APPLIED REGRESSION

An Introduction

Second Edition

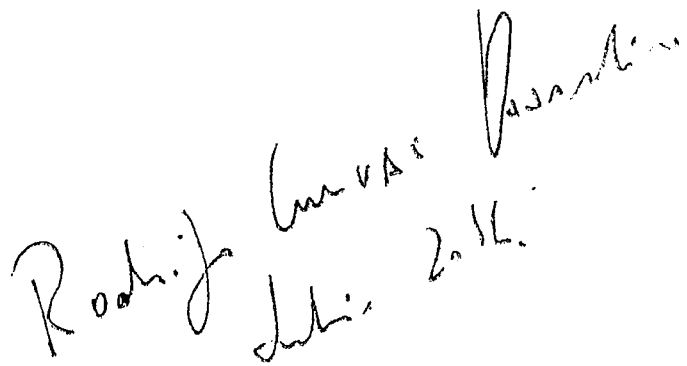
Colin Lewis-Beck
Michael Lewis-Beck

Series: Quantitative Applications in the Social Sciences

Series: Quantitative Applications in the Social Sciences

APPLIED REGRESSION

Second Edition



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 *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 *Schroeder/Sjoquist/Stephan*
58. Randomized Response and Related Methods, 2nd Edition *Fox*
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 *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 *Ellason*
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 *James E. 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 *Pampel*
133. Translating Questionnaires and Other Research Instruments: Problems and Solutions *Behling/Law*
134. Generalized Linear Models: A Unified Approach *Gill*
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 *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 *Gilbert*
154. Social Network Analysis, 2nd Edition *Knoke/Yang*
155. Spatial Regression Models *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 *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*

APPLIED REGRESSION

An Introduction

Second Edition

Colin Lewis-Beck

Iowa State University

Michael S. Lewis-Beck

University of Iowa

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

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.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Lewis-Beck, Michael S.

Applied regression : an introduction. — Second edition / Colin Lewis-Beck, Michael S. Lewis-Beck.

pages cm
Includes bibliographical references and index.

ISBN 978-1-4833-8147-3 (pbk. : alk. paper)

1. Regression analysis. I. Title.
HA31.3.L48 2016
519.5'36—dc23 2015011813

This book is printed on acid-free paper.

SUSTAINABLE FORESTRY INITIATIVE
Certified Chain of Custody
Promoting Sustainable Forestry
www.sfiprogram.org
SFI-01268
SFI label applies to text stock

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

Acquisitions Editor: Helen Salmon
eLearning Editor: Katie Bierach
Editorial Assistant: Anna Villarruel
Production Editor: Kelly DeRosa
Copy Editor: Gillian Dickens
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Scott Oney
Indexer: William Ragsdale
Cover Designer: Candice Harman
Marketing Manager: Nicole Elliott

CONTENTS

SERIES EDITOR'S INTRODUCTION

Invented more than 200 years ago, apparently independently by the German mathematician Carl Friedrich Gauss and the French mathematician Adrien-Marie Legendre, the method of least squares occupies a central place in statistical methods. Linear least squares regression not only is very widely employed in research but also furnishes a basis for much of applied statistics. Many statistical models — generalized linear models, linear and generalized linear mixed-effects models, survival regression models, and linear structural equation models, to name a few of the more prominent — represent direct generalizations of linear regression; and computation for statistical models often involves least squares fitting — for example, the use of iterated weighted least squares to compute maximum likelihood estimates for generalized linear models. Both for its direct application and for its many generalizations, a sound background in linear least squares regression is fundamental to the study of statistics.

In *Applied Regression*, Colin and Michael Lewis-Beck provide a thorough primer in linear least squares regression, introducing the method from first principles. They attend to practical details of regression analysis; to the statistical model underlying inference in linear regression and to violations of the assumptions of the model; and — most important – to the interpretation of results and the interplay between statistical modeling and the substance of social research.

There is clearly a need for a brief, accessible, and nontechnical treatment of regression analysis, and the first edition of this monograph was one of the most widely read in the QASS series. I expect that this new, expanded, and extensively revised edition will be similarly well received.

On a personal note, I am particularly pleased to be able to assist in the publication of this monograph because I have known Michael Lewis-Beck since we were both graduate students at the University of Michigan many years ago, and I have subsequently had the pleasure of becoming acquainted with his son, Colin.

—*John Fox*
Series Editor

PREFACE

In this second edition of *Applied Regression: An Introduction*, we maintain our firm commitment to the method of ordinary least squares (OLS). We are not alone in our defense of OLS. Peter Kennedy (2008), author of a leading econometrics text, observed the following: "The central role of the OLS estimator in the econometrician's catalog is that of a standard against which all other estimators are compared. The reason for this is that the OLS estimator is extraordinarily popular" (p. 43). This popularity was recently affirmed in a methodological content analysis of the articles in the three leading general political science journals, with the finding that "OLS is by far the most popular method" (Krueger & Lewis-Beck, 2008, p. 3). This is not surprising, since OLS is the analytic tool of the classical linear regression model.

As Jan Kmenta (1997), author of our favorite econometrics book, reminds us, "The need for familiarity with the basic principles of statistical inference and with the fundamentals of econometrics has not diminished. . . . Most econometric problems can be characterized as situations in which some of the basic assumptions of the classical regression model are violated" (pp. v–vi). In our monograph, we pay special attention to these basic regression assumptions. Also, in the writing, we are inspired again by Kmenta (1997) and his "philosophy of making everything as simple and clear as possible" (p. vii). It is our hope that readers agree that we have realized this goal. Indeed, if readers are interested in further analyzing or replicating the results presented in this monograph, the datasets are available for download through the SAGE website.

ACKNOWLEDGMENTS

There are many people to thank for help on this project. At SAGE, we would especially like to thank Helen Salmon (Acquisitions Editor) and John Fox (Series Editor). Their general enthusiasm for this second edition, coupled with specific useful suggestions for improvement, are greatly appreciated. Also, we wish to take note of the insightful comments made by the many reviewers of our proposal. In addition, we acknowledge particular teachers, including Frank M. Andrews, John DiNardo, Lawrence B. Mohr, and Jan Kmenta, whose wisdom has abided. Furthermore, we are grateful to the Inter-University Consortium for Political and Social Research (ICPSR) Summer Program in Quantitative Methods of Social Research, University of Michigan, and its past and current directors, William G. Jacoby and Saundra K. Schneider, who have given such strong support to this monograph. Thanks as well to the following SAGE reviewers who provided valuable feedback and suggestions: Charles R. Boehmer, University of Texas at El Paso; Tim Liao, University of Illinois; Michael Ornstein, York University; Carl L. Palmer, Illinois State University; Walter J. Stone, University of California, Davis; Matt Vogel, University of Missouri–St. Louis; and Herbert F. Weisberg, The Ohio State University. Last, but not least, we express gratitude to our students, in Iowa, Michigan, and around the world. They have taught us what we need to teach.

ABOUT THE AUTHORS

Colin Lewis-Beck is a PhD candidate in Statistics at Iowa State University. He holds a BA from Middlebury College and a dual MPP/MA in Public Policy and Applied Statistics from the University of Michigan. While at Michigan, he received an Outstanding Teaching Award from the Department of Statistics. Also, he has worked as a teaching assistant and a computer consultant, during multiple summers at the Inter-University Consortium for Political and Social Research (ICPSR) Summer Program, University of Michigan. His research experiences in statistics are varied, including serving as a statistician in the Economic Analysis and Statistics Division of the OECD (Paris) and at STATinMED, a health outcomes research firm in Ann Arbor, Michigan. His interests are applied statistics related to social science research, causal inference, and spatial statistics. Mr. Lewis-Beck has coauthored papers on the quality of life and work productivity, modeling health care costs, and technology use in educational performance.

Michael S. Lewis-Beck is F. Wendell Miller Distinguished Professor of Political Science at the University of Iowa and holds a PhD from the University of Michigan. His interests are comparative elections, election forecasting, political economy, and quantitative methodology. He has been designated the fourth most cited political scientist since 1940, in the field of methodology. Professor Lewis-Beck has authored or coauthored more than 240 articles and books, including *Applied Regression: An Introduction, Data Analysis: An Introduction, Economics and Elections: The Major Western Democracies, Forecasting Elections, The American Voter Revisited,* and *French Presidential Elections*. He has served as an Editor of the *American Journal of Political Science*, the SAGE *QASS* series (the green monographs) in quantitative methods, and *The SAGE Encyclopedia of Social Science Research Methods*. Currently, he is Associate Editor of *International Journal of Forecasting* and Associate Editor of *French Politics*. In the spring of 2012, he held the position of Paul Lazersfeld University Professor at the University of Vienna. During the fall of 2012, he was

Visiting Professor at the Center for Citizenship and Democracy, University of Leuven (KU Leuven), Belgium. In the spring of 2013, Professor Lewis-Beck was Visiting Scholar, Centennial Center, American Political Science Association, Washington, DC. During the fall of 2013, he served as Visiting Professor, Faculty of Law and Political Science, Universidad Autónoma de Madrid, Spain. In the spring of 2014, he was Visiting Scholar, Department of Political Science, University of Göteborg, Sweden. In the fall of 2014, he served as a Visiting Professor at LUISS University, Rome. At present, he is coauthoring a book on how Latin Americans vote.

CHAPTER 1. BIVARIATE REGRESSION: FITTING A STRAIGHT LINE

Social researchers often inquire about the relationship between two variables. Numerous examples come to mind. Do men participate more in politics than women? Is the working class more liberal than the middle class? Are Democratic members of Congress bigger spenders of the taxpayer's dollar than Republicans? Are changes in the unemployment rate associated with changes in the president's popularity at the polls? These are specific instances of the common query, "What is the relationship between variable x and variable y?" One answer comes from bivariate regression, a straightforward technique that involves fitting a line to a scatter of points.

Exact Versus Inexact Relationships

Two variables, x and y, may be related to each other exactly or inexactly. In the physical sciences, variables frequently have an exact relationship to each other. The simplest such relationship between an *independent variable* (the "cause"), labeled x, and a *dependent variable* (the "effect"), labeled y, is a straight line, expressed in the formula

$$y = b_0 + b_1 x$$

where the values of the coefficients, b_0 and b_1, determine, respectively, the precise height and steepness of the line. Thus, the coefficient b_0 is referred to as the *intercept,* and the coefficient b_1 is referred to as the *slope*. The hypothetical data in Table 1.1, for example, indicate that y is linearly related to x by the following equation:

$$y = 5 + 2x$$

This straight line is fitted to these data in Figure 1.1a. we note that for each observation on x, one and only one y value is possible. When, for instance, x equals 1, y must equal 7. If x increases one unit in value, then y necessarily increases by precisely two units. Hence, knowing the x score, the y score can be perfectly predicted. A real-world example with which we are all familiar is

$$y = 32 + 9/5x$$

where temperature in Fahrenheit (y) is an exact linear function of temperature in Celsius (x).

In contrast, relationships between variables in the social sciences are almost always inexact. Practically speaking, this inexactness comes from different sources, such as faulty measures, missing observations, or improperly stated relationships. The equation for a linear relationship between two social science variables would be written, more realistically, as

$$y = b_0 + b_1x + e$$

where e is the error term, or disturbance as it is sometimes called, and represents this inexact component. A simple linear relationship for social science data is pictured in Figure 1.1b. The equation for these data happens to be the same as that for the data of Table 1.1, except for the addition of the error term,

$$y = 5 + 2x + e$$

Table 1.1 Perfect Linear Relationship Between x and y

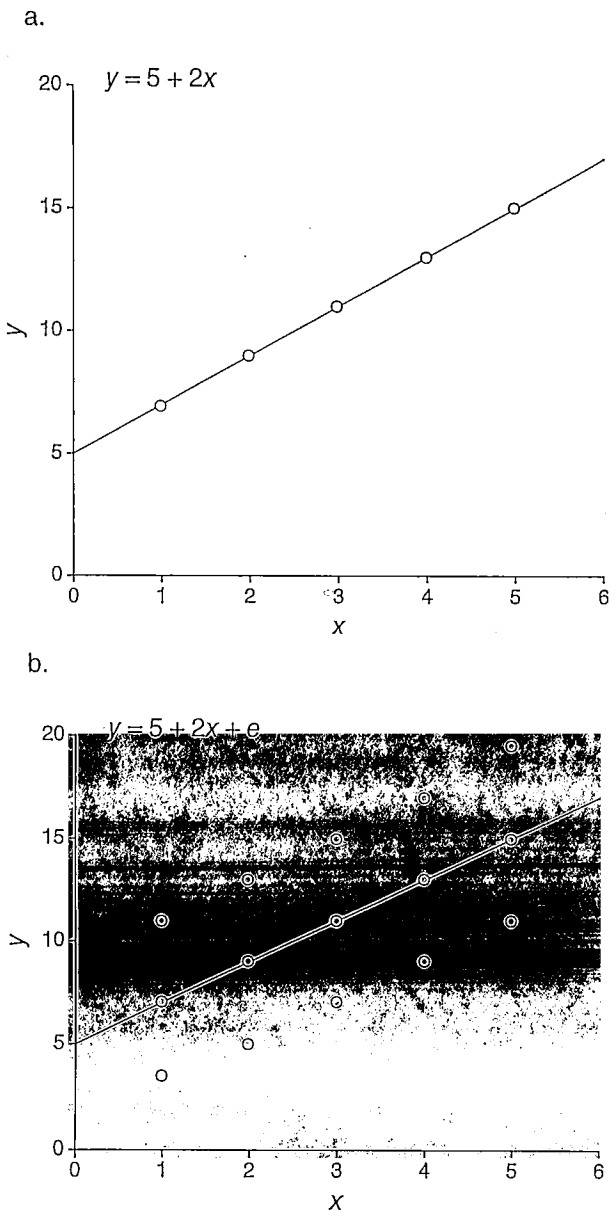
$y = 5 + 2x$	
x	y
0	5
1	7
2	9
3	11
4	13
5	15

The error term acknowledges that the prediction equation by itself, written as follows,

$$\hat{y} = 5 + 2x$$

does not perfectly predict y. (The $\hat{y}$, read y-hat, distinguishes the predicted y from the observed y.) Every y value does not fall exactly on the line. Thus, with a given x, there may occur more than one y. For example, with

Figure 1.1 (a-b) Exact and Inexact Linear Relationships Between x and y



$x = 1$ (as in Figure 1.1b), we see there is a $y = 7$, as predicted, but also there is a $y = 11$. In other words, knowing x, we do not always know y.

This inexactness is not surprising. If, for instance, x = number of elections voted in (since the last presidential election), and y = campaign contributions (in dollars), we would not expect everyone who voted in, say, three elections to contribute exactly the same amount to campaigns. Still, we would anticipate that someone voting three times would likely contribute more than someone voting one time and less than someone voting five times. Put another way, a person's campaign contribution is likely to be a linear function of electoral participation, plus some error, which is the situation described in Figure 1.1b.

The Least Squares Principle

In postulating relationships among social science variables, we commonly assume linearity, as described above. For example, in the simple two variable case, we assume the observations follow, or fall along, a straight line. Of course, this assumption is not always correct. But its adoption, at least as a starting point, might be justified on several grounds. First, numerous real relationships have been found empirically to be essentially linear. Second, the linear specification is generally the most parsimonious. Third, our theory is often so weak that we are not at all sure what the nonlinear specification would be. Fourth, inspection of the data themselves may fail to suggest a clear alternative to the straight-line model. (All too frequently, in a plot of x versus y, the figure may look like nothing so much as a large chocolate chip cookie.) Below, we focus on establishing a linear relationship between variables. Nevertheless, we should always be alert to the possibility that a relationship is actually nonlinear, following a curve of some sort. (In Chapter 4, we explicitly model the possibility that a relationship is nonlinear.)

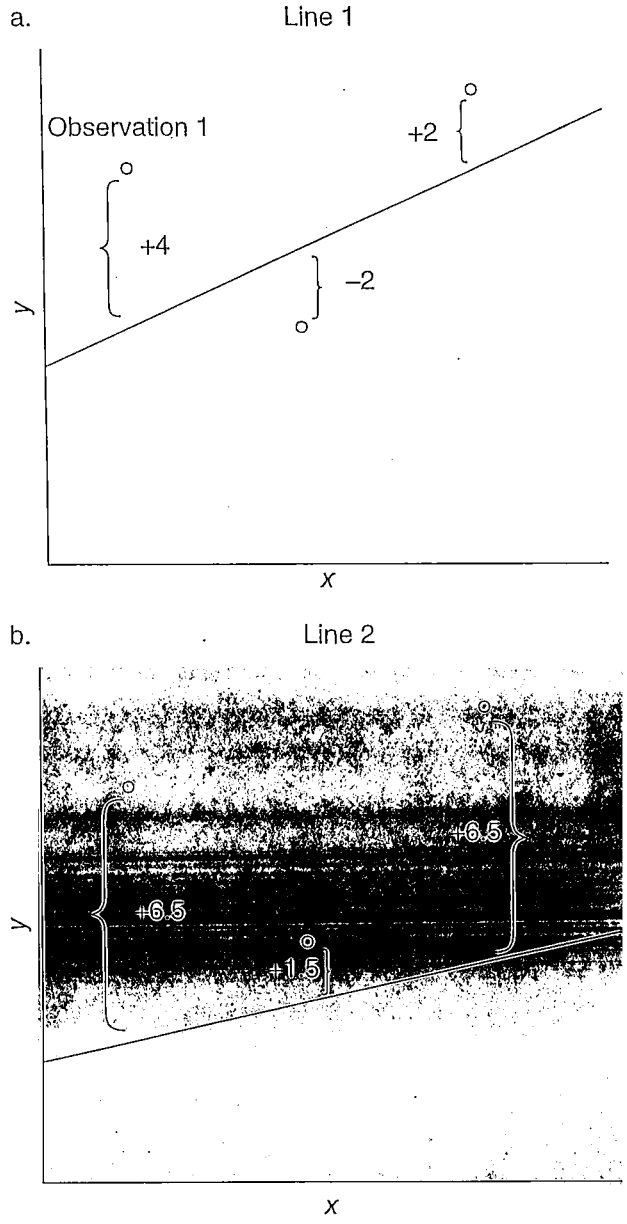
Given that we want to relate y to x with a straight line, the question arises as to which, of all possible straight lines, we should choose. For the data plotted in Figure 1.2a, we have sketched in freehand the line 1, defined by this prediction equation:

$$\hat{y} = b_{01} + b_{11}x$$

One observes that the line does not predict perfectly; for example, the vertical distance from Observation 1 to the line is four units. The *prediction error* for this Observation 1 (e_1), or any other observation, i, can be calculated as follows:

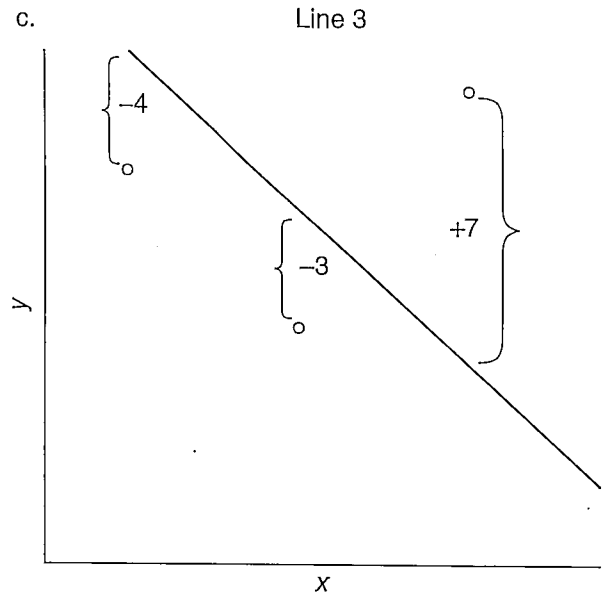
$$\text{prediction error} = e_i = \text{observed} - \text{predicted} = y_i - \hat{y}_i$$

Figure 1.2 (a–d) Straight Lines Fit to the Same Scatter of Points



(Continued)

Figure 1.2 (Continued)



Summing the prediction error for all the observations would yield a total prediction error (TPE), *total prediction error* $= \Sigma_{i=1}^{3}(y_i - \hat{y}_i) = (+4 - 2 + 2) = 4$. Clearly, line 1 fits the data better than freehand line 2 (see Figure 1.2b), represented by the equation

$$\hat{y} = b_{02} + b_{12}x$$

(*TPE for line 2* = 14.5). However, there are a vast number of straight lines besides line 2 with which line 1 could be compared. Does line 1 reduce prediction error to the minimum, or is there some other line that could do better? Obviously, we cannot possibly evaluate all the freehand straight lines that could be sketched to describe the relationship. Instead, we rely on calculus to discover the values of b_0 and b_1, which generate the line with the lowest prediction error. (Interestingly, calculus was discovered independently by mathematicians Newton and Leibnitz, working at about the same time in the 1600s.)

Before presenting this solution, however, it is necessary to modify somewhat our notion of prediction error. Note that line 3 (see Figure 1.2c), indicated by the equation,

$$\hat{y} = b_{03} + b_{13}x$$

provides a fit that is clearly less good than line 1. Nevertheless, the $TPE = 0$ for line 3. This example reveals that TPE is an inadequate measure of error, because the positive errors cancel out the negative errors (here, $-4 - 3 + 7 = 0$). One way to overcome this problem of opposite signs is to square each prediction error. (Taking the absolute value of the prediction errors is another option. However, it fails to account adequately for large errors and is computationally unwieldy. Furthermore, it makes inference problematic.) Our goal, then, becomes one of selecting the straight line that *minimizes the sum of the squares of the prediction errors* (SSE):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Through the use of calculus, it can be shown that this sum of squares is at a minimum, or "least," when the coefficients b_0 and b_1 are calculated as follows:

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These values of b_0 and b_1 are our "least squares" estimates.[1] (For a proof of the least squares solution that does not require the use of calculus, see the Appendix. The least squares method was initially arrived at by French mathematician Legendre and German mathematician Gauss, both practicing around 1800.)

Returning to the data used in the freehand examples (Figure 1.2a–c), we now apply least squares to estimate the best-fitting line, as shown in Figure 1.2d. A quick visual inspection shows that the least squares line is closer to the data than our freehand lines. Moreover, we know mathematically the property of least squares guarantees the prediction error is minimized. No other line can improve upon the least squares fit. It should also be noted that the sum of the error terms is 0 for the least squares fitted line. This is a mathematical consequence of the least squares criterion: $\sum_{i=1}^{n} e_i = 0$. (The other restriction implied by least squares is the values of the independent variable, x, must be uncorrelated with the error terms: $\sum_{i=1}^{n} e_i x_i = 0$. Using these two constraints, an interested reader can algebraically derive the same least squares solutions for the intercept and slope coefficients as shown above.)

At this point, it is appropriate to apply this least squares principle in a research example. Suppose we are studying income differences among local government employees in Riverview, a hypothetical medium-size Midwestern city. Exploratory interviews suggest a relationship between income and education. Specifically, those employees with more formal training appear to receive better pay. In an attempt to verify whether this is so, we gather relevant data. (Note that the word *data* is a plural word. Thus, it is correct to say, for example, "the data *are* gathered." It is incorrect to say that "the data *is* gathered.")

The Data

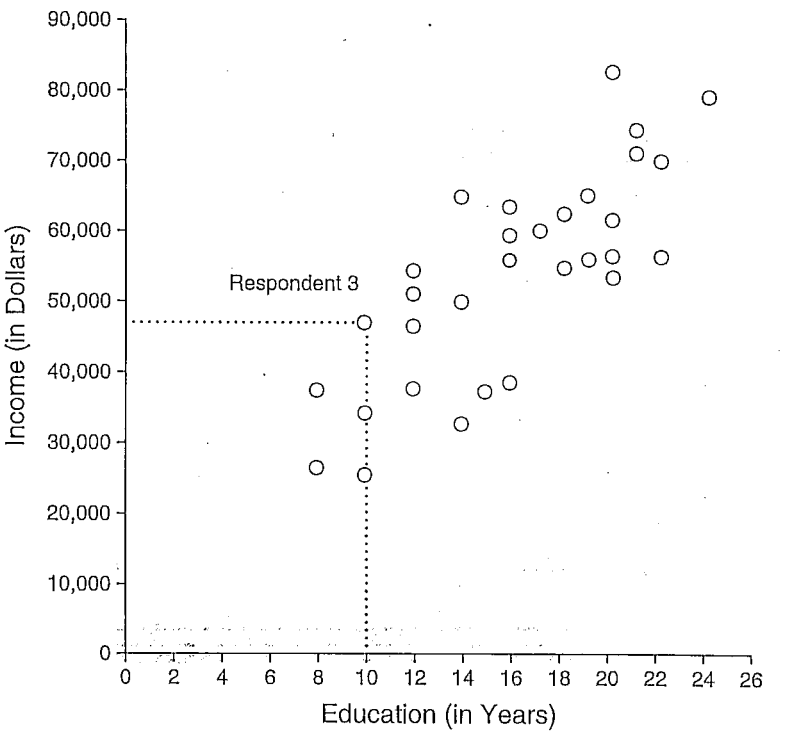
We do not have the time or money to interview the entire population: all 320 employees on the city payroll. Therefore, we decide to interview a *simple random sample* of 32, selected from the personnel list that the city clerk kindly provided.[2] (The personnel list totals 320 employees and defines the population of city employees. Our sample from this population can be represented by a lowercase "n," so we can write $n = 32$.) The data obtained on the current annual income (labeled variable y) and the number of years of formal education (labeled variable x) of each respondent are given in Table 1.2.

The Scatterplot

From simply reading the numbers in Table 1.2, it is difficult to tell whether there is a relationship between education (x) and income (y). However, the

picture becomes clearer when the data are arranged in a *scatterplot.* In Figure 1.3, education scores are plotted along the x-axis and income scores along the y-axis. Every respondent is represented by a point, located where a perpendicular line from his or her x value intersects a perpendicular line from his or her y value. (Recall from high school geometry that this is called a Cartesian coordinate.) For example, the dotted lines in Figure 1.3 fix the position of Respondent 3, who has an income of $47,034 and 10 years of education.

Figure 1.3 Scatterplot of Education and Income



Visual inspection of this scatterplot suggests the relationship is essentially linear. That is, the points huddle around a rising line that is easy to imagine, with more years of education leading to more dollars of income. Given the actual data, we can write the model as

$$y_i = b_0 + b_1 x_i + e_i \qquad i = 1,\ldots,32$$

Table 1.2 Data on Education and Income

Respondent	Education (in years) x	Income (in dollars) y
1	8	26,430
2	8	37,449
3	10	34,182
4	10	25,479
5	10	47,034
6	12	37,656
7	12	50,265
8	12	46,488
9	12	52,480
10	14	32,631
11	14	49,968
12	14	64,926
13	15	37,302
14	16	38,586
15	16	55,878
16	16	59,499
17	16	55,782
18	16	63,471
19	17	60,068
20	18	54,840
21	18	62,466
22	19	56,019
23	19	65,142
24	20	56,343
25	20	54,672
26	20	61,629
27	20	82,726
28	21	71,202
29	21	73,542
30	22	56,322
31	22	70,044
32	24	79,227

where y = respondent's annual income (in dollars), x = respondent's formal education (in years), b_0 = intercept, b_1 = slope, and e = error.

Fitting this equation by least squares yields

$$\hat{y} = 11{,}321 + 2{,}651x$$

which indicates the straight line that best fits this scatter of points (see Figure 1.4). This prediction equation is commonly referred to as a *bivariate regression equation* (or a simple regression). Furthermore, we say dependent (or outcome) variable y has been "regressed" on independent (or explanatory) variable x. And we say that this regression equation has been estimated using *ordinary least squares* (OLS for short).

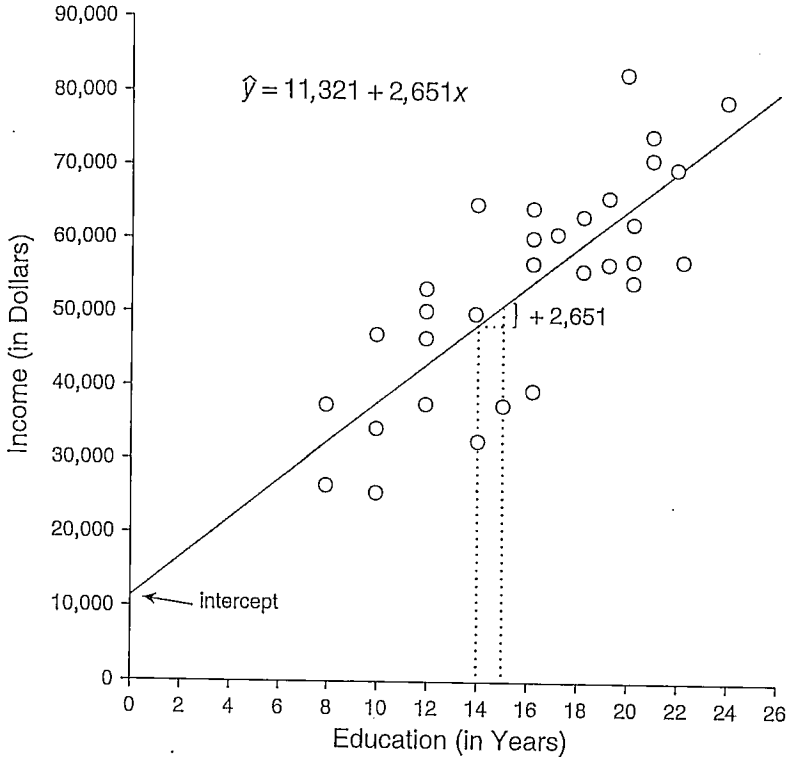
The Slope

Interpretation of the estimates is uncomplicated. Let us first consider the estimate of the slope, b_1. *The slope estimate indicates the average change in y associated with a unit change in x.* In our Riverview example, the slope estimate, 2,651, says that a 1-year increase in an employee's amount of formal education is associated with an average annual income increase of $2,651. Put another way, we expect an employee with, say, 15 years of education to have an income that is $2,651 more than an employee having only 14 years of education. We can see how the slope dictates the change in y for a unit change in x by studying Figure 1.4, which locates the expected values of y, given $x = 14$ and $x = 15$, respectively. (It is also important to recognize that the slope is a fixed value. That is, a 1-year increase in education has the same marginal effect on income for all values of x.)[3]

Note that the slope tells us only the *average* change in y that accompanies a unit change in x. The relationship between social science variables is inexact; that is, there is always error. For instance, we would not suppose that an additional year of education for any particular Riverview employee would be associated with an income rise of exactly $2,651. However, when we look at a large number of employees who have managed to acquire this extra year of schooling, the average of their individual income gains would be about $2,651.

The slope estimate suggests the average change in y *caused by* a unit change in x. Of course, this causal language may be inappropriate. The regression of y on x might support your notion of the causal process, but it

Figure 1.4 The Regression Line for the Income and Education Data



cannot establish it. To appreciate this critical point, realize that it would be a simple matter to apply OLS to the following regression equation:

$$x = b_0 + b_1 y + e$$

where now x = the *dependent* variable, and y = the *independent* variable. Obviously, such a computational exercise would not suddenly reverse the causal order of x and y in the real world. The correct causal ordering of the variables is determined outside the estimation procedure. In practice, it is based on theoretical considerations, research design, good judgment, and past research. With regard to our Riverview example, the actual causal relationship of these variables does seem to be reflected in our original model; that is, shifts in education appear likely to cause shifts in income,

but the view that changes in income cause changes in formal years of education is implausible, at least in this instance. Thus, it is only somewhat adventuresome to conclude that a 1-year increase in formal education *causes* an increase in income of $2,651, on average. (If the researcher favors a more cautious use of language here, he or she might substitute the phrase *leads to* for the word *causes*.)

The Intercept

The intercept, b_0, is so called because it indicates the point where the regression line "intercepts" the y-axis. It estimates the average value of y when x equals zero. Thus, in our Riverview example, the intercept estimate suggests that the expected income for someone with no formal education would be $11,321. This particular estimate highlights worthwhile cautions to observe when interpreting the intercept. First, one should be leery of making a prediction for y based on an x value beyond the range of the data. In this example, the lowest level of educational attainment is eight years; therefore, it is risky to extrapolate to the income of someone with zero years of education. Quite literally, we would be generalizing beyond the realm of our experience, and so may be way off the mark. (For instance, the relationship between education and income could change to a steep downward curve for individuals with less than 8 years of education.) If we are actually interested in those with no education, then we would do better to gather data on them.

A second problem may arise if the intercept has a negative value. Then, when $x = 0$, the predicted y would necessarily equal the negative value. Often, however, in the real world it is impossible to have a score on y that is below zero; for example, a Riverview employee could not receive a negative income. In such cases, the intercept is "nonsense," if taken literally. Its utility would be restricted to ensuring mathematically that a prediction "comes out right." It is a constant that must always be added on to the slope component, "b_1x," for y to be properly predicted. Drawing on an analogy from the economics of the firm, the intercept represents a "fixed cost" that must be included along with the "varying costs" determined by other factors, in order to calculate "total cost."

Prediction

Knowing the intercept and the slope, we can predict y for a given x value. For instance, if we encounter a Riverview city employee with 10 years of

schooling, then we would predict his or her income would be $37,831, as the following calculations show:

$$\begin{aligned}\hat{y} &= 11{,}321 + 2{,}651x \\ &= 11{,}321 + 2{,}651(10) \\ &= 11{,}321 + 26{,}510 \\ \hat{y} &= 37{,}831\end{aligned}$$

In our research, we might be primarily interested in prediction, rather than explanation. That is, we may not be directly concerned with identifying the variables that cause the dependent variable under study; instead, we may want to locate the variables that will allow us to make accurate guesses about the value of the dependent variable. For instance, in studying elections, we may simply want to predict winning candidates, not caring much about why they win. Of course, predictive models are not completely distinct from explanatory models. A good explanatory model may predict fairly well. Similarly, an accurate predictive model is often based on causal variables, or their surrogates. In developing a regression model, the research question dictates whether one emphasizes prediction or explanation. It is safe to conclude that, generally, social scientists stress explanation rather than prediction.

Assessing Explanatory Power: The R^2

We want to know how powerful an explanation (or prediction) our regression model provides. More technically, how well does the regression equation account for variation in the dependent variable? A preliminary judgment comes from visual inspection of the scatterplot. The closer the regression line to the points, the better the equation "fits" the data. While such "eyeballing" is an essential first step in determining the "goodness of fit" of a model, we obviously need a more formal measure, which the *coefficient of determination* (R^2) gives us.

We begin our discussion by considering the problem of predicting y. If we *only* have observations on y, then the best prediction for y is generally the estimated mean of y. Obviously, guessing this average score for each case will result in many poor predictions. However, knowing the values of x, our predictive power can be improved, provided that x is related to y. The question, then, is how much does this knowledge of x improve our prediction of y?

Figure 1.5 is a scatterplot, with a regression line fitted to the points. Consider prediction of a specific case, y_1. Ignoring the x score, the best

guess for the y score would be the mean, $\bar{y}$. There is a good deal of error in this guess, as indicated by the deviation of the actual score from the mean, $y_1 - \bar{y}$. However, by using our knowledge of the relationship of x to y, we can improve this prediction. For the particular value, x_1, the regression line predicts the dependent variable is equal to $\hat{y}_1$, which is a clear improvement over the previous guess. Thus, the regression line has managed to account for some of the deviation of this observation from the mean; specifically, it "explains" the portion, $\hat{y}_1 - \bar{y}$. Nevertheless, our regression prediction is not perfect but rather is off by the quantity $y_1 - \hat{y}_1$; this deviation is left "unexplained" by the regression equation. In brief, the deviation of y_1 from the mean can be grouped into the following components:

$$(y_1 - \bar{y}) = \text{total deviation of } y_1 \text{ from the mean, } \bar{y}$$

$$(\hat{y}_1 - \bar{y}) = \text{explained deviation of } y_1 \text{ from } \bar{y}$$

$$(y_1 - \hat{y}_1) = \text{unexplained deviation of } y_1 \text{ from } \bar{y}$$

We can calculate these deviations for each observation in our study. If we first square the deviations, then sum them, we obtain the complete components of variation for the dependent variable:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{total sum of squared deviations (TSS)}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{regression (explained) sum of squared deviations (RSS)}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{error (unexplained) sum of squared deviations (ESS)}$$

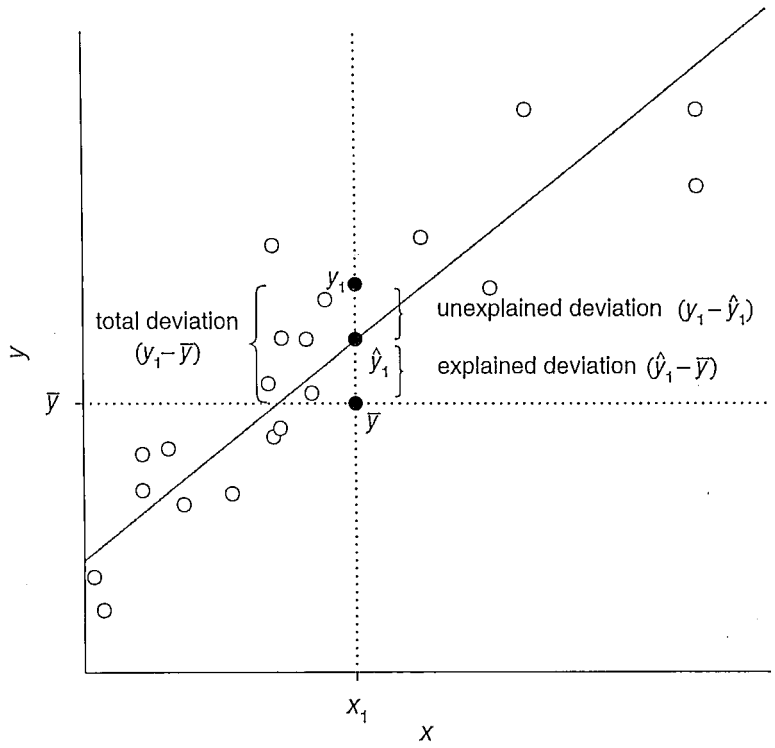
Expanding out the total sum of squared deviations term, we can derive

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

The TSS indicates the total variation in the dependent variable that we would like to explain. This total variation can be divided into two parts: the part accounted for by the regression equation (RSS) and the part the regression equation cannot account for, ESS. (We recall that the least squares procedure guarantees that this error component is at minimum.) Clearly, the larger RSS is relative to TSS, the better. This notion forms the basis of the R^2 measure:

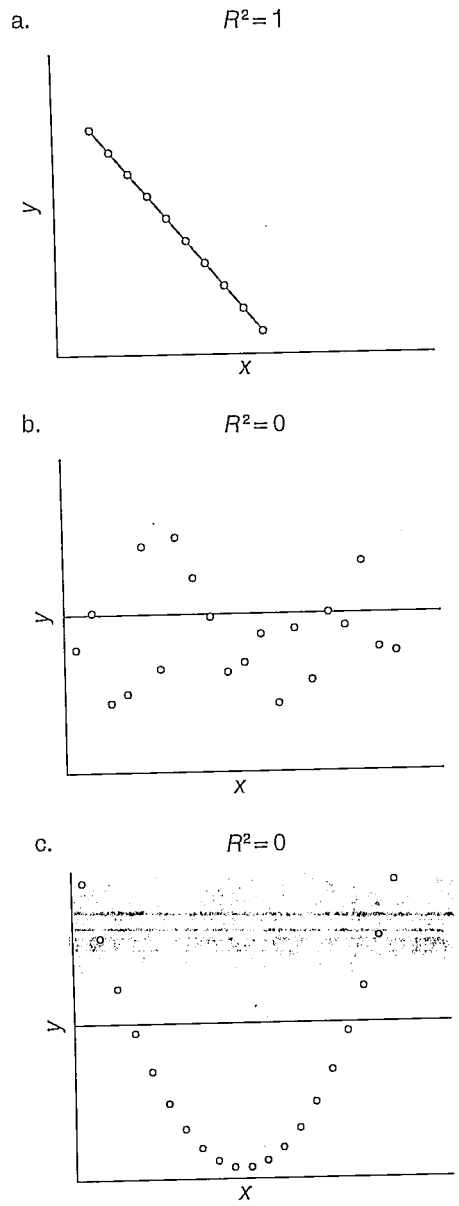
$$R^2 = \text{RSS/TSS}$$

Figure 1.5 Components of Variation in y



The coefficient of determination, R^2, indicates the linear explanatory power of the bivariate regression model. It records the proportion of variation in the dependent variable "explained" or "accounted for" by the independent variable. The possible values of the measure range from "+1" to "0." At the one extreme, when $R^2 = 1$, the independent variable completely accounts for variation in the dependent variable. All observations fall on the regression line, so knowing x enables the prediction of y without error. Figure 1.6a provides an example where $R^2 = 1$. At the other extreme, when $R^2 = 0$, the independent variable accounts for no linear variation in the dependent variable. The knowledge of x is no help in predicting y, for the two variables are totally independent of each other. Figure 1.6b gives an example where $R^2 = 0$ (note that the slope of the line also equals zero). Generally, R^2 falls between these two extremes. Then, the closer R^2 is to 1,

Figure 1.6 (a–c) Examples of the Extreme Values of the R^2



the better the fit of the regression line to the points, and the more variation in y is explained by x. In practice, when evaluating a fitted model, what constitutes a good R^2 very much depends on the discipline and type of data being analyzed. There is no universal threshold for a meaningful R^2 value. In the hard sciences, R^2 values above .90 are common, while in the social sciences, an R^2 value of .30 could be of note, especially if the data are from public opinion surveys. In our Riverview example, $R^2 = .62$. Thus, we could say that education, the independent variable, accounts for an estimated 62% of the variation in income, the dependent variable.

In regression analysis, we are virtually always pleased when the R^2 is high, because it indicates we are accounting for a large portion of the variation in the phenomenon under study. Furthermore, a very high R^2 (say about .9) is almost essential if our predictions are to be accurate. (In practice, it is difficult to attain an R^2 of this magnitude. Thus, quantitative social scientists are generally cautious in making predictions.) However, a sizable R^2 does not necessarily mean we have a *causal* explanation for the dependent variable; instead, we may have provided merely a *statistical* explanation. In the Riverview case, suppose we regressed current income, y, on income of the previous year, y_{t-1}. Our revised equation would be as follows:

$$y = b_0 + b_1 y_{t-1} + e$$

The R^2 for this new equation could be quite large (above .9), but it would not really tell us what causes income to vary; rather, it offers merely a tracking, a statistical explanation. The original equation, where education was the independent variable, provides a more convincing causal explanation of income variation, despite the lower R^2 of .62.

Even if estimation yields an R^2 that is rather small (say below .2), disappointment need not be inevitable, for it can be informative. It may suggest that the linear assumption of the R^2 is incorrect. If we turn to the scatterplot, we might discover that x and y actually have a close relationship, but it is nonlinear. For instance, the curve (a parabola) formed by connecting the points in Figure 1.6c illustrates a perfect relationship between x and y (e.g., $y = x^2$), but $R^2 = 0$. Suppose, however, that we rule out nonlinearity. Then, a low R^2 can still reveal that x does help explain y but contributes a rather small amount to that explanation. Finally, of course, an extremely low R^2 (near 0) offers very useful information, for it implies that y has virtually no linear dependency on x.

A final point on the interpretation of R^2 deserves mention. Suppose we estimate the *same* bivariate regression model for two samples from different populations, labeled 1 and 2. (For example, we wish to compare the income-education model from Riverview with the income-education model from Flatburg.) The R^2 for Sample 1 could differ from the R^2 for Sample 2,

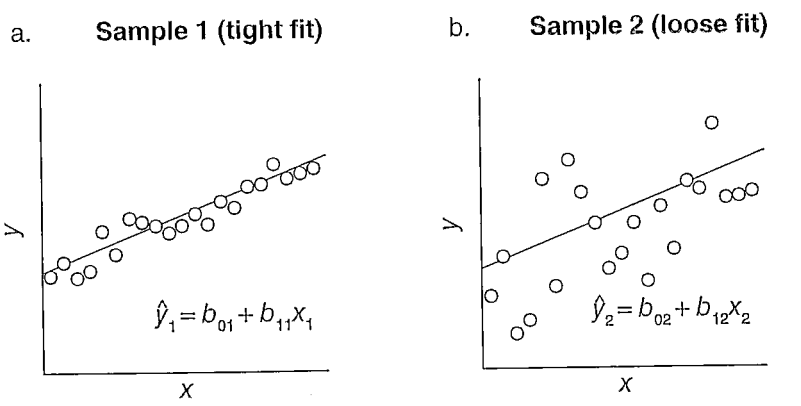
even though the parameter estimates for each were exactly the same. It simply implies that the structural relationship between the variables is the same ($b_{01} = b_{02}$; $b_{11} = b_{12}$), but it is less predictable in Population 2. In other words, the same equation provides the best possible fit for both samples but, in the second instance, is less satisfactory as a total explanation of the dependent variable. Visually, this is clear. We can see, in comparing Figure 1.7a and 1.7b, that the points are clustered more tightly around the regression line of Figure 1.7a, indicating the model fits those data better. Thus, the independent variable, x, appears a more important determinant of y in Sample 1 than in Sample 2.

Figure 1.7 (a-b) Tight Fit Versus Loose Fit of a Regression Line

$$R_1^2 > R_2^2$$

$$b_{01} = b_{02}$$

$$b_{11} = b_{12}$$



R^2 Versus r

The relationship between the coefficient of determination, R^2, and the estimate of the correlation coefficient, r, is straightforward:

$$R^2 = r^2$$

This equality suggests a possible problem with r, which is a commonly used measure of the strength and direction of a linear association, developed

by Karl Pearson.[4] That is, r can inflate the importance of the relationship between x and y. For instance, a correlation of .5 implies to the unwary reader that one half of y is being explained by x, since a perfect correlation is 1.0. Actually, though, we know that the $r = .5$ means that x explains only 25% of the variation in y (because $r^2 = .25$), which leaves fully three fourths of the variation in y unaccounted for. (The r will equal the R^2 only at the extremes, when $r = \pm 1$ or 0.) By relying on r rather than R^2, the impact of x on y can be made to seem much greater than it is. Hence, to assess the strength of the relationship between the independent variable and the dependent variable, the R^2 is the preferred measure.

Last, it should be noted there is a connection between r and the slope coefficient, b_1, in the bivariate regression setting. We can estimate the slope from the correlation coefficient between x and y using the alternative formula

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

Note that the correlation coefficient is *standardized,* with a range of ±1 (perfect negative, or positive, linear association between x and y). Also, if we first standardize x and y, the correlation coefficient will equal the slope.[5] For instance if $r_{xy} = -.30$, we can say a one-unit standard deviation increase in x will on average be associated with a −.30 standard deviation decrease for y. We are often interested, though, in making interpretations on the scale of the original data. Multiplying r_{xy} by the ratio of the standard deviation of y over the standard deviation of x will return to us the raw unstandardized coefficient, b_1, that we get from OLS.

Notes

1. $\bar{x}$, read x-bar, is an estimate of the sample mean, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

2. Statistical tests for making inferences from a sample to a population, such as the significance test, are based on a simple random sample (SRS). In our Riverview example, we could select a sample of 32 by using a random-number generator where the probability of selection is the same for all 320 employees. Practically speaking, we might apply the Systematic Selection Procedure, which simply means selecting the sample randomly from a list. This generally works well, barring a random start that taps into a relevant cycle (e.g., every tenth person is a manager).

3. Recall from high school algebra that slope is also defined as $b_1 = \frac{Rise(\Delta y)}{Run(\Delta x)}$

4. See especially his seminal papers that came out in the early 1900s, in *Biometrica* (e.g., Pearson, 1913). One formula for the sample correlation coefficient between x and y is

$$r_{xy} = s_{xy}/s_x s_y$$

where

$$s_{xy} = \text{covariance}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

and

$$s_x = standard\ deviation_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$s_y = standard\ deviation_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

5. A standardized variable (also known as a z-score) is computed by subtracting the mean from each observation and dividing by the variable's standard deviation. For a sample, $z_i = \frac{x_i - \bar{x}}{s_x}$

CHAPTER 2. BIVARIATE REGRESSION: ASSUMPTIONS AND INFERENCES

Recall that the foregoing regression results from the Riverview study are based on a *sample* of the city employees ($n = 32$). Since we wish to make accurate inferences about the actual *population* values of the intercept and slope, this bivariate regression model should meet certain assumptions. For the population, the bivariate regression model is

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

where the Greek letters indicate it is the population equation, and we have included the subscript, i, which refers to the ith observation. With the sample, we calculate

$$y_i = b_0 + b_1 x_i + e_i$$

To infer accurately the true population values, β_0 and β_1, from these sample values, b_0 and b_1, we need to satisfy necessary conditions. Inspired by the Gauss-Markov theorem, these assumptions affirm what can be called the "classical linear regression model." Different texts state these assumptions in slightly different ways, the differences depending mostly on what the author takes for granted. (For a useful review of these different statements, see Larocca [2005].) We rely heavily on the fine econometric work of Kmenta (1997, chaps. 7–10) for our formulation.

The Regression Assumptions

1. No specification error
   a. y_i is the dependent variable, x_i the independent variable
   b. No relevant independent variables have been excluded
   c. No irrelevant independent variables have been included
   d. The form of the relationship between y_i and x_i is linear
2. No measurement error
   a. The variables are quantitatively measured
   b. The variables x_i and y_i are accurately measured

3. A well-behaved error term, ε_i
   a. Zero mean: $E(\varepsilon_i) = 0$
      i. For each observation, the *expected value* of the error term is zero. (We use the symbol $E(\ )$ for expected value, which, for a random variable, is simply equal to its mean.)
   b. Homoscedasticity: $E\left(\varepsilon_i^2\right) = \sigma^2$
      i. The variance of the error term is constant for all values of x_i
   c. No autocorrelation: $E(\varepsilon_i \varepsilon_j) = 0$ $(i \neq j)$
      i. The error terms are uncorrelated
   d. The error term is uncorrelated with the independent variable: $E(\varepsilon_i x_i) = 0$
   e. Normality
      i. The error term, ε_i, is normally distributed[1]

When Assumptions 1 to 3d are met, desirable estimators of the population parameters, β_0 and β_1, will be obtained; technically, they will be the "best linear unbiased estimates," BLUE. (An unbiased estimator correctly estimates the population parameter, on average, i.e., $E(b_1) = \beta_1$.) For instance, if we repeatedly draw samples from the population, each time recalculating b_1, we would expect the average of all these b_1's to equal β_1. If the normality assumption (3e) also holds, they will still be the "best unbiased estimates," and we can carry out significance tests on them to determine how likely it is that the population parameter values differ from zero. Below, we consider each assumption in more detail.

The first assumption, absence of specification error, is critical. In sum, it asserts that the theoretical model embodied in the equation is correct. That is, x actually influences y and not vice versa. Furthermore, the functional form of the relationship conforms to a straight line. Finally, no "causal" variables have been improperly excluded or included. Let us examine the Riverview example for specification error. Visual inspection of the shape of the scatterplot (see Figure 1.3), along with the $R^2 = .62$, indicates that the relationship is essentially linear. However, it seems likely that relevant variables have been excluded, for factors besides education undoubtedly influence income. These other variables should be identified and brought into the equation, both to provide a more complete explanation and to assess the impact of education after additional forces are taken into account. (We take up this task in the next chapter.) The remaining aspect of specification error, inclusion of an irrelevant variable, argues that education might

not really be associated with income. One way to evaluate this possibility is to perform a test for statistical significance. Of course, that would not be a sufficient test. In certain cases, strong theory may dictate the presence of the variable, even in the face of a weak test result.

The need for the second assumption, no measurement error, is self-evident. If our variables are measured inaccurately, then our estimates are likely to be inaccurate. (This calls to mind the old adage "garbage in, garbage out.") For instance, with the Riverview case, suppose that in the measurement of the education variable, the respondents tended to report the number of years of schooling they would *like* to have had, rather than the number of years of schooling they *actually* had. If we were to use such a variable to indicate actual years of schooling, it would contain error, and the resulting regression coefficient would not accurately reflect the impact of actual education on income. When the analyst cannot safely rule out the possibility of measurement error, then the magnitude of the estimation problem depends on the nature and location of the error. If only the dependent variable is measured with error, then the least squares estimates should remain unbiased, at least if the error is "random." However, if the independent variable is measured with error, then the least squares estimates will be biased. In this circumstance, we have an "errors-in-variables" model, and solutions are problematic. The most oft-cited approach is *instrumental variables estimation,* but it cannot promise the restoration of unbiased parameter estimates (although the property of consistency might be achieved).

The third set of assumptions involves the error term. The initial one, a zero mean, is of little concern because, regardless, the least squares estimate of the slope is unchanged. It is true that, if this assumption is not met, the intercept estimate will be biased. Nevertheless, since the intercept estimate is often of secondary interest in social science research, this potential source of bias is rather unimportant.

Violating the assumption of homoscedasticity is more serious. While the least squares estimates continue to be unbiased, the significance tests and confidence intervals will be wrong. Let us examine Figure 1.4 from the Riverview study. Homoscedasticity would appear to be present, because the error variance appears more or less constant across the values of x; that is, the points snuggle in a band of roughly equal width above and below the regression line. If, instead, the points fanned out from the regression line as the value of x increased, the assumption would not hold, and a condition of *heteroscedasticity* would prevail. One recommended solution for this condition is a *weighted least squares* procedure. (Diagnosis of heteroscedasticity is discussed further when the analysis of residuals is considered.)

The assumption of no autocorrelation means that the error corresponding to an observation is not correlated with any of the errors for the other

observations. When autocorrelation is present, the least squares parameter estimates are still unbiased; however, the significance tests and confidence intervals are invalid. Commonly, significance tests will be much more likely to indicate that a coefficient is statistically significant when in fact it is not. Autocorrelation more frequently appears with *time-series* variables (repeated observations on the same unit through time) than with *cross-sectional* variables (unique observations on different units at the same point in time, as with our Riverview study). With time-series data, the no autocorrelation assumption requires that error for an observation at an earlier time is not related to errors for observations at a later time. If we conceive of the error term in the equation as, in part, a summary of those explanatory variables that have been left out of the regression model, then no autocorrelation implies that those forces influencing y in, say, Year 1 are independent of those forces influencing y in Year 2.[2] This assumption, it should be obvious, is often untenable. (The special problems of time-series analysis have generated an extensive literature; for a good introduction, see Enders, 2010.)

The next assumption, that the independent variable is uncorrelated with the error term, can be difficult to meet in nonexperimental research. Typically, we cannot freely set the values of x like an experimenter in a lab but rather must merely observe values of x as they present themselves in society. (That is to say, the x values are "stochastic" rather than "non-stochastic.") If this observed nonexperimental x variable is related to the error term, then the least squares parameter estimates will be biased. The simplest way to test for this violation is to evaluate the error term as a collection of excluded explanatory variables, each of which might be correlated with x. Thus, in the Riverview case, the error term would include the determinants of income other than education, such as gender of the respondent. If the explanatory variable of education is correlated with the explanatory variable of gender, but this latter variable is excluded from the equation, then the slope estimate for the education variable in the bivariate regression will be biased. For example, suppose men have higher education on average than women; then b_1 will be too large because the education variable is allowed to account for some income variation that is actually the product of gender differences. The obvious remedy, which we come to employ, is the incorporation of the missing explanatory variables into the model. (If, for some reason, an explanatory variable cannot be so incorporated, then we must trust the assumption that, as part of the error term, it is uncorrelated with the independent variable actually in the model.)

The last assumption is the normal distribution of the error term. Since we cannot observe the actual error terms (ε_i), we have to inspect their estimates, known as "residuals" (e_i), to check this assumption.[3] There are graphical as

well as statistical tests to determine if a variable conforms to a normal bell-shaped curve, with 95% of the observations falling within 2 standard deviations, plus or minus, of the mean. A visual inspection of a histogram is a quick and simple check; however, it can be misleading as the shape is affected by the scaling of the plot (e.g., changing the bin widths could change the shape a good deal). A more rigorous graphical test is a normal probability plot. The idea behind a normal probability plot is to compare the sample percentiles of the data (in this case, residuals) with theoretical percentiles from a standard normal distribution.[4] If the data are normally distributed, the sample percentiles should match the theoretical percentiles—and a scatterplot of the two will have points that roughly follow a straight line. With regard to the Riverview example, we can plot a histogram and normal probability plot of the residuals side by side (Figure 2.1a-b). In

Figure 2.1 (a-b) Checking Normality of Residuals From Riverview Simple Regression

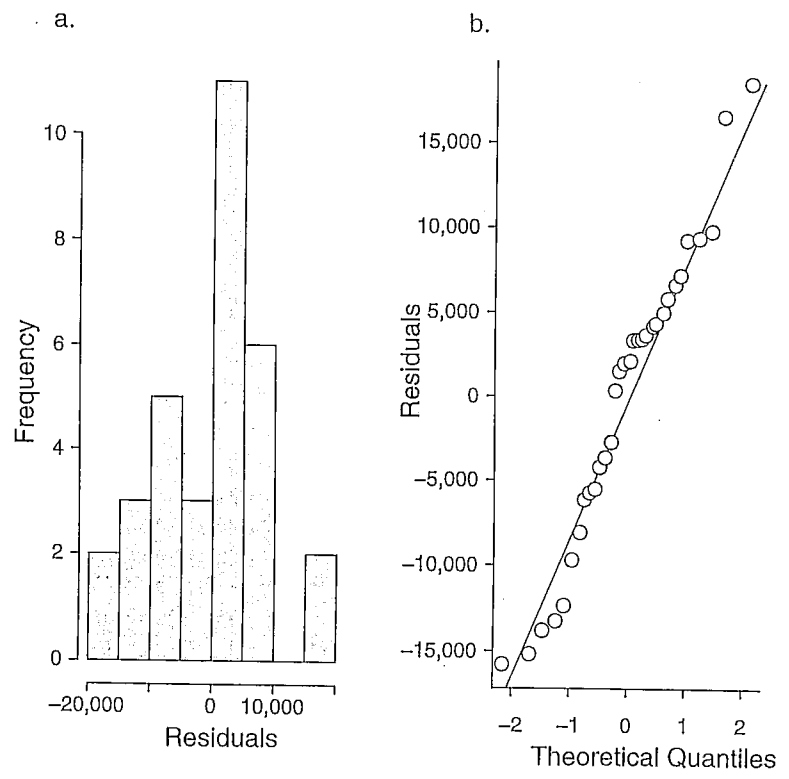


Figure 2.1a, the histogram appears roughly normal. In Figure 2.1b, the more formal probability plot gives a clearer picture. The residuals closely track the line, so suggesting normality. Experience looking at different probability plots will make it easier to diagnose violations of the normality assumption.

For a rigorous statistical test, a common test available in most statistical software programs is a Shapiro-Wilk test for normality.[5] The null hypothesis is that the residuals are normally distributed. Running the test on the residuals from the Riverview example, we fail to reject the null hypothesis at the .05 level since the p-value is quite large (p-value = .35 > .05). (More on p-values and significance tests is coming up in the next section.) Thus, it appears that Assumption 3e is satisfied. In practice, a combination of graphical and statistical tests should be employed when testing for normality.

There is some disagreement in the statistical literature over how serious violations of the regressions assumptions actually are. At one extreme, some researchers interpret their regression analyses as if the assumptions hold little practical importance. At the other extreme, some researchers feel that violations of the assumptions can render the regression results almost useless. Leamer and Leonard (1983), in a careful essay, remind us of the fragility of regression estimates. Other analysts believe that the classical linear regression assumptions can be evaluated along a continuum from 1 to 10, running from "not met at all" to "perfectly met" (Lewis-Beck, 2004, p. 938).

In fact, some are more resistant to violation than others. The normality assumption, for instance, can be ignored when the sample size is large enough, for then the Central-Limit Theorem can be invoked. (The Central-Limit Theorem indicates that the distribution of a sum of independent variables, which we can conceive of the error term as representing, approaches normality as sample size increases, irrespective of the nature of the distributions in the population.) Other violations affect only the standard errors and not unbiasedness. As an example, heteroscedasticity and autocorrelation yield invalid standard errors, but the parameter estimates remain unbiased (Long & Ervin, 2000). These violations, then, can be viewed as relatively minor. By way of contrast, the presence of specification error, such as the exclusion of a relevant variable, creates rather serious estimation problems that can be relieved only by introduction of the omitted variable. Those who wish to gain a fuller understanding of these assumption debates should consult, in addition to the efforts just cited, the treatments available in standard econometrics texts. Listing these books in order of increasing difficulty, we would recommend Woolridge (2012), Kennedy (2008), and Kmenta (1997). Furthermore, an outstanding text, by a leading sociologist, is Fox (2008).

Confidence Intervals and Significance Tests

Because social science data invariably consist of samples, we worry whether our regression coefficients actually have values of zero in the population. Specifically, is the slope (or the intercept) estimate significantly different from zero? (Of course, we could test whether the parameter estimate was significantly different from some number other than zero; however, we generally do not know enough to propose such a specific value.) Formally, we face two basic hypotheses: the null and an alternative. The *null hypothesis* states that x is not linearly associated with y; therefore, the slope is zero in the population. An *alternative hypothesis* states that x is linearly associated with y; therefore, the slope is *not* zero in the population. In summary, we have

$$H_0: \beta_1 = 0 \text{ (null hypothesis)}$$
$$H_1: \beta_1 \neq 0 \text{ (alternative hypothesis)}$$

To test these hypotheses, an interval can be constructed around the slope estimate, b_1. How do we know the distributional shape of b_1 to make a precise interval? It can be shown that the slope estimate, b_1, can be reexpressed as a linear combination of the response variable, y. Because linear combinations of normal variables are also normal, it follows that b_1 will be normally distributed and centered at the true population value, β_1 (Figure 2.2).[6]

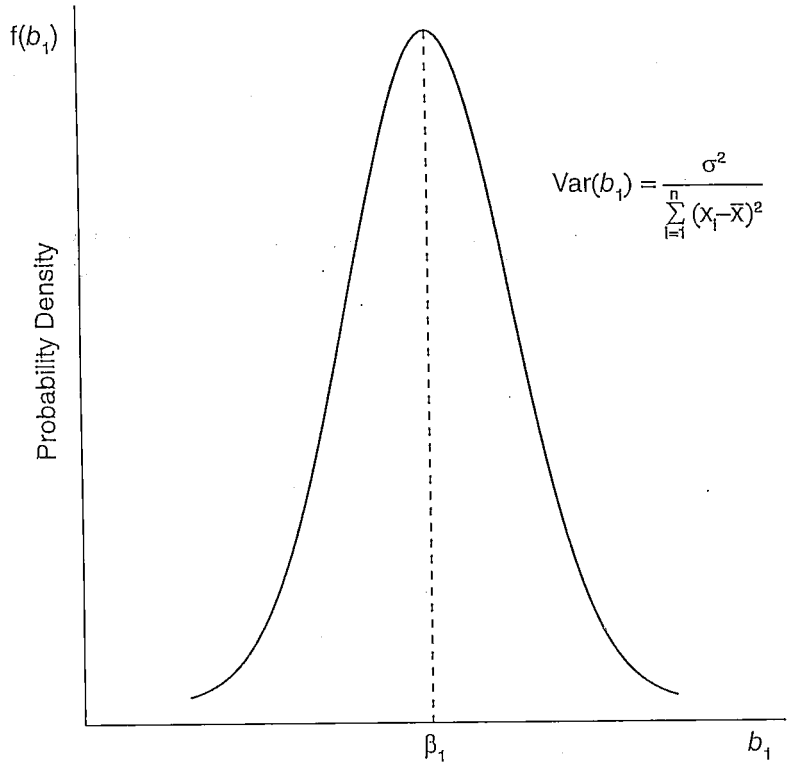
Relying on the Empirical Rule for normal distributions, we can now construct a two-sided, 95% *confidence interval* around our slope estimate, where z is the standard normal distribution:[7]

$$b_1 \pm z_{.975}\, s.e.(b_1)$$

If the value of zero does *not* fall within this interval, we reject the null hypothesis and accept the alternative hypothesis, with 95% confidence. Put another way, we could conclude that the slope, β_1, is significantly different from zero, at an alpha .05 level. (The level of *statistical significance* associated with a particular confidence interval can be determined simply by subtracting the confidence level from unity, for example, 1 − .95 = .05.)

To apply this confidence interval, we must understand the terms of the formula. These are easy enough. The term $s.e.(b_1)$ is an estimate of the standard deviation of the slope estimate, b_1, and is commonly referred to as

Figure 2.2 Distribution of the Slope Estimate, b_1



the *standard error*. It is a useful measure of the dispersion of our slope estimate. The formula for this standard error is

$$s.e.(b_1) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$= \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Expressing the formula as above reveals three important factors that determine the standard deviation of the slope. In the numerator, we have an estimate of the standard deviation of the error term (s_e), which can be seen

as an average prediction error for the model. (We discuss this further below.) Being in the numerator, as the prediction error increases, so does the standard deviation of the slope estimate. In the denominator, we have the sample size (n) as well as the variability of the independent variable, x. Therefore, as the size of the denominator increases, the standard error will decrease. One way to raise the denominator is to increase the sample size—more observations provide a more precise estimate of the slope. Another option is to increase the variability of the independent variable. If the x's are clustered together, adding an additional measurement will have a larger pull on the slope than if observations are spread out across a range of values. If possible, when designing an experiment, a researcher should strive to collect data on many different levels of the independent variable. This will improve the reliability of the slope estimate.

Because *s.e.* (b_1) is an estimate (we seldom actually know the true standard deviation of the slope), it is technically incorrect to use the normal curve to construct a confidence interval for β_1. Therefore, we replace the z distribution with the t distribution. Here we can use the t distribution with $(n-2)$ degrees of freedom. (The t distribution is quite similar to the normal distribution, especially as n becomes large, say greater than 30.) Statistical tables for the t distribution are available in many textbooks, and statistical computing software will provide exact t values.

The last component in the confidence interval formula is the subscript, ".975." This merely indicates that we are employing a 95% confidence interval but with two sides. A two-sided test means that the hypothesis about the effect of x on y is nondirectional; for example, the above alternative hypothesis, H_1, is sustained if b_1 is either significantly negative or significantly positive.

Suppose we now construct a two-sided 95% confidence interval around the regression coefficients in our Riverview study. We have

$$\hat{y} = 11{,}321 + 2{,}651x$$
$$(6{,}123) \quad (370)$$

where the figures in parentheses are the standard errors of the parameter estimates. Given the sample size is 32,

$$t_{n-2;.975} = t_{32-2;.975} = t_{30;.975} = 2.04$$

according to a t table. Therefore, the two-sided 95% confidence interval for β_1 is

$$b_1 \pm t_{n-2;.975}\, s.e.\,(b_1) = 2{,}651 \pm 2.04(370) = (2{,}651 \pm 755)$$

Thus, we are 95% confident that the true value of the population slope, β_1, is between \$1,896 and \$3,406.[8] Since the value of zero does not fall within

the interval, we reject the null hypothesis. We conclude that the slope β_1 is significantly different from zero, at the .05 level.

In the same fashion, we can construct a confidence interval for the intercept, β_0. Continuing the Riverview example,

$$b_0 \pm t_{n-2;.975}\, s.e.\ (b_0) = 11{,}321 \pm 2.04(6{,}123) = (11{,}321 \pm 12{,}491)$$

In this case, the two-sided 95% confidence band for the intercept does contain zero. We fail to reject the null hypothesis and declare that the intercept, β_0, is not significantly different from 0 at the .05 level. Graphically, this means we fail to reject the possibility that the regression line cuts the origin. However, as mentioned earlier, since there are no individuals with 0 years of education, the intercept is not substantively interpretable.

Besides providing significance tests, confidence intervals also allow us to present our parameter estimates as a range. In a bivariate regression equation, b_1 is a *point estimate*; that is, it is a specific value. The confidence band, in contrast, gives us an *interval estimate,* indicating that the slope in the population, β_1, lies within a range of values. We may well choose to stress the interval estimate over the point estimate. For example, in our Riverview study, the point estimate of β_1 is \$2,651. This is our best guess, but in reporting the results, we might prefer to say merely that a year increase in education is associated with an increase of "more or less \$2,651" a year in income. Estimating a confidence interval permits us to formalize this caution; we could assert, with 95% certainty, that a 1-year increase in education is associated with an income increase from \$1,896 to \$3,406.

In the Riverview investigation, we have rejected, with 95% confidence, the null hypothesis of no relationship between income and education. Still, we know that there is a 5% chance we are wrong. If, in fact, the null hypothesis is correct but we reject it, we commit a *Type I error*. In an effort to avoid Type I error, we could employ a 99% confidence interval, which broadens the acceptance region for the null hypothesis. The formula for a two-sided 99% confidence interval for β_1 is as follows:

$$b_1 \pm t_{n-2;.995}\, s.e.\ (b_1)$$

Applying the formula to the Riverview example,

$$2{,}651 \pm 2.75(370) = (2{,}651 \pm 1{,}018)$$

These results provide some evidence that we have not committed a Type I error. This broader confidence interval does not contain the value of zero. We continue to reject the null hypothesis but with greater confidence.

Furthermore, we can say that the slope estimate, b_1, is statistically significant at the .01 level. (This effort to prevent Type I error involves a trade-off, for the risk of *Type II error,* accepting the null hypothesis when it is false, is inevitably increased. Type II error is discussed below.)

The One-Sided Test

Thus far, we have concentrated on a two-sided test of the form

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

Often, though, our acquaintance with the phenomena under study suggests the sign of the slope. In such a circumstance, a one-sided test might be more reasonable. Taking the Riverview case, we would not expect the sign of the slope to be negative, for that would mean additional education actually decreased income.[9] Therefore, a more realistic set of hypotheses here might be

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 > 0$$

Applying a one-sided 95% confidence interval yields

$$\beta_1 > (b_1 - t_{n-2;.95}\, s.e.(b_1)) = 2{,}651 - 1.70(370) = (2{,}651 - 629) = 2{,}022$$

The lower boundary of the interval is above zero. Therefore, we reject the null hypothesis and conclude the slope is positive, with 95% confidence.

Once the level of confidence is fixed, it is "easier" to find statistical significance with a one-sided test, as opposed to a two-sided test. (The two-sided confidence interval is more likely to capture zero. For instance, the lower bounds in the Riverview case for the two-sided and one-sided tests, respectively, are \$1,896 and \$2,022.) This makes intuitive sense, for it takes into account the researcher's prior knowledge, which may rule out one of the sides from consideration.

Significance Testing: A Rule of Thumb

Recall the formula for the two-tailed 95% confidence interval for β_1:

$$b_1 \pm t_{n-2;.975}\, s.e.(b_1)$$

If this confidence interval does not contain zero, we conclude that b_1 is significant at the .05 level. We see that this confidence interval will not contain zero if, when b_1 is positive,

$$b_1 - t_{n-2;.975}\, s.e.(b_1) > 0$$

or, when b_1 is negative,

$$b_1 + t_{n-2;.975}\, s.e.(b_1) < 0$$

These requirements may be restated as

$$\frac{b_1}{s.e.(b_1)} > t_{n-2;.975}, \text{ when } b_1 \text{ is positive,}$$

or

$$\frac{b_1}{s.e.(b_1)} < t_{n-2;.975}, \text{ when } b_1 \text{ is negative.}$$

In brief, these requirements can be written

$$\left|\frac{b_1}{s.e.(b_1)}\right| > t_{n-2;.975}$$

which says that when the absolute value of the parameter estimate, b_1, divided by its standard error, $s.e.(b_1)$, surpasses the t distribution value, $t_{n-2;.975}$, we reject the null hypothesis. Thus, a significance test at the .05 level, two-tailed, can be administered by examining this ratio. The test is simplified further when one observes that, for almost any sample size, the value in the t distribution approximates 2. For example, if the sample size is only 20, then $t_{20-2;.975} = t_{18;.975} = 2.10$. In contrast, if the sample is of infinite size, $t_{\infty;.975} = 1.96$, which is the same multiplier from the normal distribution. This narrow range of values given by the t distribution leads to the following rule of thumb. If

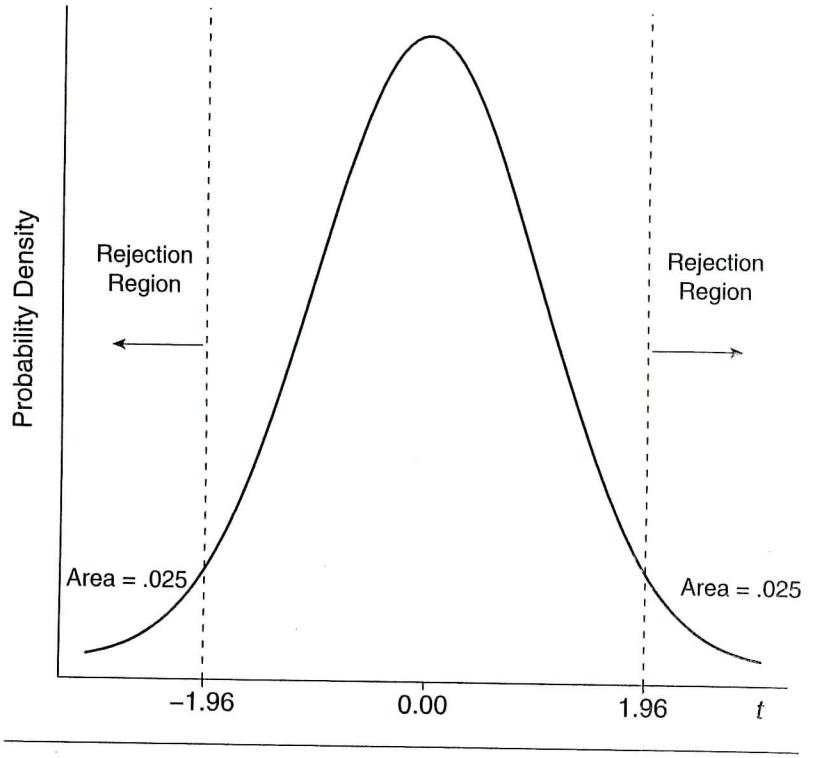
$$\left|\frac{b_1}{s.e.(b_1)}\right| > 2$$

then the parameter estimate, b_1, is significant at the .05 level, two-tailed.

This t ratio, or t-test statistic as it is commonly called, is routinely printed in the regression output of statistical software. In addition, along with the test statistic, a probability value (p-value) will be provided. A p-value is the probability of observing a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.[10] The smaller the p-value, the less likely the test statistic would be observed under the null. We reject the null hypothesis if the p-value is less than our level of statistical significance. For example, if the p-value = .02 and our level of statistical significance = .05, then we would reject the null. For a given confidence level (e.g., 95%), the decision to reject or fail to reject the null hypothesis will be the same if the t ratio is compared with a critical value (t = 1.96) or the p-value is compared with a .05 alpha level (Figure 2.3).

Rather than simply reporting whether the null hypothesis was rejected, providing a p-value is useful as it quantifies the probability of seeing the observed test statistic. It also allows the reader selection of his or her own level of significance in determining whether to reject the null hypothesis. The standard significance level is .05—that is, reject the null if the p-value is less than or equal to .05. However, certain disciplines might require a

Figure 2.3 t Distribution ($n = \infty$) for a Two-Sided Hypothesis Test, $\alpha = .05$



higher level of evidence to reject the null hypothesis (e.g., setting the significance level to .001). Below is the bivariate regression model from our Riverview example, with the reported t ratios and p-values.[11]

$$\hat{y} = 11{,}321 + 2{,}651x$$

t-statistic	(1.85)	(7.17)
p-value	(.074)	(< .001)

A quick glance at the t statistic reveals it exceeds 2 for the slope, and the p-value is less than .05; we can immediately conclude that b_1 is statistically significant at the .05 level. The intercept, however, has a p-value > .05. Thus, we fail to reject the null hypothesis that the population intercept is zero at a significance level of .05.

Reasons Why a Parameter Estimate May Not Be Significant

There are many reasons why a parameter estimate may be found not significant. Let us assume, to narrow the field somewhat, that our data compose a probability sample and that the variables are correctly measured. Then, if b_1 turns out not to be significant, the most obvious reason is that x is not a cause of y. However, suppose we doubt this straightforward conclusion. The following is a partial list of reasons why we might fail to uncover statistical significance, even though x is related to y in fact:

(1) Inadequate sample size

(2) Type II error

(3) Specification error

(4) Restricted variance in x

Below, these four possibilities are evaluated in order. (A fifth possibility is high multicollinearity, which we will consider in our discussion of multiple regression.)

As sample size increases, a given coefficient is more likely to be found significant. For instance, the b_1 value in the bivariate regression of the Riverview example would not be significant (.05) if based on only five cases but is significant with $n = 32$. This suggests it may be worthwhile for a researcher to gather more observations, for it will be easier to detect a relationship between x and y in the population, if one is present. In fact, with a very large sample, statistical significance can be uncovered even if

b_1 is substantively quite small. (For very large samples, such as public opinion surveys of several thousand respondents, significance may actually be "too easy" to find, since tiny coefficients can be statistically significant. In this situation, the analyst might prefer to rely primarily on a substantive judgment of the importance of the coefficient. That is, ask if the coefficient is "substantively significant.")

Let us suppose that sample size is fixed and turn to the problem of choosing a significance level, as it relates to Type II error. In principle, we could set the significance test at any level between 0 and 1. In practice, however, most social scientists employ the .05 or .01 levels. To avoid the charges of arbitrariness or bias, we normally select one of these conventional standards before analysis begins. For instance, suppose prior to our investigation, we decide to employ the .01 significance level. Upon analysis, we find b_1 is not significant at this .01 level, but we observe that it is significant at the less demanding level of .05. We might be loath to accept the null hypothesis as dictated by the .01 test, especially since theory and prior research indicate that x does influence y. Technically, we worry that we are committing a Type II error, accepting the null when it is false. In the end, we may prefer to accept the results of the .05 test. (In this particular case, given the strength of theory and previous research, perhaps we should have initially set the significance test at the less demanding .05 level.)

Aside from Type II error, b_1 may not appear significant because the equation misspecifies the relationship between x and y. Perhaps the raw relationship follows a curve, rather than a straight line, as assumed by the regression model. First, this curvilinearity should be detectable in the scatterplot. To establish the statistical significance of the relationship in the face of this curvilinearity, regression analysis might still be applied, but the variables would have to be properly transformed. (We pursue an example of such a transformation at the end of this chapter.)

Finally, a parameter estimate may not be found significant because the variance in x is restricted. Look again at the formula for the standard error of b_1,

$$s.e.(b_1) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

As mentioned earlier, we see that as the dispersion of x about its mean decreases, the value of the denominator decreases, thereby increasing the standard error of b_1. Other things being equal, a larger standard error makes statistical significance more difficult to achieve, as the t-ratio formula

makes clear. The implication is that b_1 may not be significant simply because there is too little variation in x. (The degree of variation in x is easily checked by evaluating its standard deviation.) In such a circumstance, the researcher may seek to gather more extreme observations on x before making any firm conclusions about whether it is significantly related to y.

The Prediction Error for y

In regression analysis, the difference between the observed and the estimated value of the dependent variable, $y_i - \hat{y}_i$, equals the prediction error for that case. The variation of all these prediction errors around the regression line can be estimated from a sample as follows:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

This s_e is called the *standard error of estimate of y;* that is, the estimated standard deviation of the actual y from the predicted y. This quantity is also referred to as the *Root Mean Squared Error,* or RMSE for short. We see that the formulation provides something close to an average prediction error for the model.[12] When prediction is without error, RMSE = 0. If the researcher's goal is merely prediction, the RMSE can be a good metric to compare the accuracy of various models. However, unlike R^2, which is bounded at both ends (1 = perfect linearity, 0 = no linearity), RMSE has no upper bound: It is dependent on the units of measurement and the size of the prediction error. Thus, by itself, RMSE is not typically a useful measure of model fit.

The Root Mean Squared Error is also used to make confidence intervals for y at a fixed x value. The specific x we are interested in making a prediction for is denoted by x^* in the equation below. A 95% confidence interval for y is constructed as follows:

$$\hat{y} \pm t_{n-2;.975} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

This equation reveals that the width of the prediction interval depends additionally on the variation in x and the sample size, n. If we predict y at the mean ($x^* = \bar{x}$), the numerator in the last term under the square root becomes zero. In other words, the width of the confidence interval is minimized at the

mean. This matches intuition: We would expect our prediction to be more accurate at the center of the data than at values far away from the mean. We can also see that as n increases, the confidence interval narrows. This provides another way for the researcher to reduce the width of the confidence interval: Collect more data to increase the sample size.

Let us take an example. In the Riverview study, we would predict someone with 10 years of education to have an average income of

$$\hat{y} = 11{,}321 + 2{,}651(10) = 37{,}831$$

How accurate is this prediction? For $x = 10$, we have this 95% confidence interval ($s_e = 8{,}978$):

$$37{,}831 \pm t_{30;.975}(8{,}978)\sqrt{\frac{1}{32} + \frac{(10-16)^2}{590}} = 37{,}831 \pm 5{,}569$$

According to this confidence interval, we are 95% confident that a city employee with 10 years of education has an average annual income between $32,262 and $43,400. We can see this confidence interval graphically in Figure 2.4. The fitted regression line is in solid black; the dashed lines are the upper and lower 95% confidence bands. As mentioned above, it is clear the width of the confidence interval is most narrow around the mean ($x = 16$) and widens in ranges of education with fewer observations.

A last point merits mention. The above confidence interval provides a confidence interval for the average or expected value of y given x. However, if we are interested in predicting a new value of y, an additional correction is required. When constructing a prediction interval for a future y, the interval will be wider since it must also account for variation in y around its mean. The formula for constructing a prediction interval is readily available (see Gujarati & Porter, 2009, pp. 126–129).

Analysis of Residuals

The prediction errors from a regression model, $e_i = y_i - \hat{y}_i$, are, as we know, also called *residuals*. Analysis of these residuals can help us detect the violation of certain regression assumptions. Residuals can be first plotted against the fitted values of the model, $\hat{y}$. It is also useful to examine scatterplots of the residuals against independent variables (x_i). In a visual inspection of the residuals, we hope to observe a healthy pattern similar to that in Figure 2.5a; that is, the points appear scattered randomly about in a steady band above and below the zero line. (To be clear, this zero line is equivalent to the

impression receives quantitative confirmation. A simple sign count reveals a roughly even balance around the line.

The Riverview data are from government workers in a midsized Midwestern town. Suppose as one last step, we want to test the external validity of our findings (education has a positive effect on income); therefore, we collect data on workers in a larger city. Once again, we take a random sample ($n = 32$) of workers (from a variety of public employment sectors) and regress income on education. We get the following model output:

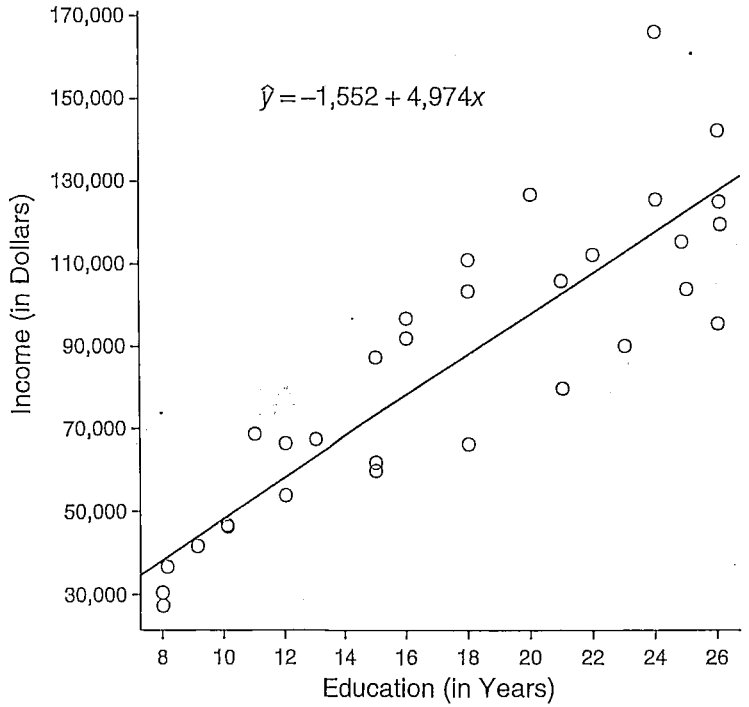
$$\hat{y} = -1{,}552 + 4{,}974x$$

t-statistic	(–0.18)	(10.22)
p-value	(0.86)	(<.001)

$R^2 = .78 \qquad n = 32 \qquad s_e = 17{,}250$

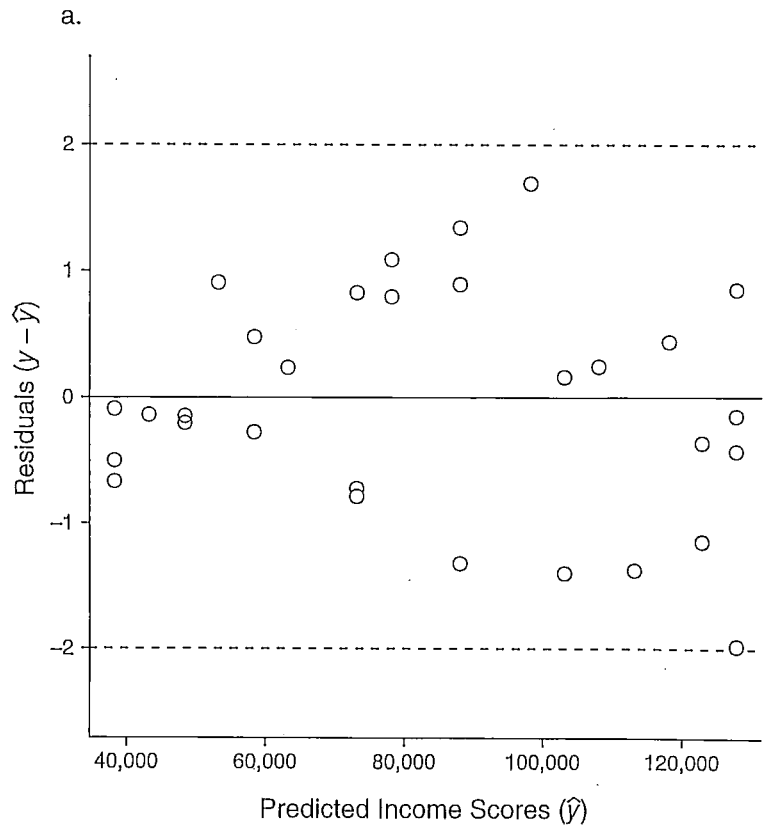
(Original Large City Model)

Figure 2.8 Regression Line for Large City Sample Data



From looking at Figure 2.8, it appears the relationship between education and income is still linear. The coefficient on education is higher now ($4,974), but that makes intuitive sense—we would expect returns on education to be higher in a larger metropolitan area. A closer look at the scatterplot, however, reveals that as education increases, there is more variability around the regression line. In a larger city, people with more education may have additional job opportunities that come with increasingly varying salaries. A plot of the residuals highlights this funnel pattern (Figure 2.9a).

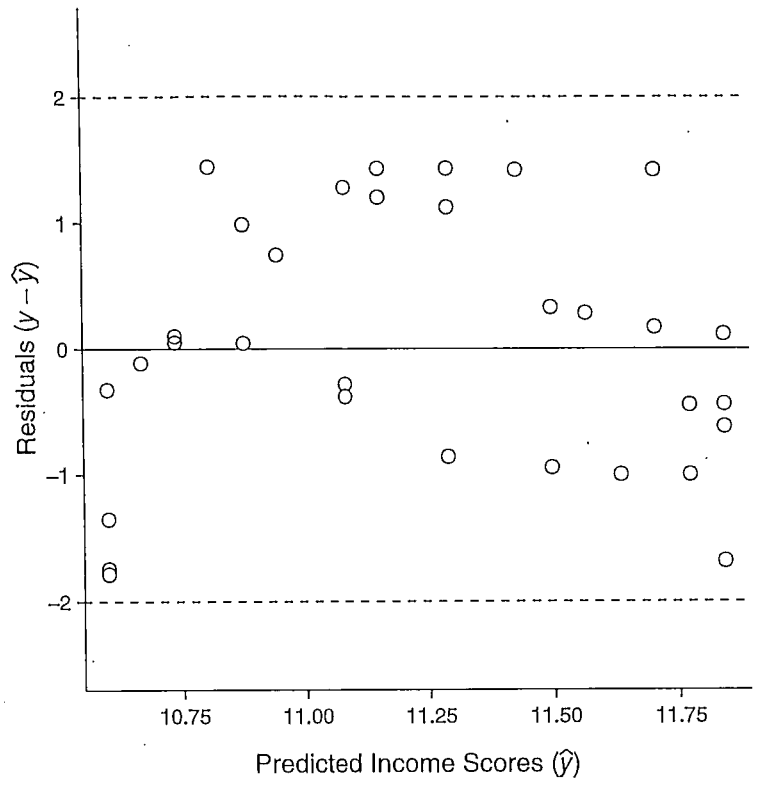
Figure 2.9 (a-b) Heteroscedastic Residuals



(Continued)

(Continued)

b.



As discussed earlier, violations of homoscedasticity make statistical inference problematic. The coefficients are still unbiased, and if the model is solely for prediction, heteroscedasticity can be ignored. However, if we want to calculate *p*-values and test the coefficients for statistical significance, Assumption 3b is necessary. The good news: The linear model is flexible, and numerous transformations are available to stabilize the variance. A common transformation is the natural-log (ln) transformation.[13] The log transformation is useful because it is simple to implement and gives a nice interpretation of the coefficients in terms of percent change. Either the independent, dependent, or both variables can be transformed. Here we will take the log of the dependent variable, income. After fitting a model to the transformed *y*, we get the following output:

$$\ln(\hat{y}) = 10.01 + 0.07x$$

t-statistic	(83.84)	(10.52)
p-value	(<.001)	(<.001)

$R^2 = .79$ $n = 32$ $s_e = 0.23$

(Logged Large City Model)

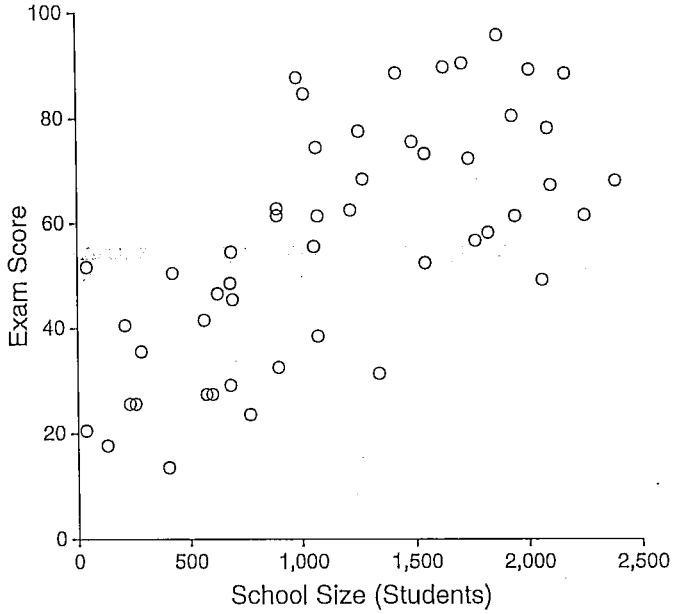
The estimated slope for education is $b_1 = 0.07$. Recall this coefficient was estimated on the ln(*y*), which changes the interpretation of b_1: A one-unit increase in education increases salary by an average of 7%. As a final check, we examine the residuals after the transformation (Figure 2.9b). The clear heteroscedastic fan pattern has disappeared, and the points now appear as white noise scattered around a mean of zero. The homoscedasticity assumption of the error terms no longer seems violated. Furthermore, comparing the *t* ratios of the two models, original ($t = 10.22$) versus logged ($t = 10.52$), we see that the latter value attains a slightly greater magnitude, making the inference of statistical significance a bit more secure. However, in this case, the gain is so slight that the analyst might ask whether it is worth returning to the use of the original unlogged *y* because of interpretation gains. (This is a typical practical decision research workers must make.)

The Effect of School Size on Educational Performance: A Bivariate Regression Example

It is time to apply what we have learned to some data that address an important social issue. A current public policy controversy concerns the impact of school size on educational outcomes (Leithwood & Jantzi, 2009). Our specific research question is, "Does school size have an effect on educational test scores?" To provide an answer, we have simulated a data set on variables expected to be relevant. While the data set itself is constructed, it does speak to a real problem and in a pedagogical way that we hope leads to further, applied research. The data consist of a random sample of 50 public high schools ($n = 50$) across the country. Each school evaluated its student body performance using the same measurement procedure. With respect to the dependent variable, we have a standardized test score (range: 0–100), averaged per high school. With respect to the independent variable, we have school size, measured as the total number of students enrolled at the beginning of the school year.

Before fitting a regression line, let us first examine a scatterplot of the data. The relationship appears to be positive and somewhat linear (Figure 2.10).

Figure 2.10 Scatterplot of School Size and Exam Score



A bivariate regression of test scores, y, on school size, x, yields the following:

$$\hat{y} = 28.76 + 0.02x$$

t-statistic	(83.84)	(10.52)
p-value	(<.001)	(<.001)

$R^2 = .48 \quad n = 50 \quad s_e = 16.35$

Although the coefficient on school size appears numerically small, it is statistically significant at an alpha level of .001. Rather than interpreting the marginal effect of adding one additional student ($b_1 = 0.02$), a more intuitive interpretation is to think of adding 100 students to a high school. The model predicts the average high school test score will increase by 2 points ($100b_1$) per 100 additional students. It might be unrealistic, though, to suspect that the relationship between school size and educational test scores is strictly linear. At some level, a school might reach a saturation point where the benefits from additional students diminish or become

negative. Incorporating a quadratic term into the model is one way to test this hypothesis. Despite requiring linearity, OLS can incorporate a variety of nonlinear relationships through transformations of the raw variables, transformations that effectively linearize the relationship. (We will explore this further in Chapter 4.) Reestimating the equation, but with an additional x^2 term, yields

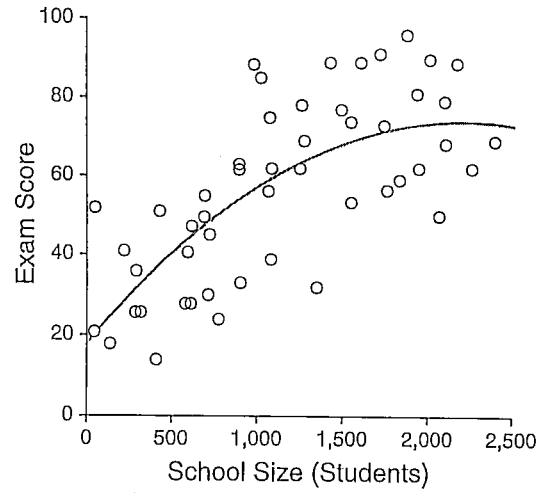
$$\hat{y} = 17.85 + 0.05x - .000012x^2$$

t-statistic	(2.56)	(3.74)	(–2.06)
p-value	(.01)	(.001)	(.045)

$R^2 = .53 \quad n = 50 \quad s_e = 15.83$

Our explanation of the exam scores has improved. The quadratic term is statistically significant, and the R^2 has improved to .53. In the earlier equation, when y is predicted with just school size, the average error is 16.35. The RMSE is now reduced to 15.83 in our revised model. While the size coefficient on the squared coefficient is not immediately interpretable, the negative sign indicates that at some point, as school size increases, the effect of school size on exam scores begins to diminish.[14] From looking at a fitted line for the new model (Figure 2.11), it appears that at around 1,500

Figure 2.11 Quadratic Fit



students, the effect of school size becomes flat, even slightly negative. Of course, although school size is an important determinant in predicting exam scores, it is not the only one, as we will discover in the next chapter.

Notes

1. Combined with Assumptions 3a and 3b, the normal distribution of the errors is often written as $\varepsilon_i \sim N(0, \sigma^2)$.
2. One might wonder why omitted explanatory variables are not simply incorporated into the equation, thus solving for autocorrelation and specification error at the same time. Lamentably, this straightforward solution is not possible when these variables are either unknown or unmeasured.
3. Although the error terms and residuals appear interchangeable, they have different properties. The error term captures the difference between the observed and true (unobserved) response value (i.e., $y_{obs(i)} - y_{pop(i)} = \varepsilon_i$), while the residual is the difference between the sample and prediction estimate (i.e., $y_{obs(i)} - y_{hat(i)} = e_i$).
4. The term *percentile* is probably most familiar from standardized exam results. If an individual scores in the 80th percentile, it means he or she scored higher than 80% of all test takers.
5. There are many different statistical packages available, with varying degrees of difficulty in terms of user friendliness. We refrain from recommending specific software as these programs are updated frequently and wax and wane in popularity. Currently, the most common statistical software packages, in order of increasing complexity, are as follows: SPSS, Stata, SAS, and R.
6. The numerator in the variance of b_1 (σ^2) denotes the population error variance. As mentioned earlier, we never see the true errors, so we estimate σ^2 using the sample estimator denoted as s_e^2.
7. The Empirical Rule states that for normally distributed variables, approximately 95% of all observations will fall within plus or minus 2 standard deviations of the mean. The exact value is 1.96 standard deviations, which we get from calculating the value of $z_{.975}$ from a normal table.
8. It is possible the true population slope is not in this interval. Confidence intervals are a *frequentist* idea based on repeated sampling. For example, if 100 confidence intervals for β_1 are constructed based on different samples, statistical theory says approximately 95 will contain the true population slope. We are only making a confidence interval based on one sample, but we are hoping (confident) that our interval is one of the 95 that contains the true population slope, β_1.
9. This idea of using a prior belief about a probability event (e.g., assuming the slope must be positive if it is not zero) is an elementary application of Bayes' Theorem.
10. Mathematically, we define the p-value as follows: $p = Probability(|T| \geq |t|)$, where t is our observed t-statistic, and T is a random variable for all possible t ratios.

11. When reporting output from a regression model, it is convention to give a measure of statistical significance for each coefficient; that can be a t ratio, a standard error, or a p-value.
12. Although the formulation looks like an average, we divided by $n-2$ rather than the sample size, n. As the sample size increases, this difference becomes trivial. We subtract 2 from n because two *degrees of freedom* are lost when we estimate the two model parameters, β_0 and β_1. The estimation of these two parameters means that two residuals are not "free" to be independent; rather, they are fixed by the estimation process. In general, when we estimate a model from a sample, we must adjust for degrees of freedom. For more on the topic, see Lewis-Beck (2004, pp. 243–244).
13. The natural-log transformation is also useful for handling skewed distributions or turning a multiplicative relationship into a linear one. Another useful logarithmic transformation is the common log, which is to the base 10. Its advantage is that a unit change in this logged x can be interpreted as the expected change in raw y when raw x changes tenfold.
14. As seen in Figure 2.11, with the added quadratic term, the slope is no longer constant. How do we interpret b_1? It is the rate of change (slope) when $x = 0$. With the help of calculus, it is straightforward to find the slope at different values of x (school size). Frequently, researchers are interested in the slope at the average of the independent variable, $\bar{x}$. Here, the slope at the average school size (1,146 students) is $b_1 - 2b_2\bar{x} = .05 - .000012(1{,}146) = 0.04$.

CHAPTER 3. MULTIPLE REGRESSION: THE BASICS

With multiple regression, we can incorporate more than one independent variable into an equation. This is useful in two ways. First, it almost inevitably offers a fuller explanation of the dependent variable, since few phenomena are products of a single cause. Second, the effect of a particular independent variable is more certain, for the possibility of distorting influences from the other independent variables is removed. The procedure is a straightforward extension of bivariate regression. Parameter estimation and interpretation follow the same principles. Likewise, the significance test and the R^2 are parallel. Furthermore, the bivariate regression assumptions necessary for best linear unbiased estimates (BLUE) are carried over to the multivariate case. The technique of multiple regression has great range, and its mastery will enable the researcher to analyze almost any set of quantitative data.

The General Equation

In the general multiple regression equation, the dependent variable is seen as a linear function of more than one independent variable,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

where the subscripts from 1 to k identify the independent variables. The elementary three-variable case, which we shall be using below, is written

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and suggests that y is determined by x_1 and x_2, plus an error term.

To estimate the parameters, we again employ the least squares principle, minimizing the sum of the squares of the prediction errors (SSE):

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

For the three-variable model, this estimated least squares equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

The least squares combination of values for the coefficients (b_0, b_1, b_2) yields less prediction error than other possible combinations of values.

Hence, the least squares equation fits the set of observations better than any other linear equation. However, it can no longer be represented graphically with a simple straight line fitted to a two-dimensional scatterplot. Rather, we must imagine fitting a *plane* to a three-dimensional scatter of points. The location of this plane, of course, is dictated by the values of b_0, b_1, and b_2, which are determined by calculus. For most of us, it is impossible to visualize the fitting of equations with more than three variables. It might work, if the fourth variable were time (i.e., we could imagine a plane in a three-dimensional box moving through time). However, for the general case, with k independent variables, our visual geometry fails. Then, we must rely on the mathematics of the fitting, which requires conceiving of adjusting a k-dimensional hyperplane to $(k+1)$–dimensional scatter.

For purposes of illustration, let us look at a simple three-variable model from our Riverview study. On the basis of our earlier work, we believe income is related to education. But we know that education is not the only factor influencing income. Another factor is undoubtedly seniority. In most occupations, the longer one is on the job, the more money one makes. This seems likely to be so in Riverview city government. Therefore, our explanation for income differences should be improved if we move from our bivariate regression model to this multiple regression model:

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

where y = income (in dollars), x_1 = education (in years), x_2 = seniority (in years), and e = error. The least squares estimates for the parameters are as follows:

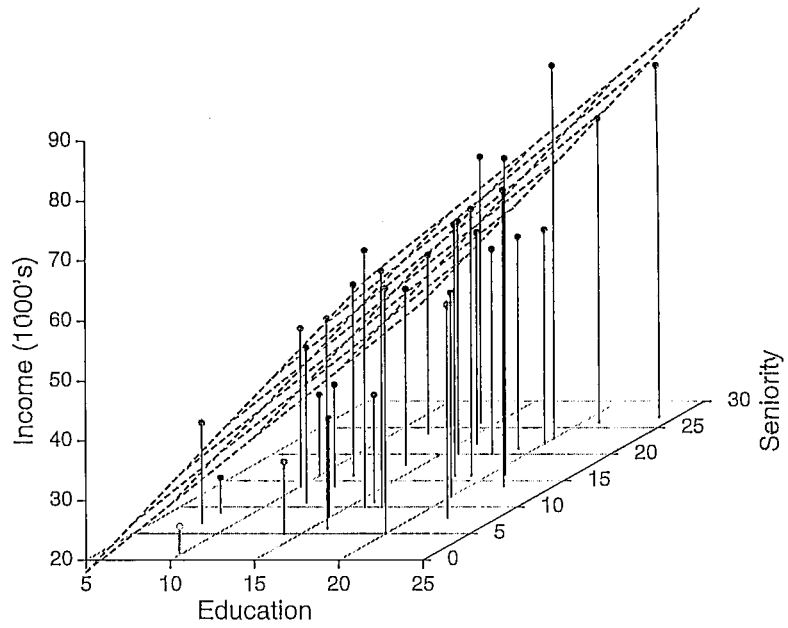
$$\hat{y} = 6{,}769 + 2{,}252x_1 + 739x_2$$

Graphically, we can see the fitted model in Figure 3.1 with education and income on the x- and y-axes and seniority on the z-axis. This graph is simply an extension of our scatterplot from Chapter 1 (Figure 1.4), but with one additional variable (dimension) added. The dashed plane that cuts through the data is the least squares fit. No other plane that passes through this three-dimensional cloud of points achieves a lower SSE.

Interpreting the Parameter Estimates

The interpretation of the intercept, which merely extends the bivariate case, need not detain us: b_0 = the average value of y when each independent variable equals zero. The interpretation of the slope, however, requires more

Figure 3.1 Three-Dimensional Plot of Education, Seniority, and Income



attention: b_k = the average change in y associated with a unit change in x, *when the other independent variables are held constant.* By this means of control, we are able to separate out the effect of x_k itself, free of any distorting influences from the other independent variables. Such a slope is sometimes called a *partial slope, or partial regression coefficient.* In the above Riverview example, partial slope b_2 estimates that a 1-year increase in seniority is associated with an average income rise of $739, assuming the employee's amount of education remains constant. In other words, a city worker can expect this annual salary increment, independent of any personal effort at educational improvement. Nevertheless, according to b_1, acquiring an additional year of schooling would add to an employee's income, regardless of the years of seniority accumulated. That is, an extra year of education will augment income an average of $2,252, beyond the benefits that come from seniority.

To appreciate fully the interpretation of the partial slope, one must grasp how multiple regression "holds constant" the other independent variables. First, it involves *statistical control* rather than *experimental control.* For instance, in our Riverview study, if we were able to exercise experimental

control, we might hold everyone's education at a constant value, say 10 years, and then record the effect on income of assigning respondents different amounts of seniority. To assess the effect of education on income, a similar experiment could be carried out. If such manipulation were possible, we could begin to analyze the effects of seniority and education, respectively, by running two separate bivariate regressions, one on each experiment. However, since such experimental control is out of the question, we have to rely on the statistical control multiple regression provides. We can show how this statistical control operates to separate the effect of one independent variable from the others by examining the formula for a partial slope.

We confine ourselves to the following three-variable model, the results of which are generalizable:

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

Let us explicate the b_1 estimation. Assuming $r_{12} \neq 0$, each independent variable can be accounted for, at least in part, by the other independent variables. That is, for example, x_1 can be written as a linear function of x_2,

$$x_1 = c_1 + c_2x_2 + u$$

Supposing x_1 is not perfectly predicted by x_2, there is error, u. Hence, the observed x_1 can be expressed as the predicted x_1, plus error:

$$x_1 = \hat{x}_1 + u$$

where $\hat{x}_1 = c_1 + c_2x_2$. The error, u, is the portion of x_1 that the other independent variable, x_2, cannot explain,

$$u = x_1 - \hat{x}_1$$

This component, u, thus represents a part of x_1 that is completely separate from x_2.

By the same steps, we can also isolate the portion of y that is linearly independent of x_2:

$$\begin{aligned} y &= d_1 + d_2x_2 + v \\ &= (d_1 + d_2x_2) + v \\ y &= \hat{y} + v \end{aligned}$$

The error, v, is that portion of y that cannot be accounted for by x_2,

$$v = y - \hat{y}$$

This component, v, then, is that part of y that is unrelated to x_2.

These two error components, u and v, are joined in the following formula for b_1:

$$b_1 = \frac{\sum_{i=1}^{n}(u_i)(v_i)}{\sum_{i=1}^{n}u_i^2} = \frac{\sum_{i=1}^{n}(x_{1i} - \hat{x}_{1i})(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(x_{1i} - \hat{x}_{1i})^2}$$

In words, *b_1 is determined by x_1 and y values that have been freed of any linear influence from x_2.* In this way, the effect of x_1 is separated from the effect of x_2. The formula, generally applicable for any partial slope, should be familiar, for we saw a special version of it in the bivariate case, where

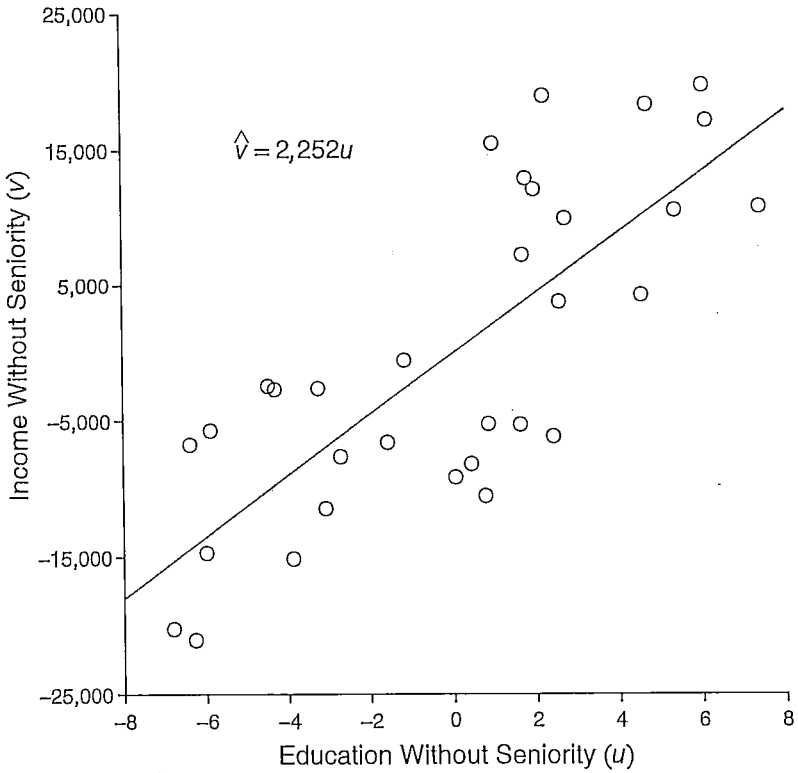
$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

A useful graph to visualize the partial effect of an independent variable on the response is an *added-variable plot,* which plots u against v as defined above. Just as a scatterplot of x versus y in simple regression allows us to visualize the correlation between one independent variable and the response, an added-variable plot allows us to see the relation between an independent variable and y while accounting for all additional explanatory variables in the model. Returning to the Riverview study, we can make an added-variable plot to see the partial effect of education (x_1) on income.

As exemplified in Figure 3.2, there still appears to be a positive correlation between education and income even after we remove the effect of seniority. Moreover, if we fit a regression line to this scatterplot, the slope coefficient is 2,252. This is the same value as b_1: the coefficient on education in the three-variable regression model fit at the beginning of the chapter. On the other hand, if there was no effect of education on income after including seniority in the model, the added-variable plot would look like a random scatter of points. In general, when building a regression model, added-variable plots are a useful way to determine graphically whether to include a new independent variable into the model (hence the name "added variable").

While the statistical control of multiple regression is weaker than experimental control, it still has great value. The careful introduction of additional

Figure 3.2 Added-Variable Plot for Education



variables into an equation permits greater confidence in our findings. For instance, the bivariate regression model of the Riverview study suggested that education is a determinant of income. However, this conclusion is open to challenge. That apparent bivariate relationship could be spurious, a product of the common influence of another variable on education and income. For example, an antagonist might argue that the observed bivariate relationship is actually caused by seniority, for those with more years on the job are those with more education, as well as higher pay. An implication is that if seniority were "held constant," education would be exposed as having no effect on income. Multiple regression permits us to test this hypothesis of spuriousness. From the above least squares estimates, we discovered that education still has an apparent effect, even after taking the influence of seniority into account. Hence, through actually bringing this third variable

into the equation, we are able to rule out a hypothesis of spuriousness and thereby strengthen our belief that education affects income.

Confidence Intervals and Significance Tests

The procedure for confidence intervals and significance tests carries over from the bivariate case. Suppose we wish to know whether the partial slope, β_1, from our three-variable equation for the Riverview study is significantly different from zero. Again, we confront the null hypothesis, which says there is no relationship in the population, and the alternative hypothesis, which says there is a relationship in the population. Let us construct a two-tailed 95% confidence interval around the partial slope estimate to test these hypotheses:

$$b_1 \pm t_{n-3;.975}\, s.e.\ (b_1)$$

Note that the only difference between this formula and the bivariate formula is the number of degrees of freedom. Here, we have one less degree of freedom, $(n-3)$ instead of $(n-2)$, because we have one more independent variable. In general, the degrees of freedom of the t distribution equal $(n-k-1)$, where n = sample size and k = number of independent variables. Applying the formula,

$$2{,}252 \pm t_{29;.975}\, s.e.(b_1) = 2{,}252 \pm 2.045(335) = 2{,}252 \pm 685$$

We are 95% confident that the value of the partial slope in the population is between \$1,567 and \$2,937. Because the value of zero is not captured within this band, we reject the null hypothesis. We state that the partial slope estimate, β_1, is statistically significant at the .05 level.

A second approach to the significance testing of β_1 would be examination of the t ratio,

$$\frac{b_1}{s.e.(b_1)} = \frac{2{,}252}{335} = 6.72$$

We observe that the value of this t ratio exceeds the t distribution value, $t_{n-3;.975}$. That is,

$$6.72 > 2.045$$

Therefore, we conclude that β_1 is statistically significant at the .05 level.

A useful preliminary means of significance testing is to use the rule of thumb, which claims statistical significance at the .05 level, two-tailed, for any coefficient whose t ratio exceeds 2 in absolute value. Below is the three-variable Riverview equation, with the t statistics in parentheses:

$$\hat{y} = 6{,}769 + 2{,}252x_1 + 739x_2$$
$$t\text{-statistic} \quad (1.26) \quad (6.73) \quad (3.52)$$

An examination of these t ratios, with this rule of thumb in mind, instantly reveals that the coefficient estimates of education and seniority (β_1, β_2) are significant at the .05 level.

The R^2

To assess the goodness of fit of a multiple regression equation, we employ the R^2, now referred to as the *coefficient of multiple determination*. Once again,

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{regression (explained) sum of squares}}{\text{total sum of squares}}$$

The R^2 for a multiple regression equation indicates the proportion of variation in y "explained" by all the independent variables. In the above three-variable Riverview model, $R^2 = .74$, indicating that education and seniority together account for 74% of the variance in income. This multiple regression model clearly provides a more powerful explanation for income differences than the bivariate regression model, where $R^2 = .62$.

Obviously, it is desirable to have a high R^2, for it implies a more complete explanation of the phenomenon under study. Nevertheless, if a higher R^2 were the only goal, then one could simply add independent variables to the equation. That is, an additional independent variable cannot lower the R^2 and is virtually certain to increase it at least somewhat. In fact, if independent variables are added until their number equals $n - 1$, then $R^2 = 1.0$. This "perfect" explanation is of course nonsense and amounts to no more than a mathematical necessity, which occurs because the degrees of freedom have been exhausted. In sum, rather than entering variables primarily to enhance R^2, the analyst must be guided by theoretical considerations in deciding which variables to include.

One check against an R^2 that is inflated simply from the addition of extraneous independent variables comes from calculation of the *adjusted* R^2. This statistic, routinely reported in statistical packages, reduces the magnitude of the R^2 according to the degrees of freedom it uses up.[1] Since the degrees of freedom exhausted is a direct function of the number of independent variables added, the adjusted R^2 offers a worthwhile correction against overfitting a model. Take our Riverview example, where we have used three degrees of freedom. We can report that $R^2_{\text{adj}} = .72$, which is .02 points less than the unadjusted R^2 of .74. This shows that the raw, uncorrected R^2 exaggerates the model fit some. Having said this, the analyst must be aware that this correction does not represent a cure-all against overfitting. In the case of large samples (e.g., $n > 100$), it will not reduce this fit statistic much (because the degrees of freedom spent becomes trivial compared with the total sample size).

Predicting y

A multiple regression equation can be used for prediction as well as explanation. Let us predict the income of a Riverview city employee who has 10 years of education and has been on the job 15 years:

$$\begin{aligned}\hat{y} &= 6{,}769 + 2{,}252x_1 + 739x_2 \\ &= 6{,}769 + 2{,}252(10) + 739(15) \\ &= 6{,}769 + 33{,}605 \\ \hat{y} &= 40{,}374\end{aligned}$$

Constructing a confidence interval by hand in the multiple regression setting is more complicated. Most statistical software, however, will provide a confidence interval after the user specifies values of the covariates for prediction. For a city employee with 10 years of education and 15 years of job experience, we get the following 95% confidence interval: [35,396; 45,343]. This confidence interval indicates that we are 95% confident that a municipal employee with 10 years of education and 15 years of seniority will earn on average between $35,396 and $45,343. While this prediction is more accurate than that generated by the bivariate regression equation, it is still far from precise.

The model is even less useful for forecasting beyond its range of experience. Certainly, we could plug in any values for x_1 and x_2 and produce a

prediction for y. However, the worth of the forecast diminishes as these x_1 and x_2 values depart from the actual range of variable values in the data. For instance, it would be risky to predict the income of a city worker with 2 years of education and 35 years of seniority, for no one in the data set registered such extreme scores. Possibly, at such extreme values, the linearity of the relationships would no longer exist. Then, any prediction based on our linear model would be quite wide of the mark.

Dummy Variables

Regression analysis encourages the use of variables whose amounts can be measured with numeric precision, that is, *interval variables*. A classic example of such a variable is income. Individuals can be ordered numerically according to their quantity of income, from the lowest to the highest. Thus, we can say that Catherine's income of $112,000 is larger than Bill's income of $56,000; in fact, it is exactly twice as large. Of course, not all variables are measured at a level that allows such precise comparison. Nevertheless, these noninterval variables are candidates for incorporation into a regression framework, through the employment of *dummy variables* (also referred to as *indicator variables*).

Many noninterval, or qualitative, variables can be considered *dichotomies*—for example, gender (male, female), race (Black, White), or marital status (single, married). Dichotomous independent variables do not cause the regression estimates to lose any of their desirable properties. Because they have two categories, they manage to "trick" least squares, entering the equation as an interval variable with just two values. (To convince yourself it is not really a "trick," recall that the average of a variable scored 0 or 1 will be a proportion, a precise interpretable number.) It is useful to examine how such "dummy" variables work. Suppose we argue that a person's income is predicted by race, as well as education, in this bivariate regression,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

where y = income, x_1 = education, and x_2 = race (0 = Black, 1 = White). If $x_2 = 0$, then

$$\hat{y} = b_0 + b_1 x_1 + b_2(0)$$

$$\hat{y} = b_0 + b_1 x_1$$

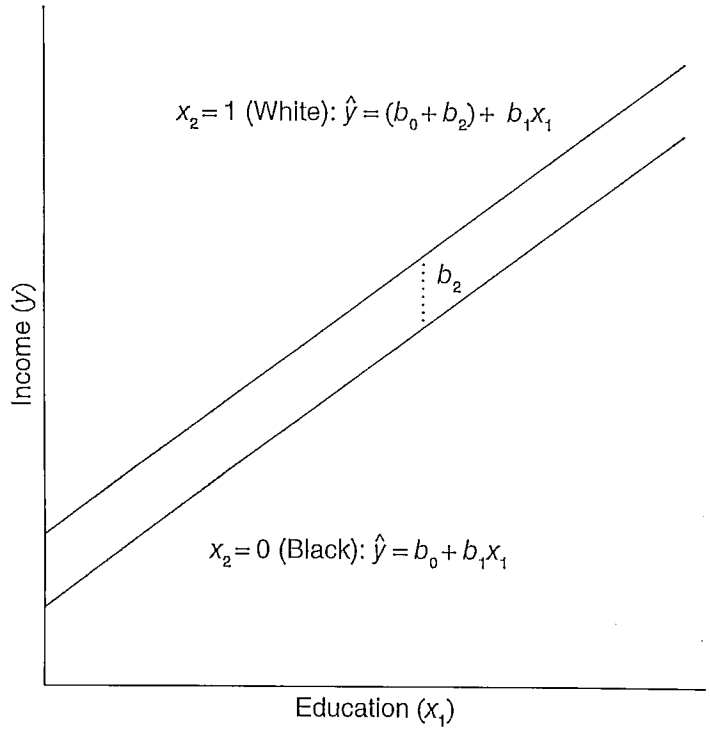
is the prediction of the mean income for Blacks. If $x = 1$, then

$$\hat{y} = b_0 + b_1 x_1 + b_2(1)$$

$$\hat{y} = (b_0 + b_2) + b_1 x_1$$

is the prediction of the mean income for Whites. Grouping the intercept together with b_2 in the above equation, we see the estimate, b_2, reports the difference in average income between Blacks and Whites. If the sign of b_2 is positive, it tells us Whites have higher incomes on average; if the coefficient is negative, it indicates Whites have lower average incomes than Blacks. Graphically, we can see b_2 as a shift up (or down) in the intercept of the regression line (Figure 3.3). Note also that this shift is constant:

Figure 3.3 Dummy Variable for Race Added to the Riverview Model



Regardless of an individual's level of education, we expect Whites to make on average b_2 more dollars than Blacks. As always, the t-test statistic of b_2 measures its statistical significance.

Obviously, not all noninterval, or qualitative, variables are dichotomous. Noninterval variables with multiple categories are of two basic types: *ordinal* and *nominal*.[2] With an ordinal variable, cases can be ordered in terms of amount but not with numeric precision. Attitudinal variables are commonly of this kind. For example, in a survey of the electorate, respondents may be asked to evaluate their political interest, ranking themselves as "not interested," "somewhat interested," or "very interested." We can say that Respondent A, who chooses "very interested," is more interested in politics than Respondent B, who selects "not interested," but we cannot say numerically how much more. Ordinal variables, then, only admit of a ranking from "less to more." The categories of a nominal variable, in contrast, cannot be so ordered. The variable of religious affiliation is a good example. The categories of Protestant, Catholic, Jewish, or Muslim, for example, represent personal attributes that yield no meaningful ranking.

Noninterval independent variables with multiple categories, whether ordinal or nominal, can be incorporated into the multiple regression model through the dummy variable technique. Let us explore an example. Suppose the dollars an individual contributes to a political campaign are a function of the above-mentioned ordinal variable, political interest. Then, a correct model would be

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

where y = campaign contributions (in dollars); x_1 = a dummy variable, scored 1 if "somewhat interested," 0 if otherwise; x_2 = a dummy variable, scored 1 if "very interested," 0 if otherwise; and e = error.

Observe that there are only *two* dummy variables to represent the trichotomous variable of political interest. If there were three dummy variables, then the parameters could not be uniquely estimated. That is, a third dummy, x_3 (scored 1 if "not interested," 0 if otherwise), would be an exact linear function of the others, x_1 and x_2. (Consider that when the score of any respondent on x_1 and x_2 is known, it would always be possible to predict his or her x_3 score. For example, if a respondent has values of 0 on x_1 and 0 on x_2, then he or she is necessarily "not interested" in politics and would score 1 on x_3.) This describes a situation of perfect multicollinearity, in which estimation cannot proceed (this problem is discussed further in the next chapter). To avoid such a dummy variable trap, which is easy to fall into, we memorize this rule: *When a noninterval variable has G categories, use G − 1 dummy variables to represent it.*

A question now arises as to how to estimate the campaign contributions of this excluded group, those who responded "not interested." Their average campaign contribution is estimated by the intercept of the equation. That is, for someone who is "not interested," the prediction equation reduces to

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= b_0 + b_1(0) + b_2(0) \\ \hat{y} &= b_0\end{aligned}$$

Thus, the intercept estimates the average campaign contribution of someone who is "not interested" in politics.

This estimated contribution, b_0, for the "not interested" category serves as a base for comparing the effects of the other categories of political interest. The prediction equation for someone in the category "somewhat interested" reduces to

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= b_0 + b_1(1) + b_2(0) \\ \hat{y} &= b_0 + b_1\end{aligned}$$

Hence, the partial slope estimate, b_1, indicates the difference in mean campaign contributions between those "somewhat interested" and those "not interested," that is, $(b_0 + b_1) - b_0 = b_1$.

For the last category, "very interested," the prediction equation reduces to

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= b_0 + b_1(0) + b_2(1) \\ \hat{y} &= b_0 + b_2\end{aligned}$$

Thus, the partial slope estimate, b_2, points out the difference in average campaign contributions between the "very interested" and the "not interested." Given the hypothesis that heightened political interest increases campaign contributions, we would expect that $b_2 > b_1$.

A data example will increase our appreciation of the utility of dummy variables. Suppose, with the Riverview study, it occurs to us that the

income received from working for city government might be determined by the employee's political party affiliation (Democrat, Republican, or independent) and by gender, as well as by education and seniority. In that case, the proper specification of the model becomes

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + e$$

where y = income (in dollars); x_1 = education (in years); x_2 = seniority (in years); x_3 = gender of respondent (0 = female, 1 = male); x_4 = a dummy variable scored 1 if independent, 0 otherwise; x_5 = a dummy variable scored 1 if Republican, 0 otherwise; and e = error.

The variable political party has three categories. Thus, applying the G – 1 rule, we had to formulate 3 – 1 = 2 dummy variables. We chose to construct one for independents (x_4) and one for Republicans (x_5), which left Democrats as the base category. The selection of a base category is entirely up to the analyst. Here, we selected Democrats as the standard for comparison because we guessed they would have the lowest income, with independents and Republicans having successively higher incomes.

Least squares yields the following parameter estimates:

$$\hat{y} = 5{,}093 + 2{,}153x_1 + 707x_2 + 8{,}084x_3 - 805x_4 + 2{,}558x_5$$

t-statistic	(1.07)	(7.33)	(3.95)	(3.13)	(–0.27)	(0.77)
p-value	(0.29)	(<.001)	(<.001)	(.004)	(0.79)	(0.45)

$R^2 = .84$ Adj. $R^2 = .81$ $n = 32$ $s_e = 6{,}356$

First, we note that the parameter estimates from our prior specification remain almost unchanged. Furthermore, from the p-value, we see that the average income of independents is not significantly different (.05 level) from the average income of Democrats, once the effects of education, seniority, and gender are controlled. (Put another way, b_4 does not add significantly to the intercept, b_0.) Likewise, the average income of Republicans is found not to differ significantly from that of the Democrats. We must conclude that, contrary to our expectation, political party affiliation does not influence the income of Riverview municipal employees. Our four-variable model, which now includes gender, stands as the preferred specification.

Through use of the dummy variable technique, the inclusion into our multiple regression equation of the noninterval variable, political party, poses no problem. Some researchers would argue that this variable could be inserted into our regression equation directly, bypassing the dummy

variable route. The argument is that an ordinal variable is a candidate for regression, even though the distances between the categories are not exactly equal. This is a controversial point of view. In brief, the advocate's primary defense is that, in practice, the conclusions are usually equivalent to those generated by more correct techniques (i.e., the application of dummy variable regression or ordinal-level statistics). A secondary argument is that multiple regression analysis is so powerful, compared with ordinal-level techniques, that the risk of error is acceptable. We cannot resolve this debate here. However, we can provide a practical test by incorporating political party into the Riverview equation as an ordinal variable.

At first blush, political party affiliation may appear as strictly nominal. Nevertheless, political scientists commonly treat it as ordinal. We can say, for example, that an independent is "more Republican" than a Democrat, who is "least Republican" of all. Hence, we can order the categories in terms of their "distance" from Republicans. This order is indicated in the following numeric code (Democrat = 0, independent = 1, Republican = 2), which ranks the categories along this dimension of "Republicanism." This code provides each respondent a score on a political party variable, x_4, which we now enter into the Riverview equation. Least squares yields the following estimates:

$$\hat{y} = 3{,}951 + 2{,}163x_1 + 686x_2 + 9{,}034x_3 + 1{,}176x_4$$

t-statistic	(0.88)	(7.42)	(3.89)	(3.94)	(0.71)
p-value	(0.39)	(<.001)	(<.001)	(<.001)	(0.48)

$R^2 = .84$ Adj. $R^2 = .81$ $n = 32$ $s_e = 6{,}316$

where y = income; x_1 = education; x_2 = seniority; x_3 = gender; x_4 = political party affiliation, scored 0 = Democrat, 1 = independent, or 2 = Republican; and the statistics are defined as above.

The estimates for the coefficients of our original variables are little changed. Also, political party affiliation is shown to have no statistically significant impact on employee's income ($p > .05$). Thus, in this particular case, regression analysis with an ordinal variable arrives at the same conclusion as the more proper regression analysis with dummy variables.

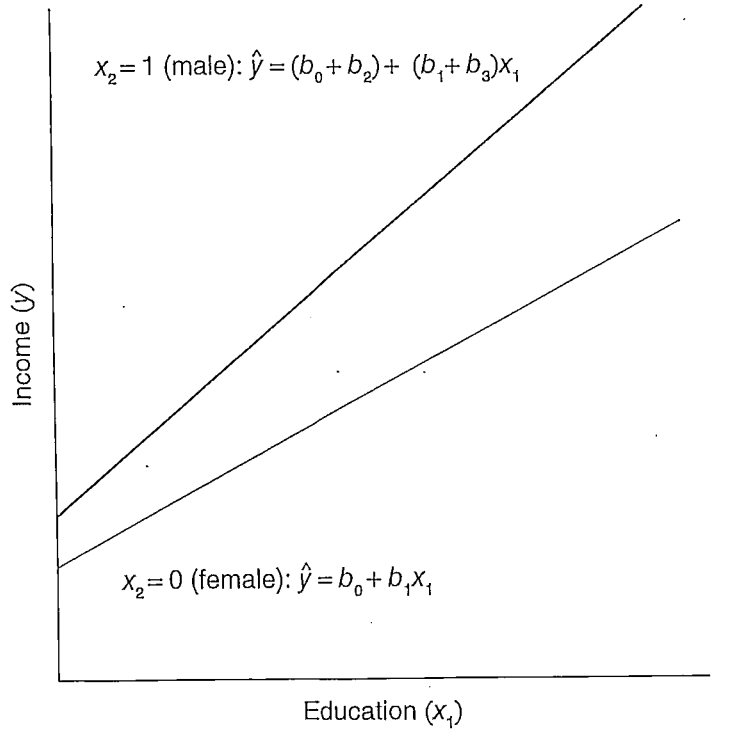
The Possibility of Interaction Effects

Thus far, we have assumed that effects are *additive*. That is, y is determined, in part, by x_1 *plus* x_2, not x_1 *times* x_2. This additivity assumption

dominates applied regression analysis and is frequently justified. However, it is not a necessary assumption. Let us explore an example.

In the previous example, we used the variable of gender of respondent as a candidate for inclusion in the Riverview income equation. The question is, should the gender variable enter additively or as an interaction? It might be argued that gender is involved interactively with education. In general, an *interaction effect* exists when the impact of one independent variable depends on the value of another independent variable. Specifically, perhaps the effect of education is dependent on the gender of the employee, with education yielding a greater financial return for men. We can see this hypothesized relationship graphically in Figure 3.4. Besides the different intercept, as education increases, the slope for men is steeper than for women.

Figure 3.4 Hypothesized Interaction Model for Gender and Education



Formally, this particular interaction model is as follows (we ignore the other variables for the moment):

$$y = b_0 + b_1x_1 + b_2x_2 + b_3(x_1x_2) + e$$

where y = income (in dollars), x_1 = education (in years), x_2 = gender of respondent (0 = female, 1 = male), x_1x_2 = an interaction variable created by multiplying x_1 times x_2, and e = error. The least squares estimates for this model are

$$\hat{y} = 8{,}330 + 2{,}574x_1 + 9{,}295x_2 + 23(x_1x_2)$$

t-statistic	(1.11)	(5.58)	(0.87)	(0.04)
p-value	(0.27)	(<.001)	(0.38)	(0.97)

$R^2 = .74$ Adj. $R^2 = .72$ $n = 32$ $s_e = 7{,}757$

These results indicate that the income returns from education are not significantly different for men and women. The *p*-value on the interaction coefficient is 0.97, which is much larger than an alpha level of .05. Thus, we fail to reject the null hypothesis that the coefficient of the interaction term is zero. Note also that the dummy variable for gender alone is no longer statistically significant, even though it was in the previous model, which included political party affiliation. This appears a consequence of multicollinearity, a problem not uncommon with interaction models. (In this case, we have gender entered twice in the interaction model: once by itself and again multiplicatively with education.[3])

Now that we have rejected the hypothesis of an interaction effect, let's fit the alternative strictly additive model, where the variables are defined as before. Estimating this model yields

$$\hat{y} = 8{,}141 + 2{,}586x_1 + 9{,}667x_2$$

t-statistic	(1.54)	(8.22)	(3.55)
p-value	(0.13)	(<.001)	(.001)

$R^2 = .74$ Adj. $R^2 = .72$ $n = 32$ $s_e = 7{,}622$

These estimates suggest that education and gender have significant, independent effects on income. Holding education fixed, we would expect males to have an average salary that is $9,667 greater than females. It seems that the data are more congruent with the additive model, which is more in keeping with a "discrimination" theory of income determination; that is, other things being equal, society pays women less solely because they are women.

A Four-Variable Model: Overcoming Specification Error

Incorporating the gender variable additively into our model, along with education and seniority, leads to the following equation for income differences in Riverview:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

where y = income (in dollars), x_1 = education (in years), x_2 = seniority (in years), x_3 = gender of respondent (0 = female, 1 = male), and e = error. Theoretically, this four-variable model is much more complete than the initial two-variable model. It asserts that income is a linear additive function of three factors: education, seniority, and gender. Estimating this multiple regression model with least squares yields

$$\hat{y} = 4{,}309 + 2{,}230x_1 + 670x_2 + 8{,}775x_3$$

t-statistic	(0.97)	(8.14)	(3.87)	(3.90)
p-value	(0.34)	(<.001)	(<.001)	(<.001)

$R^2 = .83$ Adj. $R^2 = .81$ $n = 32$ $s_e = 6{,}261$

These estimates tell us a good deal about what affects income in Riverview city government. The pay of a municipal employee is significantly influenced by years of education, amount of seniority, and gender. (Each of the independent variables has a t-test statistic greater than 2, indicating statistical significance at the .05 level.) These three factors largely determine income differences within this population. In fact, more than three quarters of the variation in income is explained by these variables ($R^2_{adj} = .81$). The differences caused are not inconsequential. For each year of education, $2,230 is added to income, on average. An extra year of seniority contributes another $670. Male workers can expect $8,775 more than female workers, even if the women have the same education and seniority. The cumulative impact of these variables can create sizable income disparities. For example, a male with a college education and 10 years of seniority would expect to make $55,464; in contrast, a female with a high school degree and just starting work could expect to earn only $31,069.

Inclusion of relevant variables, that is, seniority and gender, beyond the education variable, has markedly diminished specification error, helping ensure that our estimates are BLUE. (To refresh yourself on the meaning of specification error, review the discussion of assumptions in Chapter 2.) In particular, the estimate of the education coefficient, which equaled 2,651 in the bivariate model, has been reduced. The comparable estimate in this four-variable model, $b_1 = 2{,}230$, indicates that the true impact of an additional year of education is approximately $400 less than estimated in the original bivariate equation.

For certain models, it is fairly easy to detect the direction of bias resulting from the exclusion of a relevant variable. Suppose the real world is congruent with the following model:

$$y = b_0 + b_1x_1 + b_2x_2 + e \text{ (correct model)}$$

but we mistakenly estimate

$$y = b_0 + b_1x_1 + e^* \text{ (incorrect model)}$$

where $e^* = (b_2x_2 + e)$. By excluding x_2 from our estimation, we have committed specification error. Assuming that x_1 and x_2 are correlated, as they almost always are, the slope estimate, b_1, will be biased. This bias is inevitable, for the independent variable, x_1, and the error term, e^*, are correlated, thus violating an assumption necessary for regression to yield desirable estimators. (We see that $r_{x_1e^*} \neq 0$, because $r_{x_1x_2} \neq 0$, and x_2 is a component of e^*.) The direction of the bias of b_1 in the estimated model is determined by (1) the sign of b_2 and (2) the sign of the correlation, r_{12}. If b_2 and r_{12} have the same sign, then the bias of b_1 is positive; if not, then the bias is negative.

It happens that the direction of bias in the somewhat more complicated Riverview case accords with these rules. As noted, the bias of b_1 in the bivariate equation of the Riverview study is positive, accepting the specification and estimation of the four-variable model. The presence of this positive bias follows the above guidelines: (1) The sign of b_2 (and b_3) is positive, and (2) the sign of r_{12} (and r_{13}) is positive; therefore, the bivariate estimate of b_1 must be biased upward. Part of the variance in y that x_1 is accounting for should be explained by x_2 and x_3, but these variables are not in the equation. Thus, some of the impact of x_2 and x_3 on y is erroneously assigned to x_1.

The formulation of rules for the detection of bias implies that it is possible to predict the consequences of a given specification error. For instance, the analyst is able to foresee the direction of bias coming from the exclusion of a certain variable. With simpler models, such as those treated here, such insight might be attainable. However, for models that include several variables and face several candidates for inclusion, the direction of bias is not readily foreseeable. In this more complex situation, the analyst is better served by immediate attention to proper specification of the model.

Notes

1. We can see how the adjusted R^2 "penalizes" for adding extraneous variables to the model by looking at the following relationship between adjusted R^2 and R^2:
$$R^2_{adj} = R^2 - \left(1 - R^2\right)\frac{k}{n-k-1}$$
2. For a discussion of preferred measures of association, when the level of measurement varies, see Lewis-Beck (1995, chap. 4).
3. Including the variable gender by itself, and as an interaction term, is important as it preserves the principal of marginality. It is good practice for models with interaction terms to include all variables that make up the interaction term as individual covariates. Including interaction terms alone can lead to spurious conclusions about the coefficients.

CHAPTER 4. MULTIPLE REGRESSION: SPECIAL TOPICS

In this final chapter, we consider selected topics in multiple regression analysis that merit special consideration: the multicollinearity problem, the relative importance of independent variables, nonlinearity, and the proper presentation of research findings. As a parting note, we offer other topics that might be pursued, after the reader has absorbed the material of this monograph.

The Multicollinearity Problem

For multiple regression to produce the "best linear unbiased estimates," it must meet the bivariate regression assumptions, plus one additional assumption: the absence of *perfect multicollinearity*. That is, none of the independent variables is perfectly correlated with another independent variable or linear combination of other independent variables. For example, with the following multiple regression model,

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

perfect multicollinearity would exist if

$$x_2 = c_0 + c_1x_1$$

for x_2 is a perfect linear function of x_1 (i.e., $R^2 = 1.0$). When perfect multicollinearity exists, it is impossible to arrive at a unique solution for the least squares parameter estimates. Any effort to calculate the partial regression coefficients, numerically or analytically, will fail. Thus, the presence of perfect multicollinearity is immediately detectable. Furthermore, in practice, it is obviously quite unlikely to occur. However, *high multicollinearity* commonly perplexes the users of multiple regression.

With nonexperimental social science data, the independent variables are virtually always intercorrelated, that is, multicollinear. When this condition becomes extreme, serious estimation problems often arise. The general difficulty is that parameter estimates become unreliable. The magnitude of the partial slope estimate in the present sample may differ considerably from its magnitude in the next sample. Hence, we have little confidence that a particular slope estimate accurately reflects the impact of x on y in the

population. Obviously, because of such imprecision, this partial slope estimate cannot be usefully compared with other partial slope estimates in the equation to arrive at a judgment of the relative effects of the independent variables. Finally, an estimated regression coefficient may be so unstable that it fails to achieve statistical significance, even though x is actually associated with y in the population.

Venn diagrams can help illustrate high multicollinearity and the arising estimation difficulties.[1] Figure 4.1a-b represents regressing y on two variables, x_1 and x_2. Each circle represents a variable's variation. The red and blue sections are the portion of y uniquely explained by x_1 and x_2, respectively. The orange region is the part of y unexplained by x_1 and x_2. When estimating the coefficients b_1 and b_2, only the information in regions red and blue are used. The black section is discarded: It represents variation in y explained by both x_1 and x_2; however, because x_1 and x_2 overlap (i.e., are correlated), it is impossible to disentangle completely their shared variance that contributes to the explanation of y.[2] As mentioned, rarely will two independent variables be completely independent and have no overlap, especially in the social sciences. There will likely be some correlation illustrated by the black section in Figure 4.1a. With mild multicollinearity, there is still enough unique variation in the red and blue sections to get a precise estimate of b_1 and b_2.

However, as we see in Figure 4.1b, when there is high multicollinearity, the area of the red and blue sections shrinks—that is, the unique variation of y explained by x_1 and x_2 is reduced. The slope estimates will still be unbiased. But because we have little information to estimate b_1 and b_2, the slope coefficients will be unstable. In the extreme case of perfect multicollinearity, the two circles, x_1 and x_2, would exactly overlap, leaving zero information to estimate unique coefficients.

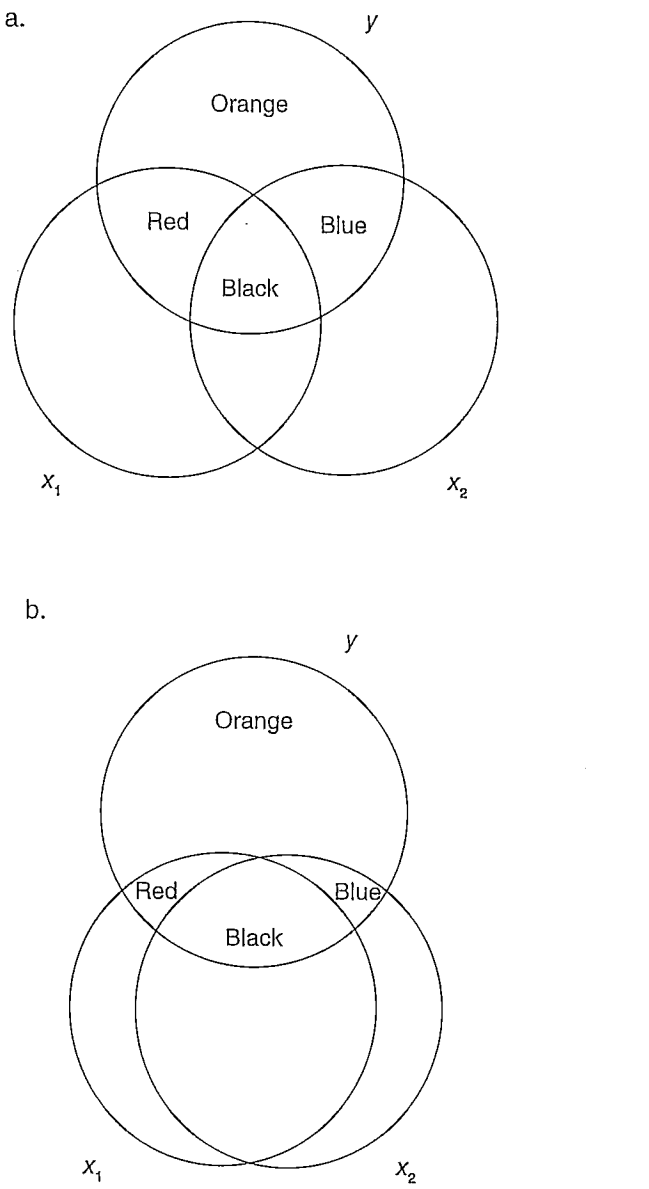
High multicollinearity creates these estimation problems because it produces large variances for the slope estimates and, consequently, large standard errors. Recalling the formula for a confidence interval (95%, two-sided),

$$b_j \pm t_{n-k-1;.975}\, s.e.(b_j) \qquad j = 1,\ldots, k$$

we recognize that a larger standard error, $s.e.(b_j)$, will widen the range of values that b_j might take on. Reviewing the formula for the t-statistic,

$$t\text{-stat} = \frac{b_j}{s.e.\left(b_j\right)}$$

Figure 4.1 (a-b) Venn Diagrams of Mild and High Multicollinearity



we observe that a larger $s.e.(b_j)$ makes it more difficult to achieve statistical significance (e.g., more difficult to exceed the value of 2, which indicates statistical significance at the .05 level, two-tailed).

We can see how large variances occur with high multicollinearity by examining this variance formula,

$$\text{variance}\, b_j = s_{b_j}^2 = \frac{s_u^2}{\sum_{i=1}^{n} v_{i,j}^2}$$

where s_u^2 is the variance of the error term in the multiple regression model, and v_j^2 is the squared residual from the regression of the jth independent variable, x_j, on the rest of the independent variables in the model. Hence,

$$v_j = x_j - \hat{x}_j$$

If these other independent variables are highly predictive of x_j, then x_j and $\hat{x}_j$ will be very close in value, and so v_j will be small. Therefore, the denominator in the above variance formula will be small, yielding a large variance estimate for b_j.

Of course, when analysts find that a partial regression coefficient is statistically insignificant, they cannot simply dismiss the result on grounds of high multicollinearity. Before such a claim can be made, high multicollinearity must be demonstrated. Let us first look at common symptoms of high multicollinearity, which may alert the researcher to the problem. Then, we will proceed to techniques of diagnosis. One rather sure symptom of high multicollinearity is a substantial R^2 for the equation but a lack of statistically significant coefficients. A second, weaker, signal is regression coefficients that change greatly in value when independent variables are dropped or added to the equation. A third, still less sure, set of symptoms involves suspicion about the magnitudes of the coefficients. A coefficient may be regarded as unexpectedly large (small), either in itself or relative to another coefficient in the equation. It may even be so large (or small) as to be rejected as nonsensical. A fourth alert is a coefficient with the "wrong" sign. Obviously, this last symptom is feeble, for knowledge of the "right" sign is often lacking.

The above symptoms might provide the watchful analyst hints of a multicollinearity problem. However, by themselves, they cannot establish that the problem exists. For diagnosis, we must look directly at the intercorrelation of the independent variables. A frequent practice is to examine the bivariate correlations among the independent variables, looking for coefficients of about .8 or larger. Then, if none is found, one goes on to conclude that multicollinearity is not a problem. While suggestive, this approach is

unsatisfactory, for it fails to take into account the relationship of an independent variable with *all* the other independent variables. It is possible, for instance, to find no large bivariate correlations, although one of the independent variables is a nearly perfect linear combination of the remaining independent variables. This possibility points to the preferred method of assessing multicollinearity: *Regress each of the k independent variables on all the other independent variables*. When any of the R_j^2 from these equations is near 1.0, there is high multicollinearity. In fact, the largest of these auxiliary R_j^2, as they are called, serves as an indicator of the amount of multicollinearity that exists. Standard statistical software can run all possible auxiliary regressions of the independent variables and report a statistic called the variance inflation factor (VIF). It measures how much the variance of the regression coefficients is inflated compared with the noninflated baseline of linearly independent predictors. For each of the j predictor variables, the VIF is calculated as follows:

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

where R_j^2 is the coefficient of determination from regressing x_j on all the other predictor variables. The "best-case" scenario is if the VIF is 1, then x_j is linearly independent of the other covariates. A good rule of thumb is if $VIF_j \geq 10$, then multicollinearity may be a problem. (From the formula for VIF, we can see that it will increase as the number of independent variables increases. This is another way of saying that the coefficient estimates will tend to become more unstable. A message here, for practicing researchers, is to keep the model specification as parsimonious as possible.)

Let us apply what we have learned about multicollinearity to the four-variable Riverview model:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

where y = income (in dollars), x_1 = education (in years), x_2 = seniority (in years), x_3 = gender of respondent (0 = female, 1 = male), and e = error. The estimates for this model, which we have already examined, reveal no symptoms of a multicollinearity problem. That is, the coefficients are all significant, and their signs and magnitudes are reasonable. Therefore, we would anticipate that the above multicollinearity test would produce an R_j^2 far from unity. Regressing each independent variable on all the others yields

$$\hat{x}_1 = 12.79 + 0.21x_2 + 0.17x_3 \quad R^2 = .11$$

$$\hat{x}_2 = 5.72 + 0.53x_1 + 1.33x_3 \qquad R^2 = .12$$

$$\hat{x}_3 = 0.28 + 0.002x_1 + 0.007x_2 \quad R^2 = .01$$

These R_j^2 show that these independent variables are intercorrelated in the Riverview sample, as we would expect with data of this type. But we observe that the largest coefficient of multiple determination, $R^2 = .12$, lies a good distance from 1.0. We could also confirm these results by looking at the VIFs for each coefficient:

	Education	Seniority	Gender
VIF	1.13	1.14	1.01

We see that none of the VIFs are anywhere close to 10. Our conclusion is that multicollinearity is not a problem for the partial slope estimates in the Riverview multiple regression model.

The results do not always turn out so well. What can we do if high multicollinearity is detected? Unfortunately, none of the possible solutions is wholly satisfactory. In general, we must make the best of a bad situation. The standard prescription is to increase our information by enlarging the sample. As noted in an earlier chapter, the bigger the sample size, the greater the chance of finding statistical significance, other things being equal. Realistically, however, the researcher is usually unable to increase the sample. Also, multicollinearity may be severe enough that even a large n will not provide much relief.

Assuming the sample size is fixed, other strategies have to be implemented. One is to combine those independent variables that are highly intercorrelated into a single indicator. If this approach makes conceptual sense, then it can work well. Suppose, for example, a model that explains political participation (y) as a function of income (x_1), race (x_2), Internet use (x_3), television watching (x_4), and newspaper reading (x_5). On one hand, it seems sensible to combine the highly intercorrelated variables (x_3, x_4, x_5) into an index of media involvement. On the other hand, it is not sensible to combine the income and race variables, even if they are highly related.

Suppose our variables are "apples and oranges," making it impractical to combine them. In the face of high multicollinearity, we cannot reliably separate the effects of the involved variables. Still, the equation may have value if its use is restricted to prediction. That is, it might be employed to predict y for a given set of values on *all* the x's (e.g., when $x_1 = 2, x_2 = 4, \ldots x_k = 3$), but not to interpret the independent effect on y of a change in the value of a *single* x. Usually, this prediction strategy is uninteresting, for the goal is generally explanation, in which we talk about the impact of a particular x on y.

A last technique for combatting multicollinearity is to discard the offending variable(s). Let us explore an example. Suppose we specify the following elementary multiple regression model:

$$y = b_0 + b_1x_1 + b_2x_2 + e \text{ (Model I)}$$

Lamentably, however, we find that x_1 and x_2 are so highly related ($r_{12} = .9$) that the least squares estimates are unable reliably to assess the effect of either. An alternative is to drop one of the variables, say x_2, from the equation, and simply estimate this model:

$$y = b_0 + b_1x_1 + e^* \text{ (Model II)}$$

A major problem with this procedure, of course, is its willful commission of specification error. Assuming Model I is the correct explanatory model, we know the estimate for b_1 in Model II will be biased. A revision that makes this technique somewhat more acceptable is to estimate yet another equation, now discarding the other offending variable (x_1),

$$y = b_0 + b_2x_2 + e^{**} \text{ (Model III)}$$

If the Model II and Model III estimates are evaluated, along with those of Model I, then the damage done by the specification error can be more fully assessed.

High Multicollinearity: An Example

To grasp more completely the influences of high multicollinearity, it is helpful to explore another example. Let us expand on our school policy example from the previous chapter. But now we will examine the findings with an eye to the multicollinearity issue. After collecting more potential explanatory variables on educational outcomes, we formulate a multiple regression model, arriving at the following estimates:

$$\hat{y} = 75.42 + 0.045x_1 - .000011x_1^2 + 3.68x_2 - .786x_3 - .000345x_4$$

t-statistic	(2.70)	(6.68)	(−4.10)	(1.62)	(−3.75)	(−1.06)
p-value	(0.01)	(<.001)	(<.001)	(0.11)	(<.001)	(0.32)

$R^2 = .89$ Adj. $R^2 = .88$ $n = 50$ $s_e = 7.62$

where y = average high school test score, x_1 = school size (total number of students), x_1^2 = school size squared, x_2 = an indicator if a school is in a rural area (1 = rural, 0 = not rural), x_3 = poverty measure (percentage of students

receiving subsidized lunch), and x_4 = average household income (averaged over the school district).

These results suggest that school size has a positive effect on test scores but becomes negative once a certain size is reached. Test scores at first appear higher in rural areas, but the difference is not statistically significant. Both poverty and income have a negative sign. The more students on subsidized school lunch, the lower average test scores appear; however, income also has an inverse relationship—higher levels of family income are associated with lower educational performance. Such a conclusion becomes much less certain when we inspect the multicollinearity in the data. Let us diagnose the level of multicollinearity by calculating the VIF for each coefficient.

	School Size	School Size Squared	Rural	Poverty	Income
VIF	16.86	16.80	1.07	22.62	22.84

Obviously, extreme multicollinearity is present. The purpose of the model is not prediction, so we cannot ignore the problem. How might it be corrected? Let us first examine the school size variable. Multicollinearity is a problem with the school size variable because we have the same variable in the model twice, once as x_1 and again as x_1^2. A strategy to deal with polynomial terms is to center them—that is, subtract the mean from the predictor and use the deviations, and squared deviations, as new variables in the model.[3] This transformation is useful as it eliminates high correlation between the two variables. It also, however, slightly changes the interpretation of the coefficients.[4] As for the poverty and income variable, it makes sense they would be highly correlated: They are both measuring in some way the wealth of the district. One possibility would be to combine them into an average measure of economic well-being. Another would be to discard one of the offending variables. Suppose we are more interested in the school district level of wealth since local property taxes fund the schools. We decide to remove poverty from the equation and add the centered school size variable. After reestimating, we get the following:

$$\hat{y} = 7.92 + .0189x_1 - .0000098x_1^2 + 5.14x_2 + .00089x_4$$

t-statistic	(26.05)	(9.75)	(–3.16)	(2.03)	(10.58)
p-value	(0.14)	(<.001)	(.002)	(.048)	(<.001)

$R^2 = .86$ Adj. $R^2 = .85$ $n = 50$ $s_e = 8.63$

where definitions are the same as above.

According to these new estimates, *all* the variables have a statistically significant impact. The sign on income has also changed from negative to positive, which makes intuitive sense: The more wealthy the school district, the higher the average test scores. How reliable are these new estimates? One check is to recalculate the level of multicollinearity. Calculating the VIF for the coefficients in the new equation, we get the following:

	Centered School Size	Centered School Size Squared	Rural	Income
VIF	1.08	1.04	1.05	1.08

We observe that all of these VIFs are quite close to 1, indicating that multicollinearity has ceased to be problematic. The revised parameter estimates would appear much more reliable than the contrary ones generated with the offending variables. Hopefully, this rather dramatic example brings home the perils of high multicollinearity.

The Relative Importance of the Independent Variables

We sometimes want to evaluate the relative importance of the independent variables in determining *y*. An obvious procedure is to compare the magnitudes of the partial slopes. However, this effort is often thwarted by the different measurement units and variances of the variables. Suppose, for example, the following multiple regression equation predicting annual dollars contributed to political campaigns as a function of an individual's age and income,

$$\hat{y} = 78 + 12x_1 + .020x_2$$

where y = campaign contributions (in dollars), x_1 = age (in years), and x_2 = income (in dollars). The relative influence of income and age on campaign contributions is difficult to assess, for the measurement units are not comparable, that is, dollars versus years. One solution is to *standardize* the variables, reestimate, and evaluate the new coefficients. (Some statistical software automatically provides the standardized coefficients along

with the unstandardized coefficients.) Any variable is standardized by converting its scores into standard deviation units from the mean. For the above variables, then,

$$y^* = \frac{y - \bar{y}}{s_y}, \quad x_1^* = \frac{x_1 - \bar{x}_1}{s_{x_1}}, \quad x_2^* = \frac{x_2 - \bar{x}_2}{s_{x_2}}$$

where the asterisk, *, indicates the variable is standardized.

Reformulating the model with these variables yields

$$\hat{y}^* = b_1^* x_1^* + b_2^* x_2^*$$

(Note that standardization forces the intercept to zero.) The standardized partial slope, or standardized regression coefficient, is sometimes designated with "b^*".[5]

The standardized regression coefficient corrects the unstandardized regression coefficient by the ratio of the standard deviation of the independent variable to the standard deviation of the dependent variable:

$$b_j^* = b_j \frac{s_{x_j}}{s_y}$$

As we saw at the end of Chapter 1, in the case of the bivariate regression model, the standardized regression coefficient equals the simple correlation between the two variables. That is, assuming the model,

$$y = b_0 + b_1 x + e$$

then,

$$b^* = b_1 \frac{s_x}{s_y} = r$$

However, this equality does not hold for a multiple regression model. (Only in the unique circumstance of no multicollinearity would $b^* = r$ with a multiple regression model.)

The standardized partial slope estimate, or standardized regression coefficient, indicates *the average standard deviation change in y associated with a standard deviation change in x when the other independent variables are held constant.* Suppose the standardized partial slopes for the above campaign contribution equation are as follows:

$$\hat{y}^* = .15x_1^* + .45x_2^*$$

For example, $b_2^* = .45$ says that a 1 standard deviation change in income is associated with a .45 standard deviation change in campaign contributions, on the average, with age held constant. Let us consider the meaning of this interpretation more fully. Assuming x_2 is normally distributed, then a 1 standard deviation income rise for persons at, say, the mean income would move them into a higher income bracket, above which only about 16% of the population resided (recall the Empirical Rule for normally distributed variables). We see that this strong manipulation of x_2 does not result in an equally strong response in y, for b_2^* is far from unity. Still, campaign contributions do tend to climb by almost one half of a standard deviation. In contrast, a considerable advance in age (a full 1 standard deviation increase) elicits a very modest increment in contributions (only .15 of a standard deviation). We conclude that the impact of income, as measured in standard deviation units, is greater than the impact of age, likewise measured. Indeed, it seems that the effect of income on campaign contributions is three times that of age (.45/.15 = 3).

The ability of standardization to ensure the comparability of measurement units guarantees its appeal when the analyst is interested in the relative effects of the independent variables. However, difficulties can arise if one wishes to make comparisons across samples. This is because, in estimating the same equation across samples, the value of the standardized slope, unlike the value of the unstandardized slope, can change merely because the variance of x changes. In fact, the larger (smaller) the variance in x, the larger (smaller) the standardized regression coefficient, other things being equal. (To understand this, consider again the standardized regression coefficient formula,

$$b_j^* = b_j \frac{s_{x_j}}{s_y}$$

We see that as s_{x_j}, the numerator of the fraction, increases, the magnitude of b_j^* must necessarily increase.)

As an example, suppose that the above campaign contributions model was developed from a U.S. sample, and we wished to test it for another Western democracy, say Sweden. Our standardized regression coefficients from this hypothetical sample of the Swedish electorate might be

$$\hat{y}^* = .18x_1^* + .22x_2^*$$

where the variables are defined as above. Comparing b_2^* (United States) = .45 with b_2^* (Sweden) = .22, we are tempted to conclude that the effect of

income in Sweden is about one half its effect in the United States. However, this inference may well be wrong, given that the standard deviation of income in the United States is greater than the standard deviation of income in Sweden. That is, the wider spread of incomes in the United States may be masking the more equal effect a unit income change actually has in both countries, that is, $b_2(\text{United States}) \cong b_2(\text{Sweden})$. To test for this possibility, we must of course examine the unstandardized regression coefficients, which we suppose to be the following:

$$\hat{y} = 83 + 10.5x_1 + .018x_2$$

When these unstandardized Swedish results are compared with the unstandardized U.S. results, they suggest that, in reality, the effect of income (always measured in dollars) on campaign contributions is essentially the same in both countries ($.02 \cong .018$). In general, when the variance in x diverges from one sample to the next, it is preferable to base any cross-sample comparisons of effect on the unstandardized partial slopes.

Extending the Regression Model: Nonlinearity

Regarding the dependent variable, the assumption thus far has been that it is a linear function of the independent variables. This assumption has utility for different reasons. First, the great body of social science research that has accumulated generally conforms to this linear assumption. Second, when it has been tried, it is hard to do better than the linear model. Of course, the analyst should be alert to the possibility that a relationship under study may be nonlinear and model it as such. We saw a glimpse of this at the end of Chapter 2 with the educational test score model.

To the novice reader, the class of linear models implies the relationship between x and y must be a straight line. This is an unnecessary and overly restrictive assumption. For OLS, the only requirement is the model parameters enter linearly into the model.[6] There are no restrictions on the actual data. The predictors and dependent variable can be transformed to model complex relationships that take on a variety of shapes. If the observed (raw) relationship between x and y appears nonlinear, then the variable(s) may be transformed to achieve linearity. Once the transformation is made, OLS can be run without violating Assumption 1d. Let us look at a few theoretical examples.

In the first chapter, we examined the form of the relationship between education (x_1) and income (y), assuming it was linear. Given that linearity, a one-unit change in x_1 would yield a fixed change in y, irrespective of

where the x_1 score falls. But, it may be possible that the relationship takes a nonlinear form, at least in another sample. In Figure 4.2, we draw three nonlinear relationships beside the simple straight-line relationship (Figure 4.2a). It may be that y is a *logarithmic* function of x_1, like the curve Figure 4.2b. If so, a unit change in x_1 affects y but less and less as the score on x_1 increases. For instance, a rise in education from 20 to 21 years would have a positive effect on salary but not as great as the change from 10 to 11 years. There are other examples where the impact of x_1 is not constant. It may be that y is a *hyperbolic* function of x_1, like in Figure 4.2c. Given that scenario, x_1 keeps a positive impact, but that damps down soon, as it moves toward zero. Going back to the education and income example, a unit change in education from 29 to 30 years would barely have an effect. A *parabolic* model for y is another type, such as sketched in Figure 4.2d. In this case, as x_1 goes up, y goes up to a point, after which more gains in x_1 lead to declines in y. We saw this relationship with the school size variable: Once a school gets too crowded, adding extra students seems to lower average test scores.

These four types of relationships can be expressed mathematically in the equations below:

Linear: $y = b_0 + b_1 x_1$

Logarithmic: $y = b_0 + b_1 \ln(x_1)$

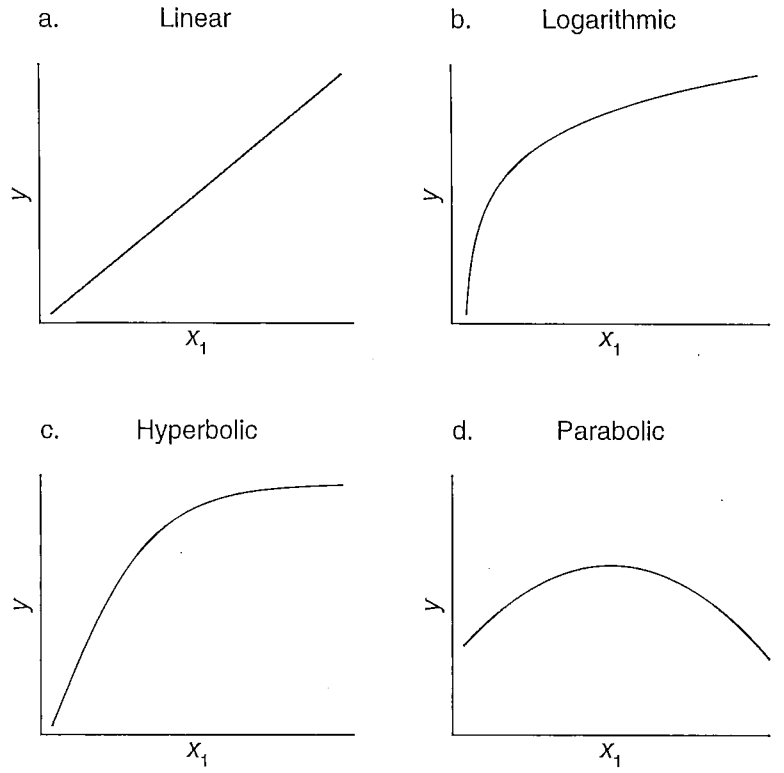
Hyperbolic: $y = b_0 - b_1\left(\frac{1}{x_1}\right)$

Parabolic: $y = b_0 + b_1 x_1 - b_2 x_1^2$

The first model illustrates a linear relationship between raw (observed) x_1 and y (see Figure 4.2a). The subsequent models conform to the curves in Figure 4.2b–d, due to the transformation of x_1. (In turn, we have a natural log, a reciprocal, and a square transformation of x_1.) Assuming that correct curvilinear specification between raw x_1 and y, the transformation will make for a linear relationship. For instance, supposing the observed relationship between raw x_1 and y to be hyperbolic, then the observed relationship between the reciprocal transformation of x_1 (i.e., $\frac{1}{x_1}$) and y is linear. OLS, then, even given its linearity assumption, can be properly used on the transformed equation.

In deciding whether to model a relationship as linear or nonlinear, theory should be heavily relied upon. Unfortunately, though, the signals from

Figure 4.2 (a–d) Different Forms of the Relationship Between x_1 and y



theory are sometimes contradictory. For example, you might believe a relationship is nonlinear, while a critic argues strongly that it is linear. Comparing different models in terms of R^2_{adj} or RMSE is one way to help select the best model. Also, examining the test statistics on the transformed xs will aid in determining if a nonlinear specification is preferred. Finally, and most important, the model specification itself should be made secure, as discussed in the previous chapter.

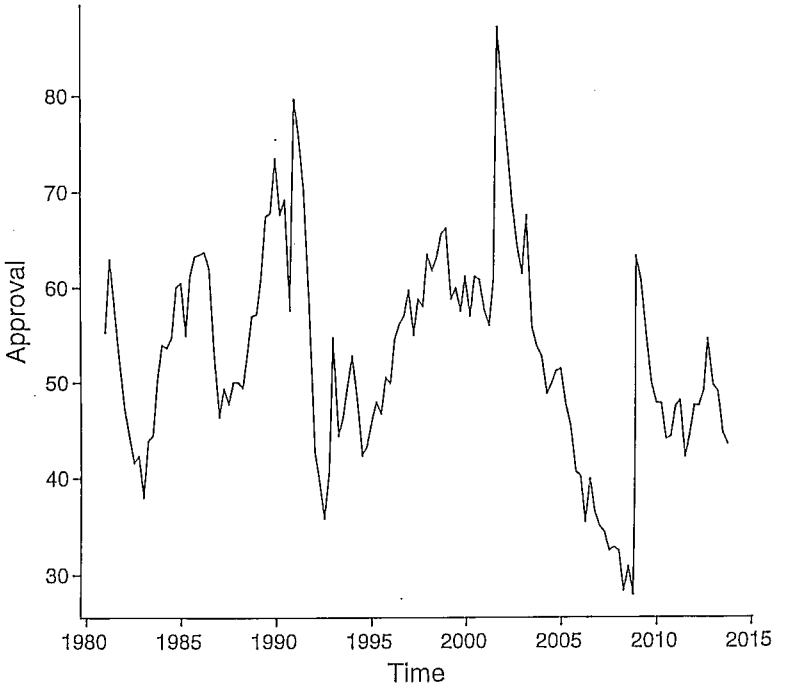
Determinants of Presidential Popularity: A Multiple Regression Example

Let us now turn to a new applied example, from the field of political science. A common question for political scientists is the following: Does

the economy have an effect on presidential popularity? (See the literature reviews in Bellucci & Lewis-Beck, 2011; Lewis-Beck & Stegmaier, 2013.) Since President Truman, the Gallup Poll has collected monthly public opinion data on presidential approval. For our model, we averaged the data to make it quarterly, focusing on the time period from 1981 through 2013. (This use of quarterly data reduces the noise from measurement error.) A time series of the data is presented below with percentage of approval on the y-axis and quarter of the year on the x-axis (Figure 4.3).

The time series fluctuates but appears to be reasonably stable across time: The variance does not explode, and the mean seems to be centered around a 55% approval rating. (Substantively, this is interesting, as it suggests that the public, on average, supports the president.) We want to know the effect of changes in unemployment and inflation, as well as the incumbency effect—that is, are presidents more or less popular during their second term? Prior research indicates it takes two quarters for changes in unemployment to affect voter attitudes, so unemployment will be lagged

Figure 4.3 Presidential Approval Time Series



two quarters (this is denoted with the subscript, $t-2$; in general, time-series data are subscripted to indicate the time period). To capture inflation, percent change in the consumer price index (CPI) will be added as an independent variable. All economic data came from the St. Louis Federal Reserve's FRED Database. OLS is applied to the following specification:

$$y_t = b_0 + b_1 x_{1,t-2} + b_2 x_{2,t} + b_3 x_{3,t} + e_t$$

where y_t = presidential approval (percent approving), $x_{1,t-2}$ = unemployment rate (lagged two quarters), $x_{2,t}$ = percent change in CPI, $x_{3,t}$ = second term dummy variable (0 = first term, 1 = second term), and e_t = error (all variables are measured quarterly). We get the following model output:

$$\hat{y}_t = 72.25 - 2.51x_{1,t-2} - 0.72x_{2,t} - 7.20x_{3,t}$$

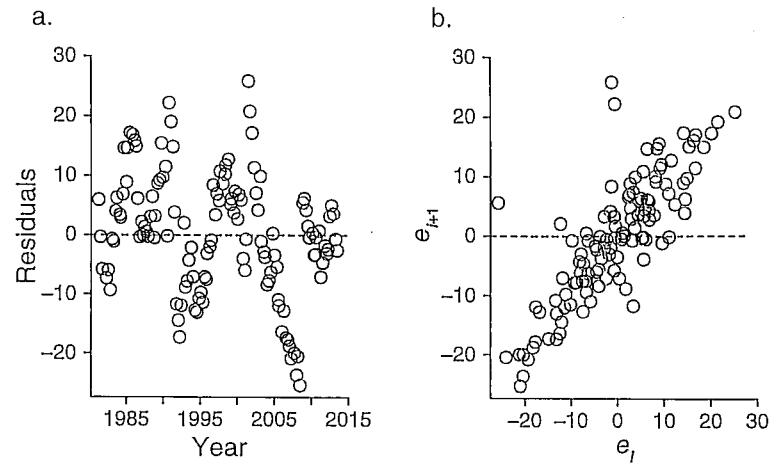
t-statistic	(15.89)	(–4.25)	(–.42)	(–3.53)
p-value	(<.001)	(<.001)	(0.67)	(<.001)

$R^2 = .15$ Adj. $R^2 = .13$ $n = 132$ $s_e = 10.30$

It appears that higher unemployment and being a second-term president have a statistically significant negative effect on presidential popularity. Higher inflation also appears associated with a lower approval rating, but the *p*-value is .67, so failing a conventional .05 statistical significance test.

Before interpreting the estimates, we need to check the assumptions. Recall Assumption 3c, which states that the error terms must be uncorrelated. Observations are typically independent when working with cross-sectional data. With time-series data, however, correlation between data points is common. Intuitively, this makes sense; for example, presidential popularity in one quarter is likely closely linked to popularity the next quarter. There are graphical as well as statistical tests to check for this autocorrelation. The first method is to make an index plot of the residuals (Figure 4.4a). If no autocorrelation is present, the residuals will appear as random noise across the time index. In this case, there are a few subtle patterns—long runs of residuals above or below zero. A second plot to detect autocorrelation is examining the residuals over successive time points (e_t vs. e_{t+1}). If the residuals are uncorrelated, the plot should look like random noise centered at zero. In Figure 4.4b, we see this is not the case. There seems to be a strong positive correlation between successive residuals. A final more formal statistical test comes from calculation of the Breusch-Godfrey statistic.

Figure 4.4 (a-b) Residual Plots for Time-Series Data



The Breusch-Godfrey (BG) test is a general statistical test that checks for autocorrelation between the error terms. The intuition behind the test is straightforward. We regress the observed residuals from the model (e_t) on the independent variables, as well as successive lagged residuals (e_{t-1}, e_{t-2}, etc.). If the R^2 from this model is large enough, there exists correlation between the error terms.[7] The null hypothesis of the BG test is no autocorrelation. Therefore, if we reject the null, it suggests autocorrelation is a problem. Most statistical packages offer a Breusch-Godfrey test. When we test the residuals from the current model, we get a *p*-value < .001. Hence, we reject the null hypothesis of no autocorrelation. This matches the conclusion from the graphical tests in Figure 4.4 (a-b).

There are many remedial measures for dealing with autocorrelation. If testing the significance of the coefficients is of no interest, we could ignore the problem—the estimates will still be unbiased even if the standard errors are incorrect. If we want to make substantive conclusions about the significance of the independent variable coefficients, however, a correction is required. One method is to add a lagged dependent variable to the right-hand side of the model. From a theory perspective, this makes sense: Presidential approval is not completely random from quarter to quarter, which may be because attitudes about the president tend to reinforce themselves over time. Also, there are likely many unobserved factors from one quarter that help shape public opinion of the president in the next quarter—in addition to the observed variables we have in the model.

Therefore, we incorporate the one quarter lagged approval dependent variable into the model as follows:

$$y_t = b_0 + b_1 x_{1,t-2} + b_2 x_{2,t} + b_3 x_{3,t} + b_4 y_{t-1} + e_t$$

Let us now estimate the presidential approval model following this new specification. We get the following output:

	$\hat{y}_t = 17.47 - 0.78x_{1,t-2} - 1.90x_{2,t} - 2.65x_{3,t} + 0.81y_{t-1}$				
t-statistic	(4.14)	(–2.22)	(–1.96)	(–2.24)	(16.40)
p-value	(<.001)	(0.02)	(0.05)	(0.02)	(<0.001)

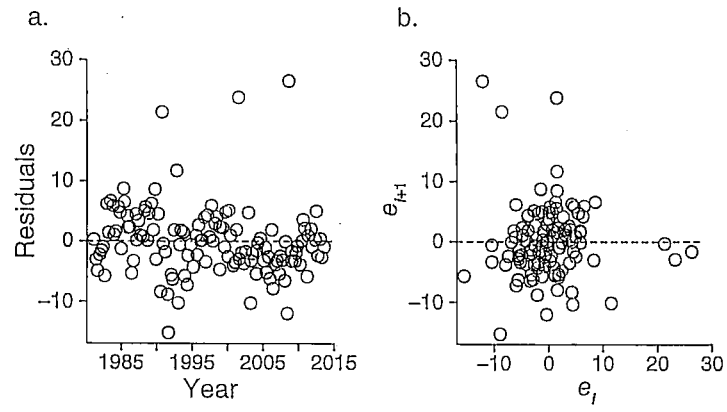
$R^2 = .73$ Adj. $R^2 = .72$ $n = 132$ $s_e = 5.82$

Unemployment and inflation still have a negative effect on the president's popularity. Unemployment is statistically significant at the .05 level. A one-unit positive change (i.e., a 1 percentage point increase) in unemployment two quarters ago reduces the current approval rating by approximately 0.8 percentage points. Inflation is still negative but now also statistically significant, with a 1 percentage point increase in inflation yielding an almost 2 percentage point decrease in popularity. The incumbent variable is also still negative and statistically significant. We can interpret this as the popularity cost of governing for a second term, which is estimated at nearly 3 percentage points. The new variable, lagged approval, is positive, $b_4 = 0.81$, and the t-statistic is greater than 2. Thus, it appears that there is significant association in approval from one quarter to the next. Last, the R^2_{adj} is now .72. Our model is explaining a much greater proportion of the variation in presidential approval. As a final check, let us look at the residuals to make sure the autocorrelation is no longer present.

The index plot (Figure 4.5a) looks like random scatter around zero.[8] We can see in Figure 4.5b that the strong linear pattern present in Figure 4.4b is gone. It appears the autocorrelation has been corrected. To confirm, we run a BG test on the residuals one more time: The statistical test reports a p-value of .85. We fail to reject the null hypothesis of no autocorrelation.[9] Assumption 3c is satisfied.

There are many other solutions to handling the unique problems encountered when modeling time-series data (e.g., generalized least squares). Hopefully, this example shows some of the complications and violations of OLS assumptions that come when working with repeated measures data.

Figure 4.5 (a-b) Residuals Plots for Final Presidential Approval Model



Presentation of Regression Results in a Research Paper

Throughout this monograph, we have presented equations horizontally, running from left to right, as is traditionally done in statistics and econometrics textbooks. This format works well for teaching purposes. However, it is not necessarily the most efficient or readable way to present them in a research paper, where often the author will have several specifications of a model with the same dependent variable. In such cases, a common format is vertical, running from top to bottom (usually beginning with a dependent variable header, then going through the slope estimates, and ending with fit statistics and the sample size). In this way, equations are lined up, one after the other. In Table 4.1, we see an example, which draws on the presidential popularity data analysis just carried out. We see the two equations that were analyzed, with their supporting statistics. (Parentheses below the coefficients contain the standard errors.) Here two equations are sufficient, but it can be readily seen that three, four, or even five model specifications could be incorporated, in sequence.

What Next?

The workhorse of nonexperimental social science research remains the classical linear regression model, estimated with ordinary least squares (Krueger & Lewis-Beck, 2008). Comprehension of the material in this

Table 4.1 Presidential Popularity Models, Quarterly Data, Years 1981–2013

	1	*2*
Unemployment Rate (*t*–2)	–2.51**	–0.78*
	(0.592)	(0.351)
Percent Change in CPI	–0.72	–1.90
	(1.718)	(0.974)
Second Term Dummy	–7.20**	–2.65*
	(2.038)	(1.185)
Presidential Popularity (*t*–1)		0.81**
		(0.049)
Constant	72.25**	17.47**
	(4.546)	(4.216)
R-Squared	0.15	0.73
Adj. *R*-Squared	0.13	0.72
Root MSE	10.3	5.82
Observations	132	132

SOURCE: Gallup Polling Data and St. Louis Federal Reserve's FRED Database. Both time-series are quarterly (averaged from monthly measures).

NOTE: Models 1 and 2 estimated with OLS.

*$p < .05$, **$p < .01$.

monograph should permit the reader to use regression analysis widely and easily. Once the regression assumptions, which we have spelled out, are met, the analyst can have considerable confidence in making inferences from OLS about how the real world works. But ordinary least squares, versatile as it is, has limits, and these limits have much to do with measurement issues. Two such issues are worth mentioning here. The first concerns the precision of the dependent variable; the second concerns the exogenous status of the independent variables.

Let us take the first concern. As explained, when the independent variables have less than interval precision, such as ordinal and nominal variables, they can be "dummied up" and desirable estimation properties of OLS saved. However, the picture shows less bright when the dependent variable has categories, especially multiple categories. If it is a simple dichotomy (i.e., with two categories), then OLS can provide unbiased, but less efficient, estimates. That loss of efficiency means that, rather than least squares, maximum likelihood techniques (MLE), such as binary logit or probit, are

generally preferred. If the dependent variable has several ordered categories, then an ordinal logistic regression should be considered. When the dependent variable has multiple, unordered categories, then a multinomial logistic regression is called for.

Let us turn to the second concern, that of questionable exogeneity of one or more independent variables. Outside of a scientific laboratory, truly randomized controlled experiments are difficult, if not impossible, to conduct. For instance, it would be unethical to randomly assign one group of individuals to smoke and compare them with a random cohort of nonsmokers. Social scientists usually have to rely on observational data where the explanatory variables are not exogenous (e.g., randomly assigned). This makes causal inference difficult as there may be many confounding factors correlated with the independent variables, as well as the dependent variable. If x is not truly exogenous, then the OLS slope estimates will be biased. A common strategy for overcoming this exogeneity problem is instrumental variables. The leading technique here is called two-stage least squares (2SLS) and involves rendering the offending x variables effectively exogenous, through the use of available "instruments." Instrumental variables (IV) are covariates that are correlated with the offending explanatory variables but uncorrelated with the error term. In the smoking example, cigarette price could be an instrumental variable if the dependent variable was lung cancer. (Cigarette price is correlated with smoking but would likely not have an effect on health outcomes.) Happily, a firm grasp of classical regression analysis, as explicated here, will speed the student's mastery of this technique, as well as the MLE single-equation techniques just outlined.

Notes

1. Venn diagrams applied to regression analysis are called a Ballentine; they were invented by Cohen and Cohen (1975). For a more detailed description, see Kennedy (2008, pp. 45–47).
2. There are automated atheoretical techniques, such as *stepwise regression,* which attempt to partition variance. Such techniques virtually guarantee biased parameter estimates and should be avoided.
3. A variable is centered as follows: $z_i = x_i - \bar{x}$. In this case, after the variable school size is centered, both z_i and z_i^2 are entered into the model as covariates.
4. Rather than interpreting the intercept as the average test score when school size equals 0, we now interpret it as the average test score for the average school size of $\bar{x} = 1{,}146$, as in this example. The slope coefficient on school size is interpreted at the mean, too. A one-person increase in school size (for the average school with 1,146 students) will result in a b_1 increase in average test score. Centering is helpful from a substantive perspective as well since the interpretation of the slope at $x_1 = 0$ has no real meaning—no actual school has zero students.

5. The standardized regression coefficient has sometimes been called a *beta weight* and represented with the Greek symbol for beta, β. However, this makes for confusion, as β routinely is used to designate the population slope.
6. An example of a model that is linear in the parameters is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. In other words, each parameter is raised to the first power. An example of a model that is nonlinear in parameters is $y = \beta_0 + \beta_1 x_1^{\beta_2} + \varepsilon$.
7. Recall from Chapter 2 that by design, the independent variables x will be uncorrelated with the error term, as estimated by the residuals. This is a mathematical property of OLS. Therefore, if the R^2 from the BG model is high, the explanatory power must be coming from the successive residuals.
8. The curious reader may notice three large positive outlier residuals in Figure 4.5a. These correspond to major shocks, or "rally around the flag" events, to the country: The first corresponds to the first Gulf War, the second to the attacks of 9-11, and the third the election of Barack Obama.
9. Another, common, autocorrelation test is the Durbin-Watson (DW) test. However, it is not appropriate when there is a lagged dependent variable on the right-hand side of the equation, as is the case here. Fortunately, the BG test is a more general test than DW and is appropriate here.

APPENDIX

A derivation of the least squares estimates without the use of calculus:

Recall, a quadratic function with one variable has the form

$$f(x) = ax^2 + bx + c$$

where a, b, and c are constants, and we assume $a > 0$.

Using basic algebra, we can rewrite this equation as

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 + \left(c - \frac{b^2}{4a}\right)$$

To minimize this function in terms of x, we can focus on the first term since x does not appear in the second. Because this is a squared term, it is never negative and will be minimized when set to zero. Thus, to minimize the whole equation, we simply need to solve

$$x + \frac{b}{2a} = 0$$

or

$$x = \frac{-b}{2a}$$

Graphically, we can think of this solution as the vertex of an upward-facing parabola.

Returning to the least squares equation, we can apply the same technique to the following minimization problem:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Plugging in the equation for $\hat{y}$, we can rewrite this as

$$SSE = \sum_{i=1}^{n} ((y_i - b_0) - b_1 x_i)^2$$

Using the distributive law, we can expand this to

$$SSE = \sum_{i=1}^{n}(y_i - b_0)^2 - 2b_1\sum_{i=1}^{n}x_i(y_i - b_0) + b_1^2\sum_{i=1}^{n}x_i^2$$

$$= b_1^2\sum_{i=1}^{n}x_i^2 + 2b_0b_1\sum_{i=1}^{n}x_i + nb_0^2 - 2b_1\sum_{i=1}^{n}x_iy_i - 2b_0\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}y_i^2$$

We can now see that the SSE is simply a quadratic function in terms of either b_0 or b_1. We can find the lowest point of each of these parabolas by combining like terms and finding the vertex as derived above.

$$f(b_0) = nb_0^2 + (2b_1\sum_{i=1}^{n}x_i - 2\sum_{i=1}^{n}y_i)b_0 + (b_1^2\sum_{i=1}^{n}x_i^2 - 2b_1\sum_{i=1}^{n}x_iy_i + \sum_{i=1}^{n}y_i^2)$$

$$f(b_1) = (\sum_{i=1}^{n}x_i^2)b_1^2 + (2b_0\sum_{i=1}^{n}x_i - 2\sum_{i=1}^{n}x_iy_i)b_1 + (nb_0^2 - 2b_0\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}y_i^2)$$

Using the general minimization solution ($x = \frac{-b}{2a}$) each of these quadratic equations is minimized at

$$b_0 = \frac{\sum_{i=1}^{n}y_i - b_1\sum_{i=1}^{n}x_i}{n}$$

$$b_1 = \frac{\sum_{i=1}^{n}x_iy_i - b_0\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2}$$

With two equations and two unknown variables, we can use substitution (and the trick $n\bar{x} = \sum_{i=1}^{n}x_i$) to rewrite these in the same form as shown in the text.

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

REFERENCES

Andersen, R. (2008). *Modern methods for robust regression* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-152). Thousand Oaks, CA: Sage.

Bellucci, P., & Lewis-Beck, M. S. (2011). A stable popularity function? Cross-national analysis. *European Journal of Political Research, 50,* 190–211.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillside, NJ: Laurence Erlbaum.

Enders, W. (2010). *Applied econometric time series* (3rd ed.). New York, NY: John Wiley.

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

Gujarati, D., & Porter, D. (2009). *Basic econometrics* (5th ed.). New York, NY: McGraw-Hill.

Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Oxford, UK: Blackwell.

Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.

Krueger, J., & Lewis-Beck, M. S. (2008). Is OLS dead? *The Political Methodologist, 15,* 2–4.

Larocca, R. T. (2005). Reconciling conflicting Gauss-Markov assumptions in the classical linear regression model. *Political Analysis, 13,* 188–207.

Leamer, E. E., & Leonard, H. (1983). Reporting the fragility of regression estimates. *Review of Economics and Statistics, 65,* 306–347.

Leithwood, K., & Jantzi, D. (2009). A review of empirical evidence about school size effects: A policy perspective. *Review of Educational Research, 79*(1), 464–490.

Lewis-Beck, M. S. (1995). *Data analysis: An introduction* (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-103). Thousand Oaks, CA: Sage.

Lewis-Beck, M. S. (2004a). Degrees of freedom. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The Sage encyclopaedia of social science research methods* (Vol. 1, pp. 243–244). Thousand Oaks, CA: Sage.

Lewis-Beck, M. S. (2004b). Regression. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The Sage encyclopaedia of social science research methods* (Vol. 3, pp. 935–938). Thousand Oaks, CA: Sage.

Lewis-Beck, M. S., & Stegmaier, M. (2013). The V-P function revisited: A survey of the literature on vote and popularity functions after over 40 years. *Public Choice, 157,* 367–385.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54,* 217–224.

Pearson, K. (1913). On the probable error of a coefficient of correlation as found from a fourfold table. *Biometrika, 9*(1–2), 22–27.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western CENGAGE Learning.

INDEX