data according to the model

$$Y_i = 1 + x_{i1} + 2 \operatorname{sign}(x_{i2})|x_{i2}|^{1/3} + \varepsilon_i \quad (i = 1, \dots, 100),$$

where x_{i1} and x_{i2} are sampled from independent $N(0, 1)$ distributions, and the errors ε_i are sampled from $N(0, \sigma^2)$, where $\sigma = 0.25$. This may be done in S-PLUS or R by the code fragment

```
eps<-rnorm(100, sd=0.25)
x1<-rnorm(100)
x2<-rnorm(100)
y<- 1 + x1 + 2*sign(x2)*abs(x2)^(1/3) + eps
```

- (b) Construct the partial residual plot for x_2 . How well does the curve in the plot reveal the transformation required to linearize the regression surface? Use the code

```
reg.stuff<-lm(y~x1+x2)
betahat2<-coef(reg.stuff)[3]
estar2<-betahat2*x2 + residuals(reg.stuff)
plot(x2, estar2)
```

- (c) Now construct another set of data, the same as before but with x_2 having correlation $\rho = 0.95$ with x_1 . Use the same code, but define x_2 with the lines

```
rho<-0.95
x2<-rho*x1 + sqrt(1-rho^2)*rnorm(100)
```

- (d) Construct the partial residual plot for the new x_2 . What do you notice?

3. Construct a set of data as in Exercise 2 but with

$$Y_i = 1 + x_{i1} + \varepsilon_i \quad (i = 1, \dots, 100),$$

so that the coefficient β_2 is zero. Set the correlation ρ between x_1 and x_2 at 0.999. What does the partial residual plot suggest about the importance of x_2 in the model? Generate a new set with $\rho = 0$. What does the plot now suggest? What does this example teach you about using partial residual plots to assess the importance of adding variables?

4. Repeat Exercise 3 using added variable plots instead of partial residual plots. Do added variable plots have a problem with correlated data? Use the R code

```
resy<-residuals(lm(y~x1))
```

```
resx2<-residuals(lm(x2~x1))
plot(resy,resx2)
```

to draw the added variable plot.

5. Construct a set of data as in Exercise 2(a) but with x_2 being generated by

$$x_{i2} = x_{i1}^2 + u_i \quad (i = 1, \dots, 100),$$

where the u_i 's are $N(0, 0.1)$. This will create a situation where the partial residual plot would be expected to fail, according to the arguments of Cook [1993] since $E[x_2|x_1] = x_1^2$, which is far from linear. Does the plot, in fact, fail? Repeat with x_2 a linear function of x_1 . What happens now?

10.4 NONCONSTANT VARIANCE AND SERIAL CORRELATION

10.4.1 Detecting Nonconstant Variance

We will assume that the usual model

$$Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$$

holds, with the errors ε_i being independent and normally distributed with zero mean. However, suppose that in place of the standard assumption

$$\text{var}[\varepsilon_i] = \sigma^2 \quad (i = 1, \dots, n),$$

we have, instead,

$$\text{var}[\varepsilon_i] = \sigma_i^2,$$

where the variances σ_i^2 may depend either on the mean $E[Y_i] = \mathbf{x}'_i \boldsymbol{\beta}$, and possibly other parameters, or on a vector of (possibly additional) explanatory variables \mathbf{z}_i . As discussed at the end of Section 3.10, the least squares estimates of $\boldsymbol{\beta}$ may not be efficient if the variances σ_i^2 are not equal. We need to check the variances for equality and, if necessary, use more efficient estimation methods.

In this section we assume that

$$\sigma_i^2 = w(\mathbf{z}_i, \lambda), \tag{10.27}$$

where \mathbf{z}_i is a vector of known explanatory variables for the i th observation and w is a variance function with the property that for some λ_0 , $w(\mathbf{z}, \lambda_0)$ does not depend on \mathbf{z} . In (10.27), the form of w is known, but the value of λ is not. For example, a popular choice is

$$w(\mathbf{z}, \lambda) = \exp(\mathbf{z}' \lambda),$$

where $\mathbf{z} = (1, z_1, \dots, z_k)$. In this case $\lambda_0 = (\lambda_0, 0, \dots, 0)'$ and, for this value,

$$\text{var}[Y_i] = w(\mathbf{z}, \lambda_0) = e^{\lambda_0} = \sigma^2,$$

say. It should be noted that the form (10.27) does not include the case where the variance function is a function of the mean. This case is considered in Section 10.4.2.

The assumptions made above completely specify the distribution of the responses, so we can test if the variances are homogeneous by testing the parametric hypothesis $H_0 : \lambda = \lambda_0$. In this section we discuss some ways of doing this, along with some informal graphical methods that can be used to detect nonconstant variances.

The residuals \mathbf{e} from a least squares fit contain information about the variances. If $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$, then

$$\begin{aligned}\text{Var}[\mathbf{e}] &= \text{Var}[(\mathbf{I}_n - \mathbf{H})\mathbf{e}] \\ &= (\mathbf{I}_n - \mathbf{H})\Sigma(\mathbf{I}_n - \mathbf{H}),\end{aligned}$$

so that

$$\text{var}[e_i] = (1 - h_i)^2 \sigma_i^2 + \sum_{k:k \neq i} h_{ik}^2 \sigma_k^2. \quad (10.28)$$

Usually, $h_{ik} \ll h_i$ for $k \neq i$, so that very often large variances are indicated by large residuals, although this will not be the case for high leverage points. We note that $E[e_i] = 0$, so that $\text{var}[e_i] = E[e_i^2]$.

The quantities

$$b_i = \frac{e_i^2}{1 - h_i}$$

are useful when making various kinds of graphics displays. When the variances are all equal to σ^2 , say, then

$$\begin{aligned}E[b_i] &= (1 - h_i)\sigma^2 + \sum_{k:k \neq i} \frac{h_{ik}^2}{1 - h_i} \sigma^2 \\ &= \sigma^2,\end{aligned}$$

since the idempotence of $\mathbf{I}_n - \mathbf{H}$ implies that

$$(1 - h_i)^2 + \sum_{k:k \neq i} h_{ik}^2 = 1 - h_i. \quad (10.29)$$

Thus, when the variances are in fact constant, the b_i 's have constant expectation.

We note that even if the variances are unequal, the fitted values $\hat{Y}_i = \mathbf{x}'_i \hat{\beta}$ still have expectation $E[Y_i]$. However, observations with large means often have large variances also. Thus, plotting the b_i 's (or equivalently, the squared internally Studentized residuals) against the fitted values should result in a

wedge-shaped display if the variances increase with the means. A typical display is shown in Figure 10.3(a). Another popular plot involves plotting the b_i 's versus the explanatory variables; this is interpreted in the same way. A smoother passed through the plot may reveal the relationship between the means and the variances. Alternatively, the raw residuals may be plotted versus the fitted values. In this case, a fan-shaped pattern, as shown in Figure 10.3(b), indicates variances increasing with the means.

Other plots have been proposed based on particular choices of the function w . Expanding w in a Taylor series, we get

$$w(\mathbf{z}_i, \boldsymbol{\lambda}) \approx w(\mathbf{z}_i, \boldsymbol{\lambda}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)' \dot{w}(\mathbf{z}_i, \boldsymbol{\lambda}_0),$$

where $\dot{w} = \partial w / \partial \boldsymbol{\lambda}$. Then, using (10.28) and (10.29), we obtain

$$\begin{aligned} E[b_i] &= (1 - h_i)w(\mathbf{z}_i, \boldsymbol{\lambda}) + \sum_{k:k \neq i} \frac{h_{ik}^2 w(\mathbf{z}_i, \boldsymbol{\lambda})}{1 - h_i} \\ &\approx w(\mathbf{z}_i, \boldsymbol{\lambda}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)' \left\{ (1 - h_i)\dot{w}(\mathbf{z}_i, \boldsymbol{\lambda}) + \sum_{k:k \neq i} \frac{h_{ik}^2 \dot{w}(\mathbf{z}_i, \boldsymbol{\lambda})}{1 - h_i} \right\}. \end{aligned}$$

Cook and Weisberg [1983] suggest plotting b_i versus the quantities in the braces $\{\}$. Departures from a constant variance show up as an approximately linear trend. A disadvantage of this plot is that the b_i do not have constant

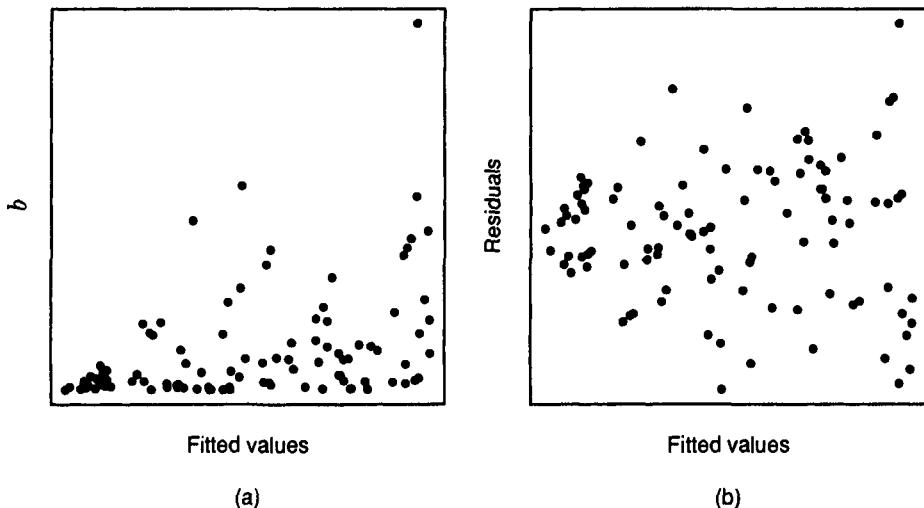


Fig. 10.3 Patterns resulting from the variances being a function of the means.

variance. Verbyla [1993] observes that the quantities $\log[b_i/(1-h_i)]$ do have an approximately constant variance, and moreover, have approximately a linear regression on \mathbf{z}_i if the variances σ_i^2 are constant. The added variable plots described in Section 10.3.3 can therefore be used to assess the linearity of this relationship. If all plots are linear, the variances are constant.

These graphical methods can be backed up by formal tests of the hypothesis that $\lambda = \lambda_0$, based on standard asymptotic tests. Under the normal distributional assumptions outlined above, the log likelihood $l(\beta, \lambda)$ is

$$\begin{aligned} l(\beta, \lambda) &= c - \frac{1}{2} \{ \log \det(\Sigma) + (\mathbf{y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \} \\ &= c - \frac{1}{2} \left\{ \sum_{i=1}^n \log w_i + \sum_{i=1}^n \frac{(y_i - \mathbf{x}'_i \beta)^2}{w_i} \right\}, \end{aligned} \quad (10.30)$$

where $w_i = w(\mathbf{z}_i, \lambda)$ and c is a constant. Using A.8.1, the score function is given by

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \mathbf{X}' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{X}' \Sigma^{-1} \varepsilon, \end{aligned} \quad (10.31)$$

and

$$\frac{\partial l}{\partial \lambda} = -\frac{1}{2} \left[\sum_{i=1}^n \left\{ \frac{1}{w_i} - \frac{(y_i - \mathbf{x}'_i \beta)^2}{w_i^2} \right\} \frac{\partial w_i}{\partial \lambda} \right]. \quad (10.32)$$

The maximum likelihood estimates can be obtained by the following algorithm.

Algorithm 10.2

Step 1: Put $\lambda = \lambda_0$.

Step 2: Compute $\hat{\beta}$ as $(\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y}$, using a weighted least squares program.

Step 3: Solve $\partial l / \partial \lambda|_{\beta=\hat{\beta}} = 0$ for λ .

Step 4: Repeat steps 2 and 3 until convergence.

EXAMPLE 10.3 In the special case $w(\mathbf{z}, \lambda) = \exp(\mathbf{z}' \lambda)$, where \mathbf{z} is the vector $(1, z_1, \dots, z_k)$, step 3 can be implemented as a least squares calculation. For this $w, \partial w_i / \partial \lambda = w_i \mathbf{z}_i$ so that

$$\frac{\partial l}{\partial \lambda} = -\frac{1}{2} \sum_{i=1}^n \left(1 - \frac{\varepsilon_i^2}{w_i} \right) \mathbf{z}_i. \quad (10.33)$$

Let $d_i = \varepsilon_i^2/w_i$, $\mathbf{d} = (d_1, \dots, d_n)'$, and let \mathbf{Z} be the matrix whose i th row is \mathbf{z}_i . Then (10.33) can be written

$$\frac{\partial l}{\partial \lambda} = \frac{1}{2} \mathbf{Z}'(\mathbf{d} - \mathbf{1}_n).$$

To calculate the information matrix, note that

$$\begin{aligned} \text{Var}\left[\frac{\partial l}{\partial \beta}\right] &= \text{Var}[\mathbf{X}'\Sigma^{-1}\varepsilon] \\ &= \mathbf{X}'\Sigma^{-1}\text{Var}[\varepsilon]\Sigma^{-1}\mathbf{X} \\ &= \mathbf{X}'\Sigma^{-1}\mathbf{X} \end{aligned}$$

and

$$\begin{aligned} \text{Var}\left[\frac{\partial l}{\partial \lambda}\right] &= \text{Var}\left[\frac{1}{2}\mathbf{Z}'(\mathbf{d} - \mathbf{1}_n)\right] \\ &= \frac{1}{4}\mathbf{Z}'\text{Var}[\mathbf{d}]\mathbf{Z}. \end{aligned}$$

Since ε_i has variance w_i , the elements of \mathbf{d} are independently and identically distributed as χ_1^2 with variance 2, so that $\text{Var}[\mathbf{d}] = 2\mathbf{I}_n$, and hence $\text{Var}[\partial l/\partial \lambda] = \frac{1}{2}\mathbf{Z}'\mathbf{Z}$. Moreover, since $E[\varepsilon_i\varepsilon_j^2] = 0$ ($i \neq j$) and $E[\varepsilon_i^3] = 0$, we have $\text{Cov}[\varepsilon, \mathbf{d}] = \mathbf{0}$ and

$$\begin{aligned} \text{Cov}\left[\frac{\partial l}{\partial \beta}, \frac{\partial l}{\partial \lambda}\right] &= \frac{1}{2}\mathbf{X}'\Sigma^{-1}\text{Cov}[\varepsilon, \mathbf{d}]\mathbf{Z} \\ &= \mathbf{0}. \end{aligned}$$

Thus, from (3.19) the (expected) information matrix is

$$\mathbf{I}(\beta, \lambda) = \begin{pmatrix} \mathbf{X}'\Sigma^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{Z}'\mathbf{Z} \end{pmatrix}.$$

To solve the likelihood equations we can use Fisher scoring (cf. A.14). The updating equations are (m denotes the iteration)

$$\begin{aligned} \hat{\beta}_{(m+1)} &= (\mathbf{X}'\Sigma_{(m)}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{(m)}^{-1}\mathbf{Y}, \\ \lambda_{(m+1)} &= \lambda_{(m)} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{d} - \mathbf{1}_n), \\ \Sigma_{(m+1)} &= \text{diag}[w(\mathbf{z}_1, \lambda_{(m)}), \dots, w(\mathbf{z}_n, \lambda_{(m)})]. \end{aligned} \quad (10.34)$$

We note that (10.34) can be written as

$$\mathbf{Z}'\mathbf{Z}\lambda_{(m+1)} = \mathbf{Z}'(\mathbf{d} - \mathbf{1}_n + \mathbf{Z}\lambda_{(m)}),$$

which are the normal equations for a formal regression of $\mathbf{d} - \mathbf{1}_n + \mathbf{Z}\lambda_{(m)}$ on \mathbf{Z} , and so can be solved using a least squares program. This form of Algorithm 10.2 is due to Aitkin [1987]. \square

Testing $\lambda = \lambda_0$

We can test this hypothesis using a likelihood ratio (LR) test or a score test, as described, for example, in Cox and Hinkley [1974]. For the LR test, consider maximizing the log-likelihood (10.30) under the hypothesis $H_0 : \lambda = \lambda_0$. Then $w_i(\mathbf{z}, \lambda_0) = \sigma^2$, and the likelihood is just the usual regression likelihood discussed in Section 3.5, which is maximized when β is the ordinary least squares estimate $\hat{\beta}_{OLS}$ and $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n e_i^2$. The maximum value of the log-likelihood (10.30) under the null hypothesis is, from (4.1),

$$l(\hat{\beta}_{OLS}, \hat{\sigma}^2) = c - \frac{1}{2}[n \log \hat{\sigma}^2 + n],$$

while the unrestricted maximum is $l(\hat{\beta}, \hat{\lambda})$, where $\hat{\beta}$ and $\hat{\lambda}$ are calculated using Algorithm 10.2. The LR test statistic is then

$$LR = -2[l(\hat{\beta}_{OLS}, \hat{\sigma}^2) - l(\hat{\beta}, \hat{\lambda})],$$

which, asymptotically, has a χ_k^2 distribution under H_0 .

Alternatively, we can use a score test. This has the advantage that we do not have to calculate the unrestricted maximum likelihood estimates. For a general weight function w , we have from (10.32) that

$$\frac{\partial l}{\partial \lambda} = -\frac{1}{2} \sum_{i=1}^n \left(\frac{1}{w_i} - \frac{\varepsilon_i^2}{w_i^2} \right) \frac{\partial w_i}{\partial \lambda}. \quad (10.35)$$

Writing \mathbf{D} for the matrix whose i th row is $\partial w_i / \partial \lambda$, setting $\mathbf{u} = (u_1, \dots, u_n)$, where $u_i = e_i^2 / \hat{\sigma}^2$ and noting that $\hat{w}_i = \hat{\sigma}^2$, we get

$$\begin{aligned} \left. \frac{\partial l}{\partial \lambda} \right|_{H_0} &= -\frac{1}{2} \sum_{i=1}^n \frac{\hat{\sigma}^2 - e_i^2}{\hat{\sigma}^4} \frac{\partial w_i}{\partial \lambda} \\ &= \frac{1}{2\hat{\sigma}^2} \mathbf{D}'(\mathbf{u} - \mathbf{1}_n). \end{aligned} \quad (10.36)$$

Differentiating and taking expectations (see Exercises 10c, No. 1, at the end of Section 10.4.4) we see that the information matrix under the hypothesis is

$$\begin{aligned} \mathbf{I}(H_0) &= \begin{pmatrix} \mathbf{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\lambda\lambda} \end{pmatrix} \\ &= \begin{pmatrix} \hat{\sigma}^{-2} \mathbf{X}' \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \hat{\sigma}^{-4} \mathbf{D}' \mathbf{D} \end{pmatrix}. \end{aligned} \quad (10.37)$$

The score test statistic is

$$\left(\left. \frac{\partial l}{\partial \lambda} \right|_{H_0} \right)' \{ \mathbf{I}_{\lambda\lambda} \}^{-1} \left(\left. \frac{\partial l}{\partial \lambda} \right|_{H_0} \right).$$

By substituting (10.36) and (10.37), the score statistic can be written

$$\frac{1}{2}(\mathbf{u} - \mathbf{1}_n)' \mathbf{D} (\mathbf{D}' \mathbf{D})^{-1} \mathbf{D}' (\mathbf{u} - \mathbf{1}_n).$$

We note that apart from the factor $1/2$, the expression above is the regression sum of squares for the formal regression of $\mathbf{u} - \mathbf{1}_n$ on \mathbf{D} , so it can be calculated easily with a regression program. This result is due to Cook and Weisberg [1983]. If $w(\mathbf{z}, \lambda) = \exp(\mathbf{z}' \lambda)$ (see Example 10.3), then $\mathbf{D} = \hat{\sigma}^2 \mathbf{Z}$ under H_0 , so that the test statistic is $\frac{1}{2}(\mathbf{u} - \mathbf{1}_n)' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{u} - \mathbf{1}_n)$.

Alternatives to these tests can be based on the *restricted likelihood*, which is obtained by considering the marginal distribution of a set of linear contrasts $\mathbf{Q}' \mathbf{Y}$ that are orthogonal to $\mathcal{C}(\mathbf{X})$ and thus do not involve β . If we take \mathbf{Q} to be a n by $n - p$ matrix whose columns form an orthonormal basis for $\mathcal{C}(\mathbf{X})^\perp$, the orthogonal complement of $\mathcal{C}(\mathbf{X})$, then (by B.1.3), $\mathbf{Q}' \mathbf{Q} = \mathbf{I}_n$ and $\mathbf{Q} \mathbf{Q}' = \mathbf{I}_n - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$. Using (1.7) and Theorem 2.2, we see that $\mathbf{Q}' \mathbf{Y}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{Q}' \Sigma \mathbf{Q}$ (since $\mathbf{Q}' \mathbf{X} \beta = \mathbf{0}$). The marginal (restricted) log likelihood based on this density is therefore

$$c - \frac{1}{2} \{ \log \det(\mathbf{Q}' \Sigma \mathbf{Q}) + \mathbf{y}' \mathbf{Q} (\mathbf{Q}' \Sigma \mathbf{Q})^{-1} \mathbf{Q}' \mathbf{y} \},$$

which depends on λ but not on β [since $\mathbf{Q}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{Q}'\mathbf{y}$]. Using Exercises 10c, No. 2, at the end of Section 10.4.4, we get

$$\mathbf{Q} (\mathbf{Q}' \Sigma \mathbf{Q})^{-1} \mathbf{Q}' = \Sigma^{-1} - \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \quad (10.38)$$

and

$$\det(\mathbf{Q}' \Sigma \mathbf{Q}) = \det(\Sigma) \det(\mathbf{X}' \Sigma^{-1} \mathbf{X}) / \det(\mathbf{X}' \mathbf{X}), \quad (10.39)$$

so that we can write the restricted log likelihood as

$$\begin{aligned} l_R(\lambda) &= c - \frac{1}{2} \{ \log \det(\Sigma) + \log \det(\mathbf{X}' \Sigma^{-1} \mathbf{X}) - \log \det(\mathbf{X}' \mathbf{X}) \\ &\quad + \mathbf{y}' [\Sigma^{-1} - \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1}] \mathbf{y} \}. \end{aligned} \quad (10.40)$$

In the case where $w(\mathbf{z}, \lambda) = \exp(\mathbf{z}' \lambda)$, Verbyla [1993] shows that the score function corresponding to this likelihood is $\frac{1}{2} \mathbf{Z}' (\Sigma^{-1} \mathbf{d} - \mathbf{1}_n + \mathbf{g})$ and the information matrix is $\frac{1}{2} \mathbf{Z}' \mathbf{V} \mathbf{Z}$. In these formulas, the elements of the vector \mathbf{g} are the diagonal elements of the hat matrix

$$\mathbf{G} = \Sigma^{-1/2} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1/2}$$

corresponding to the weighted regression, and \mathbf{V} has elements $v_{ij} = g_{ij}$ for $i \neq j$ and $v_{ii} = (1 - g_{ii})^2$. The restricted maximum likelihood (REML) estimate $\bar{\lambda}$ of λ is also obtained by Fisher scoring (A.14), using the updating equation

$$\bar{\lambda}_{(m+1)} = \bar{\lambda}_{(m)} + (\mathbf{Z}' \mathbf{V} \mathbf{Z})^{-1} \mathbf{Z}' (\Sigma^{-1} \mathbf{d} - \mathbf{1}_n + \mathbf{g}).$$

Under the null hypothesis $\lambda = \lambda_0$, we have $\Sigma = \sigma^2 \mathbf{I}_n$ and the estimates reduce to those for ordinary least squares. The REML estimate of λ is therefore

$\bar{\lambda}_0 = (\log \hat{\sigma}^2, 0, \dots, 0)'$, and the restricted likelihood has maximum value

$$\begin{aligned} l_R(\bar{\lambda}_0) &= c - \frac{1}{2}[n \log \hat{\sigma}^2 + \log \det(\hat{\sigma}^{-2} \mathbf{X}' \mathbf{X}) - \log \det(\mathbf{X}' \mathbf{X}) \\ &\quad + \hat{\sigma}^{-2} \mathbf{y}' (\mathbf{I}_n - \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') \mathbf{y}] \\ &= c - \frac{1}{2}[(n-p) \log \hat{\sigma}^2 + n]. \end{aligned}$$

The LR test statistic is $-2[l_R(\bar{\lambda}_0) - l_R(\bar{\lambda})]$ and is asymptotically χ_k^2 under the null hypothesis [k degrees of freedom because $\mathbf{z} = (1, z_1, \dots, z_k)'$].

The score test based on the restricted likelihood uses the statistic

$$\frac{1}{2}(\mathbf{u} - \mathbf{1}_n - \mathbf{h})' \mathbf{Z}(\mathbf{Z}' \mathbf{V}_0 \mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{u} - \mathbf{1}_n - \mathbf{h}),$$

where \mathbf{h} is the vector of ordinary hat matrix diagonals and \mathbf{V}_0 is defined like \mathbf{V} except that the elements of \mathbf{H} are used instead of \mathbf{G} . This is because \mathbf{G} reduces to \mathbf{H} under the null hypothesis of equal variances.

Restricted likelihoods were introduced by Patterson and Thompson [1971] in the context of variance component models. Several other variations of these tests have been proposed, and Lyon and Tsai [1996] describe and compare these. They found that the score test based on the full likelihood is better than the LR test, but the former is sensitive to departures from normality. The LR test is anticonservative even for normal errors. The score test based on the restricted likelihood has better power than that based on the full likelihood, at least for the cases considered by Lyon and Tsai.

10.4.2 Estimating Variance Functions

We saw in the preceding section how to use variance functions to construct tests for variance homogeneity using likelihood techniques. In this section we concentrate on how to allow for heterogeneity when estimating the regression coefficients.

If the weights are known, we saw in Section 3.10 that the optimal estimate is the weighted least squares (WLS) estimate

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Y}, \quad (10.41)$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. In practice, the weights are unknown, but the estimate (10.41) serves as a useful benchmark. If the variances are known functions of unknown parameters, say

$$\sigma_i^2 = w(\mathbf{z}_i, \lambda, \boldsymbol{\beta}), \quad (10.42)$$

we can use methods similar to those in the preceding section. These variances, however, are a little more general than those in the preceding section, since they can now depend on the regression coefficients as well. This is to allow for the possibility that the variance depends on the mean, and possibly on other parameters, as in the power relationship

$$\sigma_i^2 = \lambda_1 (\mathbf{x}'_i \boldsymbol{\beta})^{\lambda_2},$$

where the variance is proportional to a power of the mean. This, of course, assumes that the means are all positive.

Given (10.42), equation (10.41) suggests the following general algorithm for computing an estimate of β .

Algorithm 10.3

Step 1: Obtain an estimate β^* of β : for example, by using the ordinary least squares estimate.

Step 2: With β fixed at β^* , obtain an estimate λ^* of λ .

Step 3: Compute $\hat{\Sigma}$ as $(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{Y}$, where

$$\hat{\Sigma} = \text{diag}[w(\mathbf{z}_1, \lambda^*, \beta^*), \dots, w(\mathbf{z}_n, \lambda^*, \beta^*)],$$

using a weighted least squares program.

Step 4: Repeat steps 2 and 3 until convergence.

To describe the algorithm fully, we need to give more detail for step 2. Carroll and Ruppert [1988] and Carroll and Davidian [1987] have several suggestions, which we discuss below.

Pseudolikelihood

This involves applying the likelihood methods of the preceding section, treating β^* as fixed. The term *pseudolikelihood* refers to the fact that β^* is not the maximum likelihood estimate (MLE), so that this process is not the same as a full likelihood analysis. However, Carroll and Davidian show that if the variance does not depend on the mean, then the algorithm will converge to the MLE, so that the method is exactly equivalent to the method described in the preceding section.

Regressing Residuals

Treating $e_i \approx \varepsilon_i$, then $E[e_i^2] \approx \text{var}[\varepsilon_i] = w(\mathbf{z}_i, \lambda, \beta^*)$, and we can obtain an estimate of λ by solving the nonlinear least squares problem

$$\min_{\lambda} \sum_{i=1}^n [e_i^2 - w(\mathbf{z}_i, \lambda, \beta^*)]^2.$$

Since $\text{var}[e_i^2] \approx \text{var}[\varepsilon_i^2] = 2w(\mathbf{z}_i, \lambda, \beta)^2$ (by A.13.2) when the data are normally distributed, an alternative method is to solve the weighted least squares problem

$$\min_{\lambda} \sum_{i=1}^n \frac{[e_i^2 - w(\mathbf{z}_i, \lambda, \beta^*)]^2}{w(\mathbf{z}_i, \lambda, \beta^*)^2}.$$

Using squared residuals makes this rather nonrobust, so alternatives using the absolute values of the residuals (or even other transformations of the residuals) can be used. Carroll and Davidian [1987] give further details.

Methods Based on Replication

Suppose that for each distinct vector observation \mathbf{x}_i of the explanatory variables, there are several independent responses, so that we can write the model as

$$Y_{ij} = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_{ij} \quad (j = 1, \dots, n_i, i = 1, \dots, m), \quad (10.43)$$

where the ε_{ij} are independent $N(0, \sigma_i^2)$ errors. An obvious approach is to estimate the error variances σ_i^2 by the sample variances

$$S_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2.$$

However, as pointed out by Carroll and Cline [1988], this is very inaccurate unless the n_i 's are quite large. If $\hat{\boldsymbol{\beta}}_{SV}$ denotes the estimate of $\boldsymbol{\beta}$ obtained by using the inverses of the sample variances as weights, Carroll and Cline show that $\hat{\boldsymbol{\beta}}_{SV}$ has variance approximately $(n_0 - 3)/(n_0 - 5)$ times that of $\hat{\boldsymbol{\beta}}_{WLS}$ in the case where $n_i = n_0$ for all i , provided that $n_0 \geq 6$. For $n_0 = 10$, the increase in variance is 40%. For small n_0 , the situation is even worse: For $n_0 < 6$ and normally distributed data, the relative efficiency of $\hat{\boldsymbol{\beta}}_{SV}$ compared with $\hat{\boldsymbol{\beta}}_{WLS}$ is zero.

A better method is to estimate the variances by

$$\tilde{S}_i^2 = \sum_{j=1}^{n_i} e_{ij}^2 / n_i,$$

where e_{ij} is the ordinary least squares residual, and then calculate the weighted estimate with these weights. For $n_0 \geq 3$, the relative efficiency is positive (Fuller and Rao [1978]; Carroll and Cline [1988]). We can also consider the estimate obtained by iterating this process. If the ordinary least squares (OLS) estimate is sufficiently inefficient compared with the WLS estimate, then iterating will produce a better estimate.

Variance is a Function of the Mean

If the variance is a smooth function of the mean, we can do better. Suppose that

$$\sigma_i^2 = w(\mathbf{x}'_i \boldsymbol{\beta}),$$

where the function w is known. Then apply the following:

Algorithm 10.4

Step 1: Obtain an estimate $\hat{\beta}$ of β : for example, by using the ordinary least squares estimate.

Step 2: Calculate $\hat{\Sigma} = \text{diag}[w(\mathbf{x}'_1\hat{\beta}), \dots, w(\mathbf{x}'_n\hat{\beta})]$ using the current estimate of β .

Step 3: Recompute $\hat{\beta}$ as $(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{Y}$ using a weighted least squares program.

Step 4: Repeat steps 2 and 3 until convergence.

Carroll and Ruppert [1982] show that the estimate produced by this algorithm has the same asymptotic efficiency as $\hat{\beta}_{WLS}$, which means that in this case there is no cost in not knowing the weights, provided that w is known. Remarkably, the same is true if w is unknown but smooth. Carroll [1982a] shows that if we plot the squared residuals from a least squares fit versus the OLS fitted values and smooth the plot, we can do just as well. Let \hat{w}_i be the smoothed value of e_i^2 . Then if we use weighted least squares with weights equal to $1/\hat{w}_i$, we obtain an estimate whose asymptotic efficiency relative to $\hat{\beta}_{WLS}$ is 100%.

10.4.3 Transforming to Equalize Variances

As an alternative to explicitly modeling the variance function, we can transform the response in order to make the variances more homogeneous. We want to find an increasing transformation f such that

$$f(Y_i) = \mathbf{x}'_i\beta + \varepsilon_i,$$

where the ε_i have equal variance. If the variances in the untransformed model are increasing functions of the mean, this will often be successful. Suppose that

$$\text{var}[Y_i] = w(\mu_i),$$

where $\mu_i = \mathbf{x}'_i\beta$. Assuming that w is known, we get from A.13.4 that

$$\begin{aligned} \text{var}[f(Y)] &\approx \left(\frac{df}{d\mu} \right)^2 \text{var}[Y] \\ &= \left(\frac{df}{d\mu} \right)^2 w(\mu). \end{aligned}$$

It follows that the variances of the transformed responses $f(Y_i)$ will be approximately constant if we choose f so that $(df/d\mu)^2 w(\mu)$ is constant or,

equivalently, if we choose

$$f(\mu) = \int \frac{d\mu}{w(\mu)^{1/2}}.$$

EXAMPLE 10.4 Suppose that the responses Y_i follow Poisson distributions with means μ_i , so that $w(\mu) = \mu$. Then

$$f(\mu) = \int \frac{d\mu}{\mu^{1/2}} \propto \mu^{1/2}.$$

We expect $Y_1^{1/2}, \dots, Y_n^{1/2}$ to have approximately equal variances. \square

EXAMPLE 10.5 Suppose that the responses Y_i follow binomial(m_i, p_i) distributions with means $\mu_i = m_i p_i$, so that $w(\mu) = \mu(1 - \mu/m)$. Then

$$f_i(\mu) = \int \frac{d\mu}{\sqrt{\mu(1 - \mu/m_i)}} \propto \sin^{-1}\{(\mu/m_i)^{1/2}\},$$

so that $\sin^{-1}\{(Y_1/m_1)^{1/2}\}, \dots, \sin^{-1}\{(Y_n/m_n)^{1/2}\}$ have approximately equal variances. \square

We note that these two transformations provide a way of handling count data, provided that the counts are not too close to zero [or too close to m in the case of binomial(m, p) data].

If w is not known, we can experiment by transforming the responses Y_i using a power transformation of the form (10.19). We can transform with a suitable power less than 1, and then use the diagnostic plots described in Section 10.4.1 to see if the variances have been made equal. We proceed by reducing the power until the wedge effect in the plot of squared residuals versus fitted values disappears.

10.4.4 Serial Correlation and the Durbin–Watson Test

If the data are collected sequentially in time, then successive errors may be correlated. If this is the case, a time sequence plot of e_i against time order, which is often the same as a plot of e_i versus i , may show up the presence of any correlation between time consecutive e_i . The two plots in Figure 10.4 show the patterns due to positive and negative correlations, respectively. For positively correlated errors, a residual tends to have the same sign as its predecessor, while for negatively correlated errors, the signs of the residuals tend to alternate. One other useful plot consists of dividing up time-ordered residuals into consecutive pairs and plotting one member of the pair against the other. Serially correlated data show up as a linear trend in the plot.

There are a number of ways to test for correlation in a time series. For example, the simplest test is the *runs test* based on the sequence of signs of the

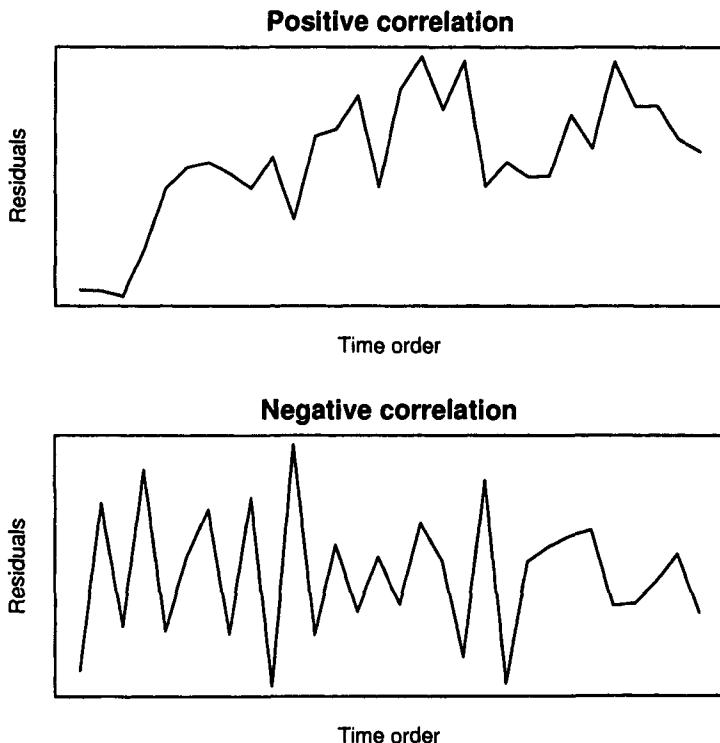


Fig. 10.4 Plots of residuals against time for positive and negatively correlated errors. Negatively correlated errors give a more jagged plot.

time-ordered residuals (see Brunk [1965: p. 354] for an excellent discussion), although this test is only approximate, as the residuals are slightly correlated even when the errors are uncorrelated. However, perhaps the most popular test for serial correlation is the d -test proposed by Durbin and Watson [1950, 1951, 1971]; this test is described below.

Suppose that the errors ε_i follow a first-order autoregressive model; that is, $\varepsilon_i = \rho \varepsilon_{i-1} + \delta_i$, where the δ_i are independently and identically distributed as $N(0, \sigma^2)$. Let

$$D = \frac{\sum_{i=2}^n (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^n \varepsilon_i^2} = \frac{\mathbf{e}' \mathbf{A} \mathbf{e}}{\mathbf{e}' \mathbf{e}}, \quad (10.44)$$

say. Then Durbin and Watson [1971] showed that the critical region $D < d_\alpha$ for testing the null hypothesis $H_0: \rho = 0$ against the one-sided alternative $H_1: \rho > 0$ has certain optimal properties; for example, it is the locally most powerful invariant critical region. Unfortunately, when H_0 is true, the null distribution of D depends on the data matrix \mathbf{X} , so that d_α has to be specially computed for each \mathbf{X} .

However, Durbin and Watson give several approximate procedures that seem to work very well in practice. In the first instance, they proved in their 1950 paper that $D_L \leq D \leq D_U$, where the distributions of D_U and D_L do not depend on \mathbf{X} ; the significance points of these distributions are tabulated in their 1951 paper for different n and k' ($= p - 1$), and in Koerts and Abrahamse [1969: pp. 176–178]. They also showed that when

H_0 is true, $R = \frac{1}{4}D$ can be approximated satisfactorily by a beta random variable with the same mean and variance; that is, the null density function of R is approximately beta(p_0, q_0) (cf. A.13.6), where

$$\begin{aligned} p_0 + q_0 &= \frac{E[D]\{4 - E[D]\}}{\text{var}[D]} - 1, \\ p_0 &= \frac{1}{4}(p_0 + q_0)E[D], \end{aligned}$$

and $E[D]$ and $\text{var}[D]$ are given by equations 3.1 to 3.4 in their 1971 paper. On the basis of these approximations they suggest the following procedure: Let d be the observed value of D , let α be the size of the test, and let $d_{L\alpha}$ and $d_{U\alpha}$ be the lower tail α significance points for the distributions of D_L and D_U , respectively. If $d < d_{L\alpha}$, reject H_0 ; if $d > d_{U\alpha}$, accept H_0 ; and if $d_{L\alpha} \leq d \leq d_{U\alpha}$, evaluate

$$\int_0^d f(r) dr$$

numerically and accept or reject H_0 according as this integral is greater or less than α . (Package computer programs are generally available for computing the beta distribution function.)

To test for negative correlation, that is, use the alternative hypothesis $H_1 : \rho < 0$, we simply use the statistic $4 - D$; the quantity $4 - d$ may now be treated as though it is the observed value of a D statistic to be tested for positive correlation. Two-sided tests for $H_1 : \rho \neq 0$ are obtained by combining the two one-sided tests and using a significance level of $\frac{1}{2}\alpha$ in each case.

Durbin and Watson [1971: p. 18] give one other approximate test procedure based on the critical value $d_\alpha = a + bd_{V\alpha}$, where a and b are calculated so that D and $a + bD_U$ have the same mean and variance.

Chatfield [1998] notes that in addition to the inconclusive nature of the d -test, the x -variables may, in practice, include lagged y -values. When n is large and p is small, he suggests using an estimate of ρ , namely,

$$r_1 = \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2} \quad \left(\approx 1 - \frac{D}{2} \right),$$

which is approximately $N(0, 1)$ when $\rho = 0$.

EXERCISES 10c

1. Differentiate (10.35) and take expectations to derive (10.37).

2. (a) Let $\mathbf{G} = \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$, and let \mathbf{Q} be an orthonormal basis for $\mathcal{C}(\mathbf{X})^\perp$, so that $\mathbf{Q}\mathbf{Q}' = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Verify that $\mathbf{G}'\Sigma\mathbf{Q} = \mathbf{0}$ and that $\mathbf{G}'\Sigma\mathbf{G} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$.
 - (b) Let $\mathbf{M} = (\Sigma^{1/2}\mathbf{Q}, \Sigma^{1/2}\mathbf{G})$. Show that \mathbf{M} is of full rank, so that $\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}' = \mathbf{I}_n$.
 - (c) Use (b) to prove (10.38).
 - (d) Use the fact that $[\det(\mathbf{M})]^2 = \det(\mathbf{M}'\mathbf{M})$ to prove (10.39).
3. (a) Suppose that $\mathbf{Y} = (Y_1, \dots, Y_n)'$, where the Y_i 's are independently and identically distributed as $N(\mu, \sigma^2)$. Find the restricted likelihood for $\mathbf{Q}'\mathbf{Y}$.
 - (b) Show that for this example the REML estimate of σ^2 is the usual unbiased estimate $S^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. Thus in this case the REML estimate is unbiased but the MLE is not.
4. Suppose that Y_1, \dots, Y_n are gamma random variables with density functions

$$f_i(y) = \frac{1}{\Gamma(r)\lambda_i^r} y^{r-1} \exp(-y/\lambda_i),$$

so that $E[Y_i] = r\lambda_i = \mu_i$, say, and $\text{var}[Y_i] = r^{-1}\mu_i^2$. Find a transformation that will make the variances of the Y_i 's approximately equal.

10.5 DEPARTURES FROM NORMALITY

10.5.1 Normal Plotting

We saw in Section 9.5 that the robustness of the F -test against departures from normality depends very much on the distribution of the explanatory variables, with the test being very robust if the explanatory variables are approximately normally distributed. F -tests can, however, sometimes be misleading, although the effects of nonnormality are not as a rule as great as those caused by inhomogeneity of variance or serial correlation. Still, nonnormality can often be an issue.

Nonnormal errors are detected by a normal plot of residuals, as shown in Figure 10.5. The normal plot is constructed by first ordering the least squares residuals, obtaining the order statistics $e_{(1)} \leq \dots \leq e_{(n)}$ and then plotting these against selected quantiles of the standard normal distribution. If the regression assumptions are satisfied, then, by (10.1), \mathbf{e} has a singular distribution $N_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$, which is approximately $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Hence the residuals are approximately a random sample from a $N(\mathbf{0}, \sigma^2)$ distribution, so that (see, e.g., Chambers et al. [1983]),

$$E[e_{(i)}] \approx \sigma \xi(\alpha_i),$$

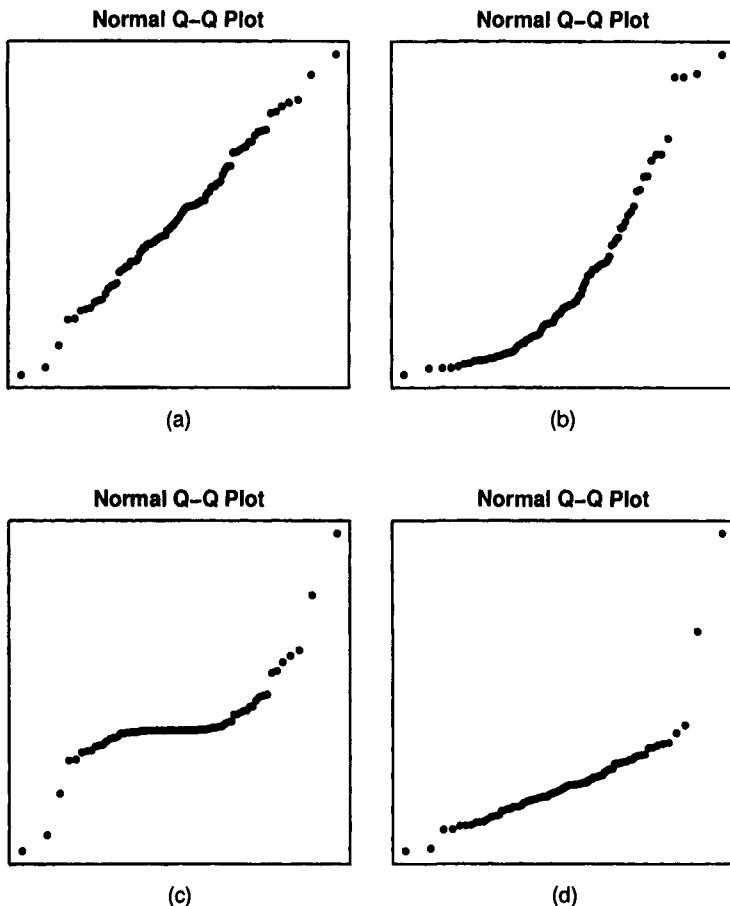


Fig. 10.5 Normal plots of residuals: (a) No indication of non-normality. (b) Skewed errors. (c) Heavy-tailed errors. (d) Outliers.

where $\alpha_i = (i - 0.5)/n$ and $\xi(\alpha)$ is the α -quantile of the standard normal distribution [with density $\phi(z)$] defined by

$$\int_{-\infty}^{\xi(\alpha)} \phi(z) dz = \alpha.$$

It follows that if we plot the normal quantiles versus the ordered residuals, normal errors will show up as a roughly straight plot. As illustrated in Figure 10.5, skewed errors show up as a curve, heavy-tailed errors as an S-shape, and outliers as isolated points at the end of the plot. Of course, the residuals are

not exactly a random sample from a normal distribution, but this seems not to matter in practice. Studentized residuals can be used if desired.

10.5.2 Transforming the Response

If the normal plot reveals evidence of nonnormality, the standard remedy is to transform the response. A popular family of transformations is the Box–Cox family (10.19) introduced in Section 10.3.2. We note that this family assumes that the response variables have positive values. Box and Cox introduced the transformations to remedy several types of regression problems, so that all the regression assumptions (means linear in the explanatory variables, homogeneous variances, and normal errors) would be satisfied. Therefore, their method assumes that there is a transformation parameter λ such that

$$Y_i^{(\lambda)} = g(Y_i, \lambda) = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where $g(Y, \lambda)$ is given by (10.19). Under this assumption, the likelihood function for the original observations \mathbf{Y} is

$$(2\pi\sigma^2)^{-(1/2)n} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta}) \right\} |J|, \quad (10.45)$$

where for each $y_i > 0$,

$$|J| = \left| \prod_{i=1}^n \frac{dy_i^{(\lambda)}}{dy_i} \right| = \prod_{i=1}^n y_i^{\lambda-1},$$

is the absolute value of the Jacobian. For λ fixed, (10.45) is the likelihood corresponding to a standard least squares problem, except for the constant factor J . From (3.18), the maximum value of this likelihood function is $(2\pi\hat{\sigma}^2)^{-(1/2)n} e^{-(1/2)n} |J|$, where

$$n\hat{\sigma}^2 = \mathbf{y}^{(\lambda)'} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \mathbf{y}^{(\lambda)} = \text{RSS}(\lambda; \mathbf{y}),$$

say. Hence, apart from a constant, the maximum log likelihood is

$$L_{\max}(\lambda) = -\frac{1}{2}n \log \{\text{RSS}(\lambda; \mathbf{y})\} + (\lambda - 1) \sum_{i=1}^n \log y_i. \quad (10.46)$$

Box and Cox suggest plotting $L_{\max}(\lambda)$ against λ for a trial series of values and reading off the maximizing value $\hat{\lambda}$. The transformation using this λ is then applied to the responses and the data are replotted to check the normality.

A more accurate value of $\hat{\lambda}$ can be obtained by solving the equations $dL_{\max}(\lambda)/d\lambda = 0$ (see equation (12) in their 1964 paper, or equation (9) in Schlesselman [1971]); some properties of $\hat{\lambda}$ are discussed further in Draper and Cox [1969]. Box and Cox also discuss the estimation of λ from a Bayesian viewpoint.

There is an obvious difficulty associated with these transformations. For the transformation to make sense, that is, to be monotone and be defined for all λ , the response data must come from a distribution supported on the positive half-axis. The only transformation of the form (10.19) that maps onto the whole real line is that with $\lambda = 0$ (i.e., the log transformation). Thus, if $\lambda \neq 0$, we cannot possibly achieve exact normality. The best we can do is to pick a transformation parameter λ to make the transformed distribution as close to normal as possible.

A justification for the Box–Cox procedure along these lines has been given by Hernandez and Johnson [1980]. Suppose that the transformed data have joint density $h_\lambda(\mathbf{y})$, and define $f(\mathbf{y}; \boldsymbol{\beta}, \sigma)$ to be the density of a multivariate normal with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance–covariance matrix $\sigma^2\mathbf{I}_n$. A convenient measure of the difference between $h_\lambda(\mathbf{y})$ and $f(\mathbf{y}; \boldsymbol{\beta}, \sigma)$ is given by the Kullback–Leibler discrepancy

$$I(f, h) = \int h_\lambda(\mathbf{y}) \{\log h_\lambda(\mathbf{y}) - \log f(\mathbf{y}; \boldsymbol{\beta}, \sigma)\} d\mathbf{y}, \quad (10.47)$$

described more fully in Section 12.3.3. If we select λ_* , $\boldsymbol{\beta}_*$, and σ_*^2 to minimize this discrepancy, then Hernandez and Johnson show that the minimizing values for $\boldsymbol{\beta}_*$ and σ_*^2 are given by

$$\begin{aligned} \boldsymbol{\beta}_* &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_{\lambda_*}[\mathbf{Y}] \\ \sigma_*^2 &= n^{-1}\{E_{\lambda_*}[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] + \text{tr}(\text{Var}_{\lambda_*}[\mathbf{Y}]\mathbf{P})\}, \end{aligned}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The value of λ_* cannot be established unless g_λ is known. However, Hernandez and Johnson show that the estimate of λ obtained by maximizing (10.46) is a consistent estimate of λ_* , thus providing an asymptotic justification of the Box–Cox procedure.

Other families of transformations can be used. John and Draper [1980] suggest the family

$$g(y, \lambda) = \begin{cases} \text{sign}(y)\{(|y| + 1)^\lambda - 1\}/\lambda, & \lambda \neq 0, \\ \text{sign}(y) \log(|y| + 1), & \lambda = 0. \end{cases}$$

This is well defined for all values of y , and for each λ we see that g is a one-to-one monotone transformation of the real line onto itself. These transformations seem to work well on data that are symmetrically distributed but with long tails, whereas the Box–Cox transformation is better at dealing with skewed data. John and Draper give examples where the Box–Cox transformation fails to induce normality, but the John–Draper transformation works very well. A further family is discussed by Yeo and Johnson [2000].

A feature of the Box–Cox technique is that the need to estimate the transformation parameter λ inflates the variances of the elements of $\hat{\boldsymbol{\beta}}$ considerably, compared to when $\boldsymbol{\beta}$ is estimated on a known scale. This has been established by Bickel and Doksum [1981]. Carroll [1980, 1982b] came to the

same conclusion, and noted that, in addition, the estimate of the transformation parameter can be very badly affected by outliers. Cook and Wang [1983] suggested a diagnostic for this problem, and Carroll [1980] and Carroll and Ruppert [1985] give robust methods of estimating λ .

Box and Cox [1982] and Hinkley and Rungger [1984] vigorously disputed the criticism of Bickel and Doksum, arguing that since β is specific to an estimated scale, its unconditional properties have little scientific relevance. Rather, a scale should be selected, and then the analysis should be carried out assuming the scale is the true one, so that β can be interpreted with respect to a fixed scale. In other words, the analysis should be performed conditionally on $\hat{\lambda}$. Carroll and Ruppert [1981] observe that in the case of prediction, the predicted response can be expressed in the original scale by back-transforming a prediction made in the transformed scale. In this case no question of scientific irrelevance arises, and there is indeed a small cost (in terms of increased prediction error) in not knowing the correct scale. Atkinson [1985] discusses this issue further and considers other families of transformations.

10.5.3 Transforming Both Sides

If $E[Y_i] = \mathbf{x}'_i\beta$, but the errors are nonnormal and/or heteroscedastic, then transforming the response will destroy the linear form for the mean. To avoid this problem, Carroll and Ruppert [1984, 1988] introduced the *transform-both-sides* model, which takes the form

$$g(Y_i, \lambda) = g(\mathbf{x}'_i\beta, \lambda) + \varepsilon_i,$$

where for some transformation family $g(y, \lambda)$ and some value of λ , the errors ε_i are normally distributed with constant variance σ^2 . Carroll and Ruppert consider more general mean structures than the one considered here, but the results are similar.

If $g(y, \lambda)$ is the Box–Cox family, the parameters can be estimated easily by maximum likelihood, as when the response alone is transformed. In the present case, the log likelihood takes the form

$$l(\beta, \sigma^2, \lambda) = c - \frac{1}{2} \left\{ n \log \sigma^2 + \sigma^{-2} \sum_{i=1}^n [g(y_i, \lambda) - g(\mathbf{x}'_i\beta)]^2 \right\} + (\lambda - 1) \sum_{i=1}^n \log y_i.$$

Differentiating with respect to σ^2 and equating the derivative to zero, we see that for fixed β and λ the log likelihood is maximized by

$$\hat{\sigma}^2(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^n [g(y_i, \lambda) - g(\mathbf{x}'_i\beta)]^2,$$

and the maximum value of the log likelihood for fixed β and λ is

$$l_{\text{MAX}}(\beta, \lambda) = c - \frac{1}{2} n \log \hat{\sigma}^2(\beta, \lambda) + n + (\lambda - 1) \sum_{i=1}^n \log y_i.$$

This can be maximized by Fisher scoring (A.14) to obtain estimates of β and the transformation parameter λ . Alternatively, if we put $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$, we see that the log likelihood, up to a constant, can be written as

$$-\frac{n}{2} [\log \hat{\sigma}^2(\beta, \lambda) - 2\lambda \log \dot{y}] = -\frac{n}{2} \log \left[\frac{\hat{\sigma}^2(\beta, \lambda)}{\dot{y}^{2\lambda}} \right],$$

which is maximized by minimizing

$$\sum_{i=1}^n \left[\frac{g(y_i, \lambda) - g(\mathbf{x}'_i \beta)}{\dot{y}^\lambda} \right]^2.$$

This can be done using a nonlinear least squares program. Standard errors can be calculated from the information matrix in the usual way. Unlike transforming the response alone, when estimating β there does not seem to be much, if any, cost in accuracy in having to estimate the transformation parameter λ .

This technique can be used to make the error distribution closer to normal and also to stabilize the variances. However, it is not necessarily true that the same transformation will achieve both these objectives. If the transformation that makes the error distribution closer to normal does not also make the variances more homogeneous, it may be necessary to use weighting to achieve the latter objective. Carroll and Ruppert [1988: Chapter 5] give further details.

EXERCISES 10d

1. Let $Z_{(1)}, \dots, Z_{(n)}$ be the order statistics calculated from a random sample from a $N(0, 1)$ distribution. Then (see, e.g., David [1981], Azzalini [1996: p. 301]) the density of $Z_{(i)}$ is

$$\frac{n!}{(i-1)! (n-i)!} \Phi(z)^{i-1} [1 - \Phi(z)]^{n-i} \phi(z),$$

where Φ and ϕ are, respectively, the distribution and density functions of the standard normal distribution.

- (a) By making a transformation, show that

$$E[Z_{(i)}] = B(i, n - i + 1)^{-1} \int_0^1 \Phi^{-1}(y) y^{i-1} (1 - y)^{n-i} dy,$$

where $B(i, n - i + 1)$ is the beta function (A.13.6).

- (b) Using the definition of the integral, show that

$$E[Z_{(i)}] \approx \sum_{j=1}^n \Phi^{-1}[(j - 0.5)/n] w_{ij},$$

where

$$w_{ij} = B(i, n - i + 1)^{-1} \int_{(j-1)/n}^{j/n} y^{i-1}(1-y)^{n-i} dy.$$

- (c) Explain why the weights w_{ij} sum to 1.
 - (d) Comment on the accuracy of the approximation
- $$E[Z_{(i)}] \approx \Phi^{-1}[(i - 0.5)/n].$$
2. Suppose that $g(y, \lambda)$ is a family of monotone-increasing transformations such that $g(Y_i, \lambda)$ is $N(\mathbf{x}'_i \boldsymbol{\beta}, \sigma^2)$.
- (a) Show that the log likelihood for Y_1, \dots, Y_n is
- $$c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n [g(y_i, \lambda) - \mathbf{x}'_i \boldsymbol{\beta}]^2 + \sum_{i=1}^n \log \left| \frac{\partial g(y_i, \lambda)}{\partial y_i} \right|.$$
- (b) Show that in the case of the John–Draper transformation, the log likelihood is
- $$c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n [g(y_i, \lambda) - \mathbf{x}'_i \boldsymbol{\beta}]^2 + (\lambda - 1) \sum_{i=1}^n \log(1 + |y_i|).$$
3. Show that for fixed λ , the values of $\boldsymbol{\beta}$ and σ^2 that minimize (10.47) are

$$\boldsymbol{\beta}_* = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' E_{\lambda_*} [\mathbf{Y}]$$

and

$$\sigma_*^2 = n^{-1} \{ E_{\lambda_*} [\mathbf{Y}' (\mathbf{I}_n - \mathbf{P}) \mathbf{Y}] + \text{tr}(\text{Var}_{\lambda_*} [\mathbf{Y}] \mathbf{P}) \},$$

where E_λ denotes expectation with respect to h_λ and $\mathbf{P} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$.

10.6 DETECTING AND DEALING WITH OUTLIERS

10.6.1 Types of Outliers

In Section 9.4 we identified two kinds of outlier in regression data. First, there may be a big difference between the explanatory vector \mathbf{x}_i for the i th case and the center of the x -data. In other words, \mathbf{x}_i may be an outlier in the p -dimensional space occupied by the rows of the regression matrix; we referred to such a point as a high-leverage point. Second, there may be a large difference between the response Y_i and the mean $\mathbf{x}'_i \boldsymbol{\beta}$ predicted by the model, and we called such a point an outlier.

We shall see that outliers that are not high-leverage points do not have a very strong influence on the fitted regression plane unless they are very large. The situation is very different when outliers are also high-leverage points. For this reason, it is important to identify the high-leverage points and determine if they are having an undue influence on the fit.

Since $Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$, $\hat{Y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + e_i$, and $\hat{\boldsymbol{\beta}}$ estimates $\boldsymbol{\beta}$, it is reasonable to treat the residual e_i as an estimate of the error ε_i and examine the residuals for extreme values. We can either use the raw residuals e_i , or the Studentized residuals r_i or t_i , which are adjusted to be identically distributed. Standard graphical plots for examining univariate data such as normal plots or box plots can be used to check the residuals for extreme values. We then identify as outliers the points having large residuals. Since the externally Studentized residuals have t_{n-p-1} distributions in the absence of outliers, a reasonable definition of "large" is a point for which $|t_i| > 2$.

The diagnostic approach described above works well provided that the data point in question does not have high leverage. If it does, we cannot expect the corresponding residual to reveal the presence of an outlier. To see this, suppose that the i th response is recorded as $Y_i - \Delta_i$ rather than Y_i , so that $Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \Delta_i + \varepsilon_i$, which we can interpret as the error being changed by an amount Δ_i . Let Δ be the vector that has i th element Δ_i and the rest zero. Then, since $\mathbf{H}\mathbf{X} = \mathbf{X}$,

$$\begin{aligned} E[\mathbf{e}] &= E[(\mathbf{I}_p - \mathbf{H})\mathbf{Y}] \\ &= (\mathbf{I}_p - \mathbf{H})E[\mathbf{Y}] \\ &= (\mathbf{I}_p - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \Delta) \\ &= (\mathbf{I}_p - \mathbf{H})\Delta, \end{aligned}$$

so that $E[e_i] = (1 - h_i)\Delta_i$. Thus if the \mathbf{x}_i is close to $\bar{\mathbf{x}}$ [cf. (10.11)] so that h_i is small, we can expect the residual to reflect the outlier quite well, as $E[e_i]$ will be close to Δ_i . On the other hand, if the data point has high leverage, then, as explained in Section 10.2, the hat matrix diagonal will be close to 1, and the residual will tend to be smaller than Δ_i .

If least squares residuals cannot reveal outliers for high-leverage points, what can be done? First, we need a reliable means of identifying high-leverage points. The hat matrix diagonals are reasonable measures of leverage, since they can be interpreted in terms of Mahalanobis distances, as explained in Section 10.2. Since the average hat matrix diagonal is p/n , an arbitrary but reasonable definition of a high-leverage point is one satisfying $h_i > 2p/n$ (see Belsley et al. [1980], Atkinson [1985]).

Whether or not a point has high leverage depends on the explanatory variables included in the regression. For example, if a particular case has an extreme value for one explanatory variable, say x_1 , but not for the other variables, then the case will have high leverage. However, if variable x_1 is dropped from the model, then the case will no longer have high leverage (cf. Miscellaneous Exercises 10, No. 2). Unfortunately, hat matrix diagonals

are themselves subject to the effects of high-leverage points and do not always give a reliable indication of leverage. This point is discussed further in Section 10.6.2.

Having identified a point as having high leverage, how can we tell if the point is an outlier? There are two standard solutions. The first involves assessing the effect that a data point has on the regression by deleting the point and refitting the regression. If the regression quantities (estimated coefficients, fitted values, standard errors, and so on) change markedly, then the point is called a *high-influence point*, and is probably an outlier. In Figure 10.6(a), where the point A is not an outlier, it is clear that deleting A and refitting will cause little change in the fitted regression. In Figure 10.6(b), where A is an outlier, the reverse is true; deleting A and refitting will cause a large change. Motivated by this example, we can consider calculating key regression quantities, such as regression coefficients and fitted values, but leaving out each data point in turn and noting the resulting change. Leave-one-out diagnostics are discussed further in Section 10.6.3.

The second solution uses a robust fitting method that is not affected by the high-leverage points, resulting in residuals that better identify outliers, and is described in Section 10.6.5.

Having tentatively identified the outliers, what action should we take? If the outliers are due to mistakes in recording data, clearly the mistakes should be corrected if possible, or else the points put aside. If the outlying data are genuine, they should not just be ignored, as they represent unexpected

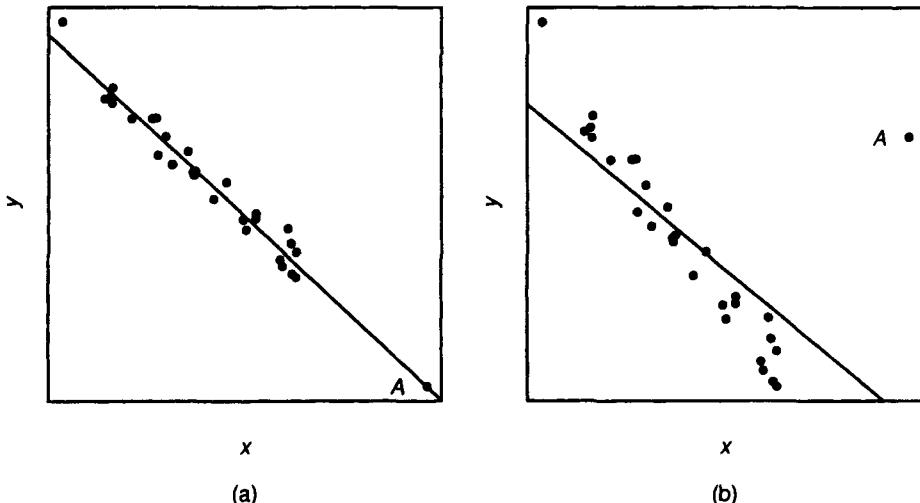


Fig. 10.6 Effect of high-leverage points on least squares fits.

and possibly important information on the relationship between explanatory variables and the response. Sometimes outliers are due to model failure and disappear when a more appropriate model is fitted. On the other hand, we do not want to fit models whose characteristics are determined by one or two points. If we are going to fit the model by least squares, we will want to remove high-influence points before fitting. This does not mean that they should be discarded from consideration.

We can, of course, use robust methods. However, there is considerable evidence that residuals from robust fits can be misleading when we are trying to detect curvature in the regression surface (cf. Cook et al. [1992]; McKean et al. [1993]).

10.6.2 Identifying High-Leverage Points

As discussed in Section 10.6.1, the hat matrix diagonals can be written in terms of Mahalanobis distances, which are in turn functions of the sample mean and covariance matrix. These two measures are very nonrobust and will be affected badly by high-leverage points. We have a situation where the diagnostics used to identify the high-leverage points are undermined by the very points they are designed to detect. The effect of a high-leverage point is to increase the elements of the covariance matrix, and this reduces the value of the Mahalanobis distance between a “reduced” row \mathbf{x}_i and the mean row of the regression matrix. This is illustrated in Figure 10.7. The ellipse shows a contour of points having the same Mahalanobis distance from the average row $\bar{\mathbf{x}}$. The points in the upper right corner of the plot distort the distances so that points A and B seem to be equally outlying, so that A is not identified as a high-leverage point. The effect of this is that high-leverage points are not identified as such, due to the influence of other high-leverage points. This phenomenon is known as *masking* (Rousseeuw and van Zomeren [1990]). To combat this, Rousseeuw and van Zomeren suggest using a modified Mahalanobis distance

$$\text{MD}_i = \{[\mathbf{x}_i - T(\mathbf{X})]' \mathbf{C}(\mathbf{X})^{-1} [\mathbf{x}_i - T(\mathbf{X})]\}^{1/2}, \quad (10.48)$$

where $T(\mathbf{X})$ and $\mathbf{C}(\mathbf{X})$ are robust estimates of location and covariance for the reduced rows (minus the first unit element) of \mathbf{X} . They suggest using the *minimum volume ellipsoid* (MVE), which is defined (Rousseeuw and Leroy [1987]) as the p -dimensional ellipsoid of minimum volume that contains 50% of the n reduced points. $T(\mathbf{X})$ is taken as the center of the ellipsoid.

The MVE is calculated using the method of elemental regressions (cf. Section 11.12.3). For a $(p+1)$ -subset J of cases (data points), we calculate the mean vector $\bar{\mathbf{x}}_J$ and the sample variance-covariance matrix \mathbf{C}_J . We then consider the ellipsoid

$$\{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}}_J)' \mathbf{C}_J^{-1} (\mathbf{x} - \bar{\mathbf{x}}_J) \leq m\},$$

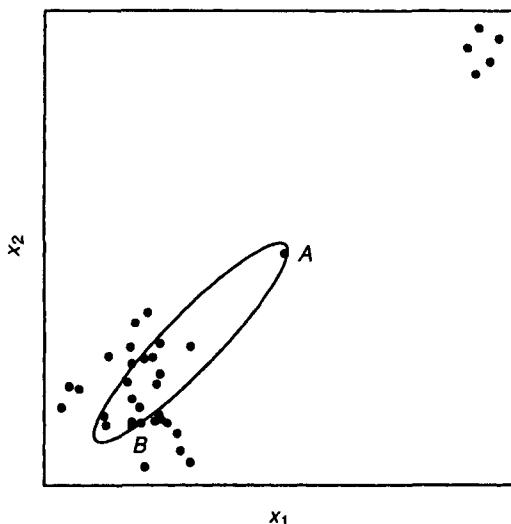


Fig. 10.7 Effect of high-leverage points on hat matrix diagonals.

where m is chosen to make the ellipsoid include 50% of the observations. Clearly, we must take $m = m_J^2$, where

$$m_J^2 = \text{median}_i (\mathbf{x}_i - \bar{\mathbf{x}}_J)' \mathbf{C}_J^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_J).$$

By Example 5.1 (Section 5.1.3), the resulting ellipsoid has volume equal to $k_p m_J^p \det(C_J)^{1/2}$, where k_p depends only on p . We repeatedly sample subsets J of observations and calculate the resulting volume. The MVE is taken to be the ellipsoid having the smallest volume of all those sampled.

The n quantities $(\mathbf{x}_i - \bar{\mathbf{x}}_J)' \mathbf{C}_J^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_J)$ would be approximately independent χ_p^2 if the (reduced) \mathbf{x} 's were multivariate normal. It follows that m_J^2 is approximately the median of independent χ_p^2 variables and is therefore a consistent estimate of $\chi_{p,0.50}^2$. Thus, to achieve consistency under the assumption of multivariate normality, we estimate $\mathbf{C}(\mathbf{X})$ by $m_J^2 \mathbf{C}_J / \chi_{p,0.50}^2$.

An alternative to the use of the MVE is the *minimum covariance determinant estimate* (MCD). This is based on an ellipsoid using the mean and covariance matrix calculated from an $(n-h)$ -subset of the data, where h is the number of outliers. The subset chosen is the one for which the determinant of the resulting covariance matrix is a minimum. In practice, h is unknown. It is important not to underestimate h , so that the choice $h = n - [(n+p+1)/2]$ is often used. Hawkins [1994b] gives an algorithm to compute the MCD estimate.

10.6.3 Leave-One-Out Case Diagnostics

The amount of computation required to calculate the changes in the regression quantities when points are deleted is greatly reduced by exploiting Theorem 10.1 in Section 10.2. We now describe a number of diagnostics.

Change in Estimated Regression Coefficients

If $\hat{\beta}(i)$ is the least squares estimate of β calculated with the i th case deleted, then (Theorem 10.1)

$$\hat{\beta} - \hat{\beta}(i) = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_i}. \quad (10.49)$$

We note that the change is proportional to the size of the residual but is greatly inflated if h_i is close to 1, showing how high-leverage points can have a big influence on the estimated regression coefficients.

The measure (10.49), called DFBETA, was introduced by Belsley et al. [1980] and is often “standardized” to aid interpretation. Let $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$; then \mathbf{C} is called the *catcher matrix* (Velleman and Welsch [1981]). The j,i element of \mathbf{C} is $c_{ji} = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i]_j$, so that the j th element of DFBETA is $c_{ji}e_i/(1 - h_i)$. Also,

$$\begin{aligned} \sum_i c_{ji}^2 &= \sum_i [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i]_j^2 \\ &= \sum_i [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1}]_{jj} \\ &= [(\mathbf{X}'\mathbf{X})^{-1} \sum_i \mathbf{x}_i \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1}]_{jj} \\ &= [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}]_{jj} \\ &= (\mathbf{X}'\mathbf{X})_{jj}^{-1}. \end{aligned}$$

We can standardize the $(j+1)$ th element DFBETA by dividing by an estimate of the standard error of $\hat{\beta}_j$, namely, $S(i)[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,j+1}^{1/2} = S(i)(\sum_i c_{j+1,i}^2)^{1/2}$, resulting in a standardized difference

$$\frac{\hat{\beta}_j - \hat{\beta}(i)_j}{S(i)(\sum_i c_{j+1,i}^2)^{1/2}}. \quad (10.50)$$

This measure is called DFBETAS by Belsley et al. [1980]. In terms of the Studentized residual t_i , it can be written as

$$\begin{aligned} \text{DFBETAS}_{ij} &= \frac{\hat{\beta}_j - \hat{\beta}(i)_j}{S(i)(\sum_i c_{j+1,i}^2)^{1/2}} \\ &= \frac{c_{j+1,i} e_i}{S(i)(\sum_i c_{j+1,i}^2)^{1/2}(1 - h_i)} \\ &= \frac{c_{j+1,i}}{(\sum_i c_{j+1,i}^2)^{1/2}} \frac{t_i}{(1 - h_i)^{1/2}}. \end{aligned} \quad (10.51)$$

If the i th data point is not an outlier and does not have high leverage, we expect that the Studentized residual will satisfy $|t_i| < 2$ and that Y_i will not have a large effect on the value of $\hat{\beta}_j$. Since $\hat{\beta}_j = \sum_{i=1}^n c_{j+1,i} Y_i$, this means that $c_{j+1,i}$ is small in relation to $(\sum_{i=1}^n c_{j+1,i}^2)^{1/2}$ or, very approximately, to $n^{-1/2}$. This suggests that a suitable cutoff value for detecting large residuals using DFBETAS_{ij} is $2/\sqrt{n}$.

Change in Fitted Values

The change in the i th fitted value is

$$\begin{aligned}\mathbf{x}'_i \hat{\beta} - \mathbf{x}'_i \hat{\beta}(i) &= \frac{\mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i e_i}{1 - h_i} \\ &= \frac{h_i e_i}{1 - h_i}\end{aligned}$$

and is called DFFITS. Standardized by the estimated standard deviation $S(i)h_i^{1/2}$ of $\mathbf{x}'_i \hat{\beta}$, it becomes

$$\begin{aligned}\text{DFFITSS}_i &= \frac{h_i^{1/2} e_i}{S(i)(1 - h_i)} \\ &= t_i \left(\frac{h_i}{1 - h_i} \right)^{1/2},\end{aligned}\quad (10.52)$$

by (10.4). For points that are not outliers and do not have high leverage, with high probability $|t_i| < 2$ and h_i will not be too far from the average value p/n . Then (10.52) suggests using the cutoff $2\sqrt{p/(n-p)}$ or even $2\sqrt{p/n}$.

Covariance Ratio

The estimated variance-covariance matrix of $\hat{\beta}$ is $S^2(\mathbf{X}' \mathbf{X})^{-1}$, and its *leave-one-out version* is $S(i)^2[\mathbf{X}(i)' \mathbf{X}(i)]^{-1}$. A convenient scalar measure of the change is the ratio

$$\frac{\det\{S(i)^2[\mathbf{X}(i)' \mathbf{X}(i)]^{-1}\}}{\det[S^2(\mathbf{X}' \mathbf{X})^{-1}]},$$

which is called the COVRATIO. To simplify this formula, we can use the result (10.9) to get

$$\frac{S^2}{S(i)^2} = \frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}. \quad (10.53)$$

Also, using A.9.7, we get

$$\begin{aligned}\det[\mathbf{X}(i)' \mathbf{X}(i)] &= \det(\mathbf{X}' \mathbf{X} - \mathbf{x}_i \mathbf{x}'_i) \\ &= \det(\mathbf{X}' \mathbf{X})[1 - \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i] \\ &= \det(\mathbf{X}' \mathbf{X})(1 - h_i),\end{aligned}\quad (10.54)$$

so that

$$\begin{aligned}\text{COVRATIO} &= \frac{\det\{S(i)^2[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}\}}{\det[S^2(\mathbf{X}'\mathbf{X})^{-1}]} \\ &= \left(\frac{S(i)^2}{S^2}\right)^p \frac{\det\{[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}\}}{\det[(\mathbf{X}'\mathbf{X})^{-1}]} \\ &= \left(\frac{n-p-1}{n-p} + \frac{t_i^2}{n-p}\right)^{-p} (1-h_i)^{-1},\end{aligned}$$

by (10.53).

To calibrate this measure, Belsley et al. [1980] consider two extreme cases. The first is when the externally Studentized residual is large, say $|t_i| > 2$, but the case has minimum leverage, with $h_i = n^{-1}$. [Recall that when the regression contains a constant term, then $h_i \geq n^{-1}$ by (10.12).] Then, for this case,

$$\begin{aligned}\text{COVRATIO} &= \frac{n}{n-1} \left(1 + \frac{t_i^2 - 1}{n-p}\right)^{-p} \\ &\approx 1 - \frac{p(t_i^2 - 1)}{n}\end{aligned}$$

when n is large and much greater than p . Therefore if $|t_i| > 2$,

$$\begin{aligned}\text{COVRATIO} &\approx 1 - \frac{p(t_i^2 - 1)}{n} \\ &\leq 1 - \frac{3p}{n}.\end{aligned}$$

The second is the opposite situation where the Studentized residual is small but the case has high leverage. Taking the extreme case $t_i = 0$ and $h_i > 2p/n$, then

$$\begin{aligned}\text{COVRATIO} &= \left(1 - \frac{1}{n-p}\right)^{-p} (1-h_i)^{-1} \\ &\geq \left(1 - \frac{1}{n-p}\right)^{-p} \left(1 - \frac{2p}{n}\right)^{-1} \\ &\approx \left(1 + \frac{p}{n}\right) \left(1 - \frac{2p}{n}\right)^{-1} \\ &\approx 1 + \frac{3p}{n},\end{aligned}$$

ignoring higher-order terms. Thus, cases having $|\text{COVRATIO} - 1| > 3p/n$ are considered to have high influence.

Cook's D

From (5.18) we see that a $100(1 - \alpha)\%$ confidence ellipsoid for β is

$$\{\mathbf{b} : (\mathbf{b} - \hat{\beta})' \mathbf{X}' \mathbf{X} (\mathbf{b} - \hat{\beta}) \leq pS^2 F_{p,n-p}^\alpha\}.$$

Cook [1977] suggested measuring the distance of $\hat{\beta}(i)$ from $\hat{\beta}$ by using the measure

$$D_i = \frac{(\hat{\beta}(i) - \hat{\beta})' \mathbf{X}' \mathbf{X} (\hat{\beta}(i) - \hat{\beta})}{pS^2} \quad (10.55)$$

based on the confidence ellipsoid. He suggested flagging as suspicious those points for which $D_i > F_{p,n-p}^{0.10}$.

Using (10.7), we can write Cook's D as [cf. (10.3)]

$$D_i = \frac{e_i^2 h_i}{(1 - h_i)^2 pS^2} = r_i^2 \frac{h_i}{p(1 - h_i)},$$

where r_i is the i th internally Studentized residual. Thus a point will have a large Cook's D if it has a large Studentized residual or is a high-leverage point. We note that apart from the constant divisor p , Cook's D is very similar to the square of DFITTS, differing only in the use of an internally rather than externally Studentized residual. For another interpretation of Cook's D in terms of influence functions, see Cook and Weisberg [1982: Chapter 3], and Exercises 10e, No. 5, at the end of Section 10.6.5.

Andrews and Pregibon Statistic

Consider the augmented matrix $\mathbf{X}_A = (\mathbf{X}, \mathbf{Y})$. Then

$$\begin{aligned} \det(\mathbf{X}_A' \mathbf{X}_A) &= \det \begin{pmatrix} \mathbf{X}' \mathbf{X} & \mathbf{X}' \mathbf{Y} \\ \mathbf{Y}' \mathbf{X} & \mathbf{Y}' \mathbf{Y} \end{pmatrix} \\ &= \det(\mathbf{X}' \mathbf{X}) \det[\mathbf{Y}' (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}] \\ &= \det(\mathbf{X}' \mathbf{X}) \times \text{RSS}, \end{aligned}$$

by A.9.5. Now consider deleting the i th row of \mathbf{X}_A . Using the same argument, we get

$$\det[\mathbf{X}_A(i)' \mathbf{X}_A(i)] = \det[\mathbf{X}(i)' \mathbf{X}(i)] \times \text{RSS}(i),$$

where $\text{RSS}(i)$ is the residual sum of squares computed from the $n - 1$ cases, excluding the i th. By (10.9),

$$\text{RSS}(i) = \text{RSS} - \frac{e_i^2}{1 - h_i}$$

and from (10.54), $\det[\mathbf{X}(i)' \mathbf{X}(i)] = \det(\mathbf{X}' \mathbf{X})(1 - h_i)$. The Andrews-Pregibon statistic (Andrews and Pregibon [1978]) is the ratio

$$\text{AP}(i) = \frac{\det[\mathbf{X}_A(i)' \mathbf{X}_A(i)]}{\det(\mathbf{X}_A' \mathbf{X}_A)},$$

which, using the arguments above, can be written as

$$\frac{\text{RSS}(i)}{\text{RSS}} \times \frac{\det[\mathbf{X}(i)'\mathbf{X}(i)]}{\det(\mathbf{X}'\mathbf{X})} = \left[1 - \frac{e_i^2}{\text{RSS}(1-h_i)} \right] (1-h_i). \quad (10.56)$$

In terms of the internally Studentized residuals r_i , we have

$$\text{AP}(i) = \left(1 - \frac{r_i^2}{n-p} \right) (1-h_i).$$

Note that since $(n-p)^{-1}r_i^2$ has a beta $[\frac{1}{2}, \frac{1}{2}(n-p-1)]$ distribution [cf. (10.10) and the following discussion], the first factor of $\text{AP}(i)$ is beta $[\frac{1}{2}(n-p-1), \frac{1}{2}]$.

The Andrews-Pregibon statistic also has an interpretation as a hat matrix diagonal. Let $h_{i,A}$ be the i th diagonal element of the hat matrix based on \mathbf{X}_A rather than \mathbf{X} . Then (see Exercises 10e, No. 4) $\text{AP}(i) = 1 - h_{i,A}$.

10.6.4 Test for Outliers

Suppose that we wish to test if a fixed set of k observations contain outliers, assuming that the remaining $n-k$ cases are “clean.” Arrange the data so that the clean observations come first, followed by the k possibly outlying observations. We will use the *outlier shift model*

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon, \quad (10.57)$$

where \mathbf{Z} is a matrix of the form

$$\mathbf{Z} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_k \end{pmatrix}$$

and γ is a k -vector containing the shifts for the possibly outlying observations.

We use the theory of Section 3.7.1 to test $\gamma = \mathbf{0}$. Let \mathbf{H} be the hat matrix for the regression

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad (10.58)$$

corresponding to $\gamma = \mathbf{0}$, and let

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix},$$

where \mathbf{H}_{11} is $n-k$ by $n-k$. Partition the residuals $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ conformably with \mathbf{H} as $\mathbf{e} = (\mathbf{e}'_1, \mathbf{e}'_2)'$. By Theorem 3.6, the least squares estimate of γ is

$$\begin{aligned} \hat{\gamma} &= [\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}]^{-1} \mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\ &= \left[(\mathbf{0}, \mathbf{I}_k)(\mathbf{I}_n - \mathbf{H}) \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_k \end{pmatrix} \right]^{-1} (\mathbf{0}, \mathbf{I}_k) \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \\ &= (\mathbf{I}_k - \mathbf{H}_{22})^{-1} \mathbf{e}_2. \end{aligned}$$

Further, from Theorem 3.6(iii), the numerator of the F -test for testing $\gamma = \mathbf{0}$ is

$$\begin{aligned}\hat{\gamma}' \mathbf{Z}' (\mathbf{I}_n - \mathbf{H}) \mathbf{Y} &= \hat{\gamma}' (\mathbf{0}, \mathbf{I}_k) \mathbf{e} \\ &= \hat{\gamma}' \mathbf{e}_2 \\ &= \mathbf{e}_2' (\mathbf{I}_k - \mathbf{H}_{22})^{-1} \mathbf{e}_2.\end{aligned}$$

Hence the F -test for $\gamma = \mathbf{0}$ is

$$F = \frac{\mathbf{e}_2' (\mathbf{I}_k - \mathbf{H}_{22})^{-1} \mathbf{e}_2 / k}{[\text{RSS} - \mathbf{e}_2' (\mathbf{I}_k - \mathbf{H}_{22})^{-1} \mathbf{e}_2] / (n - p - k)},$$

where RSS is the residual sum of squares for model (10.58).

When $k = 1$, we can test if the i th observation is an outlier. Then

$$\mathbf{e}_2' (\mathbf{I}_k - \mathbf{H}_{22})^{-1} \mathbf{e}_2 = \frac{e_i^2}{1 - h_i}$$

and the F -test reduces to [cf. (10.10)]

$$\frac{(n - p - 1)e_i^2}{(1 - h_i)\text{RSS} - e_i^2}, \quad (10.59)$$

which is distributed as $F_{1, n-p-1}$ when the i th observation is not an outlier; a significant value suggests otherwise, as e_i^2 is large.

We can also write (10.59) in terms of Studentized residuals. Using the relationship

$$r_i^2 = \frac{e_i^2}{S^2(1 - h_i)},$$

we see that (10.59) can be written as [cf. (10.10)]

$$\frac{(n - p - 1)r_i^2}{n - p - r_i^2} \quad (= t_i^2),$$

which is increasing in $|r_i|$.

10.6.5 Other Methods

Masking and Swamping

Hat matrix diagonals and the leave-one-out diagnostics described above are useful tools but can sometimes fail to identify outliers and high-leverage points. Consider the situation depicted in Figure 10.8(a). Taken singly, points A and B are not influential, since removing either will have little effect on the fitted line, because the remaining point continues to attract the line. However, taken as a pair, they are influential, since deleting both will have

a marked effect on the fitted line. In this case we say that points *A* and *B* *mask* each other. Single-case diagnostics cannot identify influential points occurring in clusters, where a point in the cluster is masked by the others. Possible remedies for masking are to use a robust fit such as least median squares to identify the cluster, or to use leave-many-out diagnostics. Both these remedies are discussed further below.

Another phenomenon is *swamping*, where points that are not outlying can be mistaken as such. Consider the situation shown in Figure 10.8(b). Point *A* will have a large residual, since *B* is attracting the line away from *A*. Deleting point *A* also has a big impact on the line. However, *A* is not an outlier. Swamping does not occur when robust fitting methods are used.

Leave-Many-Out

As we have seen, diagnostics based on the deletion of single cases will not identify a point as an outlier when it is masked by others. The obvious solution to this problem is to calculate *leave-d-out* diagnostics, but these have some drawbacks. Although this idea is practical for small values of *d*, the computational burden quickly becomes excessive, since we must calculate $\binom{n}{d}$ separate diagnostics for each regression quantity examined. Also, they will be effective in identifying the situation shown in Figure 10.8 only if all the points in a cluster are deleted. Thus, if the cluster contains *d* points, we must calculate leave-*d*-out diagnostics. This is a problem since *d* is not known

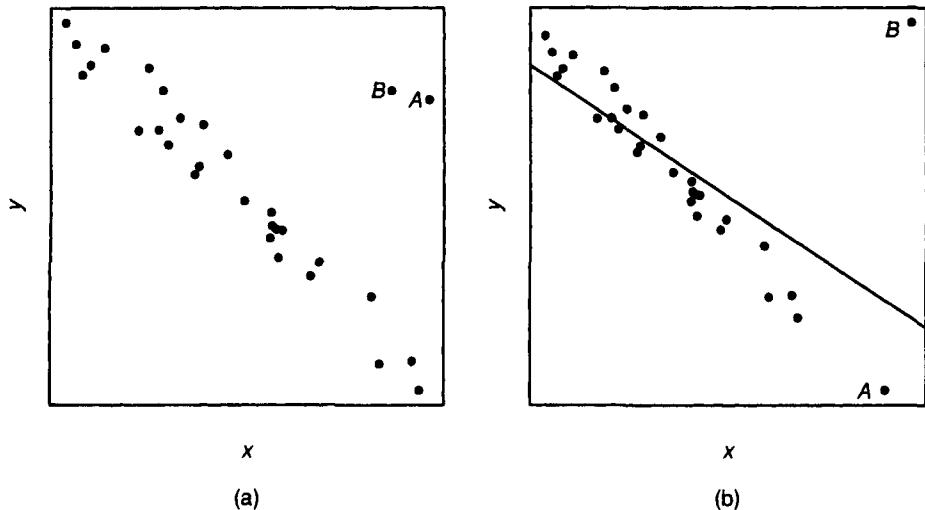


Fig. 10.8 Masking and swamping.

in advance. In any event, the procedure is computationally feasible only for small values of d .

Let D be a d -subset of cases, and let $\mathbf{X}(D)$ be the $(n-d) \times p$ submatrix of \mathbf{X} corresponding to the rows of the cases not in D , and let \mathbf{X}_D be the $d \times p$ submatrix corresponding to the rows of the cases in D . Then, using A.9.3 with $\mathbf{B} = \mathbf{I}_d$, we get

$$\begin{aligned} [(\mathbf{X}(D)'\mathbf{X}(D))]^{-1} &= (\mathbf{X}'\mathbf{X} - \mathbf{X}'_D\mathbf{X}_D)^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{X}_D(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

where $\mathbf{H}_D = \mathbf{X}_D(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D$. If $\hat{\beta}(D)$ denotes the vector of least squares estimates computed from the data with the cases in D deleted, then (Exercises 10e, No. 2)

$$\hat{\beta}(D) = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D, \quad (10.60)$$

where \mathbf{e}_D are the elements of \mathbf{e} from the cases in D .

Similarly, the difference in the fitted values for the cases in D is

$$\begin{aligned} \hat{Y}_D - \hat{Y}_D(D) &= \mathbf{X}_D\hat{\beta} - \mathbf{X}_D\hat{\beta}(D) \\ &= \mathbf{H}_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D. \end{aligned}$$

Also, the analog of the COVRATIO can be shown to be

$$\left\{ \left[\frac{n-p-d}{n-p} + \frac{\mathbf{e}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D}{n-p} \right] \det(\mathbf{I}_p - \mathbf{H}_D) \right\}^{-1}.$$

For the leave- d -out analog of the Andrews-Pregibon statistic, see Exercises 10e, No. 3. Further details on these diagnostics may be found in Belsley et al. [1980].

Using Residuals from Robust Fits

We remarked in Section 10.6.1 that when a point has high leverage, outliers cannot be diagnosed by examining the corresponding least squares residuals. However, this will not be the case if we use a fitting method such as least median squares (LMS) that resists the effect of outliers at high-leverage points. For these fitting methods, the size of the errors *will* be reflected in the size of the residuals. However, due to the problems with LMS regression noted in Section 3.13.2, points having large residuals should be classed as outliers only provisionally and should be subject to further checks. Atkinson [1986] advocates this approach. Once the outliers (say, k in all) corresponding to the large LMS residuals have been identified, we can apply an outlier test as described in Section 10.6.4 to check if any of the k points are real outliers. Atkinson also advocates examining the suspect points using add-one-back diagnostics, which look at the change in the regression based on the "good" points, as the suspect points are added back one at a time. A simple one-at-a-time test can also be used as follows.

Suppose that $\hat{\beta}_G$ and S_G^2 are the usual estimates of β and σ^2 , calculated from the “good” data. If \mathbf{X} is the corresponding $(n-k) \times p$ regression matrix, and (\mathbf{x}_i, Y_i) one of the suspect data points, then the statistic (cf. Section 5.3.1)

$$\frac{Y_i - \mathbf{x}'_i \hat{\beta}_G}{S_G(1 + \mathbf{x}'_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i)^{1/2}}$$

has a t_{n-k-p} distribution when (\mathbf{x}_i, Y_i) is not an outlier. Since this test is repeated k times, some adjustment of the significance level should be made, for example by using a Bonferroni adjustment with a significance level of α/k .

Several other authors describe systematic ways of examining a set of suspect observations to identify outliers. Dempster and Gasko-Green [1981] suggest ordering the observations according to the criteria discussed above, and deleting observations sequentially. They give two rules which indicate when to stop deleting points. Hadi and Simonoff [1993] describe a sequential procedure that modifies an initial set of suspect observations.

Hawkins et al. [1984] describe a method for identifying outliers using elemental regressions (cf. Section 11.12.3). Their method relies on the fact that residuals $e_{iJ} = Y_i - \mathbf{x}'_i \hat{\beta}_J$ from elemental regressions based on clean subsets J will be good estimates of γ_i in the outlier-shift model (10.57). On the other hand, residuals from regressions based on sets J containing outliers will not. Taking a median of all the e_{iJ} over all elemental sets J should provide a good estimate of γ_i , since the median should resist the effects of the contaminated e_{iJ} corresponding to the sets J containing outliers. Hawkins et al. [1984] also consider a form of weighted median, and their paper should be consulted for details.

EXERCISES 10e

- Consider a linear function $\mathbf{d}'\beta$ of β . Show that the change in the estimate $\mathbf{d}'\hat{\beta}$ when the i th observation is deleted is

$$\mathbf{d}'\hat{\beta}(i) - \mathbf{d}'\hat{\beta} = (\mathbf{C}'\mathbf{d})_i e_i / (1 - h_i),$$

where \mathbf{C} is the catcher matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- Show that

$$\mathbf{Y}_D - \mathbf{X}_D \hat{\beta}(D) = (\mathbf{I} - \mathbf{H}_D)^{-1} [\mathbf{Y}_D - \mathbf{X}_D \hat{\beta}(D)].$$

- Suppose that we delete a set D of observations. Show that the “delete d ” version

$$\frac{\det[\mathbf{X}_A(D)'\mathbf{X}_A(D)]}{\det(\mathbf{X}_A \mathbf{X}_A)}$$

of the Andrews-Pregibon statistic can be written as

$$\det(\mathbf{I}_d - \mathbf{H}_D) \left[1 - \frac{\mathbf{e}'_D (\mathbf{I}_d - \mathbf{H}_D)^{-1} \mathbf{e}_D}{\text{RSS}} \right].$$

4. Let $h_{i,A}$ be the i th diagonal element of the hat matrix based on (\mathbf{X}, \mathbf{Y}) rather than \mathbf{X} . Show that $AP(i) = 1 - h_{i,A}$. Hint: Use the results of Section 10.6.3.
5. The influence curve of the least squares estimate $\hat{\beta}$ was derived in Example 3.22. A sample version of the IC (cf. Section 3.13.3) can be defined by taking $\mathbf{z}_0 = (\mathbf{x}_i, Y_i)$, $t = -1/(n-1)$, F as the e.d.f. \hat{F}_n , and by considering the difference $[T(F_t) - T(F)]/t$ rather than the derivative. This leads to the *sample influence curve* SIC_i , given by

$$SIC_i = -(n-1) \left\{ T \left[(n-1)^{-1} (n\hat{F}_n - \delta_i) \right] - T(\hat{F}_n) \right\},$$

where δ_i puts mass 1 at (\mathbf{x}_i, Y_i) .

- (a) Show that $(n-1)^{-1}(n\hat{F}_n - \delta_i)$ is the empirical distribution function calculated from the remaining $(n-1)$ points when the i th observation has been deleted from the sample.
- (b) Hence show that for the least squares estimate functional T ,

$$SIC_i = (n-1)(\hat{\beta}(i) - \hat{\beta}).$$

- (c) The quantity SIC_i is a vector. We can obtain an overall scalar measure by considering the quantity

$$D_i(\mathbf{M}, c) = c^{-1} (SIC_i)' \mathbf{M} (SIC_i),$$

where \mathbf{M} is a positive-definite matrix and c is a positive constant. Show that if we choose $\mathbf{M} = \mathbf{X}'\mathbf{X}$ and $c = pS^2/(n-1)^2$, we obtain Cook's D .

For more details on the influence function approach to outlier detection, see Cook and Weisberg [1982, Chapter 3] and Chatterjee and Hadi [1988, Chapter 5].

10.7 DIAGNOSING COLLINEARITY

We saw in Section 9.7.3 that the existence of an almost linear relationship between the columns of the regression matrix is equivalent to \mathbf{R}_{zz} having one or more small eigenvalues. This is also equivalent to at least one of the variance inflation factors being large. Thus, we can detect collinearity by examining the eigenvalues and the variance inflation factors. If we have a collinearity problem, some of the eigenvalues will be small, and some of the variance inflation factors will be large.

The condition number $\kappa(\mathbf{X}^*)$ of the centered and scaled regression matrix \mathbf{X}^* is a single diagnostic that will allow us to screen the data for large VIFs

and small eigenvalues. The square of the condition number is an upper bound for the VIFs, and for the reciprocal of the smallest eigenvalue. To see this, let \mathbf{c}_i be a vector whose i th element is 1 and the rest are zero. Then, from A.7.2,

$$1 = \mathbf{c}'_i \mathbf{R}_{xx} \mathbf{c}_i \leq \max_{\|\mathbf{c}\|=1} \mathbf{c}' \mathbf{R}_{xx} \mathbf{c} = \lambda_{MAX}. \quad (10.61)$$

Let $\mathbf{T} = (t_{ij})$ be the matrix that diagonalizes \mathbf{R}_{xx} ($= \mathbf{X}^{*'} \mathbf{X}^*$). Then, by (9.56) and (10.61), since $\sum_l t_{jl}^2 = 1$ for all j , it follows that

$$\begin{aligned} VIF_j &= \text{var}[\hat{\gamma}_j]/\sigma^2 \\ &= \sum_l t_{jl}^2 \lambda_l^{-1} \\ &\leq \sum_l t_{jl}^2 \lambda_{MIN}^{-1} \\ &\leq \lambda_{MIN}^{-1} \\ &\leq \lambda_{MAX} \lambda_{MIN}^{-1} \\ &= \kappa(\mathbf{X}^*)^2, \end{aligned}$$

by (9.60). Thus, if the condition number is small, there will be no estimated regression coefficients with large variances. On the other hand, if the condition number is large, at least one eigenvalue must be small, since by Exercises 9d, No. 4 at the end of Section 9.7.5, the largest eigenvalue of \mathbf{R}_{xx} is less than $p - 1$. In this case we should examine the eigenvalues. The eigenvectors corresponding to the small eigenvalues should also be examined to determine the linear combinations that are causing the trouble.

We can also calculate the *variance proportions* (Belsley et al. [1980: p. 106]), which are the proportions of each variance due to each eigenvalue. The variance proportion for variable j and eigenvalue λ_r is the quantity

$$\frac{t_{jr}^2 \lambda_r^{-1}}{\sum_{l=1}^{p-1} t_{jl}^2 \lambda_l^{-1}}.$$

Proportions near unity indicate that most of the variance of a particular estimated coefficient is due to a single small eigenvalue.

10.7.1 Drawbacks of Centering

As discussed in Section 9.7.4, the condition number of \mathbf{X}^* gives an indication of how small changes in \mathbf{X}^* will affect the regression coefficients. If the condition number is small, a small change in \mathbf{X}^* will result in only a small change in $\hat{\gamma}$. However, we may be more interested in how a small change in the *original data* affects the regression coefficients. Under certain circumstances, a *small* change in the original regression matrix \mathbf{X} may cause a *large* change in the centered and scaled regression matrix \mathbf{X}^* . This in turn will cause a large change in $\hat{\gamma}$. The diagnostics based on \mathbf{R}_{xx} cannot detect this.

We will see in Section 11.7 that the matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ may not be computed accurately if the coefficient of variation of one or more of the columns of \mathbf{X} is small. A similar kind of instability applies here. If the elements in the j th column have a small coefficient of variation, CV_j , a small change in the j th column of \mathbf{X} can produce a large change in $s_j^2 = \sum_i (x_{ij} - \bar{x}_j)^2$ and hence in \mathbf{R}_{xx} . This, in turn, can produce large changes in the estimated regression coefficients.

To demonstrate this effect, let $\mathbf{P} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}'_n$, and suppose that the j th column $\mathbf{x}^{(j)}$ of the original regression matrix is perturbed by a small vector $\boldsymbol{\delta}^{(j)} = (\delta_{1j}, \dots, \delta_{nj})'$, where $\|\boldsymbol{\delta}\| < \epsilon \|\mathbf{x}^{(j)}\|$. Then, using the same arguments and notation as in Section 9.7.4, and noting that $\boldsymbol{\delta}^{(j)'}\mathbf{P}\boldsymbol{\delta}^{(j)} = \|\mathbf{P}\boldsymbol{\delta}^{(j)}\|^2 \leq \|\boldsymbol{\delta}^{(j)}\|^2$ (by Example 9.5 in Section 9.7.4), the relative change in s_j^2 ($= \|\mathbf{P}\mathbf{x}^{(j)}\|^2$) is

$$\begin{aligned} \frac{\|\mathbf{P}(\mathbf{x}^{(j)} + \boldsymbol{\delta}^{(j)})\|^2 - \|\mathbf{P}\mathbf{x}^{(j)}\|^2}{\|\mathbf{P}\mathbf{x}^{(j)}\|^2} &= \frac{|2\boldsymbol{\delta}^{(j)'}\mathbf{P}\mathbf{x}^{(j)} + \boldsymbol{\delta}^{(j)'}\mathbf{P}\boldsymbol{\delta}^{(j)}|}{\|\mathbf{P}\mathbf{x}^{(j)}\|^2} \\ &\leq \frac{2\|\boldsymbol{\delta}^{(j)}\| \|\mathbf{P}\mathbf{x}^{(j)}\| + \|\boldsymbol{\delta}^{(j)}\|^2}{\|\mathbf{P}\mathbf{x}^{(j)}\|^2} \quad (\text{by A.4.11}) \\ &= \frac{2\|\boldsymbol{\delta}^{(j)}\|}{\|\mathbf{P}\mathbf{x}^{(j)}\|} + \frac{\|\boldsymbol{\delta}^{(j)}\|^2}{\|\mathbf{P}\mathbf{x}^{(j)}\|^2} \\ &= \frac{2\epsilon \|\mathbf{x}^{(j)}\|}{s_j} + \frac{\epsilon^2 \|\mathbf{x}^{(j)}\|^2}{s_j^2} \\ &= 2\epsilon(1 + CV_j^{-2})^{1/2} + \epsilon^2(1 + CV_j^{-2}), \end{aligned} \tag{10.62}$$

by (9.71), where $CV_j = s_j/(n^{1/2}|\bar{x}_j|)$. This bound shows that provided CV_j is not too small, small changes in the original data will not produce much relative change in \mathbf{R}_{xx} . However, if the perturbations are such that $\boldsymbol{\delta}^{(j)'}\mathbf{1}_n = 0$ and $\boldsymbol{\delta}^{(j)'}\mathbf{x}^{(j)} = 0$, then $\mathbf{P}\boldsymbol{\delta}^{(j)} = \boldsymbol{\delta}^{(j)}$, $\boldsymbol{\delta}^{(j)'}\mathbf{x}^{(j)} = 0$ and the second term in (10.62) is attained. This shows that if the data in a column have a small CV, then small changes in the column can result in very large relative changes in s_j^2 , and hence in \mathbf{R}_{xx} .

The inability of diagnostics based on \mathbf{R}_{xx} to detect instabilities due to small values of the CV_j 's has been pointed out by Belsley [1984], who gives an example in which the centered and scaled data has condition number 1 (i.e., the columns of the regression matrix are perfectly orthogonal) but the CV_j 's are small. Using this example, he shows how a small perturbation in the original data causes a large change in the centered data and hence a large change in the regression coefficients.

Belsley advocates using the condition number of a scaled, but not centered, version of \mathbf{X} as a composite diagnostic that will reveal both small eigenvalues

in \mathbf{R}_{xx} and small CV_j 's. To see why this is so, consider the scaling

$$\check{x}_{ij} = \frac{x_{ij}}{\{\sum_{i=1}^n x_{ij}^2\}^{1/2}} \quad (j = 0, \dots, p-1),$$

so that the columns of \mathbf{X} , including the first corresponding to the constant term, have all been scaled to have unit length. Let the resulting matrix be $\check{\mathbf{X}} = (n^{-1/2}\mathbf{1}_n, \check{\mathbf{x}}^{(1)}, \dots, \check{\mathbf{x}}^{(p-1)})$. Also let $\check{\lambda}_{\text{MIN}}$ and $\check{\lambda}_{\text{MAX}}$ be the minimum and maximum eigenvalues of the scaled (but not centered) matrix $\check{\mathbf{X}}'\check{\mathbf{X}}$. Note that, by the scaling and the Cauchy-Schwartz inequality, all the elements of $\check{\mathbf{X}}'\check{\mathbf{X}}$ are less than 1 in magnitude.

For the eigenvalues of $\check{\mathbf{X}}'\check{\mathbf{X}}$ to be an effective diagnostic tool, they need to reflect both causes of instability: namely, small values of CV_j and small eigenvalues of \mathbf{R}_{xx} . We will show that if any of the CV_j 's or eigenvalues of \mathbf{R}_{xx} are small, then $\check{\mathbf{X}}'\check{\mathbf{X}}$ must have a small eigenvalue, which we need to check for.

We begin by showing that $\lambda_{\text{MIN}} \geq \check{\lambda}_{\text{MIN}}$, where λ_{MIN} is the smallest eigenvalues of \mathbf{R}_{xx} and $\check{\lambda}_{\text{MIN}}$ is the smallest eigenvalue of $\check{\mathbf{X}}'\check{\mathbf{X}}$. This will prove that if \mathbf{R}_{xx} has a small eigenvalue, so must $\check{\mathbf{X}}'\check{\mathbf{X}}$. Now any linear combination of the columns of the centered and scaled regression matrix \mathbf{X}^* can be expressed as a linear combination of the columns of $\check{\mathbf{X}}$, since

$$\begin{aligned} \mathbf{X}^* \mathbf{c} &= c_1 \mathbf{x}^{*(1)} + \dots + c_{p-1} \mathbf{x}^{*(p-1)} \\ &= c_1 (\mathbf{x}^{(1)} - \bar{x}_1 \mathbf{1}_n) / s_1 + \dots + c_{p-1} (\mathbf{x}^{(p-1)} - \bar{x}_{p-1} \mathbf{1}_n) / s_{p-1} \\ &= -(c_1 \bar{x}_1 / s_1 + \dots + c_{p-1} \bar{x}_{p-1} / s_{p-1}) \mathbf{1}_n \\ &\quad + c_1 \mathbf{x}^{(1)} / s_1 + \dots + c_{p-1} \mathbf{x}^{(p-1)} / s_{p-1} \\ &= \check{c}_0 n^{-1/2} \mathbf{1}_n + \check{c}_1 \check{\mathbf{x}}_1 + \dots + \check{c}_{p-1} \check{\mathbf{x}}_{p-1} \\ &= \check{\mathbf{X}} \check{\mathbf{c}}, \end{aligned} \tag{10.63}$$

say, where

$$\check{c}_0 = -\sqrt{n}(c_1 \bar{x}_1 / s_1 + \dots + c_{p-1} \bar{x}_{p-1} / s_{p-1})$$

and

$$\check{c}_j = c_j \left(\sum_{i=1}^n x_{ij}^2 \right)^{1/2} / s_j.$$

Now let $\mathbf{c} = (c_1, \dots, c_{p-1})'$ be the eigenvector having unit length corresponding to λ_{MIN} , and $\check{\mathbf{c}}$ be the $p+1$ vector calculated from \mathbf{c} as described above. Then, by A.7.2,

$$\begin{aligned} \lambda_{\text{MIN}} &= \mathbf{c}' \mathbf{R}_{xx} \mathbf{c} \\ &= \mathbf{c}' \mathbf{X}^* \mathbf{X}^* \mathbf{c} \\ &= \check{\mathbf{c}}' \check{\mathbf{X}}' \check{\mathbf{X}} \check{\mathbf{c}} \\ &\geq \check{\lambda}_{\text{MIN}} \|\check{\mathbf{c}}\|^2. \end{aligned}$$

But

$$\begin{aligned}\|\check{\mathbf{c}}\|^2 &= n(c_1\bar{x}_1/s_1 + \cdots + c_{p-1}\bar{x}_{p-1}/s_{p-1})^2 \\ &\quad + c_1^2\|\mathbf{x}^{(1)}\|^2/s_1^2 + \cdots + c_{p-1}^2\|\mathbf{x}^{(p-1)}\|^2/s_{p-1}^2 \\ &\geq c_1^2 + \cdots + c_{p-1}^2 \\ &= 1,\end{aligned}$$

since $\|\mathbf{x}^{(j)}\|^2/s_j^2 = 1 + CV_j^{-2} \geq 1$. Thus $\lambda_{\text{MIN}} \geq \check{\lambda}_{\text{MIN}}$.

Next we show that if CV_j is small for some j , then $\check{\mathbf{X}}'\check{\mathbf{X}}$ has a small eigenvalue. Let \mathbf{c}_j be a vector with first element equal to $-\bar{x}_j n^{1/2}$, j th element equal to $\|\mathbf{x}^{(j)}\|$, and the other elements zero. Then, by A.7.2,

$$\begin{aligned}\check{\lambda}_{\text{MIN}} &= \min_{\mathbf{c}} \frac{\mathbf{c}'\check{\mathbf{X}}'\check{\mathbf{X}}\mathbf{c}}{\mathbf{c}'\mathbf{c}} \\ &\leq \frac{\mathbf{c}'_j\check{\mathbf{X}}'\check{\mathbf{X}}\mathbf{c}_j}{\mathbf{c}'_j\mathbf{c}_j} \\ &= \frac{\|\check{\mathbf{X}}\mathbf{c}_j\|^2}{\mathbf{c}'_j\mathbf{c}_j} \\ &= \frac{\sum_i(x_{ij} - \bar{x}_j)^2}{\|\mathbf{x}^{(j)}\|^2 + n\bar{x}_j^2} \\ &= \frac{\sum_i(x_{ij} - \bar{x}_j)^2}{\sum_i x_{ij}^2 - n\bar{x}_j^2 + 2n\bar{x}_j^2} \\ &= \frac{CV_j^2}{CV_j^2 + 2},\end{aligned}$$

so that if CV_j is small, so is the smallest eigenvalue $\check{\lambda}_{\text{MIN}}$.

Since the largest eigenvalue of $\check{\mathbf{X}}'\check{\mathbf{X}}$ is greater than 1 (see Exercises 10f, No. 1), then a small CV_j or small eigenvalue of \mathbf{R}_{xx} must necessarily result in a large condition number for $\check{\mathbf{X}}$. For this reason, $\kappa(\check{\mathbf{X}})$ is a better diagnostic than $\kappa(\mathbf{X}^*)$.

10.7.2 Detection of Points Influencing Collinearity

The collinearity diagnostics introduced in Section 10.7, like many other aspects of regression, are vulnerable to the effects of high-leverage points. We illustrate the problems that can arise with a simple example.

Suppose that we have a regression with $p = 3$ and two centered and scaled explanatory variables x_1^* and x_2^* . Figure 10.9(a) shows a plot of x_2^* versus x_1^* . Due to the high-leverage point A , the correlation between the two explanatory variables will not be large, so the diagnostics discussed in the previous sections will not reveal any problems. However, if the point A is removed,

the correlation will increase dramatically. Consequently, the VIFs and the condition number of \mathbf{X}^* will become large.

A reverse situation can also occur: Consider the plot in Figure 10.9(b). The correlation is high but is reduced to almost zero if the high-leverage point A is removed. These two examples indicate how high-leverage points can have a strong effect on collinearity, as well as other aspects of the regression. Indeed, as Gunst and Mason [1985] point out, the effects can be arbitrarily large. We need to be able to detect situations where the collinearity is influenced by a single point.

The ordinary regression influence diagnostics (in particular, the hat matrix diagonals) will give an indication of possible trouble. A more focused diagnostic is to compute the condition number $\kappa_{(-i)}$ of the regression matrix with the i th row deleted. Comparison of $\kappa_{(-i)}$ with the condition number of the full matrix will indicate any points causing a marked change in the condition number. Note that it is absolute rather than relative change that is important; a 50% change in a small condition number is not as relevant as a 50% change in a large one.

10.7.3 Remedies for Collinearity

We begin by emphasizing that centering and scaling do not “cure” collinearity. Scaling merely changes the units of measurement and leads to measuring variability on a different scale. Centering merely reparameterizes the regres-

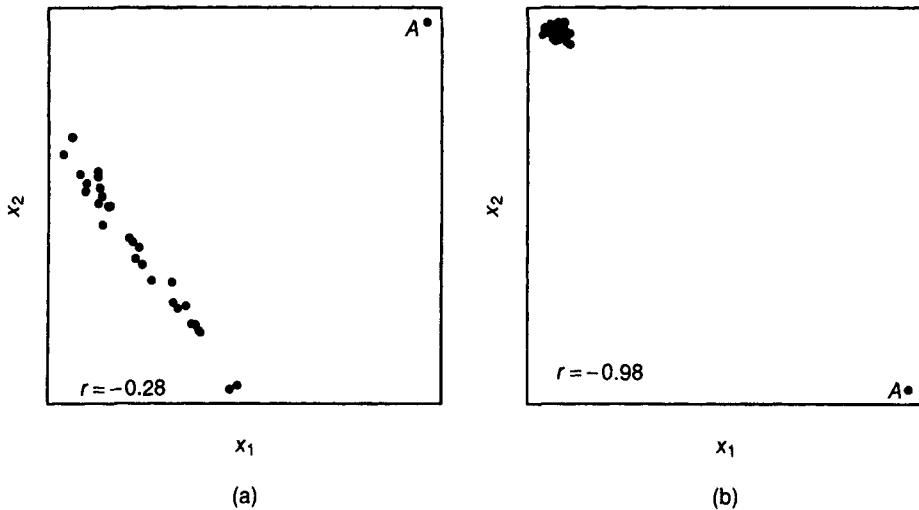


Fig. 10.9 Effect of outliers on collinearity diagnostics

sion surface, substituting a parameter that can be estimated accurately for one that cannot, while the estimated surface (i.e., the set of fitted values) remains the same. (An overall measure of the accuracy of the fitted values is $E\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2$, which is just the ME considered in Chapter 12. This is invariant under centering and scaling, and under more general transformations; see Smith and Campbell [1980] and Exercises 10f, No. 2.)

Centering and scaling lead to simple formulas, and allow us to measure regression coefficients on a common scale, that of the units of the response. Note that we have not advocated centering and scaling the response: There is no advantage in doing so and the usual distribution theory does not apply to the centered and scaled Y .

There are three ways to overcome the effects of collinearity. Since collinearity arises when the data are deficient, an obvious solution is to collect fresh data to repair the deficiencies in the regression matrix. Obviously, this is not always possible, for example when the explanatory variables are strongly correlated in the population from which the data are drawn.

Another remedy is simply to discard variables until the remaining set is not collinear. For example, if there is a single approximate linear relationship between the variables, discarding a single variable will solve the problem. However, Belsley [1991: p. 301] disparages this approach, on the grounds that the variable deleted may well have a relationship with the response. He points out that collinearity is a property of the observed data, and not of the underlying relationship between the variables. For more on the general problem of selecting which variables to include in a regression, see Chapter 12.

A final remedy, which we will see in Chapter 12 can greatly aid prediction, is to abandon the use of least squares and use a biased estimation method such as ridge regression.

Ridge Regression

Ridge regression was introduced by Hoerl and Kennard [1970a,b] as a way of dealing with collinear data in an estimation context rather than in prediction. Consider the centered and scaled regression model (9.49). The *ridge estimate* of γ is defined as

$$\hat{\gamma}(k) = (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-1}\mathbf{X}^{*'}\mathbf{Y}, \quad (10.64)$$

where k is a positive parameter, whose value must be specified. When $k = 0$, the ridge estimate reduces to the least squares estimate.

The ridge estimate has an interpretation as a Bayes estimate. We first write the centered and scaled model as [cf. (9.50)]

$$\mathbf{Y} = \mathbf{X}_s\delta + \epsilon,$$

where $\mathbf{X}_s = (\mathbf{1}, \mathbf{X}^*)$ and $\delta = (\alpha_0, \gamma')'$. In Section 3.12 we saw that if δ has a $N_p(\mathbf{m}, \sigma^2\mathbf{V})$ prior distribution, then the posterior mean of δ is

$$(\mathbf{X}'_s\mathbf{X}_s + \mathbf{V}^{-1})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'_s\mathbf{Y}).$$

If we choose $\mathbf{m} = \mathbf{0}$ and

$$\mathbf{V} = \begin{bmatrix} c^{-1} & \mathbf{0} \\ \mathbf{0} & k^{-1}\mathbf{I}_{p-1} \end{bmatrix},$$

then from Exercises 10f, No. 3) the posterior mean (the Bayes estimate) of γ is exactly the ridge estimate (10.64).

We can also interpret the ridge estimate as the least squares estimate that results if the data are augmented. As we saw above, one way to remedy the effect of collinearity is to collect new data. Imagine augmenting the observed data $(\mathbf{X}^*, \mathbf{Y})$ with new data $(k\mathbf{I}_{p-1}, \mathbf{0})$. Then the least squares estimate calculated from the augmented data

$$\begin{pmatrix} \mathbf{X}^* & \mathbf{Y} \\ k\mathbf{I}_{p-1} & \mathbf{0} \end{pmatrix}$$

is just the ridge estimate (10.64).

Thus the ideas of biased estimation (at least in the case of ridge estimation) are connected with the idea of observing extra (notional) data and with Bayes estimation. It follows that the ridge approach will be a good method if the $N_{p-1}(\mathbf{0}, \sigma^2 k^{-1} \mathbf{I}_{p-1})$ prior for γ is appropriate, but not otherwise. This is considered in more detail by Smith and Campbell [1980], Draper and Smith [1998: p. 391], and in Section 12.9.2.

We have defined the ridge estimate in terms of the centered and scaled regression model. Some writers do not assume that the data have been centered and scaled, and define the estimate in terms of the original regression matrix \mathbf{X} rather than the centered and scaled matrix \mathbf{X}^* . However, this leads to estimates that are not scale invariant, unlike the least squares estimate (cf. Brown [1977] and Smith and Campbell [1980] for a discussion).

The ridge estimate can be written (using the notation $\mathbf{X}^*\mathbf{X}^* = \mathbf{R}_{xx}$) as

$$\begin{aligned} \hat{\gamma}(k) &= (\mathbf{R}_{xx} + k\mathbf{I}_{p-1})^{-1} \mathbf{X}^{*\prime} \mathbf{Y} \\ &= (\mathbf{R}_{xx} + k\mathbf{I}_{p-1})^{-1} \mathbf{R}_{xx} \mathbf{R}_{xx}^{-1} \mathbf{X}^{*\prime} \mathbf{Y} \\ &= (\mathbf{R}_{xx} + k\mathbf{I}_{p-1})^{-1} \mathbf{R}_{xx} \hat{\gamma} \\ &= (\mathbf{I}_{p-1} + k\mathbf{R}_{xx}^{-1})^{-1} \hat{\gamma} \\ &= \mathbf{C} \hat{\gamma}, \end{aligned}$$

say. Thus, the ridge estimate is clearly biased, as $\mathbf{C} \neq \mathbf{I}_{p-1}$. We can study its accuracy for different values of k by examining its mean-squared error (MSE), given by

$$\begin{aligned} \text{MSE} &= E[\|\hat{\gamma}(k) - \gamma\|^2] \\ &= E[\|\mathbf{C} \hat{\gamma} - \gamma\|^2] \\ &= E[\|\mathbf{C}(\hat{\gamma} - \gamma) + (\mathbf{C} - \mathbf{I}_{p-1})\gamma\|^2] \\ &= E[\|\mathbf{C}(\hat{\gamma} - \gamma)\|^2] + \|(\mathbf{C} - \mathbf{I}_{p-1})\gamma\|^2. \end{aligned} \tag{10.65}$$

Since $\hat{\gamma} - \gamma$ has mean zero and variance-covariance matrix $\sigma^2 \mathbf{R}_{\mathbf{zz}}^{-1}$, it follows by Theorem 1.5 that the first term in (10.65) is $\sigma^2 \text{tr}(\mathbf{R}_{\mathbf{zz}}^{-1} \mathbf{C}' \mathbf{C})$. To simplify this, we use the spectral decomposition $\mathbf{R}_{\mathbf{zz}} = \mathbf{T} \Lambda \mathbf{T}'$ (A.1.4). Then $\mathbf{R}_{\mathbf{zz}}^{-1} = \mathbf{T} \Lambda^{-1} \mathbf{T}'$, $\mathbf{I}_n = \mathbf{T} \mathbf{T}'$ and $\mathbf{C} = \mathbf{T} \mathbf{D} \mathbf{T}'$, where \mathbf{D} is also diagonal with diagonal elements $\lambda_j / (k + \lambda_j)$. Thus the first term of (10.65) is

$$\begin{aligned}\sigma^2 \text{tr}(\mathbf{R}_{\mathbf{zz}}^{-1} \mathbf{C}' \mathbf{C}) &= \sigma^2 \text{tr}(\mathbf{T} \Lambda^{-1} \mathbf{T}' \mathbf{T} \mathbf{D} \mathbf{T}' \mathbf{T} \mathbf{D} \mathbf{T}') \\ &= \sigma^2 \text{tr}(\mathbf{T} \Lambda^{-1} \mathbf{D}^2 \mathbf{T}') \\ &= \sigma^2 \text{tr}(\Lambda^{-1} \mathbf{D}^2 \mathbf{T}' \mathbf{T}) \\ &= \sigma^2 \text{tr}(\Lambda^{-1} \mathbf{D}^2) \\ &= \sigma^2 \sum_{j=1}^{p-1} \frac{\lambda_j}{(k + \lambda_j)^2}.\end{aligned}$$

If we set $\alpha = \mathbf{T}' \gamma$, then the second term in (10.65) is

$$\begin{aligned}\gamma' (\mathbf{C} - \mathbf{I}_{p-1})' (\mathbf{C} - \mathbf{I}_{p-1}) \gamma &= \alpha' (\mathbf{D} - \mathbf{I}_{p-1}) (\mathbf{D} - \mathbf{I}_{p-1}) \alpha \\ &= \sum_{j=1}^{p-1} \frac{\alpha_j^2 k^2}{(k + \lambda_j)^2},\end{aligned}$$

so that the MSE is

$$\text{MSE} = \sum_{j=1}^{p-1} \frac{\alpha_j^2 k^2 + \sigma^2 \lambda_j}{(k + \lambda_j)^2}. \quad (10.66)$$

The derivative of this with respect to k is

$$\sum_{j=1}^{p-1} \frac{2\lambda_j(\alpha_j^2 k - \sigma^2)}{(k + \lambda_j)^3},$$

which is negative for small positive values of k , so that for k sufficiently small, MSE decreases as k increases.

In principle, to find the value of k leading to the smallest MSE, we need only solve

$$\sum_{j=1}^{p-1} \frac{2\lambda_j(\alpha_j^2 k - \sigma^2)}{(k + \lambda_j)^3} = 0.$$

However, the minimizing value obviously depends on the unknown parameters α and σ^2 . We can substitute the estimates $\hat{\alpha} = \mathbf{T}' \hat{\gamma}$ and S^2 from a least squares fit and solve the resulting equation. Unfortunately, there is no guarantee that the resulting estimate of k will lead to a smaller MSE than least squares. However, in a simulation study, Dempster et al. [1977] reported that this method, which they call SRIDG, performed well.

They also advocated using another method, called RIDGM, which uses the fact (cf. Exercises 10f, No. 4) that the $\hat{\alpha}_j$'s have independent $N(0, \sigma^2(k^{-1} +$

$\lambda_j^{-1})$) distributions when γ has the $N_{p-1}(0, \sigma^2 k^{-1} I_{p-1})$ prior discussed above. The quantities $\hat{\alpha}_j^2 k \lambda_j / \{\sigma^2(k + \lambda_j)\}$ then all have independent χ_1^2 distributions, with expectation 1, which implies that

$$\sum_{j=1}^{p-1} \frac{\hat{\alpha}_j^2 k \lambda_j}{\sigma^2(k + \lambda_j)} \quad (10.67)$$

has expectation $p - 1$. Thus, substituting S^2 for σ^2 in (10.67), and equating to $p - 1$, gives an equation that can be solved for k . This yields the estimate RIDGM, which was found to be slightly better than SRIDG in the study by Dempster et al.

If the centered and scaled explanatory variables are orthogonal, then the matrix R_{xx} is an identity matrix and the eigenvalues are all 1. In this case the SRIDG method reduces to solving the equation

$$\frac{2 \sum_{j=1}^{p-1} (\hat{\alpha}_j^2 k - S^2)}{(k + 1)^3} = 0,$$

which has solution

$$\begin{aligned} \hat{k} &= \frac{(p - 1)S^2}{\sum_{j=1}^{p-1} \hat{\alpha}_j^2} \\ &= \frac{(p - 1)S^2}{\|\hat{\gamma}\|^2}. \end{aligned} \quad (10.68)$$

The last expression follows from the fact that $\hat{\alpha} = T'\hat{\gamma}$ and T is orthogonal and consequently preserves the lengths of vectors. The estimate (10.68) is due to Hoerl et al. [1975]. In another simulation study, Hoerl et al. [1986] report that it gives good results even in non-orthogonal regressions, being much superior to least squares.

In their original paper, Hoerl and Kennard proposed a graphical method for choosing k . They recommended plotting the components of the vector $\hat{\gamma}(k)$ against k and choosing a value of k for which the coefficients are not changing rapidly and have "sensible" signs. This plot, called the *ridge trace*, has no objective basis and has been disparaged by most subsequent writers on ridge regression.

Many other estimates of k have been proposed in the literature, and many published studies compare the estimates using simulation. Some of these studies have been criticized as favoring one estimate or the other, so no clear estimate has emerged as superior (cf. Draper and Van Nostrand [1979]). It should be noted that some of these studies use a different criterion from MSE. We discuss how to estimate k using prediction error as a criterion in Section 12.9.2.

Principal Component Regression

Principal component (PC) regression is another biased estimation method, based on the spectral decomposition of \mathbf{R}_{xx} . Let $\mathbf{t}_1, \dots, \mathbf{t}_{p-1}$ be the columns of the orthogonal matrix \mathbf{T} in the spectral decomposition of \mathbf{R}_{xx} . Then for the centered and scaled model (9.49), we can write the least squares estimate (LSE) as

$$\begin{aligned}\hat{\gamma} &= \mathbf{R}_{xx}^{-1} \mathbf{X}^{*\prime} \mathbf{Y} \\ &= \mathbf{T} \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \mathbf{Y} \\ &= \sum_{l=1}^{p-1} \mathbf{t}_l Z_l\end{aligned}\quad (10.69)$$

where $\mathbf{Z} = (Z_1, \dots, Z_{p-1})'$ is the vector $\mathbf{Z} = \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \mathbf{Y}$. The name principal component regression comes from the fact that \mathbf{Z} is the LSE when \mathbf{Y} is regressed on the principal components $\mathbf{X}^* \mathbf{T}$ of \mathbf{X}^* . This follows from the fact that

$$(\mathbf{T}' \mathbf{X}^{*\prime} \mathbf{X}^* \mathbf{T})^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \mathbf{Y} = (\mathbf{T}' \mathbf{R}_{xx} \mathbf{T})^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \mathbf{Y} = \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \mathbf{Y} = \mathbf{Z}.$$

Hence

$$\begin{aligned}E[\mathbf{Z}] &= \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} E[\mathbf{Y}] \\ &= \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} (\alpha_0 \mathbf{1}_n + \mathbf{X}^* \boldsymbol{\gamma}) \\ &= \Lambda^{-1} \mathbf{T}' \mathbf{T} \Lambda \mathbf{T}' \boldsymbol{\gamma} \\ &= \mathbf{T}' \boldsymbol{\gamma} \\ &= \boldsymbol{\alpha}\end{aligned}$$

and

$$\begin{aligned}\text{Var}[\mathbf{Z}] &= \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \text{Var}[\mathbf{Y}] \mathbf{X}^* \mathbf{T} \Lambda^{-1} \\ &= \Lambda^{-1} \mathbf{T}' \mathbf{X}^{*\prime} \sigma^2 \mathbf{I}_n \mathbf{X}^* \mathbf{T} \Lambda^{-1} \\ &= \sigma^2 \Lambda^{-1} \Lambda \Lambda^{-1} \\ &= \sigma^2 \Lambda^{-1}.\end{aligned}$$

From (10.69) we have

$$\text{Var}[\hat{\gamma}] = \sigma^2 \mathbf{R}_{xx}^{-1} = \sigma^2 \mathbf{T}' \Lambda^{-1} \mathbf{T} = \sigma^2 \sum_{l=1}^{p-1} \mathbf{t}_l \mathbf{t}_l' \lambda_l^{-1}. \quad (10.70)$$

Assuming that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{p-1}$, this suggests dropping the terms in the sum (10.70) corresponding to small eigenvalues, and considering an estimate of the form

$$\hat{\gamma}(r) = \sum_{l=1}^r \mathbf{t}_l Z_l, \quad (10.71)$$

where the integer r is chosen so that the eigenvalues $\lambda_{r+1}, \dots, \lambda_p$ are all “small” (i.e., of a smaller order of magnitude than the others).

To calculate the MSE, we write

$$\gamma = \mathbf{T}\alpha = \sum_{l=1}^{p-1} t_l \alpha_l,$$

so that

$$\hat{\gamma}(r) - \gamma = \sum_{l=1}^r t_l (Z_l - \alpha_l) - \sum_{l=r+1}^p t_l \alpha_l.$$

Exploiting the fact that the vectors t_l are orthogonal, we get

$$\|\hat{\gamma}(r) - \gamma\|^2 = \sum_{l=1}^r (Z_l - \alpha_l)^2 + \sum_{l=r+1}^{p-1} \alpha_l^2,$$

so

$$E[\|\hat{\gamma}(r) - \gamma\|^2] = \sigma^2 \sum_{l=1}^r \lambda_l^{-1} + \sum_{l=r+1}^p \alpha_l^2.$$

The estimate relies on the trade-off between the sizes of the small eigenvalues, which are known, and the corresponding α_j 's, which are not. Estimating α_j with $\hat{\alpha}_j$ is not very accurate for the α_j corresponding to the small eigenvalues, since $\text{Var}[\hat{\alpha}_j] = \sigma^2 \lambda_j^{-1}$. Several authors (e.g., Dempster et al. [1977], Hoerl et al. [1986]) have reported simulations comparing PC regression with ridge and have found it inferior. Gunst and Mason [1977] describe a study where PC regression performed well but under restrictive circumstances that unduly favored PC over ridge.

Another drawback of PC regression is that it is quite possible for the discarded components to be the ones having the strongest relationship with the response. In an extreme case we might have $\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$, so that in forming the PC estimate, we discard all the principal components that are related to the mean response and retain only variables that contribute nothing to the regression. Further discussion of this point may be found in Hadi and Ling [1998]. A comparison of ridge and PC regression may be found in Frank and Friedman [1993].

EXERCISES 10f

1. Show that the largest eigenvalue $\check{\lambda}_{\text{MAX}}$ of $\check{\mathbf{X}}'\check{\mathbf{X}}$ satisfies

$$1 \leq \check{\lambda}_{\text{MAX}} \leq p.$$

2. Show that

$$E[\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2] = \sigma^2 + E[\|\mathbf{X}^*\hat{\gamma} - \mathbf{X}^*\gamma\|^2].$$

3. Prove that with the priors of Section 10.7.3, the posterior mean of γ is just the ridge estimate (10.64).
4. Prove that with the priors of Section 10.7.3, the prior distribution of $\hat{\alpha}$ is $N_{p-1}[\mathbf{0}, \sigma^2(\Lambda^{-1} + k^{-1}\mathbf{I}_{p-1})]$.

MISCELLANEOUS EXERCISES 10

1. Consider a regression with regression matrix \mathbf{X} and LSE $\hat{\beta}$. Suppose that a new case with data (\mathbf{x}, Y) is added. Show that the residual sum of squares is increased by an amount $e^2/(1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})$, where $e = Y - \mathbf{x}'\hat{\beta}$.
2. Consider deleting a variable x_j from a regression. Let \mathbf{H} and \mathbf{H}_j be the hat matrices with and without the corresponding column $\mathbf{x}^{(j)}$ included.

(a) Show that

$$\mathbf{H} = \mathbf{H}_j + \frac{(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{H}_j)}{\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}}.$$

(b) Let $\eta_{ij} = [(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}]_i$. Show that if h_i and $h_i^{(j)}$ are the diagonal elements of \mathbf{H} and \mathbf{H}_j , then

$$h_i = h_i^{(j)} + \frac{\eta_{ij}^2}{\sum_k \eta_{kj}^2}. \quad (10.72)$$

- (c) Explain how the second term on the right of (10.72) can be interpreted as the leverage of the i th case in the added variable plot for variable j . How can you use added variable plots to determine which variables contribute to the leverage of observation j ?
3. The correlation that exists between the residuals in a regression complicates the detection of departures from the standard regression model. For this reason, various proposals for independent and identically distributed “residual-like” quantities have been proposed. For example, Theil [1965, 1968] (see also Grossman and Styan [1972]) introduced BLUS residuals $\hat{\epsilon}$ having the following properties:
 - (i) $\hat{\epsilon} = \mathbf{AY}$ for some $(n-p) \times n$ matrix \mathbf{A} (i.e., linear).
 - (ii) $\text{Var}[\hat{\epsilon}] = \sigma^2 \mathbf{I}_{n-p}$ (i.e., scalar).
 - (iii) $E[\hat{\epsilon}] = \mathbf{0}$ (i.e., unbiased).
 - (iv) \mathbf{A} is chosen to minimize $E[||\hat{\epsilon} - \epsilon_1||^2]$, where ϵ_1 is some fixed $(n-p)$ -dimensional subvector of ϵ (i.e., best).

Thus, the BLUS residuals are the $N_{n-p}(\mathbf{0}, \sigma^2 \mathbf{I}_{n-p})$ linear combination of \mathbf{Y} that best approximates some $(n - p)$ -subset of the errors.

- (a) Assume that \mathbf{X} is of full rank p . Let $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ be an $n \times n$ orthogonal matrix such that the columns of \mathbf{Q}_1 are an orthonormal basis for $C(\mathbf{X})$. (For example, \mathbf{Q} could be the matrix arising in the QR decomposition discussed in Section 11.3.)

Show that any matrix \mathbf{A} that satisfies (ii) and (iii) above must be of the form $\mathbf{T}\mathbf{Q}'_2$ for some $(n - p) \times (n - p)$ orthogonal matrix \mathbf{T} .

- (b) Let \mathbf{J} be an $(n - p) \times n$ submatrix of \mathbf{I}_n such that $\boldsymbol{\varepsilon}_1 = \mathbf{J}\boldsymbol{\varepsilon}$. Show that

$$E[||\hat{\mathbf{e}} - \boldsymbol{\varepsilon}_1||^2] = 2\sigma^2[n - p - \text{tr}(\mathbf{T}\mathbf{Q}'_2\mathbf{J}')].$$

- (c) Let $\mathbf{Q}'_2\mathbf{J}' = \mathbf{U}\Delta\mathbf{V}'$ be the singular value decomposition of $\mathbf{Q}'_2\mathbf{J}'$. (cf. A.12). Show that

$$\text{tr}(\mathbf{T}\mathbf{Q}'_2\mathbf{J}') \leq \text{tr}(\Delta)$$

with equality if and only if $\mathbf{T} = \mathbf{V}\mathbf{U}'$. Hence show that $\mathbf{A} = \mathbf{V}\mathbf{U}'\mathbf{Q}'_2$. Hint: Show that $\text{tr}[(\mathbf{V} - \mathbf{T}\mathbf{U})\Delta(\mathbf{V} - \mathbf{T}\mathbf{U})'] = 2[\text{tr}(\Delta) - \text{tr}(\mathbf{T}\mathbf{Q}'_2\mathbf{J}')]$.

11

Computational Algorithms for Fitting a Regression

11.1 INTRODUCTION

In this chapter we describe the algorithms that are commonly used to fit a linear regression model to a set of data. Fitting the model involves calculating the following quantities:

- The regression coefficients $\hat{\beta}$
- The residual sum of squares and the estimate S^2 of the error variance
- The estimated variance-covariance matrix of $\hat{\beta}$
- The fitted values and residuals
- The hat matrix diagonals
- Test statistics for various hypothesis tests.

We also discuss algorithms for adding and removing variables in a regression model. Most of the discussion deals with fitting by least squares, but we also briefly discuss methods for fitting robust regressions.

11.1.1 Basic Methods

Three basic methods are used to calculate least-squares regression fits. The first forms the sum of squares and cross-products (SSCP) matrix $\mathbf{X}'\mathbf{X}$ and bases most of the calculations on this matrix. Then either Gaussian elimination, the sweep operator, or the Cholesky decomposition is used to solve

the normal equations. Related regression quantities can be generated at the same time by suitably augmenting the SSCP matrix before applying these algorithms.

Alternatively, we can avoid ever forming $\mathbf{X}'\mathbf{X}$ and work directly with \mathbf{X} . The second method uses the QR decomposition of \mathbf{X} to obtain a Cholesky factor directly without ever forming $\mathbf{X}'\mathbf{X}$. The QR decomposition can be calculated by one of three algorithms, the modified Gram–Schmidt algorithm, Householder reflections, or Givens rotations. We describe these in detail below.

The third method uses the singular value decomposition (SVD). This is the most computationally expensive approach, but also the most numerically stable.

11.2 DIRECT SOLUTION OF THE NORMAL EQUATIONS

In this section we examine methods for direct solution of the normal equations. These must first be formed by calculating $\mathbf{X}'\mathbf{Y}$ and the sum of squares and cross products (SSCP) matrix $\mathbf{X}'\mathbf{X}$.

11.2.1 Calculation of the Matrix $\mathbf{X}'\mathbf{X}$

We will assume that our model has a constant term so that the first column of \mathbf{X} consists of 1's. Then, assuming that there are n cases and $p - 1$ variables, the SSCP matrix is

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_{i1} & \cdots & \sum x_{ip-1} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{ip-1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum x_{ip-1} & \sum x_{ip-1}x_{i1} & \cdots & \sum x_{ip-1}^2 \end{pmatrix}. \quad (11.1)$$

Calculating this matrix requires approximately np^2 arithmetic operations, so is quite computationally expensive. We make about a 50% saving in computation by exploiting the symmetry of $\mathbf{X}'\mathbf{X}$. Note that it is customary to count *floating-point operations* (additions, subtractions, multiplications, and divisions) or *flops* when assessing the cost of a computation; see Section 11.8.2 for more details.

The calculation should be done in double precision, since round-off error in formation of the SSCP matrix can result in unacceptable loss of accuracy when calculating the solution to the normal equations. For more detail on this point, see Section 11.7.

11.2.2 Solving the Normal Equations

Assuming that we have accurately formed $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$, we now need to solve the normal equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

(We use \mathbf{b} instead of $\hat{\beta}$ to avoid notational problems because of β_0 .)

Gaussian Elimination

Consider a square nonsingular $p \times p$ matrix \mathbf{A} . Gaussian elimination (GE) is a scheme for reducing \mathbf{A} to upper triangular form (i.e., all entries below the diagonal are zero) by repeated application of the following operation:

Transform row i by subtracting a multiple of row k from row i .

We illustrate for the case when $p = 4$. The reduction can be achieved in three ($= p - 1$) steps, where each step consists of a series of operations of the type above. The algorithm follows.

Algorithm 11.1

Step 1: For $i = 2, 3, 4$, subtract a_{i1}/a_{11} times row 1 from row i . This results in a matrix $\mathbf{A}^{(1)}$ that has all its subdiagonal elements in column 1 equal to zero.

Step 2: For $i = 3, 4$, subtract $a_{i2}^{(1)}/a_{22}^{(1)}$ times row 2 from row i . This results in $\mathbf{A}^{(2)}$, which has its subdiagonal elements in columns 1 and 2 equal to zero.

Step 3: For $i = 4$, subtract $a_{i3}^{(2)}/a_{33}^{(2)}$ times row 3 from row i . This results in $\mathbf{A}^{(3)}$, which has all its subdiagonal elements equal to zero. $\mathbf{A}^{(3)}$ is the desired upper triangular matrix.

Note that this requires that a_{11} , $a_{22}^{(1)}$, and $a_{33}^{(2)}$ all be nonzero. This will be the case if \mathbf{A} is nonsingular.

The steps above can be described in terms of matrix multiplication. Step 1 is equivalent to premultiplication by the matrix

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -a_{21}/a_{11} & 1 & 0 & 0 \\ -a_{31}/a_{11} & 0 & 1 & 0 \\ -a_{41}/a_{11} & 0 & 0 & 1 \end{pmatrix},$$

step 2 to premultiplication by

$$\mathbf{M}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -a_{32}^{(1)}/a_{22}^{(1)} & 1 & 0 \\ 0 & -a_{42}^{(1)}/a_{22}^{(1)} & 0 & 1 \end{pmatrix},$$

and step 3 to premultiplication by

$$\mathbf{M}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -a_{43}^{(2)}/a_{33}^{(2)} & 1 \end{pmatrix}.$$

We can express this process as

$$\mathbf{M}_3 \mathbf{M}_2 \mathbf{M}_1 \mathbf{A} = \mathbf{U},$$

where \mathbf{U} is an upper triangular matrix. The product $\mathbf{M} = \mathbf{M}_3 \mathbf{M}_2 \mathbf{M}_1$ of lower triangular matrices with unit diagonal elements is again a lower triangular matrix with unit diagonals.

In general, to solve the normal equations, we apply $p - 1$ GE steps to the matrix $\mathbf{X}'\mathbf{X}$ and the same steps to the right-hand side, $\mathbf{X}'\mathbf{Y}$. This amounts to multiplying both sides of the normal equations by a lower triangular matrix \mathbf{M} which is the product of matrices such as \mathbf{M}_1 , \mathbf{M}_2 , \mathbf{M}_3 above. The result is

$$\mathbf{Ub} = \mathbf{c},$$

where \mathbf{U} is upper triangular. Upper triangular systems of linear equations can be solved easily by *back-substitution* using the equations

$$b_p = c_p/u_{pp}$$

and

$$b_{p-j} = (c_{p-j} - u_{p-j,p-j+1}b_{p-j+1} - \cdots - u_{p-j,p}b_p)/u_{p-j,p-j}$$

for $j = 1, 2, \dots, p - 1$. From a computational point of view, it is convenient to produce \mathbf{U} and \mathbf{c} together by joining $\mathbf{X}'\mathbf{Y}$ to $\mathbf{X}'\mathbf{X}$ and applying the $p - 1$ Gaussian elimination steps to $(\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{Y})$. We note that $\mathbf{M}\mathbf{X}'\mathbf{X} = \mathbf{U}$, or

$$\mathbf{X}'\mathbf{X} = \mathbf{M}^{-1}\mathbf{U} = \mathbf{LU}, \quad (11.2)$$

say, where \mathbf{L} is lower triangular. This is called the *LU decomposition* of $\mathbf{X}'\mathbf{X}$.

The GE algorithm can also be used to compute the residual sum of squares. If we apply the algorithm to the augmented matrix

$$\mathbf{X}'_A \mathbf{X}_A = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}, \quad (11.3)$$

where $\mathbf{X}_A = (\mathbf{X}, \mathbf{Y})$, we get

$$\begin{pmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{m}' & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{U} & \mathbf{c} \\ \mathbf{0}' & d \end{pmatrix}, \quad (11.4)$$

say. The matrix on the left is lower triangular with unit diagonals and the matrix on the right is upper triangular. Multiplying the matrices and equating blocks gives

$$\mathbf{M}\mathbf{X}'\mathbf{X} = \mathbf{U}, \quad (11.5)$$

$$\mathbf{M}\mathbf{X}'\mathbf{Y} = \mathbf{c}, \quad (11.6)$$

$$\mathbf{m}'\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{Y} = d, \quad (11.7)$$

$$\mathbf{m}'\mathbf{X}'\mathbf{X} + \mathbf{Y}'\mathbf{X} = \mathbf{0}'. \quad (11.8)$$

Thus from (11.5) and (11.6), the matrix \mathbf{U} and \mathbf{c} are the same as those resulting from the previous calculation using $(\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{Y})$. Solving the last two equations, (11.7) and (11.8), gives

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

and

$$d = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y},$$

so that d is the residual sum of squares. Note also that augmenting the matrix effectively performs the back-substitution, as we get the regression coefficients as a bonus.

The accuracy of the Gaussian elimination algorithm can be improved by *pivoting*. If at any stage of the algorithm the divisors $a_{22}^{(1)}, a_{33}^{(2)}, \dots$, etc. (which are called *pivots*) are small, then very large elements can be introduced into the upper triangle \mathbf{U} . These large elements will lead to inaccuracies when the solution is computed by back-substitution. Golub and Van Loan [1996: p. 107] give a simple example of this phenomenon, and Parlett [1980: p. 44] provides some interesting comments. An effective solution is to interchange rows at the beginning of each stage of the computation, so that the largest possible pivot is used. Specifically, at the beginning of the j th stage, we interchange row j with row j' , where

$$a_{j'j}^{(j-1)} = \max_{l \geq j} a_{lj}^{(j-1)}.$$

This strategy, known as *partial pivoting*, is usually enough to avoid the problem of large elements in the upper triangle \mathbf{U} . A more elaborate strategy which involves interchanging columns as well as rows is described in Golub and Van Loan [1996]. If no pivot is sufficiently large, the matrix \mathbf{X} is probably rank deficient.

The GE algorithm is not much used in fitting single regressions, since the Cholesky decomposition provides a more efficient method that exploits the

fact that the SSCP matrix is positive-definite. However, GE plays a role in the Furnival and Wilson method for fitting all possible regressions, which is discussed in Section 12.8.2. In this method, the GE algorithm is used to fit a submodel in the following way: Suppose that we partition \mathbf{X} as $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ and we want to fit the model $(\mathbf{X}_1, \mathbf{X}_3)$. We can, of course, fit the submodel by applying the GE algorithm to $(\mathbf{X}_1, \mathbf{X}_3)$ instead of \mathbf{X} . We will show that this is equivalent to applying the GE algorithm to the whole of \mathbf{X} , but skipping the GE steps corresponding to the variables in \mathbf{X}_2 .

Assuming that \mathbf{X}_1 has r_1 columns, apply r_1 GE steps to $(\mathbf{X}, \mathbf{Y})'(\mathbf{X}, \mathbf{Y})$. Schematically, we get

$$\begin{aligned} & \left(\begin{array}{cccc} \mathbf{M}_{(1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_{(2)} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_{(3)} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{M}_{(4)} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{array} \right) \left(\begin{array}{cccc} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{X}_2 & \mathbf{X}'_1 \mathbf{X}_3 & \mathbf{X}'_1 \mathbf{Y} \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{X}_2 & \mathbf{X}'_2 \mathbf{X}_3 & \mathbf{X}'_2 \mathbf{Y} \\ \mathbf{X}'_3 \mathbf{X}_1 & \mathbf{X}'_3 \mathbf{X}_2 & \mathbf{X}'_3 \mathbf{X}_3 & \mathbf{X}'_3 \mathbf{Y} \\ \mathbf{Y}' \mathbf{X}_1 & \mathbf{Y}' \mathbf{X}_2 & \mathbf{Y}' \mathbf{X}_3 & \mathbf{Y}' \mathbf{Y} \end{array} \right) \\ &= \left(\begin{array}{cccc} * & * & * & * \\ \mathbf{0} & * & * & * \\ \mathbf{0} & * & * & * \\ \mathbf{0} & * & * & * \end{array} \right). \end{aligned} \quad (11.9)$$

Here $\mathbf{M}_{(1)}$ is a product of $(r_1 - 1)$ \mathbf{M}_i -type matrices which reduces $\mathbf{X}'_1 \mathbf{X}_1$ to upper triangular form. Multiples of the rows of $\mathbf{X}'_1 \mathbf{X}_1$ have also been subtracted successively from $\mathbf{X}'_2 \mathbf{X}_1$, $\mathbf{X}'_3 \mathbf{X}_1$, and $\mathbf{Y}' \mathbf{X}_1$ to reduce these matrices to zero. For example, $\mathbf{M}_{(2)} \mathbf{X}'_1 \mathbf{X}_1 + \mathbf{X}'_2 \mathbf{X}_1 = \mathbf{0}$. Therefore, equating blocks, we see that

$$\mathbf{M}_{(j)} = -\mathbf{X}'_j \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \quad (j = 2, 3, 4), \quad (\mathbf{X}_4 \equiv \mathbf{Y}).$$

Thus, the right-hand side of (11.9) is of the form

$$\left(\begin{array}{cccc} * & * & * & * \\ \mathbf{0} & \mathbf{X}'_2 (\mathbf{I} - \mathbf{P}_1) \mathbf{X}_2 & * & * \\ \mathbf{0} & * & \mathbf{X}'_3 (\mathbf{I} - \mathbf{P}_1) \mathbf{X}_3 & \mathbf{X}'_3 (\mathbf{I} - \mathbf{P}_1) \mathbf{Y} \\ \mathbf{0} & * & \mathbf{Y}' (\mathbf{I} - \mathbf{P}_1) \mathbf{X}_3 & \mathbf{Y}' (\mathbf{I} - \mathbf{P}_1) \mathbf{Y} \end{array} \right),$$

where $\mathbf{P}_1 = \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$, so that after r_1 steps, the bottom-right element $\mathbf{Y}' (\mathbf{I} - \mathbf{P}_1) \mathbf{Y}$ is the RSS from fitting the model \mathbf{X}_1 . Now skip the GE steps for the variables in \mathbf{X}_2 , resuming the GE steps for the variables in \mathbf{X}_3 . These steps do not involve the rows and columns corresponding to \mathbf{X}_2 in any way, so the result is exactly the same as if we applied the full GE algorithm to $(\mathbf{X}_1, \mathbf{X}_3)$ rather than \mathbf{X} . The same argument applies if we partition \mathbf{X} into more parts.

It follows that any submodel can be fitted by successively pivoting out the variables present in the submodel, simply omitting the steps corresponding to the absent variables. For example, if we have four variables and we want to fit the model corresponding to the first and third variables, we simply pivot

out the first variable in step 1, skip step 2, pivot out the third variable in step 3, and skip step 4. The RSS for this fit will be the bottom right element in the result.

Sweeping

Sweeping (Goodnight [1979]) is a variation of Gaussian elimination that allows for the simultaneous computation of $\hat{\beta}$, $(\mathbf{X}'\mathbf{X})^{-1}$ and the residual sum of squares. Sweeping also provides an efficient way to add or delete variables from a regression fit and is discussed further in Sections 11.6.2 and 12.8.

The *sweep operator* is applied to a specific row (or column) of a square matrix \mathbf{A} . A sweep on row (or column) r is the transformation of \mathbf{A} to \mathbf{A}^* , where $\mathbf{A}^* = (a_{ij}^*)$ is defined by

$$\begin{aligned} a_{rr}^* &= \frac{1}{a_{rr}}, \\ a_{ir}^* &= -\frac{a_{ir}}{a_{rr}} \quad (i \neq r), \\ a_{rj}^* &= \frac{a_{rj}}{a_{rr}} \quad (j \neq r), \\ a_{ij}^* &= a_{ij} - \frac{a_{ir}a_{rj}}{a_{rr}} \quad (i \neq r, j \neq r). \end{aligned}$$

It follows directly from the definition that:

- (a) Sweeping is reversible (i.e., sweeping twice on row r is equivalent to no sweep).
- (b) Sweeping is commutative (i.e., sweeping on row r and then on row s is equivalent to sweeping on s and then sweeping on r).

The relevance of the sweep operator to regression calculations is that sweeping the augmented $(p+1) \times (p+1)$ matrix

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}$$

successively on columns $1, 2, \dots, p$ yields the matrix

$$\begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} & \hat{\beta} \\ -\hat{\beta} & \text{RSS} \end{pmatrix}.$$

This result is proved in Section 11.6.2.

Cholesky Decomposition

If \mathbf{A} is a $p \times p$ positive-definite matrix, there exists a unique upper triangular matrix \mathbf{R} with positive diagonal elements such that (A.4.10)

$$\mathbf{A} = \mathbf{R}'\mathbf{R}. \tag{11.10}$$

The decomposition (11.10) is called the *Cholesky decomposition* of \mathbf{A} , and \mathbf{R} is called the *Cholesky factor* of \mathbf{A} . Equating coefficients in (11.10) leads to the following algorithm.

Algorithm 11.2

Step 1: Set

$$r_{11} = \sqrt{a_{11}}, \quad (11.11)$$

$$r_{1j} = \frac{a_{1j}}{r_{11}} \quad (j = 2, 3, \dots, p). \quad (11.12)$$

Step 2: For $i = 2, 3, \dots, p - 1$, set

$$\begin{aligned} r_{ij} &= 0 \quad (j = 1, \dots, i - 1), \\ r_{ii} &= \left(a_{ii} - \sum_{l=1}^{i-1} r_{li}^2 \right)^{1/2}, \\ r_{ij} &= \frac{a_{ij} - \sum_{l=1}^{i-1} r_{li} r_{lj}}{r_{ii}} \quad (j = i + 1, \dots, p). \end{aligned} \quad (11.13)$$

Step 3: Set

$$r_{pp} = \left(a_{pp} - \sum_{l=1}^{p-1} r_{li}^2 \right)^{1/2}. \quad (11.14)$$

The diagonal elements are positive if we take the positive square root in (11.11), (11.13) and (11.14).

To apply this decomposition to regression calculations, we first calculate the Cholesky decomposition $\mathbf{R}'\mathbf{R}$ of $\mathbf{X}'\mathbf{X}$, which, by A.4.6, is positive-definite provided that \mathbf{X} has full column rank. Once we have calculated \mathbf{R} , we write the normal equations as

$$\mathbf{R}'\mathbf{R}\mathbf{b} = \mathbf{X}'\mathbf{Y},$$

and then solve

$$\mathbf{R}'\mathbf{z} = \mathbf{X}'\mathbf{Y}$$

for \mathbf{z} and

$$\mathbf{R}\mathbf{b} = \mathbf{z}$$

for \mathbf{b} . Since \mathbf{R} is upper triangular, these equations are easily solved by back-substitution. The RSS is given by

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - (\mathbf{R}\hat{\beta})'\mathbf{R}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{z}'\mathbf{z}. \end{aligned}$$

This calculation can be done efficiently by calculating the Cholesky decomposition of the augmented matrix

$$\mathbf{X}'_A \mathbf{X}_A = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}.$$

The Cholesky factor is of the form

$$\mathbf{R}_A = \begin{pmatrix} \mathbf{R} & \mathbf{z} \\ \mathbf{0}' & d \end{pmatrix}, \quad (11.15)$$

where $d = \sqrt{\text{RSS}}$ (by Exercises 11a, No. 4). To obtain $(\mathbf{X}'\mathbf{X})^{-1}$, we use

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{R}^{-1}(\mathbf{R}^{-1})',$$

where $(t_{ij}) = \mathbf{R}^{-1}$ is calculated by back-substitution using the formulas

$$\begin{aligned} t_{ii} &= \frac{1}{r_{ii}} \quad (i = 1, \dots, p), \\ t_{ij} &= 0 \quad (i > j), \\ t_{ij} &= -\frac{\sum_{l=i}^{j-1} t_{il} r_{lj}}{r_{jj}} \quad (i < j). \end{aligned}$$

EXERCISES 11a

- Calculate algebraically the solution to the system of linear equations corresponding to the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 + \epsilon & 2 & 2 \\ 1 & 2 & 1 & 3 \end{pmatrix}$$

using Gaussian elimination with and without partial pivoting. Show that if partial pivoting is not used, the solution x_3 is computed as

$$-\frac{2 - 1/\epsilon}{1/\epsilon},$$

and as $1 - 2\epsilon$ if partial pivoting is used. Comment on the accuracy of these two methods if ϵ is very small.

- Show that applying one Gaussian elimination step to the matrix $\mathbf{X}'\mathbf{X}$ yields the matrix

$$\begin{pmatrix} n & n\bar{x} \\ \mathbf{0} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{pmatrix},$$

where $\tilde{\mathbf{X}}$ is the centered version of \mathbf{X} .

3. Verify (11.15) and hence show that $\sqrt{\text{RSS}}$ is given by the $(p+1, p+1)$ th element of \mathbf{R}_A .
4. If $\mathbf{R}'\mathbf{R}$ is the Cholesky decomposition of $\mathbf{X}'\mathbf{X}$, show that

$$\det(\mathbf{X}'\mathbf{X}) = \prod_{i=1}^p r_{ii}^2.$$

11.3 QR DECOMPOSITION

Consider an $n \times p$ matrix \mathbf{A} of rank p with columns $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(p)}$. We can construct an orthonormal basis $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_p$ using the following algorithm.

Algorithm 11.3

Step 1: Set $\mathbf{q}_1 = \mathbf{a}^{(1)} / \|\mathbf{a}^{(1)}\|$ and $j = 2$.

Step 2: Set

$$\begin{aligned}\mathbf{w}_j &= \mathbf{a}^{(j)} - (\mathbf{a}^{(j)'}\mathbf{q}_1)\mathbf{q}_1 - \cdots - (\mathbf{a}^{(j)'}\mathbf{q}_{j-1})\mathbf{q}_{j-1}, \\ \mathbf{q}_j &= \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}.\end{aligned}$$

Step 3: Repeat step 2 for $j = 3, \dots, p$.

Using induction, it is easy to prove that for $j = 2, 3, \dots, p$, \mathbf{q}_j is orthogonal to $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_{j-1}$ and that $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_p$ form an orthonormal basis for $C(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j)})$.

Note that at any stage, \mathbf{w}_j cannot be a zero vector. If it were, then $\mathbf{a}^{(j)}$ could be expressed as a linear combination of $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_{j-1}$ and hence of $\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j-1)}$. This would contradict the assumption that the rank of \mathbf{A} was p .

This method of constructing an orthonormal basis is called the *Gram-Schmidt algorithm*. If we put $\mathbf{R} = (r_{lj})$ where

$$r_{lj} = \begin{cases} \mathbf{a}^{(j)'}\mathbf{q}_l & (l < j), \\ \|\mathbf{w}_j\| & (l = j), \\ 0 & (l > j) \end{cases}$$

and $\mathbf{Q}_p = (\mathbf{q}_1, \dots, \mathbf{q}_p)$, an $n \times p$ matrix, then

$$\mathbf{A} = \mathbf{Q}_p \mathbf{R}. \quad (11.16)$$

Equation (11.16) is called the *QR decomposition* and expresses \mathbf{A} as the product of a matrix \mathbf{Q}_p with orthonormal columns and a $p \times p$ upper triangular matrix \mathbf{R} with positive diagonal elements. Moreover, from (11.16) we get

$$\mathbf{A}'\mathbf{A} = \mathbf{R}'\mathbf{Q}_p'\mathbf{Q}_p\mathbf{R} = \mathbf{R}'\mathbf{R},$$

since $\mathbf{Q}_p'\mathbf{Q}_p = \mathbf{I}_p$. Hence \mathbf{R} is the unique Cholesky factor of $\mathbf{A}'\mathbf{A}$. Since \mathbf{R} is nonsingular, being upper triangular with positive diagonal elements, we have $\mathbf{Q}_p = \mathbf{A}\mathbf{R}^{-1}$, so \mathbf{Q}_p is unique as well.

The QR decomposition as described above involves the calculation of square roots, which are required to convert the orthogonal vectors \mathbf{z}_j into the orthonormal vectors \mathbf{q}_j . To avoid the square roots, we can use the following modified algorithm.

Algorithm 11.4

Step 1: Set $\mathbf{w}_1 = \mathbf{a}^{(1)}$ and $j = 2$.

Step 2: Set

$$\mathbf{w}_j = \mathbf{a}^{(j)} - \frac{\mathbf{a}^{(j)'}\mathbf{w}_1}{\|\mathbf{w}_1\|^2}\mathbf{w}_1 - \cdots - \frac{\mathbf{a}^{(j)'}\mathbf{w}_{j-1}}{\|\mathbf{w}_{j-1}\|^2}\mathbf{w}_{j-1}.$$

Step 3: Repeat step 2 for $j = 3, 4, \dots, p$.

The same argument used in proving the Gram–Schmidt algorithm shows that for $j = 1, \dots, p$, $\mathbf{w}_1, \dots, \mathbf{w}_j$ form an orthogonal basis for $C(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j)})$. Thus if we define an upper triangular matrix $\mathbf{U} = (u_{lj})$ by

$$u_{lj} = \begin{cases} (\mathbf{a}^{(j)'}\mathbf{w}_l)/\|\mathbf{w}_l\|^2 & (l < j), \\ 1 & (l = j), \\ 0 & (l > j), \end{cases}$$

and put $\mathbf{W}_p = (\mathbf{w}_1, \dots, \mathbf{w}_p)$, then

$$\mathbf{A} = \mathbf{W}_p\mathbf{U}.$$

To see the relationship between $\mathbf{W}_p\mathbf{U}$ and $\mathbf{Q}_p\mathbf{R}$, we first note that

$$\mathbf{D} = \mathbf{W}_p'\mathbf{W}_p = \text{diag}(\|\mathbf{w}_1\|^2, \dots, \|\mathbf{w}_p\|^2)$$

and, for $l < j$,

$$u_{lj} = \frac{\mathbf{a}^{(j)'}\mathbf{w}_l}{\|\mathbf{w}_l\|^2} = \frac{\mathbf{a}^{(j)'}\mathbf{q}_j}{\|\mathbf{w}_l\|} = \frac{r_{lj}}{r_{ll}}.$$

Then $\mathbf{U} = \mathbf{D}^{-1/2}\mathbf{R}$ or $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{U}$, and $\mathbf{Q}_p = \mathbf{W}_p\mathbf{D}^{-1/2}$, so that $\mathbf{Q}_p\mathbf{R} = \mathbf{W}_p\mathbf{U}$. Finally, we note that if $\mathbf{A} = \mathbf{X}$, then

$$\mathbf{X}'\mathbf{X} = \mathbf{U}'\mathbf{W}_p'\mathbf{W}_p\mathbf{U} = \mathbf{U}'\mathbf{D}\mathbf{U}. \quad (11.17)$$

11.3.1 Calculation of Regression Quantities

We can obtain the quantities needed for a regression analysis very conveniently from a QR (or WU) decomposition of the augmented matrix (\mathbf{X}, \mathbf{Y}) . In some computer implementations of the QR decomposition, the matrix \mathbf{Q}_p is written over the matrix \mathbf{X} , so it is convenient to express the regression quantities in terms of \mathbf{Q}_p and \mathbf{R} .

Writing the decomposition in partitioned form,

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{Q}_p, \mathbf{q}) \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ \mathbf{0} & d \end{pmatrix}, \quad (11.18)$$

and multiplying the partitioned matrices on the right, we get

$$\mathbf{X} = \mathbf{Q}_p \mathbf{R}, \quad (11.19)$$

$$\mathbf{Y} = \mathbf{Q}_p \mathbf{r} + d \mathbf{q}. \quad (11.20)$$

Premultiplying (11.20) by \mathbf{Q}'_p gives $\mathbf{Q}'_p \mathbf{Y} = \mathbf{Q}'_p \mathbf{Q}_p \mathbf{r} + d \mathbf{Q}'_p \mathbf{q} = \mathbf{r}$, since \mathbf{Q}_p and \mathbf{q} are orthogonal. Similarly, premultiplying (11.20) by \mathbf{q}' gives $d = \mathbf{q}' \mathbf{Y}$.

To compute $\hat{\beta}$, we note that

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y} \\ &= (\mathbf{R}' \mathbf{R})^{-1} \mathbf{R}' \mathbf{Q}' \mathbf{Y} \\ &= \mathbf{R}^{-1} \mathbf{r}. \end{aligned}$$

Thus we get $\hat{\beta}$ by solving the triangular system $\mathbf{R}\mathbf{b} = \mathbf{r}$. The fitted values are

$$\mathbf{X}\hat{\beta} = (\mathbf{Q}_p \mathbf{R})(\mathbf{R}^{-1} \mathbf{r}) = \mathbf{Q}_p \mathbf{r},$$

and from (11.19), the residuals are

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{Q}_p \mathbf{r} = d \mathbf{q}.$$

The residual sum of squares is

$$\text{RSS} = \mathbf{e}' \mathbf{e} = d^2 \mathbf{q}' \mathbf{q} = d^2.$$

Retaining \mathbf{Y} allows the fitted values to be calculated as $\mathbf{Y} - d \mathbf{q}$.

One advantage of algorithms that explicitly form \mathbf{Q}_p is that calculation of the hat matrix diagonals is simple. We have

$$\mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' = \mathbf{Q}_p \mathbf{R}(\mathbf{R}' \mathbf{Q}'_p \mathbf{Q}_p \mathbf{R})^{-1} \mathbf{R}' \mathbf{Q}'_p = \mathbf{Q}_p \mathbf{R} \mathbf{R}^{-1} (\mathbf{R}')^{-1} \mathbf{R}' \mathbf{Q}'_p = \mathbf{Q}_p \mathbf{Q}'_p,$$

so the hat matrix diagonals are just the squared lengths of the rows of \mathbf{Q}_p .

Similar formulas apply if we calculate the square-root-free WU decomposition. If we decompose (\mathbf{X}, \mathbf{Y}) as

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{W}, \mathbf{w}) \begin{pmatrix} \mathbf{U} & \mathbf{u} \\ \mathbf{0} & 1 \end{pmatrix} \quad (11.21)$$

and multiply out the blocks, we get

$$\mathbf{X} = \mathbf{WU} \quad (11.22)$$

and (see Exercises 11b, No. 3)

$$\mathbf{U}\hat{\boldsymbol{\beta}} = \mathbf{u}. \quad (11.23)$$

The fitted values and residuals are

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{w}, \quad (11.24)$$

$$\mathbf{e} = \mathbf{w}, \quad (11.25)$$

so that the residual sum of squares is $\|\mathbf{e}\|^2 = \|\mathbf{w}\|^2$.

11.3.2 Algorithms for the QR and WU Decompositions

We now describe in some detail the three commonly used algorithms for calculation of the QR and WU decompositions.

Modified Gram–Schmidt Algorithm

It is tempting to use (11.16) to calculate the QR decomposition. However, it has been well demonstrated (see, e.g., Björck [1996: p. 63]) that the Gram–Schmidt algorithm is not numerically accurate, and the vectors \mathbf{Q} computed using it are far from orthogonal, due to the accumulation of round-off error. Although the Gram–Schmidt algorithm remains of theoretical importance, and serves to demonstrate the existence of the QR decomposition, it is not a practical recipe for calculation. However, a variation of the method [*the modified Gram–Schmidt algorithm (MGSA)*] has excellent numerical properties. We will describe a square-root-free version of the algorithm that calculates the WU decomposition.

The square-root-free version of the MGSA converts an $n \times p$ matrix \mathbf{A} of rank p into a set of orthogonal vectors in $p - 1$ stages. At the j th stage, the columns $1, \dots, j$ are left alone and the columns $j + 1, \dots, p$ are transformed. The algorithm is as follows.

Algorithm 11.5

Step 1: Set $j = 1$.

Step 2: For $l = j + 1, \dots, p$, replace $\mathbf{a}^{(l)}$ by $\mathbf{a}^{(l)} + v_{jl}\mathbf{a}^{(j)}$, where $v_{jl} = -\mathbf{a}^{(l)'}\mathbf{a}^{(j)}/\|\mathbf{a}^{(j)}\|^2$.

Step 3: Repeat step 2 for $j = 2, \dots, p - 1$.

In this algorithm we begin with $\mathbf{a}^{(1)}$, and then multiples of $\mathbf{a}^{(1)}$ are added to each of $\mathbf{a}^{(2)}, \mathbf{a}^{(3)}, \dots, \mathbf{a}^{(p)}$ so that the new vectors, $\mathbf{a}^{*(2)}, \mathbf{a}^{*(3)}, \dots, \mathbf{a}^{*(p)}$, say, are perpendicular to $\mathbf{a}^{(1)}$. Then multiples of $\mathbf{a}^{*(2)}$ are added to each of $\mathbf{a}^{*(3)}, \dots, \mathbf{a}^{*(p)}$ so that the new vectors $\mathbf{a}^{**3}, \dots, \mathbf{a}^{**p}$ are perpendicular to $\mathbf{a}^{*(2)}$. Since \mathbf{a}^{**3} is a linear combination of $\mathbf{a}^{*(2)}$ and $\mathbf{a}^{*(3)}$, it will also be perpendicular to $\mathbf{a}^{(1)}$. Thus at the end of the j th stage, the first j columns are mutually orthogonal, are a basis for the first j columns of the original matrix \mathbf{A} , and are orthogonal to columns $j+1, \dots, p$ (cf. Exercises 11b, No. 4).

The j th stage is equivalent to postmultiplication by the matrix

$$\mathbf{V}_j = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & v_{j,j+1} & \cdots & v_{jp} \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix}.$$

Thus, we can represent the MGSA as

$$\mathbf{AV}_1 \cdots \mathbf{V}_{p-1} = \mathbf{W},$$

where \mathbf{W} is the result of applying the $p-1$ steps of the algorithm to \mathbf{A} . Now put

$$\mathbf{V} = \mathbf{V}_1 \cdots \mathbf{V}_{p-1},$$

so that $\mathbf{AV} = \mathbf{W}$. By direct multiplication, the matrix \mathbf{V} is

$$\mathbf{V} = \begin{pmatrix} 1 & v_{12} & v_{13} & \cdots & v_{1,p} \\ & 1 & v_{23} & \cdots & v_{2,p} \\ & & 1 & \cdots & v_{3,p} \\ & & & \ddots & v_{p-1,p} \\ & & & & 1 \end{pmatrix}.$$

Since \mathbf{V} is an upper triangular matrix with unit diagonal elements, it is non-singular and $\mathbf{V}^{-1} = \mathbf{U}$ must also be upper triangular with unit diagonal elements (see Exercises 11b, No. 2). Thus $\mathbf{A} = \mathbf{WU}$, which must be the unique \mathbf{WU} decomposition of \mathbf{A} . Note that the MGSA computes \mathbf{W} directly, but \mathbf{U} only indirectly by computing its inverse \mathbf{V} .

Applying the algorithm to the augmented matrix (\mathbf{X}, \mathbf{Y}) gives

$$(\mathbf{X}, \mathbf{Y}) \begin{pmatrix} \mathbf{V} & \mathbf{v} \\ \mathbf{0} & 1 \end{pmatrix} = (\mathbf{W}, \mathbf{w}). \quad (11.26)$$

Comparing (11.26) and (11.21), we see that

$$\begin{pmatrix} \mathbf{V} & \mathbf{v} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{U} & \mathbf{u} \\ \mathbf{0} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix},$$

so that $\mathbf{V} = \mathbf{U}^{-1}$, as expected, and $\mathbf{Vu} + \mathbf{v} = \mathbf{0}$. From (11.23) we get

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{U}^{-1}\mathbf{u} \\ &= \mathbf{Vu} \\ &= -\mathbf{v}.\end{aligned}$$

As before, the residuals \mathbf{e} are just the vector \mathbf{w} .

By augmenting the matrix (\mathbf{X}, \mathbf{Y}) further, we obtain a very compact algorithm. Consider the augmented matrix

$$\left(\begin{array}{cc} \mathbf{X} & \mathbf{Y} \\ \mathbf{I}_p & \mathbf{0} \end{array} \right); \quad (11.27)$$

then

$$\begin{aligned}\left(\begin{array}{cc} \mathbf{X} & \mathbf{Y} \\ \mathbf{I}_p & \mathbf{0} \end{array} \right) \left(\begin{array}{cc} \mathbf{V} & \mathbf{v} \\ \mathbf{0} & 1 \end{array} \right) &= \left(\begin{array}{cc} \mathbf{X}\mathbf{V} & \mathbf{X}\mathbf{v} + \mathbf{Y} \\ \mathbf{V} & \mathbf{v} \end{array} \right) \quad (11.28) \\ &= \left(\begin{array}{cc} \mathbf{X}\mathbf{U}^{-1} & \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ \mathbf{U}^{-1} & \mathbf{v} \end{array} \right) \\ &= \left(\begin{array}{cc} \mathbf{W} & \mathbf{e} \\ \mathbf{U}^{-1} & -\hat{\boldsymbol{\beta}} \end{array} \right).\end{aligned}$$

This yields the regression coefficients and residuals directly. The covariance matrix of $\hat{\boldsymbol{\beta}}$ is computed using the equations $\mathbf{D} = \mathbf{W}'\mathbf{W}$ and $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{U}^{-1}\mathbf{D}^{-1}\mathbf{U}^{-1'}$, which follows from (11.17).

Now consider the p stages of the MGSA that were applied to (\mathbf{X}, \mathbf{Y}) . [There are p stages since (\mathbf{X}, \mathbf{Y}) has $p+1$ columns.] The multiplication in (11.28) is the result of applying the same p stages to the augmented matrix (11.27). Note, however, that the multipliers v_{jl} are those from the MGSA applied to (\mathbf{X}, \mathbf{Y}) , not (11.27). In other words, when calculating the v_{jl} 's in step 2, we use only the columns of (\mathbf{X}, \mathbf{Y}) to compute the inner products.

Approximately $2np^2 + 2p^3$ flops are required to compute the decomposition (11.28), which makes it roughly twice as expensive as the methods based on forming the SSCP matrix. The decomposition (11.26) takes about $2np^2$ flops.

Using Householder Transformations

A Householder transformation is a matrix of the form

$$\mathbf{H} = \mathbf{I} - \frac{1}{\gamma} \mathbf{u} \mathbf{u}', \quad (11.29)$$

where $\gamma = \frac{1}{2} \|\mathbf{u}\|^2$. It follows immediately that \mathbf{H} is symmetric and orthogonal (since $\mathbf{H}^2 = \mathbf{I}$) and represents a reflection [as $\det(\mathbf{H}) = -1$, by A.9.7]. These matrices, introduced by Householder [1958], have a variety of uses in numerical linear algebra and provide an efficient way of computing the QR decomposition. Their key property in this regard is the following: Given two

vectors \mathbf{x} and \mathbf{y} of equal length, the Householder matrix \mathbf{H} corresponding to $\mathbf{u} = \mathbf{x} - \mathbf{y}$ satisfies

$$\mathbf{H}\mathbf{x} = \mathbf{y}. \quad (11.30)$$

This is because when $\|\mathbf{x}\| = \|\mathbf{y}\|$,

$$\begin{aligned} \frac{1}{\gamma} \mathbf{u}' \mathbf{x} &= \frac{2(\mathbf{x} - \mathbf{y})' \mathbf{x}}{\|\mathbf{x} - \mathbf{y}\|^2} \\ &= \frac{2\|\mathbf{x}\|^2 - 2\mathbf{y}' \mathbf{x}}{\|\mathbf{x} - \mathbf{y}\|^2} \\ &= \frac{\|\mathbf{x}\|^2 - 2\mathbf{y}' \mathbf{x} + \|\mathbf{y}\|^2}{\|\mathbf{x} - \mathbf{y}\|^2} \\ &= \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\|\mathbf{x} - \mathbf{y}\|^2} \\ &= 1, \end{aligned}$$

so that

$$\mathbf{H}\mathbf{x} = \mathbf{x} - \frac{1}{\gamma} \mathbf{u}\mathbf{u}' \mathbf{x} = \mathbf{x} - \mathbf{u} = \mathbf{y}. \quad (11.31)$$

In particular, if $\mathbf{y} = (\|\mathbf{x}\|, 0, 0, \dots, 0)'$, then the Householder matrix corresponding to $\mathbf{u} = \mathbf{x} - \mathbf{y}$ satisfies

$$\mathbf{H}\mathbf{x} = \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

so we can choose \mathbf{H} to “zero out” the elements of \mathbf{x} other than the first. Note that when calculating the first element, $x_1 - \|\mathbf{x}\|$, of the Householder vector \mathbf{u} , we have to be careful not to lose significant digits in the subtraction. If $x_1 \leq 0$, there is no problem, and we can compute u_1 as $x_1 - \|\mathbf{x}\|$. However, if x_1 is positive and large compared with the other elements of \mathbf{x} , then computing u_1 as $x_1 - \|\mathbf{x}\|$ may lead to loss of significant figures. In this case we compute u_1 as

$$u_1 = -\frac{x_2^2 + \cdots + x_n^2}{x_1 + \|\mathbf{x}\|}$$

since

$$\begin{aligned} x_1 - \|\mathbf{x}\| &= \frac{x_1^2 - \|\mathbf{x}\|^2}{x_1 + \|\mathbf{x}\|} \\ &= -\frac{x_2^2 + \cdots + x_n^2}{x_1 + \|\mathbf{x}\|}. \end{aligned}$$

The quantity γ is calculated as $\frac{1}{2}(u_1^2 + x_2^2 + \cdots + x_n^2)$.

Multiplying a vector \mathbf{x} by a Householder matrix \mathbf{H} corresponding to a vector \mathbf{u} is easy, since

$$\mathbf{H}\mathbf{x} = (\mathbf{I}_n - \gamma^{-1}\mathbf{u}\mathbf{u}')\mathbf{x} = \mathbf{x} - k\mathbf{u},$$

where $k = (\mathbf{u}'\mathbf{x})/\gamma$.

Now consider an $n \times p$ matrix \mathbf{A} of rank p with columns $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(p)}$. If we choose \mathbf{H}_1 to zero out all but the first element of $\mathbf{a}^{(1)}$, premultiplying \mathbf{A} by \mathbf{H}_1 gives

$$\mathbf{H}_1\mathbf{A} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ 0 & & & \\ \vdots & & \mathbf{A}_1 & \\ 0 & & & \end{pmatrix},$$

say, where $r_{11} = \|\mathbf{a}^{(1)}\|$. Here, $\|\mathbf{a}^{(1)}\| \neq 0$, for otherwise \mathbf{A} would have a zero column and hence not be of rank p . Next consider a matrix of the form

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{K}_2 & \\ 0 & & & \end{pmatrix},$$

where \mathbf{K}_2 is an $(n-1) \times (n-1)$ Householder matrix chosen to zero out all but the first element of the first column of \mathbf{A}_1 . We get

$$\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1p} \\ 0 & r_{22} & r_{23} & \cdots & r_{2p} \\ 0 & 0 & & & \\ \vdots & \vdots & & \mathbf{A}_2 & \\ 0 & 0 & & & \end{pmatrix}.$$

Once again $r_{22} \neq 0$. If this is not the case, then the second column of $\mathbf{H}_2\mathbf{H}_1\mathbf{A}$ would be linearly dependent on the first. This would contradict the fact that $\text{rank}(\mathbf{H}_2\mathbf{H}_1\mathbf{A}) = \text{rank}(\mathbf{A}) = p$ (by A.2.2).

Continuing on in this way, we get

$$\mathbf{H}_p\mathbf{H}_{p-1} \cdots \mathbf{H}_1\mathbf{A} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix},$$

where \mathbf{R} is $p \times p$ upper triangular with positive diagonal elements. Setting $\mathbf{Q} = (\mathbf{H}_p\mathbf{H}_{p-1} \cdots \mathbf{H}_1)' = \mathbf{H}_1\mathbf{H}_2 \cdots \mathbf{H}_p$ yields the QR decomposition

$$\mathbf{A} = \mathbf{Q} \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}, \quad (11.32)$$

since \mathbf{Q} is orthogonal, being a product of orthogonal matrices. We see that in contrast to the QR decomposition calculated by the MGSA, the matrix \mathbf{Q} is $n \times n$. Writing

$$\mathbf{Q} = (\mathbf{Q}_p, \mathbf{Q}_{n-p}),$$

where \mathbf{Q}_{n-p} is $n \times (n - p)$, gives

$$\mathbf{A} = (\mathbf{Q}_p, \mathbf{Q}_{n-p}) \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}_p \mathbf{R}, \quad (11.33)$$

which is the decomposition produced by the MGSA. The version (11.33) is sometimes called the *thin QR decomposition* (see, e.g., Golub and Van Loan [1996: p. 230]). In the *fat* version (11.32), we cannot say that the last $n - p$ columns \mathbf{Q}_{n-p} of \mathbf{Q} are unique.

In contrast to the MGSA, when using Householder transformations the matrix \mathbf{Q} (or \mathbf{W}) is not formed explicitly. Rather, if we need to multiply a vector \mathbf{a} by \mathbf{Q} , we use

$$\mathbf{Q}\mathbf{a} = \mathbf{H}_1 \mathbf{H}_2 \cdots \mathbf{H}_p \mathbf{a}. \quad (11.34)$$

When using Householder transformations, the p Householder vectors must be stored. This is usually done by overwriting \mathbf{A} with \mathbf{R} and storing the Householder vectors (minus their first elements, which will not fit) in the locations in

$$\begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \quad (11.35)$$

that are zero. The initial elements need to be stored separately.

The Householder algorithm requires around $2np^2 - \frac{2}{3}p^3$ flops, a bit less than the MGSA. A proof may be found in Trefethen and Bau [1997: p. 74].

The regression quantities are computed differently when using Householder transformations, reflecting the fact that it is \mathbf{R} that is stored rather than \mathbf{Q}_p as in the MGSA.

Algorithm 11.6

Step 1: Set $j = 1$ and put $\mathbf{A} = (\mathbf{X}, \mathbf{Y})$.

Step 2: (Calculate the Householder vector.) Put $\mathbf{u} = (0, 0, \dots, 0, a_{jj}, a_{j+1,j}, \dots, a_{n,j})'$, and recalculate u_j as $u_j - \|\mathbf{u}\|$ if $u_j < 0$ and as $-(u_j^2 + \dots + u_n^2)/(u_j + \|\mathbf{u}\|)$ if $u_j > 0$. Calculate γ as $\frac{1}{2}\|\mathbf{u}\|^2$ using the updated value of u_j .

Step 3: (Update \mathbf{A} .) Multiply columns $j, \dots, p+1$ of \mathbf{A} by the Householder matrix corresponding to the vector \mathbf{u} computed in step 2. That is, for $l = j, \dots, p+1$, replace the l th column $\mathbf{a}^{(l)}$ of \mathbf{A} by $\mathbf{a}^{(l)} - k\mathbf{u}$, where $k = (\mathbf{a}^{(l)'}\mathbf{u})/\gamma$.

Step 4: Repeat steps 2 and 3 for $j = 2, \dots, p+1$.

This yields the matrix

$$\begin{pmatrix} \mathbf{R} & \mathbf{r}_1 \\ \mathbf{0} & \mathbf{r}_2 \end{pmatrix}$$

in the decomposition

$$(\mathbf{X}, \mathbf{Y}) = (\mathbf{Q}_p, \mathbf{Q}_{n-p}) \begin{pmatrix} \mathbf{R} & \mathbf{r}_1 \\ \mathbf{0} & \mathbf{r}_2 \end{pmatrix}, \quad (11.36)$$

where \mathbf{r}_1 and \mathbf{r}_2 have p and $n-p$ elements, respectively, and $\mathbf{r}_2 = (d, 0, \dots, 0)'$. The algorithm also calculates the Householder vectors needed to compute (11.34). Multiplying out and equating blocks gives

$$\begin{aligned} \mathbf{X} &= \mathbf{Q}_p \mathbf{R}, \\ \mathbf{Y} &= \mathbf{Q}_p \mathbf{r}_1 + \mathbf{Q}_{n-p} \mathbf{r}_2. \end{aligned}$$

To calculate $\hat{\beta}$, we note that

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y} \\ &= (\mathbf{R}' \mathbf{Q}'_p \mathbf{Q}_p \mathbf{R})^{-1} \mathbf{R}' \mathbf{Q}'_p (\mathbf{Q}_p \mathbf{r}_1 + \mathbf{Q}_{n-p} \mathbf{r}_2) \\ &= (\mathbf{R}' \mathbf{R})^{-1} \mathbf{R}' \mathbf{r}_1 \quad [\text{since } \mathbf{Q}'_p \mathbf{Q}_{n-p} = 0] \\ &= \mathbf{R}^{-1} \mathbf{r}_1, \end{aligned} \quad (11.37)$$

so that $\hat{\beta}$ is the solution of $\mathbf{R}\hat{\beta} = \mathbf{r}_1$, which can be solved by back-substitution.

The residuals are

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\beta} \\ &= \mathbf{Q}_p \mathbf{r}_1 + \mathbf{Q}_{n-p} \mathbf{r}_2 - \mathbf{Q}_p \mathbf{R} \mathbf{R}^{-1} \mathbf{r}_1 \\ &= \mathbf{Q}_{n-p} \mathbf{r}_2 \\ &= (\mathbf{Q}_p, \mathbf{Q}_{n-p}) \begin{pmatrix} \mathbf{0} \\ \mathbf{r}_2 \end{pmatrix}. \end{aligned} \quad (11.38)$$

Thus the residuals are computed by premultiplying $(\mathbf{0}', \mathbf{r}'_2)'$ by \mathbf{Q} using (11.34). If \mathbf{Y} is retained, the fitted values are best calculated by $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{e}$. Otherwise, they can be calculated (cf. Exercises 11b, No. 6) by multiplying $(\mathbf{r}'_1, \mathbf{0}')'$ by \mathbf{Q} . The residual sum of squares is given by

$$\begin{aligned} \text{RSS} &= \|\mathbf{e}\|^2 \\ &= (\mathbf{0}', \mathbf{r}'_2) \mathbf{Q}' \mathbf{Q} (\mathbf{0}', \mathbf{r}'_2)' \\ &= (\mathbf{0}', \mathbf{r}'_2) (\mathbf{0}', \mathbf{r}'_2)' \\ &= d^2. \end{aligned} \quad (11.39)$$

Using Givens Transformations

A Givens transformation is a matrix of the form

$$\mathbf{G}_{ih} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \quad \begin{array}{l} \text{row } h \\ \text{row } i, \end{array} \quad (11.40)$$

where

$$c^2 + s^2 = 1.$$

If θ satisfies $\cos \theta = c$ and $\sin \theta = s$, then multiplication of a vector \mathbf{x} by \mathbf{G}_{ih} rotates \mathbf{x} clockwise through an angle θ in the h, i coordinate plane. We note that \mathbf{G}_{ih} is an orthogonal matrix and \mathbf{G}'_{ih} is a rotation in the opposite direction. For this reason, Givens transformations are sometimes called *planar rotations*.

Transforming a vector \mathbf{x} by \mathbf{G}_{ih} changes only the h th and i th elements of \mathbf{x} . If $\mathbf{v} = \mathbf{G}_{ih}\mathbf{x}$, these are given by

$$v_h = cx_h + sx_i, \quad (11.41)$$

$$v_i = -sx_h + cx_i. \quad (11.42)$$

Clearly, we can choose c and s to make v_i zero. This happens if we choose

$$c = \frac{x_h}{\sqrt{x_h^2 + x_i^2}}, \quad s = \frac{x_i}{\sqrt{x_h^2 + x_i^2}}.$$

In this case we have

$$\frac{s}{c} = \frac{x_i}{x_h}. \quad (11.43)$$

If $x_h = 0$ and $x_i > 0$, $\theta = \frac{1}{2}\pi$ and \mathbf{G}_{ih} simply interchanges these two elements. The sign of the new h th element is also changed if $x_i < 0$. In summary, if s and c are defined as described above, the net effect of multiplying by \mathbf{G}_{ih} is to reduce the i th element of \mathbf{x} to zero.

Now let \mathbf{A} be $n \times p$ of rank p , and consider multiplying \mathbf{A} by a sequence of Givens transformations $\mathbf{G}_{21}, \mathbf{G}_{31}, \dots, \mathbf{G}_{n,1}$, choosing c and s each time to zero out the second, third, ..., n th elements of the first column of \mathbf{A} . Multiplying further by the sequence $\mathbf{G}_{32}, \mathbf{G}_{42}, \dots, \mathbf{G}_{n,2}$ will zero out the third, fourth, ..., n th elements of the second column, leaving the first column unchanged. Continuing in this way, we eventually reduce \mathbf{A} to the form

$$\left(\begin{array}{c} \mathbf{R} \\ \mathbf{0} \end{array} \right).$$

An alternative (see Gentleman [1973]) is to zero out the 2,1 element, then the 3,1 and 3,2 elements and so on, thus operating on \mathbf{A} row by row rather than column by column.

Each Givens transformation is represented by the corresponding values of c and s . Using a clever device due to Stewart [1976] (see also Björck [1996: p. 55]), these can be coded as a single number. These numbers can be stored by overwriting \mathbf{A} by \mathbf{R} and storing the coded numbers in the locations of (11.35) occupied by zeros.

We see that Givens rotations provide an alternative to using Householder reflections. The relative merits of the two approaches are discussed below and also in Section 11.8.

Fast Rotators

If we use the Givens method to reduce an $n \times p$ matrix to upper triangular form, the formulas (11.41) and (11.42) must be evaluated approximately $\frac{1}{2}np^2 - \frac{1}{6}p^3$ times. Since four multiplications and two additions are required each time the formulas are evaluated, this is a total of about $3np^2 - p^3$ flops, so that the Givens method as described above requires about 50% more operations than the MGSA and Householder algorithms. However, we can use a modified Givens transformation, called a *fast rotator*, which requires only two multiplications and two additions per evaluation. A fast rotator is a matrix having one of the forms

$$\left(\begin{array}{cccccc} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & \alpha & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\eta & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{array} \right) \quad \begin{matrix} \text{row } h \\ \text{row } i \end{matrix} \quad (11.44)$$

or

$$\left(\begin{array}{cccccc} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \alpha & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -1 & \cdots & \eta & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{array} \right) \quad \begin{matrix} \text{row } h \\ \text{row } i. \end{matrix} \quad (11.45)$$

Such transformations change only two entries when multiplying a vector, as do Givens rotations, but require only two multiplications instead of four. They are related to Givens rotations in the following way. Let $\tilde{\mathbf{D}} = \text{diag}(\tilde{d}_1, \dots, \tilde{d}_n)$ be a given diagonal matrix, and let \mathbf{G}_{ih} be the Givens transformation (11.40).

Then there is a fast rotator \mathbf{M} of type (11.44) and a diagonal matrix $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ such that

$$\mathbf{G}_{ih} = \mathbf{D}\mathbf{M}\tilde{\mathbf{D}}^{-1}. \quad (11.46)$$

Direct calculation shows that

$$d_l = \tilde{d}_l \quad (l \neq i, l \neq h), \quad (11.47)$$

$$d_i = c\tilde{d}_i, \quad (11.48)$$

$$d_h = c\tilde{d}_h, \quad (11.49)$$

$$\alpha = t/r, \quad (11.50)$$

$$\eta = rt, \quad (11.51)$$

where $t = s/c$ and $r = \tilde{d}_h/\tilde{d}_i$. A similar result holds for fast rotators (11.45); for the corresponding equations, see Exercises 11b, No. 7.

Next, we show how the regression quantities can be calculated using a sequence of fast rotators, which use half the number of multiplications and no square roots. Let $\mathbf{G}_1, \dots, \mathbf{G}_m$ be the sequence of Givens rotations used to reduce (\mathbf{X}, \mathbf{Y}) to the form

$$\begin{pmatrix} \mathbf{R} & \mathbf{r}_1 \\ \mathbf{0} & \mathbf{r}_2 \end{pmatrix}.$$

We are using a single index for the Givens transformations for notational simplicity.

Now define a sequence of fast rotators as follows: Let $\mathbf{D}_0 = \mathbf{I}_n$ and $\mathbf{G}_l = \mathbf{D}_l \mathbf{M}_l \mathbf{D}_{l-1}^{-1}$, $l = 1, 2, \dots, m$, where \mathbf{D}_l and \mathbf{M}_l are derived from \mathbf{G}_l and \mathbf{D}_{l-1} as described above. Then for $l = 1, 2, \dots, m$,

$$\begin{aligned} \mathbf{G}_l \mathbf{G}_{l-1} \cdots \mathbf{G}_1 (\mathbf{X}, \mathbf{Y}) &= (\mathbf{D}_l \mathbf{M}_l \mathbf{D}_{l-1}^{-1})(\mathbf{D}_{l-1} \mathbf{M}_{l-1} \mathbf{D}_{l-2}^{-1}) \times \cdots \\ &\quad \times (\mathbf{D}_1 \mathbf{M}_1 \mathbf{D}_0^{-1})(\mathbf{X}, \mathbf{Y}) \\ &= \mathbf{D}_l \mathbf{M}_l \mathbf{M}_{l-1} \cdots \mathbf{M}_1 (\mathbf{X}, \mathbf{Y}) \\ &= \mathbf{D}_l \mathbf{B}_l, \end{aligned}$$

say, where $\mathbf{B}_l = \mathbf{M}_l \mathbf{M}_{l-1} \cdots \mathbf{M}_1 (\mathbf{X}, \mathbf{Y})$. We note that

$$\mathbf{B}_l = \mathbf{M}_l \mathbf{B}_{l-1} \quad (11.52)$$

and

$$\mathbf{D}_m \mathbf{B}_m = \begin{pmatrix} \mathbf{R} & \mathbf{r}_1 \\ \mathbf{0} & \mathbf{r}_2 \end{pmatrix}. \quad (11.53)$$

Now partition \mathbf{B}_m and \mathbf{D}_m conformably as

$$\mathbf{B}_m = \begin{pmatrix} \mathbf{B} & \mathbf{b}_1 \\ \mathbf{0} & \mathbf{b}_2 \end{pmatrix} \quad (11.54)$$

and

$$\mathbf{D}_m = \begin{pmatrix} \mathbf{D}_{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{(2)} \end{pmatrix}. \quad (11.55)$$

Equating blocks in (11.53), we see that $\mathbf{D}_{(1)}\mathbf{B} = \mathbf{R}$, $\mathbf{D}_{(1)}\mathbf{b}_1 = \mathbf{r}_1$, and $\mathbf{D}_{(2)}\mathbf{b}_2 = \mathbf{r}_2$. From these relations and (11.37), it follows that $\hat{\beta}$ satisfies the equation $\mathbf{D}_{(1)}\mathbf{B}\hat{\beta} = \mathbf{D}_{(1)}\mathbf{b}_1$, which is equivalent to

$$\mathbf{B}\hat{\beta} = \mathbf{b}_1. \quad (11.56)$$

By (11.39), the residual sum of squares is

$$\begin{aligned} \text{RSS} &= \|\mathbf{r}_2\|^2 \\ &= \mathbf{b}'_2 \mathbf{D}_{(2)}^2 \mathbf{b}_2. \end{aligned}$$

The matrix \mathbf{Q} of the QR decomposition is $\mathbf{Q} = \mathbf{G}'_1 \cdots \mathbf{G}'_m = \mathbf{M}'_1 \cdots \mathbf{M}'_m \mathbf{D}_m$, so that the vector of residuals is

$$\begin{aligned} \mathbf{Q} \begin{pmatrix} \mathbf{0} \\ \mathbf{r}_2 \end{pmatrix} &= \mathbf{M}'_1 \cdots \mathbf{M}'_m \mathbf{D}_m \begin{pmatrix} \mathbf{0} \\ \mathbf{D}_{(2)}\mathbf{b}_2 \end{pmatrix} \\ &= \mathbf{M}'_1 \cdots \mathbf{M}'_m \begin{pmatrix} \mathbf{0} \\ \mathbf{D}_{(2)}^2 \mathbf{b}_2 \end{pmatrix}. \end{aligned}$$

These results show that to compute the regression quantities, we need only know the matrices \mathbf{B}_l , \mathbf{M}_l , and \mathbf{D}_l^2 . We now describe an algorithm for computing these.

To compute \mathbf{M}_l using (11.47)–(11.51), we require \mathbf{G}_l and \mathbf{D}_{l-1} . The matrix \mathbf{G}_l is determined by the elements of the matrix $\mathbf{G}_{l-1} \cdots \mathbf{G}_1(\mathbf{X}, \mathbf{Y}) = \mathbf{D}_{l-1}\mathbf{B}_{l-1}$. Put $\mathbf{D}_{l-1} = \text{diag}(\tilde{d}_1, \dots, \tilde{d}_n)$, and $\mathbf{B}_{l-1} = (\tilde{b}_{kj})$, so that the k, j element of $\mathbf{D}_{l-1}\mathbf{B}_{l-1}$ is $\tilde{d}_k \tilde{b}_{kj}$. Suppose that \mathbf{G}_l reduces the i, h element of $\mathbf{D}_{l-1}\mathbf{B}_{l-1}$ to zero. Then, from (11.43), we must have

$$t = \frac{s}{c} = \frac{\tilde{d}_i \tilde{b}_{ih}}{\tilde{d}_h \tilde{b}_{hh}}.$$

The elements of \mathbf{M}_l are calculated using (11.50) and (11.51), and $\mathbf{D}_l^2 = \text{diag}(d_1^2, \dots, d_n^2)$ is calculated from (11.47)–(11.49) using the formulas

$$\begin{aligned} d_l^2 &= \tilde{d}_l^2 \quad (l \neq i, l \neq j), \\ d_i^2 &= \tilde{d}_i^2 / (1 + t^2), \\ d_h^2 &= \tilde{d}_h^2 / (1 + t^2). \end{aligned}$$

Finally, since $\mathbf{B}_l = \mathbf{M}_l \mathbf{B}_{l-1}$, $\mathbf{B}_l = (b_{ij})$ is calculated by

$$\begin{aligned} b_{hj} &= \tilde{b}_{hj} + \alpha \tilde{b}_{ij} \quad (j = h, h+1, \dots, p), \\ b_{ij} &= \tilde{b}_{ij} - \eta \tilde{b}_{hj} \quad (j = h+1, \dots, p), \\ b_{ih} &= 0. \end{aligned}$$

The other elements of \mathbf{B}_l are the same as the corresponding elements of \mathbf{B}_{l-1} . These calculations have exact parallels using fast rotators of the form (11.45). At each update we have the option of using either form of fast rotator. As we saw above, the update using the rotator (11.44) involves updating the diagonal matrix by multiplying by c , which could be small. A sequence of updates using a small factor may eventually result in underflow in the calculation.

If the fast rotator (11.45) is used, the corresponding factor is s , so a good method for avoiding underflow is to use (11.44) if $c^2 > s^2$ and (11.45) otherwise. There may still be a problem with underflow if too many updates are done. Tests should be done to detect this, and rescaling performed if necessary; this simply involves multiplying \mathbf{D}_l by a suitable scale factor to increase its size, and dividing \mathbf{B}_l by the same factor. The need to check for underflow diminishes to some extent the gains made in using fast rotators. Björck [1996: p. 57] suggests that the overall gain in multiplications may be a factor of 1.4 to 1.6 rather than the theoretical 2. A variation on fast rotators which removes the need to rescale is described in Anda and Park [1994].

EXERCISES 11b

1. Let $\mathbf{q}_1, \dots, \mathbf{q}_p$ be the vectors constructed in Algorithm 11.3. Show that $\mathbf{q}_1, \dots, \mathbf{q}_j$ is an orthonormal basis for $\mathcal{C}(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j)})$ for $j = 1, 2, \dots, p$.
2. Show that an upper triangular matrix with positive diagonal elements is nonsingular. Also, prove that the inverse of an upper triangular matrix with unit diagonal elements is also upper triangular with unit diagonal elements.
3. Derive formulas (11.23), (11.24), and (11.25).
4. Prove that at the beginning of stage j of the MGSA, the columns $\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j)}$ are orthogonal, are orthogonal to $\mathbf{a}^{(j+1)}, \dots, \mathbf{a}^{(p)}$, and span the first j columns of the original matrix.
5. Show that a product of orthogonal matrices is orthogonal.
6. Show that if \mathbf{Q}_p , \mathbf{Q}_{n-p} and \mathbf{r}_1 are as defined in (11.36), the fitted values $\hat{\mathbf{Y}}$ are given by

$$\hat{\mathbf{Y}} = (\mathbf{Q}_p, \mathbf{Q}_{n-p}) \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{0} \end{pmatrix}.$$

7. Let \mathbf{G}_{ih} be given by (11.40) and $\tilde{\mathbf{D}} = \text{diag}(\tilde{d}_1, \dots, \tilde{d}_n)$. Show that there is a diagonal matrix $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ and a fast rotator of the form

(11.45) such that $\mathbf{G}_{ih} = \mathbf{D}\tilde{\mathbf{M}}\tilde{\mathbf{D}}^{-1}$. Show that

$$\begin{aligned} d_l &= \tilde{d}_l \quad (l \neq i, l \neq h), \\ d_i &= s\tilde{d}_h, \\ d_h &= s\tilde{d}_i, \\ \alpha &= r/t, \\ \eta &= 1/rt, \end{aligned}$$

where $t = s/c$ and $r = \tilde{d}_h/\tilde{d}_i$.

11.4 SINGULAR VALUE DECOMPOSITION

Suppose that \mathbf{X} is $n \times p$ of rank r . Then (see A.12) there is an $n \times n$ orthogonal matrix \mathbf{U} and a $p \times p$ orthogonal matrix \mathbf{V} such that

$$\mathbf{U}'\mathbf{X}\mathbf{V} = \begin{pmatrix} \Sigma \\ \mathbf{0} \end{pmatrix},$$

where

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \\ 0 & & & \sigma_p \end{pmatrix}.$$

The quantities σ_j are the *singular values* of \mathbf{X} . They satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$$

and are unique. The equation

$$\mathbf{X} = \mathbf{U} \begin{pmatrix} \Sigma \\ \mathbf{0} \end{pmatrix} \mathbf{V}' \tag{11.57}$$

is called the *singular value decomposition* (SVD). It is sometimes written in the thin form

$$\mathbf{X} = \mathbf{U}_p \Sigma \mathbf{V}',$$

where \mathbf{U}_p is the first p columns of \mathbf{U} .

11.4.1 Regression Calculations Using the SVD

Assume that $\text{rank}(\mathbf{X}) = p$, so that the diagonal elements of Σ are all positive. (The case where \mathbf{X} is possibly rank deficient is discussed in Section 11.9.2.)

The regression quantities can be calculated in terms of the thin SVD as follows. Substituting $\mathbf{X} = \mathbf{U}_p \Sigma \mathbf{V}'$ into the normal equations and using $\mathbf{U}'_p \mathbf{U}_p = \mathbf{I}_p$, we get

$$\mathbf{V} \Sigma^2 \mathbf{V}' \hat{\boldsymbol{\beta}} = \mathbf{V} \Sigma \mathbf{U}'_1 \mathbf{Y},$$

which reduces to

$$\hat{\boldsymbol{\beta}} = \mathbf{V} \Sigma^{-1} \mathbf{U}'_p \mathbf{Y},$$

since \mathbf{V} and Σ are nonsingular. To calculate the RSS, we partition $\mathbf{U} = (\mathbf{U}_p, \mathbf{U}_{n-p})$ and use the identity $\mathbf{U}\mathbf{U}' = \mathbf{I}_n$. Then $\mathbf{I}_n - \mathbf{U}_p \mathbf{U}'_p = \mathbf{U}_{n-p} \mathbf{U}'_{n-p}$, and direct substitution shows that the projection onto $\mathcal{C}(\mathbf{X})$ is

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{U}_p \mathbf{U}'_p.$$

Finally,

$$\begin{aligned} \text{RSS} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{U}_p \mathbf{U}'_p)\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{U}_{n-p} \mathbf{U}'_{n-p})\mathbf{Y} \\ &= \|\mathbf{U}'_{n-p} \mathbf{Y}\|^2. \end{aligned}$$

11.4.2 Computing the SVD

Computation of the SVD consists of two phases. First reduce \mathbf{X} to the form

$$\left(\begin{array}{c} \mathbf{B} \\ \mathbf{0} \end{array} \right),$$

where \mathbf{B} is a *bidiagonal matrix* of the form

$$\mathbf{B} = \left(\begin{array}{cccccc} * & * & 0 & 0 & \cdots & 0 \\ 0 & * & * & 0 & \cdots & 0 \\ 0 & 0 & * & * & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & & & * & * & \\ 0 & & & 0 & * & \end{array} \right)$$

by a series of Householder transformations. Then use the fact that if $\sigma_1, \dots, \sigma_p$ are the singular values of \mathbf{B} , the eigenvalues of the symmetric matrices $\mathbf{B}'\mathbf{B}$ and $\mathbf{B}\mathbf{B}'$ are just $\sigma_1^2, \dots, \sigma_p^2$, and the eigenvectors are the columns of the orthogonal matrices in the SVD.

There are many excellent algorithms for finding the eigenvalues and eigenvectors of symmetric tridiagonal matrices such as $\mathbf{B}'\mathbf{B}$ and $\mathbf{B}\mathbf{B}'$. Discussion of these is beyond our scope but excellent accounts can be found in Björck [1996], Golub and Van Loan [1996], Parlett [1980], and Watkins [1991]. These algorithms are adapted to the SVD case to avoid explicitly forming $\mathbf{B}'\mathbf{B}$ and $\mathbf{B}\mathbf{B}'$.

The reduction of \mathbf{X} to the required form proceeds as follows. First, premultiply \mathbf{X} by a Householder transformation chosen to zero out the first row of \mathbf{X} , except the first element. Then postmultiply the result by another Householder transformation, which leaves the first element of row 1 unchanged, and zeros out the first row except for the first two elements.

Next, premultiply by a further Householder transformation to zero out the second column below the diagonal, then postmultiply to zero out the second row, starting at the third element. This leaves the first row unchanged. Proceeding in this way, after $2p - 2$ transformations \mathbf{X} is reduced to the desired form, and we obtain

$$\mathbf{X} = \mathbf{H}_1 \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} \mathbf{H}_2,$$

where \mathbf{H}_1 and \mathbf{H}_2 are products of Householder transformations. From the eigenvalue analysis, we obtain the SVD $\mathbf{B} = \tilde{\mathbf{U}}'\Sigma\tilde{\mathbf{V}}$. Combining these gives the thin SVD of \mathbf{X} .

If n is much bigger than p , it is more efficient to carry out a QR decomposition of \mathbf{X} and then calculate the SVD of the resulting \mathbf{R} according to the recipe above. This produces either the fat or thin SVD according as we start with the fat or thin QR.

11.5 WEIGHTED LEAST SQUARES

Consider calculating the estimate $\hat{\beta}_w$ that minimizes the weighted least squares criterion

$$(\mathbf{Y} - \mathbf{X}\mathbf{b})' \mathbf{W} (\mathbf{Y} - \mathbf{X}\mathbf{b}), \quad (11.58)$$

where \mathbf{W} is a diagonal matrix whose diagonal elements are positive. By the results of Section 3.10, the weighted least squares estimate is the solution of the equations

$$\mathbf{X}' \mathbf{W} \mathbf{X} \mathbf{b} = \mathbf{X}' \mathbf{W} \mathbf{Y}. \quad (11.59)$$

The simplest method of solving (11.59) is to introduce the matrix $\mathbf{X}_w = \mathbf{W}^{1/2} \mathbf{X}$ and the corresponding vector $\mathbf{Y}_w = \mathbf{W}^{1/2} \mathbf{Y}$. Then (11.59) reduces to

$$\mathbf{X}_w' \mathbf{X}_w \mathbf{b} = \mathbf{X}_w' \mathbf{Y}_w, \quad (11.60)$$

which can be solved by the methods described in previous sections of this chapter.

This simple method works well, provided that the diagonal elements of \mathbf{W} do not vary too much in magnitude. If they do, the standard methods of solution may not be accurate, as the matrix \mathbf{X}_w can be ill-conditioned. (The condition of a matrix is discussed in Sections 9.7.4 and 11.8.3.) We note that the condition number of \mathbf{X}_w (see Exercises 11c, No. 1, at the end of Section 11.8.5) satisfies

$$\kappa(\mathbf{X}_w) \leq \kappa(\mathbf{W}) \kappa(\mathbf{X}), \quad (11.61)$$

so that if the elements of \mathbf{W} differ markedly in magnitude, the condition number of \mathbf{X}_w may be very much bigger than that of \mathbf{X} . The standard methods of calculating the ordinary least squares estimate can be modified to give better accuracy in this case. For more details, see Björck [1996: p. 165] and the references cited there.

11.6 ADDING AND DELETING CASES AND VARIABLES

Often, we need to modify a fitted regression by adding or deleting cases or variables. We may want to delete cases as part of the diagnostic techniques discussed in Chapter 10, for example, if some cases are identified as outliers. Alternatively, we may want to add cases (or blocks of cases) if new data become available, or if data are arriving sequentially. Most times, we will just refit the model after adjusting the input data. However, in some cases we may wish to modify an existing fit. We discuss methods for doing this below.

In the context of model selection, we often want to fit a large number of different models, and this needs to be done efficiently if the number of models is large. Efficient ways of adding and deleting variables by modifying an existing fit are also discussed below. We begin by looking at theoretical updating formulas, based on the partitioned inverse formula and the Sherman–Morrison formula. We then discuss updating methods based on these formulas. These methods may not be accurate if the problems are ill-conditioned, so we close the section with a description of how the more accurate methods based on the QR decomposition can be used to modify regressions.

11.6.1 Updating Formulas

Adding and Deleting Cases

Suppose that we add a new row to the data matrix (\mathbf{X}, \mathbf{Y}) to obtain a new data matrix

$$\begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{x}' & y \end{pmatrix}.$$

The new SSCP matrix is $(\mathbf{X}'\mathbf{X} + \mathbf{xx}')$, a rank 1 update of the original. From A.9.4, its inverse is given by

$$(\mathbf{X}'\mathbf{X} + \mathbf{xx}')^{-1} = (\mathbf{X}'\mathbf{X})^{-1} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{xx}'(\mathbf{X}'\mathbf{X})^{-1}}{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}.$$

Using this result we find that the new vector of estimated regression coefficients is (cf. Theorem 10.1 in Section 10.2)

$$\begin{aligned} \hat{\beta}_{\text{NEW}} &= (\mathbf{X}'\mathbf{X} + \mathbf{xx}')^{-1}(\mathbf{X}'\mathbf{Y} + y\mathbf{x}) \\ &= \hat{\beta}_{\text{OLD}} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}(y - \mathbf{x}'\hat{\beta}_{\text{OLD}})}{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}. \end{aligned}$$

This suggests the following algorithm. First solve the equation

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{x}$$

for \mathbf{b} , and then calculate

$$\hat{\beta}_{\text{NEW}} = \hat{\beta}_{\text{OLD}} + \frac{\mathbf{b}(y - \mathbf{x}'\hat{\beta}_{\text{OLD}})}{1 + \mathbf{x}'\mathbf{b}}.$$

Deleting a case was discussed in Section 10.6.3. From (10.49) with the obvious sign changes, we get

$$\hat{\beta}_{\text{NEW}} = \hat{\beta}_{\text{OLD}} - \frac{\mathbf{b}(y - \mathbf{x}'\hat{\beta}_{\text{OLD}})}{1 - \mathbf{x}'\mathbf{b}},$$

where \mathbf{X} is now the original data matrix from which a row \mathbf{x} is removed. We note that $\mathbf{x}'\mathbf{b} = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$ is the hat matrix diagonal corresponding to the row \mathbf{x}' .

Adding and Deleting Variables

The partitioned inverse formula A.9.1 provides a neat way of describing the effect of adding variables to a regression. Suppose that \mathbf{X} is $n \times p$ of rank p and we want to add a new variable \mathbf{z} .

As in Section 3.7, let $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{R} = \mathbf{I}_n - \mathbf{P}$, $\mathbf{l} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$, and $\mathbf{m} = (\mathbf{z}'\mathbf{R}\mathbf{z})^{-1}$. Then, by Theorem 3.6, the new augmented vector of least squares estimates is of the form

$$\hat{\beta}_A = \begin{pmatrix} \hat{\beta} - \mathbf{l}\mathbf{z}'\mathbf{R}\mathbf{Y}\mathbf{m} \\ \mathbf{z}'\mathbf{R}\mathbf{Y}\mathbf{m} \end{pmatrix}, \quad (11.62)$$

and the projection \mathbf{P}_A onto $C(\mathbf{X}, \mathbf{z})$ is

$$\mathbf{P}_A = \mathbf{P} + \mathbf{R}\mathbf{z}\mathbf{z}'\mathbf{R}\mathbf{m}. \quad (11.63)$$

The ease with which variables can be added suggests that if a regression model has more than one variable, then the variables should be brought in one at a time, using (11.62). This is, in fact, exactly the basis of the sweep operation introduced in Section 11.2.2 and discussed further in Section 11.6.2. Using Theorem 3.6, these results can be easily generalized to the case where several columns are added at once.

11.6.2 Connection with the Sweep Operator

The formulas in the preceding section are the basis of the sweep algorithm introduced in Section 11.2.2. We now verify the claim made there: namely, that sweeping the augmented matrix

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}$$

successively on columns $1, 2, \dots, p$ yields the matrix

$$\begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} & \hat{\beta} \\ -\hat{\beta} & \text{RSS} \end{pmatrix}. \quad (11.64)$$

The proof is by induction on the number of sweeps. First, partition the data matrix as $(\mathbf{x}, \mathbf{X}, \mathbf{Y})$, where \mathbf{x} is the first variable. The corresponding SSCP matrix is

$$\begin{pmatrix} \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{X} & \mathbf{x}'\mathbf{Y} \\ \mathbf{X}'\mathbf{x} & \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{x} & \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}. \quad (11.65)$$

Now sweep on the first column. The result is

$$\begin{pmatrix} 1/\mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{X}/\mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{Y}/\mathbf{x}'\mathbf{x} \\ -\mathbf{X}'\mathbf{x}/\mathbf{x}'\mathbf{x} & \mathbf{X}'\mathbf{X} - (\mathbf{X}'\mathbf{x})(\mathbf{x}'\mathbf{X})/\mathbf{x}'\mathbf{x} & \mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{x})(\mathbf{x}'\mathbf{Y})/\mathbf{x}'\mathbf{x} \\ -\mathbf{Y}'\mathbf{x}/\mathbf{x}'\mathbf{x} & \mathbf{Y}'\mathbf{X} - (\mathbf{Y}'\mathbf{x})(\mathbf{x}'\mathbf{X})/\mathbf{x}'\mathbf{x} & \mathbf{Y}'\mathbf{Y} - (\mathbf{Y}'\mathbf{x})(\mathbf{x}'\mathbf{Y})/\mathbf{x}'\mathbf{x} \end{pmatrix}. \quad (11.66)$$

We recognize $\mathbf{x}'\mathbf{Y}/\mathbf{x}'\mathbf{x}$ as the least squares estimate $\hat{\beta}$ of the regression coefficient when a regression through the origin on a single explanatory variable \mathbf{x} is fitted. The projection onto $C(\mathbf{x})$ is

$$\mathbf{P}_1 = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' = \frac{\mathbf{x}\mathbf{x}'}{\mathbf{x}'\mathbf{x}},$$

so that the residual sum of squares from this regression, RSS_1 say, is given by

$$\text{RSS}_1 = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \frac{(\mathbf{Y}'\mathbf{x})(\mathbf{x}'\mathbf{Y})}{\mathbf{x}'\mathbf{x}}.$$

Thus we can write the result of the first sweep as

$$\begin{pmatrix} 1/\mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{X}/\mathbf{x}'\mathbf{x} & \hat{\beta} \\ -\mathbf{X}'\mathbf{x}/\mathbf{x}'\mathbf{x} & \mathbf{X}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X} & \mathbf{X}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} \\ -\hat{\beta} & \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X} & \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} \end{pmatrix}. \quad (11.67)$$

We see that sweeping on the first column has produced all the regression quantities for the regression on \mathbf{x} . Ignoring the middle row and column of (11.67), we get (11.64) for a single explanatory variable.

Now let \mathbf{X}_1 represent the first r columns of the regression matrix, \mathbf{x} the $(r+1)$ th column, and \mathbf{X}_2 the remaining columns $r+2$ through p , so that the complete data matrix is $(\mathbf{X}_1, \mathbf{x}, \mathbf{X}_2, \mathbf{Y})$. Now assume that after sweeping the first r columns of the SSCP matrix derived from this data matrix, we get the matrix

$$\begin{pmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1} & (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{x} & (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2 & (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{Y} \\ -\mathbf{x}'\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{x}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{x} & \mathbf{x}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 & \mathbf{x}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} \\ -\mathbf{X}'_2\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{x} & \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 & \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} \\ -\mathbf{Y}'\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1} & \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{x} & \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 & \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} \end{pmatrix} \quad (11.68)$$

containing the regression quantities for the regression on \mathbf{X}_1 . We will show that a further sweep on the $(r + 1)$ th column produces the regression on $(\mathbf{X}_1, \mathbf{x})$. This will show that if the result is true for r sweeps, it must be true for $(r + 1)$ steps, which proves the induction step.

To examine the effect of the $(r + 1)$ th sweep on the upper left four blocks of (11.68), we write these four blocks as

$$\begin{pmatrix} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} & \mathbf{l} \\ -\mathbf{l}' & m^{-1} \end{pmatrix}.$$

If we interchange the rows and then the columns of the matrix above we get

$$\begin{pmatrix} m^{-1} & -\mathbf{l}' \\ \mathbf{l} & (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \end{pmatrix},$$

which is now in the same format as (11.65) (without its middle row and column, and apart from a sign change). Hence sweeping on the first column of the matrix above will give us [cf. (11.66)]

$$\begin{pmatrix} m & -m\mathbf{l} \\ -m\mathbf{l}' & (\mathbf{X}'_1 \mathbf{X}_1)^{-1} + m\mathbf{l}\mathbf{l}' \end{pmatrix}.$$

Interchanging the rows and then the columns back again finally gives the result of the $(r + 1)$ th sweep, namely,

$$\begin{pmatrix} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} + m\mathbf{l}\mathbf{l}' & -m\mathbf{l} \\ -m\mathbf{l}' & m \end{pmatrix}, \quad (11.69)$$

where $m = [\mathbf{x}'(\mathbf{I} - \mathbf{P}_1)\mathbf{x}]^{-1}$, $\mathbf{l} = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{x}$, and $\mathbf{P}_1 = \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1$. By (3.25) (or A.9.1), this is exactly

$$\begin{pmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{x} \\ \mathbf{x}' \mathbf{X}_1 & \mathbf{x}' \mathbf{x} \end{pmatrix}^{-1}.$$

Similar formulas apply to the other blocks. For example, arguing as in Section 11.6.1 [cf. (11.63)], we see that the bottom-right block is transformed by the sweep to

$$\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y} - m(\mathbf{x}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y})^2 = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_A)\mathbf{Y}, \quad (11.70)$$

where \mathbf{P}_A is the projection onto $\mathcal{C}(\mathbf{X}_A)$, and $\mathbf{X}_A = (\mathbf{X}_1, \mathbf{x})$. The complete result of the sweep is

$$\begin{pmatrix} (\mathbf{X}'_A \mathbf{X}_A)^{-1} & (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_2 & (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{Y} \\ -\mathbf{X}'_2 \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} & \mathbf{X}'_A (\mathbf{I}_n - \mathbf{P}_A) \mathbf{X}_2 & \mathbf{X}'_A (\mathbf{I}_n - \mathbf{P}_A) \mathbf{Y} \\ -\mathbf{Y}' \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} & \mathbf{Y}' (\mathbf{I}_n - \mathbf{P}_A) \mathbf{X}_2 & \mathbf{Y}' (\mathbf{I}_n - \mathbf{P}_A) \mathbf{Y} \end{pmatrix}. \quad (11.71)$$

Thus, by induction, we have proved that after sweeping successively on the first r columns, the matrix

$$\begin{pmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{X}_1' \mathbf{X}_2 & \mathbf{X}_1' \mathbf{Y} \\ \mathbf{X}_2' \mathbf{X}_1 & \mathbf{X}_2' \mathbf{X}_2 & \mathbf{X}_2' \mathbf{Y} \\ \mathbf{Y}' \mathbf{X}_1 & \mathbf{Y}' \mathbf{X}_2 & \mathbf{Y}' \mathbf{Y} \end{pmatrix} \quad (11.72)$$

is reduced to the matrix

$$\begin{pmatrix} (\mathbf{X}_1' \mathbf{X}_1)^{-1} & (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2 & (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{Y} \\ -\mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} & \mathbf{X}_1' (\mathbf{I} - \mathbf{P}_1) \mathbf{X}_2 & \mathbf{X}_1' (\mathbf{I} - \mathbf{P}_1) \mathbf{Y} \\ -\mathbf{Y}' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} & \mathbf{Y}' (\mathbf{I} - \mathbf{P}_1) \mathbf{X}_2 & \mathbf{Y}' (\mathbf{I} - \mathbf{P}_1) \mathbf{Y} \end{pmatrix}. \quad (11.73)$$

Thus, after the r th sweep, the matrix (11.73) contains the components of the regression of \mathbf{Y} on the columns of the $n \times r$ matrix \mathbf{X}_1 . Ignoring the middle row and column, we get the equivalent of (11.64).

Sweeping provides an efficient method for successively fitting a sequence of regressions, and is particularly useful in the calculation of all possible regressions discussed in Section 12.8.1. However, the set of variables to be swept in has to be specified in advance. Also, as we discuss in Section 11.8.3, the sweep operator, along with Gaussian elimination, is somewhat vulnerable to round-off error. Next, we discuss a more accurate but less efficient method, based on the QR decomposition.

11.6.3 Adding and Deleting Cases and Variables Using QR

We have seen in Section 11.3 how, given the QR decomposition of (\mathbf{X}, \mathbf{Y}) , we can easily and efficiently calculate the regression coefficients and the residual sum of squares. Thus, to modify a regression, we merely need to modify the QR decomposition. We show how in this section.

Adding a Row

Suppose that our data matrix is modified by the addition of m rows, corresponding to data on m new cases. Using a convenient change in notation, we want to modify the QR decomposition

$$(\mathbf{X}, \mathbf{Y}) = \mathbf{Q} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} \\ 0 & \mathbf{r}_{22} \\ 0 & 0 \end{pmatrix}$$

to get the decomposition of the new data matrix,

$$\begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{X}_m & \mathbf{Y}_m \end{pmatrix}.$$

Consider

$$\begin{pmatrix} \mathbf{Q}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{X}_m & \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} \\ 0 & \mathbf{r}_{22} \\ 0 & 0 \\ \mathbf{X}_m & \mathbf{Y}_m \end{pmatrix},$$

so that

$$\Pi \begin{pmatrix} Q' & 0 \\ 0 & I_m \end{pmatrix} \cdot \begin{pmatrix} X & Y \\ X_m & Y_m \end{pmatrix} = \begin{pmatrix} R_{11} & r_{12} \\ 0 & r_{22} \\ X_m & Y_m \\ 0 & 0 \end{pmatrix},$$

where Π is a permutation matrix (see A.5). Now using Givens transformations to zero out the first, second,..., p th columns of X_m , we get

$$G\Pi \begin{pmatrix} Q' & 0 \\ 0 & I_m \end{pmatrix} \cdot \begin{pmatrix} X & Y \\ X_m & Y_m \end{pmatrix} = \begin{pmatrix} \tilde{R}_{11} & \tilde{r}_{12} \\ 0 & \tilde{r}_{22} \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

where G is the product of the Givens transformations.

Deleting a Row

Suppose that our original data matrix is

$$\begin{pmatrix} x' & y \\ X & Y \end{pmatrix}$$

with QR decomposition

$$\begin{pmatrix} x' & y \\ X & Y \end{pmatrix} = Q \begin{pmatrix} R \\ 0 \end{pmatrix},$$

and we want to delete the first row. Consider augmenting the data matrix with an initial column whose first element is 1 and the rest zero. Then

$$Q' \begin{pmatrix} 1 & x' & y \\ 0 & X & Y \end{pmatrix} = \begin{pmatrix} q_1 & R \\ q_2 & 0 \end{pmatrix},$$

where (q'_1, q'_2) is the first row of Q . Now, starting at the bottom, apply Givens rotations to zero out the first column except for the first element; the second-to-last element is used to zero out the last element, and so on. The result is

$$\begin{pmatrix} \alpha & w' \\ 0 & \tilde{R} \\ 0 & 0 \end{pmatrix},$$

where \tilde{R} is upper triangular. Let G' be the product of these rotations, and put $\bar{Q} = QG$. Then

$$\bar{Q}' \begin{pmatrix} 1 & x' & y \\ 0 & X & Y \end{pmatrix} = \begin{pmatrix} \alpha & w' \\ 0 & \tilde{R} \\ 0 & 0 \end{pmatrix},$$

so that the first row of $\bar{\mathbf{Q}}$ is $(\alpha, \mathbf{0}')$. Thus $\bar{\mathbf{Q}}$ is of the form

$$\begin{pmatrix} \alpha & \mathbf{0}' \\ \mathbf{q} & \tilde{\mathbf{Q}} \end{pmatrix}.$$

Now $\bar{\mathbf{Q}}$ is orthogonal, being the product of orthogonal matrices, so that

$$\mathbf{I} = \bar{\mathbf{Q}} \bar{\mathbf{Q}}' = \begin{pmatrix} \alpha & \mathbf{0}' \\ \mathbf{q} & \tilde{\mathbf{Q}} \end{pmatrix} \begin{pmatrix} \alpha & \mathbf{q}' \\ \mathbf{0} & \tilde{\mathbf{Q}}' \end{pmatrix}.$$

Equating blocks, we see that $\alpha^2 = 1$, $\mathbf{q} = \mathbf{0}$, and that $\tilde{\mathbf{Q}}$ is orthogonal. Moreover,

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{x}' & \mathbf{y} \\ \mathbf{0} & \mathbf{X} & \mathbf{Y} \end{pmatrix} &= \begin{pmatrix} \alpha & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{Q}} \end{pmatrix} \cdot \begin{pmatrix} \alpha & \mathbf{w}' \\ \mathbf{0} & \tilde{\mathbf{R}} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \alpha^2 & \alpha \mathbf{w}' \\ \mathbf{0} & \tilde{\mathbf{Q}} \begin{pmatrix} \tilde{\mathbf{R}} \\ \mathbf{0} \end{pmatrix} \end{pmatrix}, \end{aligned}$$

so that

$$(\mathbf{X}, \mathbf{Y}) = \tilde{\mathbf{Q}} \begin{pmatrix} \tilde{\mathbf{R}} \\ \mathbf{0} \end{pmatrix}$$

is the required QR decomposition.

Deleting a Column

Suppose that we want to delete the column \mathbf{x} from the decomposition

$$\mathbf{Q}'(\mathbf{X}_1, \mathbf{x}, \mathbf{X}_2, \mathbf{Y}) = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} & \mathbf{R}_{13} & \mathbf{r}_{14} \\ \mathbf{0} & \mathbf{R}_{22} & \mathbf{R}_{23} & \mathbf{r}_{24} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_{33} & \mathbf{r}_{34} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & r_{44} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Deleting \mathbf{x} gives

$$\mathbf{Q}'(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{13} & \mathbf{r}_{14} \\ \mathbf{0} & \mathbf{R}_{23} & \mathbf{r}_{24} \\ \mathbf{0} & \mathbf{R}_{33} & \mathbf{r}_{34} \\ \mathbf{0} & \mathbf{0} & r_{44} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

We can then apply Householder or Givens transformations as required to zero out the appropriate elements to reduce the matrix above to upper triangular form.

Adding a Column

Suppose that we have the decomposition

$$(\mathbf{X}, \mathbf{Y}) = \mathbf{Q} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} \\ \mathbf{0} & r_{22} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and we want the decomposition of $(\mathbf{X}, \mathbf{x}, \mathbf{Y})$. First, consider the decomposition of $(\mathbf{X}, \mathbf{Y}, \mathbf{x})$. Let $\mathbf{q} = \mathbf{Q}'\mathbf{x}$, so that

$$\begin{aligned} \mathbf{Q}'(\mathbf{X}, \mathbf{Y}, \mathbf{x}) &= [\mathbf{Q}'(\mathbf{X}, \mathbf{Y}), \mathbf{q}] \\ &= \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} & \mathbf{q}_1 \\ \mathbf{0} & r_{22} & q_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{q}_3 \end{pmatrix}, \end{aligned}$$

say. If we interchange the last two columns, we get

$$\mathbf{Q}'(\mathbf{X}, \mathbf{x}, \mathbf{Y}) = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{q}_1 & \mathbf{r}_{12} \\ \mathbf{0} & q_2 & r_{22} \\ \mathbf{0} & \mathbf{q}_3 & \mathbf{0} \end{pmatrix}.$$

Finally, we can apply Householder transformations to zero out \mathbf{q}_3 .

11.7 CENTERING THE DATA

We saw in Section 9.7 how centering and scaling the regression data has certain benefits. In this section we describe some algorithms for calculating the centered sum of squares and cross-products matrix. To compute the elements

$$c_{jj'} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \quad (11.74)$$

of the matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, several options are available. Chan et al. [1983] give a comprehensive discussion of the issues involved, and the following is a summary of their results. The algorithms available include the following.

- (a) The *textbook algorithm*. This computes the uncorrected sums of squares and cross products simultaneously and then calculates

$$c_{jj'} = \sum_{i=1}^n x_{ij}x_{ij'} - \frac{1}{n} \sum_{i=1}^n x_{ij} \sum_{i=1}^n x_{ij'}.$$

(b) The *updating algorithm*. Let

$$\begin{aligned} T_{k,j} &= \sum_{i=1}^k x_{ij}, \\ M_{k,j} &= T_{k,j}/k, \\ c_{k,j,j'} &= \sum_{i=1}^k (x_{ij} - M_{k,j})(x_{ij'} - M_{k,j'}), \end{aligned}$$

so that $c_{jj'} = c_{n,j,j'}$. These quantities are computed using the formulas

$$\begin{aligned} M_{k,j} &= M_{k-1,j} + (x_{kj} - M_{k-1,j})/k, \\ c_{k,j,j'} &= c_{k-1,j,j'} + \frac{k-1}{k}(x_{kj} - M_{k-1,j})(x_{kj'} - M_{k-1,j'}). \end{aligned}$$

(c) The *two-pass algorithm*. This computes the means \bar{x}_j on a first pass through the data, and then the $c_{jj'}$ on a second pass, using (11.74).

The first two algorithms require only one pass through the data. If the data are too numerous to fit into the high-speed memory of a computer, this can be an advantage. However, as we shall see below, these methods can be inaccurate, particularly the first. The two-pass algorithm is less efficient but more accurate.

The accuracy of the algorithms depends on the *condition number* κ_j of each column, defined by

$$\kappa_j = \left\{ \frac{\sum_{i=1}^n x_{ij}^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \right\}^{1/2}. \quad (11.75)$$

The condition number is related to the coefficient of variation of the column by the equation $\kappa_j^2 = 1 + CV_j^{-2}$ (cf. Section 9.7.4). Thus, the condition number is large if the variation in a column is small relative to the mean.

The relative errors (at least of the diagonals) of the computed elements \hat{c}_{jj} satisfy

$$\left| \frac{c_{jj} - \hat{c}_{jj}}{c_{jj}} \right| \leq C(n, \kappa)u, \quad (11.76)$$

where the quantities $C(n, \kappa)$ are given in Table 11.1 and u is the unit round-off of the computer, a measure of computational accuracy which is discussed further in Section 11.8.3. From Table 11.1, we see that the textbook method is unlikely to give accurate results if the condition number is large. The two-pass method, on the other hand, is very accurate, since u is very small.

The condition of the columns can be improved by subtracting a constant d_j from each column. This does not change the value of $c_{jj'}$ but changes the condition from κ_j to

$$\tilde{\kappa}_j = \left\{ \frac{\sum_{i=1}^n (x_{ij} - d_j)^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \right\}^{1/2},$$

Table 11.1 Quantities $C(n, \kappa)$ for various methods

Method	$C(n, \kappa)$
Textbook	$n\kappa^2 u$
Updating	$n\kappa u$
Two-pass	$nu + n^2\kappa^2 u^2$

which will be much smaller than κ_j if d_j is close to \bar{x}_j . Possibilities for d_j for the one-pass algorithms include “eyeballing” the data or using the first observation (i.e., putting $d_j = x_{1j}$).

If we use the two-pass algorithm, we can use the computed means as the d_j . This results in a condition very close to 1; it would be exactly 1 if the computations were exact. If we use the textbook algorithm with the data centered in this way, we get a very accurate algorithm, corresponding to a $C(n, \kappa)$ of $nu(1 + n^2\kappa^2 u^2)$, where κ is the condition of the uncentered data.

As a further refinement, we can replace the standard summations by *pairwise summations*, where the data are summed recursively. More detail on pairwise summation is given by Chan et al. [1983] and Higham [1996: p. 92]. The effect is to substitute $\log n$ for n in the quantities $C(n, \kappa)$, resulting in more accurate calculations.

To sum up, the two-pass algorithm is usually quite satisfactory for even ill-conditioned data, and the textbook algorithm applied to the mean-centered data is very accurate indeed. If n is very large and the two-pass algorithm is deemed too expensive, a constant should be subtracted from each column, and if necessary, pairwise summation should be used.

11.8 COMPARING METHODS

When comparing methods for regression calculations, several factors are important. These are:

Resources: How much memory does a method require?

Efficiency: How much time does a method require?

Accuracy: To what extent is the method affected by round-off error? How close are the *computed* regression quantities to the *true* quantities?

We discuss each of these aspects in turn.

11.8.1 Resources

If the number of cases greatly exceeds the number of variables, then the methods based on the $(p+1) \times (p+1)$ augmented SSCP matrix will require

much less computer storage than those based on the QR decomposition. The data for each case can be read in from external storage and the SSCP matrix accumulated, requiring only $(p + 1)(p + 2)/2$ storage locations. By contrast, working with the matrix (\mathbf{X}, \mathbf{Y}) and using the QR decomposition will require at least $n(p + 1)$ storage locations. It is for this reason that some computer packages that are designed to handle large amounts of data (such as the SAS regression program PROC REG) base their calculations on the SSCP matrix. This involves some sacrifice, as we shall see below, since methods based on the SSCP matrix are less accurate than those which use the QR decomposition. On the other hand, packages designed for more intensive exploration of smaller data sets, such as R or S-PLUS, rely principally on QR methods. (The Householder method is the default in the S-PLUS function `lm()`.) For smaller data sets, the requirement that \mathbf{X} be stored is of course not a difficulty.

11.8.2 Efficiency

Traditionally, the time it takes to execute an algorithm has been measured by counting the number of arithmetic operations required. The drawbacks of this approach have been noted by several authors. For example, Golub and Van Loan [1996: p. 19] remark that

Flop counting is a necessarily crude approach to the measuring of program efficiency, since it ignores subscripting, memory traffic, and the countless other overheads associated with program execution... We must not infer too much from flop counts... Flop counting is just a "quick and dirty" accounting method that captures only one of the several dimensions of the efficiency issue.

Nevertheless, counting flops continues to be common practice. The main disadvantage of using flop counts as a proxy for time seems to be that in most modern computers, floating-point operations are carried out using a floating-point coprocessor. The main microprocessor has many other tasks to perform, including the integer arithmetic required to manage the loops inherent in the algorithms we have described. The net result is that the non-floating-point operations are, in terms of time, a significant part of the computational load, and most of the arithmetic will be done in parallel to these non-floating-point operations.

However, it seems that the flop count remains a reasonable proxy for the time an algorithm takes. Part of the reason for this is that the flop count is proportional to the number of times the innermost loop in an algorithm is executed, which is in turn proportional to the time required. Most of the algorithms we have described require two floating-point operations (a multiplication and an addition) each time a loop is executed, plus the overhead required to manage the iteration. Thus, if the time is proportional to the flop count, the constant of proportionality will be roughly the same for each algorithm.

EXAMPLE 11.1 Imagine performing a regression by forming the SSCP matrix and then calculating its Cholesky decomposition. As stated in Section 11.2.1, forming the SSCP matrix takes about np^2 flops, and (see the discussion later in this section) the Cholesky decomposition takes about $\frac{1}{3}p^3$ flops. Alternatively, we could use the Householder QR, which (Section 11.3.2) requires $2np^2 - \frac{2}{3}p^3$ flops. We timed the implementation of these procedures in R, using the functions supplied which are coded in C and are based on the routines in LINPACK. In Figure 11.1 we plot the elapsed time versus the flop count for these two methods, for different regression problems of varying sizes. The figure reveals that the relationship between the flop count and the elapsed time is very strong for both methods, so that time is proportional to flop count for the two methods. The constants of proportionality are quite similar, being 1.24 for the Cholesky approach and 1.09 for Householder. It seems that at least for these two algorithms, the flop counts are a reasonable basis on which to compare the execution times. \square

Encouraged by this, we now compare the flop counts for the various algorithms. The GE algorithm has a flop count of about $\frac{2}{3}p^3$, ignoring terms in p^2 or smaller. Use of partial pivoting is essential for accuracy, but unfortunately this destroys the symmetry of the matrix.

Once the upper triangle U in the LU decomposition of $\mathbf{X}'\mathbf{X}$ has been computed [cf. (11.2)], the back-substitution takes only about $\frac{1}{2}p^2$ flops, so that most of the work lies in the decomposition. Moreover, if n is much greater than p , as is typically the case in regression, the amount of computation required to

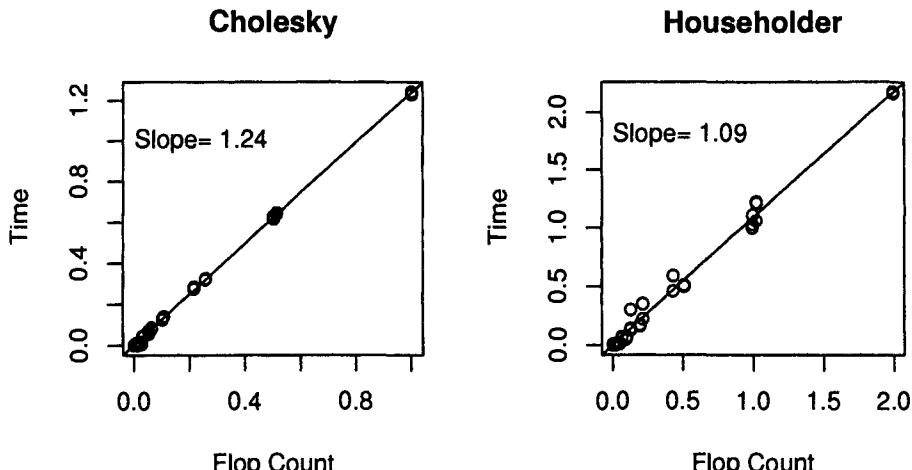


Fig. 11.1 Plot of elapsed time (seconds) versus flop count (10^8 flops) for two different regression algorithms.

solve the normal equations is small compared to the work involved in forming the SSCP matrix, which takes np^2 flops.

If we exploit the positive-definiteness of the SSCP matrix and use the Cholesky method of Section 11.2.2, the decomposition takes about $\frac{1}{3}p^3$ flops, plus a smaller number of operations for the back-substitutions. This is about one half of the work required to solve the normal equations by Gaussian elimination with partial pivoting. Thus, for problems where n is considerably greater than p , Cholesky is more efficient than Gaussian elimination and sweeping. The latter are more useful for fitting sequences of models.

The Householder method requires about $2np^2 - \frac{2}{3}p^3$ flops to form the QR decomposition. The explicit formation of \mathbf{Q} requires a further $4(n^2p - np^2 + \frac{1}{3}p^3)$ flops, but this is not usually required. If we use the basic Givens method, the cost is about 50% more than that of Householder, while using fast Givens is somewhere in between. The MGSA is very slightly more expensive than Householder, taking about $2np^2 + 2p^3$ flops.

The cost of the SVD depends on how much of the decomposition is required. If \mathbf{U} and \mathbf{V} are explicitly required [cf. (11.57)], the cost is greater, because they must be assembled from the stored Householder vectors, just as in the QR decomposition. However, the basic regression calculations require only that we be able to calculate $\mathbf{U}'\mathbf{Y}$, so that \mathbf{U} need not be formed. The cost also depends on whether a preliminary QR step is performed. If this is done, then the cost of a least squares calculation is about $2np^2 + 11p^3$ flops, slightly more than the Householder QR. If the full thin SVD is required (i.e., with \mathbf{U}_p explicitly formed), the cost rises to $6np^2 + 11p^3$. For more detail on SVD flop counts, see Golub and Van Loan [1996: p. 254].

We may summarize the discussion above by contrasting the Cholesky and GE methods (which have roughly equal efficiency) to the QR methods, which require roughly twice as much computation, with Householder requiring slightly less than fast Givens or the MGSA. Table 11.2 summarizes the flop counts to calculate the regression coefficients and the residual sum of squares for the various methods. We should not make too much of these flop counts. For all but very large problems, the amount of time required to compute the regression quantities is minute compared to the human work involved in fitting

Table 11.2 Approximate flop counts for different regression methods

Method	Flop count
Gaussian elimination	$np^2 + \frac{2}{3}p^3$
Cholesky	$np^2 + \frac{1}{3}p^3$
Householder	$2np^2 - \frac{2}{3}p^3$
Givens	$3np^2 - p^3$
MGSA	$2np^2$
SVD	$2np^2 + 11p^3$

models, examining plots and diagnostics, and refitting to obtain satisfactory fits. Questions of efficiency are only likely to have practical importance when problems are truly large. Of more importance for small to medium-sized problems is the accuracy of calculations, which we discuss next.

11.8.3 Accuracy

The analysis of accuracy is a much more complex question than merely counting storage requirements or flops. We need to distinguish between difficult problems, where the data make accurate calculations difficult, and poor algorithms, which can fail where better algorithms succeed. We begin by discussing the idea of *condition*, which is a measure of the difficulty of getting an accurate solution to the normal equations.

Roughly speaking, a regression problem is *well-conditioned* if using perfectly accurate arithmetic, small changes in the input data \mathbf{X} and \mathbf{Y} cause only small changes in the regression coefficients. Conversely, a problem is *ill-conditioned* if small changes in inputs cause large changes in outputs. As we saw in Chapter 10, this happens when the condition number of the matrix \mathbf{X} is large.

The relative change in the regression coefficients when the data are subject to small perturbations was given by (9.63) in Section 9.7. The dominant terms in this equation are κ and $\kappa^2\|\mathbf{e}\|$, so if the fit is good and the residuals are very small, the perturbations are essentially bounded by the condition number of \mathbf{X} . On the other hand, if the fit is not good, the bound is essentially the square of the condition number.

Even if the data for a problem are exact, the way that numbers are stored in modern computers means that the data actually used in calculations will usually differ slightly from the input data. Modern computers use a *floating point representation* for numbers, of the form

$$\pm m \times \eta^{t-e}, \quad (11.77)$$

where η , t , m , and e are integers. The quantities η and t are called the *base* and *precision*, respectively, and are fixed for a given computer. The quantities m and e are the *mantissa* and *exponent*. If we require that the representation be unique, we must have

$$\eta^{t-1} \leq m < \eta^t.$$

The exponent is fixed in the range $e_{\min} \leq e \leq e_{\max}$, where e_{\max} and e_{\min} are also fixed for a given computer.

For any real number x that lies between two floating-point numbers of the form (11.77), there is a floating-point number x' such that

$$\frac{|x - x'|}{|x|} < u,$$

where $u = \frac{1}{2}\eta^{1-t}$. A proof of this result can be found in Higham [1996: p. 42]. The quantity u , called the *unit round-off*, measures the relative accuracy with which numbers can be stored.

The accuracy with which arithmetic operations can be carried out is obviously crucial for least squares calculations. There is a numeric standard (IEEE 754) which prescribes that arithmetic operations be accurate to the unit round-off; that is, the relative error made in carrying out an arithmetic operation ($+, -, \times, \div, \sqrt{\cdot}$) on two floating-point numbers using computer arithmetic is less than the unit round-off. The IEEE standard also requires that the base η be 2, and that for *single precision*, the precision t be 24 and the exponent range -125 to 128, while for *double precision*, t is 53 and the exponent range is -1021 to 1024. Computers adhering to the IEEE standard include those based on Pentium chips, DEC, Hewlett-Packard, and Sun. Most modern work on the effect of round-off error in algorithms assumes that the calculations are being performed according to the IEEE standard.

It is clear from the discussion above that some perturbation of the data is to be expected. The condition number indicates the extent to which this will affect the accuracy of the estimated coefficients, if we could use accurate arithmetic. Of course, this is not possible. To assess the effect of round-off, which will depend on the algorithm used, the primary tool is *backward analysis*. In backward analysis, the computed solution is represented as the exact solution of a perturbed problem. Thus, if $\tilde{\mathbf{b}}$ is the vector of regression coefficients computed using some algorithm, then a backward analysis of the algorithm seeks to represent $\tilde{\mathbf{b}}$ as the exact solution of a perturbed problem with data $\mathbf{X} + \delta\mathbf{X}$ and $\mathbf{Y} + \delta\mathbf{Y}$. For *backward stable* algorithms, these perturbations will be small for well-conditioned problems. The perturbation result (9.63) can then be used to get a bound on the error in the computed solution. These bounds depend on the condition number and indicate that even a stable algorithm can give poor results if the problem is sufficiently ill-conditioned.

Let \mathbf{b} and $\tilde{\mathbf{b}}$ be the exact and computed solutions to the normal equations. If the regression calculations are performed using the Cholesky decomposition, then it can be shown that the backward analysis leads approximately to the bound (Higham [1996: p. 398])

$$\frac{||\mathbf{b} - \tilde{\mathbf{b}}||}{||\mathbf{b}||} \leq C\kappa^2 u, \quad (11.78)$$

where C depends on the dimensions of the problem, but not on the problem data. This bound allows for the rounding errors that occur in the formation of the SSCP matrix. The GE method has a similar bound, but the constant is very much greater, even when partial pivoting is used. Thus, Cholesky is more reliable than GE. Note, though, that partial pivoting is more accurate in practice than the large constant C would indicate (cf. Higham [1996: p. 177] for a discussion).

Conversely, if Householder transformations are used, the following can be proved. The computed solution $\tilde{\mathbf{b}}$ is the exact solution to a perturbed problem

with data $(\mathbf{X} + \mathbf{E}_X, \mathbf{Y} + \mathbf{E}_Y)$ such that [cf. (9.57)]

$$\frac{\|\mathbf{E}_X\|_2}{\|\mathbf{X}\|_2} \leq (6n - 3p + 41)n^{3/2}u \quad (= \epsilon_X, \text{ say})$$

and

$$\frac{\|\mathbf{E}_Y\|_2}{\|\mathbf{Y}\|_2} \leq (6n - 3p + 41)u \quad (= \epsilon_Y, \text{ say}).$$

These bounds will be small for all but enormous problems (cf. Lawson and Hanson [1995: p. 90ff], Björck [1996: p. 64], Björck and Paige [1994], and Higham [1996: p. 395]). Combining these bounds with the perturbation analysis (9.63), we find from the last reference that a bound on the relative error in $\tilde{\mathbf{b}}$ is given by

$$\frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|_2}{\|\mathbf{b}\|_2} \leq \frac{\kappa u}{1 - \kappa \epsilon_X} \left(\epsilon_X + \frac{\epsilon_Y \|\mathbf{Y}\|}{\|\mathbf{b}\| \|\mathbf{X}\|_2} + \epsilon_X \kappa \frac{\|\mathbf{e}\|}{\|\mathbf{b}\| \|\mathbf{X}\|_2} \right) + \epsilon_X \kappa,$$

where \mathbf{e} is the residual $\mathbf{Y} - \mathbf{X}\mathbf{b}$. The dominating parts of this bound are the terms proportional to $\kappa^2 \|\mathbf{e}\|$ and κ , which contrasts with the term in κ^2 in the case of the Cholesky method. Thus provided the regression fit is good (i.e., $\|\mathbf{e}\|$ is small), use of the Householder method should result in a more accurate calculation. These bounds can be very conservative and should be interpreted qualitatively rather than quantitatively.

The Givens and MGSA methods give results that are broadly similar in accuracy to the Householder method. It is interesting to note that the MGSA can be interpreted as the Householder method applied to an augmented matrix and so shares the desirable properties of the Householder method as a way of computing the “R” part of the QR factorization (cf. Björck and Paige [1992], and Higham [1996: p. 379] for details). However, it is not so stable for the computation of the \mathbf{Q} matrix in the QR decomposition. In fact, the columns of \mathbf{Q} may be quite far from orthogonal for ill-conditioned matrices (Björck [1996: p. 66]). Thus, when using the MGSA to solve the least squares problem, it should be implemented in the augmented form (11.18) to avoid having to multiply explicitly by \mathbf{Q}_p when computing $\mathbf{r} = \mathbf{Q}'_p \mathbf{Y}$ as in (11.18). Implemented in this way, MGSA is roughly as accurate as Householder.

The SVD has very desirable stability properties. A backward analysis (see Björck [1996: p. 89]) shows that the SVD computed using the methods described above will be the exact SVD of a matrix $\mathbf{X} + \mathbf{E}$, where $\|\mathbf{E}\|$ is a small multiple of the unit round-off. A characteristic of the SVD is that small changes in the matrix entries cause only small changes in the singular values. The Weilandt–Hoffman equality (see Golub and Van Loan [1966: p. 449]) implies that

$$|\sigma_k - \hat{\sigma}_k| \leq C \|\mathbf{E}\|_2 \tag{11.79}$$

for all the exact singular values σ_k and the computed singular values $\hat{\sigma}_k$. It follows that the SVD computes the singular values with very small absolute error.

The accuracy of the computed \mathbf{U} and \mathbf{V} [cf. (11.57)] is a more complex question. If the singular values are well separated, then small changes in \mathbf{X} will not result in large changes in \mathbf{U} and \mathbf{V} . However, if two singular values are very close, even a small change in \mathbf{X} can result in a large change in \mathbf{U} and \mathbf{V} . Despite this, in practice the SVD is the most accurate of all the methods we have discussed.

11.8.4 Two Examples

In this section we try out the numerical methods described in the previous sections on two poorly conditioned regressions. The calculations were performed in R, which uses routines based on those in LINPACK and LAPACK for its Cholesky, Householder, and SVD functions. The MGSA algorithm was programmed directly in R using the equations given in Algorithm 11.5 in Section 11.3.2. R calculates in double precision, and the calculations were performed on a machine using IEEE arithmetic.

EXAMPLE 11.2 The first data set tried was the well-known Longley data (Longley [1967]), which have been used by many writers to illustrate the effects of ill-conditioning. The data set consists of 16 observations on seven variables. The condition number of the \mathbf{X} matrix (including the constant term) is 2.38×10^7 . Despite the bad reputation of the Longley data, with its very high condition number, the Cholesky, Householder, MGSA, and SVD methods all give identical answers to seven significant figures. \square

EXAMPLE 11.3 The second data set is adapted from Trefethen and Bau [1997: p. 137] and involves fitting a function over the interval $[0,1]$ using a polynomial of degree $p - 1$ fitted at equally spaced points $0 = x_1 < x_2 < \dots < x_n = 1$, with $n = 100$. The i, j element of the SSCP matrix is [cf. equation (7.2) with $r = i - 1$ and $s = j - 1$]

$$\sum_{l=1}^n x_l^{i-1} x_l^{j-1} \approx n \int_0^1 x^{i+j-2} dx = n/(i + j - 1),$$

so that the SSCP matrix is very close to a multiple of the $p \times p$ Hilbert matrix with i, j element $1/(i + j - 1)$. This is known to be very badly conditioned, with condition number approximately $e^{3.5p}$ for large p . Suppose that we want to fit the function

$$f(t) = \exp(\sin 4t)$$

using the approximation

$$f(t) \approx \sum_{j=0}^p b_j t^j,$$

with the coefficients estimated by least squares. Suppose that we use a polynomial of degree $p = 14$. Table 11.3 shows the estimates of b_{14} calculated by the Cholesky, Householder, MGSA and SVD methods.

The correct answer to seven significant figures is 2.006787×10^3 , so that the MGSA and SVD are entirely correct. The Cholesky method has failed completely, and the Householder method is also unsatisfactory.

If we repeat the calculations for polynomials of degree p , we get the coefficients of the term in t^p , as shown in Table 11.4. The same pattern is apparent: Cholesky fails first, followed by Householder. The performance of the MGSA is remarkably good. This example may overstate the performance of the Cholesky method due to the special properties of the Hilbert matrix (see Higham [1996: p. 515] for a discussion). \square

Table 11.3 Estimated coefficient of t^{14} for various regression methods

Method	Estimated coefficient
Cholesky	-8.087709×10^2
Householder	1.512535×10^3
MGSA	2.006787×10^3
SVD	2.006787×10^3

Table 11.4 Estimated coefficient of t^p for various regression methods

Method	$p = 10$	$p = 11$	$p = 12$	$p = 13$
Cholesky	8.302081	1174.053	492.9003	-848.3019
Householder	9.570910	1296.930	-488.8949	460.9971
MGSA	9.570910	1296.930	-488.8949	-2772.221
SVD	9.570910	1296.930	-488.8949	-2772.221
Condition number	2.11×10^7	1.20×10^8	6.90×10^8	3.95×10^9

11.8.5 Summary

Generally speaking, the more accurate the method, the more computation is required. Table 11.5 lists the methods in increasing order of accuracy, ranging from the Cholesky method (the least accurate) to the SVD (the most accurate). The table entry is the flop count, expressed as a percentage of the flop count for the SVD, for various values of the ratio n/p . The table shows that there is a considerable difference between the QR and SVD methods for small values of n/p , but that this difference disappears as n/p gets large. The Cholesky method will be about half the cost of the QR methods but will be less accurate if the regression fit is good.

Table 11.5 Time required (as a % of SVD) by different methods for fitting regressions

n/p	2	3	5	10	20	30
Cholesky	16	20	25	33	44	43
Householder	22	31	44	62	77	84
MGSA	27	35	48	65	78	85
SVD	100	100	100	100	100	100

As noted above, when choosing a method, there is a trade-off between speed and accuracy. For small to medium-sized problems, the speed of modern computers means that there is very little absolute time cost in using the SVD, even if it takes five times as long as the Cholesky method. For large problems, or when many regressions are being fitted (as in all possible regressions; see Section 12.8.1) time may become a factor, and less accurate but faster methods may be preferred.

It should be stressed that if calculations are carried out in double precision according to the IEEE standard (which gives a unit round-off of about 10^{-16}), even the Cholesky method will be accurate unless the condition of the problem is very bad. The statistical difficulties resulting from the collinear explanatory variables will be a problem long before the numerical difficulties become an issue.

EXERCISES 11c

1. Prove that the condition number of \mathbf{X}_w is given by (11.61).
 2. Repeat the calculations in Example 11.3 using the function $f(t) = \cos 4t$ in place of $\exp(\sin 4t)$. Does the ranking of the methods still hold good? Sample R-code is given after Exercise 3 below.
 3. Modify the code given below to count the number of flops that are required to execute each algorithm. How accurate are the formulas given in Table 11.2?

```

#####
# MGSA function: calculates Q and R^{-1} of QR decomp
mgsa<-function(A){
  n<-dim(A)[1]
  p<-dim(A)[2]
  GG<-diag(p)
  for(i in 1:(p-1)){
    a<-A[,i]
    denom<-sum(a*a)
    G<-numeric(p)
    for(j in (i+1):p){
      G[j]<-a[j]/denom
      A[j,]<-A[j,]-G[j]*a
    }
    GG[i,(i+1):p]<-G
  }
  GG
}
```

```

num<-sum(A[,j]*a)
G[j]<- -num/denom
A[,j]<-A[,j]+a*G[j]
}
GG<-GG + outer(GG[,i],G)
}
list(W=A,G=GG)
}
#####
#Householder function: calculates R of QR decomp of A
house<-function(A){
n<-dim(A)[1]
p<-dim(A)[2]
for(j in 1:p){
  indices<-j:n
  # construct Householder vector
  u<-numeric(n)
  u[indices]<-A[indices,j]
  unorm1<-sqrt(sum(u[indices]^2))
  unorm2<-sum(u[indices[-1]]^2)
  uu<-u[j]
  u[j]<-if(uu<0)uu-unorm1 else
    -unorm2/(uu+unorm1)
  gamma<-0.5*(unorm2 + u[j]^2)
  # multiply by householder matrix
  A[indices,j]<-0
  A[j,j]<-unorm1
  if(j==p) return(A)
  for(l in (j+1):p){
    k<-sum(A[indices,l]*u[indices])/gamma
    A[indices,l]<-A[indices,l] - k*u[indices]
  }
}
A
}
#####
# Givens function: calculates R of QR decomp of A
givens<-function(A){
n<-dim(A)[1]
p<-dim(A)[2]
for(j in 1:(p-1)){
  indices<-j:p
  for(i in (j+1):n){
    sqr<-sqrt(A[j,j]^2+A[i,j]^2)
    cc<-A[j,j]/sqr
    ss<-A[i,j]/sqr
    A[j,j]<-cc*sqr
    A[i,j]<-ss*sqr
    A[i,i]<-A[i,i] - ss*ss*sqr
  }
}
A
}

```

```

    ss<-A[i,j]/sqr
    temp<-cc*A[j,indices] + ss*A[i,indices]
    A[i,indices]<- -ss*A[j,indices] + cc*A[i,indices]
    A[j,indices]<-temp
  }
}
A
}

```

11.9 RANK-DEFICIENT CASE

In this book we have mostly assumed that the design matrix \mathbf{X} has full rank, which we can always achieve by deleting variables from the regression. However, in practice we may fail to realise that we have included variables that are linearly dependent on previous variables. We need to modify our algorithms to detect linear dependencies and to delete the offending variables from the regression. In this section we discuss how this may be done.

11.9.1 Modifying the QR Decomposition

The QR decomposition described in Section 11.3 can be carried out without modification if \mathbf{X} is rank-deficient, but some of the diagonal elements of \mathbf{R} will be zero. We can arrange to have the zeros occupy the end positions on the diagonal, and the positive diagonal elements be decreasing in size, if we modify the QR algorithm to incorporate *column pivoting*. If \mathbf{X} is of rank $r < p$, the modification produces a decomposition of the form

$$\mathbf{H}\mathbf{\Xi}\mathbf{\Pi} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad (11.80)$$

where \mathbf{H} is a product of Householder transformations, $\mathbf{\Pi}$ is a product of permutation matrices that swaps columns (cf. A.5), \mathbf{R}_{11} is an $r \times r$ upper triangular matrix with positive diagonal elements, and \mathbf{R}_{12} is $r \times (p-r)$.

Algorithm 11.7

Step 1: Find the column of \mathbf{X} having greatest norm and swap it with the first. Apply a Householder transformation to zero out the first column, leaving a positive first diagonal element. The result is

$$\mathbf{H}_1\mathbf{\Xi}\mathbf{\Pi}_1 = \begin{pmatrix} r_{11}^{(1)} & \mathbf{R}_{12}^{(1)} \\ \mathbf{0} & \mathbf{R}_{22}^{(1)} \end{pmatrix}, \quad (11.81)$$

where Π_1 represents the column interchange, and $r_{11}^{(1)}$ is 1×1 with a positive element.

Step $s + 1$: Assume that we have carried out s steps, obtaining a decomposition

$$\mathbf{H}_s \mathbf{H}_{s-1} \cdots \mathbf{H}_1 \mathbf{X} \Pi_1 \cdots \Pi_s = \begin{pmatrix} \mathbf{R}_{11}^{(s)} & \mathbf{R}_{12}^{(s)} \\ \mathbf{0} & \mathbf{R}_{22}^{(s)} \end{pmatrix}, \quad (11.82)$$

where $\mathbf{R}_{11}^{(s)}$ is $s \times s$ upper triangular with positive diagonal elements and $\mathbf{R}_{12}^{(s)}$ is $s \times (p - s)$. Now we describe how to carry out the next stage. If all the columns of $\mathbf{R}_{22}^{(s)}$ have zero norm, then $\mathbf{R}_{22}^{(s)} = \mathbf{0}$, and \mathbf{X} has the desired decomposition with $r = s$. If at least one column is nonzero, find the column of $\mathbf{R}_{22}^{(s)}$ with greatest norm and swap the corresponding column of the entire matrix (11.82) with the $(s + 1)$ th column. This is equivalent to postmultiplication with a permutation matrix Π_{s+1} . Reduce the first column of the resulting new $\mathbf{R}_{22}^{(s)}$ (except the first element) to zero with a Householder transformation \mathbf{H}_{s+1} . The result is

$$\mathbf{H}_{s+1} \cdots \mathbf{H}_1 \mathbf{X} \Pi_1 \cdots \Pi_{s+1} = \begin{pmatrix} \mathbf{R}_{11}^{(s+1)} & \mathbf{R}_{12}^{(s+1)} \\ \mathbf{0} & \mathbf{R}_{22}^{(s+1)} \end{pmatrix}. \quad (11.83)$$

This is of the form desired since the $(s + 1) \times (s + 1)$ matrix $\mathbf{R}_{11}^{(s+1)}$ has positive diagonal elements, and because the product of permutation matrices is a permutation matrix.

Note that this procedure also guarantees that the diagonal elements are nonincreasing. The algorithm either proceeds for p steps or else terminates after $r < p$ steps. In the former case the result is

$$\begin{pmatrix} \mathbf{R}_{11}^{(p)} \\ \mathbf{0} \end{pmatrix},$$

and all the diagonal elements of the $p \times p$ matrix $\mathbf{R}_{11}^{(p)}$ are positive, so that \mathbf{X} is full rank. However, in the latter case, the result is

$$\begin{pmatrix} \mathbf{R}_{11}^{(r)} & \mathbf{R}_{12}^{(r)} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{R}_{11}^{(r)}$ is $r \times r$, so that the first r diagonal elements are positive and the rest are zero. In this case, \mathbf{X} is of rank r .

11.9.2 Solving the Least Squares Problem

If Π is a permutation matrix and $\hat{\beta}$ minimizes $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ as a function of \mathbf{b} , then $\Pi'\hat{\beta}$ minimizes $\|\mathbf{Y} - \mathbf{X}\Pi\mathbf{b}\|^2$. This follows because every permutation matrix is an orthogonal matrix. Hence we can work with the column-permuted version of \mathbf{X} and then apply the permutation Π to the resulting solution to recover the solution $\hat{\beta}$ to the original problem.

Using the modified decomposition

$$\mathbf{H}\mathbf{X}\Pi = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & 0 \end{pmatrix},$$

we get

$$\mathbf{X}\Pi = \mathbf{Q} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & 0 \end{pmatrix},$$

where $\mathbf{Q} = \mathbf{H}'$ is orthogonal. Now let $\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2')'$ and $\mathbf{Q}'\mathbf{Y} = (\mathbf{d}_1', \mathbf{d}_2')'$. Then, since $(\mathbf{Y} - \mathbf{a})'(\mathbf{Y} - \mathbf{a}) = (\mathbf{Y} - \mathbf{a})'\mathbf{Q}\mathbf{Q}'(\mathbf{Y} - \mathbf{a})$,

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\Pi\mathbf{b}\|^2 &= \left\| \mathbf{Y} - \mathbf{Q} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \right\|^2 \\ &= \left\| \mathbf{Q}'\mathbf{Y} - \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \right\|^2 \\ &= \left\| \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix} - \begin{pmatrix} \mathbf{R}_{11}\mathbf{b}_1 + \mathbf{R}_{12}\mathbf{b}_2 \\ 0 \end{pmatrix} \right\|^2 \\ &= \|\mathbf{d}_1 - \mathbf{R}_{11}\mathbf{b}_1 - \mathbf{R}_{12}\mathbf{b}_2\|^2 + \|\mathbf{d}_2\|^2. \end{aligned}$$

This is obviously minimized when

$$\mathbf{R}_{11}\mathbf{b}_1 + \mathbf{R}_{12}\mathbf{b}_2 = \mathbf{d}_1. \quad (11.84)$$

Equation (11.84) has infinitely many solutions; we choose the solution with $\mathbf{b}_2 = \mathbf{0}$, corresponding to deleting the last $p - r$ columns of the permuted matrix $\mathbf{X}\Pi$ from the regression. The solution is calculated by solving $\mathbf{R}_{11}\mathbf{b}_1 = \mathbf{d}_1$ in the usual way.

11.9.3 Calculating Rank in the Presence of Round-off Error

Our description of the modified QR algorithm has been in terms of exact arithmetic. In practice, due to round-off error, at no stage we will find that the norms of the columns of $\mathbf{R}_{22}^{(s)}$ are exactly zero. However, we can specify some small tolerance δ and interpret $\mathbf{R}_{22}^{(s)}$ as zero if the maximum norm is less than δ . If this happens for the first time when $s = r + 1$, it can be shown via a backward error analysis that the computed matrix \mathbf{R} is the result of applying the QR algorithm using exact arithmetic to a matrix $\mathbf{X} + \mathbf{E}$, where

$$\|\mathbf{E}\|_2 \leq \delta + c_1 u n^{1/2} \|\mathbf{X}\|_2$$

for some small constant c_1 . Thus the computed QR decomposition is that of a matrix of rank r close to \mathbf{X} .

However, we cannot guarantee that if the original \mathbf{X} is of rank $r < p$, then the computed \mathbf{R} will also have rank r . Björck [1996: p. 105] or Golub and Van Loan [1996: p. 260] give an example of a matrix with a very high condition number for which the maximum norms of all the $\mathbf{R}_{22}^{(s)}$ are not small. In practice, such behavior is rare, and the modified QR algorithm will usually (but not always) give an indication of rank deficiency. However, to be more certain, we should use the singular value decomposition.

11.9.4 Using the Singular Value Decomposition

We assume that $\text{rank}(\mathbf{X}) = r < p$. As in Section 11.4, the regression quantities can be calculated in terms of the SVD as follows. Let $\boldsymbol{\alpha} = \mathbf{V}'\mathbf{b}$, and let \mathbf{u}_i be the i th column of \mathbf{U} . Then

$$\begin{aligned}\|\mathbf{X}\mathbf{b} - \mathbf{Y}\|^2 &= \|\mathbf{U}\Sigma\mathbf{V}'\mathbf{b} - \mathbf{Y}\|^2 \\ &= \|\Sigma\mathbf{V}'\mathbf{b} - \mathbf{U}'\mathbf{Y}\|^2 \\ &= \sum_{i=1}^r [\sigma_i \alpha_i - (\mathbf{u}_i' \mathbf{Y})]^2 + \sum_{i=1}^r (\mathbf{u}_i' \mathbf{Y})^2\end{aligned}$$

is obviously minimized when $\alpha_i = \sigma_i^{-1}(\mathbf{u}_i' \mathbf{Y})$ for $i = 1, 2, \dots, r$. Moreover, since $\mathbf{b} = \mathbf{V}\boldsymbol{\alpha}$, the squared length of \mathbf{b} is then

$$\begin{aligned}\|\mathbf{b}\|^2 &= \|\mathbf{V}\boldsymbol{\alpha}\|^2 \\ &= \|\boldsymbol{\alpha}\|^2 \\ &= \sum_{i=1}^r \sigma_i^{-2} (\mathbf{u}_i' \mathbf{Y})^2 + \sum_{i=r+1}^p \alpha_i^2,\end{aligned}$$

so that if we put $\alpha_{r+1} = \dots = \alpha_p = 0$, we get the solution to the least squares problem that has smallest norm.

Since \mathbf{X} is of rank r , $\sigma_i = 0$ for $i = r+1, \dots, p$ and so by (11.79), $|\hat{\sigma}_i| < C\|\mathbf{E}\|_2$ for $i = r+1, \dots, p$. Hence at least $p-r$ of the computed singular values are very small, and the rank deficiency will be detected. This is in contrast to the QR algorithm, which cannot guarantee detection of rank deficiency.

11.10 COMPUTING THE HAT MATRIX DIAGONALS

There are two approaches to computing the hat matrix diagonals: using a Cholesky factor of $\mathbf{X}'\mathbf{X}$ or using the \mathbf{Q}_p matrix from a thin QR decomposition.

11.10.1 Using the Cholesky Factorization

Suppose that we have the factorization $\mathbf{X}'\mathbf{X} = \mathbf{R}'\mathbf{R}$, from either a Cholesky decomposition, a QR decomposition or an SVD. (In the latter case $\mathbf{R} = \Sigma\mathbf{V}'$ is a factor but not a Cholesky factor.) If \mathbf{x}_i is the i th row of \mathbf{X} , then the i th hat matrix diagonal is

$$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{x}_i'(\mathbf{R}'\mathbf{R})^{-1}\mathbf{x}_i = \|\mathbf{R}'^{-1}\mathbf{x}_i\|^2.$$

Thus, to compute h_i , we need to solve $\mathbf{R}'\mathbf{z}_i = \mathbf{x}_i$ for \mathbf{z}_i and then calculate $h_i = \|\mathbf{z}_i\|^2$. This requires about p^2 flops to solve the equation and $2p$ flops for the squared norm, so that the calculation of all the hat matrix diagonals takes about np^2 flops, ignoring terms of smaller order.

11.10.2 Using the Thin QR Decomposition

Let $\mathbf{X} = \mathbf{Q}_p\mathbf{R}$ be the thin QR decomposition of \mathbf{X} . Then the hat matrix is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{Q}_p\mathbf{R}(\mathbf{R}'\mathbf{Q}_p'\mathbf{Q}_p\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}_p' = \mathbf{Q}_p\mathbf{Q}_p',$$

so that we can compute h_i as the squared norm of the i th row of \mathbf{Q}_p . This takes $2np$ flops for all the hat matrix diagonals if \mathbf{Q}_p is available. Such is the case if the MGSA has been used for the regression calculations, but not if Householder or the SVD has been used. Computing a regression (at least the coefficients and the residual sum of squares) does not explicitly require the formation of \mathbf{Q}_p in the thin QR decomposition. If the Householder algorithm has been used, the explicit calculation of \mathbf{Q}_p takes about $4np^2 - 2p^3$ extra flops, whereas if the SVD has been used, the explicit formation of \mathbf{Q}_p (which is just \mathbf{U}_p in the thin SVD $\mathbf{X} = \mathbf{U}_p\Sigma\mathbf{V}'$) costs an extra $4np^2$ flops. Both these options are computationally quite expensive. Thus the Cholesky method is preferred unless the MGSA has been used.

The hat matrix diagonals should be calculated routinely in all regression analyses. The discussion above provides an additional argument in favor of using the MGSA in regression calculations.

11.11 CALCULATING TEST STATISTICS

We first discuss the simple case of testing that a subset of regression coefficients are zero. We can express this hypothesis as $\beta_2 = \mathbf{0}$, where $\beta' = (\beta'_1, \beta'_2)$ and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. From Section 4.3, the test statistic is

$$F = \frac{\text{RSS}_1 - \text{RSS}}{qS^2},$$

where RSS_1 and RSS are the residual sums of squares from fitting the regressions with design matrices \mathbf{X}_1 and \mathbf{X} , respectively, q is the number of

columns in \mathbf{X}_2 , and $S^2 = \text{RSS}/(n - p)$. This is most easily computed simply by fitting the two models separately. If a sequence of models is to be examined, the sweep operation can be used if the problem is not ill-conditioned, or alternatively, the QR techniques of Section 11.6.3 can be used.

To test the general linear hypothesis $\mathbf{A}\beta = \mathbf{c}$, a more complicated approach is required. The test statistic is now

$$F = \frac{\text{RSS}_H - \text{RSS}}{qS^2},$$

where $q (< p)$ is the rank of the $q \times p$ matrix \mathbf{A} , and RSS_H is the minimum value of $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$, subject to the constraint $\mathbf{Ab} = \mathbf{c}$. This form of constrained least squares is a standard problem with a large literature; see for example, Lawson and Hanson [1995: Chapters 20–22], and Björck [1996: p. 187]. These authors discuss several methods for solving this minimization problem, including the method of direct elimination, the null space method, and the method of weighting. All of these methods transform the constrained problem into an unconstrained least squares problem, which is then solved in the usual way. However, a more efficient method due to Golub and Styan [1974] makes use of the representation

$$\text{RSS}_H - \text{RSS} = (\mathbf{A}\hat{\beta} - \mathbf{c})' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c})$$

and does not require fitting another regression. Assuming that the full regression has been fitted, so that a Cholesky factor of $\mathbf{X}'\mathbf{X}$ has been calculated, the algorithm is as follows.

Algorithm 11.8

Step 1: Retrieve the Cholesky factor \mathbf{R} of $\mathbf{X}'\mathbf{X}$ (i.e., $\mathbf{X}'\mathbf{X} = \mathbf{R}'\mathbf{R}$). This will have been formed no matter how the regression was fitted.

Step 2: Put $\mathbf{G} = (\mathbf{A}\mathbf{R}^{-1})'$, so that

$$\begin{aligned} \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' &= \mathbf{A}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{A}' \\ &= \mathbf{A}\mathbf{R}^{-1}(\mathbf{A}\mathbf{R}^{-1})' \\ &= \mathbf{G}'\mathbf{G}. \end{aligned}$$

Note that \mathbf{G} can be computed cheaply by solving $\mathbf{R}'\mathbf{G} = \mathbf{A}'$ for \mathbf{G} by back-substitution.

Step 3: Calculate the thin QR decomposition $\mathbf{G} = \mathbf{Q}_q \mathbf{T}$ of the $p \times q$ matrix \mathbf{G} , where \mathbf{T} is upper triangular (i.e., $\mathbf{G}'\mathbf{G} = \mathbf{T}'\mathbf{T}$).

Step 4: Put $\mathbf{g} = \mathbf{A}\hat{\beta} - \mathbf{c}$. Then

$$\begin{aligned} \text{RSS}_H - \text{RSS} &= (\mathbf{A}\hat{\beta} - \mathbf{c})' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c}) \\ &= \mathbf{g}'\mathbf{T}^{-1}(\mathbf{T}')^{-1}\mathbf{g} \\ &= \mathbf{h}'\mathbf{h}, \end{aligned}$$

say, where $\mathbf{T}'\mathbf{h} = \mathbf{g}$. The vector \mathbf{h} is obtained by back-substitution, and the numerator of the test statistic is computed as its squared length.

If the constrained estimate $\hat{\beta}_H$ that minimizes $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ subject to $\mathbf{Ab} = \mathbf{c}$ is required, it can be calculated using (3.38). We get

$$\begin{aligned} \hat{\beta}_H &= \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}' [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1} (\mathbf{A}\hat{\beta} - \mathbf{c}) \\ &= \hat{\beta} - \mathbf{R}^{-1}\mathbf{R}^{-1'}\mathbf{A}'\mathbf{T}^{-1}\mathbf{h} \\ &= \hat{\beta} - \mathbf{R}^{-1}\mathbf{G}\mathbf{T}^{-1}\mathbf{h} \\ &= \hat{\beta} - \mathbf{R}^{-1}\mathbf{Q}_q\mathbf{h}. \end{aligned}$$

11.12 ROBUST REGRESSION CALCULATIONS

In this section we discuss some of the algorithms used to fit linear regression models using robust methods.

11.12.1 Algorithms for L_1 Regression

We recall from Section 3.13 that the L_1 estimate of β in the regression model

$$Y_i = \mathbf{x}'_i\beta + \varepsilon_i$$

is the vector \mathbf{b} , which minimizes

$$\sum_{i=1}^n |e_i(\mathbf{b})|, \quad (11.85)$$

where $e_i(\mathbf{b}) = Y_i - \mathbf{x}'_i\mathbf{b}$. The minimization problem (11.85) is usually solved using linear programming (LP) techniques. Consider the LP problem

$$\min \left(\sum_{i=1}^n u_i + \sum_{i=1}^n v_i \right) \quad (11.86)$$

subject to the constraints

$$u_i - v_i = y_i - \sum_{j=0}^{p-1} x_{ij}c_j + \sum_{j=0}^{p-1} x_{ij}d_j \quad (i = 1, \dots, n), \quad (11.87)$$

and

$$u_i \geq 0, v_i \geq 0 \quad (i = 1, \dots, n); \quad c_j \geq 0, d_j \geq 0 \quad (j = 0, \dots, p - 1). \quad (11.88)$$

This is a standard LP problem and can be solved by standard LP algorithms, such as the simplex method (cf. Fletcher [1987: Chapter 8] for a full discussion).

Next, we prove that any solution to the LP problem (11.86) minimizes the L_1 criterion (11.85). We first show that without loss of generality, we can assume that $u_i v_i = 0$. Suppose that u_i, v_i, c_j , and d_j are a solution to the LP problem. Let $u'_i = u_i - \min(u_i, v_i)$ and $v'_i = v_i - \min(u_i, v_i)$. Clearly, u'_i, v'_i, c_j, d_j satisfy the constraints (11.87) and (11.88), and

$$u'_i + v'_i \leq u_i + v_i,$$

so that

$$\sum_{i=1}^n u'_i + \sum_{i=1}^n v'_i \leq \sum_{i=1}^n u_i + \sum_{i=1}^n v_i. \quad (11.89)$$

Since u_i, v_i, c_j, d_j is a solution, the inequality (11.89) is in fact an equality, so that u'_i, v'_i, c_j, d_j also minimize (11.86) and satisfy $u'_i v'_i = 0$.

Now suppose that u_i, v_i, c_j, d_j are a solution of the LP with $u_i v_i = 0$. Let $\mathbf{b} = (b_0, \dots, b_{p-1})'$, where $b_j = c_j - d_j$. We will show that \mathbf{b} minimizes (11.86). Let \mathbf{b}^* be any other p -vector, $e_i^* = y_i - \mathbf{x}'_i \mathbf{b}^*$, $u_i^* = \max(e_i^*, 0)$, $v_i^* = \max(-e_i^*, 0)$, $c_j^* = \max(b_j^*, 0)$, and $d_j^* = \max(-b_j^*, 0)$. Then $u_i^*, v_i^*, c_j^*, d_j^*$ satisfy the constraints (11.87) and (11.88), and $u_i^* + v_i^* = |e_i^*|$. Also, since $u_i v_i = 0$, we have

$$|y_i - \mathbf{x}'_i \mathbf{b}| = |u_i - v_i| = u_i + v_i.$$

Hence, since u_i, v_i solve the LP,

$$\begin{aligned} \sum_{i=1}^n |y_i - \mathbf{x}'_i \mathbf{b}| &= \sum_{i=1}^n |u_i - v_i| \\ &= \sum_{i=1}^n u_i + \sum_{i=1}^n v_i \\ &\leq \sum_{i=1}^n u_i^* + \sum_{i=1}^n v_i^* \\ &= \sum_{i=1}^n |e_i^*| \\ &= \sum_{i=1}^n |y_i - \mathbf{x}'_i \mathbf{b}^*|, \end{aligned}$$

so that \mathbf{b} minimizes (11.86).

In practice, the particular form of the LP (11.86) –(11.88) allows the standard simplex method to be modified to achieve greater computational efficiency. Barrodale and Roberts [1974], Bartels et al. [1978] and Bloomfield and Steiger [1980] all describe algorithms that are modifications of the simplex method. Bloomfield and Steiger [1983] make comparisons between the methods and recommend the Bloomfield–Steiger algorithm.

11.12.2 Algorithms for M- and GM-Estimation

We can calculate M- and GM-estimates by a suitable sequence of weighted least squares fits. We must solve the estimating equations

$$\sum_{i=1}^n w_i(\mathbf{b})\psi\{e_i(\mathbf{b})/[s w_i(\mathbf{b})]\}\mathbf{x}_i = 0, \quad (11.90)$$

$$\sum_{i=1}^n \chi[e_i(\mathbf{b})/s] = 0. \quad (11.91)$$

We have written the weights as $w_i(\mathbf{b})$ to emphasise their dependence on the current fit, but in practice they will be functions of the current residuals and the regression matrix \mathbf{X} . For M-estimates, we set $w_i = 1$.

The algorithm is based on rewriting (11.90) by putting

$$W_i = w_i(\mathbf{b})\psi\{e_i(\mathbf{b})/[s w_i(\mathbf{b})]\}/e_i(\mathbf{b}).$$

Then (11.90) becomes

$$\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{W}\mathbf{Y}, \quad (11.92)$$

where $\mathbf{W} = \text{diag}(W_1, \dots, W_n)$. If the weights were known, we could solve (11.92) using the weighted least squares algorithm described in Section 11.5. We can deal with the unknown weights by an iterative approach. Suppose that we have a preliminary estimate of β , which could be an L_1 estimate. We can then calculate weights on the basis of this preliminary fit, solve (11.92) for an updated estimate, and iterate to convergence. Specifically, the algorithm is as follows.

Algorithm 11.9

Step 1: Obtain a preliminary estimate $\hat{\beta}^{(0)}$. Set $m = 0$.

Step 2 : Solve

$$\sum_{i=1}^n \chi[e_i(\hat{\beta}^{(m)})/s] = 0$$

for s , obtaining an estimate $\hat{\sigma}^{(m)}$ of σ .

Step 3: Put

$$W_i = w_i(\hat{\beta}^{(m)}) \psi[e_i(\hat{\beta}^{(m)}) / (\hat{\sigma}^{(m)} w_i(\hat{\beta}^{(m)}))] / e_i(\hat{\beta}^{(m)})$$

and $W^{(m)} = \text{diag}(W_1, \dots, W_n)$.

Step 4: Solve

$$\mathbf{X}' \mathbf{W}^{(m)} \mathbf{X} \mathbf{b} = \mathbf{X}' \mathbf{W}^{(m)} \mathbf{Y}$$

to obtain an updated estimate $\hat{\beta}^{(m+1)}$.

Step 5: Repeat steps 2–4 until convergence.

11.12.3 Elemental Regressions

Let J be any subset of p observations, and let \mathbf{X}_J and \mathbf{Y}_J be the corresponding submatrices of \mathbf{X} and \mathbf{Y} . An *elemental regression* (cf. Mayo and Gray [1997]) is a regression fit using just \mathbf{X}_J and \mathbf{Y}_J . Assuming that each \mathbf{X}_J has full rank, the estimated regression coefficients are $\hat{\beta}_J = \mathbf{X}_J^{-1}' \mathbf{Y}_J$ and the fit is exact.

Elemental regressions are used to obtain approximate solutions in many kinds of robust regression problems. In principle, the solution is simple: All possible $\binom{n}{p}$ elemental regressions are computed, and the one that optimizes the relevant criterion is taken as an approximate solution. Of course, in general, there is no guarantee that the true optimal solution corresponds to an elemental regression. Also, the calculation of all $\binom{n}{p}$ regressions quickly becomes computationally prohibitive as n and p increase.

It transpires that there is a solution to the L_1 problem (11.85) that is an elemental regression (cf. Bloomfield and Steiger [1983: p. 7]), although searching through all the elemental regressions is not a practical algorithm. There is also a close connection between least squares and elemental regressions, as the usual LSE can be written as a weighted linear combination of the solutions $\hat{\beta}_J$ of the elemental regressions. We have

$$\hat{\beta} = \sum_J w_J \hat{\beta}_J,$$

where $w_J = \det(\mathbf{X}_J' \mathbf{X}_J) / \det(\mathbf{X}' \mathbf{X})$. A proof may be found in Hawkins et al. [1984].

11.12.4 Algorithms for High-Breakdown Methods

Algorithms for high-breakdown methods such as least median squares and least trimmed squares also make use of elemental regressions. The original

algorithm of Rousseeuw for LMS (Rousseeuw [1984]) was based on approximating the solution using elemental regressions, as described above. In its simplest form, the algorithm computes all $\binom{n}{p}$ elemental regressions and chooses the one that minimizes the LMS criterion. Note that unlike L_1 regression, there is no guarantee that this will happen for any elemental regression. However, Hawkins [1993a] argues that this procedure will produce a minimum close to the true minimum, particularly when combined with *intercept tuning* (Hawkins [1993a]).

Computing all elemental regressions is not practical when n and p are large. An alternative is to sample elemental regressions at random. However, Portnoy [1987] cautions against this unless the number of subsets sampled, J , is very large, since insufficient sampling will produce “solutions” that are far from optimal.

The inadequacies of the elemental set method have motivated the search for exact algorithms. Stromberg [1993] presents a method that will produce an exact solution that is practical for small problems. Consider the modified LMS criterion

$$\min_{\mathbf{b}} e_{(h)}(\mathbf{b})^2 \quad (11.93)$$

discussed in Section 3.13.2. It is clear that if \mathbf{b} minimizes (11.93), it will minimize the maximum squared residual for some h -subset of the data. The fit minimizing the maximum squared residual of a set of regression data is called the *Chebyshev fit*. As noted by Stromberg [1993], a key property of the Chebyshev fit to a set of data containing n cases is that it is also the Chebyshev fit to some $(p+1)$ -subset of the cases. It follows that by calculating the Chebyshev fit to every $(p+1)$ -subset of the original set of n cases, we must eventually calculate the solution to the LMS problem.

The Chebyshev fit to a $(p+1)$ -subset can be simply found. Cheney [1966] proves that the fit is given by

$$\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}} - \kappa \mathbf{C} \mathbf{s},$$

where $\hat{\boldsymbol{\beta}}$ is the LSE of the regression using the $p+1$ cases, \mathbf{C} is the catcher matrix (cf. Section 10.6.3) for the $p+1$ cases,

$$\kappa = \frac{\sum_{i=1}^{p+1} e_i(\hat{\boldsymbol{\beta}})^2}{\sum_{i=1}^{p+1} |e_i(\hat{\boldsymbol{\beta}})|},$$

and $\mathbf{s} = (\text{sign}[e_1(\hat{\boldsymbol{\beta}})], \dots, \text{sign}[e_{p+1}(\hat{\boldsymbol{\beta}})])'$. Note that all regression quantities in these formulas refer to the regression using a $(p+1)$ -subset of cases, not the full regression.

The Stromberg algorithm will calculate an exact solution to the LMS problem, but again is practical only for small problems. What is required is an algorithm that is more efficient than random selection of elemental sets but is also practical for medium-sized problems. A variety of algorithms have been

developed to improve on random searching. Most use a randomly chosen subset of cases as a starting point, which is then refined to produce a better solution. The algorithms then repeatedly resample and refine subsets, using the best subset found after a fixed number of iterations to calculate the final estimate. The refining step makes these algorithms much more efficient than pure resampling.

The refining step depends on the estimate being calculated. For example, Hawkins [1994a] proposed an algorithm for LTS regression based on repeated resampling of h -subsets of cases. Recall that the LTS estimate minimizes the criterion

$$\sum_{i=1}^h e_{(i)}(\mathbf{b})^2,$$

where $h = [n/2] + [(p+1)/2]$. It follows that the LTS estimate is the ordinary least squares estimate calculated from some h -subset of the n cases. The Hawkins algorithm selects h -subsets at random. The subset chosen is then refined by a series of pairwise swaps, where a case in the subset is replaced by a case not in the subset. The resulting change in the residual sum of squares can be calculated easily; a formula is given in Exercises 11d, No. 1. All possible swaps are done and the swap making the biggest reduction in the RSS is recorded. The RSS of this subset is compared to the current best value of the criterion and becomes the current best value if it is smaller. This entire procedure is then repeated a fixed number of times and the best solution found is the approximate LTS estimate.

A similar algorithm has been proposed for the LMS estimate (Hawkins [1993b]), using a refinement step based on the simplex method in linear programming. The combination of a random choice of subset, followed by a refinement step, can also be used for the robust methods for estimating location and covariance for multivariate point clouds when attempting to identify high-leverage points as described in Section 10.6.2. These algorithms are called *feasible solution algorithms* since the refined subsets satisfy certain necessary conditions to be a minimum. More details may be found in Hawkins and Olive [1999]. Hawkins [1994b] describes an algorithm for the minimum covariance determinant estimate (cf. Section 10.6.2).

Ruppert [1992] proposed a similar approach for calculating S-estimates, but used refinement of the estimated coefficients rather than refining subsets of cases. In Ruppert's method an elemental set of p cases is chosen, and the exact regression calculated, resulting in an estimate $\hat{\beta}$, say. Suppose that $\hat{\beta}$ is the current best estimate, which has resulted in the current smallest value of s [cf. (3.95)]. Then Ruppert suggests examining a series of trial values of the form $t\hat{\beta} + (1-t)\hat{\beta}$ for a set of t 's equally spaced between 0 and 1. The estimate making the greatest reduction in the value of s is retained as the current best estimate, and the process is repeated. The algorithm can be made more efficient by a further refinement step once the current best value is found, and also by a clever trick that avoids having to calculate the value of

s for all the trial values. This algorithm can also be used for LMS and LTS, plus the robust covariance estimates used in outlier detection. More details can be found in Ruppert [1992].

The idea of *concentration* gives another way of refining estimates derived from randomly chosen subsets. In this method, a set of c cases is chosen at random. A regression fit using these cases is performed, and the cases corresponding to the c smallest residuals are identified. A further fit to these cases is performed, and the process is iterated until there is no change. The fit could be an L_1 , least squares or Chebyshev fit. Further discussion of concentration can be found in Hawkins and Olive [2002], and Rousseeuw and van Driessen [1999] describe a concentration algorithm for the MCD estimate.

Finally, we note that several of these methods rely on fits to subsets of cases. These may be difficult or impossible if the submatrices concerned have less than full rank. This is not usually a problem if all the variables in the regression are quantitative. However, if some of the columns of the regression matrix are indicator variables, as in analysis-of-covariance models, then this may not be the case. Hubert and Rousseeuw [1997] give a method applicable in this case. They use the minimum volume ellipsoid to identify the high-leverage points and then fit a weighted L_1 regression to all the data, using weights based on the robust Mahalanobis distance discussed in Section 10.6.2 which downweight the high-influence points. This is essentially the same as using a GM-estimate.

EXERCISES 11d

- Suppose that we fit a least squares regression to a subset J of the n available cases. Let \mathbf{X}_J be the submatrix of \mathbf{X} corresponding to the cases in J , let e_i be the i th residual from this fit, $i = 1, \dots, n$, and let $h_{ij} = \mathbf{x}'_i(\mathbf{X}'_J\mathbf{X}_J)^{-1}\mathbf{x}_j$. Show that if we drop case $i \in J$ and add case $j \notin J$, the change in the residual sum of squares is

$$\frac{e_j^2(1 - h_{ii}) - e_i^2(1 + h_{jj}) + 2e_i e_j h_{ij}}{(1 - h_{ii})(1 + h_{jj}) + h_{ij}^2}.$$

(Atkinson and Weisberg [1991])

- Using the result in Exercise 1 above, write an R function implementing Hawkins's method of calculating an approximation to the LTS estimate.
- Write an R function to implement Ruppert's algorithm to calculate an approximation to the LTS estimate. You will need to refer to Ruppert [1992] for the details of the algorithm. Devise a small simulation to compare Ruppert's algorithm to Hawkins's.

MISCELLANEOUS EXERCISES 11

1. Write a program to implement the Gram–Schmidt algorithm (Algorithm 11.3 in Section 11.3). Try out your program on the Longley data and the polynomial approximations described in Example 11.3 in Section 11.8.4. Compare your results with those in Examples 11.2 and 11.3.
2. Verify the formulas for the updating algorithm in Section 11.7.
3. Try some experiments to compare the execution time of a regression performed using the Cholesky method with one using the SVD. Do the flop counts given in the text give a reasonable indication of the relative speed of the two methods?

12

Prediction and Model Selection

12.1 INTRODUCTION

The aim of regression analysis is to discover the relationships, if any, between the response Y and the explanatory variables x_1, \dots, x_{p-1} . Ultimately, we may wish to use these relationships to make predictions about Y based on observing x_1, \dots, x_{p-1} . Alternatively, we may use the fitted model to better understand how particular variables are related to the response. In both cases, we need to choose a tentative model to fit to the data. Having fitted the model, we need to assess how well the model fits, and if necessary, modify the model to improve the fit, using the methods of Chapter 10. We can then use the fitted model to make predictions.

An important part of this process is to decide which variables to include in the model and how to construct a predictor from the variables available. In the first part of this book we assumed that the set of variables was given. In this chapter we assume that a large number K of potential explanatory variables are available. We shall discuss ways of selecting which variables to include in order to get a good model, and how to use the available variables to construct a good predictor.

If the main aim of the analysis is to make predictions, we might feel that nothing is lost by using every scrap of information available, and that we should use all the variables x_1, \dots, x_K in the construction of our predictor. However, as we show in Section 12.2, we can often do much better by discarding a proportion of the variables, particularly if K is large compared to the number of cases n . This apparently paradoxical result means that, in practice, we need to select which variables to use. Alternatively, we may abandon the

use of least squares and seek other methods of estimation that have better predictive accuracy.

The literature on model selection and prediction sometimes makes the assumption that the true model is one of the form

$$Y = \beta_0 + x_1\beta_1 + \cdots + x_K\beta_K + \varepsilon. \quad (12.1)$$

In this case a key issue is identifying the variables that are not related to the response, that is, identifying the β 's that are zero. This is model selection in its pure form. Alternatively, and more realistically, it may be felt that truth cannot be so simple, and the task is then to build a model of the form (12.1) having the best possible predictive accuracy, using the available explanatory variables as raw material. In either case, we have several options: either using least squares, possibly discarding some variables, or using alternative methods of estimation such as ridge regression.

If we want to select a subset of variables, there are two main approaches. In the first, known as *all possible regressions* (APR), we define a criterion of *model goodness*, evaluate the criterion for each possible subset of variables, and then choose the subset that optimizes the criterion. Criteria can be based on standard goodness-of-fit measures, on estimating the prediction error, on estimating the number of nonzero coefficients, or on estimating some measure of distance between the model based on the subset and the true model. These criteria, some of which make no sense if the true model is not of the form (12.1), are discussed in detail in Section 12.3. Evaluating the criteria for all possible subsets can be computationally intensive, even prohibitive, if the number of explanatory variables is large.

The second approach is to apply a sequence of hypothesis tests to the problem and attempt to identify the nonzero β 's in (12.1). These techniques, of which forward selection, backward elimination, and stepwise regression are the best known examples, obviously make the assumption that (12.1) is the true model. They are computationally much less demanding, but there is no guarantee that the models found will be optimal in terms of any of the criteria discussed in Section 12.3. These sequential testing techniques are described in Section 12.4.

An alternative to subset selection and least squares is to fit the model using all the explanatory variables, but to use a biased estimation method such as ridge regression that “shrinks” the coefficients toward zero. Some of these methods, in particular the suggestively named *garrote* and *lasso* zero some coefficients as well. Shrinkage methods are the subject of Section 12.5.

Bayesian methods of subset selection offer a different approach to the problem that has an attractive conceptual simplicity. These methods provide ways of selecting single models or, in the case of prediction, combining predictions from several models. This idea of *model averaging* can also be implemented in a non-Bayesian context, and these ideas are pursued in Section 12.6.

Very often, the same data are used to both select the model to estimate the coefficients of the chosen model and to make predictions. The standard

methods of inference covered in earlier chapters do not apply when the model has been selected using the same data as those used for model fitting. They tend to overestimate the precision of estimates and the accuracy of predictions. In Section 12.7 we look briefly at ways that the standard inferences can be modified to incorporate the uncertainty induced by the model selection phase of the analysis.

For a model with K possible explanatory variables, the APR methods for model selection mentioned above require the fitting of 2^K possible regression models. Efficient ways of doing this, and of avoiding having to examine every subset, are discussed in Section 12.8.

This chapter describes a large number of possibilities for subset selection and the construction of predictors. In the final section we compare and contrast these methods.

12.2 WHY SELECT?

We begin by considering a simple example that illustrates how discarding explanatory variables can improve prediction. The example is adapted from Linhart and Zucchini [1986: Section 6.3] and concerns the prediction of a simple exponential function.

Suppose that we have a true model of the form

$$Y_i = \exp(ax_i^2 + b) + \varepsilon_i, \quad (12.2)$$

where the ε_i are independent normal errors with mean zero and variance σ^2 . The model is not known to the data analyst, who decides to fit a polynomial of degree seven to the available data, which consists of $n = 100$ pairs (x_i, y_i) , $i = 1, 2, \dots, n$, with the x_i equally spaced from 1 to 10. The true values of a , b , and σ are $a = 0.02$, $b = 0.1$, and $\sigma = 1.5$.

Suppose that the scientific objective is to predict the value of the function at $x = 5$. In theory, it should be possible to make a good prediction, because although the functional form chosen by the analyst is not correct, the true function

$$f(x) = \exp(ax^2 + b)$$

can be approximated very well by a seven-degree polynomial. The function f is plotted in Figure 12.1(a). Also shown in Figure 12.1(b) is the difference between the function and the best-fitting seven-degree polynomial. We see that the difference is very small.

What the data analyst actually observes is a noisy version of f , described by the model (12.2). The data are plotted in Figure 12.1(c), together with the polynomial that best fits the actual data. Note that this is not the same polynomial as the one plotted in Figure 12.1(a). The polynomial fitted in Figure 12.1(c) appears to fit the current data reasonably well.

How well does this seventh-degreee polynomial model predict *future* observations? We repeatedly generated sets of 100 further observations from the

exponential model and used the polynomial model to predict $f(5)$. We also repeated the predictions using polynomials of smaller degree, so that for each simulated data set we have seven predictions, corresponding to using models of degree $1, 2, \dots, 7$. The results are shown in Figure 12.1(d) in the form of box plots, one for each of the seven models. The horizontal bar in the middle of each box plot represents the average value of the prediction, so the distance between the bar and the horizontal line [which represents the true value of $f(5)$, namely 1.822] represents the bias in the prediction. The variability of the predictions is represented by the heights of the boxes. We see that as the degree of the fitted polynomial increases, the bias decreases but the variance increases. The predictor based on a linear approximation (i.e., a polynomial of degree 1) is stable but has a large bias, since it cannot capture the curvature of the true function f . On the other hand, the seventh-degree polynomial has negligible bias, but is highly variable. This is because the flexibility of the functional form being used to fit the model allows the model to track the noise in the data too closely, resulting in unstable predictions. In fact, if we calculate the mean-squared error of the predictions, we find that the best trade-off between bias and variance occurs when we use a cubic polynomial. Thus the optimal prediction does not use all seven possible power terms, but only three.

The idea developed in this example — that basing a prediction on too few or too many variables can lead to poor predictions — can also be illustrated theoretically. We explore this in the next section.

Prediction Error and Model Error

In the prediction problem, we have an initial data set (\mathbf{x}_i, Y_i) , $i = 1, \dots, n$, consisting of n $(p + 1)$ -dimensional multivariate observations, each consisting of a response Y_i and a p -dimensional vector of explanatory variables \mathbf{x}_i . We will assume that the initial element of each \mathbf{x}_i is 1, corresponding to the constant term in the regression model.

Suppose that we want to use this data set, often referred to as the *training set*, to predict the responses Y_{0i} , $i = 1, \dots, m$, corresponding to m new vectors \mathbf{x}_{0i} . A related problem is to estimate the mean μ_{0i} of the response Y_{0i} .

Let $\mathbf{Y}' = (Y_1, \dots, Y_n)$ and $\mathbf{Y}'_0 = (Y_{01}, \dots, Y_{0m})$. We assume that \mathbf{Y} and \mathbf{Y}_0 have covariance matrices $\sigma^2 \mathbf{I}_n$ and $\sigma^2 \mathbf{I}_m$, respectively, and that \mathbf{Y} and \mathbf{Y}_0 are independent with the same probability structure; if $\mathbf{x}_i = \mathbf{x}_{j0}$, then $E[Y_i] = E[Y_{0j}]$. Also, let

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \quad \text{and} \quad \mathbf{X}_0 = \begin{bmatrix} \mathbf{x}'_{01} \\ \vdots \\ \mathbf{x}'_{0m} \end{bmatrix}.$$

Suppose that we calculate the least squares estimate $\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$ from the training set. The least squares predictor of \mathbf{Y}_0 (and also the estimate of

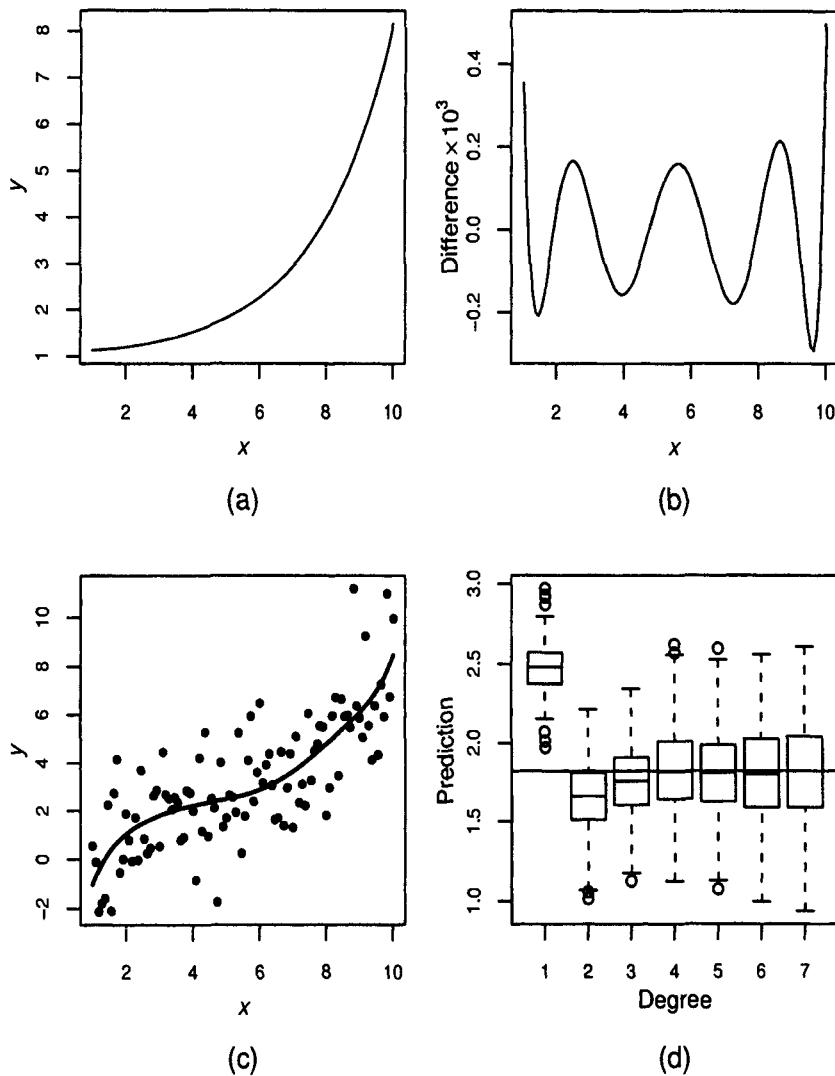


Fig. 12.1 Performance of predictors for several models: (a) function being predicted; (b) error in approximating the function by a polynomial; (c) typical data and the fitted polynomial; (d) box plots of the prediction errors.

$\mu_0 = E[Y_0]$) is $\mathbf{X}_0\hat{\beta}$. The sum of the squared prediction errors is

$$\sum_{i=1}^m (Y_{0i} - \mathbf{x}'_{0i}\hat{\beta})^2 = \|\mathbf{Y}_0 - \mathbf{X}_0\hat{\beta}\|^2, \quad (12.3)$$

which we can write as

$$\begin{aligned} \|\mathbf{Y}_0 - \mu_0 + \mu_0 - \mathbf{X}_0\hat{\beta}\|^2 &= \|\mathbf{Y}_0 - \mu_0\|^2 + \|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2 \\ &\quad + 2(\mathbf{Y}_0 - \mu_0)'(\mu_0 - \mathbf{X}_0\hat{\beta}). \end{aligned} \quad (12.4)$$

If we take expectations over the new data only, we get a quantity which we call the *prediction error* (PE), given by

$$\begin{aligned} \text{PE} &= E_{\mathbf{Y}_0} [\|\mathbf{Y}_0 - \mathbf{X}_0\hat{\beta}\|^2] \\ &= E_{\mathbf{Y}_0} [\|\mathbf{Y}_0 - \mu_0\|^2] + \|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2 \\ &= E_{\mathbf{Y}_0} \left[\sum_{i=1}^m (Y_{i0} - \mu_{i0})^2 \right] + \|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2 \\ &= m\sigma^2 + \|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2. \end{aligned} \quad (12.5)$$

In (12.5), the cross-product term vanishes because \mathbf{Y} and \mathbf{Y}_0 are independent and

$$\begin{aligned} E_{\mathbf{Y}_0}[(\mathbf{Y}_0 - \mu_0)'(\mu_0 - \mathbf{X}_0\hat{\beta})] &= (E_{\mathbf{Y}_0}[\mathbf{Y}_0] - \mu_0)'(\mu_0 - \mathbf{X}_0\hat{\beta}) \\ &= (\mu_0 - \mu_0)'(\mu_0 - \mathbf{X}_0\hat{\beta}) \\ &= 0. \end{aligned}$$

Equation (12.5) expresses the PE as the sum of a quantity $m\sigma^2$ which reflects the underlying variability of the data, and a term $\|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2$ which measures how well the linear model represented by \mathbf{X}_0 estimates the mean μ_0 of the new responses \mathbf{Y}_0 . This second term, which we will call the *model error* (ME), is the crucial quantity for measuring how well the model predicts the new responses (or, equivalently, how well the model represented by \mathbf{X}_0 estimates the mean response μ_0). We have

$$\text{PE} = m\sigma^2 + \text{ME}, \quad (12.6)$$

where

$$\text{ME} = \|\mu_0 - \mathbf{X}_0\hat{\beta}\|^2. \quad (12.7)$$

The ME depends on the regression matrix \mathbf{X}_0 .

We can get a simple formula for the ME if we take $\mathbf{X}_0 = \mathbf{X}$, which implies that $m = n$ and $\mu_0 = E[\mathbf{Y}] = \mu$, say, since we are assuming that the probability structures of the new and old data are the same. In this case, putting

$\epsilon = \mathbf{Y} - \mu$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we get

$$\begin{aligned} \text{ME} &= \|\mu - \mathbf{X}\hat{\beta}\|^2 \\ &= \|\mu - \mathbf{PY}\|^2 \\ &= \|\mu - \mathbf{P}(\mu + \epsilon)\|^2 \\ &= \|(\mathbf{I}_n - \mathbf{P})\mu - \mathbf{P}\epsilon\|^2 \\ &= \|(\mathbf{I}_n - \mathbf{P})\mu\|^2 + \|\mathbf{P}\epsilon\|^2 \\ &= \mu'(\mathbf{I}_n - \mathbf{P})\mu + \epsilon'\mathbf{P}\epsilon, \end{aligned}$$

since the cross-product term again vanishes because $\mathbf{P}^2 = \mathbf{P}$. The *expected model error* $E[\text{ME}]$ is

$$\begin{aligned} E[\text{ME}] &= \mu'(\mathbf{I}_n - \mathbf{P})\mu + E[\epsilon'\mathbf{P}\epsilon] \\ &= \mu'(\mathbf{I}_n - \mathbf{P})\mu + \sigma^2 \text{tr}(\mathbf{P}) \quad [\text{by (1.12)}] \\ &= \mu'(\mathbf{I}_n - \mathbf{P})\mu + \sigma^2 p. \end{aligned} \tag{12.8}$$

The corresponding formula for the expected PE when $\mathbf{X}_0 = \mathbf{X}$ is

$$\begin{aligned} E[\text{PE}] &= E[n\sigma^2 + \text{ME}] \\ &= (n + p)\sigma^2 + \mu'(\mathbf{I}_n - \mathbf{P})\mu, \end{aligned} \tag{12.9}$$

using (12.6) and (12.8). Now define the *total bias* and *total variance* of the predictor $\mathbf{X}\hat{\beta}$ by

$$\text{TOTAL BIAS} = \|\mu - E[\mathbf{X}\hat{\beta}]\|$$

and

$$\text{TOTAL VARIANCE} = \text{tr}(\text{Var}[\mathbf{X}\hat{\beta}]) = \sigma^2 \text{tr}(\mathbf{P}).$$

Then (see Exercises 12a, No. 1), the first term of (12.8) is just the square of the total bias, which vanishes if μ is in $C(\mathbf{X})$. Also, the second term of (12.8) is just the total variance. Thus (12.8) expresses the expected ME as the sum of a term measuring bias and a term measuring the variability of the predictor. Clearly, as new variables are added to the model, the variability increases but the bias decreases, unless the new variables are linearly dependent on the old. The best model, having smallest expected ME (and hence the smallest expected PE), will be some compromise between these conflicting requirements of small bias and low variability.

In the discussion above, the linear model (12.1) was only an approximation to the true model. Now we consider what happens if (12.1) is exactly the true model, although perhaps having some redundant variables with regression coefficients equal to zero. Specifically, suppose that we write the model in matrix form as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $\beta' = (\beta'_1, \beta'_2)$. Assume that \mathbf{X}_1 has p columns (corresponding to a submodel with $p - 1$ variables) and \mathbf{X}_2 has $K - p + 1$

columns (corresponding to the remaining variables). Also assume that β_1 and β_2 have dimensions p and $K - p + 1$, respectively.

We saw in Section 9.2.2 that including the redundant variables does have a cost; the predictor using all the explanatory variables had a larger variance than that based on \mathbf{X}_1 . On the other hand, if $\beta_2 \neq \mathbf{0}$, and we use only the subset, the predictor is biased. This raises the interesting question: Even if $\beta_2 \neq \mathbf{0}$, do we sometimes do better using the predictor based on the smaller, incorrect model? Under certain circumstances the answer is yes, as we now show. Suppose that we want to predict $\mathbf{x}'\beta$, where $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$ is a $(K+1)$ -vector. Should we use the biased predictor $\mathbf{x}'_1\tilde{\beta}_1$ based only on \mathbf{X}_1 , or the unbiased predictor $\mathbf{x}'\hat{\beta}$, where $\hat{\beta}$ is the least squares estimate based on \mathbf{X} using all the explanatory variables?

From (12.5), with $m = 1$, the expected PE of the biased predictor is

$$\begin{aligned} E[\text{PE}] &= \sigma^2 + E[(\mathbf{x}'_1\tilde{\beta}_1 - \mathbf{x}'\beta)^2] \\ &= \sigma^2 + \text{var}[\mathbf{x}'_1\tilde{\beta}_1] + (\mathbf{x}'_1 E[\tilde{\beta}_1] - \mathbf{x}'\beta)^2. \end{aligned}$$

Using the results of Section 9.2, we have

$$E[\tilde{\beta}_1] = \beta_1 + \mathbf{L}\beta_2,$$

where $\mathbf{L} = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$. Thus

$$\begin{aligned} \mathbf{x}'_1 E[\tilde{\beta}_1] - \mathbf{x}'\beta &= \mathbf{x}'_1\beta_1 + \mathbf{x}'_1\mathbf{L}\beta_2 - \mathbf{x}'_1\beta_1 - \mathbf{x}'_2\beta_2 \\ &= (\mathbf{L}'\mathbf{x}_1 - \mathbf{x}_2)' \beta_2, \end{aligned}$$

and the expected PE of the biased estimate is

$$E[\text{PE}] = \sigma^2 + \text{var}[\mathbf{x}'_1\tilde{\beta}_1] + [(\mathbf{L}'\mathbf{x}_1 - \mathbf{x}_2)' \beta_2]^2. \quad (12.10)$$

The expected PE of the unbiased predictor $\mathbf{x}'\hat{\beta}$ is

$$\begin{aligned} E[\text{PE}] &= \sigma^2 + E[(\mathbf{x}'\hat{\beta} - \mathbf{x}'\beta)^2] \\ &= \sigma^2 + \text{var}[\mathbf{x}'\hat{\beta}] \\ &= \sigma^2 + \text{var}[\mathbf{x}'_1\tilde{\beta}_1] + \sigma^2(\mathbf{L}'\mathbf{x}_1 - \mathbf{x}_2)' \mathbf{M}(\mathbf{L}'\mathbf{x}_1 - \mathbf{x}_2), \quad (12.11) \end{aligned}$$

by (9.5), where $\mathbf{M} = [\mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2]^{-1}$ and $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$. Let $\mathbf{h} = \mathbf{L}'\mathbf{x}_1 - \mathbf{x}_2$. Comparing (12.10) and (12.11), we see that the biased predictor has a smaller expected PE than the predictor based on all the variables if

$$(\mathbf{h}'\beta_2)^2 < \sigma^2 \mathbf{h}' \mathbf{M} \mathbf{h}. \quad (12.12)$$

Now \mathbf{M} is positive definite by the Lemma of Section 3.7, so by A.4.11,

$$(\mathbf{h}'\beta_2)^2 \leq \mathbf{h}' \mathbf{M} \mathbf{h} \cdot \beta'_2 \mathbf{M}^{-1} \beta_2$$

for all vectors \mathbf{h} . Thus, if $\beta'_2 \mathbf{M}^{-1} \beta_2 = \beta'_2 \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 \beta_2 < \sigma^2$, then (12.12) is true for all \mathbf{h} , and the biased predictor will have the smaller expected PE.

We see that in both the curve-fitting example and in the theory sketched above, the trade-off is the same. Using more variables leads to smaller bias and larger variance, whereas using fewer variables makes the bias larger but the variance smaller.

EXERCISES 12a

1. Show that the square of the total bias $\|\mu - E[\mathbf{X}\hat{\beta}]\|^2$ is equal to the first term in (12.8).
2. Consider the regression model

$$Y_i = \alpha + \gamma_1 x_i + \gamma_2 z_i + \varepsilon_i \quad (i = 1, \dots, n), \quad (12.13)$$

where $\sum_i x_i = \sum_i z_i = 0$ and $\sum_i x_i^2 = \sum_i z_i^2 = 1$. Suppose that we want to predict the response Y corresponding to a vector (x, z) using a predictor of the form

$$\hat{Y} = \bar{Y} + \tilde{\gamma}_1 x,$$

where $\tilde{\gamma}_1$ is the least squares estimate obtained by fitting the model

$$Y_i = \alpha + \gamma_1 x_i + \varepsilon_i \quad (i = 1, \dots, n).$$

- (a) Assuming that the model (12.13) is correct, show that the expected model error $E[ME] = E[(\hat{Y} - \alpha - \gamma_1 x - \gamma_2 z)^2]$ of the predictor \hat{Y} is given by

$$E[ME] = \sigma^2(n^{-1} + x^2) + \gamma_2^2(rx - z)^2,$$

where $r = \sum_i x_i z_i$.

- (b) Under what circumstances will \hat{Y} be a better predictor than the predictor based on both x and z ?

12.3 CHOOSING THE BEST SUBSET

We saw in Section 12.2 that using a model containing all available explanatory variables can result in poor predictions. A possible strategy to improve predictions is to use only a subset of the explanatory variables and use a least squares predictor based on the chosen subset. How do we select the subset? One possible approach, which we explore in this section, is to define a criterion that measures how well a model performs, evaluate the criterion for each subset, and pick the subset that optimizes the criterion.

A variety of criteria are in common use. We can classify them into four classes:

1. Those based on goodness-of-fit measures

2. Those based on estimating the prediction error (or equivalently the model error)
3. Those based on estimating the difference between the true distribution of the responses and the distribution specified by the model
4. Those based on approximating posterior probabilities

12.3.1 Goodness-of-Fit Criteria

We know from previous chapters that the residual sum of squares (RSS) is a measure of goodness of fit. The RSS is not a good absolute measure, since if we have two sets of variables \mathcal{S}_1 and \mathcal{S}_2 with $\mathcal{S}_1 \subset \mathcal{S}_2$, then the RSS for the model based on \mathcal{S}_2 will be smaller than that for the model based on \mathcal{S}_1 (provided that the extra variables are not linear combinations of those in \mathcal{S}_1 , cf. Exercises 3f, No. 1, at the end of Section 3.7.2). Thus, even adding variables consisting of random numbers will decrease the RSS. However, if we have two competing models, both with the same number of variables, we would certainly prefer the model with the smaller RSS.

We need a way of correcting the RSS to account for the number of variables in the model. One way of doing this is to use the estimated residual variance

$$S^2 = \text{RSS}/(n - p),$$

where p is the number of columns in the $n \times p$ regression matrix \mathbf{X} . (If a constant term is used in the model, there will be $p - 1$ variables plus the constant term, for a total of p regression coefficients.) Another criterion is the *adjusted R*², defined by

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n}{n - p}, \quad (12.14)$$

where R^2 is the coefficient of determination introduced in Section 4.4. We select the model having the largest adjusted R^2 .

This measure is motivated by testing the adequacy of the model against a “full model” consisting of all available explanatory variables. Suppose that there are K variables available and the model under consideration contains $p - 1 < K$ of these. Writing the coefficients of determination for the two models as R_p^2 and R_{K+1}^2 , the F -test for testing the adequacy of the smaller model against the larger is (cf. Exercises 4c, No. 3, at the end of Section 4.4)

$$F_p = \frac{R_{K+1}^2 - R_p^2}{1 - R_{K+1}^2} \frac{n - K - 1}{K - p + 1}, \quad (12.15)$$

so that

$$1 - R_p^2 = (1 - R_{K+1}^2) \frac{(K - p + 1)F_p + n - K - 1}{n - K - 1}.$$

Using this and (12.14) we get

$$\bar{R}_p^2 = 1 - (1 - R_{K+1}^2) \frac{n}{n - K - 1} \frac{(K - p + 1)F_p + n - K - 1}{n - p}.$$

If $F_p \geq 1$, then

$$\frac{(K - p + 1)F_p + n - K - 1}{n - p} \geq 1$$

and

$$\begin{aligned}\bar{R}_p^2 &\leq 1 - (1 - R_{K+1}^2) \frac{n}{n - K - 1} \\ &= \bar{R}_{K+1}^2.\end{aligned}$$

This motivates the use of \bar{R}^2 as a model selection criterion, since large values of F_p are evidence in favor of the K -variable model. Note also that

$$\begin{aligned}\bar{R}_p^2 &= 1 - (1 - R_p^2) \frac{n}{n - p} \\ &= 1 - \frac{\text{RSS}_p}{\text{SSY}} \frac{n}{n - p} \\ &= 1 - S_p^2 \frac{n}{\text{SSY}},\end{aligned}\tag{12.16}$$

where $\text{SSY} = \sum_i (Y_i - \bar{Y})^2$, so that the model with maximum \bar{R}^2 is also the model with minimum S^2 .

12.3.2 Criteria Based on Prediction Error

We now look at some criteria based on the idea of choosing a model that predicts well. In Section 12.2 we introduced the ME as a measure of how well a particular model predicts future data. However, because the expected ME involves the unknown mean vector μ , it cannot be used as a criterion for model selection. To get an operational criterion, we must estimate the expected ME, and we now discuss some ways of doing this.

Mallow's C_p

Let RSS_p denote the RSS that results from fitting a model with p parameters and having an $n \times p$ regression matrix \mathbf{X}_p . If $\mathbf{P}_p = \mathbf{X}_p(\mathbf{X}'_p \mathbf{X}_p)^{-1} \mathbf{X}'_p$, then, using Theorem 1.5 and (12.8), we get

$$\begin{aligned}E[\text{RSS}_p] &= E[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_p)\mathbf{Y}] \\ &= \mu'(\mathbf{I}_n - \mathbf{P}_p)\mu + (n - p)\sigma^2 \\ &= E[\text{ME}] + (n - 2p)\sigma^2,\end{aligned}\tag{12.17}$$

by (12.8), so that

$$\frac{E[\text{ME}]}{\sigma^2} = \frac{E[\text{RSS}_p]}{\sigma^2} + 2p - n.$$

If we had an estimate $\hat{\sigma}^2$ of σ^2 , we could use

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} + 2p - n \quad (12.18)$$

as an estimate of $E[\text{ME}]/\sigma^2$, the scaled expected ME. If the model fits well in the sense that μ is well approximated by vectors in $\mathcal{C}(\mathbf{X}_p)$, then the quantity $\|(\mathbf{I}_n - \mathbf{P}_p)\mu\|^2 = \mu'(\mathbf{I}_n - \mathbf{P}_p)\mu$ will be small. Hence, from (12.8), the expected value of C_p will be close to p , since

$$\begin{aligned} E[C_p] &\approx \frac{E[\text{RSS}_p]}{\sigma^2} + 2p - n \\ &= \frac{\mu'(\mathbf{I}_n - \mathbf{P}_p)\mu}{\sigma^2} + n - p + 2p - n \\ &\approx p. \end{aligned} \quad (12.19)$$

Mallows, who popularized the use of C_p (see, e.g., Mallows [1973]), suggested using the C_p plot, a plot of C_p versus p for all possible models, as a device for recognizing models for which $C_p \approx p$ (and hence fit well), and for gaining insight into which variables contribute to the regression.

It is common practice to estimate σ^2 using the full regression containing all K available explanatory variables, so that

$$\hat{\sigma}^2 = \frac{\text{RSS}_{K+1}}{n - K - 1},$$

which implies that $C_{K+1} = K + 1$. Also, using the relationship (cf. Exercises 12b, No. 1, at the end of Section 12.3.4)

$$\frac{\text{RSS}_p}{\text{RSS}_{K+1}} \frac{n - K - 1}{n - p} = \frac{1 - \bar{R}_p^2}{1 - \bar{R}_{K+1}^2},$$

we get

$$\begin{aligned} C_p &= \frac{\text{RSS}_p}{\hat{\sigma}^2} - n + 2p \\ &= \frac{\text{RSS}_p(n - K - 1)}{\text{RSS}_{K+1}} - n + 2p \\ &= \frac{(1 - \bar{R}_p^2)(n - p)}{1 - \bar{R}_{K+1}^2} - n + 2p, \end{aligned}$$

so that

$$\frac{C_p - p}{n - p} + 1 = \frac{1 - \bar{R}_p^2}{1 - \bar{R}_{K+1}^2}.$$

Hence if n is large compared to p , the smallest value of $C_p - p$ corresponds approximately to the largest \bar{R}_p^2 .

Since C_p estimates the scaled expected ME, it is often thought that choosing the model with the smallest C_p results in good predictions. However, Mallows strongly discouraged choosing the model with the smallest C_p , for reasons that explored more fully in Section 12.9. This also applies to using \bar{R}_p^2 or the equivalent S_p^2 as the basis for model choice.

Cross-Validation

If $(\mathbf{x}_{0i}, Y_{0i})$, $i = 1, \dots, m$, is a new set of data following the same model as the training data (\mathbf{x}_i, Y_i) , an obvious measure of the prediction error is

$$\frac{1}{m} \sum_{i=1}^m (Y_{0i} - \mathbf{x}'_{0i} \hat{\beta})^2, \quad (12.20)$$

where $\hat{\beta}$ is calculated using the training set and some assumed model. In practice, we do not often have the luxury of additional data. We can, however, do something similar by dividing the training set into two disjoint sets of cases, using one set to calculate $\hat{\beta}$, and the other to evaluate the prediction by calculating (12.20). Of course, accuracy is lost if too few cases are used to estimate the form of the predictor. Also, if too few cases are used to estimate the prediction error, we cannot make an accurate assessment of how good the proposed model is for prediction.

To combat this, we can select a subset D of d cases, estimate the predictor using the remaining $n - d$ cases, and then calculate (12.20). This produces an estimate $\text{PE}(D)$ of the prediction error. We then repeat the process for selected d -subsets D_1, D_2, \dots , and average the resulting estimates $\text{PE}(D_i)$. This process, called *cross-validation*, has a large literature; see Allen [1971] and Stone [1974] for early work, and George [2000] for a recent brief review.

The most popular (although, as we will see below, not the best) choice of d is $d = 1$, which involves leaving out each case in turn, and calculating the estimate of β from the remaining $(n - 1)$ cases. As in Section 10.1, we call this estimate $\hat{\beta}(i)$. The error in predicting the i th case is $Y_i - \mathbf{x}'_i \hat{\beta}(i)$, which leads to the *leave-one-out* or $\text{CV}(1)$ prediction error estimate

$$\text{CV}(1) = \frac{1}{n} \sum_{i=1}^n [Y_i - \mathbf{x}'_i \hat{\beta}(i)]^2. \quad (12.21)$$

To simplify this expression, we use (10.5) to get

$$\begin{aligned} Y_i - \mathbf{x}'_i \hat{\beta}(i) &= Y_i - \mathbf{x}'_i \left[\hat{\beta} - \frac{(\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i (Y_i - \mathbf{x}'_i \hat{\beta})}{1 - h_i} \right] \\ &= Y_i - \mathbf{x}'_i \hat{\beta} + \frac{h_i (Y_i - \mathbf{x}'_i \hat{\beta})}{1 - h_i} \\ &= \frac{Y_i - \mathbf{x}'_i \hat{\beta}}{1 - h_i}. \end{aligned} \quad (12.22)$$

As in Section 10.1, h_i is the i th hat matrix diagonal, the i th diagonal element of the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for the model under consideration. Using (12.22), the CV(1) estimate can be written

$$\text{CV}(1) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \mathbf{x}_i' \hat{\beta}}{1 - h_i} \right)^2. \quad (12.23)$$

Next, we develop an expression for the expectation of $\text{CV}(1)$ and show that it tends to overestimate the PE. If $\eta_i = \{(\mathbf{I} - \mathbf{P})\boldsymbol{\mu}\}_i$ and $\boldsymbol{\varepsilon} = \mathbf{Y} - \boldsymbol{\mu}$, then

$$\begin{aligned} Y_i - \mathbf{x}_i' \hat{\beta} &= \{\mathbf{Y} - \mathbf{X}\hat{\beta}\}_i \\ &= \{(\mathbf{I}_n - \mathbf{P})\mathbf{Y}\}_i \\ &= \{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}\}_i + \{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i \\ &= \eta_i + \{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i, \end{aligned}$$

so that

$$\begin{aligned} E[(Y_i - \mathbf{x}_i' \hat{\beta})^2] &= E[(\eta_i + \{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i)^2] \\ &= \eta_i^2 + E[\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i^2]. \end{aligned} \quad (12.24)$$

Now let \mathbf{D}_i be a diagonal matrix whose i th diagonal element is unity and the rest are zero. Then

$$\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i^2 = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P})\mathbf{D}_i(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon},$$

so that, by A.1.2 and the fact that $\mathbf{I}_n - \mathbf{P}$ is idempotent,

$$\begin{aligned} E[\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}\}_i^2] &= E[\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{P})\mathbf{D}_i(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}] \\ &= \sigma^2 \text{tr}[(\mathbf{I}_n - \mathbf{P})\mathbf{D}_i(\mathbf{I}_n - \mathbf{P})] \\ &= \sigma^2 \text{tr}[\mathbf{D}_i(\mathbf{I}_n - \mathbf{P})] \\ &= \sigma^2(1 - h_i), \end{aligned}$$

so

$$E[(Y_i - \mathbf{x}_i' \hat{\beta})^2] = \eta_i^2 + \sigma^2(1 - h_i).$$

It follows from (12.23) that

$$E[n\text{CV}(1)] = \sum_{i=1}^n \frac{\eta_i^2 + \sigma^2(1 - h_i)}{(1 - h_i)^2}. \quad (12.25)$$

Now consider the expected PE. From (12.9) we have

$$\begin{aligned} E[\text{PE}] &= (n + p)\sigma^2 + \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} \\ &= (n + p)\sigma^2 + \|(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}\|^2 \\ &= (n + p)\sigma^2 + \sum_{i=1}^n \eta_i^2 \\ &= \sum_{i=1}^n [\sigma^2(1 - h_i) + \eta_i^2], \end{aligned} \quad (12.26)$$

by (10.13). Combining (12.25) and (12.26), we get

$$\begin{aligned} E[nCV(1) - PE] &= \sum_{i=1}^n \frac{\eta_i^2 + \sigma^2(1-h_i)}{(1-h_i)^2} - \sum_{i=1}^n [\sigma^2(1+h_i) + \eta_i^2] \\ &= \sum_{i=1}^n \frac{\eta_i^2 h_i (2-h_i)}{(1-h_i)^2} + \sigma^2 \sum_{i=1}^n \frac{h_i^2}{1-h_i}. \end{aligned} \quad (12.27)$$

Since $0 \leq h_i \leq 1$, both terms in this sum are positive, and hence $nCV(1)$ tends to overestimate PE .

For $d > 1$, the calculations are numerically intensive, although there is a generalization of (12.22) that reduces the computational demands. Specifically, suppose that $D = \{i_1, \dots, i_d\}$ are the cases to be omitted. Let $\mathbf{Y}'_D = (Y_{i_1}, \dots, Y_{i_d})$ denote the estimate of β based on the cases not in D by $\hat{\beta}(D)$, let \mathbf{X}_D be the submatrix of \mathbf{X} consisting of the rows with labels in D , and let $\mathbf{H}_D = \mathbf{X}_D(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D$. Then, since we are now predicting D cases simultaneously and comparing our prediction with \mathbf{Y}_D , we have (cf. Exercises 10e, No. 2, at the end of Section 10.6.5)

$$\mathbf{Y}_D - \mathbf{X}_D \hat{\beta}(D) = (\mathbf{I} - \mathbf{H}_D)^{-1}(\mathbf{Y}_D - \mathbf{X}_D \hat{\beta}),$$

so that

$$CV(d) = \binom{n}{d}^{-1} \sum_D (\mathbf{Y}_D - \mathbf{X}_D \hat{\beta})' (\mathbf{I} - \mathbf{H}_D)^{-2} (\mathbf{Y}_D - \mathbf{X}_D \hat{\beta}),$$

where the sum is taken over all d -subsets of cases. For d more than 2 or 3, calculating $CV(d)$ is very computationally intensive. Alternatives are to select subsets of cases at random, or to select subsets based on a balanced incomplete block design; see Shao [1993] or Zhang [1993] for details. Shao and Zhang show that using $CV(d)$, where d is an appreciable fraction of n , leads to better subset selection and better predictions than does $CV(1)$.

The expressions (12.19) and (12.25) for the expectations of C_p and $CV(1)$ hold true only if the model is specified in advance. If the model is chosen on the basis of the data, say by choosing the model for which $CV(1)$ or C_p is a minimum, then the formulas no longer hold. We illustrate this important point with an example.

EXAMPLE 12.1 Suppose that the variance σ^2 is known, there is no constant term, and that the K explanatory variables are orthonormal. In this case the RSS for fitting a p -parameter submodel with regression matrix \mathbf{X}_p (which satisfies $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$) is

$$\begin{aligned} RSS_p &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}\mathbf{X}'_p\mathbf{X}_p\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \sum_{j=1}^p \hat{\beta}_j^2. \end{aligned}$$

Note that because $\mathbf{X}'_p \mathbf{X}_p = \mathbf{I}_p$, the estimated regression coefficients are not changed if variables are added or subtracted from the model (cf. Section 3.10).

If $\text{Var}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$, then $\text{Var}[\hat{\beta}] = \sigma^2 (\mathbf{X}'_p \mathbf{X}_p)^{-1} = \sigma^2 \mathbf{I}_p$.

When σ^2 is known, we can write C_p [cf. (12.18)] as

$$\begin{aligned} C_p &= \frac{\text{RSS}_p}{\sigma^2} + 2p - n \\ &= \frac{\mathbf{Y}' \mathbf{Y}}{\sigma^2} - \sum_{j=1}^p \left(\frac{\hat{\beta}_j^2}{\sigma^2} - 2 \right) - n. \end{aligned} \quad (12.28)$$

If we fit the full model with all K variables, we get K estimated regression coefficients $\hat{\beta}_1, \dots, \hat{\beta}_K$. We now order these estimates and write them as $\hat{\beta}_{(j)}$, $j = 1, \dots, K$, where

$$\hat{\beta}_{(1)}^2 \geq \dots \geq \hat{\beta}_{(K)}^2.$$

Then it is clear from (12.28) that the submodel with smallest C_p is the one including exactly the variables that satisfy $\hat{\beta}_j^2/\sigma^2 \geq 2$ (i.e., the variables whose coefficients have magnitudes exceeding the *threshold* $\sqrt{2}\sigma$). Let \hat{p} be the number of variables in this model and let $C_{\hat{p}}$ be the corresponding C_p value.

Now suppose that the true model satisfies $E[\mathbf{Y}] = \mathbf{0}$, so that the $\hat{\beta}_j$'s are independent, each having a $N(0, \sigma^2)$ distribution. Then, from (12.18), if the model is specified in advance, $E[C_p] = p$. However,

$$C_{\hat{p}} = \frac{\mathbf{Y}' \mathbf{Y}}{\sigma^2} - \sum_{j=1}^{\hat{p}} \left(\frac{\hat{\beta}_{(j)}^2}{\sigma^2} - 2 \right) - n,$$

where the upper limit of the sum is random. Since $E[\mathbf{Y}] = \mathbf{0}$,

$$E[\mathbf{Y}' \mathbf{Y}/\sigma^2] = n,$$

and hence

$$E[C_{\hat{p}}] = -E \left[\sum_{j=1}^{\hat{p}} \left(\frac{\hat{\beta}_{(j)}^2}{\sigma^2} - 2 \right) \right].$$

Since each term in the sum is nonnegative, the sum must have a nonnegative expected value, so that $E[C_{\hat{p}}] \leq 0$. Thus $E[C_{\hat{p}}] < E[C_p]$, and using $C_{\hat{p}}$ leads to an underestimate of the true model error. \square

The Little Bootstrap

A better method for estimating the PE or ME of the model selected is the *little bootstrap*, a technique introduced by Breiman in a series of papers (Breiman [1992, 1995, 1996b]). The little bootstrap is based on resampling

residuals and gives an almost unbiased method of estimating the PE of a data-selected model. It is also useful in estimating the PE of predictors constructed using shrinkage methods, so we defer further discussion until we have dealt with the latter in Section 12.5.

12.3.3 Estimating Distributional Discrepancies

Another type of criterion is based on the idea of a discrepancy between the true distribution of the data \mathbf{Y} and the distribution specified by the candidate model. If $f(\mathbf{y})$ is the density of the true distribution and $g(\mathbf{y})$ is that specified by the model, a well-known discrepancy measure is the *Kullback–Leibler discrepancy*

$$KL(f, g) = \int \log \frac{f(\mathbf{y})}{g(\mathbf{y})} f(\mathbf{y}) d\mathbf{y}.$$

Note that this is not a true distance, as $KL(f, g) \neq KL(g, f)$. Rather, it is a measure of the difference between the true fixed f and various competing models g . Note that $KL(f, g) \geq KL(f, f) = 0$ (cf. Exercises 12b, No. 3).

In practice, we are interested in a family of possible models $g(\mathbf{y}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ ranges over some parameter space, and we use the notation $KL(f, g; \boldsymbol{\theta})$ to reflect this. The true model f may or may not be of the form $g(\mathbf{y}; \boldsymbol{\theta})$ for some particular $\boldsymbol{\theta}$. In any event, we would like to choose the model corresponding to the value of $\boldsymbol{\theta}$ that minimizes $KL(f, g; \boldsymbol{\theta})$ or, equivalently, that minimizes

$$-\int \log g(\mathbf{y}; \boldsymbol{\theta}) f(\mathbf{y}) d\mathbf{y} = -E[\log g(\mathbf{Y}; \boldsymbol{\theta})]. \quad (12.29)$$

This is equal to $KL(f, g; \boldsymbol{\theta})$ up to the constant

$$\int \log \{f(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y},$$

which does not depend on g . Note that the criterion (12.29) depends on two unknowns, the parameter $\boldsymbol{\theta}$ and the unknown true model f .

To estimate $\boldsymbol{\theta}$, suppose that we have a sample \mathbf{Y} and we estimate $\boldsymbol{\theta}$ using the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ that maximizes $g(\mathbf{Y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. This leads to the modified criterion

$$\Delta = -\int \log g(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{Y})) f(\mathbf{x}) d\mathbf{x} \quad (12.30)$$

with expected value

$$E[\Delta] = -\int \int \log g(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{y})) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

The criterion Δ measures the discrepancy (up to a fixed constant depending only on f) between the unknown true model f and the best-fitting model of

the form $g(\mathbf{y}; \boldsymbol{\theta})$. However, it still depends on the unknown f . To obtain an operational criterion, we need an estimate of Δ .

The standard estimate of Δ (in fact, it estimates 2Δ) is the *Akaike information criterion* (AIC) (Akaike [1973], Burnham and Anderson [1998]), defined by

$$\text{AIC} = -2 \log g(\mathbf{Y}; \hat{\boldsymbol{\theta}}(\mathbf{Y})) + 2r, \quad (12.31)$$

where r is the dimension of the parameter vector $\boldsymbol{\theta}$. The AIC is applicable to any modeling situation. In the case of linear models, the true model $f(\mathbf{y})$ of \mathbf{Y} may be taken as multivariate normal with mean vector $\boldsymbol{\mu}$, and variance matrix $\sigma_0^2 \mathbf{I}_n$. A typical candidate model $g(\mathbf{y}; \boldsymbol{\theta})$ is also multivariate normal, with mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance matrix $\sigma^2 \mathbf{I}_n$. The regression matrix \mathbf{X} is assumed to be $n \times p$ of rank p , and the parameter vector in this case is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$.

It follows from (2.1) that for a candidate model g ,

$$-\log g(\mathbf{y}; \boldsymbol{\theta}) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The maximum likelihood estimates (MLEs) are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\text{RSS}}{n}.$$

Let \mathbf{Y}_0 have the same distribution as \mathbf{Y} and be independent of \mathbf{Y} . Then, from (12.30),

$$\begin{aligned} \Delta &= E_{\mathbf{Y}_0}[-\log g(\mathbf{Y}_0; \hat{\boldsymbol{\theta}}(\mathbf{Y}))] \\ &= \frac{n}{2} \log(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2} E_{\mathbf{Y}_0}[(\mathbf{Y}_0 - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}_0 - \mathbf{X}\hat{\boldsymbol{\beta}})] \\ &= \frac{n}{2} \log(2\pi\hat{\sigma}^2) + \frac{1}{2\hat{\sigma}^2} \{n\sigma_0^2 + \|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2\}, \end{aligned}$$

where the last step follows from Theorem 1.5. Using the fact that the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent [Theorem 3.5(iii) in Section 3.4], and setting $\lambda = \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}/\sigma_0^2$, we get

$$\begin{aligned} E[2\Delta] &= E_{\mathbf{Y}}[n \log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} \{n\sigma_0^2 + \|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2\}] \\ &= nE[\log 2(\pi\hat{\sigma}^2)] + \{n\sigma_0^2 + E[\|\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2]\}E\left[\frac{1}{\hat{\sigma}^2}\right] \\ &= nE[\log(2\pi\hat{\sigma}^2)] + (n + p + \lambda)E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] \end{aligned}$$

by (12.8). To estimate this, we use the AIC. From (12.31), and putting $r = p + 1$, we see that the AIC can be written

$$\begin{aligned} \text{AIC} &= n \log(2\pi\hat{\sigma}^2) + \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} + 2(p + 1) \\ &= n \log(2\pi\hat{\sigma}^2) + n + 2(p + 1). \end{aligned}$$

Comparing these two expressions, and using the approximation

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] = 1 + O(n^{-1}),$$

we see that up to $O(1)$ the expected value of AIC and $E[2\Delta]$ are the same, which suggests that AIC is a reasonable estimate of $E[2\Delta]$.

In the case of linear models, the true distribution f is of the form $g(y, \theta)$ for some θ if and only if μ is in $C(\mathbf{X})$, which is equivalent to $\lambda = 0$. When this holds, $n\hat{\sigma}^2/\sigma_0^2 \sim \chi_{n-p}^2$, so that using the result

$$E\left[\frac{1}{X}\right] = \frac{1}{\nu - 2} \quad (12.32)$$

for a χ_ν^2 random variable X , we get

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] = \frac{n}{n - p - 2}$$

and

$$E[2\Delta] = nE[\log(2\pi\hat{\sigma}^2)] + \frac{n(n+p)}{n - p - 2}.$$

Both 2Δ and AIC have expectations that are $O(n)$. The bias in estimating $E[2\Delta]$ by AIC is

$$E[AIC - 2\Delta] = n + 2(p+1) - \frac{n(n+p)}{n - p - 2} = 2(p+1) - \frac{2n(p+1)}{n - p - 2},$$

which is $O(1)$. Although the bias is of smaller order than $E[2\Delta]$, we can get an exactly unbiased estimate of $E[2\Delta]$ by using the modified criterion

$$AIC_c = n \log 2\pi\hat{\sigma}^2 + \frac{n(n+p)}{n - p - 2}.$$

Hurvich and Tsai [1991] give some examples where AIC_c is a much better estimate of $E[2\Delta]$ than AIC.

If μ is not in $C(\mathbf{X})$, then from (12.17), $nE[\hat{\sigma}^2] = \sigma_0^2(\lambda + n - p)$ and

$$E\left[\frac{\sigma_0^2}{\hat{\sigma}^2}\right] \approx \frac{\sigma_0^2}{E[\hat{\sigma}^2]} = \frac{n}{\lambda + n - p},$$

leading to

$$\begin{aligned} E[AIC - 2\Delta] &\approx n + 2(p+1) - \frac{(n+p+\lambda)n}{\lambda + n - p} \\ &= 2\left(p+1 - \frac{np}{\lambda + n - p}\right) \end{aligned}$$

or approximately 2 for large n , as the last term in parentheses is approximately p . Hurvich and Tsai [1991] provide exact expressions for the bias in the form of an infinite series when μ is not in $\mathcal{C}(\mathbf{X})$.

A slightly different form of the AIC arises if we assume that σ^2 is known. If this is the case, there are only p parameters, and

$$-2 \log g(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) = n \log(2\pi\sigma^2) + \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sigma^2},$$

so that up to a constant $n \log(2\pi\sigma^2)$ not depending on the model,

$$\text{AIC} = \frac{\text{RSS}_p}{\sigma^2} + 2p. \quad (12.33)$$

Thus, in this case, the AIC is very similar to C_p . If σ^2 is replaced by an estimate $\hat{\sigma}^2$, they differ by the constant n .

An obvious generalization of this version of the AIC is to consider criteria of the form

$$\frac{\text{RSS}_p}{\sigma^2} + a_n p, \quad (12.34)$$

where a_n is allowed to depend on n . The choice $a_n = \log n$ leads to a criterion known as the *Bayesian information criterion* (BIC), that is discussed further in Section 12.3.4. For other choices, see Hannan and Quinn [1979] and Zhang [1992].

Note that in the case of orthonormal explanatory variables and known σ^2 , both C_p and AIC are equivalent; by Example 12.1, choosing the model minimizing AIC is equivalent to including variables for which $\hat{\beta}_j^2/\sigma^2 > 2$. More generally, if we use the criterion (12.34), this is equivalent to including variables with $\hat{\beta}_j^2/\sigma^2 > a_n$.

The choice $a_n = 2$ can lead to too many redundant variables being included in the model. In the extreme case when all regression coefficients are zero, the quantities $\hat{\beta}_j^2/\sigma^2$ have independent χ_1^2 distributions, so that in the case of K variables, the probability of choosing the correct model is $(\Pr[\chi_1^2 \leq 2])^K = 0.8427^K$. For $K = 10$, this is 0.181, and for $K = 20$, it is 0.033. Clearly, this problem can be solved by increasing a_n . An approach to selecting the optimal value of a_n is discussed briefly in Section 12.9.

12.3.4 Approximating Posterior Probabilities

Suppose that the true mean vector $\mu = E[\mathbf{Y}]$ is in fact of the form $\mu = \mathbf{X}\boldsymbol{\beta}$, where \mathbf{X} has full rank $K + 1$, and we wish to see if one of the submodels $\mu = \mathbf{X}_p\boldsymbol{\beta}_p$ holds, where \mathbf{X}_p is a submatrix of \mathbf{X} of rank p [$< (K + 1)$]. If we take a Bayesian perspective and assign a prior probability α_p to this model, then the posterior probability of the model after observing data \mathbf{y} is proportional to

$$\alpha_p \int f_p(\mathbf{y}|\boldsymbol{\theta})\pi_p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (12.35)$$

where f_p is the conditional density of the data given the parameters and the model, and π_p is the prior for the parameter vector for the model having regression matrix \mathbf{X}_p . We then select the model with the largest posterior probability.

Let us assume that n is large and that σ^2 is known. As in Section 3.12, we will assume that conditional on σ^2 , the p -dimensional prior for β_p is multivariate normal with mean \mathbf{m} and variance matrix $\sigma^2\mathbf{V}$, and that the conditional distribution of the data given the parameters and the model is $N_n(\mathbf{X}_p\beta_p, \sigma^2\mathbf{I}_n)$. In this case the integral in (12.35) is proportional to

$$\int \exp[-Q/(2\sigma^2)] d\beta \quad (12.36)$$

where

$$Q = (\mathbf{y} - \mathbf{X}_p\beta_p)'(\mathbf{y} - \mathbf{X}_p\beta_p) + (\beta_p - \mathbf{m})'\mathbf{V}^{-1}(\beta_p - \mathbf{m}). \quad (12.37)$$

By Theorem 3.7 in Section 3.12, we can write Q as

$$Q = (\beta_p - \mathbf{m}_*)'\mathbf{V}_*^{-1}(\beta_p - \mathbf{m}_*) + (\mathbf{y} - \mathbf{X}_p\mathbf{m})'\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}_p\mathbf{m}), \quad (12.38)$$

where $\mathbf{V}_* = (\mathbf{X}'_p\mathbf{X}_p + \mathbf{V}^{-1})^{-1}$, $\mathbf{m}_* = \mathbf{V}_*(\mathbf{X}'_p\mathbf{y} + \mathbf{V}^{-1}\mathbf{m})$, and $\mathbf{W} = (\mathbf{I}_n + \mathbf{X}_p\mathbf{V}\mathbf{X}'_p)$. By (12.37) and the results of Section 2.1, the integral (12.36) is

$$(2\pi)^{p/2} \det(\mathbf{V}_*)^{1/2} \exp[-(\mathbf{y} - \mathbf{X}_p\mathbf{m})'\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}_p\mathbf{m})/(2\sigma^2)],$$

so, up to quantities that do not depend on the model, the log of the posterior probability corresponding to the model with regression matrix \mathbf{X}_p is

$$\log \alpha_p + \frac{1}{2} \log \det(\mathbf{V}_*) - (\mathbf{y} - \mathbf{X}_p\mathbf{m})'\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}_p\mathbf{m})/(2\sigma^2). \quad (12.39)$$

We note that $\log \alpha_p$ is $O(1)$, and we can approximate the other terms as follows. Assume that $\mathbf{X}'_p\mathbf{X}_p = n\Sigma_p$, where Σ_p is $O(1)$. Then

$$\begin{aligned} \det(\mathbf{V}_*) &= 1/\det(\mathbf{X}'_p\mathbf{X}_p + \mathbf{V}^{-1}) \\ &= 1/\det(n\Sigma_p + \mathbf{V}^{-1}) \\ &= 1/[n^p \det(\Sigma_p + \frac{1}{n}\mathbf{V}^{-1})], \end{aligned}$$

so that

$$\log \det(\mathbf{V}_*) = -p \log n + O(1). \quad (12.40)$$

Next, we use the fact that $\mathbf{W}^{-1} = \mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p$ (see the proof of Theorem 3.7) to approximate the third term in (12.39). We have

$$\begin{aligned} & (\mathbf{y} - \mathbf{X}_p \mathbf{m})' \mathbf{W}^{-1} (\mathbf{y} - \mathbf{X}_p \mathbf{m}) \\ &= [(\mathbf{y} - \mathbf{X}_p \hat{\beta}) - \mathbf{X}_p (\hat{\beta} - \mathbf{m})]' (\mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p) \\ &\quad \times [(\mathbf{y} - \mathbf{X}_p \hat{\beta}) - \mathbf{X}_p (\hat{\beta} - \mathbf{m})] \\ &= (\mathbf{y} - \mathbf{X}_p \hat{\beta})' (\mathbf{y} - \mathbf{X}_p \hat{\beta}) \\ &\quad - (\mathbf{X}'_p \mathbf{y} - \mathbf{X}'_p \mathbf{X}_p \hat{\beta})' \mathbf{V}_* (\mathbf{X}'_p \mathbf{y} - \mathbf{X}'_p \mathbf{X}_p \hat{\beta}) \\ &\quad + (\hat{\beta} - \mathbf{m})' \mathbf{X}'_p (\mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p) \mathbf{X}_p (\hat{\beta} - \mathbf{m}) \\ &\quad - 2(\mathbf{y} - \mathbf{X}_p \hat{\beta})' (\mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p) \mathbf{X}_p (\hat{\beta} - \mathbf{m}). \end{aligned}$$

The first term in this expression is RSS_p , the second is zero since $\hat{\beta}$ is a solution of the normal equations, and the fourth term is also zero since

$$\begin{aligned} (\mathbf{y} - \mathbf{X}_p \hat{\beta})' (\mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p) \mathbf{X}_p &= (\mathbf{y} - \mathbf{X}_p \hat{\beta})' \mathbf{X}_p (\mathbf{I}_n - \mathbf{V}_* \mathbf{X}'_p \mathbf{X}_p) \\ &= (\mathbf{X}'_p \mathbf{y} - \mathbf{X}'_p \mathbf{X}_p \hat{\beta}) (\mathbf{I}_n - \mathbf{V}_* \mathbf{X}'_p \mathbf{X}_p) \\ &= 0. \end{aligned}$$

Finally, as in the proof of Theorem 3.7, we have

$$\begin{aligned} \mathbf{X}'_p (\mathbf{I}_n - \mathbf{X}_p \mathbf{V}_* \mathbf{X}'_p) \mathbf{X}_p &= (\mathbf{V} + (\mathbf{X}'_p \mathbf{X}_p)^{-1})^{-1} \\ &= (\mathbf{V} + n^{-1} \Sigma^{-1})^{-1} \\ &= O(1), \end{aligned}$$

so that

$$(\mathbf{y} - \mathbf{X}_p \mathbf{m})' \mathbf{W}^{-1} (\mathbf{y} - \mathbf{X}_p \mathbf{m}) = \text{RSS}_p + O(1). \quad (12.41)$$

Using this result and (12.40), it follows from (12.39) that the log of the posterior probability of the model is equal to

$$-\frac{\text{RSS}_p}{2\sigma^2} - \frac{1}{2} p \log n + O(1).$$

Thus, selecting the model with largest posterior probability is asymptotically equivalent to selecting the model for which

$$\frac{\text{RSS}_p}{\sigma^2} + p \log n \quad (12.42)$$

is a minimum. The criterion (12.42) is just the BIC discussed in Section 12.3.3, first introduced by Schwarz [1978] as a device for estimating the dimension of the correct model. Although it is motivated by Bayesian ideas, the actual priors used do not explicitly form part of the criterion, so that the BIC has a non-Bayesian interpretation, similar to that of the AIC, as a goodness-of-fit

measure that penalizes models that have an excessive number of parameters. For sample sizes in excess of 7 (i.e., when $\log n > 2$), the BIC imposes a greater penalty for each extra parameter than does the AIC.

EXERCISES 12b

1. Prove that

$$\frac{\text{RSS}_p}{\text{RSS}_{K+1}} = \frac{n - K - 1}{n - p} \cdot \frac{1 - \bar{R}_p^2}{1 - \bar{R}_{K+1}^2}.$$

2. Perform a small simulation to assess how badly C_p underestimates the prediction error in Example 12.1.
3. Prove that $KL(f, g) \geq 0$. You may assume that $f(\mathbf{y}) > 0$ and $g(\mathbf{y}) > 0$ for all \mathbf{x} . Hint: Show that $\log x \leq x - 1$ for $x > 0$, and put $x = g(\mathbf{y})/f(\mathbf{y})$.
4. Obtain (12.35) using conditional probability arguments.

12.4 STEPWISE METHODS

Computing all possible regression models rapidly becomes impractical as the number of variables increases. In this case we may resort to step-by-step methods that select subsets sequentially and avoid the computational cost of fitting very large numbers numbers of models.

Suppose as in Section 12.3 that we have a regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{X} is $n \times (K + 1)$, and we want to identify the “significant” variables having nonzero regression coefficients. Suppose that we divide the K variables up into two sets: the first set consists of $p - 1$ variables that we regard as important, while the second, which contains the remaining $K - p + 1$ variables, consists of variables whose coefficients we suspect are zero. We can test if the second set contains no significant variables by using the F -statistic

$$F = \frac{\text{RSS}_p - \text{RSS}_{K+1}}{\text{RSS}_{K+1}} \frac{n - K - 1}{K - p + 1}. \quad (12.43)$$

We can think of this test as discriminating between two models, having K and $p - 1$ variables, respectively. In particular, if $K = p$, the test is

$$F = \frac{\text{RSS}_p - \text{RSS}_{p+1}}{\text{RSS}_{p+1}} (n - p - 1), \quad (12.44)$$

which tests if the addition of a specified extra variable is necessary.

12.4.1 Forward Selection

The discussion above suggests the following procedure. We start with a model containing only the constant term, compute (12.44) with $p = 1$ for all the available explanatory variables in turn and pick the variable for which (12.44) is the greatest. We then repeat the procedure for $p = 2, 3, \dots$, selecting at each stage the variable not currently included that gives the maximum value of F . We stop if the maximum F at any stage does not exceed some threshold value F_{IN} . This procedure is commonly called *forward selection* (FS).

A slight variation on forward selection is to pick at each stage the variable not currently in the model having the greatest partial correlation with the response, given the variables currently included. This is equivalent to picking the variable that maximizes the numerator of (12.44) (cf. Exercises 12c, No. 2). Thompson [1978] calls the version using the F -statistic *forward ranking*, reserving the term forward selection for the partial correlation version.

We illustrate with an example, which also compares FS with the all-possible-regressions method using the AIC criterion.

EXAMPLE 12.2 In this example we present a graphical description of forward selection and compare it with the all-possible-regressions approach to choosing a model, using the AIC as a criterion. We examine the very simple case when we have just two possible explanatory variables, x_1 and x_2 .

In this case, assuming that a constant term is always included, there are only four possible models: the null model $\{0\}$, consisting of the constant term alone, and the models $\{x_1\}$, $\{x_2\}$, and $\{x_1, x_2\}$. Consider a centered and scaled regression with these two explanatory variables. The model is

$$Y_i = \alpha_0 + \gamma_1 x_{i1}^* + \gamma_2 x_{i2}^* + \varepsilon_i. \quad (12.45)$$

For simplicity we will assume that the variance σ^2 of ε_i is known; we suppose that $\sigma^2 = 1$.

From Example 3.10 in Section 3.11.2, the least squares estimate of γ_1 is

$$\hat{\gamma}_1 = \frac{r_1 - rr_2}{1 - r^2},$$

where $r_1 = \sum_i x_{i1}^* Y_i$, $r_2 = \sum_i x_{i2}^* Y_i$, and $r = \sum_i x_{i1}^* x_{i2}^*$. A similar expression holds for $\hat{\gamma}_2$. Both estimates have variance $1/(1 - r^2)$. To test the hypothesis $\gamma_2 = 0$, we use the statistic

$$\hat{\gamma}_2(1 - r^2)^{1/2} = \frac{r_2 - rr_1}{(1 - r^2)^{1/2}},$$

which has a $N(0, 1)$ distribution under the null hypothesis, since we are assuming that the variance is 1. If we had assumed, instead, that the model was

$$Y_i = \alpha_0 + \gamma_1 x_{i1}^* + \varepsilon_i, \quad (12.46)$$

the estimate of γ_1 would be r_1 , with variance 1. The test statistic for testing $\gamma_1 = 0$ is now r_1 , whose null distribution is also standard normal. Similar remarks apply to the model having the single explanatory variable x_2^* .

Suppose that we select the model using forward selection. As demonstrated in Exercises 12c, No. 1, the first variable to be selected is the one most highly correlated with the response. Provided that the absolute value of this correlation exceeds some cutoff value c_1 , we include this variable in the model. Since the correlation is, in fact, the test statistic for testing if the coefficient is zero and the null distribution is the standard normal, a suitable value for c_1 is the 90th percentile of the standard normal distribution, or $c_1 = 1.64$.

Assuming that this cutoff is exceeded, we then have to decide if the other variable is to be included. Suppose that x_1 has been included at the first stage. Then the second variable will be included if the hypothesis $\gamma_2 = 0$ is rejected, say at the 10% level. Thus, we select model $\{x_1, x_2\}$ if $|r_2 - rr_1| > c_2$, where $c_2 = (1 - r^2)^{1/2} c_1$. The complete algorithm for selecting a model is as follows:

Algorithm 12.1

Step 1: If $\max(|r_1|, |r_2|) < c_1$, select model $\{0\}$. Otherwise,

Step 2: If $|r_1| \geq |r_2|$, $|r_1| > c_1$ and $r_2 - rr_1 < c_2$, select model $\{x_1\}$. Otherwise,

Step 3: If $|r_2| \geq |r_1|$, $|r_2| > c_1$ and $r_1 - rr_2 < c_2$, select model $\{x_2\}$. Otherwise,

Step 4: Select model $\{x_1, x_2\}$.

Note that the model finally chosen depends on Y only through r_1 and r_2 . The inequalities above partition the r_1, r_2 plane into four disjoint regions, one for each model. We thus have a nice geometrical description of the algorithm, which is shown in Figure 12.2 for the case $r = 0.5$.

Now suppose that we use AIC as a model selection criterion. Since $\sigma^2 = 1$, we can write the AIC criterion (12.33) as

$$\text{AIC} = \text{RSS}_p + 2p.$$

By Example 3.10 in Section 3.11, the RSS for the full three-parameter model is

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \mathbf{Y}' \mathbf{X}^* \hat{\boldsymbol{\gamma}} \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{1}{1-r^2} (r_1, r_2) \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \end{aligned}$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \frac{r_1^2 - 2rr_1r_2 + r_2^2}{1 - r^2}.$$

Thus, writing $\text{SSY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$, the AIC for this model is

$$\begin{aligned}\text{AIC} &= \text{SSY} - \frac{r_1^2 - 2rr_1r_2 + r_2^2}{1 - r^2} + 6 \\ &= \text{SSY} - r^* + 6,\end{aligned}$$

where $r_* = (r_1^2 - 2rr_1r_2 + r_2^2)/(1 - r^2)$. Similarly, the AICs for the other models are

$$\begin{aligned}\text{AIC} &= \text{SSY} - r_1^2 + 4, \text{ for model } \{x_1\}, \\ \text{AIC} &= \text{SSY} - r_2^2 + 4, \text{ for model } \{x_2\}, \\ \text{AIC} &= \text{SSY} + 2, \text{ for model } \{0\}.\end{aligned}$$

After some algebra, we see that picking the model with the smallest AIC is equivalent to the following algorithm.

Algorithm 12.2

- (a) Choose $\{x_1, x_2\}$ if $r_* \geq 4$, $|rr_1 - r_2| \geq \sqrt{2(1 - r^2)}$, and $|rr_2 - r_1| \geq \sqrt{2(1 - r^2)}$.
 - (b) Choose $\{x_1\}$ if $r_1^2 \geq 2$, $r_1^2 \geq r_2^2$, and $|rr_1 - r_2| < \sqrt{2(1 - r^2)}$.
 - (c) Choose $\{x_2\}$ if $r_2^2 \geq 2$, $r_2^2 > r_1^2$, and $|rr_2 - r_1| < \sqrt{2(1 - r^2)}$.
 - (d) Choose $\{0\}$ if $r_1^2 < 2$, $r_2^2 < 2$, and $r_* < 4$.
-

We see that this rule is very similar to the forward selection algorithm if we take $c_1 = \sqrt{2}$ instead of 1.64. The geometric representation of the rule is shown in Figure 12.3. Apart from the rounded corners of the square, it is identical to Figure 12.2. \square

12.4.2 Backward Elimination

As an alternative to forward selection (FS), we can start with the full model using all K variables (provided that $K+1$ is less than n) and compute (12.43) with $p = K$ for each of the K variables. We eliminate the variable having the smallest F -statistic from the model, provided that F is less than some threshold F_{OUT} . This procedure is continued until all variables are eliminated

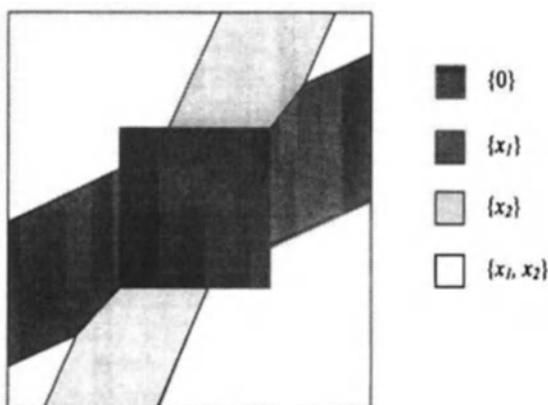


Fig. 12.2 Forward selection algorithm for two explanatory variables.

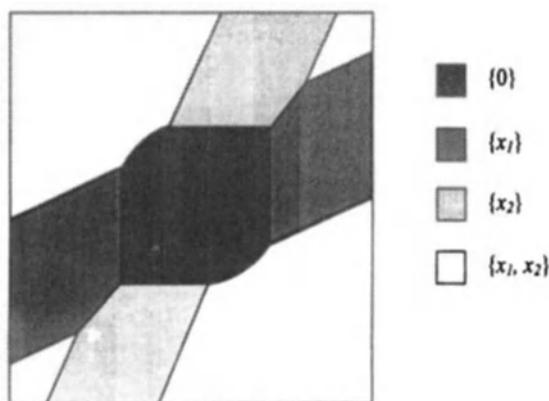


Fig. 12.3 Maximum AIC algorithm for two explanatory variables.

or the smallest F fails to be less than F_{OUT} . This procedure is called *backward elimination* (BE). It typically takes much more computation than FS when a small number of variables explain most of the variation. The thresholds F_{IN} and F_{OUT} may change as the algorithms proceed; a popular choice is to keep the formal significance level fixed, and set the thresholds to be the critical values for the corresponding F -test.

12.4.3 Stepwise Regression

A drawback of these two methods is that in the case of BE, a variable once eliminated can never be reintroduced into the regression, and in the case of FS, once included can never be removed. Also, they can give very different results on the same data (Thompson [1978]).

A method that combines FS and BE is the well-known *stepwise regression algorithm* (Efroymson [1960], Draper and Smith [1998: p. 335]), which is just FS followed by a BE step at each stage. This algorithm starts with a model consisting of the constant term alone, then performs an FS step, adding a single variable. This is followed by a BE step, removing a variable if the corresponding F is less than F_{OUT} . This combination of an FS step followed by a BE step is repeated until no further variables are introduced at the FS stage. Provided that $F_{\text{OUT}} \leq F_{\text{IN}}$, the stepwise algorithm must eventually terminate, as is shown by the following argument.

Suppose that at the beginning of an FS step, there are $p - 1$ variables in the current model, which has a residual sum of squares RSS_p . Then either no variable can be added, and the algorithm terminates, or an additional variable is added, resulting in a new residual sum of squares RSS_{p+1} say, which must satisfy

$$\frac{\text{RSS}_p - \text{RSS}_{p+1}}{\text{RSS}_{p+1}}(n - p - 1) > F_{\text{IN}}$$

or, equivalently,

$$\text{RSS}_p > \text{RSS}_{p+1} \left(1 + \frac{F_{\text{IN}}}{n - p - 1} \right) > \text{RSS}_{p+1}.$$

Now we do the BE step. Either no variable is deleted or we get a new model with $p - 1$ variables and a new RSS equal to RSS_p^* , say, where

$$\text{RSS}_p^* < \text{RSS}_{p+1} \left(1 + \frac{F_{\text{OUT}}}{n - p - 1} \right).$$

Hence

$$\text{RSS}_p^* < \text{RSS}_{p+1} \left(1 + \frac{F_{\text{OUT}}}{n - p - 1} \right) \leq \text{RSS}_{p+1} \left(1 + \frac{F_{\text{IN}}}{n - p - 1} \right) \text{RSS}_p,$$

so that at the end of each FS/BE cycle, either the algorithm has terminated or a new model is chosen whose RSS is strictly smaller. Since there are only a

finite number of models and the RSS is bounded below by zero, the algorithm must terminate.

The computations necessary to generate the sequence of models can either be based on the sweep operator discussed in Section 11.2.2 or the orthogonal decomposition methods described in Section 11.6.3. The former requires less calculation but suffers from the inaccuracy inherent in methods that form the SSCP matrix $\mathbf{X}'\mathbf{X}$ explicitly. Note that BE can be implemented as a series of Gaussian elimination steps applied to the inverse of the SSCP matrix; see Section 11.2.2 for details. No matter how the calculations are done, they will be much less computationally intensive than the all-possible-regressions methods discussed in the Section 12.4.1.

However, the stepwise methods do have some drawbacks. First, if we stop the process using an $F_{\text{IN}}/F_{\text{OUT}}$ stopping rule, only a single model is produced, whereas there may be a variety of models with similar goodness of fit that we might wish to inspect. Moreover, there is no guarantee that the chosen model will be the same as that produced by the all possible regressions (APR) methods. Berk [1978] gives an example where BE and FS agree but produce a model with an arbitrarily large difference in R^2 compared to APR methods.

Because the selection at each stage is determined by an F -test, it might be thought that the procedures find the correct subset with some specified probability. However, at each stage, the F -statistic actually used to make the decision is the maximum or minimum of a set of correlated statistics, each of which depends in complicated ways on the previous history of the procedure. Thus these F -statistics have distributions that can be very different from those considered in Chapter 4, which makes the choice of the thresholds F_{OUT} and F_{IN} somewhat arbitrary. The probability that the procedures will yield the correct subset remains unknown. Several authors (Draper et al. [1971], Kennedy and Bancroft [1971], Pope and Webster [1972]) have attempted to describe the inferential performance of isolated parts of these procedures, but no complete analysis is known.

Several variations on stepwise methods have been proposed. Broerson [1986] suggests modifying the BE step in stepwise regression by using C_p to guide the elimination of variables, and Grier [1992] uses cross-validation to compare subsets in backward elimination. Bendel and Afifi [1977] discuss alternatives to the F -statistic when using stepwise procedures.

How do stepwise methods of subset selection compare to the more computationally expensive APR methods? Both FS and BE choose a p -parameter model for each value of $p = 1, 2, \dots, K + 1$. All-possible-regression methods do the same thing, usually by selecting the p -parameter model with the smallest RSS. Obviously, the stepwise methods can do no better than the APR methods in identifying the p -parameter model with the smallest RSS. But suppose that the “correct” model has p parameters. It is quite possible that the APR method will fail to identify the correct model, but that FS (or BE) will. In stepwise methods, a good stopping rule will result in good

performance; similarly, for APR methods, a good criterion will likewise result in good performance.

EXERCISES 12c

1. Show that the first step in forward selection is equivalent to selecting the variable most highly correlated with the response.
2. Show that the variable that increases the difference $\text{RSS}_p - \text{RSS}_{p+1}$ by the greatest amount is the one having the greatest partial correlation with the response, given the variables already in the model.
3. In Example 12.2, verify Algorithm 12.2 and identify the regions corresponding to the subsets chosen by (a) backward elimination, (b) stepwise regression.

12.5 SHRINKAGE METHODS

In Sections 12.3 and 12.4 we explored ways of selecting a subset of variables in order to improve estimation of coefficients and the accuracy of predictions. When using these methods, we choose a subset of variables and then construct estimates and predictors by using least squares to fit the model selected. In this section we take a different point of view; we retain all the variables in the model but abandon the use of least squares. We shall examine the use of various “shrinkage” estimates where the least squares estimates of the regression coefficients are shrunk toward zero.

12.5.1 Stein Shrinkage

The basic idea is due to James and Stein [1961], who discuss the concept of shrinkage in the following context. Suppose that we observe \mathbf{Z} , which is assumed to have a $N_p(\mu, \sigma^2 \mathbf{I}_p)$ distribution, where $p > 2$. The obvious estimate of μ is \mathbf{Z} , which is in fact the minimum variance unbiased estimate. However, this estimate is unsatisfactory in the following sense: Its squared length $\|\mathbf{Z}\|^2$ tends to be too large, since

$$E[\|\mathbf{Z}\|^2] = \sum_{i=1}^p E[Z_i^2] \quad (12.47)$$

$$= \sum_{i=1}^p (\sigma^2 + \mu_i^2) \quad (12.48)$$

$$= p\sigma^2 + \|\mu\|^2 \quad (12.47)$$

$$> \|\mu\|^2. \quad (12.48)$$

Thus, at least some of the elements of the estimate are too large. This suggests “shrinking” the elements of \mathbf{Z} , and considering an estimate of the form $\tilde{\mu} = c\mathbf{Z}$, where $0 < c < 1$.

This estimate is biased, but it is possible to choose c so that $\tilde{\mu}$ has a smaller mean-squared error than \mathbf{Z} as an estimate of μ . Consider

$$\begin{aligned} E[||\tilde{\mu} - \mu||^2] &= \sum_{i=1}^p E[(cZ_i - \mu_i)^2] \\ &= \sum_{i=1}^p E[(c(Z_i - \mu_i) - (1 - c)\mu_i)^2] \\ &= \sum_{i=1}^p [c^2\sigma^2 + (1 - c)^2\mu_i^2] \\ &= c^2p\sigma^2 + (1 - c)^2||\mu||^2, \end{aligned} \quad (12.49)$$

which is minimized by choosing $c = ||\mu||^2/(p\sigma^2 + ||\mu||^2)$. Thus the optimal estimate can be written as

$$\tilde{\mu} = \left(1 - \frac{p\sigma^2}{p\sigma^2 + ||\mu||^2}\right) \mathbf{Z}.$$

Unfortunately, this is not a practical estimate, since it requires that we know $||\mu||$, the length of the quantity we are trying to estimate!

Suppose, however, that σ^2 is known. Then we know from (12.47) that $||\mathbf{Z}||^2$ is an unbiased estimate of $p\sigma^2 + ||\mu||^2$. This suggests using the estimate

$$\tilde{\mu} = \left(1 - \frac{p\sigma^2}{||\mathbf{Z}||^2}\right) \mathbf{Z}. \quad (12.50)$$

In fact, we can do even better than this. James and Stein [1961] showed that of all estimates of the form $(1 - b/||\mathbf{Z}||^2)\mathbf{Z}$, the best choice (in the sense of having smallest MSE) is $b = (p - 2)\sigma^2$, provided that $p > 2$. However, this is not the end of the story, since this estimate can be improved even further. We do not pursue the details here; they may be found in Efron and Morris [1973].

This James–Stein shrinkage estimate also has a nice Bayesian interpretation. Suppose that the means μ_i are i.i.d. $N(0, \sigma_0^2)$, and that, as before, conditional on the μ_i 's, the Z_i 's are independently $N(\mu_i, \sigma^2)$, where σ^2 is known. Then, using a completing the square argument as in Section 3.12, it can be shown that the posterior density of μ (i.e., the conditional density of μ given \mathbf{Z}) is $N_p((1 - \omega)\mathbf{Z}, \omega\sigma_0^2\mathbf{I}_p)$, where $\omega = \sigma^2/(\sigma^2 + \sigma_0^2)$. The Bayes estimate of μ is the posterior mean $(1 - \omega)\mathbf{Z}$.

Assuming that σ^2 is known, the Bayesian approach requires that we specify a value for σ_0 . For an alternative approach, note that (cf. Exercises 12d, No. 1) the marginal distribution of \mathbf{Z} is $N_p(\mathbf{0}, (\sigma^2 + \sigma_0^2)\mathbf{I}_p)$, so by Theorem 2.9,

$$||\mathbf{Z}||^2/(\sigma^2 + \sigma_0^2)$$

has a χ_p^2 distribution, and by (12.32),

$$E \left[\frac{\sigma^2 + \sigma_0^2}{||\mathbf{Z}||^2} \right] = \frac{1}{p-2}.$$

It follows that $(p-2)/||\mathbf{Z}||^2$ is an unbiased estimate of $1/(\sigma^2 + \sigma_0^2)$, which leads to the estimate $(1 - (p-2)\sigma^2/||\mathbf{Z}||^2)\mathbf{Z}$ of μ .

An estimate based on the posterior mean, but using the marginal distribution of the observed data to estimate the parameters of the prior distribution, is called an *empirical Bayes estimate*. We have demonstrated that the James–Stein estimate is just the empirical Bayes estimate based on the $N(0, \sigma_0^2)$ prior for the μ_i 's. More details may be found in Efron and Morris [1973]. Another clever motivation for the James–Stein estimate is given in Stigler [1990].

In the case of a regression with p orthonormal explanatory variables, known σ^2 , and no constant term, we can apply the James–Stein estimate directly by setting $\mathbf{Z} = \hat{\beta}$ and $\mu = \beta$. In the orthonormal case $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the least squares estimate $\hat{\beta}$ has an $N_p(\beta, \sigma^2 \mathbf{I}_p)$ distribution, so that the shrunken estimate

$$\tilde{\beta} = \left(1 - \frac{(p-2)\sigma^2}{||\hat{\beta}||^2} \right) \hat{\beta}$$

has the smallest MSE of any estimate of the form $(1 - c)\hat{\beta}$. If σ^2 is unknown, we can use the usual estimate S^2 for σ^2 , at the cost of losing the optimality property.

If the explanatory variables are not orthogonal, we can use the characterization of the James–Stein estimate as an empirical Bayes estimate to guide us in the choice of an estimate. Let us assume the setup of Section 3.12, where conditional on β , \mathbf{Y} is $N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ and the prior distribution of β is $N_p(\mathbf{m}, \sigma^2 \mathbf{V})$. Then we know from Section 3.12 that the posterior mean of β is of the form $\mathbf{m}_* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m})$, and this result holds if σ^2 is known or not. If we take $\mathbf{V} = \tau^{-1} \mathbf{I}_p$ and $\mathbf{m} = \mathbf{0}$, then the Bayes estimate is

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + \tau \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y},$$

which is the same form as the ridge estimate introduced in Section 10.7.3 and considered in more detail in Section 12.5.2.

If we take $\mathbf{V} = \tau^{-1}(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{m} = \mathbf{0}$, we get the estimate

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + \tau \mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = (1 + \tau)^{-1} \hat{\beta}.$$

This is often called the *James–Stein regression estimate*. We shall not consider it further, but an examination of its properties may be found in Dempster et al. [1977].

12.5.2 Ridge Regression

Ridge regression was introduced in Section 10.7.3 as a means for improving the estimation of regression coefficients when the predictors are highly correlated. Ridge methods are also effective as a means of improving the accuracy of predictions. In the previous discussion, the regression model was written in centered and scaled form. For notational convenience, we do not consider an explicitly centered and scaled model in this section, although the results hold true if \mathbf{X} is interpreted in this way. With this proviso, the ridge estimate is

$$\hat{\beta}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y},$$

which depends on the ridge parameter k , assumed to be nonnegative and not necessarily an integer. We have seen in Section 10.7.3 how the ridge estimate arises naturally as the Bayes estimate, assuming a particular prior on β . In this section we take a non-Bayesian perspective and treat k as a parameter whose value must be chosen in some way.

The ridge estimate can be written as

$$\begin{aligned}\hat{\beta}(k) &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \{\mathbf{X}'\mathbf{X}[\mathbf{I} + k(\mathbf{X}'\mathbf{X})^{-1}]\}^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= [\mathbf{I} + k(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\hat{\beta} \\ &= \mathbf{C}\hat{\beta},\end{aligned}$$

say. If we assume that $E[\mathbf{Y}] = \mathbf{X}\beta$ and use $\mathbf{X}\hat{\beta}(k)$ as a predictor of $\mathbf{X}\beta$, the expected ME is

$$\begin{aligned}E[\text{ME}] &= E\|\mathbf{X}\hat{\beta}(k) - \mathbf{X}\beta\|^2 \\ &= E\|\mathbf{X}(\mathbf{C}\hat{\beta} - \beta)\|^2 \\ &= E\|\mathbf{X}\mathbf{C}(\hat{\beta} - \beta) + \mathbf{X}(\mathbf{C} - \mathbf{I}_p)\beta\|^2 \\ &= E\|\mathbf{X}\mathbf{C}(\hat{\beta} - \beta)\|^2 + \|\mathbf{X}(\mathbf{C} - \mathbf{I}_p)\beta\|^2.\end{aligned}\quad (12.51)$$

Now let $\mathbf{X}'\mathbf{X} = \mathbf{T}\Lambda\mathbf{T}'$, where T is orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$. Setting $\alpha = \mathbf{T}'\beta$ and using the arguments of Section 10.7.3, we have

$$E[\text{ME}] = \sum_{j=1}^p \frac{\alpha_j^2 k^2 \lambda_j + \sigma^2 \lambda_j^2}{(k + \lambda_j)^2}.\quad (12.52)$$

Note the similarity with (10.66); considering the ME rather than the mean-squared error merely replaces λ_j by λ_j^2 . The derivative of (12.52) with respect to k is

$$\sum_{j=1}^p \frac{2\lambda_j^2(\alpha_j^2 k - \sigma^2)}{(k + \lambda_j)^3},$$

which is negative for small positive values of k , so there is a value of k for which the expected ME using $\mathbf{X}\hat{\beta}(k)$ is smaller than that using the predictor $\mathbf{X}\hat{\beta}(0) = \mathbf{X}\hat{\beta}$ based on the least squares estimate.

In the case of orthonormal predictor variables, we have $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$ and all the eigenvalues are unity. In this case the derivative is

$$\frac{2(kB - p\sigma^2)}{(1 + k)^3},$$

where $B = \sum_j \beta_j^2 = \sum_j \alpha_j^2$, and the minimum occurs at $k = p\sigma^2/B$.

In the general case, we need a reliable method of estimating the optimal value of k if ridge is to be superior to least squares. Note that the formulas above apply only for fixed k , not when k is data-selected. Popular methods of estimating k include cross-validation and generalized cross-validation (Golub et al. [1979]). As for subset selection, an intuitively appealing estimate of the PE of the ridge predictor is

$$\frac{1}{n} \sum_{i=1}^n [Y_i - \mathbf{x}'_i \hat{\beta}_{(-i)}(k)]^2,$$

where $\hat{\beta}_{(-i)}(k)$ is the ridge estimate calculated using ridge coefficient k but leaving out the data from the i th case. We can avoid repeatedly calculating the ridge estimates by writing this estimate as

$$\frac{1}{n} \sum_{i=1}^n \frac{[Y_i - \mathbf{x}'_i \hat{\beta}(k)]^2}{[1 - a_{ii}(k)]^2}, \quad (12.53)$$

where $a_{ii}(k)$ is the i th diagonal element of $\mathbf{A}(k) = \mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'$. The value of k is then chosen to minimize this criterion.

A variation on this is *generalized cross-validation* (GCV). Here the elements $a_{ii}(k)$ are replaced by their average $n^{-1} \text{tr } \mathbf{A}(k)$. Golub et al. [1979] discuss the relative merits of these two methods of selecting k . A more modern method is the little bootstrap, discussed in Section 12.9.

The relative merits of ridge regression versus least squares and subset selection have been endlessly debated. While we know that there is a k for which the ridge estimate has smaller MSE than least squares, no known method of estimating this k from the data will guarantee this for the chosen k .

We will see in Section 12.9 that the effectiveness of ridge depends in part on the distribution of the unknown true regression coefficients. We saw above that in the case of independent and identically distributed normal β_j 's, the ridge estimate is the Bayes estimate and is optimal in the sense of mean-squared error. In other cases, ridge will not usually be optimal, but can still be better than least squares. Many simulation studies have compared ridge to least squares; see Golub et al. [1979] and the references in Section 10.7.3 for details.

The ridge estimate can be regarded as the solution of a constrained least squares problem. Consider the problem of minimizing the sum of squares

$$\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2 \quad (12.54)$$

subject to the constraint $\sum b_j^2 \leq s$, where s is some specified constant. If $\sum_j \hat{\beta}_j^2 \leq s$, the least squares estimate solves this constrained problem. However, if $s < \sum_j \hat{\beta}_j^2$, the solution is different and is given by the solution to the Lagrangian equations (cf. Fletcher [1987], Lawson and Hanson [1995])

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{Y} + \lambda\mathbf{b} &= \mathbf{0}, \\ \sum_j b_j^2 &= s, \end{aligned}$$

which have a solution of the form

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y},$$

where λ is chosen to satisfy $\sum_j \hat{\beta}_j(\lambda)^2 = s$. Note that in the case when the explanatory variables are orthonormal, so that $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the ridge estimate is

$$\hat{\boldsymbol{\beta}}(\lambda) = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}},$$

and therefore each coefficient is shrunk by a constant factor.

By using other types of constraint, we can get a variety of alternative shrinkage estimates, which we now discuss.

12.5.3 Garrote and Lasso Estimates

Garrote Estimate

One drawback to the ridge approach is that unlike subset selection, it retains all variables in the model, thus sacrificing the possibility of a simple model with fewer variables. An alternative to ridge that also can lead to simpler models is the *garrote*, an intriguingly named technique introduced by Breiman [1995]. In the garrote, the individual least squares coefficients $\hat{\beta}_j$ are shrunk by a nonnegative quantity c_j , leading to garrote estimates $\tilde{\beta}_j = c_j \hat{\beta}_j$. The shrinkage factors are chosen to minimize the least squares criterion

$$\|\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=0}^{p-1} x_{ij} c_j \hat{\beta}_j \right)^2, \quad (12.55)$$

subject to the constraints $c_j \geq 0, j = 0, \dots, p-1$, and $\sum_{j=0}^{p-1} c_j \leq s$, where s is some specified positive constant. For $s > p$, the choice $c_j = 1, j = 0, \dots, p-1$, yields $\tilde{\beta}_j = \hat{\beta}_j$, which gives the unconstrained minimum. As s is reduced and

the garrote is drawn tighter, the shrinkage coefficients get smaller and some are even forced to zero. Thus the garrote can be regarded as a compromise between ridge (which shrinks but zeros no coefficients) and subset selection (which zeros coefficients but does not shrink).

The problem of finding the shrinkage coefficients c_j can be reduced to a standard problem in constrained least squares. We can write

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\tilde{\beta}\|^2 &= \|(\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta})\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta}\|^2 + 2(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta}) \\ &= \text{RSS} + \|\mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta}\|^2, \end{aligned} \quad (12.56)$$

since

$$(\mathbf{Y} - \mathbf{X}\hat{\beta})'\mathbf{X} = \mathbf{Y}'\mathbf{X} - \hat{\beta}'\mathbf{X}'\mathbf{X} = 0.$$

Let $a_{ij} = x_{ij}\hat{\beta}_j$, $\mathbf{A} = (a_{ij})$ and $\mathbf{c}' = (c_0, \dots, c_{p-1})$. Then $\mathbf{X}\tilde{\beta} = \mathbf{Ac}$, so from (12.56) we get

$$\|\mathbf{Y} - \mathbf{X}\tilde{\beta}\|^2 = \text{RSS} + \|\mathbf{d} - \mathbf{Ac}\|^2,$$

where $\mathbf{d} = \mathbf{X}\hat{\beta}$. We see that the vector \mathbf{c} which minimizes (12.55) subject to the constraints $c_j \geq 0$, $j = 0, \dots, p-1$ and $\sum_{j=0}^{p-1} c_j \leq s$, is the same as the vector \mathbf{c} which minimizes

$$\|\mathbf{Ac} - \mathbf{d}\|^2 \quad (12.57)$$

subject to the same constraints. The constrained minimization of (12.57) is a standard problem for which efficient algorithms are known. A full discussion may be found in Lawson and Hanson [1995: Chapter 23], and Björck [1996: p. 194].

When the explanatory variables are orthonormal, the solution of the minimization problem can be written explicitly; in this case the shrinkage coefficients are (Breiman [1995])

$$c_j = \begin{cases} 1 - \lambda/\hat{\beta}_j^2, & \lambda \leq \hat{\beta}_j^2, \\ 0, & \text{otherwise,} \end{cases}$$

where λ is determined by the equation $\sum_j c_j = s$.

As in ridge regression, we must choose the value of s , then compute the c_j 's by solving the minimization problem (12.57). The little bootstrap can be used to select s ; further details are given in Section 12.9.

A variation on the garrote (Breiman [1995, 1996b]) drops the requirement that the shrinkage coefficients be nonnegative and requires instead that they satisfy the constraint $\sum_j \beta_j^2 \leq s$ used in ridge. This gives a very similar result to ridge, except that the large least squares coefficients are not shrunk as much as the small ones. Breiman calls the previous version of the garrote the *nonnegative garrote* (nn-garrote) to distinguish it from this new version. Using a Lagrange multiplier argument very similar to that used in ridge, we can show that the shrinkage coefficients c_j are now given by

$$\mathbf{c} = (\mathbf{A}'\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}'\mathbf{d}, \quad (12.58)$$

where λ is determined by the constraint $\sum_j c_j^2 = s$. When the explanatory variables are orthonormal,

$$c_j = \frac{\hat{\beta}_j^2}{\lambda + \hat{\beta}_j^2},$$

so the large coefficients are shrunk proportionally less than the small ones. We note that like ridge, this version of the garrote does not zero any coefficients.

Lasso Estimate

Yet another shrinkage estimate is the *lasso* (Tibshirani [1996]). This estimate is the same as ridge but uses a different constraint; the criterion (12.55) is now minimized subject to the constraint $\sum_j |\beta_j| \leq s$ rather than the ridge constraint $\sum_j \beta_j^2 \leq s$. Like the nn-garrote, the lasso can zero some coefficients. Tibshirani discusses an efficient algorithm for finding the lasso estimates. In the orthonormal case, they can be written as

$$\bar{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda), & \lambda \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise,} \end{cases}$$

where λ is chosen to satisfy the constraint $\sum_j |\bar{\beta}_j| \leq s$.

In the orthonormal case, all the shrinkage methods replace the least squares coefficients $\hat{\beta}_j$ by shrunken versions $h(\hat{\beta}_j)$. The various functions h are graphed in Figure 12.4.

We note that selection based on thresholding, as described in Example 12.1, can be regarded as an extreme form of shrinkage, using the function h given by

$$h(\hat{\beta}_j) = \begin{cases} \hat{\beta}_j, & \lambda \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise.} \end{cases}$$

Which of these functions gives the best predictor? A Bayesian answer to this question is given in Section 12.9.

EXERCISES 12d

- Verify that the marginal distribution of \mathbf{Y} in the Bayesian setup of Section 12.5.1 is $N_p(\mathbf{0}, (\sigma^2 + \sigma_0^2)\mathbf{I}_p)$.
- Show that when $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, then the quantity c_j in the second version of the garrote is given by

$$c_j = \frac{\hat{\beta}_j^2}{\lambda + \hat{\beta}_j^2}.$$

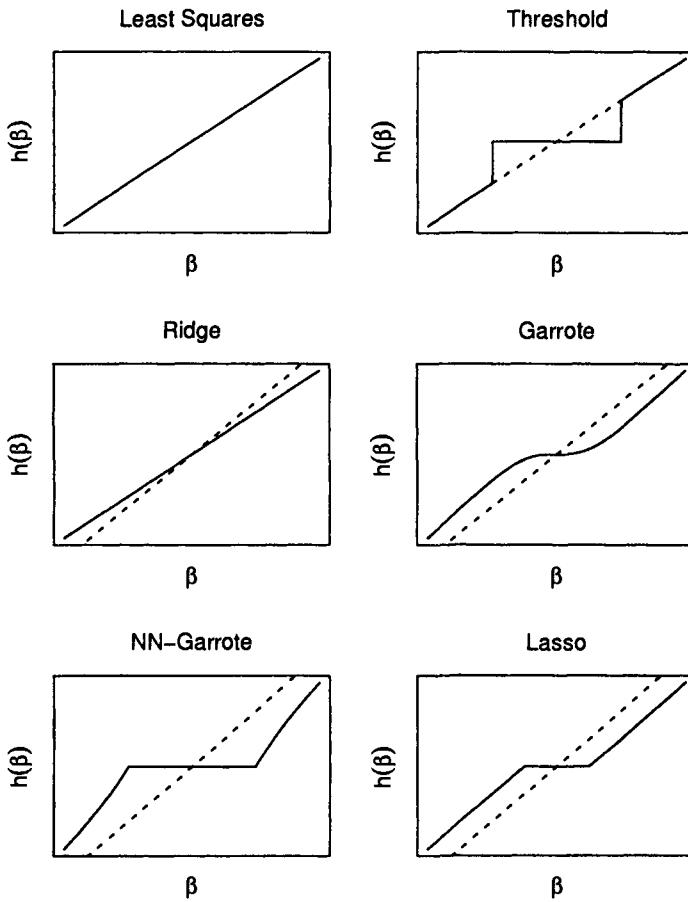


Fig. 12.4 Functions giving shrinkage estimates in the orthonormal case. The shrinkage estimate $\tilde{\beta}_j$ is $h(\hat{\beta}_j)$.

12.6 BAYESIAN METHODS

12.6.1 Predictive Densities

In contrast to the frequentist methods we have described earlier in this chapter, Bayesian methods have an appealing simplicity, as we just select the model having the greatest posterior probability. Suppose that we have m models $\mathcal{M}_1, \dots, \mathcal{M}_m$ with prior probabilities $\omega_1, \dots, \omega_m$. Conditional on \mathcal{M}_j , suppose that we have a model $f_j(y|\theta_j)$ which describes the distribution of the observations, conditional on the parameter vector θ_j . Assuming that for each

model we can specify a prior distribution $\pi_j(\boldsymbol{\theta}_j)$, we then have a *predictive density* for model \mathcal{M}_j , given by

$$f_j(\mathbf{y}|\mathcal{M}_j) = \int f_j(\mathbf{y}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j. \quad (12.59)$$

Using Bayes' theorem, we can calculate the posterior probability of \mathcal{M}_j being the true model as

$$P(\mathcal{M}_j|\mathbf{y}) = \frac{f_j(\mathbf{y}|\mathcal{M}_j)\omega_j}{\sum_l f_l(\mathbf{y}|\mathcal{M}_l)\omega_l}, \quad (12.60)$$

where the sum is taken over all m models under consideration. To select the model that is most probable a priori, we choose that model for which $P(\mathcal{M}_j|\mathbf{y})$ is a maximum. This is equivalent to choosing the model for which $f_j(\mathbf{y}|\mathcal{M}_j)\omega_j$ is a maximum, or if all models are equally likely a priori, to choosing the model with the largest predictive density, evaluated at the observations \mathbf{y} .

In the case of two models, by Bayes' theorem we have

$$\frac{P(\mathcal{M}_1|\mathbf{y})}{P(\mathcal{M}_2|\mathbf{y})} = \frac{P(\mathbf{y}|\mathcal{M}_1)\omega_1}{P(\mathbf{y}|\mathcal{M}_2)\omega_2}, \quad (12.61)$$

and if the two models are equally likely a priori, we have

$$\frac{P(\mathcal{M}_1|\mathbf{y})}{P(\mathcal{M}_2|\mathbf{y})} = \frac{P(\mathbf{y}|\mathcal{M}_1)}{P(\mathbf{y}|\mathcal{M}_2)}. \quad (12.62)$$

The quantity on the right of (12.62) is called the *Bayes factor*. Model 1 is preferred over model 2 if the Bayes factor is sufficiently large. Kass and Raftery [1995] suggest that a Bayes factor in excess of 3 should be regarded as positive evidence of model 1 over model 2, and Bayes factors in excess of 20 be regarded as strong evidence.

In the case of equal priors $\pi_j(\boldsymbol{\theta}_j)$, the predictive density behaves very much as a likelihood, except that it is obtained by integrating out parameters rather than by maximization. One crucial difference is that the likelihood ratio requires that model 1 be nested within model 2 to make sense, but the Bayes factor does not.

There are difficulties with the definition of Bayes factors when the priors are improper, since in this case they are defined only up to arbitrary constants. Atkinson [1978] objects to their use on these and other grounds. Several authors (Berger and Pericchi [1996], O'Hagen [1995], Spiegelhalter and Smith [1982]) have proposed solutions to this problem, but these are somewhat controversial (cf. the discussion in the O'Hagen paper).

Gelfand and Dey [1994] point out that even with proper priors, Bayes factors lead to *Lindley's paradox*. Suppose that we are comparing two models \mathcal{M}_1 and \mathcal{M}_2 , with \mathcal{M}_1 a submodel of \mathcal{M}_2 . As the amount of data increases, the Bayes factor leads increasingly to a preference for model \mathcal{M}_1 over \mathcal{M}_2 , no matter what the data. This is in contrast to the *F*-test, which will increasingly

prefer \mathcal{M}_2 over \mathcal{M}_1 . They also point out that using the BIC can sometimes be misleading, since the $O(1)$ terms that are omitted (cf. Section 12.3.4) can be quite large.

In the Bayesian approach, priors must be specified for both the model and the model parameters for every model under consideration. Methods for doing this are discussed by Laud and Ibrahim [1996], Draper [1995], and Garthwaite and Dickey [1992].

In general, calculating the predictive densities, and hence the Bayes factor, requires evaluating the integral in (12.59), which will be high-dimensional for most interesting models. Conventional numerical integration techniques are not much help in this situation, but recently, specialized methods have been introduced to cope with the problem. In particular, the Laplace approximation, importance sampling, and Markov chain Monte Carlo have all proved effective. Evans and Swartz [1995] give a review.

Markov chain Monte Carlo (MCMC) methods have been used by Carlin and Chib [1995] to calculate posterior model probabilities and by George and McCulloch [1993] to calculate posterior distributions for parameters. George and McCulloch assume the usual normal model, conditional on the regression coefficients β_j . The prior on β_j is then modeled as a mixture of two normals, one with very small variance and mean zero, and MCMC methods are used to obtain the posterior distributions of the mixture proportions. Coefficients for which the mixture proportion for the low variance, zero-mean component is close to 1 correspond to variables that should be deleted from the model, giving a nice Bayesian method of subset selection. Mitchell and Beauchamp [1988] describe a similar idea.

In the case of linear models, the conditional density $f_j(\mathbf{y}|\boldsymbol{\theta}_j)$ of Y , given the model \mathcal{M}_j and the parameters $\boldsymbol{\theta}_j$, is taken to be $N_n(\mathbf{X}_j\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where $\boldsymbol{\theta}_j = (\boldsymbol{\beta}'_j, \sigma)^t$. We can choose priors for $\boldsymbol{\theta}_j$ that permit the analytical evaluation of the predictive density (12.59). As in Section 3.12, we will assume that the prior on $\boldsymbol{\beta}_j$ conditional on σ^2 is $N_k(\mathbf{m}, \sigma^2\mathbf{V})$, and that the prior on σ^2 is inverse gamma with density (3.63). Using the results and notation of Section 3.12 and dropping the subscript j to simplify the notation, we see that the joint density is

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(d+n+p+2)/2} \exp[-(Q + a)/2\sigma^2].$$

Integrating this over σ^2 (cf. Section 3.12) gives

$$\begin{aligned} & (Q + a)^{-(d+n+p)/2} \\ &= [a_* + (\boldsymbol{\beta} - \mathbf{m}_*)' \mathbf{V}_*^{-1} (\boldsymbol{\beta} - \mathbf{m}_*)]^{-(d+n+p)/2} \\ &= a_*^{-(d+n+p)/2} \\ &\quad \times [1 + (d+n)^{-1} (\boldsymbol{\beta} - \mathbf{m}_*) [a_* \mathbf{V}_*/(d+n)]^{-1} (\boldsymbol{\beta} - \mathbf{m}_*)]^{-(d+n+p)/2}. \end{aligned}$$

Integrating again over β using A.13.5 gives, up to a constant not involving y ,

$$\begin{aligned} f(y|\mathcal{M}) &\propto a_*^{-(d+n+p)/2} \det[a_* \mathbf{V}_*/(d+n)]^{1/2} \\ &= a_*^{-(d+n+p)/2} a_*^{p/2} \det[\mathbf{V}_*/(d+n)]^{1/2} \\ &\propto a_*^{-(d+n)/2} \\ &= [a + (y - \mathbf{Xm})' \mathbf{W}^{-1} (y - \mathbf{Xm})]^{-(d+n)/2} \quad (12.63) \\ &\propto [1 + d^{-1} (y - \mathbf{Xm})' (a \mathbf{W}/d)^{-1} (y - \mathbf{Xm})]^{-(d+n)/2}, \end{aligned}$$

where $\mathbf{W} = \mathbf{I}_n + \mathbf{XVX}'$. Hence by A.13.5, the predictive density $f(y|\mathcal{M})$ is a multivariate t , namely, $t_n(d, \mathbf{Xm}, a\mathbf{W}/d)$.

12.6.2 Bayesian Prediction

If we observe an n -dimensional response vector \mathbf{Y} , and the associated $n \times p$ regression matrix \mathbf{X} , suppose that we want to use this training set (\mathbf{X}, \mathbf{Y}) to predict the m -vector of responses \mathbf{Y}_0 corresponding to a new $m \times p$ matrix \mathbf{X}_0 . The Bayesian solution to this problem is to calculate the *posterior predictive density*, the conditional density of \mathbf{Y}_0 given \mathbf{Y} , and use the modal value of this density as the predictor.

Assuming the same priors as before, we have the following result.

THEOREM 12.1 *With the priors and notation of the Section 12.6.1, the posterior predictive distribution (i.e., the conditional distribution of \mathbf{Y}_0 given $\mathbf{Y} = y$) is multivariate t , $t_m[d + n, \mathbf{X}_0 \mathbf{m}_*, a_0(\mathbf{I}_m + \mathbf{X}_0 \mathbf{V}_* \mathbf{X}_0')/(d + n)]$, where*

$$a_0 = a + (y - \mathbf{Xm})' (\mathbf{I}_n + \mathbf{XVX}')^{-1} (y - \mathbf{Xm}). \quad (12.64)$$

Proof. Let $\mathbf{X}'_c = (\mathbf{X}'_0, \mathbf{X}')$, $\mathbf{Y}'_c = (\mathbf{Y}'_0, \mathbf{Y}')$, and $\mathbf{W}_c = \mathbf{I}_{m+n} + \mathbf{X}_c \mathbf{V} \mathbf{X}'_c$. Then, by (12.63), the predictive density of \mathbf{Y}_c is proportional to

$$[a + (y_c - \mathbf{X}_c \mathbf{m})' \mathbf{W}_c^{-1} (y_c - \mathbf{X}_c \mathbf{m})]^{-(d+m+n)/2}.$$

Now partition the matrix \mathbf{W}_c as

$$\mathbf{W}_c = \left[\begin{array}{cc} \mathbf{I}_m + \mathbf{X}_0 \mathbf{V} \mathbf{X}'_0 & \mathbf{X}_0 \mathbf{V} \mathbf{X}' \\ \mathbf{X} \mathbf{V} \mathbf{X}'_0 & \mathbf{I}_n + \mathbf{X} \mathbf{V} \mathbf{X}' \end{array} \right] = \left[\begin{array}{cc} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{array} \right], \quad \text{say.}$$

By using the partitioned matrix inverse formula A.9.1 and completing the square (see Exercises 12e, No. 1, at the end of Section 12.6.3), we can write

$$\begin{aligned} (\mathbf{y}_c - \mathbf{X}_c \mathbf{m})' \mathbf{W}_c^{-1} (\mathbf{y}_c - \mathbf{X}_c \mathbf{m}) &= (\mathbf{y} - \mathbf{Xm})' \mathbf{W}_{22}^{-1} (\mathbf{y} - \mathbf{Xm}) \\ &\quad + (\mathbf{y}_0 - \boldsymbol{\mu}_0)' \mathbf{W}_{12}^{-1} (\mathbf{y}_0 - \boldsymbol{\mu}_0), \quad (12.65) \end{aligned}$$

where

$$\boldsymbol{\mu}_0 = \mathbf{X}_0 \mathbf{m} + \mathbf{W}_{12} \mathbf{W}_{22}^{-1} (\mathbf{y} - \mathbf{Xm})$$

and

$$\mathbf{W}_{1.2} = \mathbf{W}_{11} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{W}_{21}.$$

Since the first term on the right-hand side of (12.65) does not involve \mathbf{y}_0 , the conditional density of \mathbf{Y}_0 given $\mathbf{Y} = \mathbf{y}$, obtained by dividing the joint density of \mathbf{Y}_0 and \mathbf{Y} by the marginal density of \mathbf{Y} , is proportional to

$$[a_0 + (\mathbf{y}_0 - \boldsymbol{\mu}_0)' \mathbf{W}_{1.2}^{-1} (\mathbf{y}_0 - \boldsymbol{\mu}_0)]^{-(d+n+m)/2},$$

where a_0 is given by (12.64).

It follows from these results that the posterior predictive density of \mathbf{Y}_0 is multivariate t , $t_m[d+n, \boldsymbol{\mu}_0, a_0 \mathbf{W}_{1.2}/(d+n)]$. Thus, to complete the proof, we need only show that

$$\boldsymbol{\mu}_0 = \mathbf{X}_0 \mathbf{m}_* \quad (12.66)$$

and

$$\mathbf{W}_{1.2} = \mathbf{I}_m + \mathbf{X}_0 \mathbf{V}_* \mathbf{X}'_0, \quad (12.67)$$

where $\mathbf{V}_* = (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}$. To prove (12.66), we use the relationship $\mathbf{m}_* = \mathbf{V}_*(\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m})$ and write

$$\begin{aligned} \mathbf{X}'\mathbf{y} &= \mathbf{V}_*^{-1}\mathbf{m}_* - \mathbf{V}^{-1}\mathbf{m} \\ &= (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})\mathbf{m}_* - \mathbf{V}^{-1}\mathbf{m} \\ &= \mathbf{V}^{-1}(\mathbf{m}_* - \mathbf{m}) + \mathbf{X}'\mathbf{X}\mathbf{m}_*, \end{aligned}$$

so that

$$\begin{aligned} (\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{m}_*) &= \mathbf{y} - \mathbf{X}\mathbf{m}_* + \mathbf{X}\mathbf{V}\mathbf{X}'\mathbf{y} - \mathbf{X}\mathbf{V}\mathbf{X}'\mathbf{X}\mathbf{m}_* \\ &= \mathbf{y} - \mathbf{X}\mathbf{m}_* + \mathbf{X}\mathbf{V}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{m}_*) \\ &= \mathbf{y} - \mathbf{X}\mathbf{m}_* + \mathbf{X}(\mathbf{m}_* - \mathbf{m}) \\ &= \mathbf{y} - \mathbf{X}\mathbf{m}. \end{aligned}$$

Thus

$$(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m}) = \mathbf{y} - \mathbf{X}\mathbf{m}_*$$

and hence

$$\begin{aligned} \boldsymbol{\mu}_0 &= \mathbf{X}_0 \mathbf{m} + \mathbf{X}_0 \mathbf{V} \mathbf{X}' (\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1} (\mathbf{y} - \mathbf{X}\mathbf{m}) \\ &= \mathbf{X}_0 \mathbf{m} + \mathbf{X}_0 \mathbf{V} (\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{m}_*) \\ &= \mathbf{X}_0 (\mathbf{m} + \mathbf{m}_* - \mathbf{m}) \\ &= \mathbf{X}_0 \mathbf{m}_*, \end{aligned}$$

which proves (12.66).

To prove (12.67), we have

$$\begin{aligned} \mathbf{W}_{1.2} &= \mathbf{W}_{11} - \mathbf{W}_{12}\mathbf{W}_{22}^{-1}\mathbf{W}_{21} \\ &= \mathbf{I}_m + \mathbf{X}_0 \mathbf{V} \mathbf{X}'_0 - \mathbf{X}_0 \mathbf{V} \mathbf{X}' (\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1} \mathbf{X} \mathbf{V} \mathbf{X}'_0 \\ &= \mathbf{I}_m + \mathbf{X}_0 [\mathbf{V} - \mathbf{V} \mathbf{X}' (\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1} \mathbf{X} \mathbf{V}] \mathbf{X}'_0 \\ &= \mathbf{I}_m + \mathbf{X}_0 (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_0 \quad [\text{by A.9.3}] \\ &= \mathbf{I}_m + \mathbf{X}_0 \mathbf{V}_* \mathbf{X}'_0, \end{aligned}$$

which proves (12.67). The proof is complete. \square

Since the $t_m[d+n, \mathbf{X}\mathbf{m}_*, a_0(\mathbf{I} + \mathbf{X}_0\mathbf{V}\mathbf{X}'_0)/(d+n)]$ density has its maximum value at $\mathbf{y}_0 = \mathbf{X}_0\mathbf{m}_*$, the Bayesian predictor is $\mathbf{X}_0\mathbf{m}_*$.

12.6.3 Bayesian Model Averaging

The predictor described in Section 12.6.2 is the modal value of the posterior predictive density, and is conditional on the model chosen. We may obtain a predictor whose distribution is not conditional on a particular model by averaging over the models $\mathcal{M}_1, \dots, \mathcal{M}_m$ under consideration. As in Section 12.6.1, suppose that model \mathcal{M}_j has prior probability ω_j . We can obtain a posterior density $f(\mathbf{y}_0|\mathbf{y})$ that is not conditional on the model chosen by averaging over models. From the theorem of total probability we get

$$\begin{aligned} f(\mathbf{y}_0|\mathbf{y}) &= \frac{f(\mathbf{y}_0, \mathbf{y})}{f(\mathbf{y})} \\ &= \frac{\sum_{j=1}^m f(\mathbf{y}_0|\mathbf{y}, \mathcal{M}_j) f(\mathbf{y}|\mathcal{M}_j) \omega_j}{\sum_{j=1}^m f(\mathbf{y}|\mathcal{M}_j) \omega_j}, \end{aligned} \quad (12.68)$$

where the conditional densities $f(\mathbf{y}_0|\mathbf{y}, \mathcal{M}_j)$ and $f(\mathbf{y}|\mathcal{M}_j)$ are the posterior and prior predictive densities discussed in Section 12.6.2.

The main drawback of this approach is the amount of computation required. We must calculate the multivariate t density for each model, which will be too computationally expensive if the number of models is large. Raftery et al. [1997] offer two solutions. The first involves calculating the prior predictive densities, but avoids calculating the posteriors. The term corresponding to model j is deleted from both summations in (12.68) if (1) the prior predictive density satisfies $f(\mathbf{y}|\mathcal{M}_j)\omega_j < C \max_k f(\mathbf{y}|\mathcal{M}_k)\omega_k$ for some constant C , or, (2) there is a model $\mathcal{M}_{j'}$ contained in M_j for which $(\mathbf{y}|\mathcal{M}_j)\omega_j < f(\mathbf{y}|\mathcal{M}_{j'})\omega_{j'}$.

Raftery et al. [1997] term the first criterion *Occam's window*; it removes models that explain the data far less well than the best predicting model. The second criterion is a variation of Occam's razor; a simpler model that predicts as well as a more complex model is preferred over the complex model. They recommend a value of C of around 20, which typically reduces the number of terms in (12.68) to 25 or less. The second method described by Raftery et al. [1997] is based on MCMC techniques, which are used to approximate $f(\mathbf{y}_0|\mathbf{y})$; details may be found in their paper.

A non-Bayesian approach to model averaging has been suggested by Buckland et al. [1997] and by Burnham and Anderson [1998]. Breiman [1996a] discusses an approach to constructing weighted combinations of predictors.

EXERCISES 12e

1. Verify (12.65). *Hint:* Use the partitioned matrix inverse formula A.9.1 and multiply out the quadratic form.

2. Find the form of the posterior predictive density for straight-line regression. Assume that the explanatory variable has been centered and that the prior mean is zero. Also assume that you are predicting a single value (i.e., assume that $m = 1$).

12.7 EFFECT OF MODEL SELECTION ON INFERENCE

In regression analyses it is common practice to use the same set of data to both select and fit the model. The precision of the resulting parameter estimates is often uncritically assessed by quoting the standard errors reported by whatever statistical package has been used. These standard errors are based on the assumption that the model has been selected *a priori*, without reference to the data. If this is not the case, then the standard errors reported by the package are typically incorrect.

A related problem is the assessment of prediction error when the predictors are based on data-selected models. We deal with this in Section 12.9.

12.7.1 Conditional and Unconditional Distributions

Suppose that we have K possible explanatory variables, so that there are 2^K possible models, assuming that a constant term is always present. We want to estimate the regression coefficients $\beta = (\beta_0, \beta_1, \dots, \beta_K)'$ corresponding to the full model by first using a subset selection procedure and then fitting the selected model by least squares. Implicitly, we are estimating the coefficients of the nonselected variables by zero. Denote the resulting estimate by $\tilde{\beta}$, and the estimate that results from selecting a particular model \mathcal{M} by $\hat{\beta}(\mathcal{M})$.

Any variable selection procedure can be thought of as a partition of the set of all observation vectors \mathbf{Y} into $M = 2^K$ disjoint regions $\mathcal{R}_1, \dots, \mathcal{R}_M$, so that model \mathcal{M}_m is selected if and only if $\mathbf{Y} \in \mathcal{R}_m$. With this notation, we see that $\tilde{\beta} = \hat{\beta}(\mathcal{M}_m)$ if and only if $\mathbf{Y} \in \mathcal{R}_m$.

Three distinct distributions of these estimates are of interest here. The first is the unconditional distribution of $\tilde{\beta}$, which describes the overall performance of this estimation strategy. Second, if a particular model \mathcal{M} corresponding to region \mathcal{R} has been selected, the relevant distribution for inference is the conditional distribution of $\hat{\beta}(\mathcal{M})$ given $\mathbf{Y} \in \mathcal{R}$, and the standard error is based on the variance of this conditional distribution. Finally, if the model \mathcal{M} has been chosen *a priori* without reference to the data, the distribution of $\hat{\beta}(\mathcal{M})$ is multivariate normal, with mean and variance given by the results of Sections 9.2.1 and 9.2.2. The standard errors reported by statistical packages are based on this last distribution, with the additional assumption that the model chosen is the correct one. These reported standard errors can be very different from the correct standard errors based on the conditional or unconditional distributions. The following example should make these distinctions clear.

EXAMPLE 12.3 Consider the centered and scaled regression model with two explanatory variables discussed in Example 12.2 in Section 12.4.1. Suppose that we want to estimate the parameter γ_1 after selecting a model. Various estimates of γ_1 under different model assumptions were given in Example 12.2.

Taking the model selection into account amounts to using the estimate

$$\tilde{\gamma}_1 = \begin{cases} 0, & \text{if model } \{0\} \text{ is chosen,} \\ r_1, & \text{if model } \{x_1\} \text{ is chosen,} \\ 0, & \text{if model } \{x_2\} \text{ is chosen,} \\ (r_1 - rr_2)/(1 - r^2), & \text{if model } \{x_1, x_2\} \text{ is chosen.} \end{cases} \quad (12.69)$$

Suppose that we select the model using forward selection. If we estimate γ_1 in this way, we are using a very different estimate from that used if we pick the model a priori. If we assume a priori that model $\{x_1\}$ is the correct model, we use the estimate $\gamma_1^\dagger = r_1$. If we assume that the true model is $\{x_1, x_2\}$, we use $\hat{\gamma}_1 = (r_1 - rr_2)/(1 - r^2)$.

These estimates have quite different statistical characteristics. In Table 12.1 we list their means and variances for a selection of parameter values. These were obtained by simulation in the case of $\tilde{\gamma}_1$. It is common practice to select the model and then calculate the standard errors under the assumption that the chosen model is the true one. This amounts to using the wrong column in Table 12.1. For example, if the model selected was $\{x_1, x_2\}$, and we used the variances in the last column, we would incorrectly estimate the standard error of $\tilde{\gamma}_1$. The correct variances are in the fifth column.

We can also consider the conditional distributions of these estimates. Suppose we select a model and then calculate the LSE's for the chosen model. These are no longer unbiased. In Table 12.2, we show the conditional and a priori mean-squared errors for two different models. The conditional MSE's are calculated using the distribution of Y , conditional on the indicated model

Table 12.1 Means and variances for different estimators.

γ_1	γ_2	r	$E[\tilde{\gamma}_1]$	$\text{var}[\tilde{\gamma}_1]$	$E[\check{\gamma}_1]$	$\text{var}[\check{\gamma}_1]$	$E[\hat{\gamma}_1]$	$\text{var}[\hat{\gamma}_1]$
0	0	0.0	0.00	0.84	0.0	1	0	1.00
0	0	0.5	0.00	0.80	0.0	1	0	1.33
0	0	0.9	0.00	0.63	0.0	1	0	5.26
3	0	0.0	2.89	1.51	3.0	1	3	1.00
3	0	0.5	2.93	1.39	3.0	1	3	1.33
3	0	0.9	2.97	1.39	3.0	1	3	5.26
3	3	0.0	3.15	0.81	3.0	1	3	1.00
3	3	0.5	3.38	1.37	4.5	1	3	1.33
3	3	0.9	5.83	1.12	5.7	1	3	5.26

being selected, while the a priori MSE (equal to the a priori variance in this case) is the variance calculated assuming the model has been specified in advance. We see that, unlike the unconditional case, the conditional MSE's are smaller than the a priori variances. \square

Table 12.2 Conditional and a priori mean-squared errors

γ_1	γ_2	r	MSE		Conditional	a priori
			Assuming model $\{x_1\}$ chosen	Assuming model $\{x_1, x_2\}$ chosen		
3	0	0.0	0.74	1	0.74	1.00
3	0	0.5	0.72	1	0.97	1.33
3	0	0.9	0.76	1	1.01	5.26
3	3	0.0	0.72	1	0.73	1.00
3	3	0.5	0.94	1	0.53	1.33
3	3	0.9	0.96	1	3.48	5.26

The differences between the conditional and a priori means and variances are considered in more detail in the next subsection.

12.7.2 Bias

Suppose that we select then fit a model \mathcal{M} . Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where \mathbf{X}_1 is the regression matrix for model \mathcal{M} , and \mathbf{X}_2 is the matrix corresponding to the other variables. Suppose that the true model is

$$\mathbf{Y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \quad (12.70)$$

The least squares estimate of $\boldsymbol{\beta}_1$ based on model \mathcal{M} is

$$\hat{\boldsymbol{\beta}}(\mathcal{M}) = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{Y}. \quad (12.71)$$

From Section 9.2.1, its a priori expectation is

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}(\mathcal{M})] &= \boldsymbol{\beta}_1 + (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{X}_2 \boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \mathbf{L} \boldsymbol{\beta}_2, \end{aligned}$$

say. We reemphasize that this expectation is calculated assuming that the model has not been selected with reference to the data. The second term, $\mathbf{L} \boldsymbol{\beta}_2$, is called the *omission bias* by Miller [1990: p. 110]. If the model *has* been data-selected, the expected value is $E[\hat{\boldsymbol{\beta}}(\mathcal{M}) | \mathbf{Y} \in \mathcal{R}]$, where \mathcal{R} is the region leading

to the selection of model \mathcal{M} . The difference $E[\hat{\beta}(\mathcal{M})|\mathbf{Y} \in \mathcal{R}] - E[\hat{\beta}(\mathcal{M})]$ is called the *selection bias*. Thus,

$$E[\hat{\beta}(\mathcal{M})|\mathbf{Y} \in \mathcal{R}] = \beta_1 + \text{selection bias} + \text{omission bias}.$$

12.7.3 Conditional Means and Variances

From (12.71), we have

$$E[\hat{\beta}(\mathcal{M})|\mathbf{Y} \in \mathcal{R}] = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 E[\mathbf{Y}|\mathbf{Y} \in \mathcal{R}],$$

where the last conditional expectation is given by

$$E[Y_i|\mathbf{Y} \in \mathcal{R}] = \frac{\int_{\mathcal{R}} y_i f(\mathbf{y} - \mathbf{X}\beta; \sigma^2) dy}{\int_{\mathcal{R}} f(\mathbf{y} - \mathbf{X}\beta; \sigma^2) dy}$$

and

$$f(\mathbf{y}; \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y}\|^2\right). \quad (12.72)$$

A similar argument shows that

$$\text{Var}[\hat{\beta}(\mathcal{M})|\mathbf{Y} \in \mathcal{R}] = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \Sigma_{\mathcal{M}} \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1},$$

where $\Sigma_{\mathcal{M}} = \text{Var}[\mathbf{Y}|\mathbf{Y} \in \mathcal{R}]$. The a priori variances are calculated using the formula $\sigma^2(\mathbf{X}'_1 \mathbf{X}_1)^{-1}$. We can write the difference as

$$\begin{aligned} & \sigma^2(\mathbf{X}'_1 \mathbf{X}_1)^{-1} - (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \Sigma_{\mathcal{M}} \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \\ &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\sigma^2 \mathbf{I}_n - \Sigma_{\mathcal{M}}) \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \\ &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 (\text{Var}[\mathbf{Y}] - \text{Var}[\mathbf{Y}|\mathbf{Y} \in \mathcal{R}]) \mathbf{X}_1 (\mathbf{X}'_1 \mathbf{X}_1)^{-1}. \end{aligned} \quad (12.73)$$

Using the fact (Exercises 12f, No. 3) that $\text{Var}[\mathbf{Y}] - E[\text{Var}(\mathbf{Y}|\mathbf{Z})]$ is positive definite for any random vector \mathbf{Z} , and A.4.5, we see that the expected value of (12.73) is positive semidefinite, and typically, the package standard errors will be larger than the standard deviation of the conditional estimate (but not necessarily its root-mean-squared error).

12.7.4 Estimating Coefficients Using Conditional Likelihood

To avoid these problems, Miller [1990: p. 138] suggests using a likelihood based on the density of \mathbf{Y} conditional on $\mathbf{Y} \in \mathcal{R}$ used in Section 12.7.3. The conditional density is

$$\frac{\int_{\mathcal{R}} f(\mathbf{y} - \mathbf{X}\beta; \sigma^2) dy}{\int_{\mathcal{R}} f(\mathbf{y} - \mathbf{X}\beta; \sigma^2) dy},$$

where f is given by (12.72). The log-likelihood based on this conditional density is, up to a constant,

$$-\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 - \log \left(\int_{\mathcal{R}} f(\mathbf{y} - \mathbf{X}\beta; \sigma^2) dy \right). \quad (12.74)$$

Evaluation of the last term can be done using a Monte Carlo approach. For fixed β and σ^2 , we repeatedly generate data using the model (12.70) and subject each data set to the variable selection procedure. The integral is then estimated by the proportion of data sets that result in model \mathcal{M} (cf. Miller [1990: p. 140]). The conditional likelihood can be maximized using an algorithm that requires no derivatives, such as the Nelder–Mead method (Nelder and Mead [1965]).

12.7.5 Other Effects of Model Selection

Several authors have noted that when evaluating a model for overall significance using the test in Example 4.8 in Section 4.4, the p -values resulting from the test will be much too small if the model has been data-selected. Rencher and Pun [1980] discuss a Monte Carlo study where variable selection resulted in the null distribution of the test statistic being shifted upward, resulting in overly small p -values in the null case. Freedman [1983] found the same phenomenon. Diehr and Hoflin [1974] discuss an empirical adjustment of the null distribution to compensate for model selection. Hurvich and Tsai [1990] study the effect of model selection on confidence intervals for the regression coefficients and find that the coverages are much less than nominal.

EXERCISES 12F

1. Consider the regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (1 = 1, \dots, n),$$

where $\sum_i x_i = 0$ and $\sum_i x_i^2 = 1$. Let $r = \sum_i x_i Y_i$. Suppose that we estimate the parameter β by the estimate

$$\tilde{\beta} = \begin{cases} 0, & \text{if } |r| < c, \\ r, & \text{if } |r| \geq c, \end{cases}$$

where c is some positive constant. How does the MSE of $\tilde{\beta}$ as an estimate of β compare with that of the usual LSE?

2. Suppose that you use forward selection to choose either the null model $Y_i = \alpha + \varepsilon_i$ or the simple linear regression model in Exercise 1 above. Explain how you would evaluate the conditional density (12.74) in this situation.
3. For any pair of random vectors \mathbf{Y} and \mathbf{Z} , show that the difference $\text{Var}[\mathbf{Y}] - E[\text{Var}(\mathbf{Y}|\mathbf{Z})]$ is positive semidefinite. *Hint:* Show that the difference is $\text{Var}(E[\mathbf{Y}|\mathbf{Z}])$.

12.8 COMPUTATIONAL CONSIDERATIONS

We have seen that many criteria for subset selection can be expressed in terms of the residual sum of squares and the dimension of the model. Moreover, for subsets of fixed size, selection of the best model is equivalent to finding the model with the smallest RSS. To find the best subsets by exhaustive search, we need to calculate the RSS for all models of dimension p , $p = 1, \dots, K + 1$, where K is the number of available variables. Since there are 2^K such models (corresponding to all 2^K ways the variables may be in or out of the model, including the null model), calculation of all possible regressions is a formidable computational task if K is large. We first discuss some efficient algorithms for calculating all possible regressions and then look at some refinements that avoid the consideration of unpromising subsets. This can reduce the computation by a substantial amount.

In this section we assume that the constant term is included in every model, so that the full model, with all variables included, has $K + 1$ parameters. The regression matrix for the full model is therefore assumed to have an initial column of 1's. The symbol p will refer to the dimension of a submodel, which will have $p - 1$ variables plus a constant term.

12.8.1 Methods for All Possible Subsets

The calculations can be based either on the SSCP matrix $\mathbf{X}'\mathbf{X}$ using sweeping, or on the \mathbf{X} matrix using Givens transformations. We saw in Section 11.6.2 how a variable could be added to or removed from the regression by sweeping the augmented SSCP matrix. Recall that sweeping on a variable not in the model *adds* the variable, and sweeping on a variable already in the model *removes* it. We need to construct a series of 2^K sweeps that will generate all 2^K regressions.

The first sweep will sweep on the first column of the $(K + 2)$ by $(K + 2)$ augmented matrix (11.3), and sweeps in the constant term. The result is

$$\begin{pmatrix} \frac{1}{n} & \bar{\mathbf{x}}' & \bar{y} \\ -\bar{\mathbf{x}} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}\tilde{\mathbf{Y}} \\ -\bar{y} & \tilde{\mathbf{Y}}'\tilde{\mathbf{X}} & \tilde{\mathbf{Y}}\tilde{\mathbf{Y}} \end{pmatrix}, \quad (12.75)$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are the centered versions of \mathbf{X} (without the first column) and \mathbf{Y} . Assuming that a constant term is to be included in every model, the initial row and column of (12.75) play no further role in the calculations and can be discarded. To fit the remaining $2^K - 1$ models, we need only work with the $(K + 1)$ by $(K + 1)$ augmented centered SSCP matrix

$$\begin{pmatrix} \tilde{\mathbf{X}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}\tilde{\mathbf{Y}} \\ \tilde{\mathbf{Y}}'\tilde{\mathbf{X}} & \tilde{\mathbf{Y}}\tilde{\mathbf{Y}} \end{pmatrix}. \quad (12.76)$$

For $K = 2$, we can fit the remaining models by using the sequence 1, 2, 1, representing a sweep on variable 1, a sweep on variable 2, and a final sweep on variable 1. This fits the models $\{x_1\}$, $\{x_1, x_2\}$, and $\{x_2\}$, respectively; note that the third sweep, on x_1 , removes x_1 from the previous model $\{x_1, x_2\}$.

For $K = 3$, the sequence is 1, 2, 1, 3, 1, 2, 1 and for $K = 4$ it is 1, 2, 1, 3, 1, 2, 1, 4, 1, 2, 1, 3, 1, 2, 1. The general pattern is now clear: If S_k is the sequence of sweeps for K variables, then the sequence for $K + 1$ variables is $S_{k+1} = S_k \cup \{k + 1\} \cup S_k$. Schatzoff et al. [1968] give a formal proof that the sequence generated in this way does indeed include all the $2^K - 1$ remaining regressions.

Improving the Basic Algorithm

This algorithm in the simple form described above is due to Garside [1965], and requires $2^K - 1$ sweeps of the $(K + 1) \times (K + 1)$ centered augmented SSCP matrix. It can be improved somewhat by modifying the sweep operator to preserve the symmetry of the SSCP matrix, so that elements below the diagonal need not be calculated. This reduces the computations by around 50%. Schatzoff et al. [1968] discuss these refinements.

More dramatic improvements can be made by rearranging the calculations so that only submatrices of the $(K + 1) \times (K + 1)$ SSCP matrix need be swept at each stage. Consider the case $K = 2$. The Garside algorithm performs the sweeps in the order shown in Table 12.3. Note that the entire 3×3 matrix must be swept each time. Now consider the sequence of sweeps shown in Table 12.4, where the variables are swept in the same order as before, but a different matrix is swept. In this approach variables are never swept out, only swept in. The advantage is that at step 1, the sweep need only operate on the submatrix corresponding to variable 1 and the response; we can ignore

Table 12.3 Sequence of sweeps for the Garside algorithm

Step	Variable swept	Matrix swept	Model fitted
1	1	SSCP	{1}
2	2	Result of step 1	{1,2}
3	1	Result of step 2	{2}

Table 12.4 Modified sequence of sweeps

Step	Variable swept	Matrix swept	Model fitted
1	1	Part of SSCP	{1}
2	2	SSCP	{2}
3	1	Result of step 2	{1,2}

variable 2 entirely at step 1. We must store a copy of the SSCP matrix for use in step 2, as it is destroyed in step 1.

If this idea is applied to the case $K > 2$, substantial savings in computation result. At each stage we need only apply the sweeps to a submatrix rather than to the entire matrix as in the Garside algorithm. There is an extra storage cost, as we must store copies of matrices produced at certain stages for use in later stages, but no more than $K + 1$ matrices need be stored at any one time. This idea has been implemented by Furnival [1971] and Morgan and Tatar [1972]. We describe the latter version as it is about 50% more efficient and is very simple to program. These methods are in turn an order of magnitude more efficient than the methods of Schatzoff et al. and Garside.

The Morgan and Tatar implementation uses a modified sweep that preserves the symmetry of the matrices, so that only upper triangles need be retained. However, it can compute only the RSS for each model, whereas the Furnival method can also produce the regression coefficients and the inverse of the $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ matrix. The Morgan and Tatar algorithm is as follows.

Algorithm 12.3

Step 1: Reserve storage for $K + 1$ matrices $\mathbf{M}^{(1)}, \dots, \mathbf{M}^{(K+1)}$, where $\mathbf{M}^{(j)}$ is $j \times j$. Only the upper triangle of each matrix need be stored.

Step 2: Form the SSCP matrix as

$$\begin{bmatrix} \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} & \tilde{\mathbf{X}}'\tilde{\mathbf{Y}} \\ \tilde{\mathbf{Y}}'\tilde{\mathbf{X}} & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{bmatrix}.$$

Note that the response comes first! Copy the $j \times j$ upper triangle of the SSCP matrix into $\mathbf{M}^{(j)}$.

Step 3: For $t = 1, \dots, 2^K - 1$, identify a model with t as follows: Write t in its binary expansion

$$t = b_1 + b_2 2 + b_3 2^2 + \cdots + b_K 2^{K-1},$$

where b_j is zero or 1, and let \mathcal{M}_t be the model containing all the variables x_t for which $b_t = 1$. Let p be the position in the binary sequence $\{b_1, b_2, \dots, b_K\}$ of the first nonzero element. Then:

- (a) Sweep the matrix $\mathbf{M}^{(p+1)} [= (m_{ij})]$, say, on row $p + 1$, using the symmetric sweep

$$b_{ij} = m_{ij} - \frac{m_{i,p+1}m_{j,p+1}}{m_{p+1,p+1}} \quad (1 \leq i \leq j \leq p + 1).$$

Copy the result into a temporary matrix \mathbf{B} . Do not overwrite $\mathbf{M}^{(p+1)}$.

- (b) For $j = 1, 2, \dots, p$, copy the upper $j \times j$ upper triangle of \mathbf{B} into $\mathbf{M}^{(j)}$.
 - (c) The RSS for the model \mathcal{M}_t will now be in $\mathbf{M}^{(1)}$.
-

A formal proof of the correctness of this algorithm is given in Morgan and Tatar [1972].

12.8.2 Generating the Best Regressions

The Morgan and Tatar method is an efficient way of generating all the $2^K - 1$ nonnull regressions. However, the model selection criteria described in Section 12.3 just require finding the best-fitting model of each size: in other words, the model having the smallest RSS among all j -variable models, for $j = 1, 2, \dots, K$. In this case we can exploit a simple property of the RSS to avoid fitting most of the models and thus save a significant amount of computation. The property is this: If a model \mathcal{M}_1 is contained in model \mathcal{M}_2 (in the sense that every variable in \mathcal{M}_1 is also in \mathcal{M}_2), then we know that the RSS for \mathcal{M}_2 cannot be greater than that of \mathcal{M}_1 . This is because the RSS for model \mathcal{M}_1 is a restricted minimum, with some estimated coefficients constrained to be zero, whereas the RSS for model \mathcal{M}_2 is an unrestricted minimum. Thus, for example, if we have $K = 4$ variables and we have fitted model $\{123\}$ and obtained a RSS of 500, and we have also fitted the model $\{4\}$ with an RSS of 100, we know immediately that $\{4\}$ must be the best one-variable model, since $\text{RSS}\{1\} \geq \text{RSS}\{123\} = 500 > 100 = \text{RSS}\{4\}$, and similarly for $\text{RSS}\{2\}$ and $\text{RSS}\{3\}$. By exploiting this property of residual sums of squares, we can avoid fitting the models $\{1\}$, $\{2\}$, and $\{3\}$.

Various algorithms based on this simple idea have been proposed, including those of Hocking and Leslie [1967], LaMotte and Hocking [1970], and Furnival and Wilson [1974]. The Furnival and Wilson algorithm, an adaption of the Furnival method for all possible regressions cited in Section 12.8.1, seems to be the current best implementation of this idea and is used in the SAS procedure PROC REG. We will give a brief simplified description of this algorithm.

The first step in the algorithm is to order the variables according to the magnitude of the t -statistic used to test if the corresponding regression coefficient is zero, so that x_1 has the largest (most significant) t -statistic, and x_K the smallest. Then the models are split into two sets, the first containing all models not containing x_K , and the second containing all models that do contain x_K .

For the first set, we construct a *regression tree*, where the fitting of all possible 2^{K-1} models not containing x_K (including the null model) is represented by the paths through a binary tree. For example, if $K = 4$, to fit all $8 = 2^3$ submodels of the three variables x_1 , x_2 , and x_3 , we use the tree shown in Figure 12.5. We start from the root marked “123.” and traverse

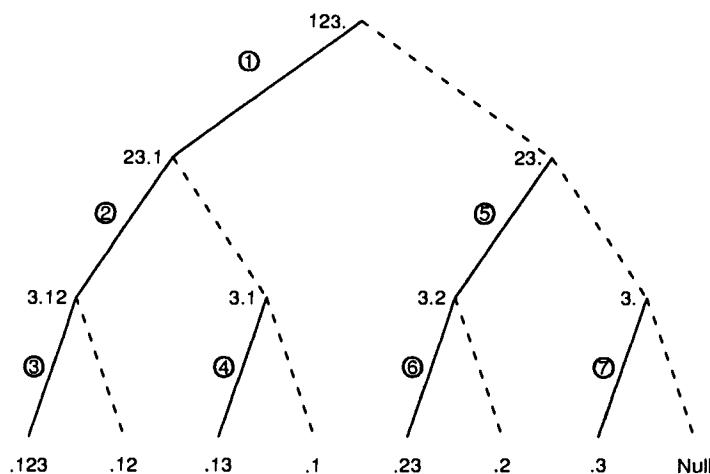


Fig. 12.5 Binary tree for fitting models with three or less variables.

down the branches. Each path down a particular branch represents a series of sweeps that results in fitting one of the 2^3 possible models. Thus, the leftmost branch represents the three sweeps “sweep in x_1 ,” “sweep in x_2 ,” “sweep in x_3 ,” which results in fitting the model {123}. Thus, the solid lines in the figure represent *actual* sweeps. The dashed lines, in contrast, represent *skipped* sweeps. For example, consider the branch starting at the root labeled “123,” moving through the nodes labelled “23.” and “3.2” and ending at “23.” This represents the series “skip the sweep on x_1 , sweep on x_2 , and sweep on x_3 ,” resulting in fitting the model {23}. The labeling of the nodes shows which variables have been swept in at any stage (the ones listed after the “.”) and those still available, not having been skipped at an earlier stage (the ones before the “.”). Note that under this scheme, variables are never swept out, only swept in. For this reason, Gaussian elimination can be used rather than sweeps, with a consequent saving in arithmetic.

This approach is, in fact, the basis of the Furnival algorithm for all possible regressions cited in Section 12.8.1. Branches of the tree are traversed in some feasible order, backing up as required to go down new branches. Intermediate results are stored as required.

The Furnival and Wilson algorithm makes use of this tree and also of a dual tree that is the same shape as the first (primary) tree, but is interpreted differently. The second tree represents fitting all the models that contain x_K . The root represents the full model, and the solid lines now represent *sweeping out* variables. As before, the dashed lines represent skipping sweeps. The

dual tree for our example with $K = 4$ is shown in Figure 12.6. The nodes are labeled differently; variables after the “.” are variables not yet swept out. Note that the nodes in each tree correspond in an obvious way. The variables in a node label in the primary tree are those either swept in or still available for sweeping in, those in a node label in the dual tree are those still available for sweeping out, plus the variable x_K , which is never swept. Each sweep in the primary tree is just a move from a particular node down the solid left branch to the node below. This corresponds to a sweep in the dual tree from the corresponding node but down the solid right branch.

Having set up these trees, we first sweep all the variables into the model by moving down the leftmost branch of the primary tree and then performing a final sweep on x_K . This fits the K models $\{1\}, \{1, 2\}, \dots, \{1, 2, 3, \dots, K\}$. We then perform the corresponding sweeps on the dual tree, for a total of $2K - 1$ sweeps.

We now traverse the primary tree, performing sweeps in order. The corresponding sweeps are also done on the dual tree. At each point we can use the information in the dual tree and the “RSS property” to decide if we need to continue down a branch.

We illustrate the procedure with our $K = 4$ example. In Table 12.5 the $2^K = 16$ models are listed, along with the RSS from an actual set of data. To select the best one-, two-, three-, and four-variable models without fitting all 15 models, we proceed as follows: We traverse the primary tree in “lexographic” order, fitting models in the sequence shown in Table 12.5. This

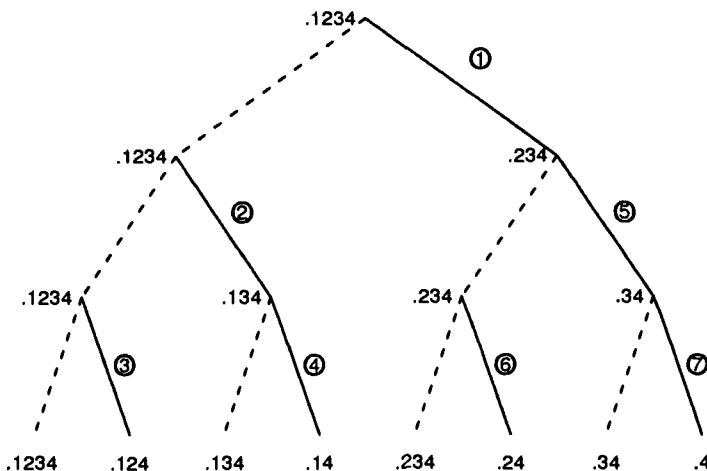


Fig. 12.6 Dual tree for $K = 4$.

Table 12.5 Models and RSS for the $K = 4$ example

Model	RSS	Model	RSS
Null	320	{4}	298
{1}	82	{1, 4}	80
{2}	186	{2, 4}	185
{1, 2}	10	{1, 2, 4}	10
{3}	229	{3, 4}	227
{1, 3}	62	{1, 3, 4}	60
{2, 3}	157	{2, 3, 4}	157
{1, 2, 3}	6	{1, 2, 3, 4}	5

amounts to always going down the left branches, backing up a minimum distance when we reach the bottom of a branch. The order is shown in Figures 12.5 and 12.6 by the numbers in the small circles. At each node we use the dual tree to decide if it is necessary to continue down a branch or if we can "bound" to a new branch.

After fitting the models {1}, {1, 2}, {1, 2, 3}, and {1, 2, 3, 4}, as discussed above, we then do the dual sweeps, resulting in fitting the further models {2, 3, 4}, {1, 3, 4}, and {1, 2, 4}. At this stage the best one-variable model is {1} with RSS 82; the best two-variable model is {1, 2} with RSS 10, and the best three-variable model is {1, 2, 3} with RSS 6. Now consider sweep 4 in the primary tree, which would fit model {1, 3}. But sweep 2 in the dual tree gave us $\text{RSS}\{134\} = 60$, so we know that $\text{RSS}\{13\} \geq \text{RSS}\{134\} = 60$. The current best two-variable model is {1, 2} with RSS 10, so we don't need to perform sweep 4.

Moving on to sweep 5 in the primary tree, we see that the models that can be fitted by going down this branch are {2} and {2, 3}. The RSS of both these models is not less than $\text{RSS}\{234\} = 157$. Since the current best one- and two- variable models have RSS 82 and 10, respectively, we don't need to go any farther down this branch, so we can skip sweeps 5 and 6.

Next we examine sweep 7. The only model fitted down this branch is {3}, but since $\text{RSS}\{3\} \geq \text{RSS}\{234\} = 157$, model {3} cannot be the best one-variable model.

We have found the best one-, two- and three- variable models using seven sweeps instead of the 15 it would have taken using all possible regressions. Even greater savings are made with larger values of K . For example, Furnival and Wilson claim that their algorithm is 15 to 50 times faster than the LaMotte-Hocking program, which is in turn much faster than the Morgan and Tatar method for all possible regressions.

The layout of the two trees ensures that as we traverse the primary tree, the necessary fits have been performed in the dual tree to get a lower bound for all the models that can be fitted by going down the current branch in the

primary tree. Furnival and Wilson call the primary tree the *branch tree* and the dual tree the *bound tree* as it determines when we can “bound” to a new branch of the primary tree.

12.8.3 All Possible Regressions Using QR Decompositions

The methods described above are fast, but being based on the SSCP matrix can be inaccurate if there are substantial correlations between the explanatory variables (cf. Section 11.8). The alternative is to use the methods based on the QR decomposition described in Section 11.6.3. The algorithms in Section 12.8.2 can be adapted to use reflections and rotations instead of sweeps and GE steps. This results in a more accurate but slower algorithm.

When using Givens rotations, it will generally be better to add and delete variables in an order different from that used in the Garside algorithm, since the computational cost of deleting variables using Givens transformations depends on their column position in the QR decomposition. For example, suppose that we have fitted a model $\{1, 2, 3, 4\}$, and have a QR decomposition of the form

$$\mathbf{Q}'[\mathbf{X}, \mathbf{y}] = \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \times \end{bmatrix}.$$

Deleting variable 4 has no computational cost, since we simply drop the fourth column. However, if we drop variable 1, we get the matrix

$$\begin{bmatrix} \times & \times & \times & \times \\ * & \times & \times & \times \\ 0 & * & \times & \times \\ 0 & 0 & * & \times \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \times \end{bmatrix},$$

and the reduction to uppertriangular form involves zeroing all the elements marked with a “*.” Thus, it is cheaper to delete variables having large indices (3 or 4, in this case) than those having low indices (1 or 2). In fact, it is cheaper to delete variables 3 and 4 simultaneously, which has no computational cost, than it is to delete variable 1, which costs three rotations.

Thus, we can perhaps do better by using a binary scheme which drops and adds more than one variable at a time rather than the sequences used in the sweep algorithms. For example, if we identify models with binary

numbers so that 0001 corresponds to {4}, 0010 corresponds to {3}, 0011 corresponds to {3, 4}, and so on, then fitting models in ascending binary order will ensure that most of the adds and deletes will involve variables with large indices, and the computational cost will be small. Miller [1990: Chapter 2] presents some calculations which suggest that the binary approach is only about 60% of the cost of the Garside ordering. This reference has more details on the QR approach to all possible regressions calculations and some accuracy comparisons with the sweep methods. However, we note that if the IEEE double-precision standard is used for computer arithmetic, the variables need to be very collinear before accuracy problems arise with the sweep methods.

EXERCISES 12g

1. Write a small computer program to calculate the residual sum of squares for all possible regressions using the Garside algorithm. What size problems can reasonably be handled?
2. Implement the Morgan and Tatar algorithm. What sort of improvements do you notice over the Garside algorithm?

12.9 COMPARISON OF METHODS

In previous sections of this chapter we have discussed a range of methods (all possible regressions, stepwise methods, shrinkage methods, and Bayesian methods) that can be used for model selection and prediction. In this section we take up the task of making some comparisons between the methods.

This is complicated by the lack of a single criterion. If our aim is subset selection, a reasonable aspect on which to focus is the ability of a method to select the “correct” subset, assuming that such a thing exists. If the aim is to make good predictions, then the methods can be evaluated in terms of prediction error.

12.9.1 Identifying the Correct Subset

Suppose that we have available explanatory variables x_1, \dots, x_K , and that the correct model is in fact

$$Y_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i \quad (i = 1, \dots, n), \quad (12.77)$$

where $\mathbf{x}'_i = (x_{i0}, \dots, x_{iK})$ and $\boldsymbol{\beta}$ is a fixed $(K+1)$ -vector of coefficients, some of which are possibly zero. We can then evaluate the methods according to how they identify the nonzero coefficients, or, in other words, select the correct subset.

There have been several investigations into this question. Most have assumed a simplified version of the problem, where the first p coefficients are

nonzero and the rest are zero. Nishi [1984] shows that as K remains fixed but n gets large, the criteria AIC, C_p , and CV(1) are all asymptotically the same, and all tend to overfit, in the sense that the probability of selecting a subset properly containing the true subset converges to a positive number rather than zero. The probability of underfitting converges to zero. In contrast, under these asymptotics, the BIC selected the true model with probability converging to 1.

Zhang [1992] considers different asymptotics, where K is allowed to converge to infinity along with n but p remains fixed. His conclusions are, however, the same: the criteria AIC, C_p and CV(1) all overfit. However, the extent of the overfitting is not very great. Zhang [1993] examines cross-validation and finds that CV(d), where $d > 1$, performs better than CV(1), provided that d is a reasonably large fraction of n . Since this implies a great deal of computation, Zhang advocates the use of cross-validation based on a balanced subset of d -subsets of cases as described in Section 12.3.2. Shao [1993] also considers cross-validation and reaches the same conclusion. The general conclusion here is that when the number of nonzero coefficients is small, the AIC-like criteria tend to overfit.

Several simulation studies (e.g., Freedman [1983] and Rencher and Pun [1980]) have been published which make the same point: namely, that even in the extreme case when all the coefficients are zero, it is common to find subsets with small numbers of variables that fit well, having small values of the RSS.

12.9.2 Using Prediction Error as a Criterion

The results quoted in Section 12.9.1 all essentially cover the case where the number of nonzero coefficients was small and the focus was on identifying this set of nonzero coefficients. Other authors have used prediction error as a criterion for evaluating the competing methods.

The most extensive investigations have been conducted by Breiman in a series of papers [1992, 1995, 1996b], and we discuss his findings in some detail. We have seen that there are several alternatives to least squares for constructing predictors. For most of these methods, the general form of the predictor is specified but the user must calibrate the predictor in some way. For example, in ridge regression, the ridge parameter k must be chosen. In subset selection, the subset must be chosen and then the predictor is calculated by applying least squares to the subset chosen. If the procedure is calibrated properly, we can expect improvements over least squares, but not otherwise.

In many techniques, this calibration is performed by estimating the prediction error and choosing the calibration parameter to minimize the estimated PE. Thus, we could use cross-validation to choose a subset or to select a value for the ridge parameter. Suppose that the PE of the method we choose depends on such a calibration parameter s , which could be a subset of variables, a ridge coefficient, or perhaps a garrote or lasso coefficient. There will be

a value of s that minimizes the PE, say s_{MIN} , with corresponding minimum prediction error $\text{PE}(s_{\text{MIN}})$. If we could evaluate $\text{PE}(s)$ for each s , we could use the method (ridge, subset selection) for which $\text{PE}(s_{\text{MIN}})$ is a minimum.

Since s_{MIN} is unknown, we have to use an estimate \hat{s}_{MIN} , which is usually the value of s which minimizes some estimate $\widehat{\text{PE}}$ of PE. Unfortunately, $\widehat{\text{PE}}(\hat{s}_{\text{MIN}})$ may be a poor estimate of $\text{PE}(\hat{s}_{\text{MIN}})$. All the results considered in previous sections about the approximately unbiased nature of criteria such as AIC and CV(1) refer to the case where s is fixed, not data dependent. They do not refer to $\widehat{\text{PE}}(\hat{s}_{\text{MIN}})$. Indeed, we saw in Example 12.1 just how bad C_p can be in the case of orthogonal explanatory variables.

When prediction is viewed in this way, there are two aspects to consider. First, there is the size of the optimal prediction error $\text{PE}(s_{\text{MIN}})$, which is a property of the prediction method. Second, how well can the method be calibrated [i.e., how close is $\text{PE}(\hat{s}_{\text{MIN}})$ to $\text{PE}(s_{\text{MIN}})$]? In other words, will estimating the PE mislead us, and lead us to a predictor that in fact performs much worse than we believe?

To explore this idea, Breiman [1996b] introduces the concept of *predictive loss* (PL), defined as

$$\text{PL} = \text{PE}(\hat{s}_{\text{MIN}}) - \text{PE}(s_{\text{MIN}}).$$

This is a property of both the prediction method and the method of estimating PE. We have

$$\text{PE}(\hat{s}_{\text{MIN}}) = \text{PL} + \text{PE}(s_{\text{MIN}}),$$

which expresses the actual PE of our chosen method as a term involving the method of PE estimation, plus a term involving just the method of prediction.

Methods of estimating PE include C_p , CV(1), CV(d), and the bootstrap. Breiman [1992, 1995, 1996b] has conducted extensive simulations of all these methods and concludes that, [with the exception of CV(d) where d is large relative to the number of cases n], none are satisfactory. He advocates instead the use of the *little bootstrap*, which has smaller predictive loss than the other methods. It can be used with subset selection and the shrinkage methods. We discuss the little bootstrap in the next section.

Breiman's simulations also indicate that subset selection has a generally small minimum PE but a big predictive loss, whereas ridge has small predictive loss but large minimum PE. The other shrinkage techniques fall in the middle and represent a good compromise choice. He also uses the idea of *stabilization*, which can be used to improve the performance of subset selection (cf. Breiman [1996b] for details).

If the number of nonzero coefficients is large, the asymptotically equivalent methods AIC, CV(1), and C_p tend to work better. Li [1987] and Shibata [1981, 1984] have explored the asymptotics in this situation. They find that under this form of asymptotics, the ratio of $E[\text{PE}(\hat{s}_{\text{MIN}})]$ to $E[\text{PE}(s_{\text{MIN}})]$ converges to 1 when these "AIC-like" techniques are used for subset selection. This implies that the expected predictive loss will be small in this situation.

However, the Breiman simulations do not support this. He finds (Breiman [1995]) that even when the number of nonzero coefficients is a reasonable proportion (say, 50%) of the total number of coefficients, the AIC-like methods still substantially underestimate the true PE. When the proportion of nonzero coefficients is much smaller, Breiman also reports substantial overfitting when using the AIC methods.

From the discussion above it is clear that the actual PE of the methods will depend on the true values of the regression coefficients. In the case of orthogonal explanatory variables, we can get some theoretical insight into how these methods compare for different distributions of the regression coefficients by using a Bayesian argument. Suppose that, conditional on β , the true model is given by (12.77). Also, suppose that the error variance σ^2 is known and that the elements of β are independently distributed with density function g . Note that since the columns of \mathbf{X} are assumed to be orthonormal, we have $\hat{\beta} = \mathbf{X}'\mathbf{Y}$. Conditional on β , $\hat{\beta}$ has independent elements; $\hat{\beta}_j$ is $N(\beta_j, \sigma^2)$, with density $\sigma^{-1}\phi[(b - \beta_j)/\sigma]$, where ϕ is the $N(0, 1)$ density. The joint distribution of β_j and $\hat{\beta}_j$ has density

$$f(b, \beta) = f_1(b|\beta)f_2(\beta) = \sigma^{-1}\phi\{(\beta_j - b_j)/\sigma\}g(\beta).$$

We have seen that in the orthonormal case $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, many of the prediction methods reduce to using the predictor $\mathbf{x}'\tilde{\beta}$, where $\tilde{\beta}_j = h(\hat{\beta}_j)$. The corresponding functions h were graphed in Figure 12.4. Also, in the orthonormal case, the ME is

$$\|\mu - \mathbf{X}\tilde{\beta}\|^2 = \|\mathbf{X}\beta - \mathbf{X}\tilde{\beta}\|^2 \quad (12.78)$$

$$= \|\beta - \tilde{\beta}\|^2, \quad (12.79)$$

so the expected ME, averaging over both $\hat{\beta}$ and β , is

$$\frac{p}{\sigma} \int \int [\beta - h(b)]^2 \phi[(\beta - b)/\sigma] g(\beta) d\beta db. \quad (12.80)$$

A standard argument (see Exercises 12h, No. 1) shows that the function h which minimizes the expected ME is the conditional expectation of β , given $\hat{\beta}$, which is

$$h(b) = E[\beta|b] = \frac{\int \beta \phi[(\beta - b)/\sigma] g(\beta) d\beta}{\int \phi[(\beta - b)/\sigma] g(\beta) d\beta}. \quad (12.81)$$

Clearly, none of the prediction methods we have considered can be better than the method based on the conditional expectation, so in general they will be suboptimal. However, in particular cases, some of these methods do coincide with the optimal choice, as we see in our next example.

EXAMPLE 12.4 Suppose that the marginal prior on β is $N(0, \sigma_0^2)$. Then, for a prediction method using transformed coefficients $\tilde{\beta} = h(\hat{\beta})$, the expected

model error is given by

$$\frac{p}{\sqrt{2\pi\sigma_0^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} [\beta - h(b)]^2 \exp\left\{-\frac{(\beta-b)^2}{2\sigma^2}\right\} \exp\left(-\frac{\beta^2}{2\sigma_0^2}\right) d\beta db. \quad (12.82)$$

Completing the square, we get

$$\begin{aligned} & \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\beta-b)^2}{2\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\beta^2}{2\sigma_0^2}\right) \\ &= \frac{1}{\sqrt{2\pi w\sigma^2}} \exp\left\{-\frac{(\beta-wb)^2}{2w\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{b^2}{2\tilde{\sigma}^2}\right) \end{aligned} \quad (12.83)$$

where $\tilde{\sigma}^2 = \sigma_0^2 + \sigma^2$ and $w = \sigma_0^2/\tilde{\sigma}^2$. Using this, from (12.82) we see that the expected ME is

$$\frac{p}{\sqrt{2\pi\tilde{\sigma}^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi w\sigma^2}} \int_{-\infty}^{\infty} [\beta - h(b)]^2 \exp\left\{-\frac{(\beta-wb)^2}{2w\sigma^2}\right\} d\beta \exp\left(-\frac{b^2}{2\tilde{\sigma}^2}\right) db.$$

Fixing b , the inner integral can be written as

$$\begin{aligned} E\{[\beta - h(b)]^2\} &= \text{var}[\beta] + [h(b) - wb]^2 \\ &= w\sigma^2 + [h(b) - wb]^2, \end{aligned}$$

so that we get, after a change of variables $z = b/\tilde{\sigma}$,

$$E[\text{ME}] = p \left\{ w\sigma^2 + \int_{-\infty}^{\infty} [h(\tilde{\sigma}z) - w\tilde{\sigma}z]^2 \phi(z) dz \right\}. \quad (12.84)$$

This shows that the function for the optimal predictor is $h(b) = wb$, which is the form for ridge regression (cf. Section 12.5.2).

We saw in Example 12.1 in Section 12.3.2 that in the orthonormal case, AIC, C_p , and BIC are equivalent to including variable j in the predictor if and only if $|\hat{\beta}_j| \geq \tau$ for some threshold τ . Since the predictors are orthogonal, the estimates $\hat{\beta}_j$ do not change if terms are deleted from the model, so this is equivalent to taking

$$h(b) = \begin{cases} b, & |b| \geq \tau, \\ 0, & |b| < \tau. \end{cases}$$

Substituting this (Miscellaneous Exercises 12, No. 4) in (12.84) gives

$$E[\text{ME}] = p\{\sigma^2 + (\sigma^2 - \sigma_0^2)[1 - 2\Phi_2(\tau/\tilde{\sigma})]\}, \quad (12.85)$$

where

$$\Phi_2(t) = \int_{-\infty}^t z^2 \phi(z) dz.$$

The function Φ_2 is increasing and has value 0.5 at zero and 1 at $+\infty$. Thus,

if $\sigma^2 > \sigma_0^2$, the expected ME is minimized when $\tau = +\infty$ (i.e., when we use the null predictor). In this case the minimum expected ME is σ_0^2 .

Conversely, if $\sigma^2 < \sigma_0^2$, then the expected ME is minimized when $\tau = 0$ (i.e., when we use least squares). In this case the minimum value of $E[ME]$ is σ^2 . Note that the choice $\tau = \sqrt{2}$ (corresponding to AIC or C_p) is never the best choice of threshold.

The expected ME for the other methods can be evaluated easily using numerical integration and (12.84). In Figure 12.7 we have plotted $E[ME]$ as a function of σ_0 , with σ set at 1. Except when σ_0 is very small (so that most of the β_j 's are zero), the shrinkage predictors are much closer to the optimal h than is the best threshold predictor. \square

EXAMPLE 12.5 Suppose that the common marginal prior is binary, with β taking on values 0 and β_0 with probabilities π and $1 - \pi$, respectively. Then

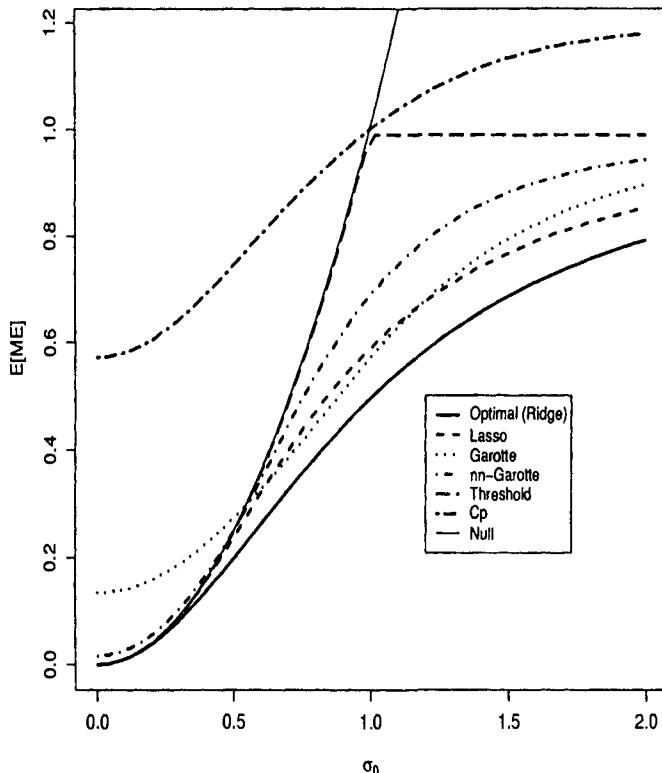


Fig. 12.7 Expected model error for various normal priors.

the optimal h is given by

$$h(b) = \frac{(1 - \pi)\beta_0\phi[(b - \beta_0)/\sigma]}{\pi\phi(b/\sigma) + (1 - \pi)\phi\{(b - \beta_0)/\sigma\}}$$

and the minimum $E[\text{ME}]$ is

$$p\pi(1 - \pi)\beta_0^2 \int_{-\infty}^{\infty} \frac{\phi(z)\phi(z - \beta_0/\sigma)}{\pi\phi(z) + (1 - \pi)\beta_0\phi(z - \beta_0/\sigma)} dz.$$

The ME for an arbitrary method is

$$p \int_{-\infty}^{\infty} \{\pi h(z\sigma)^2 + (1 - \pi)[h(\sigma z + \beta) - \beta]^2\}\phi(z) dz.$$

As in Example 12.4, this formula can be evaluated numerically for the various methods and the minimum expected ME found for each method. In Figure 12.8, we plot the minimum $E[\text{ME}]$ for each method as a function of β_0 and π for $\sigma = 1$.

When $\pi = 0.2$, i.e., when the number of null coefficients is small, ridge does well, and the threshold (subset selection) methods do poorly. Conversely, when $\pi = 0.8$, i.e., when the number of null coefficients is large, the optimal threshold method does well while ridge does very poorly. Interestingly, using C_p is bad in both cases, indicating the need for careful choice of the threshold. The optimal “gold standard” is a clear winner in both cases. The garrote methods and the lasso are roughly comparable. \square

EXAMPLE 12.6 We can combine the last two examples, and consider a prior for β in which β is zero with probability π , and distributed as $N(0, \sigma_0^2)$ with probability $1 - \pi$. Breiman [1995] compared ridge and the nn-garrote with subset selection using the optimal threshold (i.e., the optimal choice of τ). As in Example 12.5, he found that ridge does well when most of the coefficients are non-zero (when subset selection does poorly), and does badly when most of the coefficients are zero (when subset selection does well). The nn-garrote is a good compromise between these two extremes. \square

The Little Bootstrap

We now reexamine the question of estimating the PE of a prediction rule. We have seen that most of the standard methods, particularly the AIC-like methods, are not very satisfactory for estimating the PE of a prediction rule based on choosing subsets. The situation is a little better for ridge, since both CV(1) and GCV (Section 12.5.2) tend to work quite well. However, the best method known seems to be the little bootstrap, and we devote the rest of this section to a description of this technique.

As described above, most prediction methods depend on a “calibration parameter” s , and we use the prediction rule having calibration parameter

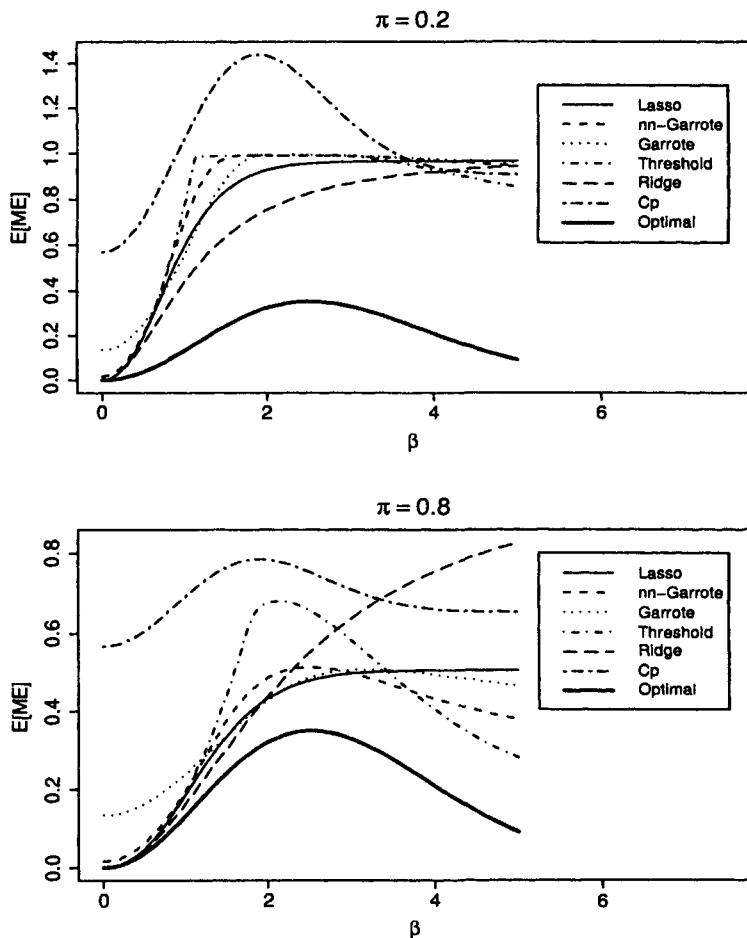


Fig. 12.8 Expected model error for various binary priors.

\hat{s}_{MIN} , which minimizes the estimated PE, $\hat{s}_{MIN} = \operatorname{argmin} \widehat{\text{PE}}(s)$. The calibration parameters and corresponding predictors for the various methods are as follows.

For subset selection: In this case s is the number of variables in the subset chosen, and we use the least squares predictor based on the best-fitting subset of size s . Alternatively, we could use forward selection or backward elimination rather than all possible regressions to choose the subset of size s .

For ridge: We use the predictor predictor $\hat{\mu}(s) = \mathbf{X}\hat{\beta}(s)$, where $\hat{\beta}(s)$ minimizes $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ subject to $\|\mathbf{b}\|^2 \leq s$.

For the nn-garrote: We use the predictor $\tilde{\mu}(s) = \mathbf{X}\hat{\beta}(s)$, where $\hat{\beta}(s)_j = c_j\hat{\beta}_j$, and the c_j 's minimize (12.57) subject to the constraints $c_j \geq 0$, $j = 0, \dots, p - 1$, and $\sum_{j=0}^{p-1} c_j \leq s$.

For the garrote: We use the predictor $\tilde{\mu}(s) = \mathbf{X}\hat{\beta}(s)$, where $\|\mathbf{c}\|^2 \leq s$.

For the lasso: We use the predictor $\tilde{\mu}(s) = \mathbf{X}\hat{\beta}(s)$, where the $\hat{\beta}(s)_j$'s minimize $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2$ subject to $\sum_{j=0}^{p-1} |b_j| \leq s$.

Suppose that our model is

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is $N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and that we use one of the predictors $\tilde{\mu}(s)$ described above. For fixed s we can estimate $\text{PE}(s)$ by the following procedure.

First note that using the same argument as in Section 12.2, we get

$$\text{PE}(s) = n\sigma^2 + \|\boldsymbol{\mu} - \tilde{\mu}(s)\|^2,$$

so that

$$\begin{aligned} \text{RSS}(s) &= \|\mathbf{y} - \tilde{\mu}(s)\|^2 \\ &= \|\boldsymbol{\varepsilon} + \boldsymbol{\mu} - \tilde{\mu}(s)\|^2 \\ &= \|\boldsymbol{\varepsilon}\|^2 + 2\boldsymbol{\varepsilon}'[\boldsymbol{\mu} - \tilde{\mu}(s)] + \|\boldsymbol{\mu} - \tilde{\mu}(s)\|^2 \\ &= \{\|\boldsymbol{\varepsilon}\|^2 + 2\boldsymbol{\varepsilon}'\boldsymbol{\mu} - n\sigma^2\} + \text{PE}(s) - 2\boldsymbol{\varepsilon}'\tilde{\mu}(s). \end{aligned}$$

Now the term in braces $\{\}$ has zero expectation, so that

$$E[\text{PE}(s)] = E[\text{RSS}(s)] + 2E[\boldsymbol{\varepsilon}'\tilde{\mu}(s)]. \quad (12.86)$$

Thus, if we could get an approximately unbiased estimate of $\boldsymbol{\varepsilon}'\tilde{\mu}(s)$, we would have an approximately unbiased estimate of $\text{PE}(s)$.

Such an estimate is furnished by the following resampling procedure.

Step 1: For a small positive number t , generate $\boldsymbol{\varepsilon}_1^*, \dots, \boldsymbol{\varepsilon}_n^*$ from $N(0, t^2\sigma^2)$.

Step 2: Compute $y_i^* = y_i + \boldsymbol{\varepsilon}_i^*$, $i = 1, \dots, n$.

Step 3: Calculate $\tilde{\mu}^*(s)$ using the same method as $\tilde{\mu}(s)$, but using the y_i^* 's instead of the y_i 's. Put $B_t(s) = \boldsymbol{\varepsilon}'\tilde{\mu}^*(s)/t^2$.

Step 4: Repeat steps 1–3 several times and compute the average $\bar{B}_t(s)$ of the resulting $B_t(s)$'s.

Then Breiman [1996b] proves that, for all the situations above, $\bar{B}_t(s)$ is an almost unbiased estimate of $E[\boldsymbol{\varepsilon}'\tilde{\mu}(s)]$, with the bias converging to zero as $t \rightarrow 0$.

For methods such as garrote and ridge, where the corresponding $\tilde{\beta}$ is a smooth function of $\hat{\beta}$, the variance of $\bar{B}_t(s)$ remains bounded as $t \rightarrow 0$. In

this case we may let $t \rightarrow 0$, which results in the *tiny bootstrap*. For ridge, the little and tiny bootstraps both have the form (Breiman [1995])

$$\widehat{PE}(s) = \text{RSS}(s) + 2\hat{\sigma}^2 \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'],$$

where $\hat{\sigma}^2$ is an estimate of σ^2 and k is the ridge coefficient corresponding to s , as in Section 12.5.2. A similar approach, using a different method to estimate $E[\boldsymbol{\epsilon}'\tilde{\boldsymbol{\mu}}(s)]$, is discussed in Tibshirani and Knight [1999].

EXERCISES 12h

1. Show that the function f which minimizes the expected ME (12.80) is the conditional expectation $E[\beta|\hat{\beta}]$ given by (12.81).
2. Draw the same diagram for Example 12.6 as those given in the text in Examples 12.4 and 12.5.
3. Show that in ridge regression when $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, the little bootstrap and GCV give the same estimate for the ridge parameter.

MISCELLANEOUS EXERCISES 12

1. Explain how you could calculate the ridge estimate $\hat{\boldsymbol{\beta}}(k)$ using a least squares program.
2. Consider a regression with two orthogonal centered and scaled explanatory variables x and z . Using the techniques of Example 12.2 in Section 12.4.1, calculate the probability of correctly identifying the true model using forward selection when the regression coefficients of x and z are both zero. Use a value of $c_1 = 1.64$ and assume that the error variance is 1.
3. Consider a regression with K explanatory variables. Let F_p be the F -statistic for testing that some specified subset of r variables can be deleted from the model, where $r = K + 1 - p$. Show that the value of C_p for this reduced model with the r variables deleted is $C_p = r(F_p - 1) + p$.
4. Verify (12.85).

Appendix A

Some Matrix Algebra

A.1 TRACE AND EIGENVALUES

Provided that the matrices are conformable:

A.1.1. $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr } \mathbf{A} + \text{tr } \mathbf{B}$.

A.1.2. $\text{tr}(\mathbf{AC}) = \text{tr}(\mathbf{CA})$.

The proofs are straightforward.

A.1.3. If \mathbf{A} is an $n \times n$ matrix with eigenvalues λ_i ($i = 1, 2, \dots, n$), then

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i \quad \text{and} \quad \det(\mathbf{A}) = \prod_{i=1}^n \lambda_i.$$

Proof. $\det(\lambda \mathbf{I}_n - \mathbf{A}) = \prod_i (\lambda - \lambda_i) = \lambda^n - \lambda^{n-1}(\lambda_1 + \lambda_2 + \dots + \lambda_n) + \dots + (-1)^n \lambda_1 \lambda_2 \dots \lambda_n$. Expanding $\det(\lambda \mathbf{I}_n - \mathbf{A})$, we see that the coefficient of λ^{n-1} is $-(a_{11} + a_{22} + \dots + a_{nn})$, and the constant term is $\det(-\mathbf{A}) = (-1)^n \det(\mathbf{A})$. Hence the sum of the roots is $\text{tr}(\mathbf{A})$, and the product $\det(\mathbf{A})$.

A.1.4. (*Principal axis theorem*) If \mathbf{A} is an $n \times n$ symmetric matrix, then there exists an orthogonal matrix $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n)$ such that $\mathbf{T}'\mathbf{A}\mathbf{T} = \Lambda$,

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$. Here the λ_i are the eigenvalues of \mathbf{A} , and $\mathbf{At}_i = \lambda_i \mathbf{t}_i$. The eigenvectors \mathbf{t}_i form an orthonormal basis for \Re_n . The factorization $\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}'$ is known as the *spectral decomposition* of \mathbf{A} .

In the next three results we assume that \mathbf{A} is symmetric.

$$\text{A.1.5. } \text{tr}(\mathbf{A}^s) = \sum_{i=1}^n \lambda_i^s.$$

$$\text{A.1.6. If } \mathbf{A} \text{ is nonsingular, the eigenvalues of } \mathbf{A}^{-1} \text{ are } \lambda_i^{-1} \text{ (} i = 1, \dots, n \text{), and hence } \text{tr}(\mathbf{A}^{-1}) = \sum_{i=1}^n \lambda_i^{-1}.$$

$$\text{A.1.7. The eigenvalues of } (\mathbf{I}_n + c\mathbf{A}) \text{ are } 1 + c\lambda_i \text{ (} i = 1, \dots, n \text{).}$$

Proof. Let $\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}'$ be the spectral decomposition of \mathbf{A} . Then $\mathbf{A}^s = \mathbf{T}'\Lambda\mathbf{T}\mathbf{T}'\Lambda\mathbf{T} \cdots \mathbf{T}'\Lambda\mathbf{T} = \mathbf{T}'\Lambda^s\mathbf{T}$. We again apply A.1.3. When \mathbf{A} is nonsingular, $\mathbf{T}'\mathbf{A}^{-1}\mathbf{T} = (\mathbf{T}'\mathbf{A}\mathbf{T})^{-1} = \Lambda^{-1}$ and the eigenvalues of \mathbf{A}^{-1} are λ^{-1} . We then apply A.1.3. Also, $\mathbf{T}'(\mathbf{I}_n + c\mathbf{A})\mathbf{T} = \mathbf{I}_n + c\mathbf{T}'\mathbf{A}\mathbf{T} = \mathbf{I}_n + c\Lambda$, which is a diagonal matrix with elements $1 + c\lambda_i$.

A.2 RANK

$$\text{A.2.1. If } \mathbf{A} \text{ and } \mathbf{B} \text{ are conformable matrices, then}$$

$$\text{rank}(\mathbf{AB}) \leq \min(\text{rank } \mathbf{A}, \text{rank } \mathbf{B}).$$

Proof. The rows of \mathbf{AB} are linear combinations of the rows of \mathbf{B} , so that the number of linear independent rows of \mathbf{AB} is less than or equal to those of \mathbf{B} ; thus $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$. Similarly, the columns of \mathbf{AB} are linear combinations of the columns of \mathbf{A} , so that $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$.

$$\text{A.2.2. If } \mathbf{A} \text{ is any matrix, and } \mathbf{P} \text{ and } \mathbf{Q} \text{ are any conformable nonsingular matrices, then } \text{rank}(\mathbf{PAQ}) = \text{rank}(\mathbf{A}).$$

Proof. $\text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{AQ}) \leq \text{rank}(\mathbf{AQ}\mathbf{Q}^{-1}) = \text{rank}(\mathbf{A})$, so that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{AQ})$, etc.

$$\text{A.2.3. Let } \mathbf{A} \text{ be any } m \times n \text{ matrix such that } r = \text{rank}(\mathbf{A}) \text{ and } s = \text{nullity}(\mathbf{A}), [\text{the dimension of } \mathcal{N}(\mathbf{A}), \text{ the null space or kernel of } \mathbf{A}, \text{ i.e., the dimension of } \{\mathbf{x} : \mathbf{Ax} = \mathbf{0}\}]. \text{ Then}$$

$$r + s = n.$$

Proof. Let $\alpha_1, \alpha_2, \dots, \alpha_s$ be a basis for $\mathcal{N}(\mathbf{A})$. Enlarge this set of vectors to give a basis $\alpha_1, \alpha_2, \dots, \alpha_s, \beta_1, \beta_2, \dots, \beta_t$ for \Re_n , n -dimensional

Euclidean space. Every vector in $\mathcal{C}(\mathbf{A})$, the column space of \mathbf{A} (the space spanned by the columns of \mathbf{A}), can be expressed in the form

$$\begin{aligned}\mathbf{Ax} &= \mathbf{A} \left(\sum_{i=1}^s a_i \alpha_i + \sum_{j=1}^t b_j \beta_j \right) \\ &= \mathbf{A} \sum_{j=1}^t b_j \beta_j \\ &= \sum_{j=1}^t b_j \mathbf{A} \beta_j \\ &= \sum_{j=1}^t b_j \gamma_j,\end{aligned}$$

say. Now suppose that

$$\sum_{j=1}^t c_j \gamma_j = \mathbf{0};$$

then

$$\mathbf{A} \left(\sum_{j=1}^t c_j \beta_j \right) = \sum_{j=1}^t c_j \gamma_j = \mathbf{0}$$

and $\sum c_j \beta_j \in \mathcal{N}(\mathbf{A})$. This is possible only if $c_1 = c_2 = \dots = c_t = 0$, that is, $\gamma_1, \gamma_2, \dots, \gamma_t$ are linearly independent. Since every vector \mathbf{Ax} in $\mathcal{C}(\mathbf{A})$ can be expressed in terms of the γ_j 's, the γ_j 's form a basis for $\mathcal{C}(\mathbf{A})$; thus $t = r$. Since $s + t = n$, our proof is complete.

A.2.4. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A}'\mathbf{A}) = \text{rank}(\mathbf{AA}')$.

Proof. $\mathbf{Ax} = \mathbf{0} \Rightarrow \mathbf{A}'\mathbf{Ax} = \mathbf{0}$ and $\mathbf{A}'\mathbf{Ax} = \mathbf{0} \Rightarrow \mathbf{x}'\mathbf{A}'\mathbf{Ax} = 0 \Rightarrow \mathbf{Ax} = \mathbf{0}$. Hence the nullspaces of \mathbf{A} and $\mathbf{A}'\mathbf{A}$ are the same. Since \mathbf{A} and $\mathbf{A}'\mathbf{A}$ have the same number of columns, it follows from A.2.3 that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A})$. Similarly, $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A}'\mathbf{A})$ and the result follows.

A.2.5. If $\mathcal{C}(\mathbf{A})$ is the column space of \mathbf{A} (the space spanned by the columns of \mathbf{A}), then $\mathcal{C}(\mathbf{A}'\mathbf{A}) = \mathcal{C}(\mathbf{A}')$.

Proof. $\mathbf{A}'\mathbf{A}\mathbf{a} = \mathbf{A}'\mathbf{b}$ for $\mathbf{b} = \mathbf{A}\mathbf{a}$, so that $\mathcal{C}(\mathbf{A}'\mathbf{A}) \subset \mathcal{C}(\mathbf{A}')$. However, by A.2.4, these two spaces must be the same, as they have the same dimension.

A.2.6. If \mathbf{A} is symmetric, then $\text{rank}(\mathbf{A})$ is equal to the number of nonzero eigenvalues.

Proof. By A.2.2, $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{T}'\mathbf{AT}) = \text{rank}(\mathbf{A})$.

- A.2.7. Any $n \times n$ symmetric matrix \mathbf{A} has a set of n orthonormal eigenvectors, and $\mathcal{C}(\mathbf{A})$ is the space spanned by those eigenvectors corresponding to nonzero eigenvalues.

Proof. From $\mathbf{T}'\mathbf{AT} = \Lambda$ we have $\mathbf{AT} = \mathbf{T}\Lambda$ or $\mathbf{At}_i = \lambda_i \mathbf{t}_i$, where $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_n)$; the \mathbf{t}_i are orthonormal, as \mathbf{T} is an orthogonal matrix. Suppose that $\lambda_i = 0$ ($i = r+1, r+2, \dots, n$) and $\mathbf{x} = \sum_{i=1}^n a_i \mathbf{t}_i$. Then

$$\mathbf{Ax} = \mathbf{A} \sum_{i=1}^n a_i \mathbf{t}_i = \sum_{i=1}^n a_i \mathbf{At}_i = \sum_{i=1}^r a_i \lambda_i \mathbf{t}_i = \sum_{i=1}^r b_i \mathbf{t}_i,$$

and $\mathcal{C}(\mathbf{A})$ is spanned by $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_r$.

A.3 POSITIVE-SEMICDEFINITE MATRICES

A symmetric matrix \mathbf{A} is said to be positive-semidefinite¹ (p.s.d.) if and only if $\mathbf{x}'\mathbf{Ax} \geq 0$ for all \mathbf{x} .

- A.3.1. The eigenvalues of a p.s.d. matrix are nonnegative.

Proof. If $\mathbf{T}'\mathbf{AT} = \Lambda$, then substituting $\mathbf{x} = \mathbf{Ty}$, we have $\mathbf{x}'\mathbf{Ax} = \mathbf{y}'\mathbf{T}'\mathbf{ATy} = \lambda_1 y_1^2 + \dots + \lambda_n y_n^2 \geq 0$. Setting $y_j = \delta_{ij}$ leads to $0 \leq \mathbf{x}'\mathbf{Ax} = \lambda_i$.

- A.3.2. If \mathbf{A} is p.s.d., then $\text{tr}(\mathbf{A}) \geq 0$. This follows from A.3.1 and A.1.3.

- A.3.3. \mathbf{A} is p.s.d. of rank r if and only if there exists an $n \times n$ matrix \mathbf{R} of rank r such that $\mathbf{A} = \mathbf{RR}'$.

Proof. Given \mathbf{A} is p.s.d. of rank r , then, by A.2.6 and A.3.1, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r, 0, \dots, 0)$, where $\lambda_i > 0$ ($i = 1, 2, \dots, r$). Let $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_r^{1/2}, 0, \dots, 0)$, then $\mathbf{T}'\mathbf{AT} = \Lambda$ implies that $\mathbf{A} = \mathbf{T}\Lambda^{1/2}\Lambda^{1/2}\mathbf{T}' = \mathbf{RR}'$, where $\text{rank}(\mathbf{R}) = \text{rank}(\Lambda^{1/2}) = r$. Conversely, if $\mathbf{A} = \mathbf{RR}'$, then $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{R}) = r$ (A.2.4) and $\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{RR}'\mathbf{x} = \mathbf{y}'\mathbf{y} \geq 0$, where $\mathbf{y} = \mathbf{R}'\mathbf{x}$.

- A.3.4. If \mathbf{A} is an $n \times n$ p.s.d. matrix of rank r , then there exists an $n \times r$ matrix \mathbf{S} of rank r such that $\mathbf{S}'\mathbf{AS} = \mathbf{I}_r$.

Proof. From

$$\mathbf{T}'\mathbf{AT} = \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix}$$

we have $\mathbf{T}_1'\mathbf{AT}_1 = \Lambda_r$, where \mathbf{T}_1 consists of the first r columns of \mathbf{T} . Setting $\mathbf{S} = \mathbf{T}_1\Lambda_r^{1/2}$ leads to the required result.

¹Some authors use the term *nonnegative-definite*.

A.3.5. If \mathbf{A} is p.s.d., then $\mathbf{X}'\mathbf{AX} = \mathbf{0} \Rightarrow \mathbf{AX} = \mathbf{0}$.

Proof. From A.3.3, $\mathbf{0} = \mathbf{X}'\mathbf{AX} = \mathbf{X}'\mathbf{RR}'\mathbf{X} = \mathbf{B}'\mathbf{B}$ ($\mathbf{B} = \mathbf{R}'\mathbf{X}$), which implies that $\mathbf{b}_i'\mathbf{b}_i = 0$; that is, $\mathbf{b}_i = \mathbf{0}$ for every column \mathbf{b}_i of \mathbf{B} . Hence $\mathbf{AX} = \mathbf{RB} = \mathbf{0}$.

A.4 POSITIVE-DEFINITE MATRICES

A symmetric matrix \mathbf{A} is said to be positive-definite (p.d.) if $\mathbf{x}'\mathbf{Ax} > 0$ for all \mathbf{x} , $\mathbf{x} \neq \mathbf{0}$. We note that a p.d. matrix is also p.s.d.

A.4.1. The eigenvalues of a p.d. matrix \mathbf{A} are all positive (proof is similar to A.3.1); thus \mathbf{A} is also nonsingular (A.2.6).

A.4.2. \mathbf{A} is p.d. if and only if there exists a nonsingular \mathbf{R} such that $\mathbf{A} = \mathbf{RR}'$.

Proof. This follows from A.3.3 with $r = n$.

A.4.3. If \mathbf{A} is p.d., then so is \mathbf{A}^{-1} .

Proof. $\mathbf{A}^{-1} = (\mathbf{RR}')^{-1} = (\mathbf{R}')^{-1}\mathbf{R}^{-1} = (\mathbf{R}^{-1})'\mathbf{R}^{-1} = \mathbf{SS}'$, where \mathbf{S} is nonsingular. The result then follows from A.4.2 above.

A.4.4. If \mathbf{A} is p.d., then $\text{rank}(\mathbf{CAC}') = \text{rank}(\mathbf{C})$.

Proof.

$$\begin{aligned}\text{rank}(\mathbf{CAC}') &= \text{rank}(\mathbf{CR}\mathbf{R}'\mathbf{C}') \\ &= \text{rank}(\mathbf{CR}) \quad (\text{by A.2.4}) \\ &= \text{rank}(\mathbf{C}) \quad (\text{by A.2.2}).\end{aligned}$$

A.4.5. If \mathbf{A} is an $n \times n$ p.d. matrix and \mathbf{C} is $p \times n$ of rank p , then \mathbf{CAC}' is p.d.

Proof. $\mathbf{x}'\mathbf{CAC}'\mathbf{x} = \mathbf{y}'\mathbf{Ay} \geq 0$ with equality $\Leftrightarrow \mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{C}'\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0}$ (since the columns of \mathbf{C}' are linearly independent). Hence $\mathbf{x}'\mathbf{CAC}'\mathbf{x} > 0$ all \mathbf{x} , $\mathbf{x} \neq \mathbf{0}$.

A.4.6. If \mathbf{X} is $n \times p$ of rank p , then $\mathbf{X}'\mathbf{X}$ is p.d.

Proof. $\mathbf{x}'\mathbf{X}'\mathbf{X}\mathbf{x} = \mathbf{y}'\mathbf{y} \geq 0$ with equality $\Leftrightarrow \mathbf{X}\mathbf{x} = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{0}$ (since the columns of \mathbf{X} are linearly independent).

A.4.7. \mathbf{A} is p.d. if and only if all the leading minor determinants of \mathbf{A} [including $\det(\mathbf{A})$ itself] are positive.

Proof. If \mathbf{A} is p.d., then

$$\det(\mathbf{A}) = \det(\mathbf{T}\Lambda\mathbf{T}') = \det(\Lambda) = \prod_i \lambda_i > 0 \quad (\text{by A.4.1}).$$

Let

$$\mathbf{A}_r = \begin{pmatrix} a_{11} & \cdots & a_{1r} \\ \cdots & \cdots & \cdots \\ a_{r1} & \cdots & a_{rr} \end{pmatrix} \quad \text{and} \quad \mathbf{x}_r = \begin{pmatrix} x_1 \\ \cdots \\ x_r \end{pmatrix};$$

then

$$\mathbf{x}'_r \mathbf{A}_r \mathbf{x}_r = (\mathbf{x}'_r, \mathbf{0}') \mathbf{A} \begin{pmatrix} \mathbf{x}_r \\ \mathbf{0} \end{pmatrix} > 0 \quad \text{for } \mathbf{x}_r \neq \mathbf{0}$$

and \mathbf{A}_r is positive definite. Hence if \mathbf{A} is $n \times n$, it follows from the argument above that $\det(\mathbf{A}_r) > 0$ ($r = 1, 2, \dots, n$). Conversely, suppose that all the leading minor determinants of \mathbf{A} are positive; then we wish to show that \mathbf{A} is p.d. Let

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{n-1}, & \mathbf{c} \\ \mathbf{c}', & a_{nn} \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} \mathbf{I}_{n-1}, & \boldsymbol{\alpha} \\ \mathbf{0}', & -1 \end{pmatrix},$$

where $\boldsymbol{\alpha} = \mathbf{A}_{n-1}^{-1} \mathbf{c}$. Then

$$\mathbf{R}' \mathbf{A} \mathbf{R} = \begin{pmatrix} \mathbf{A}_{n-1}, & \mathbf{0} \\ \mathbf{0}', & k \end{pmatrix},$$

where

$$k = \det(\mathbf{R}' \mathbf{A} \mathbf{R}) / \det(\mathbf{A}_{n-1}) = \det(\mathbf{R})^2 \det(\mathbf{A}) / \det(\mathbf{A}_{n-1}) > 0,$$

since \mathbf{R} is nonsingular. We now proceed by induction. The result is trivially true for $n = 1$; assume that it is true for matrices of orders up to $n - 1$. If we set $\mathbf{y} = \mathbf{R}^{-1} \mathbf{x}$ ($\mathbf{x} \neq \mathbf{0}$), $\mathbf{x}' \mathbf{A} \mathbf{x} = \mathbf{y}' \mathbf{R}' \mathbf{A} \mathbf{R} \mathbf{y} = \mathbf{y}'_{n-1} \mathbf{A}_{n-1} \mathbf{y}_{n-1} + k y_n^2 > 0$, since \mathbf{A}_{n-1} is p.d. by the inductive hypothesis and $\mathbf{y} \neq \mathbf{0}$. Hence the result is true for matrices of order n .

A.4.8. The diagonal elements of a p.d. matrix are all positive.

Proof. Setting $x_j = \delta_{ij}$ ($j = 1, 2, \dots, n$), we have $0 < \mathbf{x}' \mathbf{A} \mathbf{x} = a_{ii}$.

A.4.9. If \mathbf{A} is an $n \times n$ p.d. matrix and \mathbf{B} is an $n \times n$ symmetric matrix, then $\mathbf{A} - t\mathbf{B}$ is p.d. for $|t|$ sufficiently small.

Proof. The i th leading minor determinant of $\mathbf{A} - t\mathbf{B}$ is a function of t , which is positive when $t = 0$ (by A.4.7 above). Since this function is continuous, it will be positive for $|t| < \delta_i$ for δ_i sufficiently small. Let $\delta = \min(\delta_1, \delta_2, \dots, \delta_n)$; then all the leading minor determinants will be positive for $|t| < \delta$, and the result follows from A.4.7.

A.4.10. (Cholesky decomposition) If \mathbf{A} is p.d., there exists a unique upper triangular matrix \mathbf{R} with positive diagonal elements such that $\mathbf{A} = \mathbf{R}' \mathbf{R}$.

Proof. We proceed by induction and assume that the unique factorization holds for matrices of orders up to $n - 1$. Thus

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} \mathbf{A}_{n-1}, & \mathbf{c} \\ \mathbf{c}', & a_{nn} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{R}'_{n-1} \mathbf{R}_{n-1}, & \mathbf{c} \\ \mathbf{c}', & a_{nn} \end{pmatrix},\end{aligned}$$

where \mathbf{R}_{n-1} is a unique upper triangular matrix of order $n - 1$ with positive diagonal elements. Since the determinant of a triangular matrix is the product of its diagonal elements, \mathbf{R}_{n-1} is nonsingular and we can define

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{n-1}, & \mathbf{d} \\ \mathbf{0}', & \sqrt{k} \end{pmatrix},$$

where $\mathbf{d} = (\mathbf{R}'_{n-1})^{-1} \mathbf{c}$ and $k = a_{nn} - \mathbf{d}' \mathbf{d}$. Since \mathbf{R} is unique and $\mathbf{A} = \mathbf{R}' \mathbf{R}$, we have the required decomposition of \mathbf{A} provided that $k > 0$. Taking determinants,

$$\det(\mathbf{A}) = \det(\mathbf{R}' \mathbf{R}) = \det(\mathbf{R})^2 = \det(\mathbf{R}_{n-1})^2 k,$$

so that k is positive as $\det(\mathbf{A}) > 0$ (A.4.7) and $\det(\mathbf{R}_{n-1}) \neq 0$. Thus the factorization also holds for positive definite matrices of order n .

A.4.11. If \mathbf{L} is positive-definite, then for any \mathbf{b} ,

$$\max_{\mathbf{h}: \mathbf{h} \neq \mathbf{0}} \left\{ \frac{(\mathbf{h}' \mathbf{b})^2}{\mathbf{h}' \mathbf{L} \mathbf{h}} \right\} = \mathbf{b}' \mathbf{L}^{-1} \mathbf{b}.$$

Proof. For all a ,

$$\begin{aligned}0 &\leq \|(\mathbf{v} - a\mathbf{u})\|^2 \\ &= a^2 \|\mathbf{u}\|^2 - 2a\mathbf{u}' \mathbf{v} + \|\mathbf{v}\|^2 \\ &= \left(a\|\mathbf{u}\| - \frac{\mathbf{u}' \mathbf{v}}{\|\mathbf{u}\|} \right)^2 + \|\mathbf{v}\|^2 - \frac{(\mathbf{u}' \mathbf{v})^2}{\|\mathbf{u}\|^2}.\end{aligned}$$

Hence given $\mathbf{u} \neq \mathbf{0}$ and setting $a = \mathbf{u}' \mathbf{v} / \|\mathbf{u}\|^2$, we have the Cauchy-Schwartz inequality

$$(\mathbf{u}' \mathbf{v})^2 \leq \|\mathbf{v}\|^2 \|\mathbf{u}\|^2$$

with equality if and only if $\mathbf{v} = a\mathbf{u}$ for some a . Hence

$$\max_{\mathbf{v}: \mathbf{v} \neq \mathbf{0}} \left\{ \frac{(\mathbf{u}' \mathbf{v})^2}{\mathbf{v}' \mathbf{v}} \right\} = \mathbf{u}' \mathbf{u}.$$

Because \mathbf{L} is positive-definite, there exists a nonsingular matrix \mathbf{R} such that $\mathbf{L} = \mathbf{R} \mathbf{R}'$ (A.4.2). Setting $\mathbf{v} = \mathbf{R}' \mathbf{h}$ and $\mathbf{u} = \mathbf{R}^{-1} \mathbf{b}$ leads to the required result.

- A.4.12. (*Square root of a positive-definite matrix*) If \mathbf{A} is p.d., there exists a p.d. square root $\mathbf{A}^{1/2}$ such that $(\mathbf{A}^{1/2})^2 = \mathbf{A}$.

Proof. Let $\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}'$ be the spectral decomposition of \mathbf{A} (A.1.4), where the diagonal elements of Λ are positive (by A.4.1). Let $\mathbf{A}^{1/2} = \mathbf{T}\Lambda^{1/2}\mathbf{T}'$; then $\mathbf{T}\Lambda^{1/2}\mathbf{T}'\mathbf{T}\Lambda^{1/2}\mathbf{T}' = \mathbf{T}\Lambda\mathbf{T}' = \mathbf{A}$ (since $\mathbf{T}'\mathbf{T} = \mathbf{I}_n$).

A.5 PERMUTATION MATRICES

Let Π_{ij} be the identity matrix with its i th and j th rows interchanged. Then $\Pi_{ij}^2 = \mathbf{I}$, so that Π_{ij} is a symmetric and orthogonal matrix. Premultiplying any matrix by Π_{ij} will interchange its i th and j th rows so that Π_{ij} is an (elementary) permutation matrix. Postmultiplying a matrix by an elementary permutation matrix will interchange two columns.

Any reordering of the rows of a matrix can be done using a sequence of elementary permutations $\Pi = \Pi_{i_K j_K} \cdots \Pi_{i_1 j_1}$, where

$$\Pi\Pi' = \Pi_{i_K j_K} \cdots \Pi_{i_1 j_1} \Pi_{i_1 j_1} \cdots \Pi_{i_K j_K} = \mathbf{I}.$$

The orthogonal matrix Π is called a *permutation matrix*.

A.6 IDEMPOTENT MATRICES

A matrix \mathbf{P} is idempotent if $\mathbf{P}^2 = \mathbf{P}$. A symmetric idempotent matrix is called a *projection matrix*.

- A.6.1. If \mathbf{P} is symmetric, then \mathbf{P} is idempotent and of rank r if and only if it has r eigenvalues equal to unity and $n - r$ eigenvalues equal to zero.

Proof. Given $\mathbf{P}^2 = \mathbf{P}$, the $\mathbf{P}\mathbf{x} = \lambda\mathbf{x}$ ($\mathbf{x} \neq \mathbf{0}$) implies that $\lambda\mathbf{x}'\mathbf{x} = \mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{x}'\mathbf{P}^2\mathbf{x} = (\mathbf{P}\mathbf{x})'(\mathbf{P}\mathbf{x}) = \lambda^2\mathbf{x}'\mathbf{x}$, and $\lambda(\lambda - 1) = 0$. Hence the eigenvalues are 0 or 1 and, by A.2.6, \mathbf{P} has r eigenvalues equal to unity and $n - r$ eigenvalues equal to zero. Conversely, if the eigenvalues are 0 or 1, then we can assume without loss of generality that the first r eigenvalues are unity. Hence there exists an orthogonal matrix \mathbf{T} such that

$$\mathbf{T}'\mathbf{P}\mathbf{T} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \Lambda, \text{ or } \mathbf{P} = \mathbf{T}\Lambda\mathbf{T}'.$$

Therefore, $\mathbf{P}^2 = \mathbf{T}\Lambda\mathbf{T}'\mathbf{T}\Lambda\mathbf{T}' = \mathbf{T}\Lambda^2\mathbf{T}' = \mathbf{T}\Lambda\mathbf{T}' = \mathbf{P}$, and $\text{rank}(\mathbf{P}) = r$ (A.2.2).

- A.6.2. If \mathbf{P} is a projection matrix, then $\text{tr}(\mathbf{P}) = \text{rank}(\mathbf{P})$.

Proof. If $\text{rank}(\mathbf{P}) = r$, then, by A.6.1 above, \mathbf{P} has r unit eigenvalues and $n - r$ zero eigenvalues. Hence $\text{tr}(\mathbf{P}) = r$ (by A.1.3).

A.6.3. If \mathbf{P} is idempotent, so is $\mathbf{I} - \mathbf{P}$.

$$\text{Proof. } (\mathbf{I} - \mathbf{P})^2 = \mathbf{I} - 2\mathbf{P} + \mathbf{P}^2 = \mathbf{I} - 2\mathbf{P} + \mathbf{P} = \mathbf{I} - \mathbf{P}.$$

A.6.4. Projection matrices are positive-semidefinite.

$$\text{Proof. } \mathbf{x}'\mathbf{P}\mathbf{x} = \mathbf{x}'\mathbf{P}^2\mathbf{x} = (\mathbf{Px})'(\mathbf{Px}) \geq 0.$$

A.6.5. If \mathbf{P}_i ($i = 1, 2$) is a projection matrix and $\mathbf{P}_1 - \mathbf{P}_2$ is p.s.d., then

- (a) $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_2$,
- (b) $\mathbf{P}_1 - \mathbf{P}_2$ is a projection matrix.

Proof. (a) Given $\mathbf{P}_1\mathbf{x} = \mathbf{0}$, then $0 \leq \mathbf{x}'(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{x} = -\mathbf{x}'\mathbf{P}_2\mathbf{x}$. Since \mathbf{P}_2 is positive-semidefinite (A.6.4), $\mathbf{x}'\mathbf{P}_2\mathbf{x} = 0$ and $\mathbf{P}_2\mathbf{x} = \mathbf{0}$. Hence for any \mathbf{y} , and $\mathbf{x} = (\mathbf{I} - \mathbf{P}_1)\mathbf{y}$, $\mathbf{P}_2(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \mathbf{0}$ as $\mathbf{P}_1(\mathbf{I} - \mathbf{P}_1)\mathbf{y} = \mathbf{0}$. Thus $\mathbf{P}_2\mathbf{P}_1\mathbf{y} = \mathbf{P}_2\mathbf{y}$, which implies that $\mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_2$ (A.11.1). Taking transposes leads to $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2$, and (a) is proved.

$$(b) (\mathbf{P}_1 - \mathbf{P}_2)^2 = \mathbf{P}_1^2 - \mathbf{P}_1\mathbf{P}_2 - \mathbf{P}_2\mathbf{P}_1 + \mathbf{P}_2^2 = \mathbf{P}_1 - \mathbf{P}_2 - \mathbf{P}_2 + \mathbf{P}_2 = \mathbf{P}_1 - \mathbf{P}_2.$$

A.7 EIGENVALUE APPLICATIONS

A.7.1. For conformable matrices, the nonzero eigenvalues of \mathbf{AB} are the same as those of \mathbf{BA} . The eigenvalues are identical for square matrices.

Proof. Let λ be a nonzero eigenvalue of \mathbf{AB} . Then there exists \mathbf{u} ($\mathbf{u} \neq \mathbf{0}$) such that $\mathbf{ABu} = \lambda\mathbf{u}$; that is, $\mathbf{BABu} = \lambda\mathbf{Bu}$. Hence $\mathbf{BAv} = \lambda\mathbf{v}$, where $\mathbf{v} = \mathbf{Bu} \neq \mathbf{0}$ (as $\mathbf{ABu} \neq \mathbf{0}$), and λ is an eigenvalue of \mathbf{BA} . The argument reverses by interchanging the roles of \mathbf{A} and \mathbf{B} . For square matrices \mathbf{AB} and \mathbf{BA} have the same number of zero eigenvalues.

A.7.2. Let \mathbf{A} be an $n \times n$ symmetric matrix; then

$$\max_{\mathbf{x}: \mathbf{x} \neq \mathbf{0}} \left\{ \frac{(\mathbf{x}'\mathbf{Ax})}{\mathbf{x}'\mathbf{x}} \right\} = \lambda_{\text{MAX}}$$

and

$$\min_{\mathbf{x}: \mathbf{x} \neq \mathbf{0}} \left\{ \frac{(\mathbf{x}'\mathbf{Ax})}{\mathbf{x}'\mathbf{x}} \right\} = \lambda_{\text{MIN}},$$

where λ_{MIN} and λ_{MAX} are the minimum and maximum eigenvalues of \mathbf{A} . These values occur when \mathbf{x} is the eigenvector corresponding to the eigenvalues λ_{MIN} and λ_{MAX} , respectively.

Proof. Let $\mathbf{A} = \mathbf{T}\Lambda\mathbf{T}'$ be the spectral decomposition of \mathbf{A} (cf. A.1.4) and suppose that $\lambda_1 = \lambda_{\text{MAX}}$. If $\mathbf{y} \neq \mathbf{0}$ and $\mathbf{x} = \mathbf{Ty} = (t_1, \dots, t_n)\mathbf{y}$, then

$$\frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{y}'\mathbf{T}'\mathbf{A}\mathbf{T}\mathbf{y}}{\mathbf{y}'\mathbf{T}'\mathbf{T}\mathbf{y}} = \frac{\sum_i \lambda_i y_i^2}{\sum_i y_i^2} \leq \frac{\lambda_1 y'y}{y'y} = \lambda_1,$$

with equality when $y_1 = 1$, and $y_2 = y_3 = \dots = y_n = 0$, that is, when $\mathbf{x} = \mathbf{t}_1$. The second result follows in a similar fashion.

Since the ratio $\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x}$ is independent of the scale of \mathbf{x} , we can set $\mathbf{x}'\mathbf{x} = 1$, giving us

$$\max_{\mathbf{x}: \mathbf{x}'\mathbf{x}=1} (\mathbf{x}'\mathbf{A}\mathbf{x}) = \lambda_{\text{MAX}},$$

with a similar result for λ_{MIN} .

A.8 VECTOR DIFFERENTIATION

If $\frac{d}{d\beta} = \left(\frac{d}{d\beta_i} \right)$, then:

$$\text{A.8.1. } \frac{d(\beta' \mathbf{a})}{d\beta} = \mathbf{a}.$$

$$\text{A.8.2. } \frac{d(\beta' \mathbf{A} \beta)}{d\beta} = 2\mathbf{A}\beta \quad (\mathbf{A} \text{ symmetric}).$$

Proof. (1) is trivial. For (2),

$$\begin{aligned} \frac{d(\beta' \mathbf{A} \beta)}{d\beta_i} &= \frac{d}{d\beta_i} \left(\sum_i \sum_j a_{ij} \beta_i \beta_j \right) \\ &= 2a_{ii} \beta_i + 2 \sum_{j \neq i} a_{ij} \beta_j \\ &= 2 \sum_j a_{ij} \beta_j \\ &= 2(\mathbf{A}\beta)_i. \end{aligned}$$

A.9 PATTERNED MATRICES

A.9.1. If all inverses exist,

$$\begin{aligned} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{B}_{12} \mathbf{B}_{22}^{-1} \mathbf{B}_{21} & -\mathbf{B}_{12} \mathbf{B}_{22}^{-1} \\ -\mathbf{B}_{22}^{-1} \mathbf{B}_{21} & \mathbf{B}_{22}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{C}_{11}^{-1} & -\mathbf{C}_{11}^{-1} \mathbf{C}_{12} \\ -\mathbf{C}_{21} \mathbf{C}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{C}_{12} \end{pmatrix}, \end{aligned}$$

where $\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}$, $\mathbf{B}_{12} = \mathbf{A}_{11}^{-1} \mathbf{A}_{12}$, $\mathbf{B}_{21} = \mathbf{A}_{21} \mathbf{A}_{11}^{-1}$, $\mathbf{C}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}$, $\mathbf{C}_{12} = \mathbf{A}_{12} \mathbf{A}_{22}^{-1}$, and $\mathbf{C}_{21} = \mathbf{A}_{22}^{-1} \mathbf{A}_{21}$.

Proof. As the inverse of a matrix is unique, we only have to check that the matrix times its inverse is the identity matrix.

- A.9.2. Let \mathbf{W} be an $n \times p$ matrix of rank p with columns $\mathbf{w}^{(j)}$ ($j = 1, 2, \dots, p$). Then

$$(\mathbf{WW})_{jj}^{-1} = [\mathbf{w}^{(j)'}(\mathbf{I}_n - \mathbf{P}_j)\mathbf{w}^{(j)}]^{-1},$$

where $\mathbf{P}_j = \mathbf{W}^{(j)}(\mathbf{W}^{(j)'}\mathbf{W}^{(j)})^{-1}\mathbf{W}^{(j)'}$, and $\mathbf{W}^{(j)}$ is \mathbf{W} with its j th column omitted.

Proof. From the first equation of A.9.1,

$$\begin{aligned} (\mathbf{W}'\mathbf{W})^{-1} &= \begin{pmatrix} \mathbf{W}^{(p)'}\mathbf{W}^{(p)} & \mathbf{W}^{(p)'}\mathbf{w}^{(p)} \\ \mathbf{w}^{(p)'}\mathbf{W}^{(p)} & \mathbf{w}^{(p)'}\mathbf{w}^{(p)} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{F} & \mathbf{g} \\ \mathbf{g}' & h \end{pmatrix}, \end{aligned}$$

where $h = (\mathbf{W}'\mathbf{W})_{pp}^{-1} = (\mathbf{w}^{(p)'}\mathbf{w}^{(p)} - \mathbf{w}^{(p)'}\mathbf{P}_p\mathbf{w}^{(p)})^{-1}$. Let Π be the permutation matrix \mathbf{I}_n with its j th and p th columns interchanged. Then $\Pi^2 = \mathbf{I}_n$, so that Π is a symmetric orthogonal matrix, and its own inverse. Hence

$$\begin{aligned} (\mathbf{W}'\mathbf{W})^{-1} &= \Pi(\Pi\mathbf{W}'\mathbf{W}\Pi)^{-1}\Pi \\ &= \Pi \begin{pmatrix} \mathbf{F}_1 & \mathbf{g}_1 \\ \mathbf{g}_1' & h_1 \end{pmatrix} \Pi, \end{aligned}$$

where $h_1 = (\mathbf{WW})_{jj}^{-1}$. We have thus effectively interchanged $\mathbf{w}^{(p)}$ and $\mathbf{w}^{(j)}$, and $\mathbf{W}^{(p)}$ and $\mathbf{W}^{(j)}$. The result then follows.

- A.9.3. (*Sherman–Morrison–Woodbury formula*) Let \mathbf{A} and \mathbf{B} be nonsingular $m \times m$ and $n \times m$ matrices, respectively, and let \mathbf{U} be $m \times n$ and \mathbf{V} be $n \times m$. Then

$$(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\mathbf{B}(\mathbf{B} + \mathbf{B}\mathbf{V}\mathbf{A}^{-1}\mathbf{U}\mathbf{B})^{-1}\mathbf{B}\mathbf{V}\mathbf{A}^{-1}.$$

Proof. Multiply the right-hand side on the left by $\mathbf{A} + \mathbf{UBV}$ and simplify to get \mathbf{I}_n .

- A.9.4. Setting $\mathbf{B} = \mathbf{I}_n$, $\mathbf{U} = \pm \mathbf{u}$, and $\mathbf{V} = \mathbf{v}'$ in A.9.3, we have

$$(\mathbf{A} + \mathbf{uv}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uv}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}},$$

and

$$(\mathbf{A} - \mathbf{uv}')^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{uv}'\mathbf{A}^{-1}}{1 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}.$$

A.10 GENERALIZED INVERSE

A *generalized inverse* of an $m \times n$ matrix \mathbf{B} is defined to be any $n \times m$ matrix \mathbf{B}^- that satisfies the condition

$$(a) \quad \mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}.$$

Such a matrix always exists (Searle [1971: Chapter 1]). The name *generalized inverse* for \mathbf{B}^- defined by (a) is not accepted universally, although it is used fairly widely (e.g., Rao [1973], Rao and Mitra [1971a,b], Pringle and Rayner [1971], Searle [1971], Kruskal [1975]). Other names such as *conditional inverse*, *pseudo inverse*, *g-inverse*, and *p-inverse* are also found in the literature, sometimes for \mathbf{B}^- defined above and sometimes for matrices defined as variants of \mathbf{B}^- . It should be noted that \mathbf{B}^- is called “a” generalized inverse, not “the” generalized inverse, for \mathbf{B}^- is not unique. Also, taking the transpose of (a), we have

$$\mathbf{B}' = \mathbf{B}'(\mathbf{B}^-)' \mathbf{B}',$$

so that \mathbf{B}^-' is a generalized inverse of \mathbf{B}' ; we can therefore write

$$(\mathbf{B}^-)' = (\mathbf{B}')^-$$

for some $(\mathbf{B}')^-$.

If \mathbf{B}^- also satisfies three more conditions, namely,

$$(b) \quad \mathbf{B}^-\mathbf{B}\mathbf{B}^- = \mathbf{B}^-,$$

$$(c) \quad (\mathbf{B}\mathbf{B}^-)' = \mathbf{B}\mathbf{B}^-,$$

$$(d) \quad (\mathbf{B}^-\mathbf{B})' = \mathbf{B}^-\mathbf{B},$$

then \mathbf{B}^- is unique and it is called the *Moore-Penrose inverse* (Albert [1972]); some authors call it the *pseudo inverse* or the *p-inverse*. We denote this inverse by \mathbf{B}^+ .

If the regression matrix \mathbf{X} is less than full rank, the normal equations $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$ do not have a unique solution. Setting $\mathbf{B} = \mathbf{X}'\mathbf{X}$ and $\mathbf{c} = \mathbf{X}'\mathbf{Y}$, we have

$$\begin{aligned} \mathbf{c} &= \mathbf{B}\beta \\ &= \mathbf{B}\mathbf{B}^-\mathbf{B}\beta \\ &= \mathbf{B}(\mathbf{B}^-\mathbf{c}) \end{aligned}$$

and $\mathbf{B}^-\mathbf{c}$ is a solution of $\mathbf{B}\beta = \mathbf{c}$. (In fact, it can be shown that every solution of $\mathbf{B}\beta = \mathbf{c}$ can be expressed in the form $\mathbf{B}^-\mathbf{c}$ for some \mathbf{B}^- .) There are several ways of computing a suitable \mathbf{B}^- for the symmetric matrix \mathbf{B} . One method is as follows:

- (i) Delete $p - r$ rows and the corresponding columns so as to leave an $r \times r$ matrix that is nonsingular; this can always be done, as $\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}) = r$.

- (ii) Invert the $r \times r$ matrix.
- (iii) Obtain \mathbf{B}^- by inserting zeros into the inverse to correspond to the rows and columns originally deleted. For example, if

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

and \mathbf{B}_{11} is an $r \times r$ nonsingular matrix, then

$$\mathbf{B}^- = \begin{pmatrix} \mathbf{B}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

We saw above that

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{B}^- \mathbf{c} \\ &= (\mathbf{X}' \mathbf{X})^- \mathbf{X}' \mathbf{Y} \\ &= \mathbf{X}^* \mathbf{Y},\end{aligned}$$

say, is a solution of the normal equations. Because

$$\begin{aligned}(\mathbf{X}^+)' \mathbf{X}' \mathbf{X} &= (\mathbf{X} \mathbf{X}^+)' \mathbf{X} \\ &= \mathbf{X} \mathbf{X}^+ \mathbf{X} \quad [\text{condition (c)}] \\ &= \mathbf{X} \quad [\text{condition (a)}],\end{aligned}$$

we can multiply $(\mathbf{X}' \mathbf{X})(\mathbf{X}' \mathbf{X})^- (\mathbf{X}' \mathbf{X}) = \mathbf{X}' \mathbf{X}$ on the left by $(\mathbf{X}^+)'$ and obtain

$$\mathbf{X} [(\mathbf{X}' \mathbf{X})^- \mathbf{X}'] \mathbf{X} = \mathbf{X}.$$

Thus \mathbf{X}^* , the matrix in the square brackets, is a generalized inverse of \mathbf{X} , as it satisfies condition (a); using similar arguments, we find that it also satisfies (b) and (c). In fact, a generalized inverse of \mathbf{X} satisfies (a), (b), and (c) if and only if it can be expressed in the form $(\mathbf{X}' \mathbf{X})^- \mathbf{X}'$ (Pringle and Rayner [1971: p. 26]). However, any \mathbf{X}^- satisfying just (a) and (c) will do the trick:

$$\begin{aligned}\mathbf{X}' \mathbf{X} (\mathbf{X}^- \mathbf{Y}) &= \mathbf{X}' (\mathbf{X} \mathbf{X}^-) \mathbf{Y} \\ &= \mathbf{X}' (\mathbf{X} \mathbf{X}^-)' \mathbf{Y} \quad [\text{by (c)}] \\ &= \mathbf{X}' (\mathbf{X}^-)' \mathbf{X}' \mathbf{Y} \\ &= \mathbf{X}' \mathbf{Y} \quad [\text{by (a) transposed}],\end{aligned}$$

and $\mathbf{X}^- \mathbf{Y}$ is a solution of the normal equations. In particular, $\mathbf{X}^+ \mathbf{Y}$ is the unique solution which minimizes $\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}}$ (Peters and Wilkinson [1970]).

Finally, we note that $\hat{\boldsymbol{\theta}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}' \mathbf{X})^- \mathbf{X}' \mathbf{Y}$, so that by B.1.8, $\mathbf{P} = \mathbf{X} (\mathbf{X}' \mathbf{X})^- \mathbf{X}'$ is the unique matrix projecting \mathfrak{R}_n onto $\Omega = \mathcal{C}(\mathbf{X})$.

A.11 SOME USEFUL RESULTS

A.11.1. If $\mathbf{A}\mathbf{x} = \mathbf{0}$ for all \mathbf{x} , then $\mathbf{A} = \mathbf{0}$.

Proof. Setting $x_k = \delta_{ik}$ ($k = 1, 2, \dots, n$), we have $\mathbf{A}\mathbf{x} = \mathbf{a}_i = \mathbf{0}$ where \mathbf{a}_i is the i th column of \mathbf{A} .

A.11.2. If \mathbf{A} is symmetric and $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ for all \mathbf{x} , then $\mathbf{A} = \mathbf{0}$.

Proof. Setting $x_k = \delta_{ik}$ ($k = 1, 2, \dots, n$), then $a_{ii} = 0$. If we set $x_k = \delta_{ik} + \delta_{jk}$ ($k = 1, 2, \dots, n$), then $\mathbf{x}'\mathbf{A}\mathbf{x} = 0 \Rightarrow a_{ii} + 2a_{ij} + a_{jj} = 0 \Rightarrow a_{ij} = 0$.

A.11.3. If \mathbf{A} is symmetric and nonsingular, then

$$\beta'\mathbf{A}\beta - 2\mathbf{b}'\beta = (\beta - \mathbf{A}^{-1}\mathbf{b})'\mathbf{A}(\beta - \mathbf{A}^{-1}\mathbf{b}) - \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}.$$

A.11.4. For all \mathbf{a} and \mathbf{b} :

- (a) $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$.
- (b) $\|\mathbf{a} - \mathbf{b}\| \geq \|\mathbf{a}\| - \|\mathbf{b}\|$.

A.12 SINGULAR VALUE DECOMPOSITION

Let \mathbf{X} be an $n \times p$ matrix. Then \mathbf{X} can be expressed in the form

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}',$$

where \mathbf{U} is an $n \times p$ matrix consisting of p orthonormalized eigenvectors associated with the p largest eigenvalues of $\mathbf{X}\mathbf{X}'$, \mathbf{V} is a $p \times p$ orthogonal matrix consisting of the orthonormalized eigenvectors of $\mathbf{X}\mathbf{X}'$, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$ is a $p \times p$ diagonal matrix. Here $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$, called the *singular values* of \mathbf{X} , are the square roots of the (nonnegative) eigenvalues of $\mathbf{X}'\mathbf{X}$.

Proof. Suppose that $\text{rank}(\mathbf{X}'\mathbf{X}) = \text{rank}(\mathbf{X}) = r$ (A.2.4). Then there exists a $p \times p$ orthogonal matrix \mathbf{T} such that

$$\mathbf{X}'\mathbf{X}\mathbf{T} = \mathbf{T}\Lambda,$$

where $\Lambda = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2, 0, \dots, 0)$, $\sigma_i^2 > 0$. Let

$$\mathbf{s}_i = \sigma_i^{-1}\mathbf{X}\mathbf{t}_i \quad (i = 1, 2, \dots, r);$$

then $\mathbf{X}'\mathbf{s}_i = \sigma_i^{-1}\mathbf{X}'\mathbf{X}\mathbf{t}_i = \sigma_i\mathbf{t}_i$ and $\mathbf{X}\mathbf{s}_i = \sigma_i\mathbf{X}\mathbf{t}_i = \sigma_i^2\mathbf{s}_i$. Thus the \mathbf{s}_i ($i = 1, 2, \dots, r$) are eigenvectors of $\mathbf{X}\mathbf{X}'$ corresponding to the eigenvalues σ_i^2 ($i = 1, 2, \dots, r$). Now $\mathbf{s}_i'\mathbf{s}_i = 1$, and since the eigenvectors corresponding to different eigenvectors of a symmetric matrix are orthogonal,

the s_i are orthonormal. By A.2.3 and A.2.4 there exists an orthonormal set $\{s_{r+1}, s_{r+2}, \dots, s_n\}$ spanning $\mathcal{N}(\mathbf{X}\mathbf{X}') (= \mathcal{N}(\mathbf{X}'))$. But $\mathcal{N}(\mathbf{X}') \perp \mathcal{C}(\mathbf{X})$ and $s_i \in \mathcal{C}(\mathbf{X})$ ($i = 1, 2, \dots, r$) so that $\mathbf{S} = (s_1, s_2, \dots, s_n)$ is an $n \times n$ orthogonal matrix. Hence

$$\begin{aligned} (\mathbf{S}'\mathbf{X}\mathbf{T})_{ij} &= s'_i(\mathbf{X}\mathbf{t}_j) = \sigma_i s'_i s_j \quad (i = 1, 2, \dots, r), \\ &= 0 \quad (i = r + 1, \dots, n) \end{aligned}$$

and $\mathbf{S}'\mathbf{X}\mathbf{T} = \begin{pmatrix} \Sigma \\ \mathbf{0} \end{pmatrix}$. Finally,

$$\mathbf{X} = \mathbf{S} \begin{pmatrix} \Sigma \\ \mathbf{0} \end{pmatrix} \mathbf{T}' = \mathbf{U}\Sigma\mathbf{V}',$$

where \mathbf{U} is the first p columns of \mathbf{S} and $\mathbf{V} = \mathbf{T}$.

When \mathbf{X} has full rank (i.e., $r = p$), then the singular values of \mathbf{X} are $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p > 0$.

A.13 SOME MISCELLANEOUS STATISTICAL RESULTS

A.13.1. For any random variable X , $\gamma_2 \geq -2$, where $\gamma_2 = (\mu_4/\mu_2^2) - 3$.

Proof. Let $\mu = E[X]$; then

$$\begin{aligned} 0 &\leq \text{var}[(X - \mu)^2] \\ &= E[(x - \mu)^4] - \{E[(X - \mu)^2]\}^2 \\ &= \mu_4 - \mu_2^2 \\ &= \mu_2^2 \left(\frac{\mu_4}{\mu_2^2} - 3 + 2 \right) \\ &= \mu_2^2(\gamma_2 + 2) \end{aligned}$$

and $\gamma_2 + 2 \geq 0$.

A.13.2. If $X \sim N(0, \sigma^2)$, then $\text{var}[X^2] = 2\sigma^2$.

The result follows by setting $\gamma_2 = 0$ in the proof of A.13.1.

A.13.3. Let X be a nonnegative nondegenerate random variable (i.e., not identically equal to a constant). If the expectations exist, then

$$E[X^{-1}] > (E[X])^{-1}.$$

Proof. Let $f(x) = x^{-1}$ and let $\mu = E[X] (> 0$, since X is not identically zero). Taking a Taylor expansion, we have

$$f(X) = f(\mu) + (X - \mu)f'(\mu) + \frac{1}{2}(X - \mu)^2 f''(X_0),$$

where X_0 lies between X and μ . Now $f''(X_0) = 2X_0^{-3} > 0$, so that $E[(X - \mu)^2 f''(X_0)] > 0$. Hence

$$E[X^{-1}] = E[f(X)] > f(\mu) = (E[X])^{-1}.$$

A.13.4. If X is a random variable, then from A.13.3 we have

$$f(X) \approx f(\mu) + (X - \mu)f'(\mu)$$

and

$$\text{var}[f(X)] \approx E[f(X) - f(\mu)]^2 \approx \text{var}f[X](f'(\mu))^2.$$

A.13.5. (*Multivariate t-distribution*) An $m \times 1$ vector of random variables $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)'$ is said to have a multivariate *t*-distribution if its probability density function is given by

$$f(\mathbf{y}) = \frac{\Gamma(\frac{1}{2}[\nu + m])}{(\pi\nu)^{m/2}\Gamma(\frac{1}{2}\nu)\det(\Sigma)^{1/2}}[1 + \nu^{-1}(\mathbf{y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})]^{-(\nu+m)/2},$$

where Σ is an $m \times m$ positive-definite matrix. We shall write $\mathbf{Y} \sim t_m(\nu, \boldsymbol{\mu}, \Sigma)$. This distribution has the following properties:

- (a) If $\Sigma = (\sigma_{ij})$, then $(Y_r - \mu_r)/\sqrt{\sigma_{rr}} \sim t_\nu$.
- (b) $(\mathbf{Y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim F_{m,\nu}$.

A.13.6. The random variable X is said to have a beta(a, b) distribution if its density function is given by

$$f(x) = \frac{1}{B(a, b)}x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1,$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

A.14 FISHER SCORING

Consider a model with log likelihood $l(\boldsymbol{\gamma})$. Then Fisher's method of scoring for finding $\hat{\boldsymbol{\gamma}}$, the maximum-likelihood estimate of $\boldsymbol{\gamma}$, is given by the iterative process

$$\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)} - \left\{ E \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right] \right\}_{\boldsymbol{\gamma}^{(m)}}^{-1} \left(\frac{\partial l}{\partial \boldsymbol{\gamma}} \right)_{\boldsymbol{\gamma}^{(m)}}.$$

This algorithm can be regarded as a Newton method for maximizing $l(\boldsymbol{\gamma})$, but with the Hessian replaced by its expected value, the (expected) information matrix. The latter matrix is usually simpler and more likely to be positive definite because of the relationship

$$-E \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} \right] = E \left[\frac{\partial l}{\partial \boldsymbol{\gamma}} \frac{\partial l}{\partial \boldsymbol{\gamma}'} \right],$$

which holds under fairly general conditions.

Appendix B

Orthogonal Projections

B.1 ORTHOGONAL DECOMPOSITION OF VECTORS

B.1.1. Given Ω , a vector subspace of \Re_n (n -dimensional Euclidean space), every $n \times 1$ vector \mathbf{y} can be expressed uniquely in the form $\mathbf{y} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in \Omega$ and $\mathbf{v} \in \Omega^\perp$.

Proof. Suppose that there are two such decompositions $\mathbf{y} = \mathbf{u}_i + \mathbf{v}_i$ ($i = 1, 2$); then $(\mathbf{u}_1 - \mathbf{u}_2) + (\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}$. Because $(\mathbf{u}_1 - \mathbf{u}_2) \in \Omega$ and $(\mathbf{v}_1 - \mathbf{v}_2) \in \Omega^\perp$, we must have $\mathbf{u}_1 = \mathbf{u}_2$ and $\mathbf{v}_1 = \mathbf{v}_2$.

B.1.2. If $\mathbf{u} = \mathbf{P}_\Omega \mathbf{y}$, then \mathbf{P}_Ω is unique.

Proof. Given two such matrices \mathbf{P}_i ($i = 1, 2$), then since \mathbf{u} is unique for every \mathbf{y} , $(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{y} = \mathbf{0}$ for all \mathbf{y} ; hence $(\mathbf{P}_1 - \mathbf{P}_2) = \mathbf{0}$ (A.11.1).

B.1.3. The matrix \mathbf{P}_Ω can be expressed in the form $\mathbf{P}_\Omega = \mathbf{T}\mathbf{T}'$, where the columns of \mathbf{T} form an orthogonal basis for Ω .

Proof. Let $\mathbf{T} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_r)$, where r is the dimension of Ω . Expand the set of $\boldsymbol{\alpha}_i$ to give an orthonormal basis for \Re_n , namely, $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r, \boldsymbol{\alpha}_{r+1}, \dots, \boldsymbol{\alpha}_n$. Then

$$\mathbf{y} = \sum_{i=1}^n c_i \boldsymbol{\alpha}_i = \sum_{i=1}^r c_i \boldsymbol{\alpha}_i + \sum_{i=r+1}^n c_i \boldsymbol{\alpha}_i = \mathbf{u} + \mathbf{v},$$

where $\mathbf{u} \in \Omega$ and $\mathbf{v} \in \Omega^\perp$. But $\alpha'_i \alpha_j = \delta_{ij}$, so that $\alpha'_i \mathbf{y} = \mathbf{c}_i$. Hence

$$\mathbf{u} = (\alpha_1, \dots, \alpha_r) \begin{pmatrix} \alpha'_1 \mathbf{y} \\ \vdots \\ \alpha'_r \mathbf{y} \end{pmatrix} = \mathbf{T} \mathbf{T}' \mathbf{y}.$$

By B.1.2, $\mathbf{P}_\Omega = \mathbf{T} \mathbf{T}'$.

B.1.4. \mathbf{P}_Ω is symmetric and idempotent.

Proof. $\mathbf{P}_\Omega = \mathbf{T} \mathbf{T}'$, which is obviously symmetric, and

$$\mathbf{P}_\Omega^2 = \mathbf{T} \mathbf{T}' \mathbf{T} \mathbf{T}' = \mathbf{T} \mathbf{I}_r \mathbf{T}' = \mathbf{T} \mathbf{T}' = \mathbf{P}_\Omega.$$

B.1.5. $\mathcal{C}(\mathbf{P}_\Omega) = \Omega$.

Proof. Clearly, $\mathcal{C}(\mathbf{P}_\Omega) \subset \Omega$ since \mathbf{P}_Ω projects onto Ω . Conversely, if $\mathbf{x} \in \Omega$, then $\mathbf{x} = \mathbf{P}_\Omega \mathbf{x} \in \mathcal{C}(\mathbf{P})$. Thus the two spaces are the same.

B.1.6. $\mathbf{I}_n - \mathbf{P}_\Omega$ represents an orthogonal projection on Ω^\perp .

Proof. From the identity $\mathbf{y} = \mathbf{P}_\Omega \mathbf{y} + (\mathbf{I}_n - \mathbf{P}_\Omega) \mathbf{y}$ we have that $\mathbf{v} = (\mathbf{I}_n - \mathbf{P}_\Omega) \mathbf{y}$. The results above then apply by interchanging the roles of Ω and Ω^\perp .

B.1.7. If \mathbf{P} is a symmetric idempotent $n \times n$ matrix, then \mathbf{P} represents an orthogonal projection onto $\mathcal{C}(\mathbf{P})$.

Proof. Let $\mathbf{y} = \mathbf{P} \mathbf{y} + (\mathbf{I}_n - \mathbf{P}) \mathbf{y}$. Then $(\mathbf{P} \mathbf{y})' (\mathbf{I}_n - \mathbf{P}) \mathbf{y} = \mathbf{y}' (\mathbf{P} - \mathbf{P}^2) \mathbf{y} = 0$, so that this decomposition gives orthogonal components of \mathbf{y} . The result then follows from B.1.5.

B.1.8. If $\Omega = \mathcal{C}(\mathbf{X})$, then $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-} \mathbf{X}'$, where $(\mathbf{X}' \mathbf{X})^{-}$ is any generalized inverse of $\mathbf{X}' \mathbf{X}$ (i.e., if $\mathbf{B} = \mathbf{X}' \mathbf{X}$, then $\mathbf{B} \mathbf{B}^{-} \mathbf{B} = \mathbf{B}$).

Proof. Let $\mathbf{c} = \mathbf{X}' \mathbf{Y} = \mathbf{B} \beta$. Then $\mathbf{B}(\mathbf{B}^{-} \mathbf{c}) = \mathbf{B} \mathbf{B}^{-} \mathbf{B} \beta = \mathbf{B} \beta$ and $\hat{\beta} = \mathbf{B}^{-} \mathbf{c}$ is a solution of $\mathbf{B} \beta = \mathbf{c}$, that is, of $\mathbf{X}' \mathbf{X} \beta = \mathbf{X}' \mathbf{Y}$. Hence writing $\hat{\theta} = \mathbf{X} \hat{\beta}$, we have $\mathbf{Y} = \hat{\theta} + (\mathbf{Y} - \hat{\theta})$, where

$$\begin{aligned}\hat{\theta}'(\mathbf{Y} - \hat{\theta}) &= \hat{\beta}' \mathbf{X}'(\mathbf{Y} - \mathbf{X}' \mathbf{X} \hat{\beta}) \\ &= \hat{\beta}'(\mathbf{X}' \mathbf{Y} - \mathbf{X}' \mathbf{X} \hat{\beta}) \\ &= 0.\end{aligned}$$

Thus we have an orthogonal decomposition of \mathbf{Y} such that $\hat{\theta} \in \mathcal{C}(\mathbf{X})$ and $(\mathbf{Y} - \hat{\theta}) \perp \mathcal{C}(\mathbf{X})$. Since $\hat{\theta} = \mathbf{X} \hat{\beta} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-} \mathbf{X}' \mathbf{Y}$, we have that $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-} \mathbf{X}'$ (by B.1.2).

B.1.9. When the columns of \mathbf{X} are linearly independent in B.1.8, then $\mathbf{P}_\Omega = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$.

Proof. Although B.1.9 follows from B.1.8, the result can be proved directly since $\mathbf{X} = \mathbf{T}\mathbf{C}$ for nonsingular \mathbf{C} (by B.1.3) and

$$\mathbf{P}_\Omega = \mathbf{X}\mathbf{C}^{-1}(\mathbf{C}^{-1})'\mathbf{X}' = \mathbf{X}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

B.2 ORTHOGONAL COMPLEMENTS

- B.2.1. If $\mathcal{N}(\mathbf{C})$ is the null space (kernel) of the matrix \mathbf{C} , then $\mathcal{N}(\mathbf{C}) = \{\mathcal{C}(\mathbf{C}')\}^\perp$.

Proof. If $\mathbf{x} \in \mathcal{N}(\mathbf{C})$, then $\mathbf{Cx} = \mathbf{0}$ and \mathbf{x} is orthogonal to each row of \mathbf{C} . Hence $\mathbf{x} \perp \mathcal{C}(\mathbf{C}')$. Conversely, if $\mathbf{x} \perp \mathcal{C}(\mathbf{C}')$, then $\mathbf{Cx} = \mathbf{0}$ and $\mathbf{x} \in \mathcal{N}(\mathbf{C})$.

- B.2.2. $(\Omega_1 \cap \Omega_2)^\perp = \Omega_1^\perp + \Omega_2^\perp$.

Proof. Let \mathbf{C}_i be such that $\Omega_i = \mathcal{N}(\mathbf{C}_i)$ ($i = 1, 2$). Then

$$\begin{aligned} (\Omega_1 \cap \Omega_2)^\perp &= \left\{ \mathcal{N} \left(\begin{array}{c} \mathbf{C}_1 \\ \mathbf{C}_2 \end{array} \right) \right\}^\perp \\ &= \mathcal{C}(\mathbf{C}'_1, \mathbf{C}'_2) \quad (\text{by B.2.1}) \\ &= \mathcal{C}(\mathbf{C}'_1) + \mathcal{C}(\mathbf{C}'_2) \\ &= \Omega_1^\perp + \Omega_2^\perp. \end{aligned}$$

B.3 PROJECTIONS ON SUBSPACES

- B.3.1. Given $\omega \subset \Omega$, then $\mathbf{P}_\Omega \mathbf{P}_\omega = \mathbf{P}_\omega \mathbf{P}_\Omega = \mathbf{P}_\omega$.

Proof. Since $\omega \subset \Omega$ and $\omega = \mathcal{C}(\mathbf{P}_\omega)$ (by B.1.5), we have $\mathbf{P}_\Omega \mathbf{P}_\omega = \mathbf{P}_\omega$. The result then follows by the symmetry of \mathbf{P}_ω and \mathbf{P}_Ω .

- B.3.2. $\mathbf{P}_\Omega - \mathbf{P}_\omega = \mathbf{P}_{\omega^\perp \cap \Omega}$.

Proof. Consider $\mathbf{P}_\Omega \mathbf{y} = \mathbf{P}_\omega \mathbf{y} + (\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y}$. Now $\mathbf{P}_\Omega \mathbf{y}$ and $\mathbf{P}_\omega \mathbf{y}$ belong to Ω , so that $(\mathbf{P}_\Omega - \mathbf{P}_\omega)\mathbf{y} \in \Omega$. Hence the preceding equation represents an orthogonal decomposition of Ω into ω and $\omega^\perp \cap \Omega$ since $\mathbf{P}_\omega(\mathbf{P}_\Omega - \mathbf{P}_\omega) = \mathbf{0}$ (by B.3.1).

- B.3.3. If \mathbf{A}_1 is any matrix such that $\omega = \mathcal{N}(\mathbf{A}_1) \cap \Omega$, then $\omega^\perp \cap \Omega = \mathcal{C}(\mathbf{P}_\Omega \mathbf{A}'_1)$.

Proof.

$$\begin{aligned} \omega^\perp \cap \Omega &= \{\Omega \cap \mathcal{N}(\mathbf{A}_1)\}^\perp \cap \Omega \\ &= \{\Omega^\perp + \mathcal{C}(\mathbf{A}'_1)\} \cap \Omega \quad (\text{by B.2.1 and B.2.2}). \end{aligned}$$

If \mathbf{x} belongs to the right-hand side, then

$$\mathbf{x} = \mathbf{P}_\Omega \mathbf{x} = \mathbf{P}_\Omega \{(\mathbf{I}_n - \mathbf{P}_\Omega)\alpha + \mathbf{A}'_1 \beta\} = \mathbf{P}_\Omega \mathbf{A}'_1 \beta \in \mathcal{C}(\mathbf{P}_\Omega \mathbf{A}'_1).$$

Conversely, if $\mathbf{x} \in \mathcal{C}(\mathbf{P}_\Omega \mathbf{A}'_1)$, then $\mathbf{x} \in \mathcal{C}(\mathbf{P}_\Omega) = \Omega$. Also, if $\mathbf{z} \in \omega$, then $\mathbf{x}' \mathbf{z} = \beta' \mathbf{A}'_1 \mathbf{P}_\Omega \mathbf{z} = \beta' \mathbf{A}'_1 \mathbf{z} = 0$, that is, $\mathbf{x} \in \omega^\perp$. Thus $\mathbf{x} \in \omega^\perp \cap \Omega$.

- B.3.4. If \mathbf{A}_1 is a $q \times n$ matrix of rank q , then $\text{rank}(\mathbf{P}_\Omega \mathbf{A}'_1) = q$ if and only if $\mathcal{C}(\mathbf{A}'_1) \cap \Omega^\perp = \mathbf{0}$.

Proof. $\text{rank}(\mathbf{P}_\Omega \mathbf{A}'_1) \leq \text{rank } \mathbf{A}_1$ (by A.2.1). Let the rows of \mathbf{A}_1 be \mathbf{a}'_i ($i = 1, 2, \dots, q$) and suppose that $\text{rank}(\mathbf{P}_\Omega \mathbf{A}'_1) < q$. Then the columns of $\mathbf{P}_\Omega \mathbf{A}'_1$ are linearly dependent, so that $\sum_{i=1}^q c_i \mathbf{P}_\Omega \mathbf{a}'_i = \mathbf{0}$; that is, there exists a vector $\sum_i c_i \mathbf{a}'_i \in \mathcal{C}(\mathbf{A}'_1)$ that is perpendicular to Ω . Hence $\mathcal{C}(\mathbf{A}'_1) \cap \Omega^\perp \neq \mathbf{0}$, which is a contradiction. [By selecting the linearly independent rows of \mathbf{A}_1 we find that the result above is true if \mathbf{A}_1 is $k \times n$ ($k \geq q$).]

Linear Regression Analysis, Second Edition

by George A. F. Seber and Alan J. Lee

Copyright © 2003 John Wiley & Sons, Inc.

Appendix C

Tables

C.1 PERCENTAGE POINTS OF THE BONFERRONI t -STATISTIC

Tabulation of $t_{\nu}^{\alpha/(2k)}$ for different values of α , k , and ν , where

$$\text{pr}[T \geq t_{\nu}^{\alpha/(2k)}] = \frac{\alpha}{2k}$$

and T has the t -distribution with ν degrees of freedom (see Section 5.1.1).

		$\alpha = 0.05$											
$k \setminus \nu:$	5	7	10	12	15	20	24	30	40	60	120	∞	
2	3.17	2.84	2.64	2.56	2.49	2.42	2.39	2.36	2.33	2.30	2.27	2.24	
3	3.54	3.13	2.87	2.78	2.69	2.61	2.58	2.54	2.50	2.47	2.43	2.39	
4	3.81	3.34	3.04	2.94	2.84	2.75	2.70	2.66	2.62	2.58	2.54	2.50	
5	4.04	3.50	3.17	3.06	2.95	2.85	2.80	2.75	2.71	2.66	2.62	2.58	
6	4.22	3.64	3.28	3.15	3.04	2.93	2.88	2.83	2.78	2.73	2.68	2.64	
7	4.38	3.76	3.37	3.24	3.11	3.00	2.94	2.89	2.84	2.79	2.74	2.69	
8	4.53	3.86	3.45	3.31	3.18	3.06	3.00	2.94	2.89	2.84	2.79	2.74	
9	4.66	3.95	3.52	3.37	3.24	3.11	3.05	2.99	2.93	2.88	2.83	2.77	
10	4.78	4.03	3.58	3.43	3.29	3.16	3.09	3.03	2.97	2.92	2.86	2.81	
15	5.25	4.36	3.83	3.65	3.48	3.33	3.26	3.19	3.12	3.06	2.99	2.94	
20	5.60	4.59	4.01	3.80	3.62	3.46	3.38	3.30	3.23	3.16	3.09	3.02	
25	5.89	4.78	4.15	3.93	3.74	3.55	3.47	3.39	3.31	3.24	3.16	3.09	
30	6.15	4.95	4.27	4.04	3.82	3.63	3.54	3.46	3.38	3.30	3.22	3.15	
35	6.36	5.09	4.37	4.13	3.90	3.70	3.61	3.52	3.43	3.34	3.27	3.19	
40	6.56	5.21	4.45	4.20	3.97	3.76	3.66	3.57	3.48	3.39	3.31	3.23	
45	6.70	5.31	4.53	4.26	4.02	3.80	3.70	3.61	3.51	3.42	3.34	3.26	
50	6.86	5.40	4.59	4.32	4.07	3.85	3.74	3.65	3.55	3.46	3.37	3.29	
100	8.00	6.08	5.06	4.73	4.42	4.15	4.04	3.90	3.79	3.69	3.58	3.48	
250	9.68	7.06	5.70	5.27	4.90	4.56	4.4*	4.2*	4.1*	3.97	3.83	3.72	

$\alpha = 0.01$

$k \setminus \nu:$	5	7	10	12	15	20	24	30	40	60	120	∞
2	4.78	4.03	3.58	3.43	3.29	3.16	3.09	3.03	2.97	2.92	2.86	2.81
3	5.25	4.36	3.83	3.65	3.48	3.33	3.26	3.19	3.12	3.06	2.99	2.94
4	5.60	4.59	4.01	3.80	3.62	3.46	3.38	3.30	3.23	3.16	3.09	3.02
5	5.89	4.78	4.15	3.93	3.74	3.55	3.47	3.39	3.31	3.24	3.16	3.09
6	6.15	4.95	4.27	4.04	3.82	3.63	3.54	3.46	3.38	3.30	3.22	3.15
7	6.36	5.09	4.37	4.13	3.90	3.70	3.61	3.52	3.43	3.34	3.27	3.19
8	6.56	5.21	4.45	4.20	3.97	3.76	3.66	3.57	3.48	3.39	3.31	3.23
9	6.70	5.31	4.53	4.26	4.02	3.80	3.70	3.61	3.51	3.42	3.34	3.26
10	6.86	5.40	4.59	4.32	4.07	3.85	3.74	3.65	3.55	3.46	3.37	3.29
15	7.51	5.79	4.86	4.56	4.29	4.03	3.91	3.80	3.70	3.59	3.50	3.40
20	8.00	6.08	5.06	4.73	4.42	4.15	4.04	3.90	3.79	3.69	3.58	3.48
25	8.37	6.30	5.20	4.86	4.53	4.25	4.1*	3.98	3.88	3.76	3.64	3.54
30	8.68	6.49	5.33	4.95	4.61	4.33	4.2*	4.13	3.93	3.81	3.69	3.59
35	8.95	6.67	5.44	5.04	4.71	4.39	4.3*	4.26	3.97	3.84	3.73	3.63
40	9.19	6.83	5.52	5.12	4.78	4.46	4.3*	4.1*	4.01	3.89	3.77	3.66
45	9.41	6.93	5.60	5.20	4.84	4.52	4.3*	4.2*	4.1*	3.93	3.80	3.69
50	9.68	7.06	5.70	5.27	4.90	4.56	4.4*	4.2*	4.1*	3.97	3.83	3.72
100	11.04	7.80	6.20	5.70	5.20	4.80	4.7*	4.4*	4.5*		4.00	3.89
250	13.26	8.83	6.9*	6.3*	5.8*	5.2*	5.0*	4.9*	4.8*			4.11

SOURCE: Dunn [1961: Tables 1 and 2]. Reprinted with permission from the Journal of the American Statistical Association. Copyright (1961) by the American Statistical Association. All rights reserved.

*Obtained by graphical interpolation.

C.2 DISTRIBUTION OF THE LARGEST ABSOLUTE VALUE OF k STUDENT t VARIABLES

Tabulation of $u_{k,\nu,\rho}^\alpha$ for different values of ρ , α , k , and ν , where

$$\text{pr}[U \geq u_{k,\nu,\rho}^\alpha] = \alpha$$

and U is the maximum absolute value of k Student t -variables, each based on ν degrees of freedom and having a common pairwise correlation ρ (see Section 5.1.1).

		$\rho = 0.0$									
$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
		$\alpha = 0.10$									
3	2.353	2.989	3.369	3.637	3.844	4.011	4.272	4.471	4.631	4.823	5.066
4	2.132	2.662	2.976	3.197	3.368	3.506	3.722	3.887	4.020	4.180	4.383
5	2.015	2.491	2.769	2.965	3.116	3.239	3.430	3.576	3.694	3.837	4.018
6	1.943	2.385	2.642	2.822	2.961	3.074	3.249	3.384	3.493	3.624	3.790
7	1.895	2.314	2.556	2.726	2.856	2.962	3.127	3.253	3.355	3.478	3.635
8	1.860	2.262	2.494	2.656	2.780	2.881	3.038	3.158	3.255	3.373	3.522
9	1.833	2.224	2.447	2.603	2.723	2.819	2.970	3.086	3.179	3.292	3.436
10	1.813	2.193	2.410	2.562	2.678	2.741	2.918	3.029	3.120	3.229	3.368
11	1.796	2.169	2.381	2.529	2.642	2.733	2.875	2.984	3.072	3.178	3.313
12	1.782	2.149	2.357	2.501	2.612	2.701	2.840	2.946	3.032	3.136	3.268
15	1.753	2.107	2.305	2.443	2.548	2.633	2.765	2.865	2.947	3.045	3.170
20	1.725	2.065	2.255	2.386	2.486	2.567	2.691	2.786	2.863	2.956	3.073
25	1.708	2.041	2.226	2.353	2.450	2.528	2.648	2.740	2.814	2.903	3.016
30	1.697	2.025	2.207	2.331	2.426	2.502	2.620	2.709	2.781	2.868	2.978
40	1.684	2.006	2.183	2.305	2.397	2.470	2.585	2.671	2.741	2.825	2.931
60	1.671	1.986	2.160	2.278	2.368	2.439	2.550	2.634	2.701	2.782	2.884

$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\rho = 0.0$											
$\alpha = 0.05$											
3	3.183	3.960	4.430	4.764	5.023	5.233	5.562	5.812	6.015	6.259	6.567
4	2.777	3.382	3.745	4.003	4.203	4.366	4.621	4.817	4.975	5.166	5.409
5	2.571	3.091	3.399	3.619	3.789	3.928	4.145	4.312	4.447	4.611	4.819
6	2.447	2.916	3.193	3.389	3.541	3.664	3.858	4.008	4.129	4.275	4.462
7	2.365	2.800	3.056	3.236	3.376	3.489	3.668	3.805	3.916	4.051	4.223
8	2.306	2.718	2.958	3.128	3.258	3.365	3.532	3.660	3.764	3.891	4.052
9	2.262	2.657	2.885	3.046	3.171	3.272	3.430	3.552	3.651	3.770	3.923
10	2.228	2.609	2.829	2.984	3.103	3.199	3.351	3.468	3.562	3.677	3.823
11	2.201	2.571	2.784	2.933	3.048	3.142	3.288	3.400	3.491	3.602	3.743
12	2.179	2.540	2.747	2.892	3.004	3.095	3.236	3.345	3.433	3.541	3.677
15	2.132	2.474	2.669	2.805	2.910	2.994	3.126	3.227	3.309	3.409	3.536
20	2.086	2.411	2.594	2.722	2.819	2.898	3.020	3.114	3.190	3.282	3.399
25	2.060	2.374	2.551	2.673	2.766	2.842	2.959	3.048	3.121	3.208	3.320
30	2.042	2.350	2.522	2.641	2.732	2.805	2.918	3.005	3.075	3.160	3.267
40	2.021	2.321	2.488	2.603	2.690	2.760	2.869	2.952	3.019	3.100	3.203
60	2.000	2.292	2.454	2.564	2.649	2.716	2.821	2.900	2.964	3.041	3.139
$\alpha = 0.01$											
3	5.841	7.127	7.914	8.479	8.919	9.277	9.838	10.269	10.616	11.034	11.559
4	4.604	5.462	5.985	6.362	6.656	6.897	7.274	7.565	7.801	8.087	8.451
5	4.032	4.700	5.106	5.398	5.625	5.812	6.106	6.333	6.519	6.744	7.050
6	3.707	4.271	4.611	4.855	5.046	5.202	5.449	5.640	5.796	5.985	6.250
7	3.500	3.998	4.296	4.510	4.677	4.814	5.031	5.198	5.335	5.502	5.716
8	3.355	3.809	4.080	4.273	4.424	4.547	4.742	4.894	5.017	5.168	5.361
9	3.250	3.672	3.922	4.100	4.239	4.353	4.532	4.672	4.785	4.924	5.103
10	3.169	3.567	3.801	3.969	4.098	4.205	4.373	4.503	4.609	4.739	4.905
11	3.106	3.485	3.707	3.865	3.988	4.087	4.247	4.370	4.470	4.593	4.750
12	3.055	3.418	3.631	3.782	3.899	3.995	4.146	4.263	4.359	4.475	4.625
15	2.947	3.279	3.472	3.608	3.714	3.800	3.935	4.040	4.125	4.229	4.363
20	2.845	3.149	3.323	3.446	3.541	3.617	3.738	3.831	3.907	3.999	4.117
25	2.788	3.075	3.239	3.354	3.442	3.514	3.626	3.713	3.783	3.869	3.978
30	2.750	3.027	3.185	3.295	3.379	3.448	3.555	3.637	3.704	3.785	3.889
40	2.705	2.969	3.119	3.223	3.303	3.367	3.468	3.545	3.607	3.683	3.780
60	2.660	2.913	3.055	3.154	3.229	3.290	3.384	3.456	3.515	3.586	3.676

$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\rho = 0.2$											
$\alpha = 0.10$											
3	2.353	2.978	3.347	3.607	3.806	3.967	4.216	4.405	4.557	4.739	4.969
4	2.132	2.653	2.958	3.172	3.337	3.470	3.676	3.833	3.960	4.112	4.303
5	2.015	2.482	2.753	2.943	3.089	3.207	3.390	3.530	3.642	3.778	3.948
6	1.943	2.377	2.627	2.802	2.937	3.045	3.213	3.342	3.446	3.570	3.728
7	1.895	2.306	2.542	2.707	2.833	2.935	3.093	3.214	3.312	3.429	3.577
8	1.860	2.255	2.481	2.638	2.759	2.856	3.007	3.122	3.214	3.326	3.468
9	1.833	2.217	2.435	2.586	2.702	2.796	2.941	3.052	3.141	3.248	3.384
10	1.813	2.187	2.399	2.546	2.658	2.749	2.889	2.997	3.083	3.187	3.319
11	1.796	2.163	2.370	2.513	2.623	2.711	2.848	2.952	3.036	3.138	3.266
12	1.782	2.143	2.346	2.487	2.594	2.680	2.814	2.916	2.998	3.097	3.222
15	1.753	2.101	2.295	2.429	2.531	2.613	2.741	2.837	2.915	3.009	3.128
20	1.725	2.060	2.245	2.373	2.470	2.548	2.669	2.761	2.835	2.923	3.036
25	1.708	2.036	2.217	2.341	2.435	2.510	2.627	2.716	2.787	2.873	2.981
30	1.697	2.020	2.198	2.319	2.412	2.485	2.600	2.686	2.756	2.839	2.945
40	1.684	2.000	2.174	2.293	2.383	2.455	2.566	2.649	2.717	2.798	2.900
60	1.671	1.981	2.151	2.267	2.354	2.424	2.532	2.613	2.679	2.757	2.856
$\alpha = 0.05$											
3	3.183	3.946	4.403	4.727	4.976	5.178	5.492	5.731	5.923	6.154	6.445
4	2.777	3.371	3.725	3.975	4.168	4.325	4.569	4.755	4.906	5.087	5.316
5	2.571	3.082	3.383	3.596	3.760	3.893	4.102	4.261	4.390	4.545	4.742
6	2.447	2.908	3.178	3.369	3.516	3.635	3.821	3.964	4.079	4.218	4.395
7	2.365	2.793	3.042	3.218	3.353	3.463	3.634	3.766	3.872	4.000	4.163
8	2.306	2.711	2.946	3.111	3.238	3.340	3.501	3.624	3.724	3.844	3.997
9	2.262	2.650	2.874	3.031	3.151	3.249	3.402	3.518	3.613	3.727	3.873
10	2.228	2.603	2.818	2.969	3.084	3.178	3.324	3.436	3.527	3.637	3.776
11	2.201	2.565	2.774	2.919	3.031	3.122	3.263	3.371	3.458	3.564	3.698
12	2.179	2.535	2.738	2.879	2.988	3.075	3.212	3.317	3.402	3.504	3.635
15	2.132	2.469	2.660	2.793	2.895	2.977	3.105	3.203	3.282	3.377	3.499
20	2.086	2.406	2.586	2.711	2.806	2.883	3.002	3.093	3.166	3.255	3.367
25	2.060	2.370	2.543	2.663	2.754	2.828	2.942	3.029	3.099	3.183	3.291
30	2.042	2.346	2.515	2.632	2.721	2.792	2.903	2.987	3.055	3.137	3.241
40	2.021	2.317	2.481	2.594	2.679	2.748	2.855	2.936	3.001	3.097	3.179
60	2.000	2.288	2.447	2.556	2.639	2.705	2.808	2.886	2.948	3.023	3.119

		$\rho = 0.2$									
$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\alpha = 0.01$											
3	5.841	7.104	7.871	8.418	8.841	9.184	9.721	10.132	10.462	10.860	11.360
4	4.604	5.447	5.958	6.323	6.607	6.838	7.200	7.477	7.702	7.973	8.316
5	4.032	4.690	5.085	5.369	5.589	5.769	6.051	6.268	6.444	6.658	6.930
6	3.707	4.263	4.595	4.832	5.017	5.168	5.405	5.588	5.736	5.917	6.147
7	3.500	3.991	4.283	4.491	4.653	4.786	4.994	5.155	5.286	5.445	5.648
8	3.355	3.803	4.068	4.257	4.403	4.523	4.711	4.857	4.975	5.119	5.303
9	3.250	3.666	3.911	4.086	4.221	4.331	4.505	4.639	4.748	4.881	5.051
10	3.169	3.562	3.792	3.956	4.082	4.186	4.348	4.474	4.576	4.700	4.859
11	3.106	3.480	3.699	3.854	3.974	4.071	4.225	4.344	4.440	4.558	4.708
12	3.055	3.414	3.623	3.771	3.886	3.979	4.126	4.239	4.331	4.443	4.587
15	2.947	3.276	3.466	3.599	3.703	3.787	3.919	4.020	4.103	4.204	4.332
20	2.845	3.146	3.318	3.439	3.532	3.607	3.725	3.816	3.890	3.980	4.094
25	2.788	3.072	3.235	3.348	3.435	3.506	3.616	3.701	3.769	3.853	3.959
30	2.750	3.025	3.181	3.289	3.373	3.440	3.545	3.626	3.692	3.771	3.872
40	2.705	2.967	3.115	3.218	3.297	3.361	3.460	3.536	3.598	3.672	3.767
60	2.660	2.911	3.052	3.150	3.224	3.285	3.378	3.449	3.507	3.577	3.666
		$\rho = 0.4$									
$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\alpha = 0.10$											
3	2.353	2.941	3.282	3.519	3.700	3.845	4.069	4.237	4.373	4.534	4.737
4	2.132	2.623	2.905	3.101	3.250	3.370	3.556	3.696	3.809	3.943	4.113
5	2.015	2.455	2.706	2.880	3.013	3.120	3.284	3.410	3.510	3.630	3.781
6	1.943	2.352	2.584	2.745	2.867	2.965	3.117	3.233	3.325	3.436	3.575
7	1.895	2.283	2.502	2.653	2.768	2.861	3.004	3.112	3.199	3.304	3.435
8	1.860	2.233	2.442	2.587	2.697	2.786	2.922	3.026	3.109	3.208	3.334
9	1.833	2.195	2.398	2.538	2.644	2.729	2.860	2.960	3.040	3.136	3.257
10	1.813	2.166	2.363	2.499	2.602	2.684	2.812	2.909	2.986	3.079	3.196
11	1.796	2.142	2.335	2.468	2.568	2.649	2.773	2.867	2.943	3.034	3.148
12	1.782	2.123	2.312	2.442	2.541	2.620	2.742	2.834	2.908	2.996	3.108
15	1.753	2.081	2.263	2.387	2.481	2.556	2.673	2.760	2.831	2.916	3.022
20	1.725	2.041	2.216	2.334	2.424	2.496	2.606	2.690	2.757	2.837	2.938
25	1.708	2.018	2.188	2.303	2.390	2.460	2.567	2.649	2.713	2.791	2.888
30	1.697	2.003	2.169	2.283	2.368	2.437	2.542	2.621	2.684	2.760	2.856
40	1.684	1.984	2.146	2.257	2.341	2.408	2.510	2.587	2.650	2.723	2.816
60	1.671	1.965	2.124	2.233	2.315	2.379	2.479	2.554	2.615	2.686	2.776

$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\rho = 0.4$											
$\alpha = 0.05$											
3	3.183	3.902	4.324	4.620	4.846	5.028	5.309	5.522	5.693	5.898	6.155
4	2.777	3.337	3.665	3.894	4.069	4.210	4.430	4.596	4.730	4.891	5.093
5	2.571	3.053	3.333	3.528	3.677	3.798	3.986	4.128	4.243	4.381	4.555
6	2.447	2.883	3.134	3.309	3.443	3.552	3.719	3.847	3.950	4.074	4.230
7	2.365	2.770	3.002	3.164	3.288	3.388	3.543	3.661	3.756	3.870	4.014
8	2.306	2.690	2.909	3.061	3.177	3.271	3.417	3.528	3.617	3.725	3.860
9	2.262	2.630	2.839	2.984	3.095	3.184	3.323	3.429	3.513	3.616	3.745
10	2.228	2.584	2.785	2.925	3.032	3.117	3.250	3.352	3.433	3.531	3.655
11	2.201	2.547	2.742	2.877	2.980	3.063	3.192	3.290	3.369	3.464	3.583
12	2.179	2.517	2.707	2.838	2.939	3.020	3.145	3.240	3.317	3.409	3.525
15	2.132	2.452	2.632	2.756	2.850	2.927	3.042	3.133	3.205	3.291	3.400
20	2.086	2.391	2.560	2.677	2.766	2.837	2.947	3.031	3.098	3.178	3.280
25	2.060	2.355	2.520	2.631	2.718	2.786	2.891	2.971	3.036	3.113	3.211
30	2.042	2.332	2.492	2.602	2.685	2.751	2.854	2.933	2.995	3.070	3.165
40	2.021	2.304	2.459	2.565	2.646	2.711	2.810	2.885	2.945	3.018	3.110
60	2.000	2.275	2.426	2.530	2.608	2.670	2.766	2.838	2.897	2.966	3.054
$\alpha = 0.01$											
3	5.841	7.033	7.740	8.240	8.623	8.932	9.414	9.780	10.074	10.428	10.874
4	4.604	5.401	5.874	6.209	6.467	6.675	7.000	7.249	7.448	7.688	7.991
5	4.032	4.655	5.024	5.284	5.485	5.648	5.902	6.096	6.253	6.442	6.682
6	3.707	4.235	4.545	4.764	4.934	5.071	5.285	5.449	5.582	5.742	5.946
7	3.500	3.967	4.241	4.435	4.583	4.704	4.893	5.038	5.155	5.297	5.477
8	3.355	3.783	4.031	4.207	4.343	4.452	4.624	4.755	4.861	4.990	5.154
9	3.250	3.648	3.879	4.041	4.167	4.268	4.427	4.549	4.647	4.766	4.918
10	3.169	3.545	3.763	3.916	4.034	4.129	4.277	4.392	4.484	4.596	4.739
11	3.106	3.464	3.671	3.817	3.929	4.019	4.160	4.269	4.357	4.463	4.598
12	3.055	3.400	3.598	3.737	3.844	3.931	4.066	4.170	4.254	4.356	4.484
15	2.947	3.263	3.444	3.571	3.668	3.746	3.869	3.962	4.039	4.131	4.247
20	2.845	3.135	3.301	3.415	3.504	3.574	3.685	3.769	3.837	3.921	4.026
25	2.788	3.063	3.219	3.327	3.410	3.477	3.581	3.660	3.725	3.802	3.900
30	2.750	3.016	3.166	3.270	3.349	3.415	3.514	3.590	3.650	3.726	3.820
40	2.705	2.959	3.103	3.202	3.277	3.337	3.432	3.505	3.562	3.632	3.722
60	2.660	2.904	3.040	3.134	3.207	3.264	3.353	3.421	3.477	3.542	3.628

$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\rho = 0.5$											
$\alpha = 0.10$											
3	2.353	2.912	3.232	3.453	3.621	3.755	3.962	4.117	4.242	4.390	4.576
4	2.132	2.598	2.863	3.046	3.185	3.296	3.468	3.597	3.701	3.825	3.980
5	2.015	2.434	2.669	2.832	2.956	3.055	3.207	3.323	3.415	3.525	3.664
6	1.943	2.332	2.551	2.701	2.815	2.906	3.047	3.153	3.238	3.340	3.469
7	1.895	2.264	2.471	2.612	2.720	2.806	2.938	3.038	3.119	3.216	3.336
8	1.860	2.215	2.413	2.548	2.651	2.733	2.860	2.956	3.032	3.124	3.239
9	1.833	2.178	2.369	2.500	2.599	2.679	2.801	2.893	2.967	3.055	3.167
10	1.813	2.149	2.335	2.463	2.559	2.636	2.755	2.844	2.916	3.002	3.110
11	1.796	2.126	2.308	2.433	2.527	2.602	2.718	2.805	2.875	2.959	3.064
12	1.782	2.107	2.286	2.408	2.500	2.574	2.687	2.773	2.841	2.923	3.026
15	1.753	2.066	2.238	2.355	2.443	2.514	2.622	2.704	2.769	2.847	2.945
20	1.725	2.027	2.192	2.304	2.388	2.455	2.559	2.637	2.699	2.773	2.867
25	1.708	2.004	2.165	2.274	2.356	2.421	2.522	2.597	2.658	2.730	2.820
30	1.697	1.989	2.147	2.254	2.335	2.399	2.498	2.572	2.631	2.701	2.790
40	1.687	1.970	2.125	2.230	2.309	2.372	2.468	2.540	2.598	2.667	2.753
60	1.671	1.952	2.104	2.207	2.284	2.345	2.439	2.509	2.565	2.632	2.716
$\alpha = 0.05$											
3	3.183	3.867	4.263	4.538	4.748	4.916	5.176	5.372	5.529	5.718	5.953
4	2.777	3.310	3.618	3.832	3.995	4.126	4.328	4.482	4.605	4.752	4.938
5	2.57	3.03	3.29	3.48	3.62	3.73	3.90	4.03	4.14	4.26	4.42
6	2.45	2.86	3.10	3.26	3.39	3.49	3.64	3.76	3.86	3.97	4.11
7	2.36	2.75	2.97	3.12	3.24	3.33	3.47	3.58	3.67	3.78	3.91
8	2.31	2.67	2.88	3.02	3.13	3.22	3.35	3.46	3.54	3.64	3.76
9	2.26	2.61	2.81	2.95	3.05	3.14	3.26	3.36	3.44	3.53	3.65
10	2.23	2.57	2.76	2.89	2.99	3.07	3.19	3.29	3.36	3.45	3.57
11	2.20	2.53	2.72	2.84	2.94	3.02	3.14	3.23	3.30	3.39	3.50
12	2.18	2.50	2.68	2.81	2.90	2.98	3.09	3.18	3.25	3.34	3.45
15	2.13	2.44	2.61	2.73	2.82	2.89	3.00	3.08	3.15	3.23	3.33
20	2.09	2.38	2.54	2.65	2.73	2.80	2.90	2.98	3.05	3.12	3.22
25	2.060	2.344	2.500	2.607	2.688	2.752	2.852	2.927	2.987	3.059	3.150
30	2.04	2.32	2.47	2.58	2.66	2.72	2.82	2.89	2.95	3.02	3.11
40	2.02	2.29	2.44	2.54	2.62	2.68	2.77	2.85	2.90	2.97	3.06
60	2.00	2.27	2.41	2.51	2.58	2.64	2.73	2.80	2.86	2.92	3.00

$\nu \setminus k:$	1	2	3	4	5	6	8	10	12	15	20
$\rho = 0.5$											
$\alpha = 0.01$											
3	5.841	6.974	7.639	8.104	8.459	8.746	9.189	9.527	9.797	10.123	10.532
4	4.604	5.364	5.809	6.121	6.361	6.554	6.855	7.083	7.267	7.488	7.766
5	4.03	4.63	4.98	5.22	5.41	5.56	5.80	5.98	6.12	6.30	6.52
6	3.71	4.21	4.51	4.71	4.87	5.00	5.20	5.35	5.47	5.62	5.81
7	3.50	3.95	4.21	4.39	4.53	4.64	4.82	4.95	5.06	5.19	5.36
8	3.36	3.77	4.00	4.17	4.29	4.40	4.56	4.68	4.78	4.90	5.05
9	3.25	3.63	3.85	4.01	4.12	4.22	4.37	4.48	4.57	4.68	4.82
10	3.17	3.53	3.74	3.88	3.99	4.08	4.22	4.33	4.42	4.52	4.65
11	3.11	3.45	3.65	3.79	3.89	3.98	4.11	4.21	4.29	4.39	4.52
12	3.05	3.39	3.58	3.71	3.81	3.89	4.02	4.12	4.19	4.29	4.41
15	2.95	3.25	3.43	3.55	3.64	3.71	3.83	3.92	3.99	4.07	4.18
20	2.85	3.13	3.29	3.40	3.48	3.55	3.65	3.73	3.80	3.87	3.97
25	2.788	3.055	3.205	3.309	3.388	3.452	3.551	3.626	3.687	3.759	3.852
30	2.75	3.01	3.15	3.25	3.33	3.39	3.49	3.56	3.62	3.69	3.78
40	2.70	2.95	3.09	3.19	3.26	3.32	3.41	3.48	3.53	3.60	3.68
60	2.66	2.90	3.03	3.12	3.19	3.25	3.33	3.40	3.45	3.51	3.59

SOURCE: Hahn and Hendrickson, "A table of the percentage points of the distribution of k Student t variables and its applications," *Biometrika*, 1971, 58, 323-333, Tables 1, 2, and 3, by permission of the Biometrika trustees.

C.3 WORKING-HOTELLING CONFIDENCE BANDS FOR FINITE INTERVALS

The Working-Hotelling confidence band for a straight line over the interval $a \leq x \leq b$ is the region between the two curves given by Equation (6.7) in Section 6.1.3.

Tabulation of λ for different values of $n - 2$ and c , where c is given by Equation (6.9).

		$\alpha = 0.01$									
	$c \setminus n - 2:$	5	10	15	20	30	40	60	120	∞	
One point	0.0	4.03	3.16	2.95	2.84	2.75	2.70	2.66	2.62	2.58	
	0.05	4.10	3.22	2.99	2.88	2.79	2.74	2.69	2.65	2.61	
	0.1	4.18	3.27	3.03	2.93	2.83	2.78	2.73	2.69	2.64	
	0.15	4.26	3.32	3.07	2.96	2.86	2.81	2.76	2.72	2.67	
	0.2	4.33	3.36	3.11	3.00	2.89	2.84	2.80	2.75	2.70	
	0.3	4.45	3.44	3.18	3.06	2.95	2.90	2.85	2.80	2.75	
	0.4	4.56	3.50	3.24	3.11	3.00	2.95	2.89	2.84	2.79	
	0.6	4.73	3.61	3.32	3.20	3.07	3.02	2.96	2.91	2.86	
	0.8	4.85	3.68	3.39	3.25	3.13	3.07	3.01	2.95	2.90	
	1.0	4.94	3.74	3.43	3.30	3.17	3.11	3.05	2.99	2.94	
	1.5	5.05	3.81	3.50	3.36	3.22	3.16	3.10	3.04	2.98	
	2.0	5.10	3.85	3.53	3.38	3.25	3.19	3.15	3.06	3.01	
		∞	5.15	3.89	3.57	3.42	3.28	3.22	3.15	3.10	3.04

$\alpha = 0.05$

$c \setminus n - 2:$	5	10	15	20	30	40	60	120	∞	
One point	0.0	2.57	2.23	2.13	2.08	2.04	2.02	2.00	1.98	1.96
	0.05	2.62	2.27	2.17	2.12	2.08	2.06	2.03	2.02	1.99
	0.1	2.68	2.31	2.21	2.16	2.12	2.10	2.07	2.05	2.03
	0.15	2.74	2.36	2.25	2.20	2.15	2.13	2.11	2.08	2.06
	0.2	2.79	2.40	2.29	2.23	2.18	2.16	2.14	2.11	2.09
	0.3	2.88	2.47	2.35	2.30	2.42	2.22	2.19	2.17	2.15
	0.4	2.97	2.53	2.41	2.35	2.29	2.27	2.24	2.21	2.19
	0.6	3.10	2.62	2.49	2.43	2.37	2.34	2.31	2.29	2.26
	0.8	3.19	2.69	2.55	2.49	2.43	2.38	2.37	2.34	2.31
	1.0	3.25	2.74	2.60	2.53	2.47	2.44	2.41	2.38	2.35
	1.5	3.33	2.81	2.67	2.59	2.52	2.49	2.46	2.43	2.40
	2.0	3.36	2.83	2.68	2.61	2.55	2.51	2.48	2.45	2.42
	∞	3.40	2.86	2.71	2.64	2.58	2.54	2.51	2.48	2.45

 $\alpha = 0.10$

$c \setminus n - 2:$	5	10	15	20	30	40	60	120	∞	
One point	0.0	2.01	1.81	1.75	1.72	1.68	1.68	1.67	1.66	1.65
	0.05	2.06	1.85	1.79	1.76	1.73	1.72	1.70	1.69	1.68
	0.1	2.11	1.89	1.83	1.80	1.77	1.85	1.74	1.73	1.71
	0.15	2.16	1.93	1.87	1.84	1.81	1.79	1.77	1.76	1.75
	0.2	2.21	1.97	1.90	1.87	1.84	1.82	1.81	1.79	1.78
	0.3	2.30	2.04	1.97	1.93	1.90	1.88	1.87	1.85	1.84
	0.4	2.37	2.10	2.02	1.99	1.95	1.93	1.92	1.90	1.88
	0.6	2.49	2.19	2.12	2.07	2.03	2.01	1.99	1.98	1.96
	0.8	2.57	2.26	2.17	2.13	2.09	2.07	2.05	2.03	2.01
	1.0	2.62	2.31	2.22	2.17	2.13	2.11	2.09	2.07	2.05
	1.5	2.69	2.37	2.27	2.23	2.18	2.16	2.15	2.12	2.10
	2.0	2.72	2.39	2.29	2.25	2.20	2.18	2.16	2.14	2.12
	∞	2.75	2.42	2.32	2.27	2.23	2.21	2.19	2.17	2.14

SOURCE: Wynn and Bloomfield [1971: Appendix A]. Copyright (1971) by the Royal Statistical Society. All rights reserved.

Outline Solutions to Selected Exercises

EXERCISES 1a

1. Set $\mathbf{X} - \mathbf{a} = \mathbf{X} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{a}$. Then

$$\begin{aligned} E[||\mathbf{X} - \mathbf{a}||^2] &= E[\text{tr}\{(\mathbf{X} - \mathbf{a})'(\mathbf{X} - \mathbf{a})\}] \\ &= E[\text{tr}\{(\mathbf{X} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \mathbf{a})(\mathbf{X} - \boldsymbol{\mu})'\}] \quad (\text{by A.1.2}) \\ &= \text{tr } E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] \\ &= \text{tr Var}[\mathbf{X}] + \text{tr}[||\boldsymbol{\mu} - \mathbf{a}||^2]. \end{aligned}$$

2. Let $\mathbf{U} = \mathbf{X} - \mathbf{a}$, $\mathbf{V} = \mathbf{Y} - \mathbf{b}$, then $\text{Cov}[\mathbf{U}, \mathbf{V}] = E[(\mathbf{U} - E[\mathbf{U}])(\mathbf{V} - E[\mathbf{V}])']$.

3. $X_1 = Y_1$, $X_i = \sum_{j=1}^i Y_j$ and $\text{var}[X_i] = i$. For $r \leq s$,

$$\text{cov}[X_r, X_s] = \text{cov}[X_r, X_r + Y_{r+1} + \cdots + Y_s] = \text{var}[X_r] = r.$$

4. $\text{cov}[X_i, X_j] = \sigma^2 \rho |i - j|$.

EXERCISES 1b

1. Let $Q = \mathbf{X}' \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -2 \\ 0 & -2 & 1 \end{pmatrix} \mathbf{X} = \mathbf{X}' \mathbf{A} \mathbf{X}$, say. $\text{tr}(\mathbf{A} \boldsymbol{\Sigma}) = \sum_{ij} a_{ij} \sigma_{ij} = 1$ and $E[Q] = \boldsymbol{\mu}^2 + 2\boldsymbol{\mu}^2 - 4\boldsymbol{\mu}^2 + \boldsymbol{\mu}^2 + 1 = 1$.

2. $\text{var}[\bar{X}] = \sum_i \sigma_i^2 / n^2$ and $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - n\bar{X}^2$.

$$\begin{aligned} E \left[\sum_i (X_i - \bar{X})^2 \right] &= \sum_i (\sigma_i^2 + \mu^2) - n \text{var}[\bar{X}] + n(E[\bar{X}])^2 \\ &= \sum_i \sigma_i^2 - n \text{var}[\bar{X}] = n(n-1) \text{var}[\bar{X}]. \end{aligned}$$

3. (a) Either substitute $w_n = \sum_{i=1}^{n-1} w_i$ or use a Lagrange multiplier term $\lambda(\sum_i w_i - 1)$ to minimize and get $w_i \sigma_i^2 = a$, or $a = (\sum_i \sigma_i^{-2})^{-1}$. Substituting for w_i gives us $v_{\min} = a$.

(b)

$$\begin{aligned} E[(n-1)S_w^2] &= E \left[\sum_i w_i X_i^2 \right] - E[\bar{X}_w^2] \\ &= \sum_i w_i (\sigma_i^2 + \mu^2) - (\text{var}[\bar{X}_w] + \mu^2) \\ &= \sum_i w_i (1-w_i) \sigma_i^2 = na - \left(\sum_i w_i \right) a = (n-1)a. \end{aligned}$$

4. (a) $\text{var}[\bar{X}] = (\sigma^2/n)\{1 + (n-1)\rho\} \geq 0$, so that $-1/(n-1) \leq \rho \leq 1$.

(b) $E[Q] = \sigma^2 \{an + bn(1 + (n-1)\rho\} + \theta^2(an + bn^2) \equiv \sigma^2 + 0$. Hence $b = -1/\{n(n-1)(1-\rho)\}$, $a = -bn$. Thus $Q = a \sum_i X_i^2 - (a/n)(\sum X_i)^2 = a \sum (X_i - \bar{X})^2$.

5. In both cases we use Theorem 1.6 with $\mu_4 = 3\mu_2^2 = 3\sigma^2$, $\mu_3 = 0$, $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}\mathbf{1}_n = \mathbf{0}$. Thus $\text{var}[(n-1)S^2] = \text{var}[\mathbf{X}'\mathbf{A}\mathbf{X}] = 2\sigma^4 \text{tr}[\mathbf{A}^2]$.

(a) $\text{tr}[\mathbf{A}^2] = \text{tr}[\mathbf{A}] = n-1$.

(b) $2(n-1)Q = 2 \sum_{i=1}^n X_i^2 - X_1^2 - X_n^2 - 2 \sum_{i=1}^{n-1} X_i X_{i+1}$.

(c) $\text{tr}(\mathbf{A}^2) = \sum \sum a_{ij}^2 = 6n-8$. Hence $\text{var}[Q] = 2\sigma^4(6n-8)/4(n-1)^2$; ratio $= (6n-8)/(4n-4) \rightarrow \frac{3}{2}$.

EXERCISES 1c

- $\text{cov}[X+Y, X-Y] = \text{var}[X] - \text{var}[Y] = 0$. Let X and Y be independent binomial($n, \frac{1}{2}$). Then $0 = \text{pr}(X+Y=1, X-Y=0) \neq \text{pr}(X+Y=1) \text{pr}(X-Y=0)$ (i.e., not independent).
- $\text{cov}[X, Y] = p_{11} - p_{1\cdot}p_{\cdot 1} = 0$ implies that $p_{ij} = p_i \cdot p_{\cdot j}$ for all i, j [e.g., $p_{10} = p_{1\cdot} \cdot p_{\cdot 1} = p_{1\cdot}(1-p_{\cdot 1}) = p_{1\cdot}p_{\cdot 0}$].
- $E[X^{2r+1}] = 0$ because $f(x)$ is an even function. Hence

$$\text{cov}[X, X^2] = E[X^3] - E[X]E[X^2] = 0.$$

4. $f(x, y) = \frac{1}{4} = f_1(x)f_2(y)$, etc., but $f(x, y, z) \neq f_1(x)f_2(y)f_3(z)$.

MISCELLANEOUS EXERCISES 1

1. We use $E[(Z - E[Z])^2] = E[Z^2] - (E[Z])^2$ for various Z .

$$\begin{aligned} & E_Y [E\{(X - E[X|Y])^2|Y\}] + E_Y [(E[X|Y] - E[X])^2] \\ &= E_Y [E[X^2|Y] - (E[X|Y])^2] + E_Y [(E[X|Y])^2] - (E[X])^2 \\ &= E[X^2] - (E[X])^2 = \text{var}[X]. \end{aligned}$$

With vectors, $\text{Var}[\mathbf{Z}] = E[(\mathbf{Z} - E[\mathbf{Z}])(\mathbf{Z} - E[\mathbf{Z}])'] = E[\mathbf{Z}\mathbf{Z}'] - E[\mathbf{Z}]E[\mathbf{Z}]'$ and hence $\text{Var}[\mathbf{X}] = E_{\mathbf{Y}} \text{Var}[\mathbf{X}|Y] + \text{Var}_{\mathbf{Y}} E[\mathbf{X}|Y]$.

2. (a) Use $\text{var}[\mathbf{a}'\mathbf{X}] = \mathbf{a}' \text{Var}[\mathbf{X}] \mathbf{a} = 18$.

$$(b) \quad \mathbf{Y} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \mathbf{X} = \mathbf{AX}; \quad \text{Var}[\mathbf{Y}] = \mathbf{A} \text{Var}[\mathbf{X}] \mathbf{A}' = \begin{pmatrix} 12 & 15 \\ 15 & 21 \end{pmatrix}.$$

3. The nonzero covariances are $\sigma_{12}, \sigma_{23}, \dots, \sigma_{n-1,n}$.

$$\text{var}[\bar{X}] = n^{-2} \sum_{i=1}^n (\sigma_{ii} + 2 \sum_{j=1}^{n-1} \sigma_{i,i+1}) = n^{-2}(A + 2B),$$

say. Then $nE[Q_1] = (n-1)A - 2B$ and $E[Q_2] = 2A - 2B$. Solve for A and B and replace $E[Q_i]$ by Q_i , in $n^{-2}(A + 2B)$.

4. $2\frac{2}{5}$, obtained by substituting in Theorem 1.6 with $\boldsymbol{\theta} = \mathbf{0}$, $\text{tr}[\mathbf{A}^2] = \sum \sum a_{ij}^2 = 18$, $\mathbf{a}'\mathbf{a} = 12$, $\mu_4 = \frac{1}{5}$, and $\mu_2 = \frac{1}{3}$.

5. Use the formula

$$\begin{aligned} \text{Cov}[\mathbf{X}'\mathbf{AX}, \mathbf{X}'\mathbf{BX}] &= \\ &\frac{1}{2}(\text{Var}[\mathbf{X}'(\mathbf{A} + \mathbf{B})\mathbf{X}] - \text{Var}[\mathbf{X}'\mathbf{AX}] - \text{Var}[\mathbf{X}'\mathbf{BX}]). \end{aligned}$$

EXERCISES 2a

- 1.

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \text{ and } \det(\boldsymbol{\Sigma}) = 1.$$

Let $Q = 2y_1^2 + \dots + 65 = 2(y_1 - a_1)^2 + (y_2 - a_2)^2 + 2(y_1 - a_1)(y_2 - a_2)$. Equating coefficients, $a_1 = 4$ and $a_2 = 3$.

- (a) $k = 2\pi$.

- (b) $E[\mathbf{Y}] = (4, 3)'$ and $\text{Var}[\mathbf{Y}] = \boldsymbol{\Sigma}$.

2. $f(\mathbf{y}) = g(u(\mathbf{y})) |\det(\partial u_i / \partial y_j)|$. $\mathbf{U} = \mathbf{A}^{-1}\mathbf{Y} - \mathbf{c}$, so that $(\partial u_i / \partial y_j) = (a^{ij})$ with determinant $\det(\mathbf{A}^{-1}) = (\det \mathbf{A})^{-1}$.
3. (a) Using A.4.7, the leading minor determinants are positive if and only if $(1-\rho)^2(1+2\rho) > 0$ (i.e., if $\rho > -\frac{1}{2}$).
- (b) We use A.1.4. Now $\det(\Sigma - \lambda_i \mathbf{I}_2) = 0$ implies that the eigenvalues are $\lambda_1 = 1+\rho$ and $\lambda_2 = 1-\rho$. Solving $(\Sigma - \lambda_i \mathbf{I}_2)\mathbf{x} = \mathbf{0}$ for \mathbf{x} gives us the orthogonal matrix $\mathbf{T} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ of eigenvectors. Then, from A.4.12, $\Sigma^{1/2} = \mathbf{T} \text{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}) \mathbf{T}'$.

EXERCISES 2b

- $\exp[\theta_1 t_1 + \theta_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2)]$.
- Using Example 2.7, we have $\mathbf{Y}_i = (0, \dots, 1, \dots, 0)\mathbf{Y} = \mathbf{a}'_i \mathbf{Y}$, which is $N(\mathbf{a}'_i \boldsymbol{\mu}, \mathbf{a}'_i \boldsymbol{\Sigma} \mathbf{a}_i)$.
- Writing $\mathbf{Z} = \mathbf{AY}$, we find that \mathbf{Z} is multivariate normal with mean $(5, 1)'$ and variance-covariance matrix $\begin{pmatrix} 10 & 0 \\ 0 & 3 \end{pmatrix}$.
- Let $\mathbf{Z} = \mathbf{CY}$, where $\mathbf{C} = \begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \end{pmatrix}$ has independent rows. Hence \mathbf{Z} is bivariate normal with diagonal variance-covariance matrix \mathbf{CC}' .
- Let $\mathbf{Z} = (X, Y)'$ $\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then,

$$\begin{aligned} E[\exp(t_1 \bar{X} + t_2 \bar{Y})] &= E\left[\exp\left(\frac{1}{n}t_1 X + \frac{1}{n}t_2 Y\right)\right]^n \\ &= \exp\left(n \frac{\mathbf{t}'}{n} \boldsymbol{\mu} + n \frac{1}{2} \frac{\mathbf{t}'}{n} \boldsymbol{\Sigma} \frac{\mathbf{t}}{n}\right) = \exp\left(\mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \frac{\boldsymbol{\Sigma}}{n} \mathbf{t}\right), \end{aligned}$$

showing that the distribution is $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$.

- Let $X_1 = Y_1 + Y_2$ and $X_2 = Y_1 - Y_2$, so that $Y_1 = \frac{1}{2}(X_1 + X_2)$ and $Y_2 = \frac{1}{2}(X_1 - X_2)$, or $\mathbf{Y} = \mathbf{AX} \sim N_2(\mathbf{0}, \mathbf{AA}' = \text{diag}(\frac{1}{2}, \frac{1}{2}))$.
- Since the last term $g(x_i)$, say, is an odd function of x_i [i.e. $g(-x_i) = -g(x_i)$] this term vanishes when integrating over x_i .

8.

$$\begin{aligned} E\left[\exp\left(s\bar{Y} + \sum_{i=1}^n t_i(Y_i - \bar{Y})\right)\right] &= \prod_{i=1}^n E\left[\exp\left\{\left(t_i - \bar{t} + \frac{s}{n}\right) Y_i\right\}\right] \\ &= \prod_{i=1}^n \exp\left[\frac{1}{2} \left(t_i - \bar{t} + \frac{s}{n}\right)^2\right] \\ &= \exp\left(\frac{1}{2} \sum (t_i - \bar{t})^2 + \frac{s^2}{2n}\right), \end{aligned}$$

which factorizes.

9. $\mathbf{Y} = \mathbf{T}\mathbf{X}$, where \mathbf{T} is orthogonal, so that $\mathbf{Y} \sim N_3(\mathbf{0}, \mathbf{T}\mathbf{T}' = \mathbf{I}_3)$.

EXERCISES 2c

1. Yes, as the variance-covariance matrix is diagonal.

2.

$$\begin{aligned}\text{Cov}[\bar{\mathbf{Y}}, \mathbf{Y} - \mathbf{1}_n \bar{\mathbf{Y}}] &= \text{Cov}[n^{-1} \mathbf{1}'_n \mathbf{Y}, (\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}'_n) \mathbf{Y}] \\ &= n^{-1} [\mathbf{1}'_n \Sigma - n^{-1} \mathbf{1}'_n \Sigma \mathbf{1}_n \mathbf{1}'_n] = 0,\end{aligned}$$

since $\Sigma \mathbf{1}_n = (1 + (n-1)\rho) \mathbf{1}_n$; hence independence.

3. Find the joint distribution of $X_1 = Y_1 + Y_2 + Y_3$ and $X_2 = Y_1 - Y_2 - Y_3$, and set the off-diagonal elements of $\text{Var}[\mathbf{X}]$ equal to 0; then $\rho = -\frac{1}{2}$.

EXERCISES 2d

1. The m.g.f. of χ_k^2 is $(1 - 2t)^{-k/2}$; $Z_i^2 \sim \chi_1^2$. Hence

$$E \left[\exp \left(\sum_i t d_i Z_i^2 \right) \right] = \prod_i \exp[(t d_i) Z_i^2] = \prod_i (1 - 2t d_i)^{-1/2}.$$

2. (a) $E[\exp(t \mathbf{Y}' \mathbf{A} \mathbf{Y})] = \int (2\pi)^{-m/2} \exp[-\frac{1}{2} \mathbf{y}' (\mathbf{I}_n - 2t \mathbf{A}) \mathbf{y}] d\mathbf{y} = [\det(\mathbf{I}_n - 2t \mathbf{A})]^{-1/2}$, by A.4.9.
(b) Choose orthogonal \mathbf{T} such that $\det(\mathbf{I}_n - 2t \mathbf{A}) = \det(\mathbf{T}' \mathbf{T}) \det(\mathbf{I}_n - 2t \mathbf{A}) = \det(\mathbf{I}_n - 2t \mathbf{T}' \mathbf{A} \mathbf{T})$, where $\mathbf{T}' \mathbf{A} \mathbf{T} = \text{diag}(\mathbf{1}'_r, 0, \dots, 0)$ (since \mathbf{A} is idempotent with r unit eigenvalues and $n-r$ zero eigenvalues).
(c) If $\mathbf{Y} \sim N_n(\mathbf{0}, \Sigma)$, then $\mathbf{Z} = \Sigma^{-1/2} \mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. The m.g.f. of $\mathbf{Y}' \mathbf{A} \mathbf{Y} = \mathbf{Z}' \Sigma^{1/2} \mathbf{A} \Sigma^{1/2} \mathbf{Z}$ is

$$\begin{aligned}&[\det(\mathbf{I}_n - 2t \Sigma^{1/2} \mathbf{A} \Sigma^{1/2})]^{-1/2} \\ &= [\det(\Sigma^{1/2} \Sigma^{-1/2} - 2t \Sigma^{1/2} \mathbf{A} \Sigma \Sigma^{-1/2})]^{-1/2} \\ &= [\det(\mathbf{I}_n - 2t \mathbf{A} \Sigma)]^{-1/2}.\end{aligned}$$

3. If $Q = \mathbf{Y} \mathbf{A} \mathbf{Y}$, then $\mathbf{A}^2 = \mathbf{A}$ implies that $a = b = \frac{1}{2}$.
4. Writing $Q = \mathbf{Y}' \mathbf{A} \mathbf{Y}$, we find that $\mathbf{A}^2 = \mathbf{A}$ and $\text{tr}[\mathbf{A}] = 2$. Using a similar method, the result is true only for $n = 3$.
5. Using Exercise 2(a), $E[\exp(s \mathbf{Y}' \mathbf{A} \mathbf{Y} + t \mathbf{Y}' \mathbf{A} \mathbf{Y})] = E[\exp\{\mathbf{Y}'(s \mathbf{A} + t \mathbf{B}) \mathbf{Y}\}] = \det(\mathbf{I}_n - 2s \mathbf{A} - 2t \mathbf{B})^{-1/2} = \det(\mathbf{I}_n - 2s \mathbf{A} - 2t \mathbf{B} + st \mathbf{A} \mathbf{B})^{-1/2} = \det(\mathbf{I}_n - s \mathbf{A})^{-1/2} \det(\mathbf{I}_n - t \mathbf{B})^{-1/2}$ using A.4.9; hence independence.

MISCELLANEOUS EXERCISES 2

1.

$$\mathbf{Y} = \begin{pmatrix} \rho & 1 & 0 & 0 \\ \rho^2 & \rho & 1 & 0 \\ \rho^2 & \rho^2 & \rho & 1 \end{pmatrix} \begin{pmatrix} Y_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \mathbf{AZ}.$$

(a) $\text{Var}[\mathbf{Y}] = \mathbf{A} \text{diag}(\sigma_0^2, \sigma^2, \sigma^2, \sigma^2) \mathbf{A}'$.

(b) \mathbf{Y} is multivariate normal as \mathbf{Z} is multivariate normal.

2. Let $\mathbf{Z} = \mathbf{U} + \mathbf{V} = (\mathbf{B} + \mathbf{C})\mathbf{Y}$. Then $\text{Cov}[\mathbf{X}, \mathbf{Z}] = \text{Cov}[\mathbf{X}, \mathbf{U} + \mathbf{V}] = \text{Cov}[\mathbf{X}, \mathbf{U}] + \text{Cov}[\mathbf{X}, \mathbf{V}] = \mathbf{0}$. Hence \mathbf{X} and \mathbf{Z} are independent.

3. Let $U = \bar{Y} = n^{-1}\mathbf{1}_n'$ and $\mathbf{V} = (Y_1 - Y_2, Y_2 - Y_3, \dots, Y_{n-1} - Y_n)'$ = \mathbf{AY} ; then $\mathbf{A}\mathbf{1}_n = \mathbf{0}$ and $\text{Cov}[U, \mathbf{V}] = \sigma^2 n^{-1} \mathbf{1}_n' \mathbf{A}' = \mathbf{0}'$.

4.

$$\begin{aligned} M_{a\mathbf{X}+b\mathbf{Y}}(t) &= E[\exp[t'(a\mathbf{X} + b\mathbf{Y})]] \\ &= E[\exp(at'\mathbf{X})]E[\exp(bt'\mathbf{Y})] \\ &= \exp[t'(a\mu_X + b\mu_Y) + \frac{1}{2}t'(a^2\Sigma_X + b^2\Sigma_Y)t]. \end{aligned}$$

5. Choose orthogonal \mathbf{T} with first row $\mathbf{a}'/\|\mathbf{a}\|$ and let $\mathbf{Z} = \mathbf{TY}$. Then $Z_1 = 0$ and $\mathbf{Y}'\mathbf{Y} = \mathbf{Z}'\mathbf{Z} = Z_2^2 + \dots + Z_n^2$. Since the Z_i are i.i.d. $N(0, 1)$, the conditional distribution of the Z_i ($i \neq 1$) is the same as the unconditional distribution.

6. Let $X_i = Y_i - \theta$; then $\mathbf{X} \sim N_n(\mathbf{0}, \Sigma)$ and $Q = \mathbf{X}'\mathbf{AX}/(1 - \rho)$. Now, by Exercises 2d, No. 2, $M(t) = [\det(\mathbf{I}_n - 2t\mathbf{A}\Sigma/(1 - \rho))]^{-1/2} = |\mathbf{I}_n - 2t\mathbf{A}|^{-1/2}$, by Example 1.9, which does not depend on ρ . Hence Q has the same distribution as for $\rho = 0$, which is χ_{n-1}^2 .

7.

$$\begin{aligned} E \left[\sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})' \right] &= \sum_i E \{ [\mathbf{Y}_i - \theta - (\bar{\mathbf{Y}} - \theta)][\mathbf{Y}_i - \theta - (\bar{\mathbf{Y}} - \theta)]' \} \\ &= \sum_i \{ \text{Var}[\mathbf{Y}_i] - 2\text{Cov}[\bar{\mathbf{Y}}, \mathbf{Y}_i] + \text{Var}[\bar{\mathbf{Y}}] \} \\ &= \sum_i \left(\Sigma - 2\frac{1}{n}\Sigma + \frac{1}{n}\Sigma \right) = (n-1)\Sigma. \end{aligned}$$

8. Let $\mathbf{U} = \mathbf{AY}$, $\mathbf{V} = \mathbf{BY}$, and $\mathbf{W} = (\mathbf{I}_n - \mathbf{A} - \mathbf{B})\mathbf{Y}$. Then $\text{Cov}[\mathbf{U}, \mathbf{V}] = \mathbf{AB}' = \mathbf{0}$, $\text{Cov}[\mathbf{W}, \mathbf{U}] = (\mathbf{I}_n - \mathbf{A} - \mathbf{B})\mathbf{A}' = \mathbf{0}$, etc., so they are

independent by Theorem 2.5. Also, $\mathbf{I}_n - \mathbf{A} - \mathbf{B}$ is idempotent and $\mathbf{U}'\mathbf{U} = \mathbf{Y}'\mathbf{A}'\mathbf{A}\mathbf{Y} = \mathbf{Y}'\mathbf{A}\mathbf{Y}$, etc.

9. (a) First show that \mathbf{W} has the correct mean and variance-covariance matrix. Then use Theorem 2.3 to show that \mathbf{W} is normal, namely,

$$\mathbf{a}'\mathbf{W} = \mathbf{a}'_1\mathbf{X} + \mathbf{a}'_2\mathbf{Y} = \mathbf{b}'_1(X_1, Y_1) + \cdots + \mathbf{b}'_n(X_n, Y_n) = \sum_i U_i,$$

where the U_i are i.i.d. univariate normals.

- (b) Using Example 2.9, we have $E[\mathbf{X}|\mathbf{Y}] = \mu_1 \mathbf{1}_n + \rho(\sigma_1/\sigma_2)(\mathbf{Y} - \mu_2 \mathbf{1}_n)$ and $\text{Var}[\mathbf{X}|\mathbf{Y}] = \sigma_1^2(1 - \rho^2)\mathbf{I}_n$.

10. Let $\sigma_{ii} = \sigma_i^2, \sigma_{12} = \sigma_1\sigma_2\rho$. Expanding the expression gives

$$\frac{1}{1 - \rho^2} \left(\frac{\rho Y_1}{\sigma_1} - \frac{Y_2}{\sigma_2} \right)^2 = \frac{1}{1 - \rho^2} Y_3^2,$$

where $Y_3 \sim N(0, 1 - \rho^2)$.

11. $\mathbf{Y} = \begin{pmatrix} \phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & \phi & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & 0 \\ 0 & 0 & 0 & \cdots & \phi & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \cdot \\ a_n \end{pmatrix} = \mathbf{AX} \sim N_{n+1}(\mathbf{0}, \sigma^2 \mathbf{AA}')$,

where \mathbf{AA}' is tridiagonal with diagonal elements $1 + \phi^2$ and the other elements ϕ .

12. Let $Q = 2(Y_1Y_2 - Y_2Y_3 - Y_3Y_1) = \mathbf{Y}'\mathbf{AY}$. Then $\det(\lambda\mathbf{I}_n - \mathbf{A}) = 0$ gives eigenvalues of 2, -1, and -1. Using (2.10), $\sum_i \lambda_i Z_i^2 = 2Z_1^2 - Z_2^2 - Z_3^2$.
13. (a) $E[\mathbf{Z}] = \mathbf{0}$ and $\text{Var}[\mathbf{Z}] = \mathbf{I}_3$. Now $\mathbf{a}'\mathbf{Z} = a_1 Z_1 + \cdots + a_n Z_n$, where the $a_i Z_i$ are i.i.d. $N_1(0, a_i^2)$, so that $X = \mathbf{a}'\mathbf{Z} \sim N_1(0, \|\mathbf{a}\|^2)$ for all \mathbf{a} .
- (b) Now from Theorem 2.3, $\mathbf{t}'\mathbf{Y} \sim N_1(\mathbf{t}'\boldsymbol{\mu}, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$. From the m.g.f. of the univariate normal, $E[\exp(sX)] = \exp[(\mathbf{t}'\boldsymbol{\mu})s + \frac{1}{2}(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})s^2]$. Put $s = 1$ to get the m.g.f. of \mathbf{Y} .
- (c) Using Theorem 2.2, we can find the m.g.f. of $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ and thus show that the Z_i are i.i.d. $N(0, 1)$. Then from (a), the density of \mathbf{Z} is $f(\mathbf{z}) = \prod_i \phi(z_i)$, where ϕ is the $N(0, 1)$ density function. We can now obtain the density of \mathbf{Y} by the change-of-variables technique (as demonstrated by Theorem 2.1, but in reverse).
14. The last term is an odd function of each y_i so that it vanishes when y_i is integrated out.
15. (a) Q can be negative.

(b) By Exercises 2d, No. 2(a), the m.g.f. of $Q = \mathbf{Y}'\mathbf{A}\mathbf{Y} = [\det(\mathbf{I}_n - 2t\mathbf{A})]^{-1/2} = (1 - t^2)^{-1}$.

16. From Theorem 1.6, $\text{var}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = 2\text{tr}(\mathbf{A}^2) = 2\sum\sum a_{ij}^2 = 12n - 16$.

17. Set $\mathbf{X} = \Sigma^{-1/2}\mathbf{Y}$ and use Theorem 1.6.

EXERCISES 3a

- Set $\mathbf{Y} - \mathbf{X}\beta = (\mathbf{Y} - \mathbf{X}\hat{\beta}) + (\mathbf{X}\hat{\beta} - \mathbf{X}\beta)$ and show that the cross-product of the terms in parentheses is zero.
- $\mathbf{0} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}})$, i.e. $\mathbf{1}'_n(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$. Alternatively, $\partial\|\mathbf{Y} - \mathbf{X}\beta\|^2/\partial\beta_0 = 0 \Rightarrow \sum(Y - \hat{\beta}_0 - \hat{\beta}_1x_{i1} - \dots - \hat{\beta}_{p-1}x_{ip-1}) = 0$.
- $\hat{\theta} = \frac{1}{6}(Y_1 + 2Y_2 + Y_3)$, $\hat{\phi} = \frac{1}{5}(2Y_3 - Y_2)$.
- $\hat{\beta}_0 = \bar{Y}$, $\hat{\beta}_1 = \frac{1}{2}(Y_3 - Y_1)$, $\hat{\beta}_2 = \frac{1}{6}(Y_1 + 2Y_2 + Y_3)$.
- Let $x = \sin\theta$; then $\hat{w} = \sum_i T_i x_i / \sum_i x_i^2$.
- $\mathbf{P}\alpha = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\alpha = \mathbf{X}\beta$, so that $\mathcal{C}(\mathbf{P}) \subset \mathcal{C}(\mathbf{X})$. Conversely, if $\mathbf{y} = \mathbf{X}\gamma$, then $\mathbf{P}\mathbf{y} = \mathbf{y}$ and $\mathcal{C}(\mathbf{X}) \subset \mathcal{C}(\mathbf{P})$.
- $\hat{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}'\mathbf{P}(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}'(\mathbf{P} - \mathbf{P}^2)\mathbf{Y} = \mathbf{0}$.
- Substitute $\mathbf{X} = \mathbf{W}\mathbf{K}$, where \mathbf{K} is a nonsingular diagonal matrix of the form $\text{diag}(1, k_1, \dots, k_{p-1})$.

EXERCISES 3b

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$, $\hat{\beta}_1 = \sum_i Y_i(x_i - \bar{x}) / \sum(x_i - \bar{x})^2$. From $(\mathbf{X}'\mathbf{X})^{-1}$, $\text{cov}[\hat{\beta}_0, \hat{\beta}_1] = -\bar{x} / \sum(x_i - \bar{x})^2$.
- It is helpful to express the model in the form

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_m & \mathbf{0} \\ \mathbf{1}_m & \mathbf{1}_m \\ \mathbf{1}_n & -2\mathbf{1}_n \end{pmatrix} \begin{pmatrix} \theta \\ \phi \end{pmatrix} + \boldsymbol{\varepsilon}.$$

Then

$$\begin{aligned} \hat{\theta} &= \frac{1}{m(m+13n)} \left\{ (m+4n) \sum_i U_i + 6n \sum_j V_j + 3m \sum_k W_k \right\}, \\ \hat{\phi} &= \frac{1}{m(m+13n)} \left\{ (2n-m) \sum_i U_i + (m+3n) \sum_j V_j - 5m \sum_k W_k \right\}. \end{aligned}$$

3. $\mathbf{Y} = \theta \mathbf{1}_n + \boldsymbol{\varepsilon}$. The BLUE of θ is \bar{Y} .

4. Let

$$\mathbf{A} = \begin{pmatrix} \sum(x_{i1} - \bar{x}_1)^2 & \sum(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ \sum(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) & \sum(x_{i2} - \bar{x}_2)^2 \end{pmatrix}.$$

The variance-covariance matrix of $\hat{\beta}_1$ and $\hat{\beta}_2$ is $\sigma^2 \mathbf{A}^{-1}$. Hence $\text{var}[\hat{\beta}_1] = \sigma^2 \sum(x_{i2} - \bar{x}_2)^2 / |\mathbf{A}|$.

EXERCISES 3c

1. (a) From Theorem 3.4, $\text{var}[S^2] = \text{var}[\mathbf{Y}'\mathbf{R}\mathbf{Y}] / (n-p)^2 = 2\sigma^4 / (n-p)$.
 (b) Let $Q = \mathbf{Y}'\mathbf{R}\mathbf{Y}$, and use $E[Q^2] = \text{var}[Q] + (E[Q])^2$. Then $E[(\mathbf{Y}'\mathbf{R}\mathbf{Y}/(n-p+2) - \sigma^2)^2] = 2\sigma^4 / (n-p+2)$. This is the mean-squared error.
 (c) $MSE[S^2] = \text{var}[S^2] = 2\sigma^4(n-p)$.
2. $\sum_i (Y_i - \bar{Y})^2 / (n-1)$, since the diagonal elements of the associated matrix \mathbf{A} are all equal.

EXERCISES 3d

1. $Y_i = \theta + \varepsilon_i$, which is a regression model with $\beta_0 = \theta$, so that $\hat{\theta} = \bar{Y}$ and $\text{RSS} = Q$. (a) and (b) follow from Theorem 3.5.
2. $\mathbf{U} = (\mathbf{Y} - \mathbf{X}\hat{\beta}) = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}$, say. $\mathbf{V} = \mathbf{X}(\hat{\beta} - \beta) = \mathbf{X}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta\} = \mathbf{P}(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{P}\boldsymbol{\varepsilon}$. Then $\text{Cov}[(\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon}, \mathbf{P}\boldsymbol{\varepsilon}] = \sigma^2(\mathbf{I}_n - \mathbf{P})\mathbf{P} = \mathbf{0}$.

EXERCISES 3e

1. We minimize $\sum_{j=1}^k \lambda_j^{-1} + \phi(\sum_{j=1}^k \lambda_j - kc)$, where ϕ is the Lagrange multiplier. Differentiating with respect to λ_j gives us $-\lambda_j^{-2} + \phi = 0$, or $\lambda_j = \text{constant}$.
2. $\phi(x) = x^2 - \frac{2}{3}$.
3. (a) Use A.9.5, second equation, to get the result.
 (b) $\mathbf{x}'_k \mathbf{x}_k - \mathbf{x}'_k \mathbf{P}_W \mathbf{x}_k \leq \mathbf{x}'_k \mathbf{x}_k$ with equality if and only if $\mathbf{P}_W \mathbf{x}_k = \mathbf{0}$ or $\mathbf{W}' \mathbf{x}_k = \mathbf{0}$.
4. Omit the condition $\sum_i x_{ij} = 0$.

5. (a) From Exercise 3, the variance is minimized when the columns of \mathbf{X} are mutually orthogonal. This minimum variance $\sigma^2/\mathbf{x}_k' \mathbf{x}_k$ is least when every element of \mathbf{x}_k is nonzero.
- (b) For the optimum design, $\text{var}[\hat{\beta}_k] = \sigma^2/n$.

EXERCISES 3f

1. $\mathbf{Y}'\mathbf{R}\mathbf{Y} - \mathbf{Y}'\mathbf{R}_G\mathbf{Y} = \hat{\gamma}_G'\mathbf{Z}'\mathbf{R}\mathbf{Y} = \hat{\gamma}_G'\mathbf{Z}'\mathbf{R}\mathbf{Z}\hat{\gamma}_G$.
2. Use A.8 to differentiate $\mathbf{Y}'\mathbf{R}\mathbf{Y} - 2\gamma'\mathbf{Z}'\mathbf{R}\mathbf{Y} + \gamma'\mathbf{Z}'\mathbf{R}\mathbf{Z}\gamma$ with respect to γ , etc.
3. By Theorem 3.6(iv), $\text{var}[\hat{\beta}_{G,i}] - \text{var}[\hat{\beta}_i] = \sigma^2(\mathbf{LML}')_{ii} \geq 0$ since \mathbf{LML}' is positive definite (or zero when $\mathbf{X}'\mathbf{Z} = 0$).
4. $\hat{\theta} = \bar{Y}$ with RSS = $\sum(Y_i - \bar{Y})^2$.
 - (a) Use Exercise 2 and minimize $\sum(Y_i - \gamma x_i - (\bar{Y} - \gamma \bar{x}))^2$ to get $\hat{\gamma}_G = \sum_i Y_i(x_i - \bar{x}) / \sum_i (x_i - \bar{x})^2$. Then use Theorem 3.6(ii) to get $\hat{\theta}_G = \bar{Y} - \hat{\gamma}_G \bar{x}$.
 - (b) Differentiate $\sum(Y_i - \theta - \gamma x_i)^2$ with respect to θ and γ .

EXERCISES 3g

1. (a) $\hat{\alpha} = Y_1 - \bar{Y} + \frac{1}{3}\pi$, $\hat{\beta} = Y_2 - \bar{Y} + \frac{1}{3}\pi$ and $\hat{\gamma} = Y_3 - \bar{Y} + \frac{1}{3}\pi$.
 (b) $\alpha_H = \frac{1}{4}(Y_1 + Y_2 - Y_3 + \pi)$.
2. Use $\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{X}(\hat{\beta} - \hat{\beta}_H) = \mathbf{0}$.
3. The second expression in $\text{Var}[\hat{\beta}_H]$ is positive-semidefinite, so that the diagonal elements of the underlying matrix are nonnegative.
4. $\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_H = \mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_H = \frac{1}{2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\lambda}_H$. $\|\mathbf{Y} - \hat{\mathbf{Y}}_H\|^2 = \frac{1}{4}\hat{\lambda}_H'\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\hat{\lambda}_H$, etc.
5. $\mathbf{X}\mathbf{B}\alpha = \mathbf{0} \Rightarrow \mathbf{B}\alpha = \mathbf{0} \Rightarrow \alpha = \mathbf{0}$; that is, columns of \mathbf{XB} are linearly independent.

EXERCISES 3h

1. $\mathbf{X}'\mathbf{X}\hat{\beta}_1 = \mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\hat{\beta}_2$. Then $\|\mathbf{Y} - \mathbf{X}\hat{\beta}_1\|^2 = \mathbf{Y}'\mathbf{Y} - \hat{\beta}_1'\mathbf{X}'\mathbf{X}\hat{\beta}_2$.
2. $\mathbf{C}\mathbf{X}\beta = \beta$ and $\mathbf{C}\mathbf{X} = \mathbf{I}$, which implies that \mathbf{X} has full rank.

EXERCISES 3i

1. $\mathbf{a}'E[\hat{\beta}] = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \mathbf{c}'\beta$, where $\mathbf{c} = \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} \in \mathcal{C}(\mathbf{X}')$.
 2. If $\mathbf{a}_i \in \mathcal{C}(\mathbf{X}')$, then $\sum_i c_i \mathbf{a}_i \in \mathcal{C}(\mathbf{X}')$.
 3. First show that if $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$, then $\mathbf{b} + \mathbf{c}$ is a solution for all $\mathbf{c} \perp \mathcal{C}(\mathbf{X}')$ and a unique $\mathbf{b} \in \mathcal{C}(\mathbf{X}')$. Thus $\mathbf{a}'\mathbf{c}$ is invariant for all such \mathbf{c} , including $\mathbf{c} = \mathbf{0}$. Hence $\mathbf{a}'\mathbf{c} = 0$ and $\mathbf{a} \in \mathcal{C}(\mathbf{X}')$.
 4. If $\mathbf{a}' = \alpha'\mathbf{X}'\mathbf{X}$, the result follows. Conversely, given the result,
- $$E[\mathbf{a}'\hat{\beta}] = E[\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \mathbf{a}'\beta$$
- and $\mathbf{a}'\beta$ is estimable.
5. Use $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and Exercise 4.
 6. $\mathbf{a} \in \mathcal{C}(\mathbf{X}')$ for all $\mathbf{a} \in \mathfrak{N}_p$, so that $\mathcal{C}(\mathbf{X}') = \mathfrak{N}_p$.

EXERCISES 3j

1. Given $(\mathbf{I}_n - \mathbf{P})\mathbf{Z}\alpha = \mathbf{0}$, then $\mathbf{Z}\alpha \in \mathcal{C}(\mathbf{P}) = \mathcal{C}(\mathbf{X})$. But $\mathcal{C}(\mathbf{X}) \cap \mathcal{C}[\mathbf{Z}] = \mathbf{0}$, so that $\alpha = \mathbf{0}$. Hence $(\mathbf{I}_n - \mathbf{P})\mathbf{Z}$ has full rank. Then
$$\mathbf{Z}'(\mathbf{I}_n - \mathbf{P})'(\mathbf{I}_n - \mathbf{P})\mathbf{Z} = \mathbf{Z}'(\mathbf{I}_n - \mathbf{P})\mathbf{Z}.$$
2. Permuting the columns of \mathbf{X} to make the first r columns \mathbf{X}_1 we get $\mathbf{X}\Pi = (\mathbf{X}_1, \mathbf{X}_1\mathbf{K}) = \mathbf{X}_1(\mathbf{I}_r, \mathbf{K})$, where Π is an orthogonal permutation matrix (see A.5). Then $\mathbf{X} = \mathbf{X}_1\mathbf{L}_1$, where $\mathbf{L}_1 = (\mathbf{I}_r, \mathbf{K})\Pi'$.
3. We use B.3.4. Suppose that $\mathcal{C}(\mathbf{M}) \cap \Omega^\perp \neq \mathbf{0}$. Then there exists $\alpha \neq \mathbf{0}$ such that $\mathbf{M}'\alpha = (\mathbf{I}_n - \mathbf{P})\beta$ (i.e., $\mathbf{A}'\alpha = \mathbf{X}'\mathbf{M}'\alpha = \mathbf{0}$), which implies that $\alpha = \mathbf{0}$.
4. Set $\hat{\theta}_\Omega = \mathbf{X}\hat{\beta}$, $\hat{\theta}_\omega = \mathbf{X}\hat{\beta}_H$, $\mathbf{B} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$, and solve for $\hat{\beta}_H$.
5. Let β_0 be any solution of $\mathbf{A}\beta = \mathbf{c}$. Then $\mathbf{Y} - \mathbf{X}\beta_0 = \mathbf{X}(\beta - \beta_0) + \epsilon$, or $\tilde{\mathbf{Y}} = \mathbf{X}\gamma + \epsilon$ and $\mathbf{A}\gamma = \mathbf{A}\beta - \beta_0 = \mathbf{0}$. Now apply the theory to this transformed setup.

EXERCISES 3k

1. $\frac{1}{3}(2Y_1 - Y_2)$, $\frac{2}{3}\sigma^2$.
2. $\sum_i w_i Y_i / \sum_i w_i$, $\sigma^2 / \sum_i w_i$.

3. $\theta^* = (1/n) \sum_i (Y_i/i)$.
4. $\mathbf{V}^{-1} \mathbf{X} = \mathbf{V}^{-1} \mathbf{1}_n = c \mathbf{1}_n = c \mathbf{X}$, etc.
5. Reduce the model as in Section 3.10.
6. Use Lagrange multipliers to show that $\theta^* = (\mathbf{I}_n - \mathbf{V}\mathbf{A}'(\mathbf{V}\mathbf{A}'')^{-1}\mathbf{V})\mathbf{Y}$.

EXERCISES 3I

1. Since $\tilde{\mathbf{X}}'\mathbf{1}_n = 0$, $\tilde{\mathbf{P}}\mathbf{1}_n = 0$, $\tilde{\mathbf{P}}\tilde{\mathbf{Y}} = \tilde{\mathbf{P}}\mathbf{Y}$, and $\text{RSS} = \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}'\tilde{\mathbf{P}}\tilde{\mathbf{Y}}$.
2. Changing the scales of the explanatory variables does not change the fitted model. Therefore, we simply replace $\tilde{\mathbf{Y}}$ by \mathbf{Y}^* in the RSS for Exercise 1.

EXERCISES 3m

1. First, let $y = a/x^2$ and convert the integral to a Gamma function. Second, if $Q = ||y - \mathbf{X}\beta||^2$, then $\int \sigma^{-(n+1)} \exp(-(1/2\sigma^2)Q) d\sigma = (Q/2)^{-n/2} / \Gamma(n/2) \propto ||y - \mathbf{X}\beta||^{-n}$.
2. Using $Q = (n-p)s^2 + ||\mathbf{X}(\beta - \hat{\beta})||^2$, we have

$$\begin{aligned} f(\beta|y, \sigma) &= f(\beta, y, \sigma)/f(y, \sigma) \\ &\propto f(\beta, \sigma|y) \propto \exp\left(-\frac{1}{2\sigma^2}Q\right) \\ &\propto \exp\left[-\frac{1}{2\sigma^2}(\beta - \hat{\beta})' \mathbf{X}' \mathbf{X}(\beta - \hat{\beta})\right], \end{aligned}$$

which is normal. This has a mean of $\hat{\beta}$.

3. (a) $f(v) \propto v^{-1/2} dv^{1/2}/dv \propto 1/v$.
- (b) $f(\beta, v) \propto f(y|\theta)v^{-1} \propto v^{-n/2-1} \exp[-(1/2v)Q]$.
- (c) Using $Q = ||y - \mathbf{X}\beta||^2 = ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 + ||\mathbf{X}(\beta - \hat{\beta})||^2$,

$$\begin{aligned} f(v|y) &\propto \exp(-a/v) \int v^{-n/2-1} \exp\left[-\frac{1}{2v}||\mathbf{X}(\beta - \hat{\beta})||^2\right] d\beta \\ &\propto \exp(-a/v) v^{-n/2-1} v^{p/2}. \end{aligned}$$

- (d) The posterior mean is

$$\int_0^\infty v v^{-(\nu/2+1)} \exp(-a/v) dv / \int_0^\infty v^{-(\nu/2+1)} \exp(-a/v) dv.$$

Letting $x = a/v$ and integrating gives $\int_0^\infty v^{-(\nu/2+1)} \exp(-a/v) dv = a^{-(\nu/2)} \Gamma(\nu/2)$, so the posterior mean is

$$a^{-(\nu/2-1)} \Gamma[(\nu-2)/2] / a^{-\nu/2} \Gamma(\nu/2) = a/(\nu/2-1) = \text{RSS}/(n-p-2).$$

EXERCISES 3n

- The equation $\sum_i \text{sign}(|e_i/s| - 1/c) = 0$ implies that the number of residuals satisfying $|e_i/s| < 1/c$ is the same as that satisfying $|e_i/s| < 1/c$. Thus $\text{median}_i |e_i/s| = 1/c$, which implies that result.
- Let $e_{(1)} \leq e_{(2)} \leq \dots \leq e_{(n)}$. Then

$$\begin{aligned} \sum_{1 \leq i < j \leq n} |e_i - e_j| &= \sum_{1 \leq i < j \leq n} |e_{(i)} - e_{(j)}| \\ &= \sum_{1 \leq i < j \leq n} e_{(j)} - e_{(i)} \\ &= \sum_{j=2}^n (j-1)e_{(j)} - \sum_{i=1}^{n-1} (n-i)e_{(i)} \\ &= \sum_{i=1}^n (2i-n-1)e_{(i)} \\ &= 2(n+1) \sum_{i=1}^n a(i)e_{(i)} \\ &= 2(n+1) \sum_{i=1}^n a(R_i)e_i, \end{aligned}$$

where $a(i) = i/(n+1) - 0.5$.

- The left-hand side of (3.101) is the number of differences $\geq s$, so by (3.101), the number of differences $\geq s$ is approximately $3/4 \binom{n}{2}$ (i.e., s is approximately the lower quartile of the differences).

MISCELLANEOUS EXERCISES 3

- $\sum_i a_i b_i = 0$.
- Use Lagrange multipliers or show that $\mathbf{I}_n - \mathbf{P}_\Omega = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{A}'$.
- Use Section 3.7.1 with $\mathbf{X} = \mathbf{X}_1$ and $\mathbf{Z} = \mathbf{X}_2$.

$$\text{Var}[\hat{\beta}_2] = \sigma^2 \mathbf{X}'_2 (\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1) \mathbf{X}_2.$$

4. Let $\mathbf{c}'\mathbf{Y} = (\mathbf{a} + \mathbf{b})'\mathbf{Y}$ be any other linear unbiased estimate of $\mathbf{a}'\mathbf{X}\beta$ (i.e., $\mathbf{b}'\mathbf{X} = \mathbf{0}'$). Then $\text{var}[\mathbf{a}'\mathbf{Y}] + \text{var}[\mathbf{b}'\mathbf{Y}] + 2\text{cov}[\mathbf{a}'\mathbf{Y}, \mathbf{b}'\mathbf{Y}] \geq \text{var}[\mathbf{a}'\mathbf{Y}]$ with equality if and only if $\mathbf{b} = \mathbf{0}$.
5. $\text{tr Var}[\hat{\mathbf{Y}}] = \text{tr Var}[\mathbf{P}\mathbf{Y}] = \sigma^2 \text{tr } \mathbf{P} = \sigma^2 \text{rank } \mathbf{P} = \sigma^2 p$.
6. 9.95, 5.0, 4.15, 1.1.
7. $\frac{1}{8}(3Y_{1..} - Y_{1.} - Y_{..1}), \frac{1}{8}(-Y_{1..} + 3Y_{1.} - Y_{..1}), \frac{1}{8}(-Y_{1..} - Y_{1.} + 3Y_{..1})$, where $Y_{1..} = \sum_j \sum_k Y_{1jk}$, etc.
8. (a) $(\sum_i Y_i x_i)/(\sum_i x_i^2)$. (b) $(\sum_i Y_i)/(\sum_i x_i)$. (c) $(1/n) \sum_i (Y_i/x_i)$.
10. Use the identity $\mathbf{Y} - \mathbf{X}\beta = \mathbf{Y} - \mathbf{X}\beta^* + \mathbf{X}(\beta^* - \beta)$. Then

$$(\mathbf{X}\beta^* - \mathbf{X}\beta)' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta^*) = (\beta^* - \beta)' (\mathbf{X}' \mathbf{V}^{-1} \mathbf{Y} - \mathbf{X}' \mathbf{V}^{-1} \mathbf{X}\beta^*) = 0.$$
11. $\hat{\theta}_1 = \frac{1}{3}(Y_1 + Y_3), \hat{\theta}_2 = \frac{1}{3}(Y_1 + Y_2)$.
 $\hat{\theta}_1 = \frac{1}{3}(-Y_1 - 2Y_2 + 3Y_3), \hat{\theta}_2 = \frac{1}{3}(Y_1 - Y_2), \hat{\theta}_3 = Y_1 + Y_2 + Y_3$.
- 13.

$$\begin{aligned} \text{Var}[\mathbf{u}] &= \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \dots & \dots & \dots & \cdots & \dots \\ \rho^{n-1} & \rho^{n-2} & \dots & \cdots & 1 \end{pmatrix} = \sigma^2 \mathbf{V}, \quad \text{say.} \\ \text{var}[\hat{\beta}] &= \sigma^2 (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' \mathbf{V} \mathbf{x} (\mathbf{x}' \mathbf{x})^{-1} \\ &= \frac{\sigma^2}{(\mathbf{x}' \mathbf{x})^2} [\mathbf{x}' \mathbf{x} + f(\rho)] > \frac{\sigma^2}{(\mathbf{x}' \mathbf{x})}. \end{aligned}$$

14. $\mathbf{X}'\mathbf{X}$ is diagonal. Hence $\hat{\beta}_0 = \bar{Y}$, $\hat{\beta}_1 = (2/n) \sum_{t=1}^n Y_t \cos(2\pi k_1 t/n)$, $\hat{\beta}_2 = (2/n) \sum_{t=1}^n Y_t \sin(2\pi k_2 t/n)$.

EXERCISES 4a

- Using A.4, we see that $(\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$, and its inverse are all positive definite. Now use Theorem 4.1(ii).
- Proceed as in Theorem 4.1(iv) except that $\mathbf{A}\hat{\beta} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is replaced by $\mathbf{A}\hat{\beta} - \mathbf{c} = \mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)$ when $\mathbf{A}\beta = \mathbf{c}$.
- $\hat{\lambda}_H = 2[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{c})$, etc.
- From Section 4.5, $\mathbf{X}_A\gamma = (\mathbf{X}_1 - \mathbf{X}_2\mathbf{A}_2^{-1}\mathbf{A}_1)\beta_1$.
- Set $\mathbf{A} = (\mathbf{0}, \mathbf{I}_q)$.
 - $\hat{\beta}_2' \mathbf{B} \hat{\beta}_2$, where $\mathbf{B} = \mathbf{X}_2' \mathbf{R}_1 \mathbf{X}_2$ and $\mathbf{R}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$.
 - $\sigma^2 q + \beta_2' \mathbf{B} \beta_2$.

EXERCISES 4b

1. Since $\mathbf{1}_n \in \mathcal{C}(\mathbf{X})$, $(\mathbf{I}_n - \mathbf{P})\mathbf{1}_n = \mathbf{0}$, so that $(\mathbf{Y} - c\mathbf{1}_n)'(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - c\mathbf{1}_n) = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}$. The same is true for H .
 2. (a) This follows from $(\mathbf{X}'\mathbf{X})^{-1}$.
(b) The hypothesis is $H : (1, 0)\beta = 0$. Using the general matrix theory, $F = \hat{\beta}_0^2 / (\sum x_i^2 S^2 / n \sum (x_i - \bar{x})^2)$, where $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$, etc.
 - 3.
- $$\frac{(\hat{\beta}_1 - \bar{Y})^2}{S^2 \{(1/n) + [1/\sum(x_1 - \bar{x})^2]\}}.$$
4. $F = (\hat{\theta}_1 - 2\hat{\theta}_2)^2 / (\frac{7}{6}S^2)$, where $\hat{\theta}_1 = \frac{1}{2}(Y_1 - Y_3)$, $\hat{\theta}_2 = \frac{1}{6}(Y_1 + 2Y_2 + Y_3)$, and $S^2 = Y_1^2 + Y_2^2 + Y_3^2 = 2\hat{\theta}_1^2 = 6\hat{\theta}_2^2$.
 5. Using a Lagrange multiplier, $\hat{\theta}_i = Y_i - \bar{Y}$. Apply Theorem 4.1(i) with $\mathbf{A}\beta = \hat{\theta}_1 - \hat{\theta}_3$.

EXERCISES 4c

1. From (4.31), F is distributed as $F_{p-1, n-p}$; hence show that the distribution of R^2 is beta and find $E[R^2]$.
2. Let $Y_i* = Y_i/c$ and $x_{ij}* = x_{ij}k_j$, so that $\mathbf{X}^* = \mathbf{X}\mathbf{K}$, where \mathbf{K} is a nonsingular diagonal matrix $\text{diag}(1, k_1, \dots, k_{p-1})$. Apply the general theory to the starred variables and substitute.
3. (a) Use (4.30) twice. (b) $F > 0$.

EXERCISES 4d

1. $\mathbf{X}_H = (1 \ 1 \ 3)'$; $\mathbf{X}_H = \mathbf{1}_n$ ($n = n_1 + n_2$).
2. We can express the general model in the form $E[\mathbf{U}] = E \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$. Under $H : \gamma_1 = \gamma_2 = \gamma$, say, or $E[\mathbf{U}] = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \gamma$.

MISCELLANEOUS EXERCISES 4

1. Minimizing $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ subject to $\mathbf{1}_4'\boldsymbol{\theta} = 2\pi$ using a Lagrange multiplier gives us $\hat{\theta}_i = Y_i - \bar{Y} + \frac{1}{2}\pi$. Then RSS = $\sum(Y_i - \hat{\theta}_i)^2 = 4(\bar{Y} - \frac{1}{2}\pi)^2$. Under

$H, \theta_1 = \theta_3 = \phi_1, \theta_2 = \theta_4 = \pi - \phi_1$, and $\hat{\phi}_1 = \frac{1}{4}(Y_1 - Y_2 + Y_3 - Y_4 + 2\pi)$. Hence $\text{RSS}_H = (Y_1 - \hat{\phi}_1)^2 + (Y_2 - \pi + \hat{\phi}_1)^2 + (Y_3 - \hat{\phi}_1)^2 + (Y_4 - \pi + \hat{\phi}_1)^2$. Finally, $F = \frac{1}{2}(\text{RSS}_H - \text{RSS})/\text{RSS}$.

2. Use $H : (1, -1)\beta = 0$ and $F = \mathbf{A}\hat{\beta}[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\hat{\beta}/qS^2$ with $\mathbf{A} = (1, -1)$.
4. $t = \sqrt{n}(\bar{Y}_n - Y_{n+1})/S_n(1 + 1/n)^{1/2} \sim t_{n-1}$ when the means are equal, where $S_n^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

MISCELLANEOUS EXERCISES 5

1. $E[I_j] = 1 \cdot \text{pr}(I_j = 1) + 0 \cdot \text{pr}(I_j = 0) = \text{pr}(I_j = 1) = \alpha_j$. Hence $\gamma = E[\sum I_j] = \sum_j E[I_j] = \sum \alpha_j$.
2. Use a binomial expansion and show that the absolute values of successive terms after the first two ($= 1 - \alpha$) are decreasing, so that the remainder is positive.
3. Use the same method as for Miscellaneous Exercises 1, No. 1.
4. $\hat{\alpha}_0 = \bar{Y}$ and the least squares estimates of the β_j are the same. Let $\hat{\beta}' = (\hat{\alpha}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})$ and $\mathbf{v}' = (x_1 - \bar{x}_1, \dots, x_{p-1} - \bar{x}_{p-1})$. Then

$$\text{Var}[\hat{\beta}] = \sigma^2 \begin{pmatrix} \frac{1}{n} & \mathbf{0}' \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

and

$$\begin{aligned} \text{var}[\hat{Y}] &= (1, \mathbf{v}') \text{Var}[\hat{\beta}] (1, \mathbf{v}')' \\ &= \sigma^2 \left(\frac{1}{n} + \mathbf{v}' \mathbf{C} \mathbf{v} \right) \geq \frac{\sigma^2}{n}, \end{aligned}$$

since \mathbf{C} is positive definite. Equality occurs when $\mathbf{v} = \mathbf{0}$.

5. Let $\hat{Y}_{0G} = (\mathbf{x}'_0, \mathbf{z}'_0)\hat{\delta}_G$; then using an identical argument, we find that

$$\begin{aligned} \text{var}[\hat{Y}_0] &= \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + \sigma^2 (\mathbf{L}'\mathbf{x}_0 - \mathbf{z}_0)' \mathbf{M} (\mathbf{L}'\mathbf{x}_0 - \mathbf{z}_0) \\ &\geq \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0, \end{aligned}$$

because \mathbf{M} is positive definite.

6. By (5.15), $a_0\hat{\beta}_0 + a_1\hat{\beta}_1 \pm (2F_{2,n-2}^\alpha)^{1/2}\hat{v}^{1/2}$, where

$$\hat{v} = \frac{S^2 \{ a_0^2 (\sum x_i^2/n) - 2a_0 a_1 \bar{x} + a_1^2 \}}{\sum (x_i - \bar{x})^2}.$$

7. We can divide \mathbf{x} by x_0 , reducing the first element to 1, throughout the confidence interval. Change p to $p - 1$.

EXERCISES 6a

1. Follows from $Y_0 = \bar{Y} + \hat{\beta}_1(x_0 - \bar{x})$ and $(\mathbf{X}'\mathbf{X})^{-1}$.

2. The log likelihood function is

$$L(\phi, \beta_1) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_i (Y_i - \phi\beta_1 + \beta_1 x_1)^2.$$

$$\frac{\partial \log L}{\partial \phi} = 0 \Rightarrow \bar{Y} - \hat{\phi}\hat{\beta}_1 + \hat{\beta}_1 \bar{x} = 0.$$

$$\frac{\partial \log L}{\partial \beta_1} = 0 \Rightarrow \sum x_i (Y_i - \bar{Y} - \hat{\beta}_1(x_1 - \bar{x})) = 0,$$

and solve for ϕ and β_1 .

3. Apply the method of Section 6.1.2 to $U = \mathbf{a}_1'\hat{\beta} - \phi\mathbf{a}_2'\hat{\beta}$, where $\phi = \mathbf{a}_1'\beta / \mathbf{a}_2'\beta$. Let $\sigma_U^2 = (\mathbf{a}_1 - \phi\mathbf{a}_2)' \text{Var}[\hat{\beta}] (\mathbf{a}_1 - \phi\mathbf{a}_2)$ and show that

$$T = \frac{U/\sigma_U}{\sqrt{S^2/\sigma^2}} \sim t_{n-p}.$$

Then consider $T^2 = F_{1,n-p}^\alpha$ as a quadratic function of ϕ .

4. $\dot{x}_*/\hat{x}_* = \dot{\beta}_1 \hat{\beta}_1 = r^2$.

EXERCISES 6b

1. (a) $\beta_1^* = n^{-1} \sum (Y_i/x_i)$. (b) $\beta_1^* = \sum Y_i / \sum x_i$.

2. Same as equation (6.20) with $\sum x_i^2$ replaced by $\sum w_i x_i^2$, and $\tilde{\beta}$ and S^2 replaced by β^* and S_W^2 of (6.28) and (6.29), respectively.

3. Set $w_i = x_i^{-2}$.

EXERCISES 6c

- Minimize $\sum \sum (Y_{ki} - \alpha_k - \beta x_{ki})^2$ directly.
- $\hat{\beta} = \sum \sum (Y_{ki} - \bar{Y}_{..})(x_{ki} - \bar{x}_{..}) / \sum \sum (x_{ki} - \bar{x}_{..})^2$ and $\hat{\alpha} = \bar{Y}_{..} - \hat{\beta} \bar{x}_{..}$
- (a) Differentiate the sum of squares, write the normal equations for α and β_k in the form $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and reduce all the elements of

the first row of $\mathbf{X}'\mathbf{X}$, except the first, to zero by subtracting off suitable multiples of the other rows.

4. (a) Under H , $E[\mathbf{Y}] = (\mathbf{X}_1, \mathbf{z})(\beta'_1, \beta)' = \mathbf{W}\delta$, say, where $\mathbf{z} = \mathbf{X}'_2 \mathbf{1}_{p_2}$. Let $\mathbf{R}_1 = \mathbf{I}_n - \mathbf{P}_1$. Then, using Theorem 3.6,

$$\tilde{\beta} = (\mathbf{z}'\mathbf{R}_1\mathbf{z})^{-1}\mathbf{z}'\mathbf{R}_1\mathbf{Y} = \frac{\mathbf{z}'\mathbf{R}_1\mathbf{Y}}{\mathbf{z}'\mathbf{R}_1\mathbf{z}} = \frac{\mathbf{1}'_{p_2} \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y}}{\mathbf{1}'_{p_2} \mathbf{X}'_2(\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 \mathbf{1}_{p_2}}$$

and $RSS_H = \mathbf{Y}'\mathbf{R}_1\mathbf{Y} - \tilde{\beta}\mathbf{z}'\mathbf{R}_1\mathbf{Y}$.

- (b) Using Theorem 3.6(iii) with $\mathbf{Z} = \mathbf{X}_2$, we get $RSS = \mathbf{Y}'\mathbf{R}_1\mathbf{Y} - \hat{\beta}'_2 \mathbf{X}'_2 \mathbf{R}_1\mathbf{Y}$ and

$$RSS_H - RSS = \hat{\beta}'_2 \mathbf{X}'_2 \mathbf{R}_1\mathbf{Y} - \tilde{\beta}\mathbf{1}'_{p_2} \mathbf{X}'_2 \mathbf{R}_1\mathbf{Y} = (\hat{\beta}_2 - \tilde{\beta}\mathbf{1}_{p_2})' \mathbf{X}'_2 \mathbf{R}_1\mathbf{Y}.$$

- (c) We use Theorem 3.6(ii) to get $\mathbf{X}_1\hat{\beta}_{G,1} = \mathbf{X}_1\hat{\beta}_1 - \mathbf{P}_1\mathbf{X}_2\hat{\beta}$ and $\hat{\mathbf{Y}}_G = \mathbf{X}_1\hat{\beta}_{G,1} + \mathbf{X}_2\hat{\beta}_2 = \mathbf{P}_1\mathbf{Y} + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2\hat{\beta}_2$. Obtain a similar expression for $\hat{\mathbf{Y}}_H$, with $\mathbf{X}_2\hat{\beta}_2$ replaced by $\mathbf{X}'_2 \mathbf{1}_{p_2} \tilde{\beta}$, then use $RSS_H - RSS = \|\hat{\mathbf{Y}}_H - \hat{\mathbf{Y}}_G\|^2$.

MISCELLANEOUS EXERCISES 6

- Let $R_{Yx} = \sum(Y_i - \bar{Y})(x_i - \bar{x})$; then $\hat{x}_* - \bar{x} = (Y_* - \bar{Y})/(R_{Yx}/R_{xx})$ and $\tilde{x}_* - \bar{x} = (Y_* - \bar{Y})(R_{Yx}/R_{YY})$. Using $(n-2)S^2 = R_{YY} - R_{Yx}^2/R_{xx}$ we find that $F = R_{Yx}^2/(R_{YY}R_{xx})$, etc.
- Under $H : E[Y_{ki}] = b + \beta_k(x_{ki} - a)$. Obtain RSS_H by minimizing $\sum_k \sum_i [Y_{ki} - b - \beta_k(x_{ki} - a)]^2$ with respect to β_1 and β_2 . Also equivalent to shifting the origin to (a, b) and testing the hypothesis that both lines go through the origin.
- An estimate of the distance δ is $d = (\tilde{\alpha}_2 - \tilde{\alpha}_1)/\tilde{\beta}$, where $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, and $\tilde{\beta}$ are the LSEs from the parallel lines model (cf. Example 6.2). Use the method of Section 6.1.2 and consider $U = (\tilde{\alpha}_2 - \tilde{\alpha}_1) - \delta\beta = (\bar{Y}_2 - \bar{Y}_1) + \tilde{\beta}(\bar{x}_1 - \bar{x}_2 - \delta)$. Then $E[U] = 0$, and since $\text{cov}[\bar{Y}_{k\cdot}, Y_{ki} - \bar{Y}_{k\cdot}] = 0$,

$$\sigma_U^2 = \text{var}[U] = \sigma^2 \left\{ \frac{1}{n_2} + \frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2 - \delta)^2}{\sum \sum (x_{ki} - \bar{x}_{k\cdot})^2} \right\}.$$

Let $S^2 = RSS_H/(n_1 + n_2 - 3)$; then the confidence limits are the roots of the quadratic in δ given by $T^2 = F_{1,n_1+n_2-3}^\alpha$, where $T = (U/\sigma_U)/\sqrt{S^2/\sigma^2}$.

- $1/y = 1/\alpha + \sin^2 \theta (1/\beta - 1/\alpha)$. Let $x = \sin^2 \theta$, etc.

MISCELLANEOUS EXERCISES 7

1.

$$\begin{aligned}
 n &= 5, x = -2, -1, 0, 1, 2. \\
 \hat{\beta}_1 &= \frac{1}{10}(-2Y_1 - Y_2 + Y_4 + 2Y_5) = 1.65, \\
 \hat{\beta}_2 &= \frac{1}{14}(2Y_1 - Y_2 - 2Y_3 - Y_4 + 2Y_5) = -0.064, \\
 \hat{\beta}_3 &= \frac{1}{10}(-Y_1 + 2Y_2 - 2Y_4 + Y_5) = -0.167. \\
 \hat{Y} &= \bar{Y} + \hat{\beta}_1 x + \hat{\beta}_2(x^2 - 2) + \hat{\beta}_3 \frac{5}{6} \left(x^3 - \frac{68}{20}x \right) \\
 &= 13.02 + 1.65x - 0.064(x^2 - 2) - 0.167(x^3 - 3.4x). \\
 \text{RSS} &= \sum(Y_i - \bar{Y})^2 - 10\hat{\beta}_1^2 - 14\hat{\beta}_2^2 - 10\hat{\beta}_3^2 = 0.00514. \\
 F_{1,1} &= \frac{10\hat{\beta}_3^2}{\text{RSS}} = 78. H_0 : B_3 = 0 \text{ is not rejected.}
 \end{aligned}$$

2. Bias is zero, as the extra column is orthogonal to the original \mathbf{X} . $\hat{\beta}_{12} = \frac{1}{4}(Y_1 - Y_a - Y_b + Y_{ab})$.
3. Differentiating, we have $x_m = -\beta_1/2\beta_2$. Let $\hat{x}_m = -\hat{\beta}_1/2\hat{\beta}_2$ and consider $U = \hat{\beta}_1 + 2x_m\hat{\beta}_2$. Then $E[U] = 0$,

$$\sigma_U^2 = \text{var}[U] = \text{var}[\hat{\beta}_1] + 4x_m \text{cov}[\hat{\beta}_1, \hat{\beta}_2] + 4x_m^2 \text{var}[\hat{\beta}_2],$$

where $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and

$T = (U/\sigma_U)/\sqrt{S^2/\sigma^2} \sim t_{n-3}$. The confidence limits are the roots of $T^2 = F_{1,n-3}^\alpha$, a quadratic in x_m .

EXERCISES 8a

1. Use $H : \mu_1 - \mu_2 = \cdots = \mu_{I-1} - \mu_I$; then

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix},$$

of rank $I - 1$. We can also use $H : \mu_1 - \mu_I = \mu_2 - \mu_I = \cdots = \mu_{I-1} - \mu_I$

with $\mathbf{A}_1 = \begin{pmatrix} 1 & -0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$. Then $\mathbf{A} = \mathbf{A}_1 \mathbf{B}$, where \mathbf{B} is nonsingular.

2. Let $Y_{ij} - \bar{Y}_{..} = Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{..}$; then square and show that the cross-product term vanishes.
3. $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}(J_1^{-1}, \dots, J_I^{-1})$ and $\mathbf{X}'\mathbf{Y} = (J_1\bar{Y}_{1\cdot}, \dots, J_I\bar{Y}_{I\cdot})$.
4. $\sum \sum (\bar{Y}_{ij} - \bar{Y}_{i\cdot})^2 = \sum \sum Y_{ij}^2 - \sum_i Y_{i\cdot}^2 / J_i$.
5. Use Exercise 4 and the fact that the trace of a quadratic Q is the sum of the coefficients of the squared terms, then replace each random variable in the quadratic by its expected value. (a) $E[Q] = (I-1)\sigma^2 + \sum \sum (\mu_i - \sum_i J_i \mu_i / \sum_i J_i)^2$. (b) $E[Q] = (n-I)\sigma^2$.

EXERCISES 8b

1. From the text $H_{AB} \Rightarrow H_{AB3}$; the steps can be reversed so that $H_{AB3} \Rightarrow H_{AB}$. In a similar fashion the four hypotheses can be shown to be equivalent. For example, H_{AB1} implies that $\mu_{ij} - \mu_{Ij} = \mu_{ij_1} - \mu_{Ij_1}$ or $\mu_{ij} - \mu_{ij_1} = \mu_{Ij} - \mu_{Ij_1} = \mu_{i_2j} - \mu_{i_2j_1}$, which implies H_{AB} . Then, if $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, H_{AB3} implies that $(\alpha\beta)_{ij} = -(\overline{\alpha\beta})_{..} + (\overline{\alpha\beta})_{i\cdot} + (\overline{\alpha\beta})_{\cdot j}$ and

$$\mu_{ij} = (\mu - (\overline{\alpha\beta})_{..}) + (\alpha_i + (\overline{\alpha\beta})_{i\cdot}) + (\beta_j + (\overline{\alpha\beta})_{\cdot j}) = \mu^\dagger + \alpha_i^\dagger + \beta_j^\dagger.$$

2. Use the constraints to get $\bar{\mu}_i = \mu + \beta_j$, $\bar{\mu}_{..} = \mu$, etc.
3. Minimizing $\sum_{ijk} (Y_{ijk} - \mu)^2$ gives us $\mu_H = \bar{Y}_{..}$, so that $\text{RSS}_H - \text{RSS} = \|\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_H\|^2 = \sum_{ijk} (\hat{\mu}_{ij} - \hat{\mu}_H)^2 = \sum_{ij} K_{ij} (\bar{Y}_{ij\cdot} - \bar{Y}_{..})^2$, with $q = IJ - 1$. $E[Q] = (n - IJ)\sigma^2 + \sum_{ij} K_{ij} (\mu_{ij} - \sum_{ij} K_{ij} \mu_{ij} / n)^2$.

EXERCISES 8c

2. (a) For example, $\sum u_i \alpha_i = \sum u_i (A_i - \mu) = \sum u_i A_i - \mu = 0$, etc.
- (b) $A_i = \sum_j v_j \mu_{ij} = \sum_j v_j (\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}) = \mu + \alpha_i$ using (a), etc.
- (c) Suppose that one system of weights gives us $(\alpha\beta)_{ij}^\dagger = 0$. Consider another system (without \dagger 's). Then substitute $\mu_{ij} = \mu^\dagger + \alpha_i^\dagger + \beta_j^\dagger$ in A_i , B_j , and μ , and then use these expressions in $(\alpha\beta)_{ij} = \mu_{ij} - A_i - B_j + \mu$ to prove that $(\alpha\beta)_{ij} = 0$.
3. Using the weights given in the hint, we find that the decomposition (8.23) is still orthogonal, provided that we define $\bar{\varepsilon}_{i..} = \sum_j v_j \bar{\varepsilon}_{ij\cdot} = \sum_j K_{ij} \bar{\varepsilon}_{ij\cdot} / K_{..}$, $\bar{\varepsilon}_{..j} = \sum_i K_{ij} \bar{\varepsilon}_{ij\cdot} / K_{i\cdot} = \sum_i \sum_k \varepsilon_{ijk} / K_{i\cdot}$, and $\bar{\varepsilon}_{...} = \sum_i \sum_j \sum_k u_i v_j \bar{\varepsilon}_{ij\cdot} = \sum_i \sum_j \sum_k K_{ij} \bar{\varepsilon}_{ij\cdot} / K_{..} = \sum_i \sum_j \sum_k \varepsilon_{ijk} / K_{..}$, etc.

Least squares estimates are $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$, $\hat{\beta}_j = \bar{Y}_{.j..} - \bar{Y}_{...}$, and $(\hat{\alpha}\hat{\beta})_{ij} = \bar{Y}_{ij..} - \bar{Y}_{i..} - \bar{Y}_{.j..} + \bar{Y}_{...}$. The F -statistic is

$$F = \frac{\sum_i \sum_j K_{ij} (\hat{\alpha}\hat{\beta})_{ij}^2 / (I-1)(J-1)}{\sum \sum \sum (Y_{ijk} - \bar{Y}_{ij..})^2 / (K_{..} - IJ)}.$$

EXERCISES 8d

1. Substitute for $\mu_{ij} = \mu + \alpha_i + \beta_j$.
2. Show that the first-order interactions are the same for both tables.
3. $\mathbf{X}'_1 \mathbf{X}_1 \hat{\beta} = \mathbf{X}'_1 \mathbf{Y}$, $\mathbf{X}'_2 \mathbf{X}_2 \hat{\beta} = \mathbf{X}'_2 \hat{\mathbf{Y}}_2$. Adding gives $\mathbf{X}' \mathbf{X} \hat{\beta} = \mathbf{X}' (\mathbf{Y}, \hat{\mathbf{Y}}_2)$.
4. $\hat{Y}_{IJ} = \hat{\mu}_I = \bar{Y}_{I..} = (1/J)(Y_{I..} + \hat{Y}_{IJ})$, so that $\hat{Y}_{IJ} = Y_{I..}/(J-1)$, where $Y_{I..} = \sum_{j=1}^{J-1} Y_{Ij}$.

EXERCISES 8e

1. $\mathbf{Y}' \mathbf{R} \mathbf{Y} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i..})^2 = \sum_i R_{yyi} = R_{yy}$, say. Replacing Y_{ij} by $Y_{ij} - \gamma_i x_j$ and differentiating leads to $\hat{\gamma}_{i,G} = \sum_j (Y_{ij} - \bar{Y}_{i..})(x_j - \bar{x}) / \sum_j (x_j - \bar{x})^2 = R_{yx_i} / R_{xx}$, say, and $\text{RSS}_G = R_{yy} - \sum_i R_{yx_i}^2 / R_{xx}$. In a similar fashion, if $\gamma_1 = \gamma_2 = \gamma$, say, then $\hat{\gamma}_H = \sum_i R_{yx_i} / 2R_{xx}$ and $\text{RSS}_H = R_{yy} - (\sum_i R_{yx_i})^2 / 2R_{xx}$. Then $q = 1$. The next step is to construct a t -statistic. Now $R_{yx_i} = \sum_j Y_{ij}(x_j - \bar{x})$, so that $\text{var}[\hat{\gamma}_{i,G}] = \sigma^2 / R_{xx}$. Then $t = (\hat{\gamma}_1 - \hat{\gamma}_2) / S \sqrt{2/R_{xx}} \sim t_{2J-4}$, when H is true, where $S^2 = \text{RSS}_G / (2J-4)$. Then verify that $t^2 = F$.
2. (a) $\hat{\gamma}_1 = (R_{ww} R_{yz} - R_{zw} R_{yw}) / (R_{zz} R_{ww} - R_{yz}^2)$.
 (b) $\sigma^2 \begin{pmatrix} R_{zz} & R_{zw} \\ R_{zw} & R_{ww} \end{pmatrix}^{-1}$. (c) $R_{zw} = 0$.
3. $\hat{\gamma}_{ij} = R_{yzij} / R_{zzij}$, where $R_{yzij} = \sum_k (Y_{ijk} - \bar{Y}_{ij..})(z_{ijk} - \bar{z}_{ij..})$.

$$\begin{aligned} \text{RSS} &= R_{yy} - \sum_i \sum_j (R_{yzij}^2 / R_{zzij}), & \text{RSS}_H &= R_{yy} - (R_{yz}^2 / R_{zz}). \\ F &= \frac{(\text{RSS}_H - \text{RSS}) / (IJ-1)}{\text{RSS} / (IJK - 2IJ)}. \end{aligned}$$

MISCELLANEOUS EXERCISES 8

1. Show that $\text{cov}[\bar{\varepsilon}_{r..} - \bar{\varepsilon}_{..}, \varepsilon_{ij} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j..} + \bar{\varepsilon}_{...}] = 0$ (e.g., $\text{cov}[\bar{\varepsilon}_{r..}, \bar{\varepsilon}_{..}] = J^{-1} \text{var}[\bar{\varepsilon}_{r..}]$).

2. Differentiating $\sum \sum \varepsilon_{ij}^2 + \lambda \sum_i d_i \alpha_i$ with respect to μ and α_i gives us $\sum \sum (Y_{ij} - \mu - \alpha_i) = 0$ and $-2 \sum_j (Y_{ij} - \mu - \alpha_i) + \lambda d_i = 0$. Summing the second equation over i leads to $\lambda \sum d_i = 0$ or $\lambda = 0$. Then $\hat{\mu} = \sum d_i \bar{Y}_{i..} / \sum d_i$ and $\hat{\alpha}_i = \bar{Y}_{i..} - \hat{\mu}$.
3. (a) $\varepsilon_{ijk} = \bar{\varepsilon}_{...} + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{..j} - \bar{\varepsilon}_{...}) + \varepsilon_{ijk} - \bar{\varepsilon}_{ij..}$. Squaring and summing on i, j, k , the cross-product terms vanish. Hence, substituting for ε_{ijk} , we obtain $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_1 = \bar{Y}_{i..} - \bar{Y}_{...}$, $\hat{\beta}_{ij} = \bar{Y}_{ij..} - \bar{Y}_{i..}$. Zero covariance and Theorem 2.5 imply independence.

(b) Test for H_1 is

$$F = \frac{\sum \sum \sum (\bar{Y}_{ij..} - \bar{Y}_{i..})^2 / I(J-1)}{\sum \sum \sum (\bar{Y}_{ijk} - \bar{Y}_{ij..})^2 / IJ(K-1)}.$$

4. (a) $H : \begin{pmatrix} 1 & -2 & 0 & 0 \\ 0 & 2 & -3 & 0 \end{pmatrix} \mu = 0$. $F = [(Q_H - Q)/2]/[Q/(4J-4)]$, where $Q = \sum \sum (Y_{ij} - \bar{Y}_{i..})^2$, $Q_H = J\{(\bar{Y}_{1..} - 3\hat{\mu}_{3H})^2 + (\bar{Y}_{2..} - \frac{3}{2}\hat{\mu}_{3H})^2 + (\bar{Y}_{3..} - \hat{\mu}_{3H})^2\}$, and $\hat{\mu}_{3H} = (1/49)(12\bar{Y}_{1..} + 6\bar{Y}_{2..} + 4\bar{Y}_{3..})$.
- (b) Show that $\sum_{i=1}^2 \sum_{j=1}^J (\bar{Y}_{i..} - \bar{Y}_{...})^2 = (\bar{Y}_{1..} - \bar{Y}_{2..})^2 / (2/J)$.
5. (a) $\mu = \bar{\mu}_{...}$, $\alpha_i = \bar{\mu}_{i..} - \bar{\mu}_{...}$, $\beta_j = \bar{\mu}_{..j} - \bar{\mu}_{...}$, $\gamma_k = \bar{\mu}_{..k} - \bar{\mu}_{...}$.

(b) Use the decomposition

$$\begin{aligned} \varepsilon_{ijk} &= \bar{\varepsilon}_{...} + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{..j} - \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{..k} - \bar{\varepsilon}_{...}) \\ &\quad + (\varepsilon_{ijk} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{..j} - \bar{\varepsilon}_{..k} + 2\bar{\varepsilon}_{...}). \end{aligned}$$

Hence find RSS and RSS_H by inspection and obtain

$$F = \frac{\sum \sum \sum (\bar{Y}_{i..} - \bar{Y}_{...})^2 / (I-1)}{\sum \sum \sum (\bar{Y}_{ijk} - \bar{Y}_{i..} - \bar{Y}_{..j} - \bar{Y}_{..k} + 2\bar{Y}_{...})^2 / \nu},$$

where $\nu = IJK - I - J - K + 2$.

(c)

$$\begin{aligned} &\sum \sum \sum (\bar{\varepsilon}_{ij..} - \bar{\varepsilon}_{i..})^2 \\ &= \sum \sum \sum \{(\bar{\varepsilon}_{ij..} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{..j} + \bar{\varepsilon}_{...}) + (\bar{\varepsilon}_{..j} - \bar{\varepsilon}_{...})\}^2 \\ &= \sum \sum \sum (\bar{\varepsilon}_{ij..} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{..j} + \bar{\varepsilon}_{...})^2 \\ &\quad + \sum \sum \sum (\bar{\varepsilon}_{..j} - \bar{\varepsilon}_{...})^2 \end{aligned}$$

or $Q_1 = (Q_1 - Q_2) + Q_2$. Now $Q_1/\sigma^2 \sim \chi_{IJ-J}^2$, $Q_2/\sigma^2 \sim \chi_{J-1}^2$, $Q_1 - Q_2 \geq 0$, so that by Example 2.13, $(Q_1 - Q_2)/\sigma^2 \sim \chi_{IJ-I-J+1}^2$.

6. (a) Split up ε_{ijk} in the same way as μ_{ijk} and obtain an orthogonal decomposition. Hence $\hat{\mu} = \bar{Y}_{...}$, $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$, $\hat{\beta}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..}$, $\hat{\gamma}_k = \bar{Y}_{...k} - \bar{Y}_{...}$.

(b)

$$F = \frac{\sum_i \sum_j \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2 / (I-1)}{\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{...k} + \bar{Y}_{...})^2 / (IJK - IJ - K + 1)}.$$

EXERCISES 9a

- There is no bias in this case.
- Mean is zero and covariance matrix is $\sigma^2(\mathbf{I}_n - \mathbf{P})$.
- $E[\hat{Y}_0] = \mathbf{x}'_0 E[\hat{\beta}] = \mathbf{x}'_0 \boldsymbol{\beta}$. $E[\hat{Y}_{10}] = \mathbf{x}'_{10} E[\hat{\beta}_1] = \mathbf{x}'_{10} \boldsymbol{\beta}_1$. $\text{var}[\hat{Y}_0] = \sigma^2 \mathbf{x}'_0 (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}'_0$. $\text{var}[\hat{Y}_{10}] = \sigma^2 \mathbf{x}'_{10} (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{x}'_{10}$. As in Section 9.2.1, we find that $\text{var}[\hat{Y}_{10}] < \text{var}[\hat{Y}_0]$.

EXERCISES 9b

- $\mathbf{V}(\mathbf{I}_n - \mathbf{P}) = \mathbf{V} - \mathbf{V}\mathbf{P} = \mathbf{V} - [(1-\rho)\mathbf{I}_n + \rho\mathbf{1}_n\mathbf{1}'_n \mathbf{P}] = \mathbf{V} - [(1-\rho)\mathbf{P} + \rho\mathbf{1}_n\mathbf{1}'_n]$, since $\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$. Hence $\text{tr}[\mathbf{V}(\mathbf{I}_n - \mathbf{P})] = \text{tr}[\mathbf{V} - [(1-\rho)\mathbf{P} + \rho\mathbf{1}_n\mathbf{1}'_n]] = \text{tr}[\mathbf{V}] - (1-\rho) \text{tr}(\mathbf{P}) + \rho \text{tr}(\mathbf{1}_n\mathbf{1}'_n) = n - (1-\rho)p - n\rho - (n-p)(1-\rho)$, so $\text{var}[S^2] = [\sigma^2/(n-p)](n-p)(1-\rho) = \sigma^2(1-\rho)$.
- We must show that $(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^1 \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$ for each \mathbf{Y} . Write $\mathbf{Y} = \mathbf{Y}_1 + \mathbf{Y}_2$ where \mathbf{Y}_1 is in $\mathcal{C}(\mathbf{X})$ and \mathbf{Y}_2 is orthogonal to $\mathcal{C}(\mathbf{X})$. Then $\mathbf{Y}_1 = \mathbf{X}\mathbf{a}$ for some vector \mathbf{a} , so $(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^1 \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}_1 = \mathbf{a} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}_1$. To show that $(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^1 \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}_2 = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}_2$, we need only show that $\mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}_2 = 0$, since $\mathbf{X}' \mathbf{Y}_2 = 0$. But \mathbf{Y}_2 is orthogonal to $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{V}^{-1} \mathbf{X})$, so $\mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}_2 = 0$. The proof is complete.

EXERCISES 9c

- Use Exercises 4a, No. 2, at the end of Section 4.3.2.
- From (9.19), $E[Z] \sim \frac{1}{2}(f_2^{-1} - f_1^{-1})(1 + \frac{1}{2}\gamma_2 A)$. In this case $f_1 = k$, $f_2 = n - k - 1$, $\mathbf{P}_1 = \mathbf{P} - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$, and $\mathbf{P}_2 = \mathbf{I}_n - \mathbf{P}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' = [(p_{ij})]$. Hence

$$A = \frac{k^2 \sum (1 - p_{ii})^2 - (n - k - 1)^2 \sum [(1/n) - p_{ii}]^2}{k(n - k - 1)(2k - n + 1)}.$$

EXERCISES 9d

1. Put $s_{xz} = \sum_i (x_i - \bar{x})(z_i - \bar{z})$. Then

$$\begin{aligned}\mathbf{w}' \mathbf{S}^{-1} \mathbf{w} &= \frac{n}{s_{xx}s_{zz} - s_{xz}^2} (\bar{x}, \bar{z})' \begin{pmatrix} s_{zz} & -s_{xz} \\ -s_{xz} & s_{xx} \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{z} \end{pmatrix} \\ &= \frac{n}{1-r^2} \left[\frac{\bar{x}^2}{s_{xx}} - \frac{2\bar{x}\bar{z}r}{\sqrt{s_{xx}s_{zz}}} + \frac{\bar{z}^2}{s_{zz}} \right] \\ &= \frac{1}{1-r^2} \left[\frac{1}{\text{CV}_x^2} - \frac{2 \text{sign}(\bar{x}) \text{sign}(\bar{z}) r}{\text{CV}_x \text{CV}_z} + \frac{1}{\text{CV}_z^2} \right].\end{aligned}$$

2. $1 = \|\mathbf{x}^{*(j)}\|^2 = \|\mathbf{P}_j \mathbf{x}^{*(j)} + (\mathbf{I}_n - \mathbf{P}_j) \mathbf{x}^{*(j)}\|^2 = \|(\mathbf{I}_n - \mathbf{P}_j) \mathbf{x}^{*(j)}\|^2 + \|\mathbf{P}_j \mathbf{x}^{*(j)}\|^2$, so that by (9.54), $R_j^2 = \|\mathbf{P}_j \mathbf{x}^{*(j)}\|^2$. Thus $R_j^2 = 0$ if and only if $\mathbf{P}_j \mathbf{x}^{*(j)} = 0$ (i.e., if and only if $\mathbf{x}^{*(j)}$ is orthogonal to the columns of $\mathbf{X}^{*(j)}$).

3. Since $\mathbf{R}_{xz} = \mathbf{X}^{*\prime} \mathbf{X}^*$, it is enough to show that $\text{rank}(\mathbf{X}^*) = p-1$. Suppose that $\mathbf{c} = (c_1, \dots, c_{p-1})'$ is a vector with

$$c_1 \mathbf{x}^{*(1)} + \dots + c_{p-1} \mathbf{x}^{*(p-1)} = \mathbf{0}.$$

Since $\mathbf{x}^{*(j)} = (\mathbf{x}^{(j)} - \bar{x}_j \mathbf{1}_n)/s_j$, we get

$$-(c_1 \bar{x}_1/s_1 + \dots + c_{p-1} \bar{x}_1/s_{p-1}) \mathbf{1}_n + c_1 \mathbf{x}^{(1)}/s_1 + \dots + c_{p-1} \mathbf{x}^{(p-1)}/s_{p-1} = \mathbf{0}.$$

Since \mathbf{X} has full rank, $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p-1)}$ are linearly independent, so $c_1 = \dots = c_{p-1} = 0$ and hence $\text{rank}(\mathbf{X}^*) = p-1$.

4. Let \mathbf{a}_0 be a $(p-1)$ -vector with last element 1 and the rest zero. Then $\lambda_{\text{MAX}} = \max_{\mathbf{a}} \|\mathbf{X}^* \mathbf{a}\|^2 / \|\mathbf{a}\|^2 \geq \|\mathbf{X}^* \mathbf{a}_0\|^2 / \|\mathbf{a}_0\|^2 = \|\mathbf{x}^{*(p-1)}\|^2 = 1$. Also, if \mathbf{x}_i^* is the i th row of \mathbf{X}^* , then for any $(p-1)$ -vector \mathbf{a} , $\|\mathbf{X}^* \mathbf{a}\|^2 / \|\mathbf{a}\|^2 = \sum_i (\mathbf{x}_i^* \mathbf{a})^2 / \|\mathbf{a}\|^2 \leq \sum_i \|\mathbf{x}_i^*\|^2 \|\mathbf{a}\|^2 = \sum_i \sum_j x_{ij}^{*2} \|\mathbf{a}\|^2 = (p-1) \|\mathbf{a}\|^2$, so that $\lambda_{\text{MAX}} \leq p-1$. Also,

$$\begin{aligned}\text{VIF}_j &= \sum_{k=1}^{p-1} t_{ki}^2 / \lambda_k \\ &\leq \sum_{k=1}^{p-1} t_{ki}^2 / \lambda_{\text{MIN}} \\ &= 1 / \lambda_{\text{MIN}} \\ &= \lambda_{\text{MAX}} / \lambda_{\text{MIN}} \\ &= \kappa^2.\end{aligned}$$

5. $\sum_i (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_i x_i(Y_i - \bar{Y})$ so $\frac{\partial}{\partial x_i} \sum_i (x_i - \bar{x})(Y_i - \bar{Y}) = (Y_i - \bar{Y})$. $\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - (\sum_i x_i)^2/n$, so $\frac{\partial}{\partial x_i} \sum_i (x_i - \bar{x})^2 = 2x_i - 2(\sum_i x_i)/n = 2(x_i - \bar{x})$.

MISCELLANEOUS EXERCISES 9

1. $E[\hat{\beta}_0] = \beta_3 + 4\beta_2$, $E[\hat{\beta}_1] = \beta_1 + 7\beta_3$.

2. $e_i = Y_i - \bar{Y} - \hat{\beta}_1(x_{i1} - \bar{x}_1)$, where

$$\hat{\beta}_1 = \left[\sum_i Y_i(x_{i1} - \bar{x}_1) \right] / \sum_i (x_{i1} - \bar{x}_1)^2.$$

Also, $E[Y_i - \bar{Y}] = \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2)$ and

$$E[\hat{\beta}_1] = \left[\sum_i E[Y_i](x_{i1} - \bar{x}_1) \right] / \sum_i (x_{i1} - \bar{x}_1)^2 = \beta_1 + \eta\beta_2,$$

where

$$\eta = \left[\sum_i (x_{i21} - \bar{x}_1)(x_{i2} - \bar{x}_2) \right] / \sum_i (x_{i1} - \bar{x}_1)^2.$$

Thus

$$\begin{aligned} E[e_i] &= \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) - (\beta_1 + \eta\beta_2)(x_{i1} - \bar{x}_1) \\ &= [x_{i2} - \eta x_{i1} - (\bar{x}_2 - \eta \bar{x}_1)]\beta_2. \end{aligned}$$

3. From Example 4.5, $RSS_H - RSS = n_1 n_2 (\bar{U} - \bar{V})^2 / (n_1 + n_2) = \mathbf{Y}' \mathbf{P}_1 \mathbf{Y}$ and $RSS = \sum(U_j - \bar{U})^2 + \sum(V_j - \bar{V})^2 = \mathbf{Y}' \mathbf{P}_2 \mathbf{Y}$. We then use (9.18) and (9.19) with $f_1 = 1$, $f_2 = n_1 + n_2 - 2$,

$$\mathbf{P}_1' = \frac{n_1 n_2}{n_1 + n_2} \left(\frac{1}{n_1^2} \mathbf{1}'_{n_1}, \frac{1}{n_2^2} \mathbf{1}'_{n_2} \right)$$

and

$$\mathbf{P}_2' = \left[\left(1 - \frac{1}{n_1} \right) \mathbf{1}'_{n_1}, \left(1 - \frac{1}{n_2} \right) \mathbf{1}'_{n_2} \right].$$

When $n_1 = n_2$, F is quadratically balanced.

4. Let $\tilde{U} = U - \mu_U$. Then $X = \mu_U + \tilde{U} + \delta$ and $Y = \beta_0 + \beta_1 \mu_U + \beta_1 \tilde{U} + \epsilon$, so

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_U \\ \beta_0 + \beta_1 \mu_U \end{pmatrix} + \begin{pmatrix} 1 & 0 & 1 \\ \beta_1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \tilde{U} \\ \epsilon \\ \delta \end{pmatrix}.$$

$(\tilde{U}, \epsilon, \delta)$ has a multivariate normal distribution with zero mean vector and covariance matrix $\text{diag}(\sigma_U^2, \sigma_\epsilon^2, \sigma_\delta^2)$, so by Theorem 2.2, (X, Y) has a multivariate normal distribution with mean vector $(\mu_U, \beta_0 + \beta_1 \mu_U)$ and covariance matrix

$$\begin{aligned} \begin{pmatrix} 1 & 0 & 1 \\ \beta_1 & 1 & 0 \end{pmatrix} \text{diag}(\sigma_U^2, \sigma_\epsilon^2, \sigma_\delta^2) \begin{pmatrix} 1 & 0 & 1 \\ \beta_1 & 1 & 0 \end{pmatrix}' \\ = \begin{pmatrix} \sigma_U^2 + \sigma_\delta^2 & \beta_1 \sigma_U^2 \\ \beta_1 \sigma_U^2 & \beta_1^2 \sigma_U^2 + \sigma_\epsilon^2 \end{pmatrix}. \end{aligned}$$

EXERCISES 10a

1. From (10.11), $h_i = n^{-1} + (n-1)^{-1} \mathbf{M} \mathbf{D}_i$. For $p=2$, $\mathbf{M} \mathbf{D}_i = (n-1) \times (x_i - \bar{x})^2 / \sum_i (x_i - \bar{x})^2$, so that $h_i = n^{-1} + (x_i - \bar{x})^2 / \sum_i (x_i - \bar{x})^2$.
2. Follows from the change-of-variable formula. $B = \alpha F / (\alpha F + \beta)$.
3. (a) $(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\beta + \boldsymbol{\epsilon}) = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}$, so that $e_i = \mathbf{c}'_i(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}$, where \mathbf{c}_i has i th element 1 and the rest 0.

(b)

$$\begin{aligned} (n-p)^{-1} r_i^2 &= \frac{(n-p)^{-1} e_i^2}{S^2(1-h_i)} \\ &= \frac{[\mathbf{c}'_i(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}]^2}{\boldsymbol{\epsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}(1-h_i)} \\ &= \frac{\mathbf{Z}' \mathbf{Q} \mathbf{Z}}{\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}}, \end{aligned}$$

where $\mathbf{Z} = \boldsymbol{\epsilon}/\sigma$, since $\text{RSS} = \boldsymbol{\epsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}$.

- (c) Follows from $\mathbf{c}'_i(\mathbf{I}_n - \mathbf{H})\mathbf{c}_i = 1 - h_i$.
- (d) $(\mathbf{I}_n - \mathbf{H})\mathbf{Q} = \mathbf{Q}$ so $(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n - \mathbf{H} + \mathbf{Q} - 2\mathbf{Q} = \mathbf{I}_n - \mathbf{H} - \mathbf{Q}$. $\mathbf{Z}' \mathbf{Q} \mathbf{Z}$ and $\mathbf{Z}'(\mathbf{I}_n - \mathbf{H} - \mathbf{Q})\mathbf{Z}$ are independent since $\mathbf{Q}(\mathbf{I}_n - \mathbf{H} - \mathbf{Q}) = 0$. Also $\text{tr}(\mathbf{Q}) = (1-h_i)^{-1} \text{tr}[\mathbf{c}_i \mathbf{c}'_i(\mathbf{I}_n - \mathbf{H})] = (1-h_i)^{-1}(1-h_i) = 1$, so that $\text{rank}(\mathbf{Q}) = \text{tr}(\mathbf{Q}) = 1$ and $\text{rank}(\mathbf{I}_n - \mathbf{H} - \mathbf{Q}) = \text{tr}(\mathbf{I}_n - \mathbf{H} - \mathbf{Q}) = n-p-1$. Thus $\mathbf{Z}' \mathbf{Q} \mathbf{Z} / \mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}$ is of the form $\chi_1^2 / (\chi_1^2 + \chi_{n-p-1}^2)$ and so is $B(1/2, (n-p-1)/2)$.
4. $(\mathbf{I}_n - \mathbf{H})^2$ has i, i element $\sum_{j=1}^n (\delta_{ij} - h_{ij})(\delta_{ij} - h_{ij}) = (1-h_i)^2 + \sum_{j \neq i} h_{ij}$. Since $(\mathbf{I}_n - \mathbf{H})^2$ is idempotent, this is just $(1-h_i)$.

EXERCISES 10b

1. (a) Assuming that the first column of \mathbf{X} is all 1's, $\sum_{i=1}^n (x_{ij} - \bar{x}_j) \mathbf{e}_i = [\mathbf{X}'(\mathbf{I} - \mathbf{P})\mathbf{Y}]_{j+1} - \bar{x}_j [\mathbf{X}'(\mathbf{I} - \mathbf{P})\mathbf{Y}]_1 = 0$ since $(\mathbf{I} - \mathbf{P})\mathbf{X} = 0$.

- (b) Numerator of squared correlation is $[\sum_{i=1}^n (x_{ij} - \bar{x})(e_i + \hat{\beta}_j(x_{ij} - \bar{x}))]^2 = \hat{\beta}_j^2 [\sum_{i=1}^n (x_{ij} - \bar{x})^2]^2$. Denominator is $\sum_{i=1}^n (x_{ij} - \bar{x})^2 \times \sum_{i=1}^n [e_i + \hat{\beta}_j(x_{ij} - \bar{x})]^2 = \sum_{i=1}^n (x_{ij} - \bar{x})^2 [\text{RSS} + \hat{\beta}_j^2 \sum_{i=1}^n (x_{ij} - \bar{x})^2]$.
2. (b) Plotting the function $g(x) = 2 \text{sign}(x)|x|^{1/3}$ on top of the partial residual plot shows that the plot reveals the shape of g very well.
- (d) In this case the form of g is not revealed at all.
3. The partial residual plot gives the impression of a linear relationship for $r = 0.999$ but not for $r = 0$.
4. The added variable plot suggests no relationship for both $r = 0.999$ and $r = 0$.
5. Partial residual plot fails to reveal g in the first case but does in the second.

EXERCISES 10c

1. From (10.35) $\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = \frac{\partial \ell}{\partial \beta} \mathbf{X}' \Sigma^{-1} (\mathbf{Y} - \mathbf{X}\beta) = -\sigma^{-2} \mathbf{X}' \mathbf{X}$, since $\Sigma = \sigma^2 \mathbf{I}_n$ under H_0 . Also, from (10.35),

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta \partial \lambda'} &= 1/2 \sum_i (Y_i - \mathbf{x}'_i \beta)^2 / w_i^2 \frac{\partial w_i}{\partial \lambda} \\ &= \sum_i x_{ij} (Y_i - \mathbf{x}'_i \beta) / w_i^2 \frac{\partial w}{\partial \lambda}, \end{aligned}$$

which has zero expectation. Finally,

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \lambda \partial \lambda'} &= -1/2 \sum_i (1/w_i - \varepsilon_i^2/w_i^2) \frac{\partial^2 w_i}{\partial \lambda \partial \lambda'} \\ &\quad + 1/2 \sum_i (1/w_i^2 - 2\varepsilon_i^2/w_i^3) \frac{\partial w_i}{\partial \lambda} \frac{\partial w_i}{\partial \lambda} \\ &= -\frac{1}{2} \mathbf{D}' \mathbf{D} \sigma^{-4} \end{aligned}$$

since $E[1/w_i - \varepsilon_i^2/w_i^2] = 0$ and $E[1/w_i^2 - 2\varepsilon_i^2/w_i^3] = -\sigma^{-4}$.

2. (a) $\mathbf{G}' \Sigma \mathbf{Q} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \Sigma \mathbf{Q} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{Q} = \mathbf{0}$ since $\mathbf{Q}' \mathbf{X} = \mathbf{0}$, and
- $$\mathbf{G}' \Sigma \mathbf{G} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \Sigma \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1}.$$
- (b) Assume that \mathbf{X} is $n \times p$ of rank p . We need to show that the matrix $(\Sigma^{1/2} \mathbf{Q}, \Sigma^{1/2} \mathbf{G})$ is of rank n . $\mathbf{G}' \Sigma \mathbf{Q} = \mathbf{0}$, so the columns of $\Sigma^{1/2} \mathbf{Q}$

are orthogonal to the columns of $\Sigma^{1/2}\mathbf{G}$. Since $\Sigma^{1/2}$ is nonsingular and \mathbf{Q} is of rank $n - p$, $\Sigma^{1/2}\mathbf{Q}$ is of rank $n - p$. Thus, because of the orthogonality, it is enough to show that $\text{rank}(\Sigma^{1/2}\mathbf{G}) = p$. This follows by A.2.2 since $\Sigma^{1/2}\mathbf{G}$ is of the form \mathbf{AXB} with \mathbf{A} and \mathbf{B} nonsingular, and \mathbf{X} is of rank p .

(c) By (b),

$$\begin{aligned}\mathbf{I}_n &= \mathbf{M} \\ &= (\Sigma^{1/2}\mathbf{Q}, \Sigma^{1/2}\mathbf{G}) \begin{pmatrix} \mathbf{Q}'\Sigma\mathbf{Q} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Q}'\Sigma^{1/2} \\ \mathbf{G}'\Sigma^{1/2} \end{pmatrix} \\ &= \Sigma^{1/2}\mathbf{Q}(\mathbf{Q}'\Sigma\mathbf{Q})^{-1}\mathbf{Q}'\Sigma^{1/2} + \Sigma^{1/2}\mathbf{G}(\mathbf{X}'\Sigma^{-1}\mathbf{X})\mathbf{G}'\Sigma^{1/2} \\ &= \Sigma^{1/2}[\mathbf{Q}(\mathbf{Q}'\Sigma\mathbf{Q})^{-1}\mathbf{Q}' + \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{1/2}].\end{aligned}$$

Pre- and postmultiplying by $\Sigma^{1/2}$ and rearranging gives the result.

(d)

$$\begin{aligned}\det(\mathbf{M})^2 &= \det(\mathbf{M}'\mathbf{M}) \\ &= \det \left[\begin{pmatrix} \mathbf{Q}'\Sigma\mathbf{Q} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \end{pmatrix} \right] \\ &= \det(\mathbf{Q}'\Sigma\mathbf{Q})/\det(\mathbf{X}'\Sigma^{-1}\mathbf{X}).\end{aligned}$$

Also,

$$\begin{aligned}\det(\mathbf{M})^2 &= \det[\Sigma^{1/2}(\mathbf{Q}, \mathbf{G})]^2 \\ &= \det(\Sigma) \det \left[\begin{pmatrix} \mathbf{I}_n & \mathbf{Q}'\mathbf{G} \\ \mathbf{G}'\mathbf{Q} & \mathbf{G}'\mathbf{G} \end{pmatrix} \right] \\ &= \det(\Sigma) \det(\mathbf{G}'\mathbf{G} - \mathbf{G}'\mathbf{Q}\mathbf{Q}'\mathbf{G}) \\ &= \det(\Sigma)/\det(\mathbf{X}'\mathbf{X}),\end{aligned}$$

since $\mathbf{G}'\mathbf{G} - \mathbf{G}'\mathbf{Q}\mathbf{Q}'\mathbf{G} = \mathbf{G}'(\mathbf{I}_n - \mathbf{Q}\mathbf{Q}')\mathbf{G} = \mathbf{G}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}$. Combining these relations proves the result.

3. (a) From (10.40), the restricted likelihood is

$$\begin{aligned}l_R(\sigma^2) &= c - 1/2\{\log \det(\sigma^2\mathbf{I}_n) + \log \det(\mathbf{1}'\sigma^{-2}\mathbf{I}_n\mathbf{1}) \\ &\quad - \log \det(\mathbf{1}\mathbf{1}') + \sigma^{-2}\mathbf{Y}'(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\mathbf{Y}\} \\ &= c' - 1/2\{n \log \sigma^2 + \log(n\sigma^{-2}) + \sum(Y_i - \bar{Y})^2/\sigma^2\} \\ &= c'' - 1/2\{(n-1) \log \sigma^2 + \sum(Y_i - \bar{Y})^2/\sigma^2\}.\end{aligned}$$

(b) Differentiating and equating to zero gives the REML estimate $\hat{\sigma}^2 = (n-1)^{-1} \sum(Y_i - \bar{Y})^2$, the usual unbiased estimate.

4. For this example, $w(\mu) = \mu^2/r$, so that $f(\mu) = r^{1/2} \int \frac{d\mu}{\mu} = r^{1/2} \log \mu$.

EXERCISES 10d

1. (a)

$$\begin{aligned} E[Z_{(i)}] &= B(i, n - i + 1)^{-1} \int_{-\infty}^{\infty} z \Phi(z)^{i-1} [1 - \Phi(z)]^{n-i} \phi(z) dz \\ &= B(i, n - i + 1)^{-1} \int_0^1 \Phi^{-1}(y) y^{i-1} (1-y)^{n-i} dy, \end{aligned}$$

after making the change of variable $y = \Phi(z)$.

(b)

$$\begin{aligned} E[Z_{(i)}] &= B(i, n - i + 1)^{-1} \sum_{j=1}^n \int_{(j-1)/n}^{j/n} \Phi^{-1}(y) y^{i-1} (1-y)^{n-i} dy \\ &\approx \sum_{j=1}^n \Phi^{-1}[(i - 0.5)/n] w_{ij}. \end{aligned}$$

(c) The sum of the weights is the area under the beta density.

(d) Accurate if w_{ij} is small for $i \neq j$.2. (a) Let f_i be the density of Y_i . Then for some λ , the density of $g(Y_i, \lambda)$ is $N(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$, so by the change-of-variable formula

$$\frac{1}{\sigma} \phi[(u - \mathbf{x}_i' \boldsymbol{\beta})/\sigma] = f_i(g^{-1}(u)) \left/ \left| \frac{\partial g(y_i, \lambda)}{\partial y_i} \right| \right.,$$

and hence

$$f_i(y_i) = \frac{1}{\sigma} \phi\{[g(y_i, \lambda) - \mathbf{x}_i' \boldsymbol{\beta}]/\sigma\} \left| \frac{\partial g(y_i, \lambda)}{\partial y_i} \right|.$$

The log likelihood is

$$\begin{aligned} \sum_{i=1}^n \log f_i(y_i) &= -(n/2) \log \sigma^2 - (1/2\sigma^2) \sum_{i=1}^n [g(y_i, \lambda) - \mathbf{x}_i' \boldsymbol{\beta}]^2 \\ &\quad + \sum_{i=1}^n \log \left| \frac{\partial g(y_i, \lambda)}{\partial y_i} \right|. \end{aligned}$$

(b) For the John-Draper transformation, $\left| \frac{\partial g(y_i, \lambda)}{\partial y_i} \right| = (1 + |y_i|)^{\lambda-1}$, so the result follows from (a).3. Minimize $-\int h_\lambda(\mathbf{y}) [-(n/2) \log \sigma^2 - (1/2\sigma^2) \sum_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2] d\mathbf{y}$. Differentiating with respect to $\boldsymbol{\beta}$, we get $\int h_\lambda(\mathbf{y}) \sum_i x_{ij} (y_i - \mathbf{x}_i' \boldsymbol{\beta}) d\mathbf{y} = 0$

for each j , so that $\beta_* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_\lambda[\mathbf{Y}]$. Differentiating with respect to σ^2 gives $(n/2\sigma_*^2) - (1/2)\sigma^4 E[\sum_{i=1}^n (y_i - \mathbf{x}_i'\beta_*)^2] = 0$, so $\sigma_*^2 = n^{-1}E[||\mathbf{Y} - \mathbf{P}E_\lambda[\mathbf{Y}]||^2] = n^{-1}\{E_\lambda[\mathbf{Y}]'(\mathbf{I}_n - \mathbf{P})E_\lambda[\mathbf{Y}] + \text{tr}(\text{Var}_\lambda[\mathbf{Y}])\}$.

EXERCISES 10e

1. $\mathbf{d}'(\hat{\beta}(i) - \hat{\beta}) = \mathbf{d}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i / (1 - h_i) = (\mathbf{C}'\mathbf{d})_i e_i / (1 - h_i)$.

2.

$$\begin{aligned}\hat{\beta}(D) &= (\mathbf{X}(D)'\mathbf{X}(D))^{-1}\mathbf{X}(D)'\mathbf{Y}(D) \\ &= (\mathbf{X}'\mathbf{X} - \mathbf{X}'_D\mathbf{X}_D)^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{X}'_D\mathbf{Y}_D) \\ &= [(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{X}_D(\mathbf{X}'\mathbf{X})^{-1}] \\ &\quad \times (\mathbf{X}'\mathbf{Y} - \mathbf{X}'_D\mathbf{Y}_D) \\ &= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1} \\ &\quad \times [\mathbf{X}_D\hat{\beta} - \mathbf{H}_D\mathbf{Y}_D - (\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{Y}_D] \\ &= \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D.\end{aligned}$$

3.

$$\begin{aligned}\text{AP}(D) &= \frac{\det[\mathbf{X}_A(D)'\mathbf{X}_A(D)]}{\det(\mathbf{X}'_A\mathbf{X}_A)} \\ &= \frac{\det[\mathbf{X}(D)'\mathbf{X}(D)]}{\det(\mathbf{X}'\mathbf{X})} \cdot \frac{\text{RSS}(D)}{\text{RSS}}\end{aligned}$$

Using A.9.6, we get $\det[\mathbf{X}(D)'\mathbf{X}(D)] = \det(\mathbf{X}'\mathbf{X}) \times \det(\mathbf{I}_d - \mathbf{H}_D)$. Also, by (10.60),

$$\begin{aligned}\mathbf{Y} - \mathbf{X}\hat{\beta}(D) &= \mathbf{Y} - \mathbf{X}\hat{\beta} + [\mathbf{X}(\hat{\beta} - \hat{\beta}(D))] \\ &= \mathbf{e} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D,\end{aligned}$$

so that

$$\begin{aligned}\text{RSS}(D) &= ||\mathbf{Y}(D) - \mathbf{X}(D)\hat{\beta}(D)||^2 \\ &= ||\mathbf{Y} - \mathbf{X}\hat{\beta}(D)||^2 - ||\mathbf{Y}_D - \mathbf{X}_D\hat{\beta}(D)||^2 \\ &= ||\mathbf{e} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D||^2 \\ &\quad - ||\mathbf{e}_D + \mathbf{H}_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D||^2 \\ &= \text{RSS} + \mathbf{e}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{H}_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D \\ &\quad - \mathbf{e}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-2}\mathbf{e}_D \\ &= \text{RSS} - \mathbf{e}'_D(\mathbf{I}_d - \mathbf{H}_D)^{-1}\mathbf{e}_D.\end{aligned}$$

Combining these results, we get the desired expression for $\text{AP}(D)$.

4. The augmented hat matrix is

$$\begin{aligned} & (\mathbf{X}, \mathbf{Y}) \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}' \\ \mathbf{Y}' \end{pmatrix} \\ &= \mathbf{H} + (\mathbf{I}_n - \mathbf{H})\mathbf{Y}\mathbf{Y}'(\mathbf{I}_n - \mathbf{H})/\text{RSS}. \end{aligned}$$

The diagonal elements are of the form $h_i + e_i^2/\text{RSS}$. Thus, by (10.56), $\text{AP}(i) = 1 - h_{i,A}$.

5. (a) $(n-1)^{-1}(n\hat{F}_n - \delta_i) = (n-1)^{-1}(\sum_j \delta_j - \delta_i) = (n-1)^{-1} \sum_{j \neq i} \delta_j$, which is the empirical distribution function of the sample with point i deleted.
- (b) $T(\hat{F}_n) = \hat{\beta}$, so that by (a), $T[(n-1)^{-1}(n\hat{F}_n - \delta_i)] = \hat{\beta}(i)$. Thus $\text{SIC}_i = (n-1)^{-1}(\hat{\beta}(i) - \hat{\beta})$.
- (c) $c^{-1}(\text{SIC}_i)' \mathbf{M}(\text{SIC}_i) = (\hat{\beta}(i) - \hat{\beta})' \mathbf{X}'\mathbf{X}(\hat{\beta}(i) - \hat{\beta})/pS^2$, which is Cook's D .

EXERCISES 10f

1. Let \mathbf{a}_0 be a p -vector with last element 1 and the rest zero. Then $\check{\lambda}_{\text{MAX}} = \max_{\mathbf{a}} \|\check{\mathbf{X}}\mathbf{a}\|^2/\|\mathbf{a}\|^2 \geq \|\check{\mathbf{X}}\mathbf{a}_0\|^2/\|\mathbf{a}_0\|^2 = \|\check{\mathbf{x}}^{(p-1)}\|^2 = 1$. Also, if $\check{\mathbf{x}}_i$ is the i th row of $\check{\mathbf{X}}$, then for any p -vector \mathbf{a} , $\|\check{\mathbf{X}}\mathbf{a}\|^2/\|\mathbf{a}\|^2 = \sum_i (\check{\mathbf{x}}_i' \mathbf{a})^2 \leq \sum_i \|\check{\mathbf{x}}_i\|^2 \|\mathbf{a}\|^2 = \sum_i \sum_j \check{x}_{ij}^2 \|\mathbf{a}\|^2 = p\|\mathbf{a}\|^2$, so that $\check{\lambda}_{\text{MAX}} \leq p$.
2. $\mathbf{X}\hat{\beta} = \begin{pmatrix} \bar{Y} \\ \mathbf{X}^*\hat{\gamma} \end{pmatrix}$ and $\mathbf{X}\beta = \begin{pmatrix} \alpha \\ \mathbf{X}^*\gamma \end{pmatrix}$, so that $\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\|^2 = (\bar{Y} - \alpha)^2 + \|\mathbf{X}^*\hat{\gamma} - \mathbf{X}^*\gamma\|^2$. The result follows by taking expectations.
3. Posterior mean is $(\mathbf{X}_s'\mathbf{X}_s + \mathbf{V}^{-1})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}_s'\mathbf{Y})$. With the assumed priors, $\mathbf{X}_s'\mathbf{X}_s + \mathbf{V}^{-1} = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & R_{xx} \end{pmatrix} + \begin{pmatrix} c & \mathbf{0}' \\ \mathbf{0} & k\mathbf{I}_{p-1} \end{pmatrix} = \begin{pmatrix} n+c & \mathbf{0}' \\ \mathbf{0} & R_{xx} + k\mathbf{I}_{p-1} \end{pmatrix}$. Also, $\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}_s'\mathbf{Y} = \begin{pmatrix} \sum_i Y_i \\ \mathbf{X}_s^*\mathbf{Y} \end{pmatrix}$, so the posterior mean is $\begin{pmatrix} \sum_i Y_i/(n+c) \\ (R_{xx} + k\mathbf{I}_{p-1})^{-1}\mathbf{X}_s^*\mathbf{Y} \end{pmatrix}$.
4. Conditional on $\gamma, \hat{\gamma}$ has a $N_{p-1}(\gamma, \sigma^2 \mathbf{R}_{xx}^{-1})$ distribution, so that $\hat{\alpha} = \mathbf{T}'\hat{\gamma}$ has a $N_{p-1}(\alpha, \sigma^2 \mathbf{I}_{p-1})$ distribution. If γ has a $N_{p-1}(\mathbf{0}, \sigma_0^2 \mathbf{I}_{p-1})$ prior, so does $\hat{\alpha}$. By the arguments of Example 12.3, the marginal distribution of $\hat{\alpha}$ is $N_{p-1}(\mathbf{0}, (\sigma^2 + \sigma_0^2)\mathbf{I}_{p-1})$.

MISCELLANEOUS EXERCISES 10

1. Let $\tilde{\beta}$ be the new LSE, $\hat{\beta}$ the old LSE, and \mathbf{X} the old regression matrix with i th row \mathbf{x}_i . Also, set $h = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$ and $\tilde{h}_i = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$. The new RSS is $\sum_{i=1}^n(Y_i - \mathbf{x}'_i\tilde{\beta})^2 + (Y - \mathbf{x}'\tilde{\beta})^2$. Arguing as in Theorem 10.1, we get $\tilde{\beta} = \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}e/(1+h)$, so that $Y_i - \mathbf{x}'_i\tilde{\beta} = e_i - \tilde{h}_i e/(1+h)$ and the new RSS is $\sum_{i=1}^n[e_i - \tilde{h}_i e/(1+h)]^2 + e^2/(1+h)$. Multiplying out and using $\sum \tilde{h}_i e_i = 0$ and $\sum \tilde{h}_i^2 = h$ gives the result.

2. (a) Let $m = [\mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}]^{-1}$. Then using A.9.1,

$$\begin{aligned}\mathbf{H} &= (\mathbf{x}^{(j)}, \mathbf{x}^{(j)}) \begin{pmatrix} (\mathbf{x}^{(j)'}\mathbf{x}^{(j)}) & \mathbf{x}^{(j)'}\mathbf{x}^{(j)} \\ -\mathbf{x}^{(j)'}\mathbf{x}^{(j)} & \mathbf{x}^{(j)'}\mathbf{x}^{(j)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}^{(j)'} \\ \mathbf{x}^{(j)'} \end{pmatrix} \\ &= \mathbf{H}_j + m[\mathbf{H}_j \mathbf{x}^{(j)} \mathbf{x}^{(j)'} \mathbf{H}_j - \mathbf{x}^{(j)} \mathbf{x}^{(j)'} \mathbf{H}_j \\ &\quad - \mathbf{H}_j \mathbf{x}^{(j)} \mathbf{x}^{(j)'} + \mathbf{x}^{(j)} \mathbf{x}^{(j)'}] \\ &= \mathbf{H}_j + m[(\mathbf{I}_n - \mathbf{H}_j) \mathbf{x}^{(j)} \mathbf{x}^{(j)'} (\mathbf{I}_n - \mathbf{H}_j)].\end{aligned}$$

- (b) $m^{-1} = \mathbf{x}^{(j)'}(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)} = \|(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}\|^2 = \sum_k \eta_{kj}^2$. Thus $h_i = h_i^{(j)} + m\{[(\mathbf{I}_n - \mathbf{H}_j)\mathbf{x}^{(j)}]_i\}^2 = h_i^{(j)} + \eta_{ij}^2 / \sum_k \eta_{kj}^2$.

- (c) The second term is the leverage of a point in the added variable plot for variable x_j . The equation splits the leverage of a point into two parts; one due to adding variable x_j to a regression and a remainder.

3. (a) (ii) and (iii) imply that $\mathbf{A}\mathbf{A}' = \mathbf{I}_{n-p}$ and $\mathbf{A}\mathbf{X} = \mathbf{0}$, so that the columns of \mathbf{A}' are an orthonormal basis for $\mathcal{C}(\mathbf{X})^\perp$. Since the columns of \mathbf{Q}_2 are also a basis for $\mathcal{C}(\mathbf{X})^\perp$, there is an orthogonal matrix \mathbf{T} such that $\mathbf{A} = \mathbf{T}\mathbf{Q}_2'$.

(b)

$$\begin{aligned}E\|\hat{\epsilon} - \epsilon_1\|^2 &= E\|(\mathbf{T}\mathbf{Q}_2' - \mathbf{J})\epsilon\|^2 \\ &= \sigma^2 \text{tr}[(\mathbf{T}\mathbf{Q}_2' - \mathbf{J})'(\mathbf{T}\mathbf{Q}_2' - \mathbf{J})] \\ &= \sigma^2 \text{tr}[(\mathbf{T}\mathbf{Q}_2' - \mathbf{J})(\mathbf{T}\mathbf{Q}_2' - \mathbf{J}')] \\ &= 2\sigma^2[n - p - \text{tr}(\mathbf{T}\mathbf{Q}_2'\mathbf{J}')].\end{aligned}$$

- (c) $\text{tr}(\mathbf{V} - \mathbf{T}\mathbf{U})\Delta(\mathbf{V} - \mathbf{T}\mathbf{U})' \geq 0$ since Δ is positive definite. Hence $0 \leq \text{tr}(\Delta) - \text{tr}(\mathbf{T}\mathbf{Q}_2'\mathbf{J}')$, with equality when $\mathbf{T} = \mathbf{V}\mathbf{U}'$.

EXERCISES 11a

1. Without pivoting, we get

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1+\epsilon & 2 & 2 \\ 1 & 2 & 1 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \epsilon & 1 & 1 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \epsilon & 1 & 1 \\ 0 & 0 & -1/\epsilon & 2-1/\epsilon \end{pmatrix}.$$

and x_3 is calculated as $(2 - \epsilon^{-1}) / -\epsilon^{-1}$. With pivoting, we get

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1+\epsilon & 2 & 2 \\ 1 & 2 & 1 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \epsilon & 1 & 1 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & \epsilon & 1 & 1 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1-2\epsilon \end{pmatrix}$$

and x_3 is calculated as $1 - 2\epsilon$. This is more accurate if ϵ is small.

- 2.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{x}_1 & \cdots & \bar{x}_{p-1} \\ n\bar{x}_1 & \sum_i x_{i1}^2 & \cdots & \sum_i x_{i1}x_{ip-1} \\ \vdots & \vdots & \ddots & \vdots \\ n\bar{x}_{p-1} & \sum_i x_{ip-1}x_{i1} & \cdots & \sum_i x_{ip-1}^2 \end{pmatrix}.$$

Subtracting \bar{x}_j times row 1 from row $j + 1$ for $j = 1, \dots, p - 1$ gives

$$\begin{pmatrix} n & n\bar{x}_1 & \cdots & \bar{x}_{p-1} \\ 0 & \sum_i x_{i1}^2 - n\bar{x}_1^2 & \cdots & \sum_i x_{i1}x_{ip-1} - n\bar{x}_1\bar{x}_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \sum_i x_{ip-1}x_{i1} - n\bar{x}_{p-1}\bar{x}_1 & \cdots & \sum_i x_{ip-1}^2 - n\bar{x}_{p-1}^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & n\bar{x} \\ 0 & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{pmatrix}$$

using the identity $\sum_i x_{ij}x_{ij'} - n\bar{x}_j\bar{x}_{j'} = \sum_i (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$.

3. The Cholesky factor \mathbf{R}_A of $\mathbf{X}'_A \mathbf{X}_A$ is of the form

$$\begin{pmatrix} \mathbf{R}_0 & \mathbf{z}_0 \\ \mathbf{0}' & d_0 \end{pmatrix},$$

where \mathbf{R}_0 is upper triangular. The relationship $\mathbf{X}'_A \mathbf{X}_A = \mathbf{R}'_A \mathbf{R}_A$ implies that $\mathbf{X}'\mathbf{X} = \mathbf{R}'_0 \mathbf{R}_0$, $\mathbf{R}'_0 \mathbf{z}_0 = \mathbf{X}'\mathbf{Y}$, and $d_0^2 + \mathbf{z}'_0 \mathbf{z}_0 = \mathbf{Y}'\mathbf{Y}$. By the uniqueness of the Cholesky decomposition, we have $\mathbf{R}_0 = \mathbf{R}$, so that $\mathbf{z}_0 = \mathbf{z}$ and $d_0^2 = \mathbf{Y}'\mathbf{Y} - \mathbf{z}'\mathbf{z} = \text{RSS}$.

4.

$$\det(\mathbf{X}'\mathbf{X}) = \det(\mathbf{R}'\mathbf{R}) = \det(\mathbf{R})^2 = \left(\prod_{i=1}^p r_{ii} \right)^2 = \prod_{i=1}^p r_{ii}^2.$$

EXERCISES 11b

1. The proof is by induction. The result is obviously true for $j = 1$. Assume that it is true for j . We have

$$\mathbf{z}_{j+1} = \mathbf{a}^{(j+1)} - \mathbf{q}_1 \mathbf{a}^{(j+1)'} \mathbf{q}_1 - \cdots - \mathbf{q}_j \mathbf{a}^{(j+1)'} \mathbf{q}_j$$

and

$$\begin{aligned} \mathbf{z}'_{j+1} \mathbf{q}_l &= \mathbf{a}^{(j+1)'} \mathbf{q}_l - \mathbf{q}_1' \mathbf{q}_l \mathbf{a}^{(j+1)'} \mathbf{q}_1 - \cdots - \mathbf{q}_l' \mathbf{q}_j \mathbf{a}^{(j+1)'} \mathbf{q}_j \\ &= \mathbf{a}^{(j+1)'} \mathbf{q}_l - \mathbf{a}^{(j+1)'} \mathbf{q}_l \\ &= 0. \end{aligned}$$

Thus $\mathbf{q}_1, \dots, \mathbf{q}_{j+1}$ are orthonormal. These vectors span $C(\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(j+1)})$ since $\mathbf{a}^{(j+1)}$ can be expressed as a linear combination of $\mathbf{q}_1, \dots, \mathbf{q}_{j+1}$, by the expression above and the induction hypothesis.

2. Matrix has a nonzero determinant (an upper triangular matrix has its determinant equal to the product of the diagonal elements).
3. $(\mathbf{X}, \mathbf{Y}) = (\mathbf{WU}, \mathbf{Wu} + \mathbf{w})$ (i.e., $\mathbf{X} = \mathbf{WU}$, $\mathbf{Y} = \mathbf{Wu} + \mathbf{w}$), so $\mathbf{X}'\mathbf{X} = \mathbf{U}'\mathbf{W}'\mathbf{WU}$ and $\mathbf{X}'\mathbf{Y} = \mathbf{U}'\mathbf{W}(\mathbf{Wu} + \mathbf{w}) = \mathbf{U}'\mathbf{WWu}$, since $\mathbf{Ww} = 0$. The normal equations are $\mathbf{U}'\mathbf{W}'\mathbf{WU}\mathbf{b} = \mathbf{U}'\mathbf{W}'\mathbf{Wu}$, which reduce to $\mathbf{Ub} = \mathbf{u}$. Finally, $\mathbf{X}\hat{\beta} = \mathbf{WUU}^{-1}\mathbf{u} = \mathbf{Wu} - \mathbf{Y} = \mathbf{w}$, implying that $\mathbf{e} = \mathbf{w}$.
4. Suppose that after j stages, the $n \times p$ matrix \mathbf{A} with columns $\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(p)}$ has been transformed into the matrix \mathbf{W}_j with columns $\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_j^{(p)}$. In matrix terms (cf. Section 11.3.2) we have $\mathbf{AV}_1 \times \cdots \times \mathbf{V}_j = \mathbf{W}_j$ or $\mathbf{A} \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{0} & \mathbf{I}_{p-j} \end{pmatrix} = \mathbf{W}_j$, where \mathbf{V}_{11} is a $j \times j$ upper triangular matrix with unit diagonals. Thus, the first j columns of \mathbf{W}_j can be written as linear combinations of the first j columns of \mathbf{A} . Conversely, since $\mathbf{A} = \mathbf{W}_j \begin{pmatrix} \mathbf{V}_{11}^{-1} & -\mathbf{W}_{11}^{-1}\mathbf{V}_{12} \\ \mathbf{0} & \mathbf{I}_{p-j} \end{pmatrix}$, the first j columns of \mathbf{A} can be written as linear combinations of the first j columns of \mathbf{W}_j .

To prove the orthogonality part, we use induction. For $j = 1$, $\mathbf{w}_j^{(1)} = \mathbf{a}^{(1)}$ and for $\ell > 1$, $\mathbf{w}_j^{(1)'} \mathbf{w}_j^{(\ell)} = \mathbf{a}^{(1)'} (\mathbf{a}^{(\ell)} + v_{1\ell} \mathbf{a}^{(1)}) = 0$ since $v_{1\ell} = -\mathbf{a}^{(1)'} \mathbf{a}^{(\ell)} / \|\mathbf{a}^{(1)}\|^2$. Thus the result is true for $j = 1$. Now assume that the result is true after stage j . To show that $\mathbf{w}_{j+1}^{(1)}, \dots, \mathbf{w}_{j+1}^{(j+1)}$ are orthogonal, we note that $\mathbf{w}_{j+1}^{(m)} = \mathbf{w}_j^{(m)}$ for $m \leq j+1$, so $\mathbf{w}_{j+1}^{(1)}, \dots, \mathbf{w}_{j+1}^{(j+1)}$ are orthogonal by the induction hypothesis. Now consider $\mathbf{w}_{j+1}^{(m)'} \mathbf{w}_{j+1}^{(\ell)}$ for $m \leq j+1 < \ell$. We have

$$\mathbf{w}_{j+1}^{(m)'} \mathbf{w}_{j+1}^{(\ell)} = \mathbf{w}_j^{(m)'} (\mathbf{w}_j^{(\ell)} + v_{j,\ell} \mathbf{w}_j^{(j+1)}).$$

For $m < j+1$, both inner products in this last expression are zero by the induction hypothesis. For $m = j+1$, the expression is zero since $v_{j,\ell} = -\mathbf{w}_j^{(j+1)'} \mathbf{w}_j^{(\ell)} / \|\mathbf{w}_j^{(j+1)}\|^2$. The proof is complete.

5. $(\mathbf{T}_1 \mathbf{T}_2)' \mathbf{T}_1 \mathbf{T}_2 = \mathbf{T}_2' \mathbf{T}_1' \mathbf{T}_1 \mathbf{T}_2 = \mathbf{T}_2' \mathbf{T}_2 = \mathbf{I}$.

6.

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{Q}_p \mathbf{R} \mathbf{R}^{-1} \mathbf{r}_1 = \mathbf{Q}_p \mathbf{r}_1 = (\mathbf{Q}_p, \mathbf{Q}_{n-p}) \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{0} \end{pmatrix}.$$

EXERCISES 11d

1. If we delete variable i , the new RSS is $\text{RSS}^* = \text{RSS} - e_i^2 / (1 - h_{ii})$. Consider adding case j to the regression. By Miscellaneous Exercises 10, No. 1, the new RSS is $\text{RSS}^* + e_j^*{}^2 / (1 + h_{jj}^*)$, where e_j^* is the residual from the “delete i ” regression, and $h_{jj}^* = \mathbf{x}_j' (\mathbf{X}_J(i)' \mathbf{X}_J(i))^{-1} \mathbf{x}_j$. These are given by $e_j^* = e_j + h_{ij} e_i / (1 - h_{ii})$ and $h_{jj}^* = h_{jj} + h_{ij}^2 / (1 - h_{ii})$, respectively. Thus the RSS from the “add j ” regression is

$$\begin{aligned} & \text{RSS}^* + \left(e_j + \frac{h_{ij} e_i}{1 - h_{ii}} \right)^2 \Big/ \left[(1 + h_{jj}) + \frac{h_{ij}^2}{1 - h_{ii}} \right] \\ &= \text{RSS} - \frac{e_i^2}{1 - h_{ii}} + \frac{[e_j(1 - h_{ii}) + h_{ij}]^2}{(1 - h_{ii})[(1 + h_{jj})(1 - h_{ii}) + h_{ij}^2]} \\ &= \text{RSS} + \frac{e_i^2(1 + h_{jj}) - e_j^2(1 - h_{ii}) + 2e_i e_j h_{ij}}{(1 + h_{jj})(1 - h_{ii}) + h_{ij}^2}. \end{aligned}$$

EXERCISES 12a

- $\|\mu - E[\mathbf{X} \hat{\beta}]\|^2 = \|\mu - \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X} E[\mathbf{Y}]\|^2 = \|(\mathbf{I}_n - \mathbf{P})\mu\|^2 = \mu'(\mathbf{I}_n - \mathbf{P})\mu$.
- (a) $E|\bar{Y} + \tilde{\gamma}_1 x - (\alpha + \gamma_1 x + \gamma_2 z)|^2 = E[(\bar{Y} - \alpha)^2] + E[(\tilde{\gamma}_1 - \gamma_1)^2]x^2 + \gamma_2 z^2 - 2\gamma_2 xz E[\tilde{\gamma}_1 - \gamma_1]$. Now $\tilde{\gamma}_1 = \sum x_i Y_i = \sum x_i (\alpha + \gamma_1 x_i + \gamma_2 z_i +$

$\varepsilon_i) = \gamma_1 + \gamma_2 r + \sum x_i \varepsilon_i$, so $E[(\tilde{\gamma}_1 - \gamma_1)^2] = E[(\gamma_2 r + \sum x_i \varepsilon_i)^2] = \gamma_2^2 r^2 + \sigma^2$ and $E[\tilde{\gamma}_1 - \gamma_1] = \gamma_2 r$. Also, $E[(\bar{Y} - \alpha)^2] = \sigma^2/n$ so combining these shows that the expected model error is $\sigma^2[(1/n) + x^2] + \gamma_2^2(rx - z)^2$.

- (b) Using the results of Section 9.7.1 yields

$$\begin{aligned} & \text{var}[\bar{Y} + \hat{\gamma}_1 x + \hat{\gamma}_2 z] \\ &= \sigma^2/n + x^2 \text{var}[\hat{\gamma}_1] + z^2 \text{var}[\hat{\gamma}_2] + 2xz \text{cov}[\hat{\gamma}_1, \hat{\gamma}_2] \\ &= \sigma^2/n + \frac{x^2 \sigma^2}{1 - r^2} + \frac{z^2 \sigma^2}{1 - r^2} - \frac{2rxz \sigma^2}{1 - r^2} \\ &= \sigma^2 \left(\frac{1}{n} + x^2 \right) + \sigma^2 \frac{(rx - z)^2}{1 - r^2}, \end{aligned}$$

which is also the expected model error in this case. Thus the biased predictor is best if $\gamma_2^2 < \sigma^2/(1 - r^2)$.

EXERCISES 12b

1. From (12.16) we get

$$1 - \bar{R}_p^2 = \frac{\text{RSS}_p}{\text{SSY}} \frac{n}{n-p},$$

so that

$$\begin{aligned} \frac{1 - \bar{R}_p^2}{1 - \bar{R}_{K+1}^2} &= \frac{\text{RSS}_p}{\text{SSY}} \frac{n}{n-p} \cdot \frac{\text{SSY}}{\text{RSS}_{K+1}} \frac{n - K - 1}{n} \\ &= \frac{\text{RSS}_p}{\text{RSS}_{K+1}} \cdot \frac{n - K - 1}{n - p}. \end{aligned}$$

3. For all $x > 0$, $\log x \leq x - 1$, so that

$$f(y) \log \left[\frac{g(y)}{f(y)} \right] \leq g(y) - f(y).$$

Integrating both sides gives the result.

- 4.

$$\begin{aligned} P[\mathcal{M}|\mathbf{Y}] &= P[\mathcal{M}, \mathbf{Y}]/P[\mathbf{Y}] \\ &= \int P[\mathcal{M}, \mathbf{Y}, \theta] d\theta/P[\mathbf{Y}] \\ &\propto \int P[\mathbf{Y}, |\theta, \mathcal{M}] P[\theta|\mathcal{M}] P[\mathcal{M}] d\theta \\ &= \alpha_p \int f_p(\mathbf{Y}, |\theta) \pi_p(\theta) d\theta. \end{aligned}$$

EXERCISES 12c

- For $p = 1$, we have $\text{RSS}_p = \text{SSY}$ and $\text{RSS}_{p+1} = (1 - r^2)\text{SSY}$. Thus $(\text{RSS}_p - \text{RSS}_{p+1})/\text{RSS}_{p+1} = r^2/(1 - r^2)$, so the largest F corresponds to the largest r^2 .
- Consider a centered and scaled model and let \mathbf{x}_* be the new (centered and scaled) column. By Theorem 3.6, we have $\text{RSS}_p - \text{RSS}_{p+1} = [\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_p)\mathbf{x}_*]^2/\mathbf{x}_*(\mathbf{I}_n - \mathbf{P}_p)\mathbf{x}_*$, where \mathbf{P}_p is the projection for the current model. Then the square of the partial correlation is also given by $[\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_p)\mathbf{x}_*]^2/\mathbf{x}_*(\mathbf{I}_n - \mathbf{P}_p)\mathbf{x}_*$.
- (a) The regions for backward elimination are:
 - Model $\{x_1, x_2\}$: $|r_1 - rr_2| \geq c_1(1 - r^2)^{1/2}$ and $|r_2 - rr_1| \geq c_1(1 - r^2)^{1/2}$.
 - Model $\{x_2\}$: $|r_1 - rr_2| < c_1(1 - r^2)^{1/2} \leq |r_2 - rr_1|$ and $|r_2| \geq c_1$.
 - Model $\{x_1\}$: $|r_2 - rr_1| < c_1(1 - r^2)^{1/2} \leq |r_1 - rr_2|$ and $|r_1| \geq c_1$.
 - Model $\{0\}$ otherwise.
- (b) Stepwise regression is the same as forward selection in this case.

EXERCISES 12d

- The joint density of \mathbf{Y} and $\boldsymbol{\mu}$ can be written as [after completing the square; see (12.83)]

$$\frac{1}{(2\pi\tilde{\sigma}^2)^p/2} \exp\left(-\frac{\sum_i y_i^2}{2\tilde{\sigma}^2}\right) \times \frac{1}{(2\pi(1-\omega)\sigma^2)^p/2} \exp\left\{-\frac{\sum_i [\mu_i - (1-\omega)y_i]^2}{2(1-\omega)\sigma^2}\right\},$$

where $\tilde{\sigma}^2 = \sigma_0^2 + \sigma^2$. Integrating out the μ_i 's gives the marginal density of \mathbf{Y} as $N_p(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I}_p)$.

- $\mathbf{c} = (\mathbf{A}'\mathbf{A} + \lambda \mathbf{I}_p)^{-1} \mathbf{A}'\mathbf{d}$, where $\mathbf{A} = \mathbf{X} \text{diag}(\hat{\beta}_0, \dots, \hat{\beta}_{p-1}) = \mathbf{X}\mathbf{B}$, say. Then $\mathbf{c} = (\mathbf{B}'\mathbf{B} + \lambda \mathbf{I}_p)^{-1} \mathbf{B}'\hat{\beta}$ since $\hat{\beta} = \mathbf{X}'\mathbf{Y}$. Thus $c_j = \hat{\beta}_j^2 / (\hat{\beta}_j^2 + \lambda)$.

EXERCISES 12e

1.

$$\begin{aligned} \mathbf{W}_c^{-1} &= \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{W}_{1.2}^{-1} & -\mathbf{W}_{1.2}^{-1}\mathbf{W}_{12}\mathbf{W}_{22}^{-1} \\ -\mathbf{W}_{22}^{-1}\mathbf{W}_{12}\mathbf{W}_{1.2}^{-1} & \mathbf{W}_{22}^{-1} + \mathbf{W}_{22}^{-1}\mathbf{W}_{21}\mathbf{W}_{1.2}^{-1}\mathbf{W}_{12}\mathbf{W}_{22}^{-1} \end{pmatrix}. \end{aligned}$$

Thus

$$\begin{aligned} & (\mathbf{Y}_c - \mathbf{X}_c \mathbf{m})' \mathbf{W}_c^{-1} (\mathbf{Y}_c - \mathbf{X}_c \mathbf{m}) \\ &= (\mathbf{Y} - \mathbf{X}\mathbf{m})' \mathbf{W}_{22}^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{m}) + (\mathbf{Y}_0 - \boldsymbol{\mu}_0)' \mathbf{W}_{1,2}^{-1} (\mathbf{Y}_0 - \boldsymbol{\mu}_0). \end{aligned}$$

2. Assume that $\mathbf{V} = \begin{pmatrix} \tau_0^{-1} & 0 \\ 0 & \tau_1^{-1} \end{pmatrix}$. Then, for this example,

$$\mathbf{V}_* = \begin{pmatrix} (n + \tau_0)^{-1} & 0 \\ 0 & (s_{xx} + \tau_1)^{-1} \end{pmatrix},$$

$\mathbf{m}_* = [\bar{Y}/(1 + \tau_0/n), \hat{\beta}/(1 + \tau_1/s_{xx})]'$, $\mathbf{X}'_0 \mathbf{V}_* \mathbf{X}_0 = 1/(n + \tau_0) + (x - \bar{x})^2/(\tau_1 + s_{xx})$, $\mathbf{X}_0 \mathbf{m}_* = \bar{Y}/(1 + \tau_0/n) + \hat{\beta}(x - \bar{x})/(1 + \tau_1/s_{xx}) = \tilde{\alpha} + \tilde{\beta}(x - \bar{x})$, say. Thus the predictive density is proportional to

$$\left\{ 1 + \frac{[y_0 - \tilde{\alpha} - \tilde{\beta}(x - \bar{x})]^2}{a_0[1 + 1/(n + \tau_0) + (x - \bar{x})^2/(\tau_1 + s_{xx})]} \right\}^{-(n+d)/2}.$$

EXERCISES 12f

1. The MSE of the LSE is σ^2 . The MSE of $\tilde{\beta}$ is

$$\begin{aligned} E[(\tilde{\beta} - \beta)^2] &= \frac{1}{\sigma} \int_{|r|<c} \beta^2 \phi[(r - \beta)/\sigma] dr \\ &\quad + \frac{1}{\sigma} \int_{|r|\geq c} (r - \beta)^2 \phi[(r - \beta)/\sigma] dr \\ &= \sigma^2 + \frac{1}{\sigma} \int_{|r|<c} [\beta^2 - (r - \beta)^2] \phi[(r - \beta)/\sigma] dr. \end{aligned}$$

Put $\beta = \sigma t$, $c = \sigma \epsilon$. Then the MSE is $\sigma^2 + \sigma^2 \int_{|z+t|<\epsilon} [t^2 - z^2] \phi(z) dz$.

2. The conditional log likelihood is

$$\begin{aligned} & -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (y_i - \beta x_i)^2 \\ & - \log \left[\frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathcal{R}} \exp \left[\frac{1}{2\sigma^2} \sum (y_i - \beta x_i)^2 \right] dy \right], \end{aligned}$$

where $\mathcal{R} = \{y : |\sum_i x_i y_i| < c\}$. The integral is $\text{pr}[|\sum_i x_i Y_i| < c] = \frac{1}{\sigma} \int_{|r|<c} \phi[(r - \beta)/\sigma] dr$, which can be evaluated numerically.

3.

$$\begin{aligned}
 \text{Var}[\mathbf{Y}] &= \int (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' f(\mathbf{y}) d\mathbf{y} \\
 &= \int \int [(\mathbf{y} - E[\mathbf{Y}|\mathbf{z}]) + (E[\mathbf{Y}|\mathbf{z}] - \boldsymbol{\mu})] \\
 &\quad [(\mathbf{y} - E[\mathbf{Y}|\mathbf{z}]) + (E[\mathbf{Y}|\mathbf{z}] - \boldsymbol{\mu})]' f(\mathbf{y}, \mathbf{z}) d\mathbf{y} d\mathbf{z} \\
 &= \int \left\{ \int (\mathbf{y} - E[\mathbf{Y}|\mathbf{z}])(\mathbf{y} - E[\mathbf{Y}|\mathbf{z}])' f(\mathbf{y}|\mathbf{z}) d\mathbf{y} \right\} f(\mathbf{z}) d\mathbf{z} \\
 &\quad + \int (E[\mathbf{Y}|\mathbf{z}] - \boldsymbol{\mu})(E[\mathbf{Y}|\mathbf{z}] - \boldsymbol{\mu})' f(\mathbf{z}) d\mathbf{z} \\
 &\quad + \text{zero cross-product term} \\
 &= E(\text{Var}[\mathbf{Y}|\mathbf{Z}]) + \text{a positive-definite integral}.
 \end{aligned}$$

EXERCISES 12h

1. Up to a factor p/σ , the expected model error is

$$\int \int [\beta - f(b)]^2 \phi[(\beta - b)/\sigma] g(\beta) d\beta db.$$

The inner integral is $\int (\beta - f)^2 \phi[(\beta - b)/\sigma] g(\beta) d\beta$, which is $\int \beta^2 \phi g d\beta + 2f \int \beta \phi g d\beta + f^2 \int \phi g d\beta$. This is minimized when $f = \int \beta \phi g d\beta / \int \phi g d\beta$.

3. Put $\mathbf{A}(k) = \mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'$ and $\hat{\sigma}^2 = \text{RSS}/(n-p)$. Then $\text{RSS}(k) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}(k)\|^2 = \text{RSS} + \hat{\beta}'\mathbf{X}'(\mathbf{C} - \mathbf{I}_p)^2\mathbf{X}\hat{\beta}$. When $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, $\text{tr}[\mathbf{A}(k)] = p/(1+k)$ and $\mathbf{X}'(\mathbf{C} - \mathbf{I}_p)^2\mathbf{X} = k^2(1+k)^{-2}\mathbf{I}_p$. Put $B = \|\hat{\beta}\|^2$. The little bootstrap and GCV are

$$\begin{aligned}
 \text{LB} &= \text{RSS}(k) + 2\hat{\sigma}^2 \text{tr}[\mathbf{A}(k)] \\
 &= \text{RSS} + Bk^2/(1+k)^2 + 2\hat{\sigma}^2 p/(1+k)
 \end{aligned}$$

and

$$\begin{aligned}
 \text{GCV} &= n^{-1} \text{RSS}(k)/(1 - n^{-1} \text{tr}[\mathbf{A}(k)])^2 \\
 &= n[\text{RSS} + (1+k)^2 + Bk^2]/(nk + n - p)^2.
 \end{aligned}$$

Differentiating with respect to k , we see that both LB and GCV are minimized subject to $k > 0$ at $\hat{k} = p\hat{\sigma}^2/(B - p\hat{\sigma}^2)$ if $B > p\hat{\sigma}^2$, and at zero otherwise.

MISCELLANEOUS EXERCISES 12

1. Augment the data as described in Section 10.7.3 and do least squares.

2. Using the notation of Example 12.2, the correct model is chosen if $|r_1| < c_1$ and $|r_2| < c_1$. When x and z are orthogonal, r_1 and r_2 are independent $N(0, 1)$, so the probability of correct selection is $[\Phi(c_1) - \Phi(-c_1)]^2$, where Φ is the distribution function of the standard normal.

3.

$$F_p = \frac{(\text{RSS}_p - \text{RSS}_{K+1})/r}{\hat{\sigma}_{K+1}^2},$$

so that

$$\begin{aligned} rF_p &= \text{RSS}_p/\hat{\sigma}_{K+1}^2 - (n - K - 1) \\ &= C_p + n - 2p - (n - K - 1) \\ &= C_p - p + r, \end{aligned}$$

and hence $C_p = r(F_p - 1) + p$.

4. The integral in (12.84) is

$$\begin{aligned} &\int_{-\infty}^{\infty} [h(\tilde{\sigma}z) - w\tilde{\sigma}z]^2 \phi(z) dz \\ &= w^2 \tilde{\sigma}^2 \int_{|z| < \tau \tilde{\sigma}^{-1}} z^2 \phi(z) dz + (1-w)^2 \tilde{\sigma}^2 \int_{|z| \geq \tau \tilde{\sigma}^{-1}} z^2 \phi(z) dz \\ &= [w^2 - (1-w)^2] \tilde{\sigma}^2 \int_{|z| < \tau \tilde{\sigma}^{-1}} z^2 \phi(z) dz + (1-w)^2 \tilde{\sigma}^2, \end{aligned}$$

so that

$$\begin{aligned} E[\text{ME}] &= p \{ w\sigma^2 + (1-w)^2 \tilde{\sigma}^2 - [w^2 - (1-w)^2] \tilde{\sigma}^2 [1 - \Phi_2(\tau/\tilde{\sigma})] \} \\ &= p \{ \sigma^2 + (\sigma^2 - \sigma_0^2)[1 - \Phi_2(\tau/\tilde{\sigma})] \}. \end{aligned}$$

References

- Aitkin, M. (1987). Modeling variance heterogeneity in normal regression using GLIM. *Appl. Stat.*, **36**, 332–339.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Proceedings, 2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.
- Albert, A. (1972). *Regression and the Moore-Penrose Pseudoinverse*. New York: Academic Press.
- Allen, D. M. (1971). Mean squared error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–475.
- Anda, A. A. and Park, H. (1994). Fast plane rotations with dynamic scaling. *SIAM J. Matrix Anal. Appl.*, **15**, 162–174.
- Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.
- Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *J. R. Stat. Soc. B*, **40**, 85–93.
- Atiqullah, M. (1962). The estimation of residual variance in quadratically balanced least squares problems and the robustness of the F-test. *Biometrika*, **49**, 83–91.
- Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.
- Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford: Clarendon Press.
- Atkinson, A. C. (1986). Masking unmasked. *Biometrika*, **73**, 533–541.
- Atkinson, A. C. and Weisberg, S. (1991). Simulated annealing for the detection of multiple outliers using least squares and least median of squares fitting. In W.

- Stahel and S. Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics*. New York: Springer-Verlag, pp. 7–20.
- Azzalini, A. (1996). *Statistical Inference Based on the Likelihood*. New York: Chapman & Hall.
- Barrodale, I. and Roberts, F. D. K. (1974). Algorithm 478: Solution of an over-determined system of equations in the L_1 norm. *Commun. ACM*, **14**, 319–320.
- Bartels, R. H., Conn, A. R. and Sinclair, J. W. (1978). Minimization techniques for piecewise differentiable functions: The L_1 solution to an overdetermined system. *SIAM J. Numer. Anal.*, **15**, 224–241.
- Bekker, R. A., Cleveland, W. S. and Weil, G. (1988). The use of brushing and rotation for data analysis. In *Dynamic Graphics for Statistics*. Pacific Grove, CA: Wadsworth.
- Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *Am. Stat.*, **38**, 73–77.
- Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward “stepwise” regression. *J. Am. Stat. Assoc.*, **72**, 46–53.
- Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.*, **91**, 109–122.
- Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1–6.
- Berk, K. N. and Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics*, **37**, 385–398.
- Bickel, P. J. and Doksum, K. A. (1981). The analysis of transformations revisited. *J. Am. Stat. Assoc.*, **76**, 296–311.
- Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*. New York: Wiley.
- Björck, A. (1996). *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM.
- Björck, A. and Paige, C. C. (1992). Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, **13**, 176–190.
- Björck, A. and Paige, C. C. (1994). Solution of augmented linear systems using orthogonal factorizations. *BIT*, **34**, 1–26.
- Bloomfield, P. and Steiger, W. L. (1980). Least absolute deviations curve-fitting. *SIAM J. Sci. Stat. Comput.*, **1**, 290–301.
- Bloomfield, P. and Steiger, W. L. (1983). *Least Absolute Deviations: Theory, Applications and Algorithms*. Boston: Birkhäuser.
- Bohrer, R. (1973). An optimality property of Scheffé bounds. *Ann. Stat.*, **1**, 766–772.
- Bohrer, R. and Francis, G. K. (1972). Sharp one-sided confidence bounds for linear regression over intervals. *Biometrika*, **59**, 99–107.
- Bowden, D. C. (1970). Simultaneous confidence bands for linear regression models. *J. Am. Stat. Assoc.*, **65**, 413–421.

- Bowden, D. C. and Graybill, F. A. (1966). Confidence bands of uniform and proportional width for linear models. *J. Am. Stat. Assoc.*, **61**, 182–198.
- Box, G. E. P. (1966). Use and abuse of regression. *Technometrics*, **8**, 625–629.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. B*, **26**, 211–252.
- Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Am. Stat. Assoc.*, **77**, 209–210.
- Box, G. E. P. and Taio, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.
- Box, G. E. P. and Watson, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika*, **49**, 93–106.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Am. Stat. Assoc.*, **87**, 738–754.
- Breiman, L. (1995). Better subset selection using the Nonnegative Garrote. *Technometrics*, **37**, 373–384.
- Breiman, L. (1996a). Stacked regressions. *Machine Learning*, **24**, 49–64.
- Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Stat.*, **24**, 2350–2383.
- Broersen, P. M. T. (1986). Subset regression with stepwise directed search. *Appl. Stat.*, **35**, 168–177.
- Brown, P. J. (1977). Centering and scaling in ridge regression. *Technometrics*, **19**, 35–36.
- Brown, P. J. (1993). *Measurement, Regression and Calibration*. Oxford: Clarendon Press.
- Brown, M. B. and Forsythe, A. B. (1974). Robust tests for equality of variance. *J. Am. Stat. Assoc.*, **69**, 364–367.
- Brunk, H. D. (1965). *An Introduction to Mathematical Statistics*, 2nd ed. Waltham, MA: Blaisdell.
- Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, **53**, 603–618.
- Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.
- Canner, P. L. (1969). Some curious results using minimum variance linear unbiased estimators. *Am. Stat.*, **23** (5), 39–40.
- Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. R. Stat. Soc. B*, **57**, 473–484.
- Carlstein, E. (1986). Simultaneous confidence intervals for predictions. *Am. Stat.*, **40**, 277–279.
- Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. R. Stat. Soc. B*, **42**, 71–78.
- Carroll, R. J. (1982a). Adapting for heteroscedasticity in linear models. *Ann. Stat.*, **10**, 1224–1233.
- Carroll, R. J. (1982b). Two examples of transformations where there are possible outliers. *Appl. Stat.*, **31**, 149–152.

- Carroll, R. J. and Cline, D. B. H. (1988). An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika*, **75**, 35–43.
- Carroll, R. J. and Davidian, M. (1987). Variance function estimation. *J. Am. Stat. Assoc.*, **82**, 1079–1091.
- Carroll, R. J. and Ruppert, D. (1981). On prediction and the power transformation family. *Biometrika*, **68**, 609–615.
- Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Stat.*, **10**, 429–441.
- Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *J. Am. Stat. Assoc.*, **79**, 321–328.
- Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis. *Technometrics*, **27**, 1–12.
- Carroll, R. J. and Ruppert, D. (1988). *Transformations and Weighting in Regression*. New York: Chapman & Hall.
- Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Boston: Duxbury Press.
- Chan, T. F., Golub, G. H. and LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *Am. Stat.*, **37**, 242–247.
- Chang, W. H., McKean, J. W., Naranjo, J. D. and Sheather, S. J. (1999). High-breakdown rank regression. *J. Am. Stat. Assoc.*, **94**, 205–219.
- Chatfield, C. (1998). Durbin-Watson test. In P. Armitage and T. Colton, (Eds.), *Encyclopedia of Biostatistics*, Vol. 2. Wiley: New York, pp. 1252–1253.
- Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. New York: Wiley.
- Cheney, E. W. (1966). *Introduction to Approximation Theory*. New York: McGraw-Hill.
- Cleenshaw, C. W. (1955). A note on the summation of Chebyshev series. *Math. Tables Aids Comput.*, **9**, 118.
- Cleenshaw, C. W. (1960). Curve fitting with a digital computer. *Comput. J.*, **2**, 170.
- Cleenshaw, C. W. and Hayes, J. G. (1965). Curve and surface fitting. *J. Inst. Math. Appl.*, **1**, 164–183.
- Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.
- Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.
- Coakley, C. and Hettmansperger, T. P. (1993). A bounded-influence, high breakdown, efficient regression estimator. *J. Am. Stat. Assoc.*, **88**, 872–880.
- Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *J. R. Stat. Soc. Suppl.*, **5**, 171–176.
- Conover, W. J., Johnson, M. E. and 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. *J. Am. Stat. Assoc.*, **89**, 177–189.
- Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
- Cook, R. D. and Wang, P. C. (1983). Transformation and influential cases in regression. *Technometrics*, **25**, 337–343.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall.
- Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1–10.
- Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.
- Cook, R. D. and Weisberg, S. (1999). *Applied Regression including Computing and Graphics*. New York: Wiley.
- Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *J. Am. Stat. Assoc.*, **87**, 419–424.
- Cooper, B. E. (1968). The use of orthogonal polynomials: Algorithm AS 10. *Appl. Stat.*, **17**, 283–287.
- Cooper, B. E. (1971a). The use of orthogonal polynomials with equal x -values: Algorithm AS 42. *Appl. Stat.*, **20**, 208–213.
- Cooper, B. E. (1971b). A remark on algorithm AS 10. *Appl. Stat.*, **20**, 216.
- Cox, C. P. (1971). Interval estimating for X -predictions from linear Y -on- X regression lines through the origin. *J. Am. Stat. Assoc.*, **66**, 749–751.
- Cox, D. R. and Hinkley, D. V. (1968). A note on the efficiency of least squares estimates. *J. R. Stat. Soc. B*, **30**, 284–289.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of cross-validation. *Numer. Math.*, **31**, 377–403.
- Croux, C., Rousseeuw, P. J. and Hössjer, O. (1994). Generalized S-estimators. *J. Am. Stat. Assoc.*, **89**, 1271–1281.
- David, H. A. (1981). *Order Statistics*, 2nd ed. New York: Wiley.
- Davies, R. B. and Hutton, B. (1975). The effects of errors in the independent variables in linear regression. *Biometrika*, **62**, 383–391. Correction, **64**, 655.
- Davis, P. (1975). *Interpolation and Approximation*. New York: Dover.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- De Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer-Verlag.
- Dempster, A. P. and Gasko-Green, M. (1981). New tools for residual analysis. *Ann. Stat.*, **9**, 945–959.
- Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Am. Stat. Assoc.*, **72**, 77–106.
- Diehr, G. and Hoflin, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, **16**, 317–320.

- Diercx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford: Clarendon Press.
- Dodge, Y. (Ed.) (1987). *Statistical Data Analysis Based on the L₁ Norm and Related Methods*. Amsterdam: North-Holland.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. R. Stat. Soc. B*, **57**, 45–98.
- Draper, N. R. and Cox, D. R. (1969). On distributions and their transformation to normality. *J. R. Stat. Soc. B*, **31**, 472–476.
- Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed. New York: Wiley.
- Draper, N. R. and Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: Review and comments. *Technometrics*, **21**, 451–466.
- Draper, N. R., Guttman, I., and Kanemasu, H. (1971). The distribution of certain regression statistics. *Biometrika*, **58**, 295–298.
- Dunn, O. J. (1959). Confidence intervals for the means of dependent, normally distributed variables. *J. Am. Stat. Assoc.*, **54**, 613–621.
- Dunn, O. J. (1961). Multiple comparisons among means. *J. Am. Stat. Assoc.*, **56**, 52–64.
- Dunn, O. J. (1968). A note on confidence bands for a regression line over finite range. *J. Am. Stat. Assoc.*, **63**, 1029–1033.
- Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, **37**, 409–428.
- Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, **38**, 159–178.
- Durbin, J. and Watson, G. S. (1971). Testing for serial correlation in least squares regression. III. *Biometrika*, **58**, 1–19.
- Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Stat. Assoc.*, **68**, 117–130.
- Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston and H. S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, 1, 191–203.
- Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Ann. Math. Stat.*, **34**, 447–456.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Sci.*, **11**, 89–121.
- Eubank, R. L. (1984). Approximate regression models and splines. *Commun. Stat. A*, **13**, 485–511.
- Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed. New York: Marcel Dekker.
- Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Stat. Sci.*, **10**, 254–272.
- Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables. *J. Am. Stat. Assoc.*, **19**, 431–453.
- Ezekiel, M. and Fox, K. A. (1959). *Methods of Correlation and Regression Analysis*, 3rd ed. New York: Wiley.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

- Farebrother, R. W. (1990). Algorithm AS 256: The distribution of a quadratic form in normal variables. *Appl. Stat.*, **23**, 470–476.
- Farley, J. U. and Hinich, M. J. (1970). A test for shifting slope coefficient in a linear model. *J. Am. Stat. Assoc.*, **65**, 1320–1329.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: Wiley.
- Fieller, E. C. (1940). The biological standardization of insulin. *J. R. Stat. Soc. Suppl.*, **7**, 1–64.
- Fisher, R. A. and Yates, F. (1957). *Statistical Tables for Biological, Agricultural, and Medical Research*, 5th ed. London: Oliver and Boyd.
- Fletcher, R. (1987). *Practical Methods of Optimization*, 2nd ed. New York: Wiley.
- Forsythe, G. E. (1957). Generation and use of orthogonal polynomials for data-fitting with a digital computer. *J. Soc. Ind. Appl. Math.*, **5**, 74–87.
- Frank, I. E. and Friedman, J. H. (1993). A comparison of some chemometrics regression tools. *Technometrics*, **35**, 109–148.
- Freedman, D. A. (1983). A note on screening regression equations. *Am. Stat.*, **37**, 152–155.
- Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.
- Fuller, W. A. and Rao, J. N. K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Stat.*, **6**, 1149–1158.
- Furnival, G. M. (1971). All possible regressions with less computation. *Technometrics*, **13**, 403–408.
- Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, **16**, 499–511.
- Gafarian A. V. (1964). Confidence bands in straight line regression. *J. Am. Stat. Assoc.*, **59**, 182–213.
- Garside, M. J. (1965). The best subset in multiple regression analysis. *Appl. Stat.*, **14**, 196–200.
- Garthwaite, P. H. and Dickey, J. M. (1992). Elicitation of prior distributions for variable selection problems in regression. *Ann. Stat.*, **20**, 1697–1719.
- Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. B*, **56**, 501–514.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- Gentleman, W. M. (1973). Least squares computations by Givens transformations without square roots. *J. Inst. Math. Appl.*, **10**, 195–197.
- George, E. I. (2000). The variable selection problem. *J. Am. Stat. Assoc.*, **95**, 1304–1308.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.
- Ghosh, M. N. and Sharma, D. (1963). Power of Tukey's tests for non-additivity. *J. R. Stat. Soc. B*, **25**, 213–219.
- Glaser, R. E. (1983). Levene's robust test of homogeneity of variances. In N. L. Johnson and C. B. Read (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 4. New York: Wiley, pp. 608–610.

- Golub, G. H. and Styan, G. P. H. (1974). Some aspects of numerical computations for linear models. In *Proceedings, 7th Annual Symposium on the Interface*, Iowa State University, Ames, IA, pp. 189-192.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, 3rd. ed. Baltimore: Johns Hopkins University Press.
- Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215-223.
- Good, I. J. (1969). Conditions for a quadratic form to have a chi-squared distribution. *Biometrika*, **56**, 215-216.
- Good, I. J. (1970). Correction to "Conditions for a quadratic form to have a chi-squared distribution." *Biometrika*, **57**, 225.
- Goodnight, J. (1979). A tutorial on the SWEEP operator. *Am. Stat.*, **33**, 149-158.
- Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*. New York: McGraw-Hill.
- Graybill, F. A. and Bowden D. C. (1967). Linear segment confidence bands for simple linear models. *J. Am. Stat. Assoc.*, **62**, 403-408.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.
- Grier, D. A. (1992). An extended sweep operator for the cross-validation of variable selection in linear regression. *J. Stat. Comput. Simul.*, **43**, 117-126.
- Grossman, S. I. and Styan, G. P. H. (1972). Optimal properties of Theil's BLUS residuals. *J. Am. Stat. Assoc.*, **67**, 672-673.
- Gujarati, D. (1970). Use of dummy variables in testing for equality between sets of coefficients in linear regressions: A generalization. *Am. Stat.*, **24**, 18-22.
- Gunst, R. F. and Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error. *J. Am. Stat. Assoc.*, **72**, 616-628.
- Gunst, R. F. and Mason, R. L. (1985). Outlier-induced collinearities. *Technometrics*, **27**, 401-407.
- Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *Am. Stat.*, **52**, 15-19.
- Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc.*, **88**, 1264-1272.
- Hahn, G. J. (1972). Simultaneous prediction intervals for a regression model. *Technometrics*, **14**, 203-214.
- Hahn, G. J. and Hendrickson, R. W. (1971). A table of percentage points of the distribution of the largest absolute value of k Student t variates and its applications. *Biometrika*, **58**, 323-332.
- Halperin, M. and Gurian, J. (1968). Confidence bands in linear regression with constraints on the independent variables. *J. Am. Stat. Assoc.*, **63**, 1020-1027.
- Halperin, M. and Gurian, J. (1971). A note on estimation in straight line regression when both variables are subject to error. *J. Am. Stat. Assoc.*, **66**, 587-589.
- Halperin, M., Rastogi, S. C., Ho, I. and Yang, Y. Y. (1967). Shorter confidence bands in linear regression. *J. Am. Stat. Assoc.*, **62**, 1050-1067.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

- Han, C. P. (1969). Testing the homogeneity of variances in a two-way classification. *Biometrics*, **25**, 153–158.
- Handschin, E., Kohlas, J., Fiechter, A. and Scheweppe, F. (1975). Bad data analysis for power system state estimation. *IEEE Trans. Power Apparatus Syst.*, **2**, 329–337.
- Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. B*, **41**, 190–195.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York: Springer-Verlag.
- Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry. *Stat. Sci.*, **8**, 120–143.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.
- Hawkins, D. M. (1993a). The accuracy of elemental set approximations for regression. *J. Am. Stat. Assoc.*, **88**, 580–589.
- Hawkins, D. M. (1993b). The feasible solution algorithm for least median of squares regression. *Comput. Stat. Data Anal.*, **16**, 81–101.
- Hawkins, D. M. (1994a). The feasible solution algorithm for least trimmed squares regression. *Comput. Stat. Data Anal.*, **17**, 185–196.
- Hawkins, D. M. (1994b). The feasible solution algorithm for the minimum covariance determinant estimate in multivariate data. *Comput. Stat. Data Anal.*, **17**, 197–210.
- Hawkins, D. M. and Olive, D. (1999). Improved feasible solution algorithms for high-breakdown estimation. *Comput. Stat. Data Anal.*, **30**, 1–11.
- Hawkins, D. M. and Olive, D. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *J. Am. Stat. Assoc.*, **97**, 136–159.
- Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, **26**, 197–208.
- Hayes, D. G. (1969). A method of storing the orthogonal polynomials used for curve and surface fitting. *Comput. J.*, **12**, 148–150.
- Hayes, J. G. (1970). Curve fitting by polynomials in one variable. In J. G. Hayes (Ed.), *Numerical Approximation to Functions and Data*. London: Athlone Press, pp. 43–64.
- Hayes, J. G. (1974). Numerical methods for curve and surface fitting. *J. Inst. Math. Appl.*, **10**, 144–152.
- Hayter, A. J. (1984). A proof of the conjecture that the Tukey–Kramer multiple comparison procedure is conservative. *Ann. Stat.*, **12**, 61–75.
- Hernandez, F. and Johnson, R. A. (1980). The large-sample behavior of transformations to normality. *J. Am. Stat. Assoc.*, **75**, 855–861.
- Higham, N. I. (1996). *Accuracy and Stability of Numerical Algorithms*. Philadelphia: SIAM.
- Hill, R. W. (1977). *Robust regression where there are outliers in the carriers*. Ph.D. Dissertation, Harvard University, Department of Statistics.

- Hinkley, D. V. (1969a). On the ratio of two correlated normal random variables. *Biometrika*, **56**, 635–639.
- Hinkley, D. V. (1969b). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495–504.
- Hinkley, D. V. (1971). Inference in two-phase regression. *J. Am. Stat. Assoc.*, **66**, 736–743.
- Hinkley, D. V. and Rungger, G. (1984). Analysis of transformed data. *J. Am. Stat. Assoc.*, **79**, 302–308.
- Hoadley, B. (1970). A Bayesian look at inverse linear regression. *J. Am. Stat. Assoc.*, **65**, 356–369.
- Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
- Hocking, R. R. (1996). *Methods and Applications of Linear Models: Regression and Analysis of Variance*. New York: Wiley.
- Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, **9**, 531–540.
- Hocking, R. R. and Pendleton, O. J. (1983). The regression dilemma. *Comm. Statist. A*, **12**, 497–527.
- Hodges, S. D. and Moore, P. G. (1972). Data uncertainties and least squares regression. *Appl. Stat.*, **21**, 185–195.
- Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.
- Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Applications to non-orthogonal problems. *Technometrics*, **12**, 69–82.
- Hoerl, A. E., Kennard, R. W. and Baldwin K. E. (1975). Ridge regression: Some simulations. *Commun. Stat. A*, **4**, 105–124.
- Hoerl, R. W., Schuenemeyer, J. H. and Hoerl, A. E. (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics*, **28**, 369–380.
- Hogg, R. V. and Craig, A. T. (1958). On the decomposition of certain chi-square variables. *Ann. Math. Stat.*, **29**, 608–610.
- Hogg, R. V. and Craig, A. T. (1970). *Introduction to Mathematical Statistics*, 3rd ed. New York: Macmillan.
- Hössjer, O. (1994). Rank-based estimates in the linear model with high-breakdown point. *J. Am. Stat. Assoc.*, **89**, 149–158.
- Householder, A. S. (1958). A class of methods for inverting matrices. *J. Soc. Ind. Appl. Math.*, **6**, 189–195.
- Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. London: Chapman & Hall.
- Hsu, P. L. (1938). On the best unbiased quadratic estimate of the variance. *Stat. Res. Mem.*, **2**, 91–104.
- Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.
- Hubert, M. and Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors. *J. Stat. Plann. Inference*, **57**, 153–163.
- Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Am. Stat. Assoc.*, **61**, 1097–1129.

- Hudson, D. J. (1969). Least squares fitting of a polynomial constrained to be either non-negative, non-decreasing or convex. *J. R. Stat. Soc. B*, **31**, 113–118.
- Hunt, D. N. and Triggs, C. M. (1989). Iterative missing value estimation. *Appl. Stat.*, **38**, 293–300.
- Hurvich, C. M. and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *Am. Stat.*, **44**, 214–217.
- Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, **78**, 499–509.
- Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Stat.*, **43**, 1449–1458.
- James, A. T. and Wilkinson, G. N. (1971). Factorization of the residual operator and canonical decomposition of non-orthogonal factors in analysis of variance. *Biometrika*, **58**, 279–294.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Stat. Probab.*, **1**, 361–379.
- Jarrett, R. G. (1978). The analysis of designed experiments with missing observations. *Appl. Stat.*, **27**, 38–46.
- Jennrich, R. I. and Sampson, P. I. (1971). A remark on algorithm AS 10. *Appl. Stat.*, **20**, 117–118.
- John, J. A. and Draper, N. R. (1980). An alternative family of transformations. *Appl. Stat.*, **29**, 190–197.
- Johnson, D. E. and Graybill, F. A. (1972a). Estimation of σ^2 in a two-way classification model with interaction. *J. Am. Stat. Assoc.*, **67**, 388–394.
- Johnson, D. E. and Graybill, F. A. (1972b). An analysis of a two-way model with interaction and no replication. *J. Am. Stat. Assoc.*, **67**, 862–868.
- Joiner, B. L. (1981). Lurking variables: Some examples. *Am. Stat.*, **35**, 227–233.
- Jones, G. and Rocke, D. M. (1999). Bootstrapping in controlled calibration experiments. *Technometrics*, **41**, 224–233.
- Joshi, S. W. (1970). Construction of certain bivariate distributions. *Am. Stat.*, **24** (2), 32.
- Jureckova, J. (1971). Non-parametric estimate of regression coefficients. *Ann. Math. Stat.*, **42**, 1328–1338.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
- Kennedy, W. J. and Bancroft, T. A. (1971). Model building for prediction in regression based upon repeated significance tests. *Ann. Math. Stat.*, **42**, 1273–1284.
- Khatri, C. G. (1978). A remark on the necessary and sufficient conditions for a quadratic form to be distributed as chi-squared. *Biometrika*, **65**, 239–240.
- Koerts, J. and Abrahamse, A. P. J. (1969). *On the Theory and Application of the General Linear Model*. Rotterdam: Rotterdam University Press.
- Kowalski, C. (1970). The performance of some rough tests for bivariate normality before and after coordinate transformations to normality. *Technometrics*, **12**, 517–544.
- Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *J. Am. Stat. Assoc.*, **77**, 595–604.

- Kruskal, W. (1975). The geometry of generalized inverses. *J. R. Stat. Soc. B*, **37**, 272–283.
- Krutchoff, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics*, **9**, 425–439.
- Krutchoff, R. G. (1969). Classical and inverse regression methods of calibration in extrapolation. *Technometrics*, **11**, 605–608.
- Kupper, L. L. (1972). Letter to the editor. *Am. Stat.*, **26** (1), 52.
- LaMotte, L. R. and Hocking, R. R. (1970). Computational efficiency in the selection of regression variables. *Technometrics*, **12**, 83–93.
- Lane, T. P. and DuMouchel, W. H. (1994). Simultaneous confidence intervals in multiple regression. *Am. Stat.*, **48**, 315–321.
- Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781–790.
- Laud, P. W. and Ibrahim, J. G. (1996). Predictive specification of prior model probabilities in variable selection. *Biometrika*, **83**, 267–274.
- Lawson, C. L. and Hanson, R. J. (1995). *Solving Least Squares Problems*. Philadelphia: SIAM.
- Levene, H. (1960). Robust tests for the equality of variance. In I. Olkin (Ed.), *Contributions to Probability and Statistics*. Palo Alto, CA: Stanford University Press, pp. 278–292.
- Li, K.-C. (1987). Asymptotic optimality for C_p , CL, cross-validation and generalized cross-validation: Discrete index set. *Ann. Stat.*, **15**, 958–975.
- Lieberman, G. J., Miller, R. G., Jr. and Hamilton M. A. (1967). Unlimited simultaneous discrimination intervals in regression. *Biometrika*, **54**, 133–145.
- Limam, M. M. T. and Thomas, D. R. (1988). Simultaneous tolerance intervals for the linear regression model. *J. Am. Stat. Assoc.*, **83**, 801–804.
- Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York: Wiley.
- Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of use. *J. Am. Stat. Assoc.*, **62**, 819–841.
- Lyon, J. D. and Tsai, C.-L. (1996). A comparison of tests for heteroscedasticity. *Statistician*, **45**, 337–349.
- Malinvaud, E. (1970). *Statistical Methods of Econometrics*, 2nd ed. (translated by A. Silvey). Amsterdam: North Holland.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, **15**, 661–675.
- Mallows, C. L. (1975). Some topics in robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, N.J.
- Mallows, C. L. (1986). Augmented partial residual plots. *Technometrics*, **28**, 313–320.
- Mansfield, E. R. and Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *Am. Stat.*, **41**, 107–116.
- Markowski, C. A. and Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *Am. Stat.*, **44**, 322–326.
- Mayo, M. S. and Gray, J. B. (1997). Elemental subsets: The building blocks of regression. *Am. Stat.*, **51**, 122–129.
- McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *J. Am. Stat. Assoc.*, **62**, 1302–1304.

- McKean, J. W., Sheather, S. J. and Hettmansperger, T. P. (1993). The use and interpretation of residuals based on robust estimation. *J. Am. Stat. Assoc.*, **88**, 1254–1263.
- Mee, R. W. and Eberhardt, K. R. (1996). A comparison of uncertainty criteria for calibration. *Technometrics*, **38**, 221–229.
- Miller, A. (1990). *Subset Selection in Regression*. New York: Chapman & Hall.
- Miller, R. G. (1977). Developments in multiple comparisons, 1966–1976. *J. Am. Stat. Assoc.*, **72**, 779–788.
- Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd ed. New York: McGraw-Hill.
- Mitchell, J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023–1036.
- Moran, P. A. P. (1970). Fitting a straight line when both variables are subject to error. In R. S. Anderssen and M. R. Osborne (Eds.), *Data Presentation*. St Lucia, QLD, Australia: University of Queensland Press, pp. 25–28.
- Moran, P. A. P. (1971). Estimating structural and functional relationships. *J. Mult. Anal.*, **1**, 232–255.
- Morgan, J. A. and Tatar, J. F. (1972). Calculation of the residual sum of squares for all possible regressions. *Technometrics*, **14**, 317–325.
- Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: Wiley.
- Naranjo, J. D. and Hettmansperger, T. P. (1994). Bounded influence rank regression. *J. R. Stat. Soc. B*, **56**, 209–220.
- Nelder, J. A. (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. R. Soc. A*, **283**, 147–162.
- Nelder, J. A. (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. R. Soc. A*, **283**, 163–178.
- Nelder, J. A. (1994). The statistics of linear models: Back to basics. *Stat. Comput.*, **4**, 221–234.
- Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Comput. J.*, **7**, 308–313.
- Nishi, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Stat.*, **12**, 758–765.
- O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics*, Vol. 2B, *Bayesian Inference*. London: Arnold.
- O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Stat. Soc. B*, **57**, 99–138.
- Olejnik, S. F. and Algina, J. (1987). Type I error rates and power estimates of selected parametric and non-parametric tests of scale. *J. Ed. Stat.*, **12**, 45–61.
- Olshen, R. A. (1973). The conditional level of the *F*-test. *J. Am. Stat. Assoc.*, **68**, 692–698.
- Osborne, C. (1991). Statistical calibration: A review. *Int. Stat. Rev.*, **59**, 309–336.
- Parlett, B. N. (1980). *The Symmetric Eigenvalue Problem*. Englewood Cliffs, NJ: Prentice Hall.

- Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
- Payne, J. A. (1970). An automatic curve-fitting package. In J. G. Hayes (Ed.), *Numerical Approximation of Functions and Data*. London: Athlone Press, pp. 98–106.
- Pearce, S. C., Caliński, T. and Marshall, T. F. de C. (1974). The basic contrasts of an experimental design with special reference to the analysis of data. *Biometrika*, **61**, 449–460.
- Pearson, E. S. and Hartley, H. O. (1970). *Biometrika Tables for Statisticians*, 3rd ed. Cambridge: Cambridge University Press.
- Peters, G. and Wilkinson, J. H. (1970). The least squares problem and pseudoinverses, *Comput. J.*, **13**, 309–316.
- Pierce, D. A. and Dykstra, R. L. (1969). Independence and the normal distribution. *Am. Stat.*, **23** (4), 39.
- Pope, P. T. and Webster, J. T. (1972). The use of an F -statistic in stepwise regression procedures. *Technometrics*, **14**, 327–340.
- Portnoy, S. (1987). Using regression fractiles to identify outliers. In Y. Dodge (Ed.), *Statistical Data Analysis Based on the L_1 Norm and Related Methods*. Amsterdam: North Holland, pp. 345–356.
- Prentice, R. L. (1974). Degrees-of-freedom modifications for F tests based on non-normal errors. *Biometrika*, **61**, 559–563.
- Pringle, R. M. and Rayner, A. A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. London: Griffin.
- Quenouille, M. H. (1950). An application of least squares to family diet surveys. *Econometrica*, **18**, 27–44.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.*, **92**, 179–191.
- Rahman, N. A. (1967). *Exercises in Probability and Statistics*. London: Griffin.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *J. R. Stat. Soc. B*, **31**, 350–371.
- Rao, C. R. (1952). Some theorems on minimum variance estimation. *Sankhyā*, **12**, 27–42.
- Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *J. Am. Stat. Assoc.*, **65**, 161–172.
- Rao, C. R. (1972). Estimation of variance and co-variance components in linear models. *J. Am. Stat. Assoc.*, **67**, 112–115.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.
- Rao, C. R. (1974). Projectors, generalized inverses and the BLUE's. *J. R. Stat. Soc. B*, **36**, 442–448.
- Rao, C. R. and Mitra, S. K. (1971a). *Generalized Inverse of Matrices and its Applications*. New York: Wiley.
- Rao, C. R. and Mitra, S. K. (1971b). Further contribution to the theory of generalized inverse of matrices and its applications. *Sankhyā Ser. A*, **33**, 289–300.
- Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*. New York: Wiley.

- Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.
- Richardson, D. H. and Wu, D.-M. (1970). Alternative estimators in the error in variables model. *J. Am. Stat. Assoc.*, **65**, 724–748.
- Rogers, C. E. and Wilkinson, G. N. (1974). Regression, curve fitting and smoothing numerical problems in recursive analysis of variance algorithms. *J. Inst. Math. Appl.*, **10**, 141–143.
- Ronchetti, E. (1987). Bounded influence inference in regression: A review. In Y. Dodge (Ed.), *Statistical Data Analysis Based on the L₁ Norm and Related Methods*, Amsterdam: North Holland, pp. 65–80.
- Roth, A. J. (1988). Welch tests. In *Encyclopedia of Statistical Sciences*, Vol. 9, N. L. Johnson and C. B. Read (Eds.). New York: Wiley, pp. 608–610.
- Rousseeuw, P. J. (1984). Least median of squares regression. *J. Am. Stat. Assoc.*, **79**, 871–880.
- Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- Rousseeuw, P. J. and van Driesen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
- Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.*, **85**, 633–639.
- Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle and D. Martin (Eds.), *Lecture Notes in Statistics*, Vol. 26. New York: Springer-Verlag, pp. 256–272.
- Ruppert, D. (1992). Computing S-estimators for regression and multivariate location/dispersion. *J. Comput. Graph. Stat.*, **1**, 253–270.
- Savage, I. R. and Lukacs, E. (1954). Tables of inverses of finite segments of the Hilbert matrix. In O. Taussky (Ed.). *Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues*. National Bureau of Standards Applied Mathematics Series 39. Washington, DC: U.S. Government Printing Office, pp. 105–108.
- Saw, J. G. (1966). A conservative test for the concurrence of several regression lines and related problems. *Biometrika*, **53**, 272–275.
- Schatzoff, M., Tsao, R. and Feinberg, S. (1968). Efficient calculation of all possible regressions. *Technometrics*, **10**, 769–779.
- Scheffé, H. (1953). A method of judging all contrasts in the analysis of variance. *Ann. Math. Stat.*, **24**, 87–104.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *J. R. Stat. Soc. B*, **33**, 307–311.
- Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Q. J. Appl. Math.*, **4**, 45–99; 112–141.
- Schimek, M. (Ed.). (2000). *Smoothing and Regression: Approaches, Computation and Application*, New York: Wiley.
- Schumaker, L. L. (1981). *Spline Functions*. New York: Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

- Scott, A. and Smith, T. M. F. (1970). A note on Moran's approximation to Student's t . *Biometrika*, **57**, 681–682.
- Scott, A. J. and Smith, T. M. F. (1971). Interval estimates for linear combinations of means. *Appl. Stat.*, **20**, 276–285.
- Searle, S. R. (1971). *Linear Models*. New York: Wiley.
- Seber, G. A. F. (1980). *The Linear Hypothesis: A General Theory*, 2nd ed. London: Griffin.
- Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, 2nd ed. London: Griffin.
- Seber, G. A. F. (1984). *Multivariate Observations*. New York: Wiley.
- Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.
- Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Stat. Assoc.*, **88**, 486–494.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.
- Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43–49.
- Sidak, Z. (1968). On multivariate normal probabilities of rectangles. *Ann. Math. Stat.*, **39**, 1425–1434.
- Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Commun. Stat. A*, **12**, 1161–1179.
- Silvey, S. D. (1970). *Statistical Inference*. London: Penguin.
- Simonoff, J. S. (1995). *Smoothing Methods in Statistics*. New York: Springer-Verlag.
- Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Am. Stat. Assoc.*, **87**, 439–450.
- Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. *J. Am. Stat. Assoc.*, **75**, 74–81.
- Speed, F. M. and Hocking, R. R. (1976). The use of $R()$ notation with unbalanced data. *Am. Stat.*, **30**, 30–33.
- Speed, F. M., Hocking, R. R. and Hackney, O. P. (1978). Methods of analysis of linear models with unbalanced data. *J. Am. Stat. Assoc.*, **73**, 105–112.
- Speigelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and loglinear models. *J. R. Stat. Soc. B*, **44**, 377–387.
- Spjøtvoll, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Stat.*, **43**, 398–411.
- Sprent, P. (1961). Some hypotheses concerning two phase regression lines. *Biometrics*, **17**, 634–645.
- Sprent, P. (1969). *Models in Regression and Related Topics*. London: Methuen.
- Stewart, G. W. (1976). The economical storage of plane rotations. *Numer. Math.*, **25**, 137–138.
- Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Stat. Sci.*, **5**, 147–155.
- Stone, M. (1974). Cross-validatory choice and the assessment of statistical predictions (with discussion). *J. R. Stat. Soc. B*, **36**, 111–147.

- Stromberg, A. J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Sci. Comput.*, **14**, 1289–1299.
- Stromberg, A. J., Hössjer, O. and Hawkins, D. M. (2000). The least trimmed differences estimator and alternatives. *J. Am. Stat. Assoc.*, **95**, 853–864.
- Swindel, B. F. (1968). On the bias of some least-squares estimators of variance in a general linear model. *Biometrika*, **55**, 313–316.
- Swindel, B. F. and Bower, D. R. (1972). Rounding errors in the independent variables in general linear model. *Technometrics*, **14**, 215–218.
- Theil, H. (1965). The analysis of disturbances in regression analysis. *J. Am. Stat. Assoc.*, **60**, 1067–1079.
- Theil, H. (1968). A simplification of the BLUS procedure for analyzing regression disturbances. *J. Am. Stat. Assoc.*, **63**, 242–251.
- Theil, H. and Schweitzer, A. (1961). The best quadratic estimator of the residual variance in regression analysis. *Stat. Neerl.*, **15**, 19–23.
- Thompson, M. L. (1978). Selection of variables in multiple regression: Part I: A review and evaluation. Part II: Chosen procedures, computations and examples. *Int. Stat. Rev.*, **46**, 1–19, 129–146.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.
- Tibshirani, R. and Hastie, T. J. (1990). *Generalized Additive Models*. New York: Chapman & Hall.
- Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. B*, **61**, 529–546.
- Todd, J. (1954). The condition of the finite segments of the Hilbert matrix. In O. Taussky (Ed.), *Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues*, National Bureau of Standards Applied Mathematics Series 39, Washington, DC: U.S. Government Printing Office, pp. 109–116.
- Todd, J. (1961). Computational problems concerning the Hilbert matrix. *J. Res. Nat. Bur. Standards*, **65**, 19–22.
- Tong, Y. L. (1980). *Probability Inequalities in Multivariate Distributions*. New York: Academic Press.
- Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. Philadelphia: SIAM.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242.
- Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton, NJ.
- Tukey, J. W. (1954). Causation, regression and path analysis. In O. Kempthorne (Ed.), *Statistics and Mathematics in Biology*. Ames, IA: Iowa State College Press, pp. 35–66.
- Turner, M. E. (1960). Straight line regression through the origin. *Biometrics*, **16**, 483–485.
- Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *Am. Stat.*, **35**, 234–242.

- Verbyla, A. P. (1993). Modeling variance heterogeneity: Residual maximum likelihood and diagnostics. *J. R. Stat. Soc. B*, **55**, 493–508.
- Wahba, G. (1990). *Spline Methods for Observational Data*. Philadelphia: SIAM.
- Walls, R. C. and Weeks, D. L. (1969). A note on the variance of a predicted response in regression. *Am. Stat.*, **23**, 24–26.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. New York: Chapman & Hall.
- Warren, W. G. (1971). Correlation or regression: Bias or precision. *Appl. Stat.*, **20**, 148–164.
- Watkins, D. S. (1991). *Fundamentals of Matrix Computations*. New York: Wiley.
- Wedderburn, R. W. M. (1974). Generalized linear models specified in terms of constraints. *J. R. Stat. Soc. B*, **36**, 449–454.
- Wilkinson, G. N. (1970). A general recursive procedure for analysis of variance. *Biometrika*, **57**, 19–46. **23**, 377–380.
- Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, **26**, 23–32.
- Williams, E. J. (1959). *Regression Analysis*. New York: Wiley.
- Wold, S. (1974). Spline functions in data analysis. *Technometrics*, **16**, 1–11.
- Working, H. and Hotelling, H. (1929). Application of the theory of error to the interpretation of trends. *J. Am. Stat. Assoc. Suppl. (Proc.)*, **24**, 73–85.
- Wynn, H. P. and Bloomfield, P. (1971). Simultaneous confidence bands in regression analysis. *J. R. Stat. Soc. B*, **33**, 202–217.
- Yates, F. (1972). A Monte-Carlo trial on the behavior of the non-additivity test with non-normal data. *Biometrika*, **59**, 253–261.
- Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality and symmetry. *Biometrika*, **87**, 954–959.
- Zhang, P. (1992). On the distributional properties of model selection criteria. *J. Am. Stat. Assoc.*, **87**, 732–737.
- Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Stat.*, **21**, 299–313.
- Zimmerman, D (1987). Comments on “Simultaneous confidence intervals for prediction” (86V40 p279), *Am. Stat.*, **41**, 247.

Index

Abbreviated probability error rate, 120
Accuracy
 of Givens, 371
 of Householder, 371
 of MGSA, 371
 of SVD, 371
Added variable plot, 277–278
Adding an extra variable, 54, 57
 effect on confidence interval, 135
 effect on prediction interval, 135
 full rank case, 57
 less than full rank case, 65
Adding columns to the data matrix, 356
Adding rows to the data matrix, 356
Additivity
 test for, 213
Adjusted R^2 , 400
AIC, 408
 expected value of, 409
 modified, 409
 σ^2 known, 410
Akaike information criterion, 408
Algorithm
 for all possible regressions, 393, 439
 for all possible regressions via QR, 446
 for Cholesky decomposition, 336
 for high-breakdown methods, 385
 for stepwise regression, 418
Furnival, 441
Furnival and Wilson, 442

Garside, 439
Gram–Schmidt, 338
modified Gram–Schmidt, 341
Morgan and Tatar, 441
Nelder–Mead, 438
Schatzoff et al., 440
Wilkinson, 58
All possible regressions, 392, 399
 Garside algorithm for, 439
 Morgan and Tatar method for, 441
 order of generation, 440
 using sweeps for, 439
Alternative parameterizations
 in one-way classification, 190
Analysis of covariance, 222
Analysis of variance, 187
Andrews and Pregibon statistic, 309
ANOVA table
 higher-way classification, 219
 one-way classification, 192
 two-way classification, 208
 two-way classification with one
 observation per mean, 212
Apparent residuals, 58
Backfitting, 276
Back-substitution, 337
Backward analysis, 370
Backward elimination, 168, 392, 416
Backward stable algorithm, 370
Bayes estimate

- in inverse prediction, 147
- Bayes factor, 429
- Bayes' formula, 73
- Bayesian information criterion, 410, 412
- Bayesian model averaging, 392, 433
- Bayesian prediction, 392, 428, 431
- Berkson's model, 240
- Best linear unbiased estimate, 68
- Best unbiased estimate, 50
- Beta distribution, 473
- Bias**
 - due to overfitting, 230
 - due to underfitting, 228
 - in polynomial fitting, 393
 - trade-off with variance, 394
- BIC, 410, 412
- Binomial distribution, 8
- Bivariate normal density, 19
- Bivariate normal distribution, 19
- Biweight function, 90
- BLUE, 42, 68
- BLUS residuals, 327
- Bonferroni confidence interval, 121–122, 124, 129, 131, 142
 - percentage points for, 480
- Bonferroni inequality, 121
- Bonferroni method, 193
 - in straight-line regression, 140
- Bonferroni tests, 121
- Bonferroni *t*-interval, 121–122, 124, 129, 131, 142
- Bound tree, 446
- Box–Cox transformation, 276, 297
- Branch tree, 446
- Breakdown point, 82
 - of generalized M-estimate, 89
 - of LMS estimate, 83
 - of LQD estimate, 92
 - of LTS estimate, 83
 - of one-step estimate, 89
 - of S-estimate, 90
- B-spline basis, 174
- C_p , 402
- Calibration, 145
- Canonical form for *F*-test, 113
- Cauchy distribution, 13
- Cauchy–Schwartz inequality, 463
- Centering, 69
- Centering the data
 - algorithms for, 363
- CERES plot, 275
- Changeover point, 160
- Chebyshev fit, 386
- Chebyshev polynomials, 169
- Chi-square distribution, 13
 - idempotent matrices and, 28
- m.g.f. for, 13
- notation for, 2
 - of differences in two quadratic forms, 29
- Cholesky decomposition, 329, 336
- Cholesky factor, 336
- Choosing the best regression subset, 399
- Coefficient of determination, 111, 400
 - expectation of, 113
- Coincidence
 - test for, 155
- Collinearity, 249, 315
 - diagnosis of, 315
 - remedies for, 321
- Column space
 - notation for, 2
- Columns
 - linearly independent, 37
- Comparing two means, 104
- Comparing two regression models, 114
- Concurrence
 - test for, 156
- Condition number, 316
- Conditional distributions
 - for multivariate normal, 25
- Conditional expectation
 - notation for, 1
- Conditional likelihood, 437
- Confidence band
 - for regression surface, 129
 - for straight line regression, 141
 - Gafarian, 143
 - Graybill–Bowden, 142
 - Working–Hotelling, 142
- Confidence interval
 - Bonferroni, 121–122, 124, 129, 131, 142
 - for contrast, 107
 - for regression surface, 129
 - for special subset, 124
 - for straight-line regression, 139
 - for x intercept in straight-line regression, 140
 - in one-way classification, 192
 - in two-way classification, 204
 - maximum modulus, 124, 130, 142
 - simultaneous, 119
 - simultaneous for regression coefficients, 126
 - Tukey–Kramer, 193
- Confidence region, 125
- Conjugate prior, 74
- Consistency, 79
- Contrast, 187
 - confidence interval for, 107, 123
 - test for, 107
- Cook's *D*, 309
- Correlation

- multiple, 110
- Covariance ratio, 307
- Covariance
 - independence and, 14
 - notation for, 1
 - operator, 5
- COVRATIO, 307
- Cramer–Rao lower bound, 50
- Criteria
 - for variable selection, 392, 399
- Cross-validation
 - expected value of statistic, 404
 - for subset selection, 403
 - leave- d -out, 405
 - leave-one-out, 403
 - use in spline smoothing, 179
- Cubic splines, 173
- Density
 - bivariate normal, 19
 - of multivariate normal, 17
- Design matrix, 36, 187
 - less than full rank, 62
- Determinant
 - notation for, 2
- DFBETAS, 306
- DFFITS, 307
- Diagonal matrix
 - notation for, 1
- Dispersion matrix, 6
- Distribution
 - beta, 473
 - binomial, 8
 - bivariate normal, 19
 - Cauchy, 13
 - chi-square, 13
 - F , 99
 - inverted gamma, 74
 - multinomial, 8
 - multivariate t , 74, 121, 473
 - of quadratic forms, 27
 - of Studentized range, 193
 - posterior, 74
 - t , 13
- Double precision, 330
- Dual tree, 443
- Dummy variable, 4
 - in ANOVA models, 191
 - use in straight-line regression, 156
- Durbin–Watson test, 292
- Dynamic plot, 271
- Effective degrees of freedom
 - in spline smoothing, 179
- Efroymson's method, 418
- Eigenvalues
 - of correlation matrix, 255
 - of idempotent matrix, 27
- Elemental regression, 304, 314, 385
- Empirical distribution function, 84
- Equal variances
 - test for, 195
- Error variance
 - prior for, 74
 - unbiased estimate of, 44
- Estimable functions, 64
- Estimate
 - best linear unbiased, 68
 - best unbiased, 50
 - bounded influence, 88
 - garrote, 425
 - generalized least squares, 66, 94
 - generalized M-, 88
 - generalized S-, 92
 - James–Stein, 421
 - L_1 , 79
 - LAD, 79
 - lasso, 427
 - least median of squares, 78, 80
 - least squares, 35, 41, 44, 53, 55, 57, 59, 62, 77, 93, 95
 - least trimmed squares, 78, 81
 - LQD, 92
 - LTD, 93
 - M-, 77
 - MAD, 80
 - maximum likelihood, 49
 - minimum norm quadratic unbiased, 47
 - one-step GM, 89
 - ordinary least squares, 68
 - plug-in, 84
 - R-, 91
 - restricted least squares, 59
 - ridge, 423
 - robust, 77
 - S-, 90
 - shrinkage, 420
 - unique nonnegative quadratic unbiased, 45
 - weighted least squares, 67
- Estimated residual variance, 400
- Expectation
 - notation for, 1
 - operator, 5
- Expected value of quadratic forms, 9
- Experimentwise error rate, 120
- Explanatory variable, 3, 36
 - adding to model, 54, 135
 - dummy variable, 130
 - errors in, 241, 246
 - random, 5
- Exponent, 369
- Externally Studentized residual, 267
- Factor, 188

- adding to ANOVA model, 57
- levels of, 188
- Familywise error rate**, 120
- Fast rotator**, 349
- F-distribution**, 99
 - notation for, 2
- Feasible solution algorithms**, 387
- Fieller's method**, 140
- First order interaction**, 216
- Fitted values**, 38, 266
- Floating point operations**, 330
- Floating point representation**, 369
- Flop**, 330
- Flop count**
 - drawbacks of, 366
 - for Cholesky, 368
 - for Givens QR, 349
 - for Householder QR, 346, 368
 - for modified Gram-Schmidt, 343, 368
 - for SVD, 368
- Forward ranking**, 414
- Forward selection**, 168, 392, 414
- Fourth moment about the mean**, 10
- F-test**
 - calculation of, 380
 - canonical form for, 113
 - for linear hypothesis, 99
 - less than full rank case, 116
 - quadratically balanced, 236
 - robustness to non-normality, 235
- Furnival's method**, 441
- Gafarian confidence band**, 143
- Garrote**, 392, 425
 - for orthonormal regressors, 427
 - nonnegative, 426
- Gaussian elimination**, 329, 331
 - use in APR, 334
- General linear hypothesis**
 - canonical form for, 113
 - for regression model, 98
 - test of, 99–100
- Generalized cross-validation**
 - use in spline smoothing, 179
- Generalized inverse**, 38, 63, 469
 - calculation of, 469
 - definition of, 469
 - Moore-Penrose, 469
- Generalized least squares**, 66
- Generalized least squares estimate**, 94
- Generalized M-estimate**, 88
 - breakdown point of, 89
- Generalized S-estimate**, 92
- Generating the better regressions**, 442
- Givens transformation**, 348
 - fast rotator, 349
 - QR using, 348
- use in all possible regressions, 446
- use in updating regressions, 361
- Goodness-of-fit measures**, 400
- Goodness-of-fit test**, 115
- Gram-Schmidt algorithm**, 338
 - modified, 341
- Gravitation**, 3, 97
- Graybill-Bowden confidence band**, 142
- Hat matrix**, 266
- Hat matrix diagonals**, 267
 - calculation of, 379
- Helmart transformation**, 24, 191
- Hierarchical design**, 221
- Higher-way classification**
 - ANOVA table for, 219
 - hypothesis testing in, 217
- High-influence point**, 303
- High-leverage point**, 88, 233, 301
- Hilbert matrix**, 166, 372
- Householder transformation**, 343
 - use in less than full rank regression, 376
 - use in QR decomposition, 345
 - use in updating regressions, 362
- Huber Proposal 2**, 79
- Hypothesis**
 - of no interactions, 198
- Idempotent matrix**, 36
 - definition of, 28
 - eigenvalues of, 27
- Identifiability constraints**, 63
- IEEE standard**, 370
- Ill-conditioned regression problem**, 369
 - in polynomial regression, 165
- Importance sampling**, 430
- Incorrect variance matrix**, 231
- Independence**, 13
 - in bivariate normal, 14
 - in multivariate normal, 24
 - moment generating functions and, 13
 - of $\hat{\beta}$ and S^2 , 48
 - of quadratic forms, 29
- Independent variable**, 3
- Indicator variable**, 4
- Influence curve**, 83
- Influential point**, 234
- Information matrix**, 50
- Interaction**, 198
 - definition of, 216
 - in higher-way classification, 216
- Internally Studentized residual**, 267
- Inverse estimate**
 - in inverse prediction, 147
- Inverse prediction**, 145
- Inverse**
 - notation for, 2
- Inverted gamma distribution**, 74

- Jacobian, 17
- James–Stein shrinkage estimate, 421
- John–Draper transformation, 298
- Knots, 173
 - choice of, 175
 - number of, 178
- Kruskal–Wallis test, 196
- Kullback–Leibler discrepancy, 298, 407
- L_1 estimate, 79
- LAD estimate, 79
- Lagrange multiplier, 60, 98, 103, 181
- Laplace approximation, 430
- Lasso, 392, 427
 - for orthonormal regressors, 427
- Least median of squares, 78
- Least median of squares estimate, 80
- Least median of squares
 - inefficiency of, 82
 - instability of, 81
- Least squares
 - weighted, 150–151
- Least squares estimate, 35–37, 41, 44, 53, 55, 57, 59, 62, 77, 93, 95
 - distribution of, 47
 - in weighted least squares, 153
 - independent of S^2 , 48
 - influence curve of, 86
 - properties of, 42
 - unbiased, 48
 - under linear restrictions, 59
 - variance matrix of, 48
- Least squares under restrictions
 - method of orthogonal projections, 61
- Least trimmed squares, 78
- Least trimmed squares estimate, 81
- Least trimmed squares
 - inefficiency of, 82
- Levene test, 195
- Likelihood
 - connection with M-estimate, 78
 - for linear regression model, 49
- Likelihood ratio test, 98
- Linear regression model, 4
 - constant term in, 36
 - general hypothesis for, 97
 - likelihood for, 49
 - linear hypothesis for, 98
 - matrix form of, 35
- Little bootstrap, 406, 424, 453
- Local linear regression, 162
 - multidimensional, 184
- Loess, 163, 271
- Lowess, 162
- LQD estimate, 92
 - breakdown point of, 92
- LTD estimate, 93
- M-estimate, 77
 - influence curve of, 87
- MAD estimate, 80, 90
- Mahalanobis distance, 70, 261, 269
- Main effects, 217
- Mallows' C_p , 402
- Mallows weights, 89
- Mantissa, 369
- Marginal distribution, 22
- Markov chain Monte Carlo, 430
- Masking, 312
- Matrix
 - design, 36
 - dispersion, 6
 - idempotent, 36
 - information, 50
 - notation for, 1
 - projection, 36
 - regression, 36
 - sum of squares and cross-products, 329
 - variance, 6
- Maximum likelihood estimate, 49
 - in inverse prediction, 145
 - in weighted least squares, 151
- Maximum modulus confidence interval, 124, 130, 142
- Maximum modulus t -interval, 121, 124, 130, 142
- Mean
 - influence curve of, 85
 - of multivariate normal, 18
- Means
 - comparing, 104
 - comparing two, 4
- Median scores, 91
- MGSA, 341
- Minimum covariance determinant estimate, 305
- Minimum volume ellipsoid, 304
- MINQUE, 47
- Missing observations
 - in balanced design, 220
- Model
 - linear regression, 4
 - regression through the origin, 270
 - transform both sides, 299
 - for two straight lines, 97
- Model averaging, 392
- Model error, 394, 396
- Model selection, 98, 391
- Modified Gram–Schmidt algorithm, 341
 - flop count for, 343
- Moment generating function, 13
 - of chi-squared, 13
 - of multivariate normal, 20
- Moore–Penrose generalized inverse, 469

- Multidimensional smoothing, 184
 Multidimensional splines, 184
 Multinomial distribution, 8
 Multiple comparisons
 Scheffé method for, 124
 Tukey–Kramer method for, 193
 Multiple correlation coefficient, 110
 Multivariate normal, 17
 characterization of, 22
 conditional distributions, 25
 density of, 17
 extended definition of, 21
 independence and, 24
 marginal distributions, 22
 mean of, 18
 m.g.f. of, 20
 variance matrix of, 18
 Multivariate *t*-distribution, 74, 121, 431, 473
 Nelder–Mead algorithm, 438
 Nested design, 221
 Noninformative prior, 73
 Nonnegative garrote, 426
 Nonnormality
 effect on *F*-test, 235
 Norm
 definition of, 1
 of matrix, 256
 Normal distribution
 independence of sample mean and variance, 25
 notation for, 2
 Normal equations, 37, 330
 Normal plot, 295
 Notation, 1
 Null space
 notation for, 2
 Occam's window, 433
 Ohm's Law, 3
 Omission bias, 436
 One-step estimate
 breakdown point of, 89
 One-step GM-estimate, 89
 One-way classification, 188
 ANOVA table for, 192
 balanced case, 194
 confidence intervals in, 192
 F-test for, 189
 One-way layout, 188
 Orthogonal columns
 in regression matrix, 51
 regression coefficients unchanged, 51
 variance minimised when, 52
 Orthogonal complement, 40
 Orthogonal polynomials, 166
 for equally spaced x -values, 170
 generation of, 168
 statistical properties of, 166
 Orthogonal projection, 40, 61
 Orthonormal basis, 338
 Outlier, 77, 233, 301
 Outlier shift model, 310
 Overfitting, 230
 Pairwise independence, 14
 Parallelism
 test for, 155
 Partial residual, 273
 Partial residual plot, 272
 Partitioned matrices, 466
 Path of steepest ascent, 181
 Patterned matrices, 466
 Permutation matrix, 361, 376, 464
 Piecewise polynomial fitting, 173
 Pivoting
 in Gaussian elimination, 333, 367, 370
 Plug-in estimate, 84
 robustness of, 85
 Polynomial regression, 165
 choosing degree in, 168
 ill-conditioning in, 165
 use in surface fitting, 180
 Positive-definite matrix, 8, 17, 461
 Positive-semidefinite matrix, 21, 460
 Posterior predictive density, 431
 Posterior
 density function, 73
 distribution, 74
 for regression coefficients, 74
 Prediction, 391
 Prediction band
 in straight-line regression, 142
 Prediction error, 394, 396
 Prediction interval, 131
 in straight-line regression, 141
 Predictive density, 428
 Predictive loss, 449
 Predictor
 calibration of, 448
 construction of, 391
 Prior
 conjugate, 74
 for error variance, 74
 for regression coefficients, 74
 improper, 73
 noninformative, 73
 Prior information, 73
 Production function, 3
 Profile likelihood
 use in inverse prediction, 148
 Projection matrix, 36–37, 66, 464
 for centered data, 70
 for two-variable model, 72

- F*-test and, 116
- Pseudolikelihood, 289
- QR decomposition, 330, 338
 - calculation of regression quantities using, 340
 - in rank-deficient case, 376
 - use in adding and deleting cases, 360
 - use in adding and deleting variables, 360
 - using Givens transformations, 348
 - using Householder transformations, 345
 - using MGSA, 341
- Quadratic form, 9
 - chi-squared distribution of, 28
 - condition to be chi-squared, 30
 - distribution of, 27
 - independence of, 29
 - mean of, 9
 - variance of, 9
- Quadratically balanced *F*-test, 236
- R, 121
- R^2 , 111
- Random explanatory variables, 5, 240
- Randomized block design, 62
- Rank, 458
 - calculation of in presence of round-off, 378
- Regression
 - testing significance of, 112
- Regression analysis
 - aim of, 2
- Regression calculations
 - for all possible regressions, 439
 - for robust regression, 382
 - using fast rotators, 350
 - using Gaussian elimination, 332
 - using Householder transformations, 346
 - using MGSA, 342
 - using SVD, 353
- Regression coefficients
 - in straight-line regression, 107
 - MLE of, 49
 - posterior for, 74
 - prior for, 74
 - tests for, 106
- Regression matrix, 36
 - ill-conditioned in polynomial regression, 165
 - orthogonal columns in, 51
- Regression model
 - linear, 4
 - two phase, 159
- Regression splines, 173
- Regression surface, 271
 - confidence band for, 129
 - confidence interval for, 129
- Regression tree, 442
- Regression updating, 356
- Regressor, 3
- Relationships
 - between variables, 3
 - causal, 3
- REML, 287
- Residual, 38
 - apparent, 58
 - BLUS, 327
 - externally Studentized, 267
 - internally Studentized, 267
 - partial, 273
 - properties of, 266
- Residual plot, 272, 283
- Residual sum of squares, 38, 400
- Response surface, 180
- Response variable, 3, 36
- R-estimate, 91
- Restricted likelihood, 287
- Ridge estimate, 423
 - as Bayes estimate, 424
 - as solution of constrained least squares problem, 425
 - orthogonal case, 425
- Ridge parameter, 423
 - estimation of, 424
 - estimation of via cross validation, 424
 - estimation of via GCV, 424
- Ridge regression, 392, 423
- $R()$ -notation, 202
- Robust estimation, 77, 304, 313
- Round-off error, 245, 370
- RSS, 38
- σ^2 (Error variance)
 - estimation of, 44
- S-estimate, 90
 - breakdown point of, 90
- S-interval, 142
- Sample correlation, 108
- Sample mean and variance
 - independence for normal distribution, 25
- Scaling, 69, 71, 250, 318
- Scheffé interval, 124, 129–130, 142
- Scheffé's method, 123–124, 130, 142, 193
- Schweppe weights, 88
- Second order interaction, 216
- Selection bias, 437
- Shrinkage estimate, 420
- Simple block structure, 221
- Simultaneous confidence interval, 119
 - comparisons between, 124
 - relationship with *F*-test, 127
- Simultaneous confidence intervals, 124
- Simultaneous prediction interval
 - for straight-line regression, 145
- Single explanatory variable

- adding to model, 57
- Singular normal distribution, 21
- Singular value decomposition, 353, 471
- Smoother matrix, 178
- Smoothing parameter
 - in spline smoothing, 177–178
- Smoothing splines, 176, 271
- Spectral decomposition, 27
- Spline, 173
 - cubic, 173
 - multidimensional, 184
 - regression, 173
 - smoothing, 176
 - thin plate, 184
- S-PLUS, 121, 162–163, 176, 178–179, 184
- SSCP matrix, 329–330, 363
- Standard normal distribution
 - definition of, 2
- Statistical functional, 83
 - influence curve of, 85
- Statistical models, 2
- Steepest ascent, 180
- Stein shrinkage, 420
- Stepwise regression, 392, 418
- Straight-line regression, 2, 107, 139
 - Bonferroni method in, 140
 - calibration in, 145
 - comparing two lines, 154
 - confidence interval for x intercept in, 140
 - confidence interval in, 139
 - estimating coefficients in, 139
 - inverse prediction in, 145
 - least squares estimates in, 139
 - prediction band in, 142
 - prediction interval in, 141
 - regression coefficients in, 107
 - simultaneous confidence interval for, 122
 - simultaneous prediction interval for, 145
 - tests for coefficients in, 109
 - through origin, 149
- Studentized maximum-modulus method, 193
- Studentized range distribution, 193
- Subset selection, 392
 - bias due to, 436
 - effect on inference, 434
- SVD, 353
 - calculation of, 354
 - regression calculations using, 353
- Swamping, 312
- Sweep operator, 329
- Sweeping, 335
 - in all possible regressions, 440
 - use in updating, 357
- t -distribution, 13
 - notation for, 2
- t -interval
 - Bonferroni, 121–122, 124, 129, 131, 142
 - in simultaneous inference, 119
 - maximum modulus, 121, 124, 130, 142
- Tensor product basis, 184
- Test
 - Bonferroni, 121
 - Durbin–Watson, 292
 - for additivity, 213
 - for coincidence, 155
 - for concurrence, 156
 - for interactions, 213
 - for outliers, 310
 - for parallelism, 155
 - for significance of regression, 112
 - in one-way classification, 189
 - in two-way classification, 201
 - Kruskal–Wallis, 196
 - Levene (for equal variances), 195
 - of general linear hypothesis, 100
 - of goodness of fit, 115
 - Tukey (for additivity), 213
- Textbook algorithm
 - for centered SSCP matrix, 363
- Thin plate spline, 184
- Third moment about the mean, 10
- Threshold, 406
- Total bias, 397
- Total variance, 397
- Trace
 - definition of, 1
- Training set, 394
- Transformation
 - Box–Cox, 276, 297
 - John–Draper, 298
- Transpose
 - definition of, 2
- Tukey test, 213
- Tukey–Kramer confidence interval, 193
- Two-phase regression, 159
 - with known changeover point, 160
 - with unknown changeover point, 160
- Two-pass algorithm
 - for centered SSCP matrix, 364
- Two-way classification
 - ANOVA table for, 208
 - balanced, 206
 - confidence intervals in, 204
 - F -test in, 197
 - one observation per mean, 211
 - unbalanced, 197
- Two-way classification with one observation per mean
 - ANOVA table for, 212
- Type 1 procedure, 202
- Type 2 procedure, 204

- Type 3 procedure, 204
- Type 4 procedure, 204
- Underfitting, 228
- Underlying assumptions, 227
- Unit round-off, 370
- Updating algorithm
 - for centered SSCP matrix, 364
- Updating the regression model, 356
- Upper triangular matrix, 331, 339, 342
- Van der Waerden scores, 91
- Variable
 - dummy, 4
 - explanatory, 3, 36
 - independent, 3
 - indicator, 4
- response, 36
- Variance matrix, 6
 - of MLE in weighted least squares, 152
 - of multivariate normal, 18
 - singular, 8
- Variance
 - notation for, 1
- Vector
 - notation for, 1
- Vector differentiation, 466
- Weighted least squares, 67, 150–151, 355
 - with known weights, 150
 - with unknown weights, 151
- Weighted least squares estimate, 288
- Wilcoxon scores, 91
- Working–Hotelling confidence band, 142

Linear Regression Analysis, Second Edition

by George A. F. Seber and Alan J. Lee

Copyright © 2003 John Wiley & Sons, Inc.

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Second Edition*
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
- BERNARDO and SMITH · Bayesian Theory
- BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
- BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
- BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY · Probability and Measure, *Third Edition*
- BIRKES and DODGE · Alternative Methods of Regression
- BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
- BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN · Structural Equations with Latent Variables
- BOROVKOV · Ergodicity and Stability of Stochastic Processes
- BOULEAU · Numerical Methods for Stochastic Processes
- BOX · Bayesian Inference in Statistical Analysis
- BOX · R. A. Fisher, the Life of a Scientist
- BOX and DRAPER · Empirical Model-Building and Response Surfaces
- *BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
- BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
- BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
- BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
- BROWN and HOLLANDER · Statistics: A Biomedical Introduction
- BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
- BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
- CAIROLI and DALANG · Sequential Stochastic Optimization
- CHAN · Time Series: Applications to Finance
- CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
- CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
- CHERNICK · Bootstrap Methods: A Practitioner's Guide
- CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
- CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
- CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies
- CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
- *COCHRAN and COX · Experimental Designs, *Second Edition*
- CONGDON · Bayesian Statistical Modelling
- CONOVER · Practical Nonparametric Statistics, *Second Edition*
- COOK · Regression Graphics
- COOK and WEISBERG · Applied Regression Including Computing and Graphics
- COOK and WEISBERG · An Introduction to Regression Graphics
- CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
- COVER and THOMAS · Elements of Information Theory
- COX · A Handbook of Introductory Statistical Methods
- *COX · Planning of Experiments
- CRESSIE · Statistics for Spatial Data, *Revised Edition*
- CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
- DANIEL · Applications of Statistics to Industrial Experimentation
- DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Sixth Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- *DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
- DAVID · Order Statistics, *Second Edition*
- *DEGROOT, FIENBERG, and KADANE · Statistics and the Law
- DEL CASTILLO · Statistical Process Adjustment for Quality Control
- DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
- DEY and MUKERJEE · Fractional Factorial Plans
- DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
- DODGE · Alternative Methods of Regression
- *DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
- *DOOB · Stochastic Processes
- DOWDY and WEARDEN · Statistics for Research, *Second Edition*
- DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
- DRYDEN and MARDIA · Statistical Shape Analysis
- DUDEWICZ and MISHRA · Modern Mathematical Statistics
- DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, *Second Edition*
- DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
- DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
- *ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
- ETHIER and KURTZ · Markov Processes: Characterization and Convergence
- EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
- FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition, Revised; Volume II, Second Edition*
- FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
- *FLEISS · The Design and Analysis of Clinical Experiments
- FLEISS · Statistical Methods for Rates and Proportions, *Second Edition*
- FLEMING and HARRINGTON · Counting Processes and Survival Analysis
- FULLER · Introduction to Statistical Time Series, *Second Edition*
- FULLER · Measurement Error Models
- GALLANT · Nonlinear Statistical Models
- GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
- GIFI · Nonlinear Multivariate Analysis
- GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
- GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
- GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
- GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
- GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
- *HAHN and SHAPIRO · Statistical Models in Engineering
- HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
- HALD · A History of Probability and Statistics and their Applications Before 1750
- HALD · A History of Mathematical Statistics from 1750 to 1930
- HAMEL · Robust Statistics: The Approach Based on Influence Functions
- HANNAN and DEISTLER · The Statistical Theory of Linear Systems
- HEIBERGER · Computation for the Analysis of Designed Experiments
- HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
- HELLER · MACSYMA for Statisticians
- HINKELMAN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
- HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
- *HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory
Data Analysis
- HOCHBERG and TAMHANE · Multiple Comparison Procedures
- HOCKING · Methods and Applications of Linear Models: Regression and the Analysis
of Variance, *Second Edition*
- HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
- HOGG and KLUGMAN · Loss Distributions
- HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
- HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
- HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of
Time to Event Data
- HØYLAND and RAUSAND · System Reliability Theory: Models and Statistical Methods
- HUBER · Robust Statistics
- HUBERTY · Applied Discriminant Analysis
- HUNT and KENNEDY · Financial Derivatives in Theory and Practice
- HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—
with Commentary
- IMAN and CONOVER · A Modern Approach to Statistics
- JACKSON · A User's Guide to Principle Components
- JOHN · Statistical Methods in Engineering and Quality Assurance
- JOHNSON · Multivariate Statistical Simulation
- JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A
Volume in Honor of Samuel Kotz
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of
Econometrics, *Second Edition*
- JOHNSON and KOTZ · Distributions in Statistics
- JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the
Seventeenth Century to the Present
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
Volume 1, Second Edition
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
Volume 2, Second Edition
- JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
- JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
- JUREČKOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations
- JUREK and MASON · Operator-Limit Distributions in Probability Theory
- KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
- KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
- KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second
Edition*
- KASS and VOS · Geometrical Foundations of Asymptotic Inference
- KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster
Analysis
- KEDEM and FOKIANOS · Regression Models for Time Series Analysis
- KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
- KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
- KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
- KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
- KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models:
From Data to Decisions
- KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions,
Volume 1, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
- KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
- KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
- KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
- KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
- LACHIN · Biostatistical Methods: The Assessment of Relative Risks
- LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
- LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
- LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
- LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
- LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
- LAWSON · Statistical Methods in Spatial Epidemiology
- LE · Applied Categorical Data Analysis
- LE · Applied Survival Analysis
- LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
- LePAGE and BILLARD · Exploring the Limits of Bootstrap
- LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
- LIAO · Statistical Group Comparison
- LINDVALL · Lectures on the Coupling Method
- LINHART and ZUCCHINI · Model Selection
- LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
- LLOYD · The Statistical Analysis of Categorical Data
- MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
- MALLER and ZHOU · Survival Analysis with Long Term Survivors
- MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
- MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
- MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
- MARDIA and JUPP · Directional Statistics
- MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
- McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
- McFADDEN · Management of Data in Clinical Trials
- McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
- McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
- McLACHLAN and PEEL · Finite Mixture Models
- McNEIL · Epidemiological Research Methods
- MEEKER and ESCOBAR · Statistical Methods for Reliability Data
- MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
- *MILLER · Survival Analysis, *Second Edition*
- MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Third Edition*
- MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
- MUIRHEAD · Aspects of Multivariate Statistical Theory

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
- MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*
- MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
- NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- NELSON · Applied Life Data Analysis
- NEWMAN · Biostatistical Methods in Epidemiology
- OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
- OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
- OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
- PANKRATZ · Forecasting with Dynamic Regression Models
- PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- *PARZEN · Modern Probability Theory and Its Applications
- PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
- PIANTADOSI · Clinical Trials: A Methodologic Perspective
- PORT · Theoretical Probability for Applications
- POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
- PRESS · Bayesian Statistics: Principles, Models, and Applications
- PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
- PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
- PUKELSHEIM · Optimal Experimental Design
- PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
- PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
- *RAO · Linear Statistical Inference and Its Applications, *Second Edition*
- RENCHER · Linear Models in Statistics
- RENCHER · Methods of Multivariate Analysis, *Second Edition*
- RENCHER · Multivariate Statistical Inference with Applications
- RIPLEY · Spatial Statistics
- RIPLEY · Stochastic Simulation
- ROBINSON · Practical Strategies for Experimenting
- ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
- ROLSKI, SCHMIDL, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
- ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
- ROSS · Introduction to Probability and Statistics for Engineers and Scientists
- ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
- RUBIN · Multiple Imputation for Nonresponse in Surveys
- RUBINSTEIN · Simulation and the Monte Carlo Method
- RUBINSTEIN and MELAMED · Modern Simulation and Modeling
- RYAN · Modern Regression Methods
- RYAN · Statistical Methods for Quality Improvement, *Second Edition*
- SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
- *SCHEFFE · The Analysis of Variance
- SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
- SCHOTT · Matrix Analysis for Statistics
- SCHUSS · Theory and Applications of Stochastic Differential Equations
- SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
- *SEARLE · Linear Models
- SEARLE · Linear Models for Unbalanced Data
- SEARLE · Matrix Algebra Useful for Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

- SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER and LEE · Linear Regression Analysis, *Second Edition*
SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, *Second Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VIDAKOVIC · Statistical Modeling by Wavelets
WEISBERG · Applied Linear Regression, *Second Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZHOU, OBUCHOWSKI, and MCCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.