A MATHEMATICAL PRIMER FOR SOCIAL STATISTICS

Quantitative Applications in the Social Sciences

A SAGE PUBLICATIONS SERIES

1. **Analysis of Variance, 2nd Edition** *Iversen/ Norpoth*
2. **Operations Research Methods** *Nagel/Neef*
3. **Causal Modeling, 2nd Edition** *Asher*
4. **Tests of Significance** *Henkel*
5. **Cohort Analysis, 2nd Edition** *Glenn*
6. **Canonical Analysis and Factor Comparison** *Levine*
7. **Analysis of Nominal Data, 2nd Edition** *Reynolds*
8. **Analysis of Ordinal Data** *Hildebrand/Laing/ Rosenthal*
9. **Time Series Analysis, 2nd Edition** *Ostrom*
10. **Ecological Inference** *Langbein/Lichtman*
11. **Multidimensional Scaling** *Kruskal/Wish*
12. **Analysis of Covariance** *Wildt/Ahtola*
13. **Introduction to Factor** *Analysis Kim/Mueller*
14. **Factor Analysis** *Kim/Mueller*
15. **Multiple Indicators** *Sullivan/Feldman*
16. **Exploratory Data Analysis** *Hartwig/Dearing*
17. **Reliability and Validity Assessment** *Carmines/Zeller*
18. **Analyzing Panel** *Data Markus*
19. **Discriminant Analysis** *Klecka*
20. **Log-Linear Models** *Knoke/Burke*
21. **Interrupted Time Series Analysis** *McDowall/ McCleary/Meidinger/Hay*
22. **Applied Regression, 2nd Edition** *Lewis-Beck/ Lewis-Beck*
23. **Research Designs** *Spector*
24. **Unidimensional Scaling** *McIver/Carmines*
25. **Magnitude Scaling** *Lodge*
26. **Multiattribute Evaluation** *Edwards/Newman*
27. **Dynamic Modeling** *Huckfeldt/Kohfeld/Likens*
28. **Network Analysis** *Knoke/Kuklinski*
29. **Interpreting and Using Regression** *Achen*
30. **Test Item Bias** *Osterlind*
31. **Mobility Tables** *Hout*
32. **Measures of Association** *Liebetrau*
33. **Confirmatory Factor Analysis** *Long*
34. **Covariance Structure Models** *Long*
35. **Introduction to Survey Sampling, 2nd Edition** *Kalton*
36. **Achievement Testing** *Bejar*
37. **Nonrecursive Causal Models** *Berry*
38. **Matrix Algebra** *Namboodiri*
39. **Introduction to Applied Demography** *Rives/Serow*
40. **Microcomputer Methods for Social Scientists, 2nd Edition** *Schrodt*
41. **Game Theory** *Zagare*
42. **Using Published Data** *Jacob*
43. **Bayesian Statistical Inference** *Iversen*
44. **Cluster Analysis** *Aldenderfer/Blashfield*
45. **Linear Probability, Logit, and Probit Models** *Aldrich/Nelson*
46. **Event History and Survival Analysis, 2nd Edition** *Allison*
47. **Canonical Correlation Analysis** *Thompson*
48. **Models for Innovation Diffusion** *Mahajan/Peterson*
49. **Basic Content Analysis, 2nd Edition** *Weber*
50. **Multiple Regression in Practice** *Berry/Feldman*
51. **Stochastic Parameter Regression Models** *Newbold/Bos*
52. **Using Microcomputers in Research** *Madron/Tate/Brookshire*
53. **Secondary Analysis of Survey Data** *Kiecolt/ Nathan*
54. **Multivariate Analysis of Variance** *Bray/ Maxwell*
55. **The Logic of Causal Order** *Davis*
56. **Introduction to Linear Goal Programming** *Ignizio*
57. **Understanding Regression Analysis, 2nd Edition** *Schroeder/Sjoquist/Stephan*
58. **Randomized Response and Related Methods, 2nd Edition** *Fox/Tracy*
59. **Meta-Analysis** *Wolf*
60. **Linear Programming** *Feiring*
61. **Multiple Comparisons** *Klockars/Sax*
62. **Information Theory** *Krippendorff*
63. **Survey Questions** *Converse/Presser*
64. **Latent Class Analysis** *McCutcheon*
65. **Three-Way Scaling and Clustering** *Arabie/ Carroll/DeSarbo*
66. **Q Methodology, 2nd Edition** *McKeown/ Thomas*
67. **Analyzing Decision Making** *Louviere*
68. **Rasch Models for Measurement** *Andrich*
69. **Principal Components Analysis** *Dunteman*
70. **Pooled Time Series Analysis** *Sayrs*
71. **Analyzing Complex Survey Data, 2nd Edition** *Lee/Forthofer*
72. **Interaction Effects in Multiple Regression, 2nd Edition** *Jaccard/Turrisi*
73. **Understanding Significance Testing** *Mohr*
74. **Experimental Design and Analysis** *Brown/Melamed*
75. **Metric Scaling** *Weller/Romney*
76. **Longitudinal Research, 2nd Edition** *Menard*
77. **Expert Systems** *Benfer/Brent/Furbee*
78. **Data Theory and Dimensional Analysis** *Jacoby*
79. **Regression Diagnostics, 2nd Edition** *Fox*
80. **Computer-Assisted Interviewing** *Saris*
81. **Contextual Analysis** *Iversen*
82. **Summated Rating Scale Construction** *Spector*
83. **Central Tendency and Variability** *Weisberg*
84. **ANOVA: Repeated Measures** *Girden*
85. **Processing Data** *Bourque/Clark*
86. **Logit Modeling** *DeMaris*
87. **Analytic Mapping and Geographic Databases** *Garson/Biggs*
88. **Working With Archival Data** *Elder/Pavalko/Clipp*
89. **Multiple Comparison Procedures** *Toothaker*
90. **Nonparametric Statistics** *Gibbons*
91. **Nonparametric Measures of Association** *Gibbons*
92. **Understanding Regression Assumptions** *Berry*
93. **Regression With Dummy Variables** *Hardy*
94. **Loglinear Models With Latent Variables** *Hagenaars*
95. **Bootstrapping** *Mooney/Duval*
96. **Maximum Likelihood Estimation** *Eliason*
97. **Ordinal Log-Linear Models** *Ishii-Kuntz*
98. **Random Factors in ANOVA** *Jackson/Brashers*
99. **Univariate Tests for Time Series Models** *Cromwell/Labys/Terraza*
100. **Multivariate Tests for Time Series Models** *Cromwell/Hannan/Labys/Terraza*

Quantitative Applications in the Social Sciences

A SAGE PUBLICATIONS SERIES

101. **Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models** *Liao*
102. **Typologies and Taxonomies** *Bailey*
103. **Data Analysis: An Introduction** *Lewis-Beck*
104. **Multiple Attribute Decision Making** *Yoon/Hwang*
105. **Causal Analysis With Panel Data** *Finkel*
106. **Applied Logistic Regression Analysis, 2nd Edition** *Menard*
107. **Chaos and Catastrophe Theories** *Brown*
108. **Basic Math for Social Scientists: Concepts** *Hagle*
109. **Basic Math for Social Scientists: Problems and Solutions** *Hagle*
110. **Calculus** *Iversen*
111. **Regression Models: Censored, Sample Selected, or Truncated Data** *Breen*
112. **Tree Models of Similarity and Association** *Corter*
113. **Computational Modeling** *Taber/Timpone*
114. **LISREL Approaches to Interaction Effects in Multiple Regression** *Jaccard/Wan*
115. **Analyzing Repeated Surveys** *Firebaugh*
116. **Monte Carlo Simulation** *Mooney*
117. **Statistical Graphics for Univariate and Bivariate Data** *Jacoby*
118. **Interaction Effects in Factorial Analysis of Variance** *Jaccard*
119. **Odds Ratios in the Analysis of Contingency Tables** *Rudas*
120. **Statistical Graphics for Visualizing Multivariate Data** *Jacoby*
121. **Applied Correspondence Analysis** *Clausen*
122. **Game Theory Topics** *Fink/Gates/Humes*
123. **Social Choice: Theory and Research** *Johnson*
124. **Neural Networks** *Abdi/Valentin/Edelman*
125. **Relating Statistics and Experimental Design: An Introduction** *Levin*
126. **Latent Class Scaling Analysis** *Dayton*
127. **Sorting Data: Collection and Analysis** *Coxon*
128. **Analyzing Documentary Accounts** *Hodson*
129. **Effect Size for ANOVA Designs** *Cortina/Nouri*
130. **Nonparametric Simple Regression: Smoothing Scatterplots** *Fox*
131. **Multiple and Generalized Nonparametric Regression** *Fox*
132. **Logistic Regression: A Primer, 2nd Edition** *Pampel*
133. **Translating Questionnaires and Other Research Instruments: Problems and Solutions** *Behling/Law*
134. **Generalized Linear Models: A Unified Approach, 2nd Edition** *Gill/Torres*
135. **Interaction Effects in Logistic Regression** *Jaccard*
136. **Missing Data** *Allison*
137. **Spline Regression Models** *Marsh/Cormier*
138. **Logit and Probit: Ordered and Multinomial** *Models Borooah*
139. **Correlation: Parametric and Nonparametric Measures** *Chen/Popovich*
140. **Confidence Intervals** *Smithson*
141. **Internet Data Collection** *Best/Krueger*
142. **Probability Theory** *Rudas*
143. **Multilevel Modeling, 2nd Edition** *Luke*
144. **Polytomous Item Response Theory Models** *Ostini/Nering*
145. **An Introduction to Generalized Linear Models** *Dunteman/Ho*
146. **Logistic Regression Models for Ordinal Response Variables** *O'Connell*
147. **Fuzzy Set Theory: Applications in the Social Sciences** *Smithson/Verkuilen*
148. **Multiple Time Series Models** *Brandt/Williams*
149. **Quantile Regression** *Hao/Naiman*
150. **Differential Equations: A Modeling Approach** *Brown*
151. **Graph Algebra: Mathematical Modeling With a Systems Approach** *Brown*
152. **Modern Methods for Robust Regression** *Andersen*
153. **Agent-Based Models, 2nd Edition** *Gilbert*
154. **Social Network Analysis, 3rd Edition** *Knoke/Yang*
155. **Spatial Regression Models, 2nd Edition** *Ward/Gleditsch*
156. **Mediation Analysis** *Iacobucci*
157. **Latent Growth Curve Modeling** *Preacher/Wichman/MacCallum/Briggs*
158. **Introduction to the Comparative Method With Boolean Algebra** *Caramani*
159. **A Mathematical Primer for Social Statistics, 2nd Edition** *Fox*
160. **Fixed Effects Regression Models** *Allison*
161. **Differential Item Functioning, 2nd Edition** *Osterlind/Everson*
162. **Quantitative Narrative Analysis** *Franzosi*
163. **Multiple Correspondence Analysis** *LeRoux/Rouanet*
164. **Association Models** *Wong*
165. **Fractal Analysis** *Brown/Liebovitch*
166. **Assessing Inequality** *Hao/Naiman*
167. **Graphical Models and the Multigraph Representation for Categorical Data** *Khamis*
168. **Nonrecursive Models** *Paxton/Hipp/ Marquart-Pyatt*
169. **Ordinal Item Response Theory** *Van Schuur*
170. **Multivariate General Linear Models** *Haase*
171. **Methods of Randomization in Experimental Design** *Alferes*
172. **Heteroskedasticity in Regression** *Kaufman*
173. **An Introduction to Exponential Random Graph Modeling** *Harris*
174. **Introduction to Time Series Analysis** *Pickup*
175. **Factorial Survey Experiments** *Auspurg/Hinz*
176. **Introduction to Power Analysis: Two-Group Studies** *Hedberg*
177. **Linear Regression: A Mathematical Introduction** *Gujarati*
178. **Propensity Score Methods and Applications** Bai/Clark
179. **Multilevel Structural Equation Modeling** Silva/Bosancianu/Littvay
180. **Gathering Social Network Data** adams
181. **Generalized Linear Models for Bounded and Limited Quantitative Variables**, Smithson and Shou
182. **Exploratory Factor Analysis**, *Finch*
183. **Multidimensional Item Response Theory**, *Bonifay*
184. **Argument-Based Validation in Testing and Assessment**, *Chapelle*
185. **Using Time Series to Analyze Long Range Fractal Patterns**, *Koopmans*
186. **Understanding Correlation Matrices**, *Hadd and Rodgers*
187. **Rasch Models for Solving Measurement Problems**, *Engelhard and Wang*

To the memory of my mother, Diana,
the real mathematician in the family.

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

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

A MATHEMATICAL PRIMER FOR SOCIAL STATISTICS

Second Edition

John Fox
McMaster University

Quantitative Applications in the Social Sciences, Volume 159

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

Los Angeles | London | New Delhi
Singapore | Washington DC

FOR INFORMATION:

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

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

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

SAGE Publications Asia-Pacific Pte. Ltd.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Acquisitions Editor: Helen Salmon
Editorial Assistant: Elizabeth Cruz
Production Editor: Natasha Tiwari
Copy Editor: QuADS Prepress Pvt. Ltd.
Typesetter: Hurix Digital
Proofreader: Theresa Kay
Indexer: Integra
Cover Designer: Candice Harman
Marketing Manager: Victoria Velasquez

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Fox, John, 1947-author. | Sage (Firm)
Title: A mathematical primer for social statistics / John Fox, McMaster University.
Other titles: Quantitative applications in the social sciences.

Description: Second Edition. | Los Angeles : SAGE 2020. | Series: Quantitative applications in the social sciences | First edition published 2009. | Includes bibliographical references.

Identifiers: LCCN 2020031287 | ISBN 9781071833209 (Paperback : acid-free paper) | ISBN 9781071833247 (ePub) | ISBN 9781071833230 (ePub) | ISBN 9781071833223 (ePub)

Subjects: LCSH: Social sciences-Mathematics. | Social sciences-Statistical methods.

Classification: LCC H61.25 .F69 2020 | DDC 519.5-dc23

LC record available at https://lccn.loc.gov/2020031287

This book is printed on acid-free paper.
20 21 22 23 24 10 9 8 7 6 5 4 3 2 1

CONTENTS

SERIES EDITOR'S INTRODUCTION

The statistical sophistication of articles published in major social science journals has been increasing steadily over time. However, because the mathematical knowledge that social science students bring to their graduate statistics training has not always kept pace, the skills needed to fully understand, critique, and replicate these methods may be lacking. *A Mathematical Primer for Social Statistics* (2nd ed.) provides the missing foundation for those who need it and fills in the gaps for those whose training is spotty or out-of-date.

The *Primer*'s author, John Fox, is a well-known and respected expert in statistical methods. The mathematical concepts and skills needed to learn advanced social statistical methods are thus well-known to him. But perhaps as importantly, so are the areas of particular weakness among social scientists. The *Primer* is designed to address these weaknesses very specifically in order to provide the background social scientists need for the statistical methods they are likely to use. The scope is similar to that of the math camps that precede the beginning of PhD programs in economics and some political science, public policy, and sociology programs.

The *Primer* (2nd ed.) is organized around bodies of mathematical knowledge central to learning and understanding advanced statistics: the basic "language" of linear algebra, differential and integral calculus, probability theory, common probability distributions, and statistical estimation and inference. The volume concludes showing the application of mathematical concepts and operations to the familiar case, linear least-squares regression. Compared to the first edition of the *Primer*, published a decade ago, the second edition gives much more attention to visualization. It also covers some new topics—for example, an introduction to Markov-chain Monte Carlo methods. Also included is a companion website with materials that will enable readers to use the R statistical computing environment to reproduce and expand on computations presented in the volume.

The *Primer* would make an excellent text to accompany a math camp or a course designed to provide foundational mathematics needed to understand advanced statistics. It would also serve as a valuable reference for those who have completed their formal training but are still interested in learning new statistical methods. For example, those preparing to learn factor analysis or principal components analysis might benefit from a review of eigenvalues and eigenvectors (Chapter 2). Those about to dive into generalized linear models might usefully review the exponential family of distributions (Chap-

ter 5). In the process of working through an advanced text, readers might consult the *Primer* when they encounter a topic for which they need a quick refresher—for example, a Kronecker product (Chapter 1), a Lagrange multiplier (Chapter 3), or the likelihood ratio test (Chapter 6). A detailed Table of Contents as well as an Index help readers navigate the topics covered in the *Primer*, large and small.

Generations have learned from Professor Fox's many texts. In addition to the *Primer*, there are several others in the QASS Series: *Multiple and Generalized Nonparametric Regression* (Book 131); *Nonparametric Simple Regression: Smoothing Scatterplots* (Book 130); and *Regression Diagnostics*, 2nd ed. (Book 79). The *Primer* is thus in excellent company and will serve the needs of generations to come.

—*Barbara Entwisle*
Series Editor

ACKNOWLEDGMENTS

I am grateful to Barbara Entwisle, the academic editor of the Sage QASS series, and to several (originally anonymous) referees for their helpful comments and suggestions:

Scott Basinger, *University of Houston*

Victor Ferreros, *Walden University*

Scott Liebertz, *University of South Alabama*

I am also grateful to Helen Salmon, my editor at SAGE, for her continuing help and encouragement. Finally, I'd like to acknowledge support for this work from the Social Sciences and Humanities Research Council of Canada.

PREFACE

Statistics is not mathematics. Math is central to the development, communication, and understanding of statistics, but applied statistics—the kind of statistics of most interest to social scientists—is not about proving abstract theorems but about analyzing data.

Typical introductory statistics courses taught to social science students use only very basic mathematics—arithmetic, simple formulas, and the interpretation of graphs. There are good reasons for this: Most social science students have weak backgrounds in mathematics. Even more important, however, the fundamental goals of a basic statistics course (or at least what in my opinion should be the fundamental goals) are to convey the role of statistical methods in collecting and summarizing data along with the essential ideas of statistical inference. Accomplishing these goals is sufficiently challenging without drowning the big ideas in a sea of equations. I believe, incidentally, that this is the case even for students who have strong foundations in mathematics.

Once beyond the introductory level, and perhaps a second course in applied regression analysis, the situation changes: Insufficient grounding in mathematics makes it difficult to proceed in applied statistics. The good news, however, is that a relatively modest background in intermediate-level mathematics suffices for the study of a great deal of statistics. Often, all that is needed is an understanding of basic mathematical ideas, familiarity with some important facts, and an ability to read and perhaps manipulate equations. This book aims to provide that basic background.

The book originated in online appendices that I wrote for the second edition of my applied regression text (Fox, 2008, which is now in a third edition, Fox, 2016). I felt initially that some readers might prefer a printed and bound copy of the appendices to downloading them from the internet. It then occurred to me that the appendices might prove more generally useful, and ultimately I augmented them with material that was not directly relevant to my applied regression text but that is important to other statistical methods that are widely employed in the social sciences. The book, therefore, includes material not in the original appendices, and this second edition of the book includes material not in the first edition (see page xvii below).

The book covers three areas of mathematics that are of central importance to applied statistics:

- Chapters 1 and 2 takes up matrices, linear algebra, and vector geometry. Matrices, which are rectangular arrays of numbers, are a natural representation of most statistical data, and consequently, the arithmetic and algebra of matrices is the natural language for developing most statistical methods. Beyond the basic level, matrices are omnipresent in statistics, and therefore, some acquaintance with matrices is necessary for reading statistical material. The closely related areas of linear algebra and its visual representation, vector geometry, are also central to the development and understanding of many statistical methods.
- Chapter 3 introduces the basic ideas of differential and integral calculus. Here, the emphasis is on fundamental concepts and simple methods. Differential calculus is frequently used in statistics for optimization problems—that is, minimization and maximization: Think, for example, of the method of *least* squares or of *maximum*-likelihood estimation. Integral calculus figures prominently in probability theory, which is fundamentally tied to statistical modeling and statistical inference. Although the presentation of calculus in this book is elementary, I do cover topics important to statistics, such as multivariable and matrix calculus, that, while not fundamentally difficult, are often deferred to advanced treatments of the subject.
- Chapters 4, 5, and 6 develop probability theory, describe probability distributions important to statistics, and introduce statistical theory, including asymptotic distribution theory, the properties of estimators, the centrally important method of maximum likelihood, and the basics of Bayesian statistical inference. The ideas in these chapters feature prominently in applied statistics, and indeed, the three chapters represent a kind of "crash course" in some of the fundamentals of mathematical statistics.
- Chapter 7 illustrates the use of the preceding mathematics in applied statistics by briefly developing the seminal statistical method of linear least-squares regression and deriving some of its properties.

It is, all told, remarkable how far one can get in applied statistics with a modicum of mathematics—the modicum that this book supplies. This is the resource that I wish I had when I started to study statistics seriously. I hope that it will prove helpful to you, both on initial reading and as a reference.

What's New in the Second Edition

Although the material has been reorganized, the contents of the first edition of the book are included in the second edition, with small additions and modifications. There are as well a few more substantial additions to the book:

- Chapter 2 includes new material on visualizing quadratic forms using ellipses and on the QR matrix decomposition.
- Chapter 3 on calculus includes a new introduction to numerical optimization.
- Ellipses are also used in Chapter 5 to represent contours of the bivariate-normal distribution and in Chapter 7 to visualize properties of simple and multiple least-squares regression.
- Chapter 6 includes a new introduction to Markov-chain Monte Carlo (MCMC) methods for approximating probability distributions, methods that are central to modern Bayesian statistics.
- The QR and singular-value decompositions are applied in Chapter 7 to the numerically stable computation of least-squares regression coefficients.

Notation

Specific notation is introduced at various points in the text. Throughout the text, I adhere to the following general conventions, with few exceptions. [Examples are shown in brackets.]

- Known scalar constants (i.e., individual numbers, including subscripts) are represented by lowercase italic letters [a, b, x_i].
- Observable scalar random variables are represented by uppercase italic letters [X, Y_i]. Where it is necessary to make the distinction, *specific values* of random variables are represented as constants [x, y_i].
- Scalar parameters are represented by lowercase Greek letters [α, β, γ_2]. (See the Greek alphabet in Table 1.) Their estimators are generally denoted by "corresponding" italic characters [A, B, C_2], or by Greek letters with "hats" [$\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}_2$].

Table 1 The Greek Alphabet With Roman "Equivalents"

Greek Letter			*Roman Equivalent*	
Lowercase	*Uppercase*		*Phonetic*	*Other*
α	A	alpha	a	
β	B	beta	b	
γ	Γ	gamma	g, n	c
δ	Δ	delta	d	
ε	E	epsilon	e	
ζ	Z	zeta	z	
η	H	eta	e	
θ	Θ	theta	th	
ι	I	iota	i	
κ	K	kappa	k	
λ	Λ	lambda	l	
μ	M	mu	m	
ν	N	nu	n	
ξ	Ξ	xi	x	
o	O	omicron	o	
π	Π	pi	p	
ρ	P	rho	r	
σ	Σ	sigma	s	
τ	T	tau	t	
υ	Υ	upsilon	y, u	
ϕ	Φ	phi	ph	
χ	X	chi	ch	x
ψ	Ψ	psi	ps	
ω	Ω	omega	o	w

- Unobservable scalar random variables are also represented by lowercase Greek letters [ε_i].
- Vectors (one-dimensional "lists" of numbers) and matrices (rectangular tables of numbers) are represented by boldface characters—lowercase for vectors [$\mathbf{x}_1$, $\boldsymbol{\beta}$], uppercase for matrices [$\mathbf{X}$, $\boldsymbol{\Sigma}$]. In a statistical context, Roman letters are used for constants and observable random variables [$\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{X}$], and Greek letters are used for parameters and unobservable random variables [$\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\varepsilon}$]. It is occasionally convenient to show the order (number of rows and columns) of a vector or matrix below the matrix [$\underset{(n\times 1)}{\boldsymbol{\varepsilon}}$, $\underset{(n\times k+1)}{\mathbf{X}}$]. The order of an

identity matrix is given by a subscript [$\mathbf{I}_n$]. A zero matrix or vector is represented by a boldface zero [$\mathbf{0}$]; a vector of 1s is represented by a boldface $\mathbf{1}$, possibly subscripted with its number of elements [$\mathbf{1}_n$]. The transpose of a matrix is denoted by a prime [$\mathbf{X}'$], and vectors are column vectors (i.e., one-column matrices), unless they are explicitly transposed [column: $\mathbf{x}$; row: $\mathbf{x}'$].

- The symbol $\equiv$ can be read as "is defined by," or "is equal to by definition" [$\overline{X} \equiv (\sum X_i)/n$].
- The symbol $\approx$ means "is approximately equal to" [$\pi \approx 3.14159$].
- The symbol $\propto$ means "is proportional to" [$p(\alpha|D) \propto L(\alpha)p(\alpha)$].
- The symbol $\sim$ means "is distributed as" [$\varepsilon_i \sim N(0, \sigma^2)$].
- The operator $E(\ )$ denotes the expectation of a scalar, vector, or matrix random variable [$E(Y_i)$, $E(\boldsymbol{\varepsilon})$, $E(\mathbf{X})$].
- The operator $V(\ )$ denotes the variance of a scalar random variable or the variance–covariance matrix of a vector random variable [$V(\varepsilon_i)$, $V(\mathbf{b})$].
- Estimated variances or variance–covariance matrices are indicated by a circumflex ("hat") placed over the variance operator [$\widehat{V}(\varepsilon_i)$, $\widehat{V}(\mathbf{b})$].
- The operator $C(\ )$ gives the covariance of two scalar random variables or the covariance matrix of two vector random variables [$C(X, Y)$, $C(\mathbf{x}, \mathbf{y})$].
- The operators $\mathscr{E}(\ )$ and $\mathscr{V}(\ )$ denote asymptotic expectation and variance, respectively. Their usage is similar to that of $E(\ )$ and $V(\ )$ [$\mathscr{E}(B)$, $\mathscr{E}(\mathbf{b})$, $\mathscr{V}(\widehat{\boldsymbol{\beta}})$, $\widehat{\mathscr{V}}(\mathbf{b})$].
- Probability limits are specified by plim [$\text{plim}\ b = \beta$].
- Standard mathematical functions are shown in lowercase [$\cos W$ or $\cos(W)$, $\text{trace}(\mathbf{A})$]. The base of the log function is always specified explicitly [$\log_e L$, $\log_{10} X$], unless it is irrelevant [$\log 1 = 0$]. The exponential function $\exp(x)$ represents e^x.
- The summation sign $\sum$ is used to denote continued addition [$\sum_{i=1}^{n} X_i \equiv X_1 + X_2 + \cdots + X_n$]. Often, the range of the index is suppressed if it is clear from the context [$\sum_i X_i$], and the index may be suppressed as well [$\sum X_i$]. The symbol $\prod$ similarly indicates continued multiplication [$\prod_{i=1}^{n} p(Y_i) \equiv p(Y_1) \times p(Y_2) \times \cdots \times p(Y_n)$].

- The symbol ∂ denotes the partial derivative $[\partial f(x_1, x_2)/\partial x_1]$.
- To avoid awkward and repetitive phrasing in the statement of definitions and results, the words "if" and "when" are understood to mean "if and only if," unless explicitly indicated to the contrary. Terms are generally set in *italics* when they are introduced. ["Two vectors are *orthogonal* if their inner product is zero."]

Recommended Reading

The subjects addressed in this book—linear algebra, calculus, probability, and statistical theory—are larger than can be covered in depth in a 200-page book. It is my hope that the book will not only provide a basic background in these topics for students of social statistics, but also the foundation required to pursue the topics in greater depth, for example in the following sources.

There is a plethora of books on linear algebra and matrices. Most presentations develop the fundamental properties of vector spaces, but often, unfortunately, without explicit visual representation.

- Several matrix texts, including Healy (1986), Graybill (1983), Searle (1982), and Green and Carroll (1976), focus specifically on statistical applications. The last of these sources has a strongly geometric orientation.
- Davis (1965), who presents a particularly lucid and simple treatment of matrix algebra, includes some material on vector geometry (limited, however, to two dimensions).
- Namboodiri (1984) provides a compact introduction to matrix algebra (but not to vector geometry).
- Books on statistical computing, such as the classic text by Kennedy and Gentle (1980) and Monahan (2001), typically describe the implementation of matrix and linear-algebra computations on digital computers. Fieller (2016) presents a treatment of numerical matrix algebra that focuses on the *R* statistical computing environment.

There is an almost incredible profusion of introductory calculus texts, and I cannot claim to have read more than a few of them.

- Of these, my favorite brief treatment is Thompson and Gardner (1998), which was first published in the early 20th century, and which deals almost exclusively with functions of one independent variable.

- For a much more detailed introduction, including to multivariable calculus, see Stewart (2016), a very popular text that has appeared in many versions and editions.

Most more advanced treatments of calculus are either highly abstract or focus on applications in the physical sciences.

- For an extensive treatment of calculus of several variables with a social science (specifically, economic) orientation, see Binmore and Davies (2001).
- Nash (2014) provides an in-depth treatment of numerical optimization methods oriented toward the *R* statistical computing environment.

Almost any introductory text in mathematical statistics, and many econometric texts, cover probability theory, statistical distributions, and the foundations of statistical inference more formally and in greater detail than I do in this book, and there are also books that focus on each of these subjects.

- The text by Cox and Hinkley (1974) is a standard, if relatively difficult, treatment of most of the topics in Chapters 4, 5, and 6.
- A compact summary appears in Zellner (1983).
- Wonnacott and Wonnacott (1990) present insightful treatments of many of these topics at a much lower level of mathematical sophistication; I particularly recommend this source if you found the simpler parts of Chapter 4 too terse.
- A good, relatively accessible, discussion of asymptotic distribution theory appears in Theil (1971, Chapter 8).
- A general presentation of Wald, likelihood-ratio, and score tests can be found in Engle (1984).
- Lancaster (2004), Gelman and Hill (2007), and McElreath (2020) offer accessible introductions to Bayesian methods, while Gelman, Carlin, Stern, and Rubin (2013) present a more extensive treatment of the subject.
- Clear explanations of the Gibbs sampler and Hamiltonian Monte Carlo may be found in Casella and George (1992) and Neal (2011), respectively.

Website

I have prepared a website for the book, accessible at **https://tinyurl.com/Math-Primer**, including errata (if any, as they come to my attention), and a variety of materials focussed on computations using the *R* statistical computing environment. For example, I use the **matlib** package for *R* to illustrate matrix and linear-algebra computations employed in the book, such as step-by-step demonstrations of Gaussian elimination and the construction of vector diagrams.

ABOUT THE AUTHOR

John Fox is Professor Emeritus of Sociology at McMaster University in Hamilton, Ontario, Canada, where he was previously the Senator William McMaster Professor of Social Statistics. Professor Fox received a PhD in sociology from the University of Michigan in 1972 and is the author of many articles and books on statistics, including *Applied Regression Analysis and Generalized Linear Models* (3rd ed., 2016), *Using the R Commander: A Point-and-Click Interface for R* (2018), *Regression Diagnostics* (2nd ed., 2019), and, with Sanford Weisberg, *An R Companion to Applied Regression* (3rd ed., 2019). He continues to work on the development of statistical methods and their implementation in software. Professor Fox is an elected member of the R Foundation for Statistical Computing and an associate editor of the *Journal of Statistical Software*.

CHAPTER 1. MATRICES, LINEAR ALGEBRA, AND VECTOR GEOMETRY: THE BASICS

Matrices provide a natural notation for much of statistics; the algebra of linear statistical models is linear algebra; and vector geometry is a powerful conceptual tool for understanding linear algebra and for visualizing many aspects of linear models. The purpose of this chapter is to present essential concepts and results concerning matrices, linear algebra, and vector geometry. The focus is on topics that are employed widely in social statistics, and the style of presentation is informal rather than mathematically rigorous: At points, results are stated without proof; at other points, proofs are outlined; often, results are justified intuitively. Readers interested in pursuing linear algebra at greater depth might profitably make reference to one of the many available texts on the subject, each of which develops in greater detail most of the topics presented here (see, e.g., the recommended readings in the Preface, page xx).

The first section of the chapter develops elementary matrix algebra. The second and third sections introduce vector geometry and vector spaces. The final section discusses the related topics of matrix rank and the solution of linear simultaneous equations.

1.1 Matrices

1.1.1 Introducing the Actors: Definitions

A *matrix* is a rectangular table of numbers or of numerical variables;[1] for example,

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix} \tag{1.1}$$

[1] In this text, I restrict consideration to matrices composed of *real numbers*, but matrices can also have *complex numbers* as elements—that is, numbers of the form $a + bi$, where $i \equiv \sqrt{-1}$. Matrices with complex elements have few applications in statistics (e.g., in time-series analysis), although they are prominent in other fields, such as physics.

or, more generally,

$$\underset{(m\times n)}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1.2}$$

A matrix such as this with m rows and n columns is said to be of *order* m by n, written as $(m \times n)$. For clarity, I at times indicate the order of a matrix below the matrix, as in Equations 1.1 and 1.2. Each *entry* or *element* of a matrix may be subscripted by its row and column indices: a_{ij} is the entry in the ith row and jth column of the matrix $\mathbf{A}$. Individual numbers, such as the entries of a matrix, are termed *scalars*. Sometimes, for compactness, I specify a matrix by enclosing its typical element in braces; for example, $\underset{(m\times n)}{\mathbf{A}} = \{a_{ij}\}$ is equivalent to Equation 1.2.

A matrix consisting of one column is called a *column vector*; for example,

$$\underset{(m\times 1)}{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

Likewise, a matrix consisting of one row is called a *row vector*,

$$\mathbf{b}' = [b_1, b_2, \cdots, b_n]$$

In specifying a row vector, I typically place commas between its elements for clarity.

The *transpose* of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$, is formed from $\mathbf{A}$ so that the ith *row* of $\mathbf{A}'$ consists of the elements of the ith *column* of $\mathbf{A}$;[2] thus (using the

[2] Although in this book I'll consistently use a prime, as in $\mathbf{A}'$, to denote the matrix transpose, it's also common to use a superscript T, as in $\mathbf{A}^T$.

matrices in Equations 1.1 and 1.2),

$$\underset{(3\times 4)}{\mathbf{X}'} = \begin{bmatrix} 1 & 4 & 7 & 0 \\ -2 & -5 & 8 & 0 \\ 3 & -6 & 9 & 10 \end{bmatrix}$$

$$\underset{(n\times m)}{\mathbf{A}'} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

The transpose of the transpose is the original matrix: $(\mathbf{A}')' = \mathbf{A}$. I adopt the common convention that a vector is a column vector (such as $\mathbf{a}$ above) unless it is explicitly transposed (such as $\mathbf{b}'$).

A *square matrix of order n*, as the term implies, has n rows and n columns. The entries a_{ii} (i.e., $a_{11}, a_{22}, \ldots, a_{nn}$) of a square matrix $\mathbf{A}$ comprise the *main diagonal* of the matrix. The sum of the diagonal elements is the *trace* of the matrix:

$$\text{trace}(A) \equiv \sum_{i=1}^{n} a_{ii}$$

For example, the square matrix

$$\underset{(3\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 3 \\ 2 & 2 & 6 \\ 7 & 3 & -4 \end{bmatrix}$$

has diagonal elements, $-5, 2$, and -4, and $\text{trace}(\mathbf{B}) = \sum_{i=1}^{3} b_{ii} = -5 + 2 - 4 = -7$.

A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}'$, that is, when $a_{ij} = a_{ji}$ for all i and j. Consequently, the matrix $\mathbf{B}$ (above) is not symmetric, while the matrix

$$\mathbf{C} = \begin{bmatrix} -5 & 1 & 3 \\ 1 & 2 & 6 \\ 3 & 6 & -4 \end{bmatrix}$$

is symmetric. Many matrices that appear in statistical applications are symmetric—for example, correlation matrices, covariance matrices, and matrices of sums of squares and cross products.

An *upper-triangular matrix* is a square matrix with 0s below its main diagonal:

$$\underset{(n\times n)}{\mathbf{U}} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

To be clear, some of the elements on and above the main diagonal of $\mathbf{U}$ *may* be 0, but all of the elements below the diagonal *are* 0. Similarly, a *lower-triangular matrix* is a square matrix of the form

$$\underset{(n\times n)}{\mathbf{L}} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

A square matrix is *diagonal* if all entries except those on its main diagonal are 0; thus,

$$\underset{(n\times n)}{\mathbf{D}} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

For compactness, I may write $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$. A *scalar matrix* is a diagonal matrix all of whose diagonal entries are equal: $\mathbf{S} = \text{diag}(s, s, \ldots, s)$. An especially important family of scalar matrices are the *identity matrices* $\mathbf{I}$, which have 1s on the main diagonal:

$$\underset{(n\times n)}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

I write $\mathbf{I}_n$ for $\underset{(n\times n)}{\mathbf{I}}$.

Two other special matrices are the family of *zero matrices*, $\mathbf{0}$, all of whose entries are 0, and the $\mathbf{1}$ vectors, all of whose entries are 1. I write $\mathbf{1}_n$ for the column vector of 1s with n entries; for example, $\mathbf{1}_4 = [1, 1, 1, 1]'$. Although the identity matrices, the zero matrices, and the $\mathbf{1}$ vectors are *families* of matrices, it is often convenient to refer to these matrices in the singular, for example, to "*the* identity matrix.'

A *partitioned matrix* is a matrix whose elements are organized into *sub-matrices*; for example,

$$\underset{(4\times 3)}{\mathbf{A}} = \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \hline a_{41} & a_{42} & a_{43} \end{array}\right] = \left[\begin{array}{c|c} \underset{(3\times 2)}{\mathbf{A}_{11}} & \underset{(3\times 1)}{\mathbf{A}_{12}} \\ \hline \underset{(1\times 2)}{\mathbf{A}_{21}} & \underset{(1\times 1)}{\mathbf{A}_{22}} \end{array}\right]$$

where the submatrix

$$\mathbf{A}_{11} \equiv \begin{bmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

and $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are similarly defined. When there is no possibility of confusion, I omit the lines separating the submatrices. If a matrix is partitioned vertically but not horizontally, then I separate its submatrices by commas; for example, $\underset{(m\times n+p)}{\mathbf{C}} = \left[\underset{(m\times n)}{\mathbf{C}_1}, \underset{(m\times p)}{\mathbf{C}_2}\right]$.

1.1.2 Simple Matrix Arithmetic

Two matrices are *equal* if they are of the same order and all corresponding entries are equal (a definition used implicitly in the preceding section).

Two matrices may be *added* only if they are of the same order; then their sum is formed by adding corresponding elements. Thus, if $\mathbf{A}$ and $\mathbf{B}$ are of order $(m \times n)$, then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is also of order $(m \times n)$, with $c_{ij} = a_{ij} + b_{ij}$. Likewise, if $\mathbf{D} = \mathbf{A} - \mathbf{B}$, then $\mathbf{D}$ is of order $(m \times n)$, with $d_{ij} = a_{ij} - b_{ij}$. The *negative* of a matrix $\mathbf{A}$, that is, $\mathbf{E} = -\mathbf{A}$, is of the same order as $\mathbf{A}$, with elements $e_{ij} = -a_{ij}$. For example, for matrices

$$\underset{(2\times 3)}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

and

$$\underset{(2\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 2 \\ 3 & 0 & -4 \end{bmatrix}$$

we have

$$\underset{(2\times 3)}{\mathbf{C}} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} -4 & 3 & 5 \\ 7 & 5 & 2 \end{bmatrix}$$

$$\underset{(2\times 3)}{\mathbf{D}} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 5 & 10 \end{bmatrix}$$

$$\underset{(2\times 3)}{\mathbf{E}} = -\mathbf{B} = \begin{bmatrix} 5 & -1 & -2 \\ -3 & 0 & 4 \end{bmatrix}$$

Because they are element-wise operations, matrix addition, subtraction, and negation follow essentially the same rules as the corresponding scalar

arithmetic operations; in particular,

$$\begin{aligned}
\mathbf{A}+\mathbf{B} &= \mathbf{B}+\mathbf{A} \text{ (matrix addition is commutative)}\\
\mathbf{A}+(\mathbf{B}+\mathbf{C}) &= (\mathbf{A}+\mathbf{B})+\mathbf{C} \text{ (matrix addition is associative)}\\
\mathbf{A}-\mathbf{B} &= \mathbf{A}+(-\mathbf{B}) = -(\mathbf{B}-\mathbf{A})\\
\mathbf{A}-\mathbf{A} &= \mathbf{0}\\
\mathbf{A}+\mathbf{0} &= \mathbf{A}\\
-(-\mathbf{A}) &= \mathbf{A}\\
(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}'
\end{aligned}$$

The *product* of a scalar c and an $(m \times n)$ matrix $\mathbf{A}$ is an $(m \times n)$ matrix $\mathbf{B} = c\mathbf{A}$ in which $b_{ij} = ca_{ij}$. Continuing the preceding examples:

$$\underset{(2\times 3)}{\mathbf{F}} = 3 \times \mathbf{B} = \begin{bmatrix} -15 & 3 & 6 \\ 9 & 0 & -12 \end{bmatrix}$$

The product of a scalar and a matrix obeys the following rules:

$$\begin{aligned}
c\mathbf{A} &= \mathbf{A}c \text{ (commutative)}\\
\mathbf{A}(b+c) &= \mathbf{A}b+\mathbf{A}c \text{ (distributes over scalar addition)}\\
c(\mathbf{A}+\mathbf{B}) &= c\mathbf{A}+c\mathbf{B} \text{ (distributes over matrix addition)}\\
0\mathbf{A} &= \mathbf{0}\\
1\mathbf{A} &= \mathbf{A}\\
(-1)\mathbf{A} &= -\mathbf{A}
\end{aligned}$$

where $b, c, 0, 1$, and -1 are scalars, and $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{0}$ are matrices of the same order.

The *inner product* (or *dot product*) of two vectors (each with n entries), say $\underset{(1\times n)}{\mathbf{a}'}$ and $\underset{(n\times 1)}{\mathbf{b}}$, denoted $\mathbf{a}' \cdot \mathbf{b}$, is a scalar formed by multiplying corresponding entries of the vectors and summing the resulting products:

$$\mathbf{a}' \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

For example,

$$[2, 0, 1, 3] \cdot \begin{bmatrix} -1 \\ 6 \\ 0 \\ 9 \end{bmatrix} = 2(-1) + 0(6) + 1(0) + 3(9) = 25$$

Although this example is for the inner product of a row vector with a column vector, both vectors may be row vectors or both column vectors, as long as the two vectors have the same number of elements.

Two matrices **A** and **B** are *conformable for multiplication* in the order given (i.e., **AB**) if the number of *columns* of the left-hand factor (**A**) is equal to the number of *rows* of the right-hand factor (**B**). Thus **A** and **B** are conformable for multiplication if **A** is of order $(m \times n)$ and **B** is of order $(n \times p)$, where m and p are unconstrained. For example,

$$\underset{(2\times 3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times 3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \tag{1.3}$$

are conformable for multiplication but

$$\underset{(3\times 3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \underset{(2\times 3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \tag{1.4}$$

are not.

Let $\mathbf{C} = \mathbf{AB}$ be the *matrix product*; and let $\mathbf{a}_i'$ represent the ith *row* of **A** and $\mathbf{b}_j$ represent the jth *column* of **B**. Then **C** is a matrix of order $(m \times p)$ in which

$$c_{ij} = \mathbf{a}_i' \cdot \mathbf{b}_j = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Here are some examples:

$$\underset{(2\times 3)}{\begin{bmatrix} & \Longrightarrow & \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times 3)}{\begin{bmatrix} 1 & 0 & 0 \\ \Downarrow & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{bmatrix}}$$

$$= \underset{(2\times 3)}{\begin{bmatrix} 1(1)+2(0)+3(0), & 1(0)+2(1)+3(0), & 1(0)+2(0)+3(1) \\ 4(1)+5(0)+6(0), & 4(0)+5(1)+6(0), & 4(0)+5(0)+6(1) \end{bmatrix}}$$

$$= \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$$\underset{(1\times 4)}{[\beta_0, \beta_1, \beta_2, \beta_3]} \underset{(4\times 1)}{\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \underset{(1\times 1)}{[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3]}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 8 & 13 \end{bmatrix}$$
$$\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 5 & 8 \end{bmatrix} \tag{1.5}$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.6}$$

In the first of these examples, the arrows indicate how the rows of the left-hand factor are multiplied into the columns of the right-hand factor.

Matrix multiplication is associative, $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$, and distributive with respect to addition:

$$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$
$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$$

but it is not in general commutative: If $\mathbf{A}$ is $(m \times n)$ and $\mathbf{B}$ is $(n \times p)$, then the product $\mathbf{AB}$ is defined but $\mathbf{BA}$ is defined only if $m = p$ (cf., e.g., the matrices in 1.3 and 1.4 above). Even so, $\mathbf{AB}$ and $\mathbf{BA}$ are of different orders (and hence are not candidates for equality) unless $m = p$. And even if $\mathbf{A}$ and $\mathbf{B}$ are square, $\mathbf{AB}$ and $\mathbf{BA}$, though of the same order, are not necessarily equal (as illustrated in Equation 1.5). If it is the case that $\mathbf{AB} = \mathbf{BA}$ (as in Equation 1.6), then the matrices $\mathbf{A}$ and $\mathbf{B}$ are said to *commute* with one another. A scalar factor, however, may be moved anywhere within a matrix product: $c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$.

The identity and zero matrices play roles with respect to matrix multiplication analogous to those of the numbers 1 and 0 in scalar algebra:

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{}{\mathbf{I}_n} = \underset{}{\mathbf{I}_m}\ \underset{(m\times n)}{\mathbf{A}} = \mathbf{A}$$
$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times p)}{\mathbf{0}} = \underset{(m\times p)}{\mathbf{0}}$$
$$\underset{(q\times m)}{\mathbf{0}}\ \underset{(m\times n)}{\mathbf{A}} = \underset{(q\times n)}{\mathbf{0}}$$

A further property of matrix multiplication, which has no analog in scalar algebra, is that $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$—the transpose of a product is the product of

the transposes taken in the opposite order, a rule that extends to the product of several (conformable) matrices:

$$(\mathbf{AB}\cdots\mathbf{F})' = \mathbf{F}'\cdots\mathbf{B}'\mathbf{A}'$$

The *powers* of a square matrix are the products of the matrix with itself. That is, $\mathbf{A}^2 = \mathbf{AA}$, $\mathbf{A}^3 = \mathbf{AAA} = \mathbf{AA}^2 = \mathbf{A}^2\mathbf{A}$, and so on. If $\mathbf{B}^2 = \mathbf{A}$, then we call $\mathbf{B}$ a *square root* of $\mathbf{A}$, which we may write as $\mathbf{A}^{1/2}$. Unlike in scalar algebra, however, the square root of a matrix is not generally unique. Of course, even the scalar square root is unique only up to a change in sign: For example, $\sqrt{4} = \pm 2$.[3] If $\mathbf{A}^2 = \mathbf{A}$, then $\mathbf{A}$ is said to be *idempotent*. As in scalar algebra, and by convention, $\mathbf{A}^0 = \mathbf{I}$ (where the identity matrix $\mathbf{I}$ is of the same order as $\mathbf{A}$). The matrix inverse $\mathbf{A}^{-1}$ is discussed later in the chapter (Section 1.1.3), and is *not* $\{1/a_{ij}\}$.

For purposes of matrix addition, subtraction, and multiplication, the submatrices of partitioned matrices may be treated as if they were elements, as long as the factors are partitioned conformably. For example, if

$$\mathbf{A} = \left[\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{array}\right] = \left[\begin{array}{cc} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right]$$

and

$$\mathbf{B} = \left[\begin{array}{ccc|cc} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} \\ \hline b_{31} & b_{32} & b_{33} & b_{34} & b_{35} \end{array}\right] = \left[\begin{array}{cc} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right]$$

then

$$\mathbf{A}+\mathbf{B} = \left[\begin{array}{c|c} \mathbf{A}_{11}+\mathbf{B}_{11} & \mathbf{A}_{12}+\mathbf{B}_{12} \\ \hline \mathbf{A}_{21}+\mathbf{B}_{21} & \mathbf{A}_{22}+\mathbf{B}_{22} \end{array}\right]$$

Similarly, if

$$\underset{(m+n\times p+q)}{\mathbf{A}} = \left[\begin{array}{cc} \underset{(m\times p)}{\mathbf{A}_{11}} & \underset{(m\times q)}{\mathbf{A}_{12}} \\ \underset{(n\times p)}{\mathbf{A}_{21}} & \underset{(n\times q)}{\mathbf{A}_{22}} \end{array}\right]$$

and

$$\underset{(p+q\times r+s)}{\mathbf{B}} = \left[\begin{array}{cc} \underset{(p\times r)}{\mathbf{B}_{11}} & \underset{(p\times s)}{\mathbf{B}_{12}} \\ \underset{(q\times r)}{\mathbf{B}_{21}} & \underset{(q\times s)}{\mathbf{B}_{22}} \end{array}\right]$$

[3] For another kind of matrix square root, see the discussion of the Cholesky decomposition in Section 2.2.2.

then

$$\underset{(m+n\times r+s)}{\mathbf{AB}} = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11}+\mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12}+\mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right]$$

The Sense Behind Matrix Multiplication

The definition of matrix multiplication makes it simple to formulate systems of scalar equations as a single matrix equation, often providing a useful level of abstraction. For example, consider the following system of two linear equations in two unknowns, x_1 and x_2:

$$\begin{aligned} 2x_1 + 5x_2 &= 4 \\ x_1 + 3x_2 &= 5 \end{aligned}$$

These equations are linear because each additive term in the equation is either a constant (e.g., 4 on the right-hand side of the first equation) or the product of a constant and a variable (e.g., $2x_1$ on the left-hand side of the first equation). Each of the equations $2x_1 + 5x_2 = 4$ and $x_1 + 3x_2 = 5$ literally represents a line in two-dimensional (2D) coordinate space (see the review of the equations of lines and planes in Section 3.1.2). Writing the two scalar equations as a matrix equation,

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$\underset{(2\times 2)}{\mathbf{A}} \; \underset{(2\times 1)}{\mathbf{x}} = \underset{(2\times 1)}{\mathbf{b}}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

The formulation and solution of systems of linear simultaneous equations is taken up subsequently (Section 1.4.2).

1.1.3 Matrix Inverses

In scalar algebra, division is essential to the solution of simple equations. For example,

$$6x = 12$$
$$x = \frac{12}{6} = 2$$

or, equivalently,

$$\frac{1}{6} \times 6x = \frac{1}{6} \times 12$$
$$x = 2$$

where $\frac{1}{6} = 6^{-1}$ is the scalar inverse of 6.

In matrix algebra, there is no direct analog of division, but most square matrices have a *matrix inverse*. The inverse of a square matrix $\mathbf{A}$ is a square matrix of the same order, written $\mathbf{A}^{-1}$, with the property that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.[4] If a square matrix has an inverse, then the matrix is termed *nonsingular*; a square matrix without an inverse is termed *singular*.[5] If the inverse of a matrix exists, then it is unique; moreover, if for a square matrix $\mathbf{A}$, $\mathbf{AB} = \mathbf{I}$, then necessarily $\mathbf{BA} = \mathbf{I}$, and thus $\mathbf{B} = \mathbf{A}^{-1}$.

For example, the inverse of the nonsingular matrix

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

is the matrix

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}$$

as we can readily verify:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$
$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$

[4] As I will explain (Section 1.4.3), it is also possible to define *generalized inverses* for rectangular matrices and for square matrices that do not have conventional inverses.

[5] When mathematicians first encountered nonzero matrices without inverses, they found the existence of such matrices remarkable or "singular."

In scalar algebra, only the number 0 has no inverse. It is simple to show by example that there exist singular *nonzero* matrices: Let us hypothesize that $\mathbf{B}$ is the inverse of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

But

$$\mathbf{AB} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix} \neq \mathbf{I}_2$$

which contradicts the hypothesis, and $\mathbf{A}$ consequently has no inverse.

There are many methods for finding the inverse of a nonsingular square matrix. I will briefly and informally describe a procedure called *Gaussian elimination* (after the great German mathematician, Carl Friedrich Gauss, 1777–1855). Although there are methods that tend to produce more accurate numerical results when implemented on a digital computer, elimination has the virtue of relative simplicity, and has applications beyond matrix inversion (as we will see later in this chapter).

To illustrate the method of elimination, I will employ the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \tag{1.7}$$

Let us begin by adjoining to this matrix an identity matrix; that is, form the partitioned or *augmented* matrix

$$[\mathbf{A}, \mathbf{I}_3] = \left[\begin{array}{ccc|ccc} 2 & -2 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

Then let's attempt to reduce the original matrix to an identity matrix by applying operations of three sorts:

E_{I}: Multiply each entry in a row of the matrix by a nonzero scalar constant.

E_{II}: Add a scalar multiple of one row to another, replacing the other row.

E_{III}: Exchange two rows of the matrix.

E_{I}, E_{II}, and E_{III} are called *elementary row operations*.

Starting with the first row, and dealing with each row in turn, ensure that there is a nonzero entry in the diagonal position, employing a row interchange for a lower row if necessary. Then divide the row through by its diagonal element (called the *pivot*) to obtain an entry of 1 in the diagonal position. Finally, add multiples of the current row to the other rows so as to *"sweep out"* the nonzero elements in the pivot column. For the illustration:

1. Divide Row 1 by 2,
$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$
2. Subtract the new Row 1 from Row 2,
$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$
3. Subtract $4 \times$ Row 1 from Row 3,
$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \end{array}\right]$$
4. Move to Row 2; there is a 0 entry in Row 2, Column 2, so interchange Rows 2 and 3,
$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$
5. Divide Row 2 by 8,
$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$
6. Add Row 2 to Row 1,
$$\left[\begin{array}{ccc|ccc} 1 & 0 & -\frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

7. Move to Row 3; there is already a 1 in the pivot position; add $\frac{1}{2}\times$ Row 3 to Row 1,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{8} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

8. Add $\frac{1}{2}\times$ Row 3 to Row 2,

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{8} \\ 0 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

Once the original matrix is reduced to the identity matrix, the final columns of the augmented matrix contain the inverse, as we can verify for the example:

$$\left[\begin{array}{ccc} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{array}\right] \left[\begin{array}{ccc} 0 & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & 1 & 0 \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \checkmark$$

It is simple to explain why Gaussian elimination works: Each elementary row operation can be represented as multiplication on the left by an appropriately formulated square matrix. Thus, for example, to interchange the second and third rows, we can multiply on the left by[6]

$$\mathbf{E}_{\text{III}} \equiv \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right]$$

The elimination procedure applies a sequence of (say p) elementary row operations to the augmented matrix $[\underset{(n\times n)}{\mathbf{A}}, \mathbf{I}_n]$, which we can then write as

$$\mathbf{E}_p \cdots \mathbf{E}_2 \mathbf{E}_1 [\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$$

using $\mathbf{E}_i$ to represent the ith operation in the sequence. Defining $\mathbf{E} \equiv \mathbf{E}_p \cdots \mathbf{E}_2 \mathbf{E}_1$, we have $\mathbf{E}[\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$; that is, $\mathbf{EA} = \mathbf{I}_n$ (implying that $\mathbf{E} = \mathbf{A}^{-1}$),

[6] *Reader:* Show how Types I (e.g., Step 1 in the example) and II (e.g., Step 3 in the example) elementary row operations can also be represented as multiplication on the left by suitably formulated square matrices, say $\mathbf{E}_{\text{I}}$ and $\mathbf{E}_{\text{II}}$.

and $\mathbf{E}\mathbf{I}_n = \mathbf{B}$. Consequently, $\mathbf{B} = \mathbf{E} = \mathbf{A}^{-1}$. If $\mathbf{A}$ is singular, then it cannot be reduced to $\mathbf{I}$ by elementary row operations: At some point in the process, we will find that no nonzero pivot is available.

The matrix inverse obeys the following rules:

$$\begin{aligned}\mathbf{I}^{-1} &= \mathbf{I}\\ (\mathbf{A}^{-1})^{-1} &= \mathbf{A}\\ (\mathbf{A}')^{-1} &= (\mathbf{A}^{-1})'\\ (\mathbf{A}\mathbf{B})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1}\\ (c\mathbf{A})^{-1} &= c^{-1}\mathbf{A}^{-1}\end{aligned}$$

(where $\mathbf{A}$ and $\mathbf{B}$ are order-n nonsingular matrices, and c is a nonzero scalar). If $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$, and if all $d_i \neq 0$, then $\mathbf{D}$ is nonsingular and $\mathbf{D}^{-1} = \text{diag}(1/d_1, 1/d_2, \ldots, 1/d_n)$; if any of the d_i are 0, then $\mathbf{D}$ is singular. Finally, the inverse of a nonsingular symmetric matrix is itself symmetric.

1.1.4 Determinants

Each square matrix $\mathbf{A}$ is associated with a scalar called its *determinant*, written as $\det \mathbf{A}$.[7] For a (2×2) matrix $\mathbf{A}$, the determinant is $\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$. For a (3×3) matrix $\mathbf{A}$, the determinant is

$$\begin{aligned}\det \mathbf{A} =& a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31}\\ &- a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}\end{aligned}$$

Although there is a general definition of the determinant of a square matrix of order n, I find it simpler to define the determinant implicitly by specifying the following properties (or *axioms*):

D1: Multiplying a row of a square matrix by a scalar constant multiplies the determinant of the matrix by the same constant.

D2: Adding a multiple of one row to another leaves the determinant unaltered.

D3: Interchanging two rows changes the sign of the determinant.

D4: $\det \mathbf{I} = 1$.

[7] A common alternative notation for $\det \mathbf{A}$ is $|\mathbf{A}|$.

Axioms D1, D2, and D3 specify the effects on the determinant of the three kinds of elementary row operations. Because the Gaussian elimination method described previously reduces a square matrix to the identity matrix, these properties, along with axiom D4, are sufficient for establishing the value of the determinant. Indeed, the determinant is simply the product of the pivot elements, with the sign of the product reversed if, in the course of elimination, an odd number of row interchanges is employed. For the illustrative matrix $\mathbf{A}$ in Equation 1.7 (on page 12), then, the determinant is $-(2)(8)(1) = -16$, because there was one row interchange (in Step 4) and the pivots were 2, 8, and 1 (Steps 1, 5, and 7). If a matrix is singular, then one or more of the pivots are zero, and the determinant is zero. Conversely, a nonsingular matrix has a nonzero determinant.

Some additional properties of determinants (for order-n square matrices $\mathbf{A}$ and $\mathbf{B}$) are as follows:

- $\det \mathbf{A}' = \det \mathbf{A}$.
- $\det(\mathbf{AB}) = \det \mathbf{A} \times \det \mathbf{B}$.
- If $\mathbf{A}$ is nonsingular, then $\det \mathbf{A}^{-1} = 1/\det \mathbf{A}$.
- If $\mathbf{A}$ is idempotent (recall, $\mathbf{A}^2 = \mathbf{A}$), then $\det \mathbf{A} = 1$ if $\mathbf{A}$ is nonsingular or 0 if it is singular.

The third result follows from second, along with the observations that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$ and $\det \mathbf{I}_n = 1$. The fourth result also follows from the second. (*Reader:* Can you see why?)

In addition to their useful algebraic properties, determinants occasionally appear directly in statistical applications—for example, in the formula for the multivariate-normal distribution (see Section 5.2.5).

1.1.5 The Kronecker Product

Suppose that $\mathbf{A}$ is an $m \times n$ matrix and that $\mathbf{B}$ is a $p \times q$ matrix. Then the *Kronecker product* (named after the 19th-century German mathematician Leopold Kronecker) of $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} \otimes \mathbf{B}$, is defined as

$$\underset{(mp \times nq)}{\mathbf{A} \otimes \mathbf{B}} \equiv \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

The Kronecker product is sometimes useful in statistics for compactly representing patterned matrices. For example, suppose that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

is a (2×2) variance–covariance matrix (see Section 4.2.4). Then,

$$\begin{aligned} \mathbf{I}_3 \otimes \boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \\ &= \left[\begin{array}{cc|cc|cc} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{array}\right] \end{aligned}$$

Such expressions arise naturally, for example, in multivariate statistics.

Many of the properties of the Kronecker product are similar to those of ordinary matrix multiplication; in particular,

$$\begin{aligned} \mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) &= \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C} \\ (\mathbf{B} + \mathbf{C}) \otimes \mathbf{A} &= \mathbf{B} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{A} \\ (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{D} &= \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{D}) \\ c(\mathbf{A} \otimes \mathbf{B}) &= (c\mathbf{A}) \otimes \mathbf{B} = \mathbf{A} \otimes (c\mathbf{B}) \end{aligned}$$

where $\mathbf{B}$ and $\mathbf{C}$ are matrices of the same order, and c is a scalar. As well, like matrix multiplication, the Kronecker product is not commutative: In general, $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$. Additionally, for matrices $\underset{(m\times n)}{\mathbf{A}}$, $\underset{(p\times q)}{\mathbf{B}}$, $\underset{(n\times r)}{\mathbf{C}}$, and $\underset{(q\times s)}{\mathbf{D}}$,

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$$

Consequently, if $\underset{(n\times n)}{\mathbf{A}}$ and $\underset{(m\times m)}{\mathbf{B}}$ are nonsingular matrices, then

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$$

because

$$(\mathbf{A} \otimes \mathbf{B})\left(\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}\right) = (\mathbf{A}\mathbf{A}^{-1}) \otimes (\mathbf{B}\mathbf{B}^{-1}) = \mathbf{I}_n \otimes \mathbf{I}_m = \mathbf{I}_{(nm\times nm)}$$

Finally, for any matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$$

and for square matrices $\mathbf{A}$ and $\mathbf{B}$ of order m and n, respectively,

$$\begin{aligned} \text{trace}(\mathbf{A} \otimes \mathbf{B}) &= \text{trace}(\mathbf{A}) \times \text{trace}(\mathbf{B}) \\ \det(\mathbf{A} \otimes \mathbf{B}) &= (\det \mathbf{A})^m (\det \mathbf{B})^n \end{aligned}$$

1.2 Basic Vector Geometry

Considered algebraically, vectors are one-column (or one-row) matrices. Vectors also have the following geometric interpretation: The vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]'$ is represented as a directed line segment extending from the origin of an n-dimensional coordinate space to the point defined by the entries (called the *coordinates*) of the vector. Some examples of geometric vectors in 2D and 3D space are shown in Figure 1.1.

The basic arithmetic operations defined for vectors have simple geometric interpretations. To add two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is, in effect, to place the "tail" of one at the tip of the other. When a vector is shifted from the origin in this manner, it retains its length and orientation (the angles that it makes with respect to the coordinate axes); length and orientation serve to define a vector uniquely. The operation of vector addition, illustrated in two dimensions in Figure 1.2, is equivalent to completing a parallelogram in which $\mathbf{x}_1$ and $\mathbf{x}_2$ are two adjacent sides; the vector sum is the diagonal of the parallelogram, starting at the origin.

As shown in Figure 1.3, the difference $\mathbf{x}_1 - \mathbf{x}_2$ is a vector whose length and orientation are obtained by proceeding from the tip of $\mathbf{x}_2$ to the tip of $\mathbf{x}_1$. Likewise, $\mathbf{x}_2 - \mathbf{x}_1$ proceeds from $\mathbf{x}_1$ to $\mathbf{x}_2$.

The *length* of a vector $\mathbf{x}$, denoted by $||\mathbf{x}||$, is the square root of its sum of squared coordinates:

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

This result follows from the Pythagorean theorem in two dimensions,[8] as shown in Figure 1.4(a). The result can be extended one dimension at a time

[8] Recall that the Pythagorean theorem (named after the ancient Greek mathematician Pythagoras) states that the squared length of the hypotenuse (side opposite the right angle) in a right triangle is equal to the sums of squared lengths of the other two sides of the triangle.

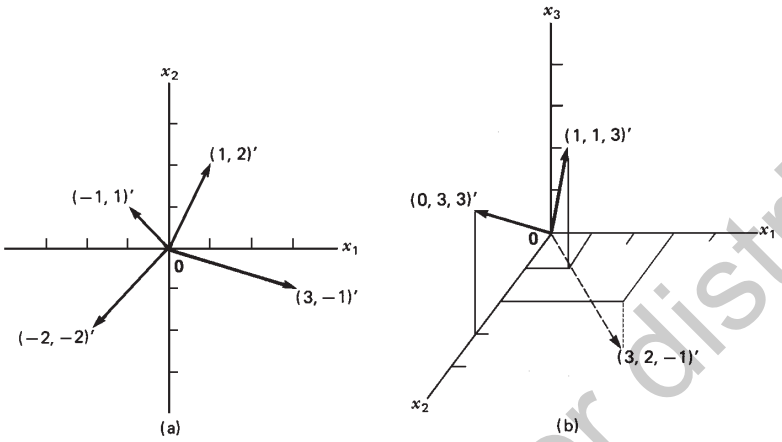

Figure 1.1 Examples of geometric vectors in (a) two-dimensional and (b) three-dimensional space. Each vector is a directed line segment from the origin ($\mathbf{0} = [0,0]'$ in two dimensions or $\mathbf{0} = [0,0,0]'$ in three dimensions) to the point whose coordinates are given by the entries of the vector.

to higher-dimensional coordinate spaces, as shown for a 3D space in Figure 1.4(b). The *distance* between two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, defined as the distance separating their tips, is given by $||\mathbf{x}_1 - \mathbf{x}_2|| = ||\mathbf{x}_2 - \mathbf{x}_1||$ (see Figure 1.3).

The product $a\mathbf{x}$ of a scalar a and a vector $\mathbf{x}$ is a vector of length $|a| \times ||\mathbf{x}||$, as is readily verified:

$$\begin{aligned} ||a\mathbf{x}|| &= \sqrt{\sum (ax_i)^2} \\ &= \sqrt{a^2 \sum x_i^2} \\ &= |a| \times ||\mathbf{x}|| \end{aligned}$$

If the scalar a is positive, then the orientation of $a\mathbf{x}$ is the same as that of $\mathbf{x}$; if a is negative, then $a\mathbf{x}$ is *collinear* with (i.e., along the same line as) $\mathbf{x}$ but in the opposite direction. The negative $-\mathbf{x} = (-1)\mathbf{x}$ of $\mathbf{x}$ is, therefore, a vector of the same length as $\mathbf{x}$ but of opposite orientation. These results are illustrated for two dimensions in Figure 1.5.

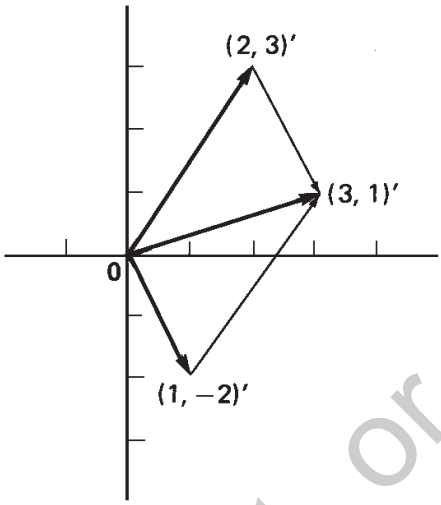

Figure 1.2 Vectors are added by placing the "tail" of one on the tip of the other and completing the parallelogram. The sum is the diagonal of the parallelogram starting at the origin.

1.3 Vector Spaces and Subspaces

The *vector space of dimension n* is the infinite set of all vectors $\mathbf{x} = [x_1, x_2, \ldots, x_n]'$; the coordinates x_i may be any real numbers. The vector space of dimension one is, therefore, the real line; the vector space of dimension two is the plane; and so on.

The *subspace* of the n-dimensional vector space that is *generated* by a set of k vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is the subset of vectors $\mathbf{y}$ in the space that can be expressed as linear combinations of the generating set:

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$$

The set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is said to *span* the subspace that it generates. Notice that each of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is a vector, with n coordinates; that is, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a set of k vectors, *not* a vector with k coordinates.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is *linearly independent* if no vector in the set can be expressed as a linear combination of other vectors:

$$\mathbf{x}_j = a_1\mathbf{x}_1 + \cdots + a_{j-1}\mathbf{x}_{j-1} + a_{j+1}\mathbf{x}_{j+1} + \cdots + a_k\mathbf{x}_k \qquad (1.8)$$

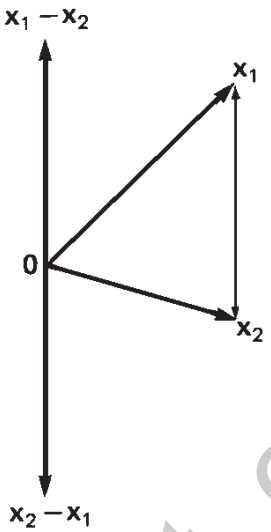

Figure 1.3 Vector differences $\mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{x}_2 - \mathbf{x}_1$.

(where some of the constants a_l can be zero). Equivalently, the set of vectors is linearly independent if there are no constants $b_1, b_2, \ldots, b_k$, not all zero, for which

$$b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k = \underset{(n\times 1)}{\mathbf{0}} \tag{1.9}$$

Equation 1.8 or 1.9 is called a *linear dependency* or *collinearity*. If these equations hold, then the vectors comprise a *linearly dependent* set. By Equation 1.8, the zero vector is linearly dependent on every other vector, inasmuch as $\mathbf{0} = 0\mathbf{x}$.

The *dimension* of the subspace spanned by a set of vectors is the number of vectors in its largest linearly independent subset. The dimension of the subspace spanned by $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ cannot, therefore, exceed the smaller of k and n. These relations are illustrated for a vector space of dimension $n = 3$ in Figure 1.6. Figure 1.6(a) shows the 1D subspace (i.e., the line) generated by a single nonzero vector $\mathbf{x}$; Figure 1.6(b) shows the 1D subspace generated by two collinear vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; Figure 1.6(c) shows the 2D subspace (the plane) generated by two linearly independent vectors $\mathbf{x}_1$ and $\mathbf{x}_2$; and Figure 1.6(d) shows the plane generated by three linearly dependent vectors $\mathbf{x}_1$, $\mathbf{x}_2$,

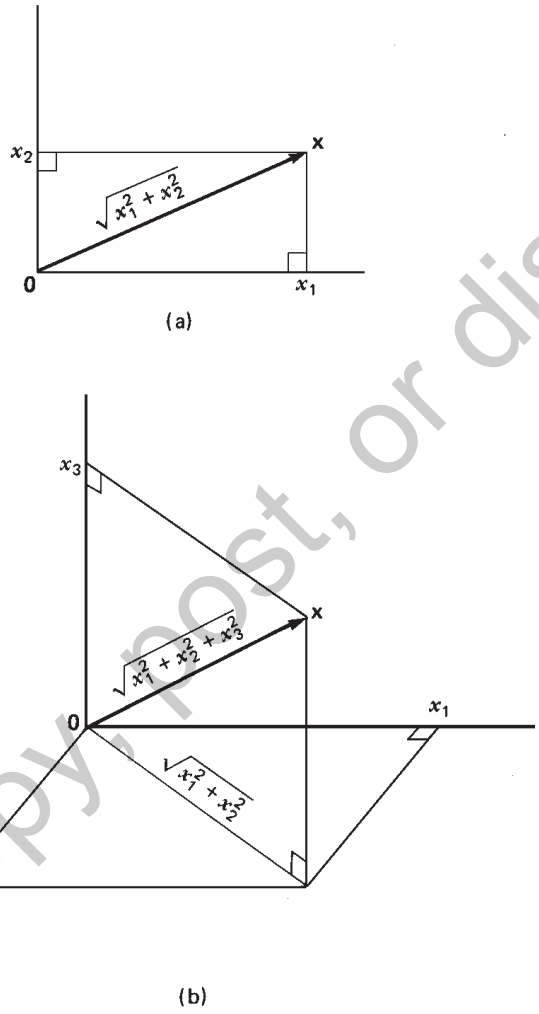

Figure 1.4 The length of a vector is the square root of its sum of squared coordinates, $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$. This result is illustrated in (a) two and (b) three dimensions.

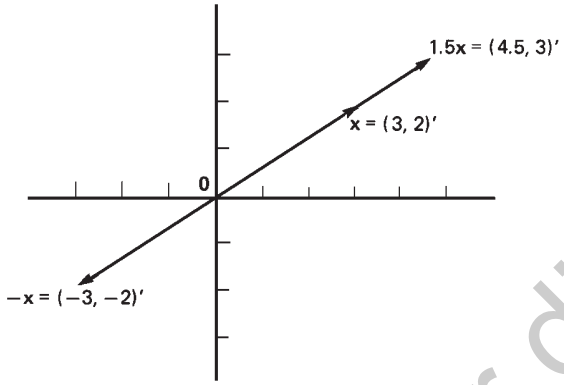

Figure 1.5 Product $a\mathbf{x}$ of a scalar and a vector, illustrated in two dimensions. The vector $a\mathbf{x}$ is collinear with $\mathbf{x}$; it is in the same direction as $\mathbf{x}$ if $a > 0$, and in the opposite direction from $\mathbf{x}$ if $a < 0$.

and $\mathbf{x}_3$, no two of which are collinear. In this last case, any one of the three vectors lies in the plane generated by the other two.

A linearly independent set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$—such as $\{\mathbf{x}\}$ in Figure 1.6(a), $\{\mathbf{x}_1, \mathbf{x}_2\}$ in Figure 1.6(c), or (say) $\{\mathbf{x}_1, \mathbf{x}_3\}$ in Figure 1.6(d)—is said to provide a *basis* for the subspace that it spans. (*Reader:* What about Figure 1.6(b)?) Any vector $\mathbf{y}$ in this subspace can be written *uniquely* as a linear combination of the basis vectors:

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k$$

The constants $c_1, c_2, \ldots, c_k$ are called the *coordinates of* $\mathbf{y}$ *with respect to the basis* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. Because $\mathbf{0} = 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 0\mathbf{x}_k$, the zero vector is included in every subspace.

The coordinates of a vector with respect to a basis for a 2D subspace can be found geometrically by the parallelogram rule of vector addition, as illustrated in Figure 1.7. Finding coordinates algebraically entails the solution of

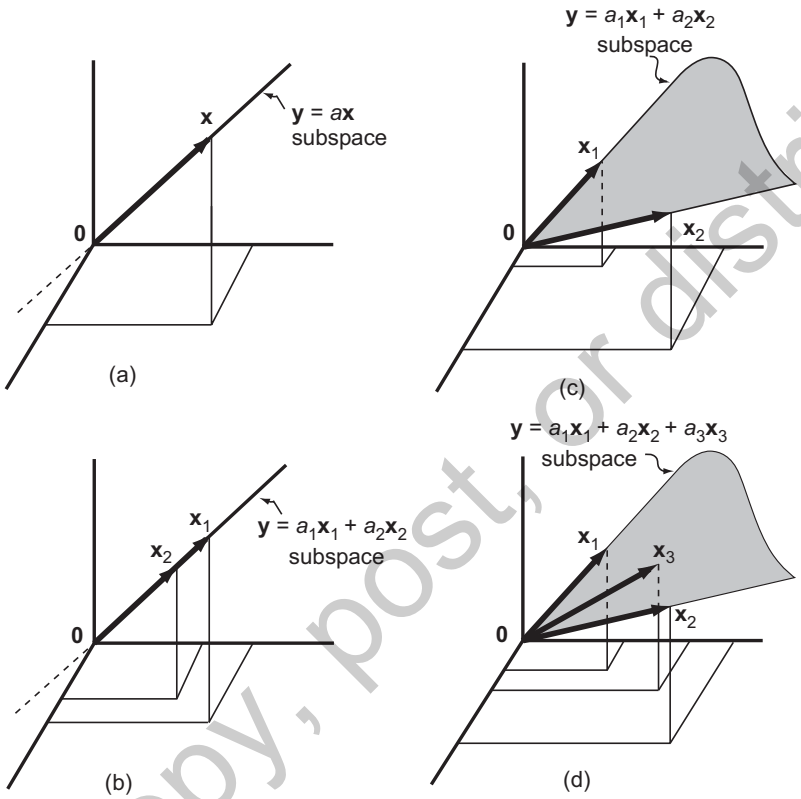

Figure 1.6 Subspaces generated by sets of vectors in three-dimensional space. (a) One nonzero vector generates a one-dimensional subspace (a line). (b) Two collinear vectors also generate a one-dimensional subspace. (c) Two linearly independent vectors generate a two-dimensional subspace (a plane). (d) Three linearly dependent vectors, two of which are linearly independent, generate a two-dimensional subspace. The lines in (a) and (b) extend infinitely, as do the planes in (c) and (d): The planes are drawn between $\mathbf{x}_1$ and $\mathbf{x}_2$ only for clarity.

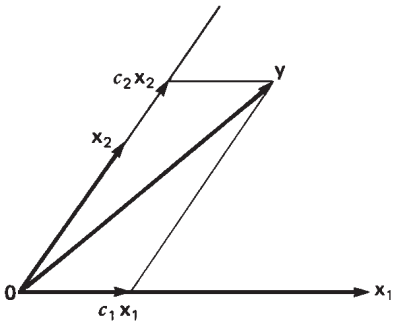

Figure 1.7 The coordinates of $\mathbf{y}$ with respect to the basis $\{\mathbf{x}_1, \mathbf{x}_2\}$ of a two-dimensional subspace can be found from the parallelogram rule of vector addition.

a system of linear simultaneous equations in which the c_js are the unknowns:

$$\begin{aligned}\underset{(n\times 1)}{\mathbf{y}} &= c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k \\ &= [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} \\ &= \underset{(n\times k)(k\times 1)}{\mathbf{X}\quad\mathbf{c}}\end{aligned}$$

When the vectors in $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ are linearly independent, the matrix $\mathbf{X}$ is of *full column rank* k, and the equations have a unique solution. The concept of rank and the solution of systems of linear simultaneous equations are taken up later in this chapter (Section 1.4.2).

1.3.1 Orthogonality and Orthogonal Projections

Recall that the inner product of two vectors is the sum of products of their coordinates:

$$\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

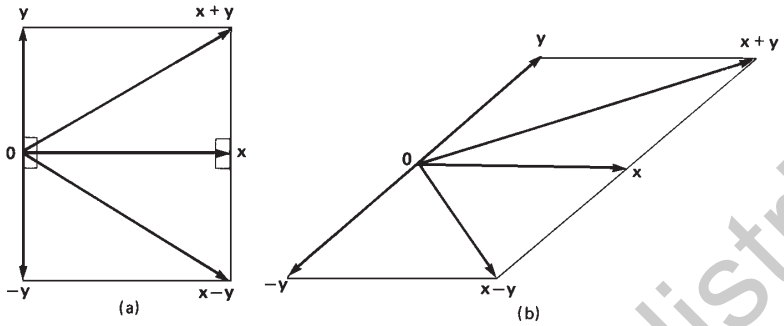

Figure 1.8 When two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, as in (a), their inner product $\mathbf{x} \cdot \mathbf{y}$ is zero. When the vectors are not orthogonal, as in (b), their inner product is nonzero.

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* (i.e., perpendicular) if their inner product is zero. The essential geometry of vector orthogonality is shown in Figure 1.8. Although $\mathbf{x}$ and $\mathbf{y}$ lie in an n-dimensional space (which cannot be visualized directly when $n > 3$), they span a subspace of dimension two, which, by convention, I make the plane of the paper.[9] When $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, as in Figure 1.8(a), the two right triangles with vertices ($\mathbf{0}$, $\mathbf{x}$, $\mathbf{x}+\mathbf{y}$) and ($\mathbf{0}$, $\mathbf{x}$, $\mathbf{x}-\mathbf{y}$) are congruent; consequently, $||\mathbf{x}+\mathbf{y}|| = ||\mathbf{x}-\mathbf{y}||$. Because the squared length of a vector is the inner product of the vector with itself ($\mathbf{x} \cdot \mathbf{x} = \sum x_i^2$), we have

$$\begin{aligned}
(\mathbf{x}+\mathbf{y}) \cdot (\mathbf{x}+\mathbf{y}) &= (\mathbf{x}-\mathbf{y}) \cdot (\mathbf{x}-\mathbf{y}) \\
\mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} &= \mathbf{x} \cdot \mathbf{x} - 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\
4\mathbf{x} \cdot \mathbf{y} &= 0 \\
\mathbf{x} \cdot \mathbf{y} &= 0
\end{aligned}$$

When, in contrast, $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal, as in Figure 1.8(b), then $||\mathbf{x}+\mathbf{y}|| \neq ||\mathbf{x}-\mathbf{y}||$, and $\mathbf{x} \cdot \mathbf{y} \neq 0$.

The definition of orthogonality can be extended to matrices in the following manner: The matrix $\underset{(n \times k)}{\mathbf{X}}$ is orthogonal if each pair of its columns

[9] It is helpful to employ this device in applying vector geometry to statistical problems, where the subspace of interest can often be confined to two or three dimensions, even though the dimension of the full vector space is typically equal to the sample size n.

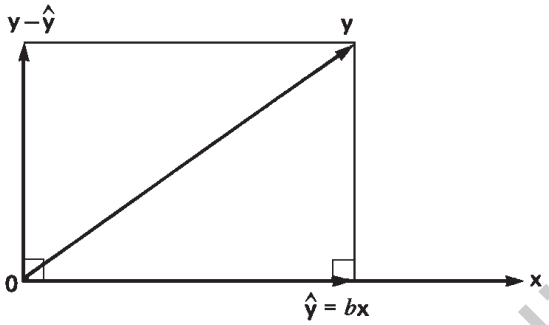

Figure 1.9 The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{y}$ onto $\mathbf{x}$.

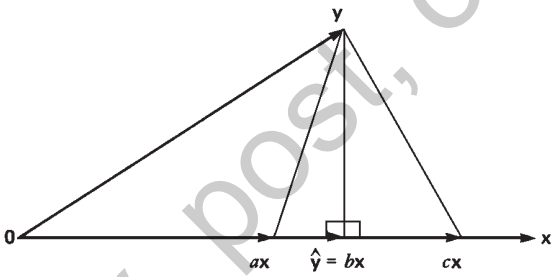

Figure 1.10 The orthogonal projection $\widehat{\mathbf{y}} = b\mathbf{x}$ is the point along the line spanned by $\mathbf{x}$ that is closest to $\mathbf{y}$.

is orthogonal—that is, if $\mathbf{X}'\mathbf{X}$ is diagonal.[10] The matrix $\mathbf{X}$ is *orthonormal* if $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

The *orthogonal projection* of one vector $\mathbf{y}$ onto another vector $\mathbf{x}$ is a scalar multiple $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{x}$ such that $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to $\mathbf{x}$. The geometry of orthogonal projection is illustrated in Figure 1.9. By the Pythagorean theorem (see Figure 1.10), $\widehat{\mathbf{y}}$ is the point along the line spanned by $\mathbf{x}$ that is closest to

[10]The i, jth entry of $\mathbf{X}'\mathbf{X}$ is $\mathbf{x}_i'\mathbf{x}_j = \mathbf{x}_i \cdot \mathbf{x}_j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are, respectively, the ith and jth columns of $\mathbf{X}$. The ith diagonal entry of $\mathbf{X}'\mathbf{X}$ is likewise $\mathbf{x}_i'\mathbf{x}_i = \mathbf{x}_i \cdot \mathbf{x}_i$, which is necessarily nonzero unless $\mathbf{x}_i = \mathbf{0}$.

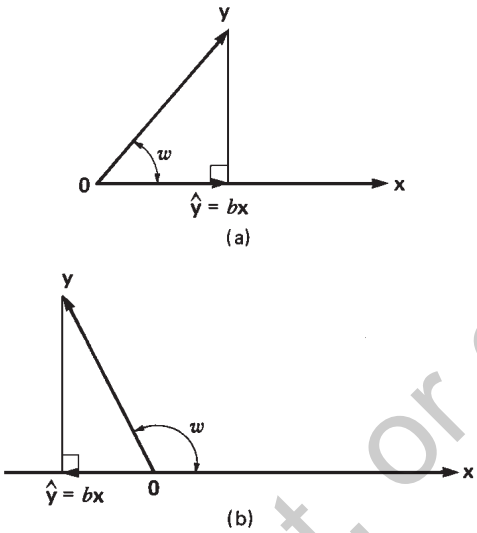

Figure 1.11 The angle w separating two vectors, $\mathbf{x}$ and $\mathbf{y}$: (a) $0^\circ < w < 90^\circ$ (when b is positive); (b) $90^\circ < w < 180^\circ$ (when b is negative).

$\mathbf{y}$. To find b, we note that

$$\mathbf{x} \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x} \cdot (\mathbf{y} - b\mathbf{x}) = 0$$

Thus, $\mathbf{x} \cdot \mathbf{y} - b\mathbf{x} \cdot \mathbf{x} = 0$ and $b = (\mathbf{x} \cdot \mathbf{y})/(\mathbf{x} \cdot \mathbf{x})$.

The orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}$ can be used to determine the angle w separating two vectors, by finding its cosine. Because the cosine function is symmetric around $w = 0$, it does not matter in which direction we measure an angle, and I will simply treat angles as positive.[11] I will distinguish between two cases:[12] In Figure 1.11(a), the angle separating the vectors is between 0° and 90°; in Figure 1.11(b), the angle is between 90° and 180°.

[11] The cosine and other basic trigonometric functions are reviewed in Section 3.1.5.

[12] By convention, we examine the smaller of the two angles separating a pair of vectors, and, therefore, never encounter angles that exceed 180°. Call the smaller angle w; then the larger angle is $360^\circ - w$. This convention is of no consequence because $\cos(360 - w) = \cos(w)$.

In the first instance,

$$\cos(w) = \frac{||\widehat{\mathbf{y}}||}{||\mathbf{y}||} = \frac{b||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}||^2} \times \frac{||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}|| \times ||\mathbf{y}||}$$

and, likewise, in the second instance,

$$\cos(w) = -\frac{||\widehat{\mathbf{y}}||}{||\mathbf{y}||} = \frac{b||\mathbf{x}||}{||\mathbf{y}||} = \frac{\mathbf{x}\cdot\mathbf{y}}{||\mathbf{x}|| \times ||\mathbf{y}||}$$

In both instances, the sign of b for the orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}$ correctly reflects the sign of $\cos(w)$. When the vectors are orthogonal (not shown in the figure), $b = 0$, $\cos(w) = 0$, and $w = 90°$; when the vectors are collinear (also not shown), $\cos(w) = 1$, and $w = 0°$.

The orthogonal projection of a vector $\mathbf{y}$ onto the subspace spanned by a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is the vector

$$\widehat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k$$

formed as a linear combination of the $\mathbf{x}_j$s such that $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to each and every vector $\mathbf{x}_j$ in the set. The geometry of orthogonal projection for $k = 2$ is illustrated in Figure 1.12. The vector $\widehat{\mathbf{y}}$ is the point closest to $\mathbf{y}$ in the subspace spanned by the $\mathbf{x}_j$s.

Placing the constants b_j into a vector $\mathbf{b}$, and gathering the vectors $\mathbf{x}_j$ into an $(n \times k)$ matrix $\mathbf{X} \equiv [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$, we have $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$. By the definition of an orthogonal projection,

$$\mathbf{x}_j \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x}_j \cdot (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \quad \text{for } j = 1, \ldots, k \tag{1.10}$$

Equivalently, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$, or $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. We can solve this matrix equation uniquely for $\mathbf{b}$ as long as the $(k \times k)$ matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, in which case $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (see Section 1.4.2 on the solution of linear simultaneous equations). The matrix $\mathbf{X}'\mathbf{X}$ is nonsingular if $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a linearly independent set of vectors, providing a basis for the subspace that it generates; otherwise, $\mathbf{b}$ is not unique.

The application of the geometry of orthogonal projections to linear least-squares regression is quite direct, and so I will explain it here (rather than in Chapter 7 on least-square regression). For example, suppose that the vector $\mathbf{x}$ in Figures 1.9 and 1.11 represents the explanatory ("independent") variable in a simple regression; the vector $\mathbf{y}$ represents the response ("dependent") variable; and both variables are expressed as deviations from their means, $\mathbf{x} = \{X_i - \overline{X}\}$ and $\mathbf{y} = \{Y_i - \overline{Y}\}$. Then $\widehat{\mathbf{y}} = b\mathbf{x}$ is the mean-deviation vector of fitted ("predicted") Y-values from the linear least-squares regression of Y

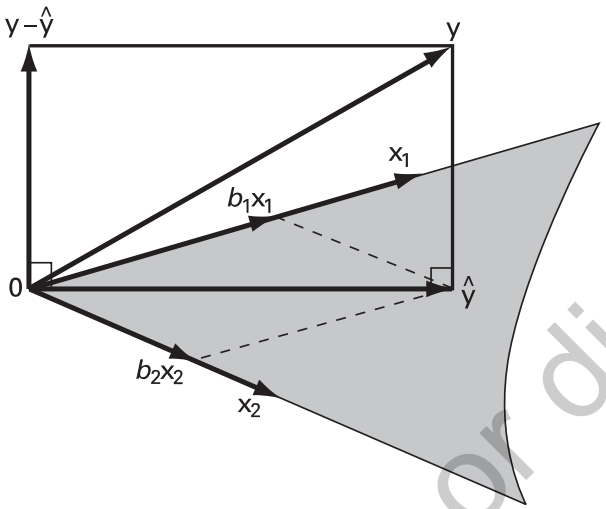

Figure 1.12 The orthogonal projection $\widehat{\mathbf{y}}$ of $\mathbf{y}$ onto the subspace (plane) spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.

on X; b is the slope coefficient for the regression; and $\mathbf{y} - \widehat{\mathbf{y}}$ is the vector of least-square residuals. By the Pythagorean theorem,

$$||\mathbf{y}||^2 = ||\widehat{\mathbf{y}}||^2 + ||\mathbf{y} - \widehat{\mathbf{y}}||^2$$

which shows the decomposition of the total sum of squares for Y into the regression and residual sums of squares—the so-called analysis of variance for the regression. The correlation r between X and Y is then the cosine of the angle w separating their mean-deviation vectors.

Suppose similarly that $\mathbf{y}$ is the mean-deviation vector for the response variable and that $\mathbf{x}_1$ and $\mathbf{x}_2$ are the mean-deviation vectors for two explanatory variables in a multiple regression. Then Figure 1.12 represents the linear least-squares regression of Y on X_1 and X_2; b_1 and b_2 are the partial regression coefficients for the two explanatory variables; $\widehat{\mathbf{y}}$ is the vector of mean-deviation fitted values for the multiple regression; the right triangle formed by the origin and the vectors $\mathbf{y}$ and $\widehat{\mathbf{y}}$ gives the analysis of variance for the multiple regression; and the cosine of the angle separating $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is the multiple-correlation coefficient R for the regression—that is, the correlation between observed and fitted Y-values.

1.4 Matrix Rank and the Solution of Linear Simultaneous Equations

1.4.1 Rank

The *row space* of an $(m \times n)$ matrix $\mathbf{A}$ is the subspace of the n-dimensional vector space spanned by the m rows of $\mathbf{A}$ (treated as a set of vectors). The *rank* of $\mathbf{A}$ is the dimension of its row space, that is, the maximum number of linearly independent rows in $\mathbf{A}$. It follows immediately that $\text{rank}(\mathbf{A}) \leq \min(m,n)$.

For example, the row space of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

consists of all vectors

$$\begin{aligned} \mathbf{x}' &= a[1,0,0] + b[0,1,0] \\ &= [a,b,0] \end{aligned}$$

for any values of a and b. This subspace is of dimension two, and thus $\text{rank}(\mathbf{A}) = 2$.

A matrix is said to be in *reduced row–echelon form* (*RREF*) if it satisfies the following criteria:

R1: All of its nonzero rows (if any) precede all of its zero rows (if any).

R2: The first nonzero entry (proceeding from left to right) in each nonzero row, called the *leading entry* in the row, is 1.

R3: The leading entry in each nonzero row after the first is to the right of the leading entry in the previous row.

R4: All other entries are 0 in a *column* containing a leading entry.

RREF is displayed schematically in the matrix in 1.11, where the asterisks represent elements of arbitrary value (i.e., they may be zero or nonzero):

$$\left[\begin{array}{ccccccccccccccc}
0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & & \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * \\
\hline
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0
\end{array}\right]
\begin{array}{l}
\\ \\ \\ \\ \\ \\ \\
\text{nonzero} \\ \text{rows} \\ \\ \\ \\
\text{zero} \\ \text{rows}
\end{array}$$

(1.11)

The rank of a matrix in RREF is equal to the number of nonzero rows in the matrix: The pattern of leading entries, each located in a column all of whose other elements are zero, ensures that no nonzero row can be formed as a linear combination of other rows.

A matrix can be placed in RREF by a sequence of elementary row operations, adapting the Gaussian elimination procedure described earlier in this chapter. For example, starting with the matrix

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

1. Divide Row 1 by -2,

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

2. Subtract $4\times$ Row 1 from Row 2,

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

3. Subtract $6\times$ Row 1 from Row 3,

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

4. Multiply Row 2 by -1,

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

5. Subtract $\frac{1}{2}\times$ Row 2 from Row 1,

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

6. Add 2× Row 2 to Row 3,

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The rank of a matrix $\mathbf{A}$ is equal to the rank of its RREF $\mathbf{A}_R$, because a zero row in $\mathbf{A}_R$ can only arise if one row of $\mathbf{A}$ is expressible as a linear combination of other rows (or if $\mathbf{A}$ contains a zero row). That is, none of the elementary row operations alters the rank of a matrix. The rank of the matrix transformed to RREF in the example is thus 2. The RREF of a nonsingular square matrix is the identity matrix, and the rank of a nonsingular square matrix is therefore equal to its order. Conversely, the rank of a singular matrix is less than its order.

I have defined the rank of a matrix $\mathbf{A}$ as the dimension of its row space. It can be shown that the rank of $\mathbf{A}$ is also equal to the dimension of its *column space*—that is, to the maximum number of linearly independent columns in $\mathbf{A}$.

1.4.2 Linear Simultaneous Equations

A system of m linear simultaneous equations in n unknowns can be written in matrix form as

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}} \tag{1.12}$$

where the elements of the coefficient matrix $\mathbf{A}$ and the right-hand-side vector $\mathbf{b}$ are prespecified constants, and $\mathbf{x}$ is a vector of unknowns. Suppose that there is an equal number of equations and unknowns—that is, $m = n$. Then if the coefficient matrix $\mathbf{A}$ is nonsingular, Equation 1.12 has the *unique solution* $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Alternatively, $\mathbf{A}$ may be singular. Then $\mathbf{A}$ can be transformed to RREF by a sequence of (say, p) elementary row operations, representable as successive multiplication on the left by elementary-row-operation matrices:

$$\mathbf{A}_R = \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{E}\mathbf{A}$$

Applying these operations to both sides of Equation 1.12 produces

$$\begin{aligned} \mathbf{E}\mathbf{A}\mathbf{x} &= \mathbf{E}\mathbf{b} \\ \mathbf{A}_R\mathbf{x} &= \mathbf{b}_R \end{aligned} \tag{1.13}$$

where $\mathbf{b}_R \equiv \mathbf{E}\mathbf{b}$. Equations 1.12 and 1.13 are *equivalent* in the sense that any solution vector $\mathbf{x} = \mathbf{x}^*$ that satisfies one system also satisfies the other.

Let r represent the rank of $\mathbf{A}$. Because $r < n$ (recall that $\mathbf{A}$ is singular), $\mathbf{A}_R$ contains r nonzero rows and $n - r$ zero rows. If any zero row of $\mathbf{A}_R$ is associated with a nonzero entry (say, b) in $\mathbf{b}_R$, then the system of equations is *inconsistent* or *overdetermined*, for it contains the self-contradictory "equation"

$$0x_1 + 0x_2 + \cdots + 0x_n = b \neq 0$$

If, on the other hand, every zero row of $\mathbf{A}_R$ corresponds to a zero entry in $\mathbf{b}_R$, then the equation system is *consistent*, and there is an infinity of solutions satisfying the system: $n - r$ of the unknowns may be given arbitrary values, which then determine the values of the remaining r unknowns. Under this circumstance, we say that the equation system is *underdetermined*.

Suppose now that there are *fewer* equations than unknowns—that is, $m < n$. Then r is necessarily less than n, and the equations are either overdetermined (if a zero row of $\mathbf{A}_R$ corresponds to a nonzero entry of $\mathbf{b}_R$) or underdetermined (if they are consistent). For example, consider the following system of three equations in four unknowns:

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

Adjoin the right-hand-side vector to the coefficient matrix

$$\left[\begin{array}{cccc|c} -2 & 0 & -1 & 2 & 1 \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

and reduce the coefficient matrix to row–echelon form:

1. Divide Row 1 by -2,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

2. Subtract 4 $\times$ Row 1 from Row 2, and subtract 6 $\times$ Row 1 from Row 3,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & -1 & 4 & 4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

3. Multiply Row 2 by -1,

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & 1 & -4 & -4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

4. Subtract $\frac{1}{2}\times$ Row 2 from Row 1, and add 2 $\times$ Row 2 to Row 3,

$$\left[\begin{array}{cccc|c} 1^{\swarrow} & 0 & 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1^{\swarrow} & -4 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

(with the leading entries marked by arrows).

Writing the result as a scalar system of equations, we get

$$\begin{aligned} x_1 + x_4 &= \tfrac{3}{2} \\ x_3 - 4x_4 &= -4 \\ 0x_1 + 0x_2 + 0x_3 + 0x_4 &= 0 \end{aligned}$$

The third equation is uninformative (it simply states that $0 = 0$), but it does confirm that the original system of equations is consistent. The first two equations imply that the unknowns x_2 and x_4 can be given arbitrary values (say x_2^* and x_4^*), and the values of x_1 and x_3 (corresponding to the leading entries) follow:

$$\begin{aligned} x_1 &= \tfrac{3}{2} - x_4^* \\ x_3 &= -4 + 4x_4^* \end{aligned}$$

and thus any vector of the form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} - x_4^* \\ x_2^* \\ -4 + 4x_4^* \\ x_4^* \end{bmatrix}$$

is a solution of the system of equations.

Now consider the system of equations

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Attaching **b** to **A** and transforming the coefficient matrix to RREF yields (as the reader may wish to verify)

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & -4 & -2 \\ 0 & 0 & 0 & 0 & 2 \end{array}\right]$$

The last "equation,"

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 2$$

is contradictory, implying that the original system of equations has no solution (i.e., is overdetermined).

Suppose, finally, that there are *more* equations than unknowns: $m > n$. If **A** is of full-column rank (i.e., if $r = n$), then $\mathbf{A}_R$ consists of the order-n identity matrix followed by $m - r$ zero rows. If the equations are consistent, they therefore have a unique solution; otherwise they are overdetermined. If $r < n$, the equations are either overdetermined (if inconsistent) or underdetermined (if consistent).

To illustrate these results geometrically, consider a system of three linear equations in two unknowns:[13]

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 &= b_3 \end{aligned}$$

Each equation describes a line in a 2D coordinate space in which the unknowns define the axes, as illustrated schematically in Figure 1.13. If the three lines intersect at a point, as in Figure 1.13(a), then there is a *unique solution* to the equation system: Only the pair of values (x_1^*, x_2^*) simultaneously satisfies all three equations. If the three lines fail to intersect at a

[13]The geometric representation of linear equations by lines (or, more generally, by linear surfaces, i.e., planes in three dimensions or hyperplanes in higher dimensions) should not be confused with the geometric vector representation taken up previously in this chapter. The graphs of linear equations in two and three dimensions are reviewed in Section 3.1.2.

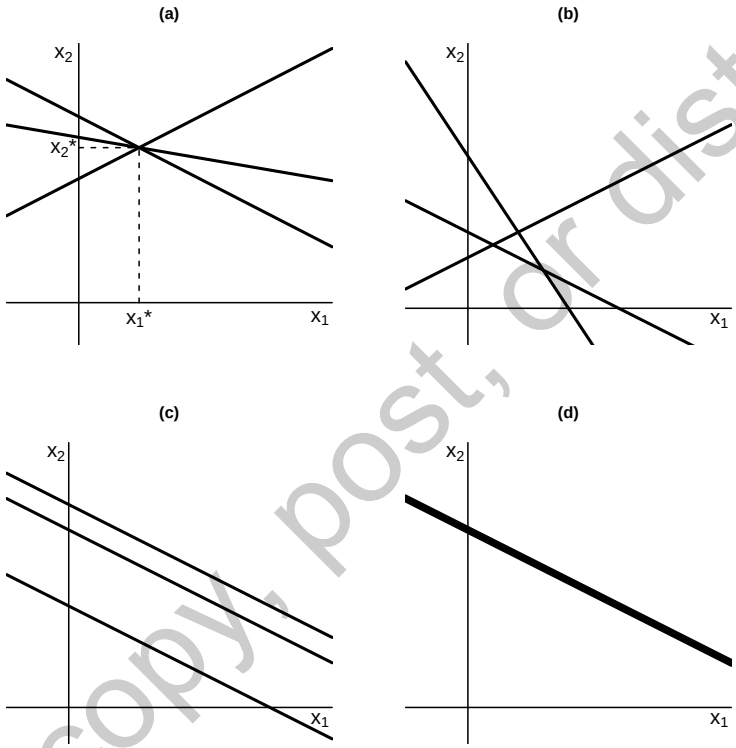

Figure 1.13 Three linear equations in two unknowns x_1 and x_2: (a) unique solution ($x_1 = x_1^*, x_2 = x_2^*$); (b) and (c) overdetermined (no solution); (d) underdetermined (three coincident lines, an infinity of solutions).

Table 1.1 Solutions of m Linear Simultaneous Equations in n Unknowns

Number of equations	$m < n$	$m = n$		$m > n$	
Rank of coefficient matrix	$r < n$	$r < n$	$r = n$	$r < n$	$r = n$
		General equation system			
Consistent	Under-determined	Under-determined	Unique solution	Under-determined	Unique solution
Inconsistent	Over-determined	Over-determined	—	Over-determined	Over-determined
		Homogeneous equation system			
Consistent	Nontrivial solutions	Nontrivial solutions	Trivial solution	Nontrivial solutions	Trivial solution

common point, as in Figures 1.13(b) and (c), then *no* pair of values of the unknowns simultaneously satisfies the three equations, which therefore are overdetermined. Last, if the three lines are coincident, as in Figure 1.13(d), then *any* pair of values on the common line satisfies all three equations, and the equations are underdetermined.

When the right-hand-side vector $\mathbf{b}$ in a system of linear simultaneous equations is the zero vector, the system of equations is said to be *homogeneous*:

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{0}}$$

The *trivial solution* $\mathbf{x} = \mathbf{0}$ always satisfies a homogeneous system, which consequently cannot be inconsistent. From the previous work in this section, we can see that nontrivial solutions exist if $\text{rank}(\mathbf{A}) < n$—that is, when the system is underdetermined.

The results concerning the solution of linear simultaneous equations developed in this section are summarized in Table 1.1.

Linear simultaneous equations have many statistical applications, such as solving for least-squares coefficients in regression analysis (see Section 7.1).

1.4.3 Generalized Inverses

As I explained previously in this chapter, only square nonsingular matrices have inverses. All matrices, however—including singular and rectangular matrices—have *generalized inverses*, which are occasionally employed in

statistical applications, such as some presentations of linear statistical models.[14]

A generalized inverse of the $(m \times n)$ matrix $\mathbf{A}$ is an $(n \times m)$ matrix $\mathbf{A}^-$ that satisfies the equation

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A} \tag{1.14}$$

We say that $\mathbf{A}^-$ is *a* generalized inverse, not *the* generalized inverse of $\mathbf{A}$, because unless $\mathbf{A}$ is square and nonsingular (in which case $\mathbf{A}^- = \mathbf{A}^{-1}$), the generalized inverse is not unique.[15]

There are many ways to find a generalized inverse of a matrix, including by Gaussian elimination. Suppose that we begin by putting the matrix $\mathbf{A}$ in RREF by a sequence of elementary row operations; we know that we can represent this process by successive multiplication on the left by suitably configured elementary-row-operations matrices (see pages 14 and 32):

$$\mathbf{E}\mathbf{A} = \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{A}_R \tag{1.15}$$

where $\mathbf{E} \equiv \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1$ is a nonsingular $(m \times m)$ matrix. Applying an analogous series of Types II and III *elementary column operations* (pivoting is unnecessary because all of the leading entries in $\mathbf{A}_R$ are already 1), we can further reduce $\mathbf{A}_R$ to the following *canonical form*:

$$\underset{(m\times n)}{\mathbf{A}_C} \equiv \mathbf{A}_R\mathbf{E}^* = \mathbf{A}_R\mathbf{E}_1^*\mathbf{E}_2^* \cdots \mathbf{E}_q^* = \begin{bmatrix} \mathbf{I}_r & \underset{(r\times n-r)}{\mathbf{0}} \\ \underset{(m-r\times r)}{\mathbf{0}} & \underset{(m-r\times n-r)}{\mathbf{0}} \end{bmatrix} \tag{1.16}$$

where $\mathbf{E}^* \equiv \mathbf{E}_1^*\mathbf{E}_2^* \cdots \mathbf{E}_q^*$ is a nonsingular $(n \times n)$ matrix; the order r of the identity matrix in the upper-left corner is the rank of $\mathbf{A}$; and any or all of the zero matrices may be absent. For example, if $\mathbf{A}$ is a square nonsingular matrix of order n then $r = n$ and none of the zero matrices are required.

Putting together Equations 1.15 and 1.16, we have

$$\mathbf{A}_C = \mathbf{E}\mathbf{A}\mathbf{E}^* \tag{1.17}$$

[14] For an extensive discussion of the role of generalized inverses in statistics, See Rao and Mitra (1971).

[15] The generalized inverse can be made unique by placing additional restrictions on it beyond Equation 1.14: For example, the *Moore–Penrose generalized inverse* $\mathbf{A}^+$ satisfies four conditions: $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$; $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$; $\mathbf{A}\mathbf{A}^+$ is symmetric; and $\mathbf{A}^+\mathbf{A}$ is symmetric. In a typical statistical application, however, one generalized inverse is as good as another.

A generalized inverse of $\mathbf{A}$ is then given by[16]

$$\mathbf{A}^- \equiv \mathbf{E}^* \mathbf{A}_C' \mathbf{E}$$

Consider, for example, the matrix

$$\mathbf{A} = \begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

Earlier in the chapter (page 32), I transformed this matrix to RREF by a sequence of elementary row operations:

$$\mathbf{A}_R = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The reduction to canonical form is completed by exchanging Columns 2 and 3, and then sweeping out the fourth column, producing

$$\mathbf{A}_C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Collecting the elementary row and column operations into matrices, we have (as the reader may wish to verify)

$$\mathbf{E} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix}$$

$$\mathbf{E}^* = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

[16] The following proof is adapted from Healy (1986, p. 40): First, $\mathbf{A}_C'$ is a generalized inverse of $\mathbf{A}_C$ (*Reader:* Check it!); second, solving Equation 1.17 for $\mathbf{A}$ produces $\mathbf{A} = \mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}$. Then,

$$\begin{aligned} \mathbf{A}\mathbf{A}^-\mathbf{A} &= (\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1})(\mathbf{E}^*\mathbf{A}_C'\mathbf{E})(\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}) \\ &= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{A}_C'\mathbf{A}_C\mathbf{E}^{*-1} \\ &= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1} \\ &= \mathbf{A} \end{aligned}$$

which establishes the result.

from which

$$\begin{aligned}\mathbf{A}^{-} &= \mathbf{E}^{*}\mathbf{A}_{C}'\mathbf{E} \\ &= \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ -2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\end{aligned}$$

is a generalized inverse of $\mathbf{A}$ (as the reader can also verify).

Now consider a system of m linear simultaneous equations in n unknowns,

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}}$$

as discussed in the preceding section, and suppose that the system of equations is consistent and underdetermined. Then

$$\mathbf{x}^{*} = \mathbf{A}^{-}\mathbf{b} \tag{1.18}$$

provides an arbitrary solution to the equations. If the equation system has a unique solution, then Equation 1.18 yields it. Finally, if the equation system is overdetermined, then the "solution" provided by Equation 1.18 will fail to satisfy the original system of equations. Thus, if the equation system is consistent, then $\mathbf{A}\mathbf{A}^{-}\mathbf{b} = \mathbf{b}$, and if the system is inconsistent, then $\mathbf{A}\mathbf{A}^{-}\mathbf{b} \neq \mathbf{b}$. The reader may wish to apply these results to the examples in the previous section.

REFERENCES

Aldrich, J. (1997). R. A. Fisher and the making of maximum-likelihood 1912–1922. *Statistical Science*, *12*, 162–176.

Barndorf-Nielsen, O. E., & Cox, D. R. (1994). *Inference and asymptotics*. Boca Raton: Chapman & Hall/CRC Press.

Binmore, K., & Davies, J. (2001). *Calculus: Concepts and methods*. Cambridge UK: Cambridge University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Casella, G., & George, E. J. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London.

Davis, P. J. (1965). *Mathematics of matrices: A first book of matrix theory and linear algebra*. New York: Blaisdell.

Engle, R. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. II, pp. 775–879). Amsterdam: North-Holland.

Fieller, N. (2016). *Basics of matrix algebra for statistics with R*. Boca Raton: Chapman & Hall/CRC Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A*, *222*, 309–368.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks CA: Sage.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.

Francis, J. G. F. (1961). The QR transformation: A unitary analogue to the LR transformation—part 1. *The Computer Journal*, *4*, 265–271.

Francis, J. G. F. (1962). The QR transformation—part 2. *The Computer Journal*, *4*, 332–345.

Friendly, M., Monette, G., & Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, *28*, 1–39.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton FL: Chapman & Hall/CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge UK: Cambridge University Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–742.

Graybill, F. A. (1983). *Introduction to matrices with applications in statistics* (2nd ed.). Belmont CA: Wadsworth.

Green, P. E., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.
Healy, M. J. R. (1986). *Matrices for statistics*. Oxford UK: Clarendon Press.
Johnston, J. (1972). *Econometric methods* (2nd ed.). New York: McGraw-Hill.
Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing*. New York: Dekker.
Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Oxford UK: Blackwell.
Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049–3082.
McCallum, B. T. (1973). A note concerning asymptotic covariance expressions. *Econometrica*, *41*, 581–583.
McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (second ed.). Boca Raton: Chapman & Hall/CRC Press.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.
Monahan, J. F. (2001). *Numerical methods of statistics*. Cambridge UK: Cambridge University Press.
Namboodiri, K. (1984). *Matrix algebra: An introduction*. Beverly Hills: Sage.
Nash, J. C. (2014). *Nonlinear parameter optimization using R tools*. Chichester UK: Wiley.
Neal, R. M. (1996). *Bayesian learning for neural networks*. New York: Springer.
Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Boca Raton FL: Chapman & Hall/CRC Press.
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, *135*, 370–384.
Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.
Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications*. New York: Wiley.
Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Boca Raton FL: Chapman & Hall/CRC.
Stewart, J. (2016). *Calculus* (eighth ed.). Boston: Cengage Learning.
Theil, H. (1971). *Principles of econometrics*. New York: Wiley.
Thompson, S. P., & Gardner, M. (1998). *Calculus made easy*. New York: St. Martin's.
Weisberg, S. (2014). *Applied linear regression* (4th ed.). Hoboken, NJ: Wiley.
Wolfram, S. (2003). *The Mathematica book* (5th ed.). Champaign IL: Wolfram Research.
Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: Wiley.
Zellner, A. (1983). Statistical theory and econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 1, pp. 67–178). Amsterdam: North-Holland.