

STATISTICAL AND ECONOMETRIC METHODS FOR TRANSPORTATION DATA ANALYSIS

SECOND EDITION

$$P_n^m(i) = \int_{\mathbf{X}} P_n(i) f(\boldsymbol{\beta} | \boldsymbol{\varphi}) d\boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{Y}$$

$$LL = \sum_{n=1}^{N_s} \sum_{i=1}^I S_m \ln \left[P_n^m(i) \right]$$

$$E(\xi | i) = \left(\frac{1}{P_i} \right) \int_{\mathbb{R}^J} E(\xi | \gamma) \prod_{j=1}^J f(\varepsilon_j) d\gamma$$

$$L(y|\beta_1,...,\beta_k,\mu_2,...,\mu_{I-1}) = \prod_{n=1}^N \prod_{i=1}^I [\Phi(\mu_i - \beta \mathbf{X}_n) - \Phi(\mu_{i+1} - \beta \mathbf{X}_n)]^{\delta_{in}}$$

$$S(t) = \int_0^\infty S(t|w)g(w)dw = \left[1 + \theta(\lambda t)^P \right]^{-1/\theta}$$

Simon P. Washington
Matthew G. Karlaftis
Fred L. Mannerling

$$\beta^T \mathbf{X}_n = \frac{1}{\eta} \ln \sum_{\forall l \neq i} \exp(\eta \beta_l \mathbf{X}_{ln})$$

$$P_n(i) = P\left(\beta_i \mathbf{X}_m + \varepsilon_m \geq \max_{\forall l \neq i} (\beta_l \mathbf{X}_{ln} + \varepsilon_{ln})\right)$$



CRC Press

Taylor & Francis Group

A CHAPMAN & HALL BOOK

$$\frac{\partial V_{in}}{\partial Inc_n} y_{in}^\theta + \frac{\partial V_{in}}{\partial p_m} = 0$$

STATISTICAL AND
ECONOMETRIC
METHODS FOR
TRANSPORTATION
DATA **A**NALYSIS
SECOND **E**DITION

STATISTICAL AND **E**CONOMETRIC **M**ETHODS FOR **T**RANSPORTATION **D**ATA **A**NALYSIS

SECOND **E**DITION

Simon P. Washington
Matthew G. Karlaftis
Fred L. Mannering



CRC Press
Taylor & Francis Group
Boca Raton London New York

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

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

© 2011 by Taylor and Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-8286-9 (Ebook-PDF)

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

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

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

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

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

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

To David, Judy, Karen, Tracy, Devon, and Samantha

Simon

To Amy, George, John, Stavriani, Nikolas

Matt

To Jill, Willa, and Freyda

Fred

Contents

Preface.....	xv
--------------	----

Part I Fundamentals

1. Statistical Inference I: Descriptive Statistics	3
1.1 Measures of Relative Standing.....	3
1.2 Measures of Central Tendency.....	4
1.3 Measures of Variability	5
1.4 Skewness and Kurtosis	9
1.5 Measures of Association	11
1.6 Properties of Estimators.....	14
1.6.1 Unbiasedness.....	14
1.6.2 Efficiency	15
1.6.3 Consistency.....	16
1.6.4 Sufficiency	16
1.7 Methods of Displaying Data	17
1.7.1 Histograms	17
1.7.2 Ogives	18
1.7.3 Box Plots	19
1.7.4 Scatter Diagrams.....	19
1.7.5 Bar and Line Charts.....	20
2. Statistical Inference II: Interval Estimation, Hypothesis Testing and Population Comparisons.....	25
2.1 Confidence Intervals.....	25
2.1.1 Confidence Interval for μ with Known σ^2	26
2.1.2 Confidence Interval for the Mean with Unknown Variance.....	28
2.1.3 Confidence Interval for a Population Proportion.....	28
2.1.4 Confidence Interval for the Population Variance.....	29
2.2 Hypothesis Testing	30
2.2.1 Mechanics of Hypothesis Testing.....	31
2.2.2 Formulating One- and Two-Tailed Hypothesis Tests....	33
2.2.3 The p -Value of a Hypothesis Test	36
2.3 Inferences Regarding a Single Population	36
2.3.1 Testing the Population Mean with Unknown Variance.....	37

2.3.2	Testing the Population Variance	38
2.3.3	Testing for a Population Proportion.....	38
2.4	Comparing Two Populations.....	39
2.4.1	Testing Differences between Two Means: Independent Samples	39
2.4.2	Testing Differences between Two Means: Paired Observations	42
2.4.3	Testing Differences between Two Population Proportions	43
2.4.4	Testing the Equality of Two Population Variances	45
2.5	Nonparametric Methods.....	46
2.5.1	Sign Test	47
2.5.2	Median Test.....	52
2.5.3	Mann–Whitney <i>U</i> Test	52
2.5.4	Wilcoxon Signed-Rank Test for Matched Pairs	55
2.5.5	Kruskal–Wallis Test.....	56
2.5.6	Chi-Square Goodness-of-Fit Test.....	58

Part II Continuous Dependent Variable Models

3.	Linear Regression	63
3.1	Assumptions of the Linear Regression Model	63
3.1.1	Continuous Dependent Variable <i>Y</i>	64
3.1.2	Linear-in-Parameters Relationship between <i>Y</i> and <i>X</i>	64
3.1.3	Observations Independently and Randomly Sampled.....	65
3.1.4	Uncertain Relationship between Variables	65
3.1.5	Disturbance Term Independent of <i>X</i> and Expected Value Zero	65
3.1.6	Disturbance Terms Not Autocorrelated	66
3.1.7	Regressors and Disturbances Uncorrelated.....	66
3.1.8	Disturbances Approximately Normally Distributed	66
3.1.9	Summary.....	67
3.2	Regression Fundamentals.....	67
3.2.1	Least Squares Estimation.....	69
3.2.2	Maximum Likelihood Estimation.....	73
3.2.3	Properties of OLS and MLE Estimators.....	74
3.2.4	Inference in Regression Analysis	75
3.3	Manipulating Variables in Regression.....	79
3.3.1	Standardized Regression Models.....	79
3.3.2	Transformations	80
3.3.3	Indicator Variables	82

3.4	Estimate a Single Beta Parameter	83
3.5	Estimate Beta Parameter for Ranges of a Variable	83
3.6	Estimate a Single Beta Parameter for $m - 1$ of the m Levels of a Variable	84
3.6.1	Interactions in Regression Models	84
3.7	Checking Regression Assumptions	87
3.7.1	Linearity	88
3.7.2	Homoscedastic Disturbances	90
3.7.3	Uncorrelated Disturbances	93
3.7.4	Exogenous Independent Variables	93
3.7.5	Normally Distributed Disturbances	95
3.8	Regression Outliers	98
3.8.1	The Hat Matrix for Identifying Outlying Observations	99
3.8.2	Standard Measures for Quantifying Outlier Influence	101
3.8.3	Removing Influential Data Points from the Regression	101
3.9	Regression Model GOF Measures	106
3.10	Multicollinearity in the Regression	110
3.11	Regression Model-Building Strategies	112
3.11.1	Stepwise Regression	112
3.11.2	Best Subsets Regression	113
3.11.3	Iteratively Specified Tree-Based Regression	113
3.12	Estimating Elasticities	113
3.13	Censored Dependent Variables—Tobit Model	114
3.14	Box–Cox Regression	116
4.	Violations of Regression Assumptions	123
4.1	Zero Mean of the Disturbances Assumption	123
4.2	Normality of the Disturbances Assumption	124
4.3	Uncorrelatedness of Regressors and Disturbances Assumption	125
4.4	Homoscedasticity of the Disturbances Assumption	127
4.4.1	Detecting Heteroscedasticity	129
4.4.2	Correcting for Heteroscedasticity	131
4.5	No Serial Correlation in the Disturbances Assumption	135
4.5.1	Detecting Serial Correlation	137
4.5.2	Correcting for Serial Correlation	139
4.6	Model Specification Errors	142
5.	Simultaneous-Equation Models	145
5.1	Overview of the Simultaneous-Equations Problem	145
5.2	Reduced Form and the Identification Problem	146

5.3	Simultaneous-Equation Estimation.....	148
5.3.1	Single-Equation Methods	148
5.3.2	System-Equation Methods.....	149
5.4	Seemingly Unrelated Equations	155
5.5	Applications of Simultaneous Equations to Transportation Data	156
	Appendix 5A. A Note on GLS Estimation.....	159
6.	Panel Data Analysis	161
6.1	Issues in Panel Data Analysis.....	161
6.2	One-Way Error Component Models.....	163
6.2.1	Heteroscedasticity and Serial Correlation	166
6.3	Two-Way Error Component Models.....	167
6.4	Variable-Parameter Models	172
6.5	Additional Topics and Extensions	173
7.	Background and Exploration in Time Series	175
7.1	Exploring a Time Series	176
7.1.1	Trend Component	176
7.1.2	Seasonal Component	176
7.1.3	Irregular (Random) Component	179
7.1.4	Filtering of Time Series	179
7.1.5	Curve Fitting.....	179
7.1.6	Linear Filters and Simple Moving Averages.....	179
7.1.7	Exponential Smoothing Filters	180
7.1.8	Difference Filter.....	185
7.2	Basic Concepts: Stationarity and Dependence.....	188
7.2.1	Stationarity.....	188
7.2.2	Dependence	188
7.2.3	Addressing Nonstationarity	190
7.2.4	Differencing and Unit-Root Testing	191
7.2.5	Fractional Integration and Long Memory.....	194
7.3	Time Series in Regression.....	197
7.3.1	Serial Correlation	197
7.3.2	Dynamic Dependence.....	197
7.3.3	Volatility	198
7.3.4	Spurious Regression and Cointegration.....	200
7.3.5	Causality	202
8.	Forecasting in Time Series: Autoregressive Integrated Moving Average (ARIMA) Models and Extensions	207
8.1	Autoregressive Integrated Moving Average Models	207
8.2	Box-Jenkins Approach	210

8.2.1	Order Selection.....	210
8.2.2	Parameter Estimation.....	212
8.2.3	Diagnostic Checking	213
8.2.4	Forecasting	214
8.3	Autoregressive Integrated Moving Average Model Extensions	218
8.3.1	Random Parameter Autoregressive Models	219
8.3.2	Stochastic Volatility Models	222
8.3.3	Autoregressive Conditional Duration Models	224
8.3.4	Integer-Valued ARMA Models	224
8.4	Multivariate Models	225
8.5	Nonlinear Models	227
8.5.1	Testing for Nonlinearity	227
8.5.2	Bilinear Models	228
8.5.3	Threshold Autoregressive Models	229
8.5.4	Functional Parameter Autoregressive Models	230
8.5.5	Neural Networks	231
9.	Latent Variable Models	235
9.1	Principal Components Analysis	235
9.2	Factor Analysis	241
9.3	Structural Equation Modeling	244
9.3.1	Basic Concepts in Structural Equation Modeling	246
9.3.2	Fundamentals of Structural Equation Modeling	249
9.3.3	Nonideal Conditions in the Structural Equation Model.....	251
9.3.4	Model Goodness-of-Fit Measures.....	252
9.3.5	Guidelines for Structural Equation Modeling.....	255
10.	Duration Models.....	259
10.1	Hazard-Based Duration Models	259
10.2	Characteristics of Duration Data	263
10.3	Nonparametric Models	264
10.4	Semiparametric Models	265
10.5	Fully Parametric Models.....	268
10.6	Comparisons of Nonparametric, Semiparametric, and Fully Parametric Models	272
10.7	Heterogeneity	274
10.8	State Dependence	276
10.9	Time-Varying Covariates	277
10.10	Discrete-Time Hazard Models	277
10.11	Competing Risk Models.....	279

Part III Count and Discrete Dependent Variable Models

11. Count Data Models	283
11.1 Poisson Regression Model	283
11.2 Interpretation of Variables in the Poisson Regression Model	284
11.3 Poisson Regression Model Goodness-of-Fit Measures.....	286
11.4 Truncated Poisson Regression Model	290
11.5 Negative Binomial Regression Model.....	292
11.6 Zero-Inflated Poisson and Negative Binomial Regression Models	295
11.7 Random-Effects Count Models	300
12. Logistic Regression	303
12.1 Principles of Logistic Regression.....	303
12.2 Logistic Regression Model.....	304
13. Discrete Outcome Models	309
13.1 Models of Discrete Data.....	309
13.2 Binary and Multinomial Probit Models.....	310
13.3 Multinomial Logit Model	312
13.4 Discrete Data and Utility Theory	316
13.5 Properties and Estimation of MNL Models.....	318
13.5.1 Statistical Evaluation	321
13.5.2 Interpretation of Findings	323
13.5.3 Specification Errors.....	325
13.5.4 Data Sampling	330
13.5.5 Forecasting and Aggregation Bias	331
13.5.6 Transferability	333
13.6 Nested Logit Model (Generalized Extreme Value Models).....	334
13.7 Special Properties of Logit Models.....	342
14. Ordered Probability Models	345
14.1 Models for Ordered Discrete Data	345
14.2 Ordered Probability Models with Random Effects	352
14.3 Limitations of Ordered Probability Models.....	358
15. Discrete/Continuous Models	361
15.1 Overview of the Discrete/Continuous Modeling Problem	361
15.2 Econometric Corrections: Instrumental Variables and Expected Value Method	363
15.3 Econometric Corrections: Selectivity-Bias Correction Term.....	366
15.4 Discrete/Continuous Model Structures	368
15.5 Transportation Application of Discrete/Continuous Model Structures	372

Part IV Other Statistical Methods

16. Random-Parameter Models	375
16.1 Random-Parameter Multinomial Logit Model (Mixed Logit Model)	375
16.2 Random-Parameter Count Models.....	381
16.3 Random-Parameter Duration Models.....	384
17. Bayesian Models	387
17.1 Bayes' Theorem	387
17.2 MCMC Sampling-Based Estimation.....	389
17.3 Flexibility of Bayesian Statistical Models via MCMC Sampling-Based Estimation.....	395
17.4 Convergence and Identifiability Issues with MCMC Bayesian Models.....	396
17.5 Goodness-of-Fit, Sensitivity Analysis, and Model Selection Criterion Using MCMC Bayesian Models	399
Appendix A Statistical Fundamentals	403
A.1 Matrix Algebra Review	403
A.1.1 Matrix Multiplication	404
A.1.2 Linear Dependence and Rank of a Matrix	406
A.1.3 Matrix Inversion (Division)	406
A.1.4 Eigenvalues and Eigenvectors.....	408
A.1.5 Useful Matrices and Properties of Matrices	409
A.1.6 Matrix Algebra and Random Variables.....	410
A.2 Probability, Conditional Probability, and Statistical Independence.....	412
A.3 Estimating Parameters in Statistical Models—Least Squares and Maximum Likelihood.....	413
A.4 Useful Probability Distributions.....	415
A.4.1 The Z Distribution	416
A.4.2 The <i>t</i> Distribution	417
A.4.3 The χ^2 Distribution	418
A.4.4 The <i>F</i> Distribution.....	419
Appendix B Glossary of Terms.....	421
Appendix C Statistical Tables	459
Appendix D Variable Transformations.....	483
D.1 Purpose of Variable Transformations	483
D.2 Commonly Used Variable Transformations.....	484
D.2.1 Parabolic Transformations.....	484

D.2.2	Hyperbolic Transformations	485
D.2.3	Exponential Functions	485
D.2.4	Inverse Exponential Functions	487
D.2.5	Power Functions.....	488
References		489
Index		511

Preface

Transportation is integral to developed societies. It is responsible for personal mobility, which includes access to services, goods, and leisure. It is also a key element in the delivery of consumer goods. Regional, state, national, and the world economies rely upon the efficient and safe functioning of transportation facilities.

Besides the sweeping influence transportation has on economic and social aspects of modern society, transportation issues pose challenges to professionals across a wide range of disciplines, including transportation engineers, urban and regional planners, economists, logisticians, psychologists, systems and safety engineers, social scientists, law enforcement and security professionals, and consumer theorists. Where to place and expand transportation infrastructure; how to safely and efficiently operate and maintain infrastructure; and how to spend valuable resources to improve mobility and access to goods, services, and health care are among the decisions made routinely by transportation-related professionals.

Many transportation-related problems and challenges involve stochastic processes that are influenced by observed and unobserved factors in unknown ways. The stochastic nature of transportation problems is largely a result of the role that people play in transportation. Transportation system users are routinely faced with decisions in contexts such as what transportation mode to use, which vehicle to purchase, whether to participate in a vanpool or telecommute, where to relocate a business, whether to support a proposed light-rail project, and whether to utilize traveler information before or during a trip. These decisions involve various degrees of uncertainty. Transportation system managers and governmental agencies face similar stochastic problems in determining how to measure and compare system measures of performance, where to invest in safety improvements, how to efficiently operate transportation systems, and how to estimate transportation demand.

As a result of the complexity, diversity, and stochastic nature of transportation problems, the analytical toolbox required of the transportation analyst must be broad. This book describes and illustrates some of the tools commonly used in transportation data analysis. Every book must achieve a balance between depth and breadth of theory and applications, given the intended audience. This book targets two general audiences. First, it serves as a textbook for advanced undergraduate, masters, and Ph.D. students in transportation-related disciplines, including engineering, economics, urban and regional planning, and sociology. There is sufficient material to cover two three-unit semester courses in analytical methods. Alternatively, a one-semester course could consist of a subset of topics covered in this book.

The publisher's Web site, www.crcpress.com, contains the datasets used to develop this book so that applied-modeling problems will reinforce the modeling techniques discussed throughout the text. To facilitate teaching from this text, the Web site also contains Microsoft PowerPoint® presentations for each of the chapters in the book. These presentations, new to the second edition, will significantly improve the adoptability of this text for college, university, and professional instructors.

The book also serves as a technical reference for researchers and practitioners wishing to examine and understand a broad range of analytical tools required to solve transportation problems. It provides a wide breadth of transportation examples and case studies covering applications in various aspects of transportation planning, engineering, safety, and economics. Sufficient analytical rigor is provided in each chapter so that fundamental concepts and principles are clear and numerous references are provided for those seeking additional technical details and applications.

Part I of the book provides statistical fundamentals (Chapters 1 and 2). This section is useful for refreshing fundamentals and for sufficiently preparing students for the following sections.

Part II of the book presents continuous dependent variable models. The chapter on linear regression (Chapter 3) devotes additional pages to introduce common modeling practice—examining residuals, creating indicator variables, and building statistical models—and thus serves as a logical starting chapter for readers new to statistical modeling. The subsection on Tobit and censored regressions is new to the second edition. Chapter 4 discusses the impacts of failing to meet linear regression assumptions and presents corresponding solutions. Chapter 5 deals with simultaneous equation models and presents modeling methods appropriate when studying two or more interrelated dependent variables. Chapter 6 presents methods for analyzing panel data—data obtained from repeated observations on sampling units over time, such as household surveys conducted several times to a sample of households. When data are collected continuously over time, such as hourly, daily, weekly, or yearly, time series methods and models are often needed and are discussed in Chapters 7 and 8. New to the second edition is explicit treatment of frequency domain time series analysis, including Fourier and wavelets analysis methods. Latent variable models, discussed in Chapter 9, are used when the dependent variable is not directly observable and is approximated with one or more surrogate variables. The final chapter in this section, Chapter 10, presents duration models, which are used to model time-until-event data as survival, hazard, and decay processes.

Part III in the book presents count and discrete dependent variable models. Count models (Chapter 11) arise when the data of interest are nonnegative integers. Examples of such data include vehicles in a queue and the number of vehicle crashes per unit time. Zero inflation—a phenomenon observed frequently with count data—is discussed in detail, and a new example and corresponding data set have been added in this second edition. Logistic

regression commonly used to model probabilities of binary outcomes, is presented in Chapter 12, and is unique to the second edition. Discrete outcome models are extremely useful in many study applications, and are described in detail in Chapter 13. A unique feature of the book is that discrete outcome models are first considered statistically, and then later related to economic theories of consumer choice. Ordered probability models (a new chapter for the second edition) are presented in Chapter 14. Discrete/continuous models are presented in Chapter 15 and demonstrate that interrelated discrete and continuous data need to be modeled as a system rather than individually, such as the choice of which vehicle to drive and how far it will be driven.

Finally, Part IV of the book contains new chapters on random-parameter models (Chapter 16) and Bayesian statistical modeling (Chapter 17). Random-parameter models are starting to gain wide acceptance across many fields of study, and this chapter provides a basic introduction to this exciting newer class of models. The chapter on Bayesian statistical models arises from the increasing prevalence of Bayesian inference and Markov Chain Monte Carlo methods (an analytically convenient method for estimating complex Bayes' models). This chapter presents the basic theory of Bayesian models, of Markov Chain Monte Carlo methods of sampling, and presents two separate examples of Bayes' models.

The appendices are complementary to the remainder of the book. Appendix A presents fundamental concepts in statistics, which support analytical methods discussed. Appendix B is an alphabetical glossary of statistical terms that are commonly used and provides a quick and easy reference. Appendix C provides tables of probability distributions used in the book, while Appendix D describes typical uses of data transformations common to many statistical methods.

While the book covers a wide variety of analytical tools for improving the quality of research, it does not attempt to teach all elements of the research process. Specifically, the development and selection of research hypotheses, alternative experimental design methodologies, the virtues and drawbacks of experimental versus observational studies, and issues involved with the collection of data are not discussed. These issues are critical elements in the conduct of research, and can drastically impact the overall results and quality of the research endeavor. It is considered a prerequisite that readers of this book are educated and informed on these critical research elements to appropriately apply the analytical tools presented herein.

Simon P. Washington
Matthew G. Karlaftis
Fred L. Mannering

Part I

Fundamentals

1

Statistical Inference I: Descriptive Statistics

This chapter examines methods and techniques for summarizing and interpreting data. The discussion begins with an examination of numerical descriptive measures. These measures, commonly known as point estimators, support inferences about a population by estimating the values of unknown population parameters using a single value (or point). The chapter also describes commonly used graphical representations of data. Relative to graphical methods, numerical methods provide precise and objectively determined values that are easily manipulated, interpreted, and compared. They permit a more careful analysis of data than more general impressions conveyed by graphical summaries. This is important when the data represent a sample from which population inferences must be made.

While this chapter concentrates on a subset of basic and fundamental issues of statistical analyses, there are countless thorough introductory statistical textbooks that can provide the interested reader with greater detail. For example, Aczel (1993) and Keller and Warrack (1997) provide detailed descriptions and examples of descriptive statistics and graphical techniques. Tukey (1977) is a classic reference on exploratory data analysis and graphical techniques. For readers interested in the properties of estimators (Section 1.7), the books by Gujarati (1992) and Baltagi (1998) are excellent, mathematically rigorous, sources.

1.1 Measures of Relative Standing

A set of numerical observations can be ordered from smallest to largest magnitude. This ordering allows the boundaries of the data to be defined and supports comparisons of the relative position of specific observations. Consider the usefulness of percentile rank in terms of measuring driving speeds on a highway section. In this case, a driver's speed is compared to the speeds of all drivers who drove on the road segment during the measurement period and the relative speed positioned within the group is defined in terms of a percentile. If, for example, the 85th percentile of speed is 63 mph, then 85% of the sample of observed drivers was driving at speeds below 63 mph and 20% were above 63 mph. A percentile is defined as that value below which lies $P\%$ of the values in the remaining sample. For sufficiently

large samples, the position of the P^{th} percentile is given by $(n + 1)P/100$, where n is the sample size.

Quartiles are the percentage points that separate the data into quarters: first quarter, below which lies one quarter of the data, making it the 25th percentile; second quarter, or 50th percentile, below which lies half of the data; third quarter, or 75th percentile point. The 25th percentile is often referred to as the lower or first quartile, the 50th percentile as the median or middle quartile, and the 75th percentile as the upper or third quartile. Finally, the interquartile range (IQR), a measure of the data spread, is defined as the numerical difference between the first and third quartiles.

1.2 Measures of Central Tendency

Quartiles and percentiles are measures of the relative positions of points within a given data set. The median constitutes a useful point because it lies in the center of the data, with half of the data points lying above and half below the median. The median constitutes a measure of the “centrality” of the observations, or central tendency.

Despite the existence of the median, by far the most popular and useful measure of central tendency is the arithmetic mean, or, more succinctly, the sample mean or expectation. The sample mean is another statistical term that measures the central tendency, or average, of a sample of observations. The sample mean varies across samples and thus is a random variable. The mean of a sample of measurements x_1, x_2, \dots, x_n is defined as

$$\text{MEAN}(X) = E[X] = \bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad (1.1)$$

where n is the size of the sample.

When an entire population is examined, the sample mean \bar{X} is replaced by μ , the population mean. Unlike the sample mean, the population mean is constant. The formula for the population mean is

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (1.2)$$

where N is the size of the population.

The mode (or modes because it is possible to have more than one) of a set of observations is the value that occurs most frequently, or the most commonly occurring outcome, and strictly applies to discrete variables

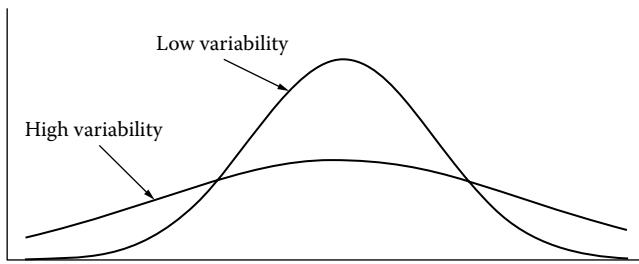
(nominal and ordinal scale variables) as well as count data. Probabilistically, it is the most likely outcome in the sample; it is observed more than any other value. The mode can also be a measure of central tendency.

There are advantages and disadvantages of each of the three central tendency measures. The mean uses and summarizes all of the information in the data in a single numerical measure, and has some desirable mathematical properties that make it useful in many statistical inference and modeling applications. The median, in contrast, is the central most (center) point of ranked data. When computing the median, the exact locations of data points on the number line are not considered; only their relative standing with respect to the central observation is required. Herein lies the major advantage of the median; it is resistant to extreme observations or outliers in the data. The mean is, overall, the most frequently applied measure of central tendency; however, in cases where the data contain numerous outlying observations the median may serve as a more reliable measure. Robust statistical modeling approaches, much like the median, are designed to be resistant to the influence of extreme observations.

If sample data are measured on the interval or ratio scale, then all three measures of centrality (mean, median, and mode) are defined, provided that the level of measurement precision does not preclude the determination of a mode. When data are symmetric and unimodal, the mode, median, and mean are approximately equal (the relative positions of the three measures in cases of asymmetric distributions is discussed in Section 1.4). Finally, if the data are qualitative (measured on the nominal or ordinal scales), using the mean or median is senseless, and the mode must be used. For nominal data, the mode is the category that contains the largest number of observations.

1.3 Measures of Variability

Variability is a statistical term used to describe and quantify the spread or dispersion of data around the center (usually the mean). In most practical situations, knowledge of the average or expected value of a sample is not sufficient to obtain an adequate understanding of the data. Sample variability provides a measure of how dispersed the data are with respect to the mean (or other measures of central tendency). Figure 1.1 illustrates two distributions of data, one that is highly dispersed and another that is relatively less dispersed around the mean. There are several useful measures of variability, or dispersion. One measure previously discussed is the IQR. Another measure is the range—the difference between the largest and the smallest observations in the data. While both the range and the IQR measure data dispersion, the IQR is more resistant to outlying observations. The two most frequently used measures of dispersion are the variance and its square root, the standard deviation.

**FIGURE 1.1**

Examples of high- and low-variability data.

The variance and the standard deviation are typically more useful than the range because, like the mean, they exploit all of the information contained in the observations. The variance of a set of observations, or sample variance, is the average squared deviation of the individual observations from the mean and varies across samples. The sample variance is commonly used as an estimate of the population variance and is given by

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (1.3)$$

When a collection of observations constitutes an entire population, the variance is denoted by σ^2 . Unlike the sample variance, the population variance is constant and is given by

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (1.4)$$

where \bar{X} in Equation 1.3 is replaced by μ .

Because calculation of the variance involves squaring differences of the raw data measurement scales, the measurement unit is the square of the original measurement scale—for example, the variance of measured distances in meters is meters squared. While variance is a useful measure of the relative variability of two sets of measurements, it is often desirable to express variability in the same measurement units as the raw data. Such a measure is the square root of the variance, commonly known as the standard deviation. The formulas for the sample and population standard deviations are given, respectively, as

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (1.5)$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (1.6)$$

Consistent with previous results, the sample standard deviation s^2 is a random variable, whereas the population standard deviation σ is a constant.

A question that frequently arises in the practice of statistics is the reason for dividing by n when computing the population standard deviation and $n - 1$ when computing the sample standard deviation. When a sample is drawn from the population, a sample variance is sought that approximates the population variance. More specifically, a statistic is desired whereby the average of a large number of sample variances calculated on samples from the population is equal to the (true) population variance. In practice, Equation 1.3 accomplishes this task. There are two explanations for this: (1) since the standard deviation utilizes the sample mean, the result has "lost" one degree of freedom; that is, there are $n - 1$ independent observations remaining to calculate the variance; (2) in calculating the standard deviation of a small sample, there is a tendency for the resultant standard deviation to be underestimated; for small samples this is accounted for by using $n - 1$ in the denominator (note that with increasing n the correction is less of a factor since, as the central limit theorem suggests, larger samples better approximate the population they were drawn from).

A mathematical theorem, widely attributed to Chebyshev, establishes a general rule by which at least $(1 - 1/k^2)$ of all observations in a sample or population will lie within k standard deviations of the mean, where k is not necessarily an integer. For the approximately bell-shaped normal distribution of observations, an empirical rule-of-thumb suggests that the following approximate percentage of measurements will fall within 1, 2, or 3 standard deviations of the mean. These intervals are given as

$$(\bar{X} - s, \bar{X} + s)$$

which contains approximately 68% of all observed values

$$(\bar{X} - 2s, \bar{X} + 2s)$$

which contains approximately 95% of all observed values, and

$$(\bar{X} - 3s, \bar{X} + 3s)$$

which contains approximately 99% of all observed values.

The standard deviation is an absolute measure of dispersion; it does not consider the magnitude of the values in the population or sample. On some

occasions, a measure of dispersion that accounts for the magnitudes of the observations (relative measure of dispersion) is needed. The coefficient of variation (CV) is such a measure. It provides a relative measure of dispersion, where dispersion is given as a proportion of the mean. For a sample, the CV is given as

$$CV = \frac{s}{\bar{X}} \quad (1.7)$$

If, for example, on a certain highway section vehicle speeds were observed with mean $\bar{X} = 45$ mph and standard deviation $s = 15$, then the CV is $s/\bar{X} = 15/45 = 0.33$. If, on another highway section, the average vehicle speed is $\bar{X} = 60$ mph and standard deviation $s = 15$, then the CV is equal to $s/\bar{x} = 15/65 = 0.23$, suggesting that the data in the first sample have higher variability.

Example 1.1

Basic descriptive statistics are sought for observed speed data on Indiana roads, ignoring for simplicity the season, type of road, highway class, and year of observation. Most commercially available software with statistical capabilities can accommodate basic descriptive statistics. Table 1.1 provides descriptive statistics for the speed data.

The descriptive statistics indicate that the mean speed in the entire sample collected is 58.86 mph, with small variability in speed observations (s is 4.41, while the CV is 0.075). The mean and median are almost equal, indicating that the distribution of the sample of speeds is fairly symmetrical. The data set contains additional information, such as the year of observation, the season (quarter), the highway class, and whether the observation was in an urban or rural area—all of which might contribute to a more complete picture of the speed characteristics in the sample. For example, Table 1.2 examines the descriptive statistics for urban versus rural roads.

TABLE 1.1
Descriptive Statistics for Speeds on
Indiana Roads

Statistic	Value
N (number of observations)	1,296
Mean	58.86
Standard deviation	4.41
Variance	19.51
CV	0.075
Maximum	72.5
Minimum	32.6
Upper quartile	61.5
Median	58.5
Lower quartile	56.4

TABLE 1.2

Descriptive Statistics for Speeds on Rural versus Urban Indiana Roads

Statistic	Rural Roads	Urban Roads
N (number of observations)	888	408
Mean	58.79	59.0
Standard deviation	4.60	3.98
Variance	21.19	15.87
CV	0.078	0.067
Maximum	72.5	68.2
Minimum	32.6	44.2
Upper quartile	60.7	62.2
Median	58.2	59.2
Lower quartile	56.4	56.15

Interestingly, although some of the descriptive statistics may seem to differ from the pooled sample examined in Table 1.1, it does not appear that the differences between mean speeds and speed variation in urban versus rural Indiana roads is important. Similar types of descriptive statistics could be computed for other categorizations of average vehicle speed.

1.4 Skewness and Kurtosis

Two useful characteristics of a frequency distribution are skewness and kurtosis. Skewness is a measure of the degree of asymmetry of a frequency distribution, and is often called the third moment around the mean or third central moment, with variance being the second moment. In general, when a probability distribution tail is larger on the right than it is on the left, it is said that the distribution is right skewed, or positively skewed. Similarly, a left-skewed (negatively skewed) distribution is one whose tail stretches asymmetrically to the left (Figure 1.2). When a distribution is right skewed, the mean is to the right of the median, which in turn is to the right of the mode. The opposite is true for left-skewed distributions. The quantity $(x_i - \mu)^3$ is made independent of the units of measurement x by dividing by σ^3 , resulting in the population skewness parameter γ_1 ; the sample estimate of this parameter, (g_1) , is calculated as

$$g_1 = \frac{m_3}{(m_2 \sqrt{m_2})} \quad (1.8)$$

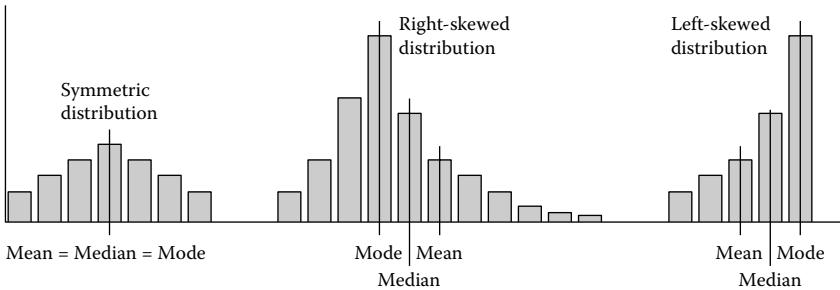


FIGURE 1.2
Skewness of a distribution.

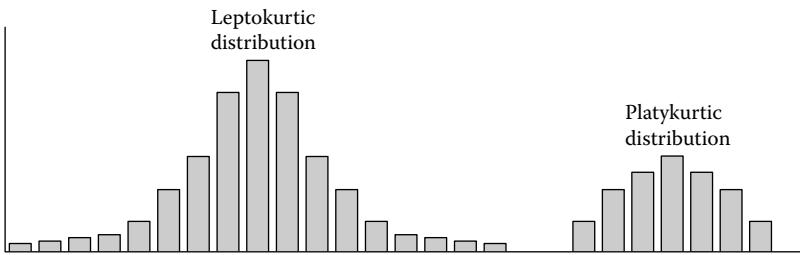


FIGURE 1.3
Kurtosis of a distribution.

where

$$m_3 = \frac{\sum_{i=1}^n (x_i - \bar{X})^3}{n}$$

$$m_2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

If a sample comes from a population that is normally distributed, then the parameter g_1 is normally distributed with mean 0 and standard deviation $\sqrt{6/n}$.

Kurtosis is a measure of the “flatness” (vs. peakedness) of a frequency distribution and is shown in Figure 1.3. The sample-based estimate is the average of $(x_i - \bar{x})^4$ divided by s^4 over the entire sample. Kurtosis (γ_2) is often called the fourth moment around the mean or fourth central moment. For the normal distribution the parameter γ_2 has a value of 3. If the parameter is larger than 3 there is usually a clustering of points around the mean (leptokurtic distribution), whereas a parameter less than 3 represents a “flatter” peak than the normal distribution (platykurtic).

The sample kurtosis parameter g_2 is often reported as standard output of many statistical software packages and is given as

$$g_2 = \gamma_2 - 3 = \frac{m_4}{m_2^2} - 3 \quad (1.9)$$

where

$$m_4 = \frac{\sum_{i=1}^n (x_i - \bar{X})^4}{n}$$

For most practical purposes, a value of 3 is subtracted from the sample kurtosis parameter so that leptokurtic sample distributions have positive kurtosis and platykurtic sample distributions have negative kurtosis.

Example 1.2

Revisiting the speed data from Example 1.1, there is interest in determining the shape of the distributions for speeds on rural and urban Indiana roads. Results indicate that when all roads are examined together their skewness is -0.05 , whereas for rural roads skewness is 0.056 and for urban roads it is -0.37 . It appears that, at least on rural roads, the distribution of speeds is symmetric, whereas for urban roads the distribution is left skewed.

Although skewness is similar for the two types of roads, kurtosis varies more widely. For rural roads the parameter has a value of 2.51 , indicating a nearly normal distribution, whereas for rural urban roads the parameter is 0.26 , indicating a relatively flat (platykurtic) distribution.

1.5 Measures of Association

So far the focus has been on statistical measures that are useful for quantifying properties of a single variable or measurement. The mean and the standard deviation, for example, convey useful information regarding the nature of the measurements related to a variable in isolation. There are, of course, statistical measures that provide useful information regarding possible relationships between variables. The correlation between two random variables is a measure of the linear relationship between them. The population linear correlation parameter ρ is a commonly used measure of how well two variables are linearly related.

The correlation parameter lies within the interval $[-1, 1]$. The value $\rho = 0$ indicates that a linear relationship does not exist between two variables. It is possible, however, for two variables with $\rho = 0$ to be nonlinearly related. When $\rho > 0$ there is a positive linear relationship between two variables,

such that when one of the variables increases in value the other variable also increases, at a rate given by the value of ρ (Figure 1.4). When $\rho = 1$ there is a “perfect” positively sloped straight-line relationship between two variables. When $\rho < 0$ there is a negative linear relationship between the two variables examined, such that an increase in the value of one variable is associated with a decrease in the value of the other, with rate of decrease ρ . Finally, when $\rho = -1$ there is a proportional negative straight-line relationship between two variables.

Correlation stems directly from another measure of association, the covariance. Consider two random variables, X and Y , both normally distributed with population means μ_X and μ_Y , and population standard deviations σ_X and σ_Y , respectively. The population and sample covariances between X and Y are defined, respectively, as follows:

$$COV_p(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{N} \quad (1.10)$$

$$COV_s(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1} \quad (1.11)$$

As Equations 1.10 and 1.11 show, the covariance of X and Y is the expected value of the product of the deviation of X and Y from their means. The covariance is positive when two variables increase together, is negative when two variables move in opposite directions, and it is zero when two variables are not linearly related.

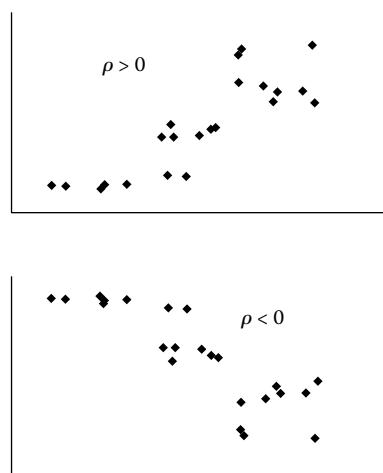


FIGURE 1.4

Positive (top) and negative (bottom) correlations between two variables.

As a measure of association, the covariance suffers from a major drawback. It is usually difficult to interpret the degree of linear association between two variables using the covariance because its magnitude depends on the magnitudes of the standard deviations of X and Y and thus is not standardized. For example, suppose that the covariance between two variables is 175: What does this say regarding the relationship between two variables? The sign, which is positive, indicates that as one increases, the other also generally increases—but the degree of correlation is hard to discern. To remedy this lack of standardization, the covariance is divided by the standard deviations to obtain a measure that is constrained to the range of values $[-1, 1]$. This measure, called the Pearson product-moment correlation parameter or correlation parameter, for short, conveys standardized information about the strength of the linear relationship between two variables. The population ρ and sample r correlation parameter of X and Y are defined, respectively, as

$$\rho = \frac{COV(X, Y)}{\sigma_X \sigma_Y} \quad (1.12)$$

$$r = \frac{COV(X, Y)}{s_X s_Y} \quad (1.13)$$

where s_X and s_Y are the sample standard deviations.

Example 1.3

Using the aviation data, the correlations between annual U.S. revenue passenger enplanements, per-capita U.S. gross domestic product (GDP), and price per gallon for aviation fuel are examined. After deflating the monetary values by the consumer price index (CPI) to 1977 values, the correlation between enplanements and per-capita GDP is 0.94, and the correlation between enplanements and fuel price –0.72.

These two correlation parameters are not surprising. One expects enplanements and economic growth to go hand in hand, while enplanements and aviation fuel price (often reflected by changes in fare price) are negatively correlated. The existence of a correlation between two variables, however, does not suggest that changes in one of the variables causes changes in value of the other. The determination of causality is a difficult question that cannot be determined by inspection of correlation parameters. To this end, consider the correlation parameter between annual U.S. revenue passenger enplanements and annual ridership of the Tacoma-Pierce Transit System in Washington State. The correlation parameter is considerably high (~0.90) indicating that the two variables move in opposite directions in nearly straight-line fashion. Nevertheless, it is safe to say that (1) neither of the variables will cause changes in value of the other and (2) the two variables are not directly related. In short, correlation does not imply causation.

The discussion on correlation has thus far focused solely on continuous variables measured on the interval or ratio scales. In some situations, however, one or both of the variables may be measured on the ordinal scale. Alternatively, two continuous variables may not satisfy the requirement of approximate normality assumed when using the Pearson product-moment correlation parameter. In such cases the Spearman rank correlation parameter, an alternative nonparametric method should be used to determine whether a linear relationship exists between two variables.

The Spearman correlation parameter is computed first by ranking the observations of each variable from smallest to largest. Then, the Pearson correlation parameter is applied to the ranks; that is, the Spearman correlation parameter is the usual Pearson correlation parameter applied to the ranks of two variables. The equation for the Spearman rank correlation parameter is given as

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (1.14)$$

where d_i , $i = 1, \dots, n$, are the differences in the ranks of x_i, y_i ; $d_i = R(x_i) - R(y_i)$.

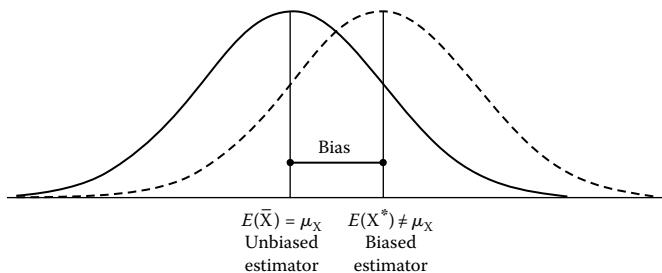
There are additional nonparametric measures of correlation between variables, including Kendall's tau. Its estimation complexity, at least when compared with Spearman's rank correlation parameter, makes it less popular in practice.

1.6 Properties of Estimators

The sample statistics computed in previous sections, such as the sample average \bar{X} , variance s^2 , and standard deviation s and others, are used as estimators of population parameters. In practice population parameters (often called parameters) such as the population mean and variance are unknown constants. In practical applications, the sample average \bar{X} is used as an estimator for the population mean μ_x , the sample variance s^2 for the population variance σ^2 , and so on. These statistics, however, are random variables and, as such, are dependent on the sample. "Good" statistical estimators of true population parameters satisfy four important properties: unbiasedness, efficiency, consistency, and sufficiency.

1.6.1 Unbiasedness

If there are several estimators of a population parameter, and if one of these estimators coincides with the true value of the unknown parameter, then this estimator is called an unbiased estimator. An estimator is said to be unbiased

**FIGURE 1.5**

Biased and unbiased estimators of the mean value of a population μ_X .

if its expected value is equal to the true population parameter it is meant to estimate. That is, an estimator, say, the sample average \bar{X} , is an unbiased estimator of μ_X if

$$E(\bar{X}) = \mu_X \quad (1.15)$$

The principle of unbiasedness is illustrated in Figure 1.5. Any systematic deviation of the estimator away from the population parameter is called a bias, and the estimator is called a biased estimator. In general, unbiased estimators are preferred to biased estimators. Practically, a statistic is said to be an unbiased estimate of a population parameter if the statistic tends to give values that are neither consistently high nor consistently low; they may not be "exactly" correct, because after all they are only an estimate, but they have no systematic source of bias. For example, saying that the sample mean is an unbiased estimate of the population mean implies that there is no distortion that will systematically overestimate or underestimate the population mean.

1.6.2 Efficiency

The property of unbiasedness is not, by itself, adequate, because there are situations in which two or more parameter estimates are unbiased. In these situations, interest is focused on which of several unbiased estimators is superior. A second desirable property of estimators is efficiency. Efficiency is a relative property in that an estimator is efficient relative to another, which means that an estimator has a smaller variance than an alternative estimator. An estimator with the smaller variance is more efficient. As is seen in Figure 1.6, both X_1 and \bar{X} are unbiased estimators of μ_X ; however, $VAR(\bar{X}) = \sigma^2/n$, while $VAR(X_1) = \sigma^2$, yielding a relative efficiency of \bar{X} relative to X_1 of $1/n$, where n is the sample size.

In general, the unbiased estimator with minimum variance is preferred to alternative estimators. A lower bound for the variance of any unbiased estimator $\hat{\theta}$ of θ is given by the Cramer–Rao lower bound and is written as (Gujarati 1992).

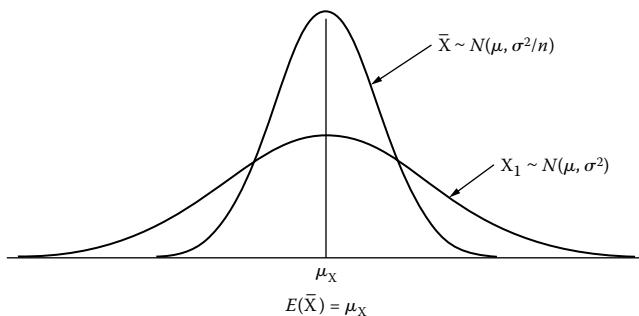


FIGURE 1.6
Comparing efficiencies.

$$VAR(\hat{\theta}) \geq \frac{1}{[nE(\partial LNf(X; \theta))/\partial \theta]^2} = \frac{-1}{[nE(\partial^2 LNf(X; \theta))]/\partial \theta^2} \quad (1.16)$$

The Cramer–Rao lower bound is only a sufficient condition for efficiency. Failing to satisfy this condition does not necessarily imply that the estimator is not efficient. Finally, unbiasedness and efficiency hold true for any finite sample n , and when $n \rightarrow \infty$ they become asymptotic properties.

1.6.3 Consistency

A third asymptotic property is that of consistency. An estimator $\hat{\theta}$ is said to be consistent if the probability of being closer to the true value of the parameter it estimates (θ) increases with increasing sample size. Formally, this says that as $n \rightarrow \infty$ $\lim P[|\hat{\theta} - \theta| > c] = 0$, for any arbitrary constant c . For example, this property indicates that \bar{X} will not differ from μ as $n \rightarrow \infty$. Figure 1.7 graphically depicts the property of consistency showing the behavior of an estimator X^* of the population mean μ with increasing sample size.

It is important to note that a statistical estimator may not be an unbiased estimator; however, it may be a consistent one. In addition, a sufficient condition for an estimator to be consistent is that it is asymptotically unbiased and that its variance tends to zero as $n \rightarrow \infty$ (Hogg and Craig 1992).

1.6.4 Sufficiency

An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates. In other words, \bar{X} is sufficient for μ if \bar{X} contains all the information in the sample pertaining to μ .

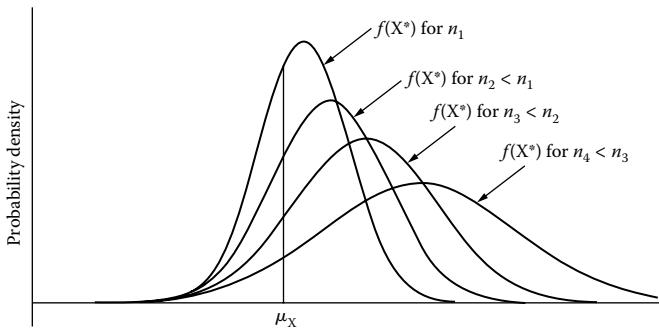


FIGURE 1.7
The property of consistency.

1.7 Methods of Displaying Data

Although the different measures described in the previous sections often provide much of the information necessary to describe the nature of the data set being examined, it is often useful to utilize graphical techniques for examining data. These techniques provide ways of inspecting data to determine relationships and trends, identify outliers and influential observations, and quickly describe or summarize data sets. Pioneering methods frequently used in graphical and exploratory data analysis stem from the work of Tukey (1977).

1.7.1 Histograms

Histograms are most frequently used when data are either naturally grouped (gender is a natural grouping, for example) or when small subgroups may be defined to help uncover useful information contained in the data. A histogram is a chart consisting of bars of various heights. The height of each bar is proportional to the frequency of values in the class represented by the bar. As seen in Figure 1.8, a histogram is a convenient way of plotting the frequencies of grouped data. In the figure, frequencies on the first (left) Y axis are absolute frequencies, or counts of the number of city transit buses in the State of Indiana belonging to each age group (data were taken from Karlaftis and Sinha 1997). Data on the second Y axis are relative frequencies, which are simply the count of data points in the class (age group) divided by the total number of data points.

Histograms are useful for uncovering asymmetries in data and, as such, skewness and kurtosis are easily identified using histograms.

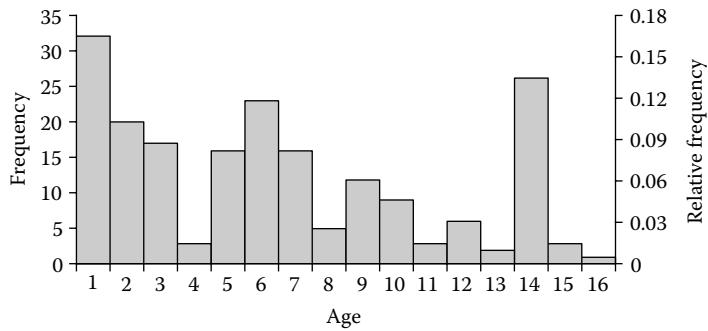


FIGURE 1.8
Histogram for bus ages in the State of Indiana (1996 data).

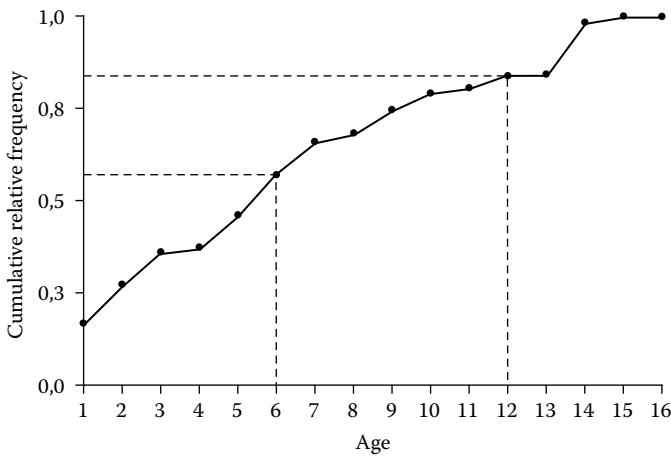


FIGURE 1.9
Ogive for bus ages in the State of Indiana.

1.7.2 Ogives

A natural extension of histograms is ogives. Ogives are the cumulative relative frequency graphs. Once an ogive such as the one shown in Figure 1.9 is constructed, the approximate proportion of observations that are less than any given value on the horizontal axis are read directly from the graph. Thus, for example, using Figure 1.9 the estimated proportion of buses that are less than 6 years old is approximately 60%, and the proportion less than 12 years old is 85%.

1.7.3 Box Plots

When faced with the problem of summarizing essential information of a data set, a box plot (or box-and-whisker plot) is a pictorial display that is extremely useful. A box plot illustrates how widely dispersed observations are and where the data are centered. This is accomplished by providing, graphically, five summary measures of the distribution of the data: the largest observation, the upper quartile, the median, the lower quartile, and the smallest observation (Figure 1.10).

Box plots are useful for identifying the central tendency of the data (through the median), identifying the spread of the data (through the IQR and the length of the whiskers), identifying possible skewness of the data (through the position of the median in the box), identifying possible outliers (points beyond the 1.5 [IQR] mark), and for comparing data sets.

1.7.4 Scatter Diagrams

Scatter diagrams are most useful for examining the relationship between two continuous variables. As examples, assume that transportation researchers are interested in the relationship between economic growth and enplane-ments, or the effect of a fare increase on travel demand. In some cases, when one variable depends (to some degree) on the value of the other variable, then the first variable, the dependent, is plotted on the vertical axis. The pattern of the scatter diagram provides information about the relationship between two variables. A linear relationship is one that is approximated well by a straight line (see Figure 1.4). A scatter plot can show a positive correlation, no correlation, and a negative correlation between two variables (Section 1.5 and Figure 1.4 analyzed this issue in greater depth). Nonlinear relationships between two variables can also be seen in a scatter diagram and typically are curvilinear. Scatter diagrams are typically used to uncover underlying relationships between variables, which can then be explored in greater depth with more quantitative statistical methods.

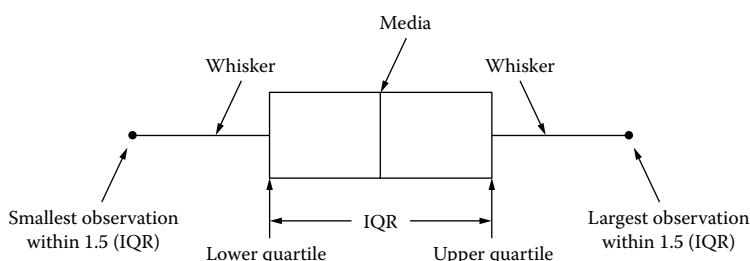


FIGURE 1.10

The box plot.

1.7.5 Bar and Line Charts

A common graphical method for examining nominal data is a pie chart. The Bureau of Economic Analysis of the U.S. Department of Commerce in its 1996 Survey of Current Business reported the percentages of the U.S. GDP accounted for by various social functions. As shown in Figure 1.11, transportation is a major component of the economy, accounting for nearly 11% of GDP in the United States. The data are nominal since the “values” of the variable, major social function, include six categories: transportation, housing, food, education, health care, and other. The pie graph illustrates the proportion of expenditures in each category of major social function.

The U.S. Federal Highway Administration (FHWA 1997) completed a report for Congress that provided information on highway and transit assets, trends in system condition, performance, and finance, and estimated investment requirements from all sources to meet the anticipated demands in both highway travel and transit ridership. One of the interesting findings of the report was the pavement ride quality of the nation’s urban highways as measured by the International Roughness Index. The data are ordinal because the “values” of the variable, pavement roughness, include five categories: very good, good, fair, mediocre, and poor. This scale, although it resembles the nominal categorization of the previous example, possesses the additional property of natural ordering between the categories (without uniform increments between the categories). A reasonable way to describe these data is to count the number of occurrences of each value and then to convert these counts into proportions (Figure 1.12).

Bar charts are a common alternative to pie charts. They graphically represent the frequency (or relative frequency) of each category as a bar rising from the horizontal axis; the height of each bar is proportional to the frequency (or relative frequency) of the corresponding category. Figure 1.13, for example, presents the motor vehicle fatal accidents by posted speed limit for

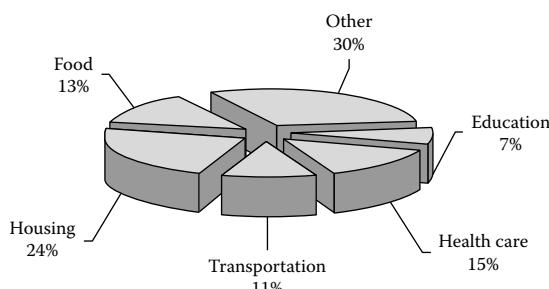
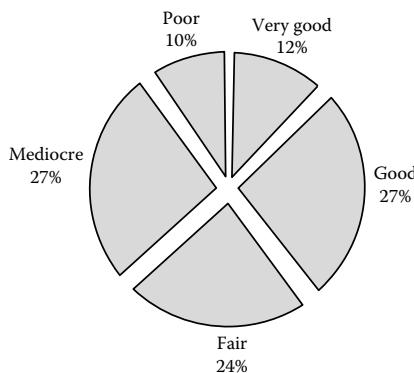
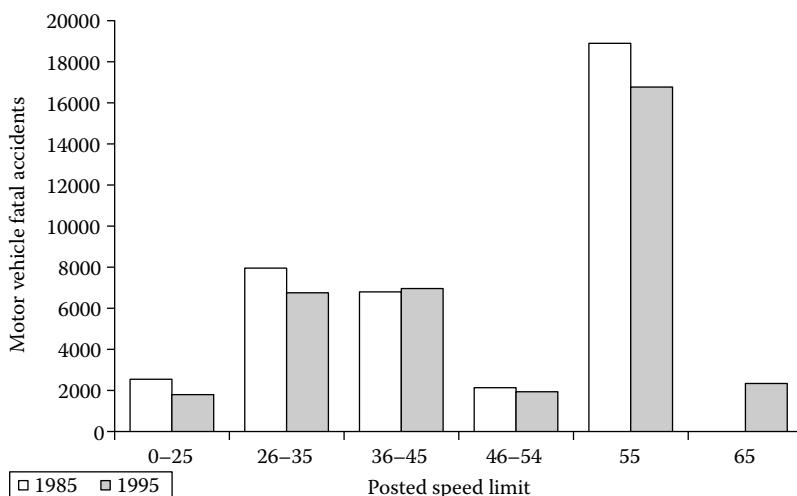


FIGURE 1.11

U.S. GDP by major social function (1995). (From U.S. DOT., *Transportation in the United States: A Review*, Bureau of Transportation Statistics, Washington, DC, 1997a.)

**FIGURE 1.12**

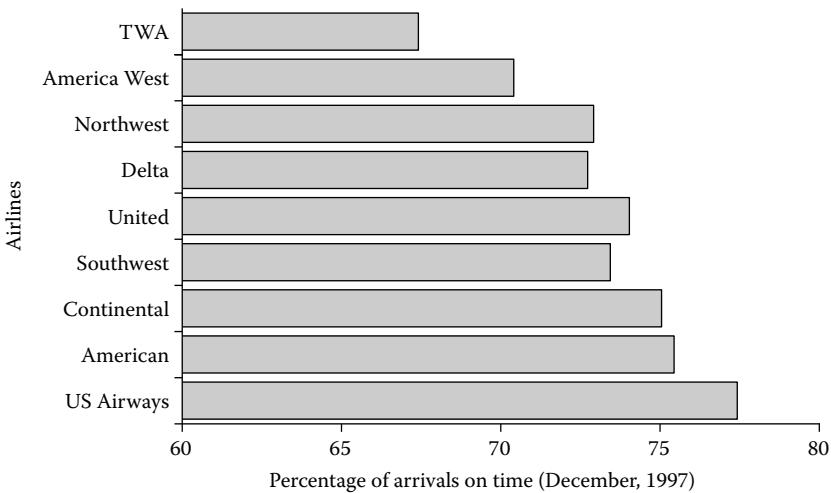
Percent miles of urban interstate by pavement roughness category. (From FHWA., *Status of the Nation's Surface Transportation System: Condition and Performance*. Federal Highway Administration, Washington, DC, 1997.)

**FIGURE 1.13**

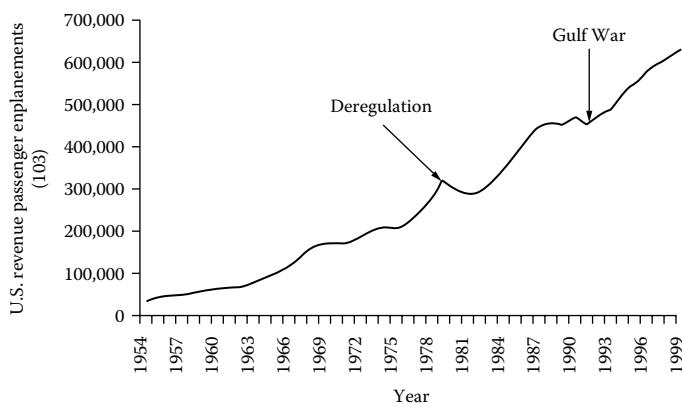
Motor vehicle fatal accidents by posted speed limit. (From U.S. DOT., *Transportation in the United States: A Review*, Bureau of Transportation Statistics, Washington, DC, 1997b.)

1985 and 1995, and Figure 1.14 presents the percent of on-time arrivals for some U.S. airlines for December, 1997.

The final graphical technique considered in this section is the line chart. A line chart is obtained by plotting the frequency of a category above the point on the horizontal axis representing that category and then joining the

**FIGURE 1.14**

Percent of on-time arrivals for December, 1997. (From Bureau of Transportation Statistics, www.bts.gov)

**FIGURE 1.15**

U.S. revenue passenger enplanements 1954 through 1999. (From Bureau of Transportation Statistics, www.bts.gov)

points with straight lines. A line chart is most often used when the categories are points in time (time-series data). Line charts are excellent for uncovering trends of a variable over time. For example, consider Figure 1.15, which represents the evolution of the U.S. air-travel market. A line chart is useful for showing the growth in the market over time. Two points of particular interest to the air-travel market, the deregulation of the market and the Gulf War, are marked on the graph.

2

Statistical Inference II: Interval Estimation, Hypothesis Testing and Population Comparisons

Scientific decisions should be based on sound analyses and accurate information. This chapter provides the theory and interpretation of confidence intervals (CIs), hypothesis tests, and population comparisons, which are statistical constructs (tools) used to ask and answer questions about the transportation phenomena under study. Despite their enormous utility, CIs are often ignored in transportation practice and hypothesis tests and population comparisons are frequently misused and misinterpreted. The techniques discussed in this chapter are used to formulate, test, and make informed decisions regarding a large number of hypotheses. Questions such as the following serve as examples. Does crash occurrence at a particular intersection support the notion that it is a hazardous location? Do traffic-calming measures reduce traffic speeds? Does route guidance information implemented via a variable message sign system successfully divert motorists from congested areas? Did the deregulation of the air-transport market increase the market share for business travel? Does altering the levels of operating subsidies to transit systems change their operating performance? To address these, and similar types of questions, transportation researchers and professionals can apply the techniques presented in this chapter.

2.1 Confidence Intervals

In practice, statistics calculated from samples such as the (sample) average \bar{X} , variance s^2 , standard deviation s , as well as others reviewed in the previous chapter are used to estimate population parameters. For example, the sample average \bar{X} is used as an estimator for the population mean μ_x , the sample variance s^2 is an estimate of the population variance σ^2 , and so on. Recall from Section 1.6 that desirable or “good” estimators satisfy four important properties: unbiasedness, efficiency, consistency, and sufficiency. However, regardless of the properties an estimator satisfies, estimates vary across samples and there is at least some probability that an estimated parameter will differ from the population parameter it is intended to estimate. Unlike the point

estimators reviewed in the previous chapter, the focus here is on interval estimates. Interval estimates allow inferences to be drawn about a population by providing an interval, a lower and upper boundary within which an unknown parameter will lie with a prespecified level of confidence. The logic behind an interval estimate is that an interval calculated using sample data contains the true population parameter with some level of confidence (the long-run proportion of times that the true population parameter interval is contained in the interval). Intervals are called CIs and are constructed across an array of levels of confidence. The lower value is called the lower confidence limit (LCL) and the upper value the upper confidence limit (UCL). The wider a CI, the more confident the researcher is that it contains the population parameter (overall confidence is relatively high). In contrast, a relatively narrow CI is less likely to contain the population parameter (overall confidence is relatively low).

All the parametric methods presented in the first four sections of this chapter make specific assumptions about the probability distributions of sample estimators, or make assumptions about the nature of the sampled populations. In particular, the assumption of an approximately normally distributed population (and sample) is usually made. As such, it is imperative that these assumptions, or requirements, be checked before applying the methods. When the assumptions are not met, then the nonparametric statistical methods—provided in Section 2.5—are more appropriate.

2.1.1 Confidence Interval for μ with Known σ^2

The central limit theorem (CLT) suggests that whenever a sufficiently large random sample is drawn from any population with mean μ and standard deviation σ , the sample mean \bar{X} is approximately normally distributed with mean μ and standard deviation σ/\sqrt{n} . It can easily be verified that this standard normal random variable Z has a 0.95 probability of being between the range of values [-1.96, 1.96] (see Table C.1 in Appendix C). A probability statement regarding Z is given as

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \quad (2.1)$$

With some basic algebraic manipulation the probability statement of Equation 2.1 is written in a different, yet equivalent form

$$\begin{aligned} 0.95 &= P\left(\frac{-1.96\sigma}{\sqrt{n}} < \bar{X} - \mu < \frac{1.96\sigma}{\sqrt{n}}\right) \\ &= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \end{aligned} \quad (2.2)$$

Equation 2.2 suggests that, with a large number of intervals computed from different random samples drawn from the population, the proportion of values of \bar{X} for which the interval $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ captures μ is 0.95. This interval is called the 95% CI estimator of μ . A shortcut notation for this interval is

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad (2.3)$$

Obviously, probabilities other than 95% are often used. For example, a 90% CI is

$$\bar{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

In general, any confidence level can be used in estimating CIs. The CI is $(1 - \alpha)$ and $Z_{\alpha/2}$ is the value of Z such that the area in each of the tails under the standard normal curve is $(\alpha/2)$. Using this notation, the CI estimator of μ is written as

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (2.4)$$

Because the confidence level is inversely proportional to the risk that the CI fails to include the actual value of μ , it generally ranges between 0.90 and 0.99, reflecting 10% and 1% levels of risk of not including the true population parameter, respectively.

Example 2.1

A 95% CI is desired for the mean vehicular speed on Indiana roads (see Example 1.1 for more details). First, the assumption of normality is checked; if this assumption is satisfied one can proceed with the analysis. The sample size is $n = 1,296$, and the sample mean is $\bar{X} = 58.86$. Suppose a long history of prior studies has shown the population standard deviation to be $\sigma = 5.5$. Using Equation 2.4, the CI is obtained

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 58.86 \pm 1.96 \frac{5.5}{\sqrt{1296}} = 58.86 \pm 0.30 = [58.56, 59.16]$$

The result indicates that the 95% CI for the unknown population parameter μ consists of lower and upper bounds of 58.56 and 59.16. This suggests that the true and unknown population parameter would lie somewhere in this interval about 95 times out of 100, on average. The CI is rather “tight,” meaning that the range

of possible values is relatively small. This is a result of the low assumed standard deviation (or variability in the data) of the population examined.

The 90% CI, using the same standard deviation, is [58.60, 59.11], and the 99% CI is [58.46, 59.25]. As the CI becomes wider, there is increasingly higher confidence that the interval contains the true unknown population parameter.

2.1.2 Confidence Interval for the Mean with Unknown Variance

In the previous section, a procedure was discussed for constructing CIs around the mean of a normal population when the variance of the population is known. In the majority of practical sampling situations, however, the population variance is rarely known and is instead estimated from the data. When the population variance is unknown and the population is normally distributed, a $(1 - \alpha)100\%$ CI for μ is given by

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad (2.5)$$

where s is the square root of the estimated variance (s^2), $t_{\alpha/2}$ is the value of the t distribution with $n - 1$ degrees of freedom (for a discussion of the t distribution, see Appendix A).

Example 2.2

Continuing with the previous example, a 95% CI for the mean speed on Indiana roads is computed, assuming that the population variance is not known and instead an estimate is obtained from the data with the same value as before. The sample size is $n = 1,296$, and the sample mean is $\bar{X} = 58.86$. Using Equation 2.3, the CI is obtained as

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 58.86 \pm 1.96 \frac{4.41}{\sqrt{1296}} = [58.61, 59.10]$$

Interestingly, inspection of probabilities associated with the t distribution (see Table C.2 in Appendix C) shows that the t distribution converges to the standard normal distribution as $n \rightarrow \infty$. Although the t distribution is the correct distribution to use whenever the population variance is unknown, when sample size is sufficiently large the standard normal distribution is used as an adequate approximation to the t distribution.

2.1.3 Confidence Interval for a Population Proportion

Sometimes, interest centers on a qualitative (nominal scale) variable rather than a quantitative (interval or ratio scale) variable. There may be interest in the relative frequency of some characteristic in a population such as, for

example, the proportion of people in a population who are transit users. In such cases, an estimate of the population proportion p whose estimator is \hat{p} has an approximate normal distribution provided that n is sufficiently large ($np \geq 5$ and $nq \geq 5$, where $q = 1 - p$). The mean of the sampling distribution \hat{p} is the population proportion p and the standard deviation is $\sqrt{pq/n}$.

A large sample $(1-\alpha)100\%$ CI for the population proportion, p is given by

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} \quad (2.6)$$

where the estimated sample proportion \hat{p} is equal to the number of “successes” in the sample divided by the sample size n and $\hat{q} = 1 - \hat{p}$.

Example 2.3

A transit planning agency wants to estimate, at a 95% confidence level, the share of transit users in the daily commute “market” (i.e., the percentage of commuters using transit). A random sample of 100 commuters is obtained and it is found that 28 people in the sample are transit users. By using Equation 2.6, a 95% CI for p is calculated as

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} = 0.28 \pm 1.96 \sqrt{\frac{(0.28)(0.72)}{100}} = 0.28 \pm 0.088 = [0.192, 0.368]$$

Thus, the agency is 95% confident that transit’s share in the daily commute ranges from 19.2% to 36.8%.

2.1.4 Confidence Interval for the Population Variance

In many situations, in traffic safety research, for example, interest centers on the population variance (or a related measure such as the population standard deviation). As an example, vehicle speeds contribute to crash probability, but speed variability on the roadway may be even more important. Speed variance, measured as differences in travel speeds on a roadway, relates to crash frequency in that a larger variance in speed between vehicles correlates with a larger frequency of crashes, especially for crashes involving two or more vehicles (Garber 1991). Large differences in speeds result in an increase in the frequency with which motorists pass one another, increasing the number of opportunities for multivehicle crashes. Clearly, vehicles traveling the same speed in the same direction do not overtake one another; therefore, they cannot collide as long as the same speed is maintained (for additional literature on the topic of speeding and crash probabilities, covering both the United States and abroad, the interested reader should consult FHWA 1997, 1998).

A $(1 - \alpha)100\%$ CI for σ^2 , assuming the population is normally distributed, is given by

$$\left[\frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2} \right] \quad (2.7)$$

where $\chi_{\alpha/2}^2$ is the value of the χ^2 distribution with $n - 1$ degrees of freedom. The area in the right-hand tail of the distribution is $\chi_{\alpha/2}^2$, while the area in the left-hand tail of the distribution is $\chi_{1-\alpha/2}^2$. The chi-square distribution is described in Appendix A, and the table of probabilities associated with the chi-square distribution is provided in Table C.3 of Appendix C.

Example 2.4

A 95% CI for the variance of speeds on Indiana roads is desired. With a sample size of 100 and a variance of 19.51 mph^2 , and using the values from the χ^2 table (Appendix C, Table C.3), one obtains $\chi_{\alpha/2}^2 = 129.56$ and $\chi_{1-\alpha/2}^2 = 74.22$. Thus, the 95% CI is given as

$$\left[\frac{99(19.51)}{129.56}, \frac{99(19.51)}{74.22} \right] = [15.05, 26.02]$$

The speed variance is, with 95% confidence, between 15.05 and 26.02. Again, the units of the variance in speed are in mph^2 .

2.2 Hypothesis Testing

Hypothesis tests are used to assess the evidence on whether a difference in a population parameter (a mean, variance, proportion, etc.) between two or more groups is likely to have arisen by chance or whether some other factor is responsible for the difference. Statistical distributions are employed in hypothesis testing to estimate probabilities of observing the sample data, given an assumption about what “should have” occurred. When observed results are extremely unlikely to have occurred under assumed conditions, then the assumed conditions are considered unlikely. In statistical terms, the hypothesis test provides the following probability:

$$P(\text{data} | \text{true null hypothesis}) \quad (2.8)$$

which is the probability of obtaining or observing the sample data conditional upon a true null hypothesis. It is not, however, the probability of the null hypothesis being true, despite the number of times it is mistakenly interpreted this way. This and other common misinterpretations of hypothesis tests are described in Washington (2000a).

Consider an example. Suppose there is interest in quantifying the impact of the repeal of the national maximum speed limit (NMSL) on average speeds on U.S. roads. Representative speed data are collected in time periods before and after the repeal. Using a simple before–after study design, the mean speeds before and after repeal of the NMSL are compared.

The data are used to assess whether the observed difference in mean speeds between the periods is explained by the natural sampling variability, or whether the observed difference is attributable to the repeal of the NMSL. Using hypothesis testing, the researcher calculates the probability of observing the increase (or decrease) in mean speeds (as reflected in the sample data) from the period before to the period after repeal, under the assumption that the repeal of the NMSL had no effect. If an observed increase in speed is extremely unlikely to have been produced by random sampling variability, then the researcher concludes that the repeal was instrumental in bringing about the change. Conversely, if the observed speed differences are not all that unusual and can rather easily be explained by random sampling variability, then it is difficult to attribute the observed differences to the effect of the repeal. At the conclusion of the hypothesis test the probability of observing the actual data is obtained, given that the repeal of the NMSL had no effect on speeds.

Because many traffic investigations are observational types of studies, it is not easy to control the many other factors that could influence speeds in the after period (besides the repeal). Such factors include changes in vehicle miles traveled (VMT), driver population, roadside changes, adjacent land-use changes, weather, and so on. It is imperative to account for, or control, these other factors to the extent possible, because lack of control could result in attributing a change in speeds to the repeal of the NMSL, when other factors were responsible for the change.

2.2.1 Mechanics of Hypothesis Testing

To formulate questions about transportation phenomena a researcher must pose two competing statistical hypotheses: a null hypothesis (the hypothesis to be nullified) and an alternative. The null hypothesis, typically denoted by H_0 , is an assertion about one or more population parameters that are assumed to be true until there is sufficient statistical evidence to conclude otherwise. The alternative hypothesis, typically denoted by H_a , is the assertion of all situations not covered by the null hypothesis. Together the null and the alternative constitute a set of hypotheses that covers all possible values of the parameter or parameters in question. Considering the NMSL repeal

previously discussed, the following pair of competing hypotheses could be formulated:

Null Hypothesis (H_0): There has not been a change in mean speeds as a result of the repeal of the NMSL

Alternative Hypothesis (H_a): There has been a change in mean speeds as a result of the repeal of the NMSL.

The purpose of a hypothesis test is to determine whether it is appropriate to reject or not to reject the null hypothesis. The test statistic is the sample statistic upon which the decision to reject, or fail to reject, the null hypothesis is based. The nature of the hypothesis test is determined by the question being asked. For example, if an engineering intervention is expected to change the mean of a sample (the mean of vehicle speeds), then a null hypothesis of no difference in means is appropriate. If an intervention is expected to change the spread or variability of data, then a null hypothesis of no difference in variances should be used. Many different types of hypothesis tests can be conducted. Regardless of the type of hypothesis test, the process is the same: the empirical evidence is assessed and will either refute or fail to refute the null hypothesis based on a prespecified level of confidence. Test statistics used in many parametric hypothesis testing applications rely upon the Z , t , F , and χ^2 distributions.

The decision to reject or fail to reject the null hypothesis may or may not be based on the rejection region, which is the range of values such that, if the test statistic falls into the range, the null hypothesis is rejected. Recall that, upon calculation of a test statistic, there is evidence either to reject or to fail to reject the null hypothesis. The phrases *reject* and *fail to reject* have been chosen carefully. When a null hypothesis is rejected, the information in the sample does not support the null hypothesis and it is concluded that it is unlikely to be true, a definitive statement. On the other hand, when a null hypothesis is not rejected, the sample evidence is consistent with the null hypothesis. This does not mean that the null hypothesis is true; it simply means that it cannot be ruled out using the observed data. It can never be proved that a statistical hypothesis is true using the results of a statistical test. In the language of hypothesis testing, any particular result is evidence as to the degree of certainty, ranging from almost uncertain to almost certain. No matter how close to the two extremes a statistical result may be, there is always a nonzero probability to the contrary.

Whenever a decision is based on the result of a hypothesis test, there is a chance that it is incorrect. Consider Table 2.1. In this classical Neyman–Pearson methodology, the sample space is partitioned into two regions. If the observed data reflected through the test statistic falls into the rejection or critical region, the null hypothesis is rejected. If the test statistic falls into the acceptance region, the null hypothesis cannot be rejected. When the null hypothesis is true, there is α percent chance of rejecting it (Type I error). When the null hypothesis is false, there is still a β percent chance of accepting

TABLE 2.1

Results of a Test of Hypothesis

Test Result	Reality	
	H_0 Is True	H_0 Is False
Reject	Type I error $P(\text{Type I error}) = \alpha$	Correct decision
Do not reject	Correct decision	Type II error $P(\text{Type II error}) = \beta$

it (Type II error). The probability of Type I error is the size of the test. It is conventionally denoted by α and called the significance level. The power of a test is the probability that it will correctly lead to rejection of a false null hypothesis, and is given as $1 - \beta$.

Because both probabilities α and β reflect probabilities of making errors, they should be kept as small as possible. There is, however, a trade off between the two. For several reasons, the probability of making a Type II error is often ignored. Also, the smaller the α , the larger the β . Thus, if α is made to be really small, the “cost” is a higher probability for making a Type II error, all else being equal. The determination of which statistical error is least desirable depends on the research question asked and the subsequent consequences of making the respective errors. Both error types are undesirable, so attention to proper experimental design before data collection and sufficiently large sample sizes will help to minimize the probability of making these two statistical errors. In practice, the probability of making a Type I error α is usually set in the range from 0.01 to 0.10 (1% and 10% error rates, respectively). The selection of an appropriate α level is based on the consequences of making a Type I error. For example, if human lives are at stake when an error is made (accident investigations, medical studies), then an α of 0.01 or 0.005 may be most appropriate. In contrast, if an error results in monies being spent for improvements (congestion relief, travel time, etc.) that might not bring about improvements, then perhaps a less stringent α is appropriate.

2.2.2 Formulating One- and Two-Tailed Hypothesis Tests

As discussed previously, the decision of whether the null hypothesis is rejected (or not) is based on the rejection region. To illustrate a two-tailed rejection region, suppose a hypothesis test is conducted to determine whether the mean speed on U.S. highways is 60 mph. The null and alternative hypotheses are formulated as follows:

$$H_0: \mu = 60$$

$$H_a: \mu \neq 60$$

If the sample mean (the test statistic in this case) is significantly different from 60, and \bar{X} falls in the rejection region, the null hypothesis is rejected. On the other hand, if \bar{X} is sufficiently close to 60, the null hypothesis cannot be rejected. The rejection region provides a range of values below or above which the null hypothesis is rejected. In practice, however, a standardized normal test statistic is employed. A standardized normal variable is constructed (Equation 2.1) based on a true null hypothesis such that

$$Z^* = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (2.9)$$

The random variable is approximately standard normally distributed ($N(0, 1)$) under a true null hypothesis. Critical values of Z , or Z_c , are defined such that $P[Z^* \geq Z_c] = P[Z^* \leq -Z_c] = \alpha/2$. The values of Z_c that correspond to different values of α are provided in Table C.1 in Appendix C; some commonly used values are shown in Table 2.2. For example, if $\alpha = 0.05$ then $Z_c = 1.96$. Using Equation 2.9 the test statistic Z^* is calculated; for this statistic the following rules apply:

1. If $|Z^*| \geq Z_c$, then the probability of observing this value (or larger) if the null hypothesis is true is α . In this case the null hypothesis is rejected in favor of the alternative.
2. If $|Z^*| < Z_c$, then the probability of observing this value (or smaller) is $1 - \alpha$. In this case the null hypothesis cannot be rejected.

The conclusions corresponding with the null and alternate hypotheses are “the population mean is equal to 60 mph” and “the population mean is not equal to 60 mph,” respectively. This example illustrates a two-sided hypothesis test.

There are many situations when, instead of testing whether a population parameter is equal to a specific value, interest may center on testing whether a parameter is greater than (or smaller than) a specific value. For example, consider a hypothesis test of whether a decrease in domestic airfare ticket prices for a certain airline have a positive impact on the market share p of the airline, currently at 7%. The null and alternative hypotheses are stated as

$$H_0: p \leq .07$$

$$H_a: p > .07$$

In these hypotheses there is an a priori belief that the direction of the anticipated effect is an increase in market share (and the hypotheses are constructed to test this specific directional effect). There could have also been

TABLE 2.2

Critical Points of Z for Selected Levels of Significance

	Level of Significance α		
	0.10	0.05	0.01
One-tailed test	± 1.28	± 1.645	± 2.326
Two-tailed test	± 1.645	± 1.96	± 2.576

an a priori belief of a decrease in market share, which would yield the following hypotheses:

$$H_0: p \geq .07$$

$$H_a: p < .07$$

Finally, nondirectional hypotheses, motivated by lack of a priori belief regarding the direction of the effect, would yield

$$H_0: p = .07$$

$$H_a: p \neq .07$$

One- and two-tailed tests have different critical points on the standard normal distribution since α is allocated to a single tail for directional tests, whereas it is divided between two tails for nondirectional tests. Table 2.2 provides critical points for the Z distribution for one-tailed and two-tailed hypothesis tests for commonly used values of α . Critical values for other levels of α are found in Table C.1.

Example 2.5

Using the data from Example 2.1, a test is conducted to assess whether the mean speed on Indiana roads is 59.5 mph at the 5% significance level. The sample size is $n = 1,296$, and the sample mean is $\mu = 58.86$. Suppose that numerous past studies have revealed the population standard deviation to be $\sigma = 5.5$. The parameter of interest is the population mean, and the hypotheses to be tested are

$$H_0: \mu = 59.5$$

$$H_a: \mu \neq 59.5$$

From Example 2.1, a 95% CI for mean speeds is [58.56, 59.16]. Because the value 59.5 mph is not within the CI, the null hypothesis is rejected. Thus, there is sufficient evidence to infer that mean speed is not equal to 59.5 mph.

An alternative method for obtaining this result is to use the standardized test statistic presented in Equation 2.9 and follow the appropriate decision rules. The test statistic is

$$Z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{58.86 - 59.5}{5.5/\sqrt{1296}} = -3.27$$

Since the test statistic $|-3.27| = 3.27$ is greater than 1.96, the critical value for a two-tailed test at the 5% level of significance, the null hypothesis is rejected. As expected, a CI and the standardized test statistic lead to identical conclusions.

2.2.3 The p -Value of a Hypothesis Test

An increasingly common practice in reporting the outcome of a statistical test is to state the value of the test statistic along with its “probability-value” or “ p -value.” The p -value is the smallest level of significance α that leads to rejection of the null hypothesis. It is an important value because it quantifies the amount of statistical evidence that supports the alternative hypothesis. In general, the more evidence that exists to reject the null hypothesis in favor of the alternative hypothesis, the larger the test statistic and the smaller is the p -value. The p -value provides a convenient way to determine the outcome of a statistical test based on any specified Type I error rate α ; if the p -value is less than or equal to α , then the null hypothesis is rejected. For example, a p -value of .031 suggests that a null hypothesis is rejected at $\alpha = 0.05$, but is not rejected at $\alpha = 0.01$. Using p -values always leads to the same conclusions as the usual test procedure given a level of significance α .

The p -value of $Z^* = 3.27$ is calculated as follows:

$$\begin{aligned} p\text{-value } (Z^* = 3.27) &= p[Z \leq -3.27 \text{ and } Z \geq 3.27] \\ &= 2p[Z \geq 3.27] = 2[1 - p[Z \leq 3.27]] \\ &= 2[1 - .99946] = .001 \end{aligned}$$

As a result, the null hypothesis $H_0: \mu = 59.5$ is rejected at the 0.05 and 0.01 levels, because the p -value of the test is .001.

2.3 Inferences Regarding a Single Population

The previous section discussed some of the essential concepts regarding statistical inference and hypothesis tests concerning the mean of a population when the variance is known. However, this test is seldom applied in practice because the population standard deviation is rarely known. The section was useful, however, because it provided the framework and mechanics that are universally applied to all types of hypothesis testing.

In this section, techniques for testing a single population are presented. The techniques introduced in this section are robust, meaning that if the samples are nonnormal the techniques are still valid provided that the samples are not extremely or significantly nonnormal. Under extreme

nonnormality, nonparametric methods should be used to conduct the equivalent hypothesis tests (see Section 2.5 for a discussion on nonparametric tests).

2.3.1 Testing the Population Mean with Unknown Variance

Applying the same logic used to develop the test statistic in Equation 2.9, a test statistic for testing hypotheses about μ given that σ^2 is not known and the population is normally distributed, is

$$t^* = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad (2.10)$$

which has a t distribution with $n - 1$ degrees of freedom.

Example 2.6

Similar to the previous example, a test of whether the mean speed on Indiana roads is 60 mph is conducted at the 5% significance level. The sample size is $n = 1,296$, and the sample mean is $\bar{X} = 58.86$. As is most often the case in practice, σ^2 is replaced with $s = 4.41$, the estimated standard deviation. The statistical hypotheses to be tested are

$$\begin{aligned} H_0: \mu &= 60 \\ H_a: \mu &\neq 60 \end{aligned}$$

Using Equation 2.10, the value of the test statistic is

$$t^* = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{58.86 - 60}{4.41/\sqrt{1296}} = -335$$

The test statistic is large, and much larger than any reasonable critical value of t (Table C.2), leading again to a rejection of the null hypothesis. That is, the evidence suggests that the probability of observing these data, if the mean speed is 60 mph, is extremely small and the null hypothesis is rejected.

Why is the test statistic so large, providing strong objective evidence to reject the null hypothesis? If the sample size were 36 instead of 1,296, the test statistic would yield 0.81, leading to the inability to reject the null hypothesis. It is the square root of n in Equation 2.10 that dominates this calculation, suggesting that larger samples yield more reliable results. This emphasizes an earlier point: lack of evidence to reject the null hypothesis does not mean that the null hypothesis is true; it may mean that the effect is too small to detect $(\bar{X} - \mu)$, the sample size is too small, or the variability in the data is too large relative to the effect.

2.3.2 Testing the Population Variance

With the same logic used to develop the CIs for σ^2 in Equation 2.7, the test statistic used for testing hypotheses about σ^2 (with s^2 the estimated variance) is

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \quad (2.11)$$

which is χ^2 distributed with $n - 1$ degrees of freedom when the population variance is approximately normally distributed with variance equal to σ^2 .

Example 2.7

A test of whether the variance of speeds on Indiana roads is larger than 20 is calculated at the 5% level of significance, assuming a sample size of 100. The parameter of interest is the population variance, and the hypothesis to be tested is

$$\begin{aligned} H_0: \sigma^2 &\leq 20 \\ H_a: \sigma^2 &> 20 \end{aligned}$$

Using results from Example 2.4, the null hypothesis cannot be rejected. Nevertheless, using Equation 2.11 the test statistic is

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{99(19.51)}{20} = 96.57$$

From Table C.3 in Appendix C, the critical value for a chi-squared random variable with 99 degrees of freedom, $\sigma^2 = 0.05$ and a right-tailed test is 129.561. As expected, the null hypothesis cannot be rejected at the 0.05 level of significance.

2.3.3 Testing for a Population Proportion

The null and alternative hypotheses for proportions tests are set up in the same manner as done previously, where the null is constructed so that p is equal to, greater than, or less than a specific value, while the alternative hypothesis covers all remaining possible outcomes. The test statistic is derived from the sampling distribution of \hat{p} and is given by

$$Z^* = \frac{\hat{p} - p}{\sqrt{pq/n}} \quad (2.12)$$

where the estimated sample proportion \hat{p} is equal to the number of “successes” observed in the sample divided by the sample size n and $q = 1 - p$.

Example 2.8

The transit planning agency mentioned in Example 2.3 tests whether the transit market share in daily commute is over 20%, at the 5% significance level. The competing hypotheses are

$$\begin{aligned} H_0: p &= 0.20 \\ H_a: p &> 0.20 \end{aligned}$$

The test statistic is

$$Z^* = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.20 - 0.28}{\sqrt{(0.28)(0.72)/100}} = -1.78$$

The test statistic is not sufficiently large to warrant the rejection of the null hypothesis.

2.4 Comparing Two Populations

An extremely useful application of statistics is comparing different samples or groups. Frequently in transportation research a comparison of quantities such as speeds, accident rates, pavement performance, travel times, bridge paint life, and so on is sought. This section presents methods for conducting comparisons between groups; that is, testing for statistically significant differences between two populations. As before, the methods presented here are for interval and ratio scale variables and are robust. For ordinal scale variables and for extreme nonnormality, nonparametric alternatives should be used.

2.4.1 Testing Differences between Two Means: Independent Samples

Random samples drawn from two populations are used to test the difference between two population means. This section presents the analysis of independent samples, and the following section presents the analysis of paired observations (or matched pairs experiments). It is assumed that large samples are used to test for the difference between two population means because when sample sizes are sufficiently large the distribution of their means is approximately normally distributed as a result of the CLT. A general rule of thumb suggests that sample sizes are large if both $n_1 \geq 25$ and $n_2 \geq 25$.

There are several hypothesis testing options. The first is a test of whether the mean of one population is greater than the mean of another population (a one-tailed test). A second is a test of whether two population means are equal, assuming lack of a prior intention to prove that one mean is greater

than the other. The most common test for the difference between two population means μ_1 and μ_2 , is the one presented below where the null hypothesis states that the two means are equal

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

The competing hypotheses for a directional test, that one population mean is larger than another, are

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

The test statistic in both hypothesis testing situations is the same. As a result of the assumption about sample sizes and approximate normality of the populations, the test statistic is Z^* , such that

$$Z^* = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (2.13)$$

where $(\mu_1 - \mu_2)$ is the difference between μ_1 and μ_2 under the null hypothesis. The expression in the denominator is the standard error of the difference between the two sample means and requires two independent samples. Recall that hypothesis tests and CIs are closely related. A large sample $(1 - \alpha)100\%$ CI for the difference between two population means $(\mu_1 - \mu_2)$, using independent random samples is

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (2.14)$$

When sample sizes are small ($n_1 \leq 25$ and $n_2 \leq 25$) and both populations are approximately normally distributed, the test statistic in Equation 2.13 has approximately a t distribution with degrees of freedom given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \quad (2.15)$$

In Equations 2.13 and 2.14 it is assumed that σ_1^2 and σ_2^2 are not equal. When σ_1^2 and σ_2^2 are equal, there is an alternative test for the difference between two population means. This test is especially useful for small samples as it allows a test for the difference between two population means without having to use the complicated expression for the degrees of freedom of the approximate t distribution (Equation 2.15). When two population variances σ_1^2 and σ_2^2 are equal, then the variances are pooled together to obtain a common population variance. This pooled variance s_p^2 is based on the sample variance s_1^2 obtained from a sample of size n_1 , and a sample variance s_2^2 obtained from a sample of size n_2 , and is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (2.16)$$

A test statistic for a difference between two population means with equal population variances is given by

$$t^* = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (2.17)$$

where the term $(\mu_1 - \mu_2)$ is the difference between μ_1 and μ_2 under the null hypothesis. The degrees of freedom of the test statistic in Equation 2.17 are $n_1 + n_2 - 2$, which are the degrees of freedom associated with the pooled estimate of the population variance s_p^2 . The CI for a difference in population means is based on the t distribution with $n_1 + n_2 - 2$ degrees of freedom, or on the Z distribution when degrees of freedom are sufficiently large. A $(1 - \alpha)100\%$ CI for the difference between two population means $(\mu_1 - \mu_2)$, assuming equal population variances is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (2.18)$$

Example 2.9

Interest is focused on whether the repeal of the NMSL had an effect on the mean speeds on Indiana roads. To test this hypothesis, 744 observations in the before period and 552 observations in the after the repeal period are used. A 5% significance level is used. Descriptive statistics show that average speeds in the before and after periods are $\bar{X}_b = 57.65$ and $\bar{X}_a = 60.48$, respectively. Further,

the variances for before and after the repeal periods are $s_b^2 = 16.4$ and $s_a^2 = 19.1$, respectively. The competing hypotheses are

$$\begin{aligned} H_0: \mu_a - \mu_b &= 0 \\ H_a: \mu_a - \mu_b &\neq 0 \end{aligned}$$

Using Equation 2.13, the test statistic is

$$Z^* = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{(60.48 - 57.65) - 0}{\sqrt{\frac{19.1}{552} + \frac{16.4}{744}}} = 11.89$$

The test statistic is much larger than 1.96, the critical value for a two-tailed test at the 5% significance level, and so the null hypothesis is rejected. This result indicates that the mean speed increased after the repeal of the NMSL and that this increase is not likely to have arisen by random chance. Using Equation 2.14, a CI is obtained

$$(\bar{X}_a - \bar{X}_b) \pm Z_{\alpha/2} \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} = 2.83 \pm 1.96 \sqrt{\frac{19.1}{552} + \frac{16.4}{744}} = [2.36, 3.29]$$

Thus, with 95% confidence the increase in speeds from the before to the after period is between 2.36 and 2.29 mph.

2.4.2 Testing Differences between Two Means: Paired Observations

Presented in this section are methods for conducting hypothesis tests and for constructing CIs for paired observations obtained from two populations. The advantage of pairing observations is perhaps best illustrated by example. Suppose that a tire manufacturer is interested in whether a new steel-belted tire lasts longer than the company's current model. An experiment could be designed such that two new-design tires are installed on the rear wheels of 20 randomly selected cars and existing-design tires are installed on the rear wheels of another 20 cars. All drivers are asked to drive in their usual way until their tires wear out. The number of miles driven by each driver is recorded so a comparison of tire life can be tested. An improved experiment is possible. On 20 randomly selected cars, one of each type of tire is installed on the rear wheels and, as in the previous experiment, the cars are driven until the tires wear out.

The first experiment results in independent samples, with no relationship between the observations in one sample and the observations in the second sample. The statistical tests designed previously are appropriate for these data. In the second experiment, an observation in one sample is paired with an observation in the other sample because each pair of "competing" tires shares the same vehicle and driver. This experiment is called a matched

pairs design. From a statistical standpoint, two tires from the same vehicle are paired to remove the variation in the measurements due to driving styles, braking habits, driving surface, and so on. The net effect of this design is that variability in tire wear caused by differences other than tire type is zero (between pairs), making it more efficient with respect to detecting differences due to tire type.

The parameter of interest in this test is the difference in means between the two populations, denoted by μ_d , and is defined as $\mu_d = \mu_1 - \mu_2$. The null and alternative hypotheses for a two-tailed test case are

$$\begin{aligned} H_0: \mu_d &= 0 \\ H_a: \mu_d &\neq 0 \end{aligned}$$

The test statistic for paired observations is

$$t^* = \frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n_d}} \quad (2.19)$$

where \bar{X}_d is the average sample difference between each pair of observations, s_d is the sample standard deviation of these differences, and the sample size n_d is the number of paired observations. When the null hypothesis is true and the population mean difference is μ_d , the statistic has a t distribution with $n - 1$ degrees of freedom. Finally, a $(1 - \alpha)100\%$ CI for the mean difference μ_d is

$$\bar{X}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}} \quad (2.20)$$

2.4.3 Testing Differences between Two Population Proportions

In this section a method for testing for differences between two population proportions and drawing inferences is described. The method pertains to data measured on a qualitative (nominal), rather than a quantitative, scale. When sample sizes are sufficiently large, the sampling distributions of the sample proportions \hat{p}_1 and \hat{p}_2 and their difference $\hat{p}_1 - \hat{p}_2$ are approximately normally distributed, giving rise to the test statistic and CI computations presented.

Assuming that sample sizes are sufficiently large and the two populations are randomly sampled, the competing hypotheses for the difference between population proportions are

$$\begin{aligned} H_0: p_1 - p_2 &= 0 \\ H_a: p_1 - p_2 &\neq 0 \end{aligned}$$

As before, one-tailed tests for population proportions could be constructed. The test statistic for the difference between two population proportions when the null-hypothesized difference is 0 is

$$Z^* = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (2.21)$$

where $\hat{p}_1 = x_1/n$ is the sample proportion for sample 1, $\hat{p}_2 = x_2/n_2$ is the sample proportion for sample 2, and \hat{p} symbolizes the combined proportion in both samples and is computed as follows:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

When the hypothesized difference between the two proportions is some constant c , the competing hypotheses become

$$\begin{aligned} H_0: p_1 - p_2 &\leq 0 \\ H_a: p_1 - p_2 &> 0 \end{aligned}$$

Equation 2.21 is revised such that the test statistic becomes

$$Z^* = \frac{(\hat{p}_1 - \hat{p}_2) - c}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad (2.22)$$

The two equations reflect a fundamental difference in the two null hypotheses. For the zero-difference null hypothesis it is assumed that \hat{p}_1 and \hat{p}_2 are sample proportions drawn from one population. For the nonzero-difference null hypothesis it is assumed that \hat{p}_1 and \hat{p}_2 are sample proportions drawn from two populations and a pooled standard error of the difference between the two sample proportions is estimated.

When constructing CIs for the difference between two population proportions, the pooled estimate is not used because the two proportions are not assumed to be equal. A large sample $(1 - \alpha)100\%$ CI for the difference between two population proportions is

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad (2.23)$$

2.4.4 Testing the Equality of Two Population Variances

Suppose there is interest in the competing hypotheses

$$\begin{aligned} H_o: \sigma_1^2 &= \sigma_2^2 \\ H_a: \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

or

$$\begin{aligned} H_o: \sigma_1^2 &\leq \sigma_2^2 \\ H_a: \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

Consider first a test of whether σ_1^2 is greater than σ_2^2 . Two independent samples are collected from populations 1 and 2, and the following test statistic is computed:

$$F_{(n_1-1, n_2-1)}^* = \frac{s_1^2}{s_2^2} \quad (2.24)$$

where $F_{(n_1-1, n_2-1)}^*$ is an F -distributed random variable with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator (for additional information on the F distribution, see Appendix A). It is important to remember to place s_1^2 in the numerator because, in a one-tailed test, rejection may occur in the right tail of the distribution only. If s_1^2 is smaller than s_2^2 , the null hypothesis cannot be rejected because the statistic value is less than 1.00.

For a two-tailed test a practical approach is to insert the larger sample variance in the numerator. Then, if a test statistic is greater than a critical value associated with a specific α , the null hypothesis is rejected at the 2α significance level (double the level of significance obtained from Table C.4 in Appendix C). Similarly, if the p -value corresponding with one tail of the F distribution is obtained, it is doubled to get the actual p -value. The equivalent F test that does not require insertion of the largest sample variance in the numerator is found in Aczel (1993).

Example 2.10

There is interest in testing whether the repeal of the NMSL had an effect on the speed variances rather than mean speeds. The test is conducted at the 10% significance level. Descriptive statistics show that s_b^2 and s_a^2 , the standard deviations

for the before and after the repeal periods, are 16.4 and 19.1, respectively. The parameter of interest is variance and the competing hypotheses are

$$H_0: \sigma_a^2 - \sigma_b^2 = 0$$

$$H_a: \sigma_a^2 - \sigma_b^2 \neq 0$$

Using a two-tailed test, the larger sample variance is placed in the numerator. The test statistic value is obtained from Equation 2.24 as

$$F_{(n_1-1, n_2-1)}^* = F_{(522, 744)}^* = \frac{s_1^2}{s_2^2} = \frac{19.1}{16.4} = 1.16$$

The critical value for $\alpha = 0.05$ ($2\alpha = 0.1$) for 522 degrees of freedom in the numerator and 744 degrees of freedom in the denominator is 1.0. As a result, the analyst rejects the null hypothesis of equality of variances in the before and after periods of the repeal of the NMSL. A correct interpretation is that random sampling alone would not likely have produced the observed difference in sample variances if the variances were equal.

2.5 Nonparametric Methods

Statistical procedures discussed previously in this chapter have focused on making inferences about specific population parameters and have relied upon specific assumptions about the data being satisfied. One assumption is that the samples examined are approximately normally distributed. For the means and variances tests discussed, data are required to be measured on a ratio or interval scale. Finally, sample sizes are required to be sufficiently large. The statistical procedures discussed in the remainder of this chapter make no assumptions regarding the underlying population parameters and are, thus, called nonparametric. Some tests do not even require assumptions to be made about the distributions of the population of interest (as such, nonparametric methods are often called distribution-free methods).

Nonparametric methods typically require fewer stringent assumptions than do their parametric alternatives, and they use less information contained in the data. If nonparametric techniques are applied to data that meet conditions suitable for parametric tests, then the likelihood of committing a Type II error increases. Thus, parametric methods are more powerful and are more likely to lead correctly to rejection of a false null hypothesis. The explanation for this is straightforward. When observations measured on interval or ratio scales are ranked, that is, when they are converted to the ordinal scale, information about the relative distance between observations is lost. Because the probability of a Type I error is fixed at some value α and because

statistical tests are generally concerned with the size of an effect (difference in means, variances, etc.), using only ranked data requires more evidence to demonstrate a sufficiently large effect and so $(1 - \alpha)$, or statistical power, is reduced. Because a nonparametric test usually requires fewer assumptions and uses less information in the data, it is often said that a parametric procedure is an exact solution to an approximate problem, while a nonparametric procedure is an approximate solution to an exact problem (Conover 1980).

Given the trade offs between parametric and nonparametric methods, some guidance on when to use nonparametric methods is appropriate. A nonparametric technique should be considered under the following conditions:

1. The sample data are frequency counts and a parametric test is not available.
2. The sample data are measured on the ordinal scale.
3. The research hypotheses are not concerned with specific population parameters such as μ and σ^2 .
4. Requirements of parametric tests such as approximate normality, large sample sizes, and interval or ratio scale data, are grossly violated.
5. There is moderate violation of parametric test requirements, as well as a test result of marginal statistical significance.

The remainder of this section presents nonparametric techniques commonly encountered in practice. There are many available nonparametric tests that are not discussed and a more comprehensive list of the available techniques is presented in Table 2.3. Readers interested in comprehensive textbooks on the subject should consult Daniel (1978), Conover (1980), and Gibbons (1985a, 1985b).

2.5.1 Sign Test

The sign test is used in a large number of situations, including cases that test for central location (mean, median) differences in ordinal data or for correlation in ratio data. But, its most common application is to identify the most preferred alternative, among a set of alternatives, when a sample of n individuals (questionnaires, for example) is used. In this case the data are nominal because the expressions of interest for the n individuals simply indicate a preference. In essence, the objective of this test is to determine whether a difference in preference exists between the alternatives compared.

Consider, for example, the case where drivers are asked to indicate their preference for receiving pretrip travel time information via the Internet versus receiving the information on their mobile phones. The purpose of the study is to determine whether drivers prefer one method over the other.

TABLE 2.3

A Survey of Nonparametric Testing Methodologies

Test	Reference
<i>Location Parameter Hypotheses</i>	
Data on One Sample or Two Related Samples (Paired Samples)	
Sign test	Arbuthnott (1910)
Wilcoxon signed-rank test	Wilcoxon (1945, 1947, 1949); Wilcoxon et al. (1972)
Data on Two Mutually Independent Samples	
Mann–Whitney <i>U</i> test	Mann and Whitney (1947)
Wilcoxon rank sum test	Wilcoxon (1945, 1947, 1949); Wilcoxon et al. (1972)
Median test	Brown and Mood (1951); Westenberg (1948)
Tukey's quick test	Tukey (1959)
Normal-scores tests	Hoeffding (1951); Terry (1952); van der Waerden (1952, 1953); van der Waerden and Nievergelt (1956)
Percentile modified rank tests	Gastwirth (1965); Gibbons and Gastwirth (1970)
Wald–Wolfowitz runs tests	Wald and Wolfowitz (1940)
Data on <i>k</i>-Independent Samples (<i>k</i> ≥ 3)	
Kruskal–Wallis One-Way Analysis of Variance Test	Kruskal and Wallis (1952); Iman et al. (1975)
Steel tests for comparison with a control	Steel (1959a, 1959b, 1960, 1961); Rhyne and Steel (1965)
Data on <i>k</i>-Related Samples (<i>k</i> ≥ 3)	
Friedman two-way analysis of variance test	Friedman (1937)
Durbin test for balanced incomplete block designs	Durbin (1951)
<i>Scale or Dispersion Parameter Hypotheses</i>	
Data on Two Mutually Independent Samples	
Siegel–Tukey test	Siegel and Tukey (1960)
Mood test	Mood (1954); Laubscher et al. (1968)
Freund–Ansari test	Freund and Ansari (1957); Ansari and Bradley (1960)
Barton–David test	David and Barton (1958)
Normal-scores test	Klotz (1962); Capon (1961)
Sukhatme test	Sukhatme (1957); Laubcher and Odeh (1976)
Rosenbaum test	Rosenbaum (1953)
Kamat test	Kamat (1956)
Percentile modified rank tests	Gastwirth (1965); Gibbons and Gastwirth (1970)
Moses ranklike tests	Moses (1963)
<i>Tests of Independence</i>	
Data on Two Related Samples	
Spearman rank correlation parameter	Spearman (1904); Glasser and Winter (1961); Kendall (1962)
Kendall τ parameter	Kendall (1962); Kaarsemaker and van Wijngaarden (1953)

TABLE 2.3 (continued)

A Survey of Nonparametric Testing Methodologies

Test	Reference
Data on k-Related Samples ($k \geq 3$)	
Kendall parameter of concordance for complete rankings	Kendall (1962)
Kendall parameter of concordance for balanced incomplete rankings	Durbin (1951)
Partial correlation	Moran (1951); Maghsoodloo (1975); Maghsoodloo and Pallos (1981)
Contingency Table Data	
Chi-square test of independence	
<i>Tests of Randomness with General Alternatives</i>	
Data on One Sample	
Number of runs test	Swed and Eisenhart (1943)
Runs above and below the median	
Runs up and down test	Olmstead (1946); Edgington (1961)
Rank von Neumann runs test	Bartels (1982)
<i>Tests of Randomness with Trend Alternatives</i>	
Time Series Data	
Daniels test based on rank correlation	Daniels (1950)
Mann test based on Kendall τ parameter	Mann (1945)
Cox–Stuart test	Cox and Stuart (1955)
<i>Slope Tests in Linear Regression Models</i>	
Data on Two Related Samples	
Theil test based on Kendall's τ	Theil (1950)
Data on Two Independent Samples	
Hollander test for parallelism	Hollander (1970)
<i>Tests of Equality of k Distributions ($k \geq 2$)</i>	
Data on Two Independent Samples	
Chi-square test	
Kolmogorov–Smirnov test	Kim and Jennrich (1973)
Wald–Wolfowitz test	Wald and Wolfowitz (1940)
Data on k-Independent Samples ($k \geq 3$)	
Chi-square test	
Kolmogorov–Smirnov test	Birnbaum and Hall (1960)

continued

TABLE 2.3 (continued)

A Survey of Nonparametric Testing Methodologies

Test	Reference
<i>Tests of Equality of Proportions</i>	
Data on One Sample	
Binomial test	
<i>Data on Two Related Samples</i>	
McNemar test	McNemar (1962)
<i>Data on Two Independent Samples</i>	
Fisher's exact test	
Chi-square test	
<i>Data on k-Independent Samples</i>	
Chi-square test	
<i>Data on k-Related Samples ($k \geq 3$)</i>	
Cochran Q test	Cochran (1950)
<i>Tests of Goodness of Fit</i>	
Chi-square test	

Letting p indicate the proportion of the population of drivers favoring the Internet, the following hypotheses are to be tested

$$\begin{aligned} H_0: p &= .50 \\ H_a: p &\neq .50 \end{aligned}$$

If H_0 cannot be rejected, there is no evidence indicating a difference in preference for the two methods of delivering pretrip information. However, if H_0 is rejected, there is evidence that driver preferences are different for the two methods. In that case, the method selected by the greater number of drivers is the preferred method. Interestingly, the logic of the sign test is quite simple. For each individual, a plus sign is used if a preference is expressed for the Internet and a minus sign if the individual expresses an interest for the mobile phone. Because the data are recorded in terms of plus or minus signs, this nonparametric test is called the sign test.

Under the assumption that H_0 is true ($p = .50$), the number of signs follows a binomial distribution with $p = .50$, and one would have to refer to the binomial distribution tables to find the critical values for this test. But, for large

samples ($n > 20$), the number of plus signs (denoted by x) is approximated by a normal probability distribution with mean and standard deviation given by

$$\begin{aligned} E(+) &= 0.50n \\ \sigma_{(+)} &= \sqrt{0.25n} \end{aligned} \quad (2.25)$$

and the large sample test statistic is given by

$$Z^* = \frac{x - E(+)}{\sigma_{(+)}} \quad (2.26)$$

Example 2.11

Consider the example previously mentioned, in which 200 drivers are asked to indicate their preference for receiving pretrip travel time information via the Internet or via their mobile phones. Results show that 72 drivers preferred receiving the information via the Internet, 103 via their mobile phones, and 25 indicated no difference between the two methods. Do the responses to the questionnaire indicate a significant difference between the two methods in terms of how pretrip travel time information should be delivered?

Using the sign test, $n = 200 - 25 = 175$ individuals were able to indicate their preferred method. Applying Equation 2.25, the sampling distribution of the number of plus signs has the following properties:

$$\begin{aligned} E(+) &= 0.50n = 0.50(175) = 87.5 \\ \sigma_{(+)} &= \sqrt{0.25n} = \sqrt{0.25(175)} = 6.6 \end{aligned}$$

In addition, with $n = 175$ the sampling distribution is assumed to be approximately normal. By using the number of times the Internet was selected as the preferred alternative ($x = 72$), the following value of the test statistic (Equation 2.25) is obtained

$$Z^* = \frac{x - E(+)}{\sigma_{(+)}} = \frac{72 - 87.5}{6.6} = -2.35$$

From the test result it is obvious that the null hypothesis of no difference in how pretrip information is provided should be rejected at the 0.05 level of significance (since $|2.35| > |1.96|$). Although the number of plus signs was used to determine whether to reject the null hypothesis that $p = .50$, one could use the number of minus signs; the test results would be the same.

2.5.2 Median Test

The median test is used to conduct hypothesis tests about a population median. Recall that the median splits a population in such a way that 50% of the values are at the median or above and 50% are at the median or below. To test for the median, the sign test is applied by simply assigning a plus sign whenever the data in the sample are above the hypothesized value of the median and a minus sign whenever the data in the sample are below the hypothesized value of the median. Any data exactly equal to the hypothesized value of the median should be discarded. The computations for the sign test are done in exactly the same way as was described in the previous section.

Example 2.12

The following hypotheses have been formulated regarding the median price of new sport utility vehicles (SUVs):

$$\begin{aligned} H_0: \text{median} &= \$35,000 \\ H_a: \text{median} &\neq \$35,000 \end{aligned}$$

In a sample of 62 new SUVs, 34 have prices above \$35,000, 26 have prices below \$35,000, and 2 have prices of exactly \$35,000. Using Equation 2.24 for the $n = 60$ SUVs with prices other than \$35,000, obtained is $E(+)=30, \sigma_{(+)}=3.87$, and, using $x = 34$ as the number of plus signs, the test statistic becomes

$$Z^* = \frac{x - E(+)}{\sigma_{(+)}} = \frac{34 - 30}{3.87} = 1.03$$

Using a two-tail test and a 0.05 level of significance, H_0 cannot be rejected. On the basis of these data, the null hypothesis that the median selling price of a new SUV is \$35,000 cannot be rejected.

2.5.3 Mann–Whitney *U* Test

The Mann–Whitney–Wilcoxon (MWW) test is a nonparametric test of the null hypothesis that probability distributions of two ordinal scale variables are the same for two independent populations. The test is sensitive to the differences in location (mean or median) between the two populations. This test was first proposed for two samples of equal sizes by Wilcoxon in 1945. In 1947, Mann and Whitney introduced an equivalent statistic that could handle unequal sample sizes as well as small samples. The MWW test is most useful for testing the equality of two population means and is an alternative to the two independent samples *t* test when the assumption of normal

population distributions is not satisfied. The null and alternative hypotheses for the Mann–Whitney U test are

H_0 : The two sample distributions are drawn from the same population.

H_a : The two sample distributions are drawn from two different populations.

The hypotheses can be formulated to conduct both one- and two-tailed tests and to examine differences in population medians. The assumptions of the test are that the samples are randomly selected from two populations and are independent. To obtain the test statistic, the two samples are combined and all the observations are ranked from smallest to largest (ties are assigned the average rank of the tied observations). The smallest observation is denoted as 1 and the largest observation is n . R_1 is defined as the sum of the ranks from sample 1 and R_2 as the sum of the ranks from sample 2. Further, if n_1 is the sample size of population 1 and n_2 is the sample size of population 2, the U statistic is defined as follows:

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 \quad (2.27)$$

The U statistic is a measure of the difference between the ranks of two samples. Based on the assumption that only location (mean or median) differences exist between two populations, a large or small value of the test statistic provides evidence of a difference in the location of the two populations. For large samples the distribution of the U statistic is approximated by the normal distribution. The convergence to the normal distribution is rapid, such that for $n_1 \geq 10$ and $n_2 \geq 10$ there is a satisfactory normal approximation. The mean of U is given by

$$E(U) = \frac{n_1 n_2}{2} \quad (2.28)$$

the standard deviation of U is given by

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad (2.29)$$

and the large sample test statistic is given by

$$Z^* = \frac{U - E(U)}{\sigma_U} \quad (2.30)$$

For large samples the test is straightforward. For small samples, tables of the exact distribution of the test statistic should be used. Most software packages have the option of choosing between a large and a small sample test.

Example 2.13

Speed data were collected from two highway locations (Table 2.4). The question is whether the speeds at the two locations are identical. The first step in the MWK test is to rank speed measurements from the lowest to the highest values. Using the combined set of 22 observations in Table 2.4, the lowest data value is 46.9 mph (sixth observation of location 2) and is thus assigned the rank of 1. Continuing the ranking yields the results of Table 2.5.

In ranking the combined data, it may be that two or more data values are the same. In that case, the tied values are given the average ranking of their positions in the combined data set. The next step in the test is to sum the ranks for each sample. The sums are given in Table 2.6. The test procedure is based on the ranks for either sample. For example, the sum of the ranks for sample 1 (location 1) is 169.5. Using $R_1 = 169.5$, $n_1 = 12$ and $n_2 = 10$, the U statistic from Equation 2.27

TABLE 2.4

Speed Measurements for Two Highway Locations

Location 1		Location 2	
Observation	Speed (mph)	Observation	Speed (mph)
1	68.4	1	55.3
2	59.7	2	52.1
3	75.0	3	57.2
4	74.7	4	59.4
5	57.8	5	50.0
6	59.4	6	46.9
7	50.3	7	54.1
8	59.1	8	62.5
9	54.7	9	65.6
10	65.9	10	58.4
11	64.1		
12	60.9		

TABLE 2.5

Ranking of Speed Measurements

Speed (mph)	Item	Assigned Rank
46.9	6th observation of location 2	1
50.0	5th observation of location 2	2
50.3	7th observation of location 1	3
.	.	.
.	.	.
.	.	.
74.7	4th observation of location 1	21
75.0	3rd observation of location 1	22

TABLE 2.6

Combined Ranking of the Speed Data in the Two Samples

Location 1			Location 2		
Observation	Speed (mph)	Rank	Observation	Speed (mph)	Rank
1	68.4	20	1	55.3	7
2	59.7	14	2	52.1	4
3	75.0	22	3	57.2	8
4	74.7	21	4	59.4	12.5
5	57.8	9	5	50.0	2
6	59.4	12.5	6	46.9	1
7	50.3	3	7	54.1	5
8	59.1	11	8	62.5	16
9	54.7	6	9	65.6	18
10	65.9	19	10	58.4	10
11	64.1	17	Sum of ranks		82.5
12	60.9	15			
Sum of ranks		169.5			

equals 28.5. Further, the sampling distribution of the U statistic has the following properties:

$$\begin{aligned}E(U) &= 60 \\ \sigma_U &= 15.17\end{aligned}$$

As noted previously, with $n_1 = 12$ and $n_2 = 10$, it is assumed that the distribution of the U statistic is approximately normal; the value of the test statistic (Equation 2.30) is obtained as

$$Z^* = \frac{U - E(U)}{\sigma_U} = \frac{28.5 - 60}{15.17} = -2.07$$

At the 0.05 level of significance the null hypothesis is rejected, suggesting that speeds at the two locations are not the same.

2.5.4 Wilcoxon Signed-Rank Test for Matched Pairs

The MWW test described in the previous section deals with independent samples; the Wilcoxon signed-rank test, presented in this section, is useful for comparing two populations, say x and y , for which the observations are paired. As such, the test is a good alternative to the paired observations t test in cases where the difference between paired observations is not normally distributed. The null hypothesis of the test is that the median difference

between the two populations is zero. The test begins by computing the paired differences $d_1 = x_1 - y_1, d_2 = x_2 - y_2, \dots, d_n = x_n - y_n$, where n is the number of pairs of observations. Then, the absolute values of the differences, $d_i, i = 1, \dots, n$, are ranked from smallest to largest disregarding their sign. Following this, the sums of the ranks of the positive and negative differences are formed; that is, the smallest observation is denoted as 1 and the largest observation is n . $\Sigma(+)$ is defined as the sum of the ranks for the positive differences and $\Sigma(-)$ the sum of the ranks for the negative differences. The Wilcoxon T statistic is defined as the smaller of the two sums of ranks as follows:

$$T = \text{MIN}[\sum(+), \sum(-)] \quad (2.31)$$

As the sample size increases, the distribution of the T statistic approximates the normal distribution; the convergence to the normal distribution is good for $n \geq 25$. The mean of the T distribution is given by

$$E(T) = \frac{n(n+1)}{4} \quad (2.32)$$

the standard deviation of T is given by

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \quad (2.33)$$

and the large sample test statistic is given by

$$Z^* = \frac{T - E(T)}{\sigma_T} \quad (2.34)$$

2.5.5 Kruskal–Wallis Test

The Kruskal–Wallis test is the nonparametric equivalent to the independent samples single-factor analysis of variance. The test is applicable to problems where data are either ranked or quantitative, samples are independent, populations are not normally distributed, and the measurement scale is at least ordinal. The Kruskal–Wallis test is identical to the MWK test for comparing k populations, where k is greater than 2. In all cases the competing hypotheses are as follows:

H_0 : All k populations have the same locations (median or mean).

H_a : At least two of the k population locations differ.

First, all observations are ranked from smallest to largest disregarding their sample of origin. The smallest observation is denoted as 1 and the largest observation is n . R_1 is defined as the sum of the ranks from sample 1, R_2 the sum of the ranks from sample 2, ..., R_k is the sum of the ranks from sample k . The Kruskal–Wallis test statistic is then defined as follows:

$$W = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1) \quad (2.35)$$

For large samples ($n_i \geq 5$), the distribution of the test statistic W under the null hypothesis is approximated by the chi-square distribution with $k-1$ degrees of freedom. That is, the null hypothesis H_0 is rejected for a given level of significance α if $W > \chi^2_{k-1;\alpha}$.

If the null hypothesis is rejected an important question arises: Which populations differ? To answer this question, suppose two populations i and j are compared. For each pair of data points from the populations the average rank of the sample is computed as

$$\bar{R}_i = \frac{R_i}{n_i} \quad \text{and} \quad \bar{R}_j = \frac{R_j}{n_j} \quad (2.36)$$

where R_i is the sum of the ranks for sample i and R_j is the sum of the ranks for sample j . The test statistic D is defined as the absolute difference between \bar{R}_i and \bar{R}_j such that

$$D = |\bar{R}_i - \bar{R}_j| \quad (2.37)$$

The test is conducted by comparing the test statistic D with the critical value c_{kw} , which is computed as

$$c_{kw} = \sqrt{\chi^2_{k-1;\alpha} \left[\frac{n(n+1)}{12} \right] \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (2.38)$$

where $\chi^2_{k-1;\alpha}$ is the critical value of the chi-squared distribution used in the original test. The comparisons of all paired values is conducted jointly at the specified level of significance α . The null hypothesis is rejected if and only if $D > c_{kw}$.

2.5.6 Chi-Square Goodness-of-Fit Test

The χ^2 test sees widespread use in a variety of transportation analyses. Its popularity stems from its versatility and its ability to help assess a large number of questions. The data used in χ^2 tests are either counts or frequencies measured across categories that may be measured on any scale. Examples include the number of accidents by accident type, number of people who fall into different age and gender categories, number of speeding tickets by roadway functional class, number of vehicles purchased per year by household type, and so on. The χ^2 distribution and associated statistical tests are common and useful. All versions of the χ^2 test follow a common five-step process (Aczel 1993):

1. Competing hypotheses for a population are stated (null and alternative).
2. Frequencies of occurrence of the events expected under the null are computed. This provides expected counts or frequencies based on some “statistical model,” which may be a theoretical distribution, an empirical distribution, an independence model, and so on.
3. Observed counts of data falling in the different cells are noted.
4. The difference between the observed and the expected counts are computed and summed. The difference leads to a computed value of the χ^2 test statistic.
5. The test statistic is compared to the critical points of the χ^2 distribution and a decision on the null hypothesis is made.

In general, the χ^2 test statistic is equal to the squared difference between the observed count and the expected count in each cell divided by the expected count and summed over all cells. If the data are grouped into k cells, let the observed count in cell i be O_i and the expected count (expected under H_0) be E_i . The summation is over all cells $i = 1, 2, \dots, k$. The test statistic is

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (2.39)$$

With increasing sample size and for a fixed number of k cells, the distribution of the χ^2 test statistic approaches the χ^2 distribution with $k - 1$ degrees of freedom provided that the expected frequencies are 5 or more for all categories.

A goodness-of-fit test assesses how well the sample distribution supports an assumption about the population distribution. For example, it is often assumed that samples follow the normal distribution; the χ^2 test is used to assess the validity of such an assumption. Other assumed distributions and assumptions can also be tested with the chi-square goodness-of-fit test (for an example in transportation see Washington et al. 1999). Most statistical software packages that compare actually observed data to hypothesized

distributions use and report the X^2 test statistic. Caution should be exercised, however, when applying the χ^2 test on small sample sizes or where cells are defined such that small expected frequencies are obtained. In these instances the χ^2 test is inappropriate and exact methods should be applied (see Mehta and Patel 1983 for details).

The χ^2 distribution is useful for other applications besides goodness-of-fit tests. Contingency tables are helpful in determining whether two classification criteria, such as age and satisfaction with transit services, are independent of each other. The technique makes use of tables with cells corresponding to cross-classification of attributes or events. The null hypothesis that factors are independent is used to obtain the expected distribution of frequencies in each of the contingency table cells. The competing hypotheses for a contingency table are as follows (the general form of a contingency table is shown in Table 2.7):

H_0 : The two classification variables are statistically independent.

H_a : The two classification variables are not statistically independent.

The test statistic X^2 of Equation 2.39 for a two-way contingency table is rewritten as follows:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2.40)$$

where the differences between observed and expected frequencies are summed over all rows and columns (r and c , respectively). The test statistic in Equation 2.40 is approximately χ^2 distributed with degrees of freedom, $df = (r - 1)(c - 1)$. Finally, the expected count in cell (i, j) , where R_j and C_i are the row and column totals, respectively, is

$$E_{ij} = \frac{R_i C_j}{n} \quad (2.41)$$

The expected counts obtained from Equation 2.41 along with the observed cell counts are used to compute the value of the X^2 statistic, which provides

TABLE 2.7

General Layout of a Contingency Table

Second Classification Category	First Classification Category			Total
	1	.	j	
i	C_{1i}	.	.	R_i
	.	.	C_{ij}	R_i
Total	C_1	.	C_j	n

objective information needed to accept or reject the null hypothesis. The χ^2 test statistic can easily be extended to three or more variables, where summation is simply extended to cover all cells in the multiway contingency table. As stated previously, caution must be applied to contingency tables based on small sample sizes or when expected cell frequencies become small; in these instances the χ^2 test statistic is unreliable, and exact methods should be used.

Contingency tables and the χ^2 test statistic are also useful for assessing whether the proportion of some characteristic is equal in several populations. A transit agency, for example, may be interested in knowing whether the proportion of people who are satisfied with transit quality of service is about the same for three age groups: under 25, 25–44, and 45 and over. Whether the proportions are equal is of paramount importance in assessing whether the three age populations are homogeneous with respect to satisfaction with the quality of service. Therefore, tests of equality of proportions across several populations are called tests of homogeneity.

Homogeneity tests are conducted similarly to previously described tests, but with two important differences. First, the populations of interest are identified before the analysis and sampling is done directly from them, unlike contingency table analysis where a sample is drawn from one population and then cross-classified according to some criteria. Second, because the populations are identified and sampled from directly, the sample sizes representing the different populations of interest are fixed. This experimental setup is called a fixed marginal totals χ^2 analysis and does not affect the analysis procedure in any way.

Part II

Continuous Dependent Variable Models

3

Linear Regression

Linear regression is one of the most widely studied and applied statistical and econometric techniques, for numerous reasons. First, linear regression is suitable for modeling a wide variety of relationships between variables. In addition, the assumptions of linear regression models are often suitably satisfied in many practical applications. Furthermore, regression model outputs are relatively easy to interpret and communicate to others, numerical estimation of regression models is relatively easy, and software for estimating models is readily available in numerous “non-specialty” software packages. Linear regression can also be overused or misused. In some cases the assumptions are not strictly met, and suitable alternatives are not known, understood, or applied. Moreover, more advanced techniques may require specialized software and knowledge to estimate.

It should not be surprising that linear regression serves as an excellent starting point for illustrating statistical model estimation procedures. Although it is a flexible and useful tool, applying linear regression when other methods are more suitable should be avoided.

This chapter illustrates the estimation of linear regression models, explains when linear regression models are appropriate, describes how to interpret linear regression model outputs, and discusses how to select among a competing set of linear regression models. Matrix algebra is used throughout the chapter to illustrate the concepts, but only to the extent necessary to illustrate the most important analytical aspects.

3.1 Assumptions of the Linear Regression Model

Linear regression is used to model a linear relationship between a continuous dependent variable and one or more independent variables. Most regression applications seek to identify a set of explanatory variables that are thought to covary with the dependent variable. In general, explanatory or “causal” models are based on data obtained from well-controlled experiments (e.g., those conducted in a laboratory), predictive models are based on data obtained from observational studies, and quality control models are based on data obtained from a process or system being controlled. Whether

explanatory variables cause or are merely associated with changes in the dependent variable depends on numerous factors and cannot be determined on the basis of statistical modeling alone.

There are numerous assumptions (or requirements) of the linear regression model. When any of the requirements are not met remedial actions should be taken, and in some cases, alternative modeling approaches adopted. The following are the assumptions of the linear regression model.

3.1.1 Continuous Dependent Variable Y

An assumption in regression is that the dependent or response variable is continuous; that is, it can take on any value within a range of values. A continuous variable is measured on either the interval or ratio scales. Although, often it is done, regression on ordinal scale response variables is problematic. For example, count variables (nonnegative integers) should be modeled with count-data regression (see Chapter 11). Modeling nominal scale dependent variables (discrete variables that are not ordered) requires discrete outcome modeling approaches, presented in Chapter 13.

3.1.2 Linear-in-Parameters Relationship between Y and X

This requirement often causes confusion. The form of the regression model requires that the relationship between variables is inherently linear—a straight-line relationship between the dependent variable Y and the explanatory variables. The simple linear regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (3.1)$$

In this algebraic expression of the simple linear regression model, the dependent variable Y_i is a function of a constant term β_0 (the point where the line crosses the Y axis) and a constant β_1 times the value x_1 of independent variable X for observation i , plus a disturbance term ε_i . The subscript i corresponds to the individual or observation, where $i = 1, 2, 3, \dots, n$. In most applications the response variable Y_i is a function of many explanatory variables. In these multivariate cases it is more efficient to express the linear regression model in matrix form, where

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \quad (3.2)$$

Equation 3.2 is the regression model in matrix form, where the subscripts depict the size of the matrices. For instance, the \mathbf{X} matrix is an $n \times p$ matrix of observations, with n the number of observations and p the number of variables measured on each observation.

The linearity requirement is not as restrictive as it first sounds. Because the scales of both the dependent and independent variables are transformed, a suitable linear relationship is often found. As discussed later in the chapter and in Appendix D, there are cases when a transformation is inappropriate, requiring a different modeling framework such as logistic regression (Chapter 12), Poisson regression (Chapter 11), survival analysis (Chapter 10), and so on.

3.1.3 Observations Independently and Randomly Sampled

Although this requirement is relaxed when remedial actions are taken, an assumption necessary to make inferences about the population of interest requires that data are randomly and independently sampled from the population. Independence requires that the probability an observation is selected is unaffected by previously selected observations. In some cases, random assignment is used in lieu of random sampling, and other sampling schemes such as stratified and cluster samples are accommodated in the regression modeling framework with corrective measures.

3.1.4 Uncertain Relationship between Variables

The difference between the equation of a straight-line and a linear regression model is the addition of a stochastic, error, or disturbance term ε_i . The disturbance term consists of several components. First, it may contain the influence of variables that were omitted from the model—assumed to be the sum of many small, individually unimportant (minor) effects, some positive and others negative. Second, it contains measurement errors in the dependent variable, or the imprecision in measuring Y, again assumed to be random. Finally, it contains random variation inherent in the underlying data-generating process.

3.1.5 Disturbance Term Independent of X and Expected Value Zero

The requirements of the disturbance term ε_i is written as follows:

$$E[\varepsilon_i] = 0 \quad (3.3)$$

and

$$VAR[\varepsilon_i] = \sigma^2 \quad (3.4)$$

Equation 3.3 says that on average, model overpredictions are equal to model underpredictions—or that model disturbances sum to zero. Equation 3.4

shows that the variance of the disturbance term σ^2 is independent across observations. This property is referred to as the homoscedasticity assumption and implies that the net effect of model uncertainty, including unobserved effects, measurement errors, and true random variation, is not systematic across observations; instead it is random across observations and across covariates. When disturbances are heteroscedastic (vary systematically across observations), then alternative modeling approaches such as weighted least squares (WLSs) or generalized least squares may be required.

3.1.6 Disturbance Terms Not Autocorrelated

This requirement is written as follows:

$$\text{COV}[\varepsilon_i, \varepsilon_j] = 0 \quad \text{if } i \neq j \quad (3.5)$$

Equation 3.5 specifies that disturbances are independent across observations. Common violations of this assumption occur when observations are repeated on individuals, so the unobserved heterogeneity portion of the disturbance term ε_i is correlated within individuals. Spatial data can also exhibit dependence across observations related to location. Observations across time, however, are perhaps the most common situation in which disturbance terms are correlated. When disturbances are correlated across observations, generalized least squares or time-series methods (Chapters 7 and 8) are often most appropriate.

3.1.7 Regressors and Disturbances Uncorrelated

This property is known as exogeneity of the regressors. When the regressors are exogenous, they are not correlated with the disturbance term. Exogeneity implies that the values of the regressors are determined by influences “outside of the model.” So Y does not directly influence the value of an exogenous regressor. In mathematical terms, this requirement translates to

$$\text{COV}[X_i, \varepsilon_j] = 0 \quad \text{for all } i \text{ and } j \quad (3.6)$$

When an important variable is endogenous (depends on Y), then alternative methods are required, such as instrumental variables, two and three stage least squares, or structural equations models.

3.1.8 Disturbances Approximately Normally Distributed

Although not a requirement for the estimation of linear regression models, the disturbance terms are required to be approximately normally distributed

TABLE 3.1

Summary of Ordinary Least Squares Linear Regression Model Assumptions

Statistical Assumption	Mathematical Expression
1. Functional form	$Y_i = \beta_0 + \beta_1 X_{1i} + e_i$
2. Zero mean of disturbances	$E[e_i] = 0$
3. Homoscedasticity of disturbances	$VAR[e_i] = \sigma^2$
4. Nonautocorrelation of disturbances	$COV[e_i, e_j] = 0 \text{ if } i \neq j$
5. Uncorrelatedness of regressor and disturbances	$COV[X_i, e_j] = 0 \text{ for all } i \text{ and } j$
6. Normality of disturbances	$e_i \sim N(0, \sigma^2)$

to make inferences about the parameters from the model. In this regard the central limit theorem enables exact inference about the properties of statistical parameters. Combined with the independence assumption, this assumption results in disturbances that are independently and identically distributed as normal (i.i.d. normal). In notation, this assumption requires

$$e_i \sim N(0, \sigma^2) \quad (3.7)$$

This approximation is valid because the disturbances for any sample are not expected to be exactly normally distributed due to sampling variation. It follows that $Y_i \approx N(\beta X_i, \sigma^2)$.

3.1.9 Summary

The analytical assumptions of the regression model are summarized in Table 3.1. The table lists the assumptions described earlier. It is prudent to consider the assumptions in the table as modeling requirements that need to be met. If the requirements are not met, then remedial measures must be taken. In some cases a remedial action is taken that enables continued modeling of the data in the linear regression modeling framework. In other cases alternative modeling methodologies must be adopted to deal effectively with the nonideal conditions. Chapter 4 discusses in detail remedial actions when regression assumptions are not met.

3.2 Regression Fundamentals

The objective of linear regression is to model the relationship between a dependent variable Y with one or more independent variables X . The ability to say something about the way X affects Y is through the parameters in the regression model—the betas. Regression seeks to provide information

and properties about the parameters in the population model by inspecting properties of the sample-estimated betas, how they behave, and what they can tell us about the sample and, thus, about the population.

The linear regression model thought to exist for the entire population of interest is

$$E[Y_i | X_i] = E[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_{p-1} X_{p-1,i}] \quad (3.8)$$

The true population model is formulated from theoretical considerations, past research findings, and postulated theories. The expected value of Y_i given covariate vector X_i is a conditional expectation. In some texts the conditional expectation notation is dropped, but it should be understood that the mean or expected value of Y_i is conditional on the covariate vector for observation i . The population model represents a theoretically postulated model whose parameter values are unknown, constant, and denoted with betas, as shown in Equation 3.8. The parameters are unknown because Equation 3.8 is based on all members of the population of interest. The parameters (betas) are constant terms that reflect the underlying true relationship between the independent variables X_1, X_2, \dots, X_{p-1} and dependent variable Y_i , because the population N is presumably finite at any given time. The true population model contains p parameters in the model, and there are n observations.

The unknown disturbance term for the population regression model (Equation 3.8) is given by

$$\varepsilon_i = Y_i - \hat{Y}_i = Y_i - E[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_{p-1} X_{p-1,i}] \quad (3.9)$$

The equivalent expressions for population models (Equations 3.8 and 3.9) in matrix form (dropping the conditional expectation and replacing expected value notation with the hat symbol) are given by

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \quad (3.10)$$

and

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \quad (3.11)$$

Regression builds on the notion that information is learned about the unknown and constant parameters (betas) of the population by using information contained in the sample. The sample is used for estimating betas—random variables that fluctuate from sample to sample—and the properties of these are used to make inferences about the true population betas. There are numerous procedures to estimate the parameters of the true population model based on the sample data, including least squares and maximum likelihood.

3.2.1 Least Squares Estimation

Least squares estimation is a commonly employed estimation method for regression applications. Often referred to as “ordinary least squares” or OLS, it represents a method for estimating regression model parameters using the sample data. OLS estimation using both standard and matrix algebra is shown using Equations 3.8 and 3.10 for the simple regression model case as the starting point.

First consider the algebraic expression of the OLS regression model shown in Equation 3.8. OLS, as one might expect, requires a minimum (least) solution of the squared disturbances. OLS seeks a solution that minimizes the function Q (the subscript for observation number is not shown)

$$\begin{aligned} Q_{\min} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)_{\min}^2 = \sum_{i=1}^n (Y_i - [\beta_0 + \beta_1 X_i])_{\min}^2 \\ &= \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)_{\min}^2 \end{aligned} \quad (3.12)$$

Those values of β_0 and β_1 that minimize the function Q are the least squares estimated parameters. Of course β_0 and β_1 are parameters of the population and are unknown, so estimators B_0 and B_1 are obtained, which are random variables that vary from sample to sample. By setting the partial derivatives of Q with respect to β_0 and β_1 equal to zero, the least squares estimated parameters B_0 and B_1 are obtained:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0 \quad (3.13)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \quad (3.14)$$

Solving the previous equations using $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the estimates of β_0 and β_1 , respectively, and rearranging terms yields

$$\sum_{i=1}^n Y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_i \quad (3.15)$$

and

$$\sum_{i=1}^n X_i Y_i = \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2 \quad (3.16)$$

One can verify that these equations are minimums by their positive second partial derivatives. As one might expect, there is a least squares normal equation generated for each parameter in the model, and so a linear regression model with $p = 5$ would yield five normal equations. Equations 3.15 and 3.16 are the least squares normal equations for the simple linear regression model. Solving simultaneously for the betas in Equations 3.15 and 3.16 yields

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.17)$$

and

$$\hat{\beta}_0 = \frac{1}{n} \left(\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i \right) = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3.18)$$

The derivation of the matrix algebra equivalent of the least squares normal equations for the simple regression case is straightforward. In the simple regression case, the expression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ consists of the following matrices:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \quad (3.19)$$

The following matrix operations are used to solve for the betas:

$$\begin{aligned} (\text{step 1}): \quad & \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} \\ (\text{step 2}): \quad & \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} \\ (\text{step 3}): \quad & (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{B} \end{aligned} \quad (3.20)$$

The beta vector is shown as $\hat{\boldsymbol{\beta}}$, since it is the estimated vector of betas for the true beta vector $\boldsymbol{\beta}$. Step 2 of Equation 3.20 is used to demonstrate the equivalence of the normal equations (Equations 3.15 and 3.16) obtained using algebraic and matrix solutions. Step 2 is rewritten as

$$\begin{bmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix}^T \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix}^T \begin{bmatrix} 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \quad (3.21)$$

The matrix products in Equation 3.21 yield

$$\begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n Y_i X_i \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \quad (3.22)$$

By writing out the equations implied by Equation 3.22, the least squares normal equations 3.15 and 3.16 are obtained. It turns out that this derivation is a general result, and step 3 of Equation 3.20 is used to obtain the least squares estimated parameters for a regression model with p parameters.

Most standard statistical software packages calculate least squares parameters. The user of software should be able to find in supporting literature and/or user's manuals the type of estimation method being used to estimate parameters in a model. There are occasions when violations of regression assumptions require remedial actions to be taken, as the properties of the least squares estimators become inefficient, biased, or both (see Chapter 1 for properties of estimators in general).

Example 3.1

A database consisting of 121 observations is available to study annual average daily traffic ($AADT$) in Minnesota. Variables available in this database, and their abbreviations, are provided in Table 3.2. A simple regression model is estimated using least squares estimated parameters as a starting point. The starter specification is based on a model that has $AADT$ as a function of $CNTYPOP$, $NUMLANES$, and $FUNCTIONALCLASS$: $AADT = \beta_0 + \beta_1(CNTYPOP) + \beta_2(NUMLANES) + \beta_3(FUNCTIONALCLASS) + \text{disturbance}$.

The model is not a deterministic model because numerous random fluctuations in $AADT$ are thought to exist from day to day as a result of fluctuations in the number of commuters, vacation travelers, out-of-area traffic, and so on. In addition, numerous other variables thought to affect $AADT$ were not collected or available, such as the spatial distribution of employment, shopping, recreation, and residential land uses, which are thought to affect local traffic volumes and $AADT$.

An initial regression model is estimated. The matrices (with observations 4 through 118 not shown) used in the least squares estimation are shown below. Note that independent variable X_1 is $CNTYPOP$, X_2 is $NUMLANES$, and variables X_3 through X_5 refer to various functional classifications (see section on indicator variables later in this chapter), with $X_3=1$ for rural interstates, $X_4=1$ for rural noninterstates, and $X_5=1$ for urban interstates; otherwise these variables are zero. For example, the first observation had an average annual daily traffic of 1,616

TABLE 3.2

Variables Collected on Minnesota Roadways

Variable No.	Abbreviation: Variable Description
1	AADT: Average annual daily traffic in vehicles per day
2	CNTYPOP: Population of county in which road section is located (proxy for nearby population density)
3	NUMLANES: Number of lanes in road section
4	WIDTHLANES: Width of road section in feet
5	ACCESSCONTROL: 1 for access controlled facility; 2 for no access control
6	FUNCTIONALCLASS: Road sectional functional classifications; 1 = rural interstate, 2 = rural noninterstate, 3 = urban interstate, 4 = urban noninterstate
7	TRUCKROUTE: Truck restriction conditions: 1 = no truck restrictions, 2 = tonnage restrictions, 3 = time of day restrictions, 4 = tonnage and time of day restrictions, 5 = no trucks
8	LOCALE: Land-use designation: 1 = rural, 2 = urban with population ≤ 50,000, 3 = urban with population > 50,000

vehicles per day, a county population of 13,404, two lanes, and was a rural non-interstate facility.

$$\mathbf{Y} = \begin{bmatrix} 1616 \\ 1329 \\ 3933 \\ \vdots \\ 14,905 \\ 15,408 \\ 1,266 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 13,404 & 2 & 0 & 1 & 0 \\ 1 & 52,314 & 2 & 0 & 1 & 0 \\ 1 & 30,982 & 2 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 459,784 & 4 & 0 & 0 & 0 \\ 1 & 459,784 & 2 & 0 & 0 & 0 \\ 1 & 43,784 & 2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

The estimated least squares regression parameters are shown in Table 3.3. The values of the estimated betas give an indication of the value of the true population parameters if the presumed model is correct. For example, for each additional 1,000 people in the local county population, there is an estimated 29 additional AADT. For each lane there is an estimated 9,954 AADT.

Urban interstates are associated with an estimated 35,454 AADT more than the functional classifications that are not included in the model (urban noninterstates). Similarly, rural noninterstates are associated with an estimated 4,128 AADT more, and rural interstates are associated with 886 AADT more on average. Using these values, it is determined that urban interstates are associated with 31,326 (35,454 minus 4,128) more AADT on average than rural noninterstates.

What remains to be examined is whether the estimated regression model is correct. In other words, are the variables entered in the model truly associated with changes in AADT, are they linearly related, and do they do an adequate job of explaining

TABLE 3.3

Least Squares Estimated Parameters
(Example 3.1)

Variable	Estimated Parameter
Intercept	-2,6234.490
CNTYPOP	0.029
NUMLANES	9,953.676
FUNCLASS1	885.384
FUNCLASS2	4,127.560
FUNCLASS3	35,453.679

variation in *AADT*? Additional theory and tools are needed to assess the viability of the entered variables and the GOF of the model to the Minnesota *AADT* data.

3.2.2 Maximum Likelihood Estimation

The previous section showed the development of the OLS estimators through the minimization of the function Q . Another popular and sometimes useful statistical estimation method is called maximum likelihood estimation, which results in the maximum likelihood estimates, or MLEs. The general form of the likelihood function is described in Appendix A as the joint density of observing the sample data from a statistical distribution with parameter vector β , such that

$$f(x_1, x_2, \dots, x_n, \beta) = \prod_{i=1}^n f(x_i, \beta) = L(\beta | X)$$

For the regression model, the likelihood function for a sample of n independent, identically, and normally distributed disturbances is given by

$$\begin{aligned} L &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{X}_i^T \boldsymbol{\beta})^2\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right] \end{aligned} \quad (3.23)$$

As is usually the case, the logarithm of Equation 3.23, or the log-likelihood, is simpler to solve than the likelihood function itself, so taking the log of L yields

$$\ln(L) = LL = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.24)$$

Maximizing the log-likelihood with respect to β and σ^2 reveals a solution for the estimates of the betas that is equivalent to the OLS estimates, that is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Thus, the MLE and OLS estimates are equivalent for the regression model with assumptions listed in Table 3.1. It turns out that the MLE for the variance is biased toward zero, and is a result of small sample bias. In other words, MLEs are borne out of asymptotic theory, and as the sample size increases, the MLE estimates are consistent.

A couple of issues are worth mentioning at this stage. First, the selection of MLE or OLS estimators depends on the regression setting. It is shown in later chapters that, in general, when the response variable and disturbances are not normally distributed, MLE is the preferred method of estimation. When regression is being performed under ideal conditions expressed in Table 3.1, OLS is the preferred method. Second, MLE requires that the distribution family of the response variable be specified a priori, whereas OLS does not.

3.2.3 Properties of OLS and MLE Estimators

As discussed in Chapter 1, statistical estimators are sought that have small (minimum) sampling variance and are unbiased. The previous section showed that the MLE and OLS estimators are equivalent for the regression model with assumptions listed in Table 3.1.

The expected value of the MLE and OLS estimators is obtained as follows:

$$\begin{aligned} E\left(\left[\mathbf{X}^T \mathbf{X}\right]^{-1} \mathbf{X}^T \mathbf{Y}\right) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X}) \beta = \beta \end{aligned} \quad (3.25)$$

Thus, the MLE and OLS estimators are unbiased estimators of the betas.

The variance of the MLE and OLS estimator for the betas in the classical linear regression model without the assumption of normality has been derived through the Gauss–Markov theorem (for proofs see Myers 1990 or Greene 1990a). The theorem states that under the classical regression assumptions (without normality), the MLE and OLS estimators achieve minimum variance of the class of linear unbiased estimators. This finding means that only through the use of biased estimators can smaller variances be achieved. If the assumption of normality is imposed (in addition to independence and $E[\varepsilon] = 0$), then the OLS and MLE estimators are minimum variance among all unbiased estimators (the linear restriction is dropped). To compute the variance of the estimator, it is useful first to write the relation between $\hat{\beta}$ and β

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon) \\ &= \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \end{aligned} \quad (3.26)$$

Then, using the fact that $VAR[\mathbf{X}] = E[(x - \mu)^2]$, $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$, $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$, and $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 \mathbf{I}$, the variance of the OLS and MLE estimator is

$$\begin{aligned}
 VAR[\hat{\boldsymbol{\beta}}] &= E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right] \\
 &= E\left[\left(\left[\mathbf{X}^T \mathbf{X}\right]^{-1} \mathbf{X}^T \boldsymbol{\epsilon}\right)\left(\left[\mathbf{X}^T \mathbf{X}\right]^{-1} \mathbf{X}^T \boldsymbol{\epsilon}\right)^T\right] \\
 &= E\left[\left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\epsilon} \boldsymbol{\epsilon}^T \mathbf{X} \left(\left[\mathbf{X}^T \mathbf{X}\right]^{-1}\right)^T\right] \\
 &= E\left[\left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\epsilon} \boldsymbol{\epsilon}^T \mathbf{X} \left(\left[\mathbf{X}^T \mathbf{X}\right]^T\right)^{-1}\right] \\
 &= \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T E\left(\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T\right) \mathbf{X} \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \\
 &= \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \\
 &= \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}
 \end{aligned} \tag{3.27}$$

Both the MLE and OLS estimates, which are equivalent for the normal theory regression model, are the best linear unbiased estimators of the underlying population parameters $\boldsymbol{\beta}$.

3.2.4 Inference in Regression Analysis

To illustrate how inferences are made with the beta parameters in classical linear regression, consider the sampling distribution of the estimated parameter $\hat{\beta}_1$. The point estimator $\hat{\beta}_1$ in algebraic terms is given in Equation 3.17, and the matrix equivalent is provided in Equation 3.20. The sampling distribution of $\hat{\beta}_1$ is the distribution of values that would result from repeated samples drawn from the population with levels of the independent variables held constant. To see that this sampling distribution is approximately normal, such that

$$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) \tag{3.28}$$

one must first show, using simple algebra (see Neter et al. 1996) that $\hat{\beta}_1$ is a linear combination of the observations, Y , such that $\hat{\beta}_1 = \sum k_i Y_i$. Since the Y_i terms are independently and normally distributed random variables and given that a linear combination of independent normal random variables is normally distributed, Equation 3.28 follows.

The variance of the OLS estimator for some parameter $\hat{\beta}_k$ was shown to be a minimum unbiased estimator in Equation 3.27. However, the true

population variance σ^2 is typically unknown, and instead is estimated with an unbiased estimate called *mean squared error* (error is equivalent term for disturbance), or MSE. MSE is an estimate of the variance in the regression model and is given as

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-p} \quad (3.29)$$

where n is the sample size and p is the number of estimated model parameters. It is straightforward to show that MSE is an unbiased estimator of σ^2 , or that $E[MSE] = \sigma^2$.

Because $\hat{\beta}_k$ is normally distributed, and β_k is a constant and is the expected value of $\hat{\beta}_k$, the quantity

$$Z^* = \frac{\hat{\beta}_k - \beta_k}{\sigma \{\hat{\beta}_k\}}$$

is a standard normal variable. In practice the true variance in the denominator is not known and is estimated using MSE. When σ^2 is estimated using MSE, the following is obtained:

$$t^* = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\frac{MSE}{\sum (X_i - \bar{X})^2}}} = \frac{\hat{\beta}_k - \beta_k}{s \{\hat{\beta}_k\}} \approx t(\alpha; n-p) \quad (3.30)$$

where α is the level of significance and $n-p$ is the associated degrees of freedom. This result is important; it enables a statistical test of the probabilistic evidence in favor of specific values of β_k .

Perhaps the most common use of Equation 3.30 is in the development of a confidence interval (CI) around the parameter β_k . A CI for the parameter β_1 is given by

$$\hat{\beta}_k \pm t\left(1 - \frac{\alpha}{2}; n-p\right) s \{\hat{\beta}_k\} \quad (3.31)$$

The correct interpretation of a CI is as follows. In a $(1 - \alpha)\%$ CI on β , the true value of β will lie within the CI $(1 - \alpha) \times 100$ times out of 100 on average for repeated samples drawn from the population with levels of the independent variables held constant from sample to sample. Thus, the CI provides the

long-run probability that the true value of β lies in the computed CI, conditioned on the same levels of X being sampled. Incorrect interpretations of CIs include statements that assign a probability to the particular experimental or observational outcome obtained and statements that exclude the conditional dependence on repeated samples at levels of the independent variables.

Often the statistic t^* in Equation 3.30 is employed to conduct hypothesis tests. A two-sided hypothesis test is set up as

$$H_0: \beta_k = 0$$

$$H_a: \beta_k \neq 0$$

The decision rule given a level of significance α is

$$\begin{aligned} & \text{If } |t^*| \leq t_{crit} \left(1 - \frac{\alpha}{2}; n-p \right), \text{ conclude } H_0 \\ & \text{If } |t^*| > t_{crit} \left(1 - \frac{\alpha}{2}; n-p \right), \text{ conclude } H_a \end{aligned} \quad (3.32)$$

where t_{crit} is the critical value of t corresponding with level of significance α and degrees of freedom $n - 2$.

A one-sided hypothesis test may be set up as follows. The null and alternative hypotheses are

$$H_0: \beta_k \leq 0$$

$$H_a: \beta_k > 0$$

The decision rules for the one-sided test are

$$\begin{aligned} & \text{If } t^* \leq t_{crit} (1 - \alpha; n-p), \text{ conclude } H_0 \\ & \text{If } t^* > t_{crit} (1 - \alpha; n-p), \text{ conclude } H_a \end{aligned} \quad (3.33)$$

where again t_{crit} is the critical value of t corresponding with level of significance α and degrees of freedom $n - 2$. Note that in the one-sided test the absolute value brackets are removed from the decision rule, making the decision rule directional.

Example 3.2

Consider again the study of AADT in Minnesota. Interest centers on the development of CIs and hypothesis tests on some of the parameters estimated in Example 3.1. Specifically, it is of interest to assess whether observed data support

the originally proposed model. Shown in Table 3.4 is expanded statistical model output.

Shown in the table are the least squares regression parameter estimates, the estimated standard errors of the estimate $s\{\hat{\beta}_k\}$, the t -values corresponding with each of the estimated parameters, and finally the p -value—the probability associated with rejection of the null hypothesis that $\hat{\beta}_k = 0$. In many standard statistical packages the default null hypothesis is the two-tailed test of $\beta_k = 0$. For example, consider the hypothesis test of whether county population (*CNTYPOP*) has a statistically significant effect on *AADT*. The null and alternative hypotheses are

$$\begin{aligned} H_0: \beta_{\text{CNTYPOP}} &= 0 \\ H_a: \beta_{\text{CNTYPOP}} &\neq 0 \end{aligned}$$

Using the two-sided decision rule (see Equation 3.32 and an α of 0.005) results in the following:

$$\text{If } \left| \frac{\hat{\beta}_{\text{CNTYPOP}} - \beta_{\text{CNTYPOP}}}{s\{\hat{\beta}_{\text{CNTYPOP}}\}} \right| = \left| \frac{0.0288 - 0}{0.0048} \right| = |6| \leq t_{\text{crit}} \left(1 - \frac{0.01}{2}; 120 - 5 \right)$$

The critical value t_{crit} for $1 - \alpha = 0.995$ and 115 degrees of freedom is 2.617. Because 6 is greater than 2.617, the null hypothesis is rejected. The standard S-Plus output also gives this information in a slightly different form (as do most other statistical packages). The output shows that $t^*, 5.9948 \approx 6$, is associated with a p -value of less than .0001. Stated another way, the probability of obtaining the parameter value of .0288 conditional on the null hypothesis being true (the parameter is equal to zero) is almost 0. There is always some small probability greater than zero, which means there is always some probability of making a Type I error.

The range of plausible values of the estimated effect of *FUNCLASS2* on *AADT* is of interest. A 99% CI is computed as

$$\begin{aligned} \hat{\beta}_{\text{FUNCLASS2}} &\pm t \left(1 - \frac{\alpha}{2}; n - p \right) s\{\hat{\beta}_{\text{FUNCLASS2}}\} \\ &\Rightarrow 4127.6 \pm 2.617(3346.4) = 4127.6 \pm 8757.5 \\ &\Rightarrow -4629.9 \leq \beta_{\text{FUNCLASS2}} \leq 12885.1 \end{aligned}$$

TABLE 3.4

Least Squares Estimated Parameters (Example 3.2)

Parameter	Parameter Estimate	Standard Error of Estimate	t-Value	P(> t)
Intercept	-26,234.490	4,935.656	-5.314	<.0001
<i>CNTYPOP</i>	0.029	0.005	5.994	<.0001
<i>NUMLANES</i>	9,953.676	1,375.433	7.231	<.0001
<i>FUNCLASS1</i>	885.384	5,829.987	0.152	.879
<i>FUNCLASS2</i>	4,127.560	3,345.418	1.233	.220
<i>FUNCLASS3</i>	35,453.679	4,530.652	7.825	<.0001

The linear relationship between *AADT* and *FUNCLASS2* (facility is a rural non-interstate) is between -4,629.9 and 12,885.1 with 99% confidence. In particular, the CI would contain the true value of the parameter in 99 out of 100 repeated samples. Thus, the true parameter value could be positive, negative, or zero. At this confidence level there is insufficient evidence to support the parameter *FUNCLASS2* as a statistically significant or meaningful variable for explaining variation in *AADT*.

Assessment of the proposed model suggests that the intercept term and the variables *CNTYPOP*, *NUMLANES*, and *FUNCLASS3* are statistically significant variables, while *FUNCLASS2* and *FUNCLASS1* are not statistically significant at the 95% confidence level.

3.3 Manipulating Variables in Regression

There are numerous motivations and techniques for manipulating variables in the regression. Standardized regression models allow for direct comparison of the relative importance of independent variables. Transformations allow regression assumptions to be met, indicator variables allow appropriate expression of nominal and ordinal scale variables, and interactions enable synergistic effects among variables to be modeled.

3.3.1 Standardized Regression Models

Often interest is focused on the relative impacts of independent variables on the response variable *Y*. Using the original measurement units of the variables will not provide an indication of which variables have largest relative impact on *Y*. For example, if two independent variables in a model describing the expected number of daily trip chains were number of children and household income, one could not directly compare the estimated parameters for these two variables, because the former is in units of marginal change in the number of daily trip chains per unit child and the latter is marginal change in the number of daily trip chains per unit of income. Number of children and household income are not directly comparable. Households may have children ranging from 0 to 8, while income may range from \$5,000 to \$500,000, making it difficult to make a useful comparison between the relative impacts of these variables.

The standardized regression model is obtained by standardizing all independent variables. Standardization strictly works on continuous variables, those measured on interval or ratio scales. New variables are created by applying the standardization formula

$$X'_1 = \frac{X_1 - \bar{X}}{s(X_1)}$$

to all continuous independent variables, where standardized variables are created with expected values equal to 0 and variances equal to 1. Thus, all continuous independent variables are measured on the same scale, so comparisons across standardized units are possible. The estimated regression parameters in a standardized regression model are interpreted as a change in the response variable per unit change of one standard deviation of the independent variable. Most statistical software packages will generate standardized regression model output as an option in OLS regression. If this option is not directly available, most packages will enable the standardization of variables, which are then used in the regression.

3.3.2 Transformations

The application of linear regression requires relationships that are linear (with some rare exceptions such as polynomial regression models). This linearity is a direct result of the linear-in-parameters restriction. It is not always expected or defensible, however, that physical, engineering, or social phenomena are best represented by linear relationships, and nonlinear relationships are often more appropriate. Fortunately, nonlinear relationships are possible to accommodate within the linear regression framework. Being able to accommodate nonlinear relationships provides a great deal of flexibility for finding defensible and suitable relationships between two or more variables. Because variable transformations are useful in the application of many statistical methods, Appendix D presents a comprehensive list of transformation options, the nature of different transformations, and the consequences. In this section two example transformations, as they might be used in regression as well as their consequences and impacts, are presented and discussed.

Some ground rules for using transformations in the linear regression modeling framework are worthy of note

1. Linear relationships between the dependent and all independent variables is a requirement of the regression modeling framework, with rare exceptions being polynomial expressions of variables such as X , X^2 , X^3 , and so on.
2. Research interest is typically focused on Y in its original measurement units, and not transformed units.
3. Transformations on Y , X , or both, are merely a way to rescale the observed data so that a linear relationship is identified.

Suppose the relationship between variables Y and X is modeled using linear regression. To obtain a linear relationship required by linear regression,

the reciprocals of both Y and X are computed such that the estimated model takes the form

$$E\left[\frac{1}{\hat{Y}}\right] = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X} + e \quad (3.34)$$

To obtain original Y measurement units and determine the nature of the underlying true model assumed by this transformation, the reciprocal of both sides is computed to obtain the following

$$\begin{aligned} \frac{1}{\left(\frac{1}{\hat{Y}}\right)} &= \frac{1}{\hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X} + e} \\ \Rightarrow \hat{Y} &= \frac{1}{\hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X} + e} \left(\frac{X}{X}\right) = \frac{X}{\hat{\beta}_0 X + \hat{\beta}_1 + e X} \\ &= \frac{X}{\alpha X + \lambda + Xe} \end{aligned} \quad (3.35)$$

By taking the reciprocal of both sides of Equation 3.34, the unknown true relationship is assumed to take the form shown in Equation 3.35, where $\beta_0 = \hat{\beta}_1$, $\beta_1 = \hat{\beta}_0$, and the disturbance term is a function of X. Important to note is that the underlying model implied by this transformation is more complex than the linear regression model assumes: linear additive parameters and an additive disturbance term.

Consider another example. Suppose that the logarithm of Y is a linear function of the inverse of X. That is, an estimated regression function takes the form

$$LN(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X} + e \quad (3.36)$$

What underlying model form do these transformations imply about the relationship between Y and X? Again, the original Y units are obtained by performing the following transformations:

$$\begin{aligned} LN(\hat{Y}) &= \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X} + e \\ \Rightarrow EXP [LN(\hat{Y})] &= EXP \left[\hat{\beta}_0 + \left(\frac{\hat{\beta}_1}{X} \right) + e \right] \\ \Rightarrow \hat{Y} &= EXP [\hat{\beta}_0] EXP \left[\frac{\hat{\beta}_1}{X} \right] EXP[e] \end{aligned} \quad (3.37)$$

In this function, $\beta_0 = \text{EXP}[\hat{\beta}_0]$, $\beta_i = \text{EXP}[\hat{\beta}_i]$, and the disturbance term is a function of β_0 , β_1 , and X and is multiplicative.

Similar derivations are suitable for a wide range of transformations often applied to dependent and independent variables. Appendix D describes a variety of typical transformations and some of their permutations. The use of transformations implies alternative relationships about the underlying true and unknown model, leading to the following points of caution:

1. The implied underlying model should be consistent with the theoretical expectations of the phenomenon being studied.
2. An additive disturbance term may be converted to a multiplicative disturbance term through the application of transformations.
3. If the disturbances do not behave well in the linear regression framework, then it is likely that the correct specification has not yet been found.
4. Sometimes it is easier and more appealing to estimate directly the underlying true model form using a nonlinear regression modeling technique than to apply a set of transformations on the raw data.

3.3.3 Indicator Variables

Often there is interest in modeling the effects of ordinal and nominal scale variables. As examples the effects of qualitative variables such as roadway functional class, gender, attitude toward transit, and trip purpose are often sought. The interpretation of nominal and ordinal scale variables in regression models is different from that for continuous variables.

For nominal scale variables, $m - 1$ indicator variables must be created to represent all m levels of the variable in the regression model. These $m - 1$ indicator variables represent different categories of the response, with the omitted level captured in the slope intercept term of the regression. Theoretically meaningless and statistically insignificant indicator variables should be removed or omitted from the regression, leaving only those levels of the nominal scale variable that are important and relegating other levels to the “base” condition (slope intercept term).

Example 3.3

Consider again the study of *AADT* in Minnesota. In Example 3.2, three indicator variables representing three different functional classes of roadways were used. The variable *FUNCLASS3* represents the subset of observations belonging to the class of facilities that are urban interstates. The example shows that urban interstates have an estimated marginal effect of 35,453.7 *AADT*; that is, urban interstates are associated with 35,453.7 more *AADT* on average than urban noninterstates (*FUNCLASS4*). The effect of this indicator variable is statistically significant, whereas the effects for *FUNCLASS1* and *FUNCLASS2* are not. In practice these findings would be

contrasted to theoretical expectations and other empirical findings. The effect of urban noninterstates (*FUNCLASS4*) is captured in the intercept term. Specifically, for urban noninterstate facilities the county population (*CNTYPOP*) is multiplied by 0.0288, then 9,953.7 is added for each lane of the facility, and then 26,234.5 (the Y-intercept term $\hat{\beta}_0$) is subtracted to obtain an estimate of *AADT*.

Ordinal scale variables, unlike nominal scale variables, are ranked, and can complicate matters in the regression. Several methods for dealing with ordinal scale variables in the regression model, along with a description of when these methods are most suitable, are provided in the next section.

3.4 Estimate a Single Beta Parameter

The assumption in this approach is that the marginal effect is equivalent across increasing levels of the variable. For example, consider a variable that reflects the response to the survey question: Do you support congestion pricing on the I-10 expressway? The responses to the question include 1 = do not support; 2 = not likely to support; 3 = neutral; 4 = likely to support; and 5 = strongly support. This variable reflects an ordered response with ordered categories that do not possess even intervals across individuals or within individuals. If a single beta parameter for this variable is estimated, the assumption is that each unit increase (or decrease) of the variable has an equivalent effect on the response. This assumption is unlikely to be valid. Thus, disaggregate treatment of ordinal variables is more appropriate.

3.5 Estimate Beta Parameter for Ranges of a Variable

Suppose that an ordinal variable had two separate effects, one across one portion of the variable's range of values, and another over the remainder of the variable. In this case, two indicator variables are created, one for each range of the variable. Consider the variable *NUMLANES* used in previous chapter examples. Although the intervals between levels of this variable are equivalent, it may be believed that a fundamentally different effect on *AADT* exists for different levels of *NUMLANES*. Thus, two indicator variables could be created such that

$$Ind_1 = \begin{cases} NUMLANES & \text{if } 1 \leq NUMLANES \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

$$Ind_2 = \begin{cases} NUMLANES & \text{if } NUMLANES > 2 \\ 0 & \text{otherwise} \end{cases}$$

These two indicator variables would allow the estimation of two parameters, one for the lower range of the variable *NUMLANES* and one for the upper range. If there was an *a priori* reason to believe that these effects would be different given theoretical, behavioral, or empirical considerations, then the regression would provide evidence on whether these separate ranges of *NUMLANES* supported separate regression parameters. Note that a linear dependency exists between these two indicator variables and the variable *NUMLANES*.

3.6 Estimate a Single Beta Parameter for $m - 1$ of the m Levels of a Variable

The third, most complex treatment of an ordinal variable is equivalent to the treatment of a nominal scale variable. In this approach $m - 1$ indicator variables are created for the m levels of the ordinal scale variable. The justification for this approach is that each level of the variable has a unique marginal effect on the response. This approach is generally applied to an ordinal scale variable with few responses and with theoretical justification. For the variable *NUMLANES*, for example, three indicator variables could be created (assuming the range of *NUMLANES* is from 1 to 4) as follows:

$$Ind_1 = \begin{cases} 1 & \text{if } NUMLANES = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Ind_2 = \begin{cases} 1 & \text{if } NUMLANES = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$Ind_3 = \begin{cases} 1 & \text{if } NUMLANES = 3 \\ 0 & \text{otherwise} \end{cases}$$

This variable is now expressed in the regression as three indicator variables, one for each of the first three levels of the variable. As before, linear regression is used to assess the evidence whether each level of the variable *NUMLANES* deserves a separate beta parameter. This is evaluated statistically using an *F* test described later in this chapter.

3.6.1 Interactions in Regression Models

Interactions in regression models represent a combined or synergistic effect of two or more variables. That is, the response variable depends on the joint

values of two or more variables. A second-order effect is an interaction between two variables, typically denoted as X_1X_2 . A third-order effect is an interaction between three variables, and so on. As a general rule, the lower the order of effect (first order being a noninteracted variable), the more influence the variable has on the response variable in the regression. In addition, higher-order effects are generally included in the regression only when their lower-order counterparts are also included. For example, the second-order interaction X_1X_2 in the regression would accompany the first-order effects X_1 and X_2 .

To arrive at some understanding of the impact of including indicator variables and interaction variables, assume that X_1 is a continuous variable in the model. Also suppose that there are three indicator variables representing three levels of a nominal scale variable, denoted X_2 through X_4 in the model. The following estimated regression function is obtained (the subscript i denoting observations is dropped from the equation for convenience)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 \quad (3.38)$$

Careful inspection of the estimated model shows that, for any given observation, at most one of the new indicator variables remains in the model. This artifact is a result of the 0–1 coding: when $X_2 = 1$ then X_3 and X_4 are equal to 0 by definition. Thus, the regression model is expanded to four different models, which represent the four different possible levels of the indicator variable, such that

$$\text{level 1: } X_2 = X_3 = X_4 = 0$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$\text{level 2: } X_2 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X_1$$

$$\text{level 3: } X_3 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3 = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 X_1$$

$$\text{level 4: } X_4 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_4 X_4 = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 X_1$$

Inspection of these regression models reveals an interesting result. First, depending on which of the indicator variables is coded as 1, the slope of the regression line with respect to X_1 remains fixed, while the Y-intercept parameter changes by the amount of the parameter of the indicator variable. Thus, indicator variables, when entered into the regression model in this simple form, represent marginal adjustments to the Y-intercept term.

Suppose that each indicator variable is thought to interact with the variable X_1 . That is, each level of the indicator variable has a unique effect on Y when interacted with X_1 . To include these interactions in the model, the former model is revised to obtain

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_2 X_1 + \hat{\beta}_6 X_3 X_1 + \hat{\beta}_7 X_4 X_1 \quad (3.39)$$

The difference between this regression model and Equation 3.38 is that each indicator is entered in the model twice: as a stand-alone variable and interacted with the continuous variable X_1 . This more complex model is expanded to the following set of regression equations:

$$\text{level 1: } X_2 = X_3 = X_4 = 0$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$\text{level 2: } X_2 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_5 X_2 X_1 = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_5) X_1$$

$$\text{level 3: } X_3 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3 + \hat{\beta}_6 X_3 X_1 = (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_6) X_1$$

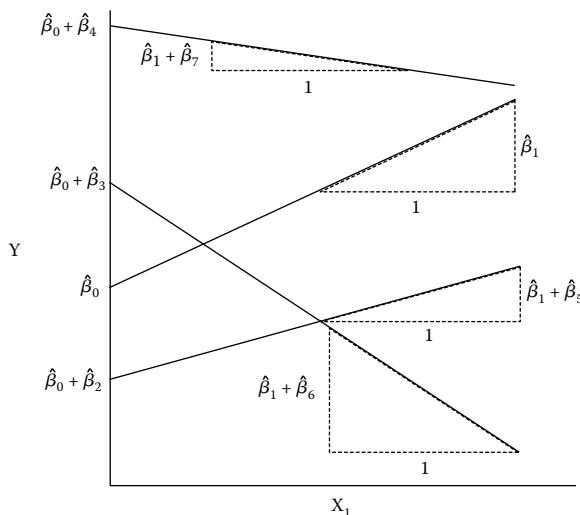
$$\text{level 4: } X_4 = 1$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_4 X_4 + \hat{\beta}_7 X_4 X_1 = (\hat{\beta}_0 + \hat{\beta}_4) + (\hat{\beta}_1 + \hat{\beta}_7) X_1$$

Each level of the indicator variable now has an effect on both the Y -intercept and slope of the regression function with respect to the variable X_1 . The net effect of this model is that each level of the nominal scale variable corresponds with a separate regression function with both a unique slope and Y -intercept. Figure 3.1 shows the graphical equivalent of the regression function for the model previously described with three levels of an indicator coded, allowing both individual slopes and Y -intercepts to vary for the three levels and the one base condition. The estimated regression parameters $\hat{\beta}_2$ through $\hat{\beta}_7$ are positive or negative, and the signs and magnitudes of the regression functions depend on their magnitudes relative to the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. The parameters $\hat{\beta}_2$ through $\hat{\beta}_4$ represent estimated marginal effects of the variables X_2 through X_4 on the Y -intercept, while the parameters $\hat{\beta}_5$ through $\hat{\beta}_7$ represent estimated interaction effects of the variables X_2 through X_4 with the continuous variable X_1 .

When interactions are between two or more continuous variables the regression equation becomes more complicated. Consider the following regression model, where variables X_1 and X_2 are continuous:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

**FIGURE 3.1**

Regression function with four levels of indicator variable represented.

Because the variables X_1 and X_2 do not represent a 0 or 1 as in the previous case, the model cannot be simplified as done previously. Notice that the response variable in this model is dependent not only on the individual values X_1 and X_2 , but also on the joint value X_1 times X_2 . This model contains an interaction term, which accounts for the fact that X_1 and X_2 do not act independently on Y . The regression function indicates that the relationship between X_1 and Y is dependent on the value of X_2 , and conversely, that the relation between X_2 and Y is dependent on the value of X_1 . The nature of the dependency is captured in the interaction term, X_1X_2 . There are numerous cases when the effect of one variable on Y depends on the value of one or more independent variables. Interaction terms often represent important aspects of a relationship and should be included in the regression. Although they may not explain a great deal of the variability in the response, they may represent important theoretical aspects of the relation and may also represent highly significant (statistically) effects.

3.7 Checking Regression Assumptions

Regression assumptions should be thought of as requirements. Refer to Table 3.1 to recall the requirements of the OLS regression model. Assumption one requires a linear-in-parameters relationship between the response and predictor variables. Assumption two is met as a consequence of OLS estimation.

Assumption three requires that the disturbances have the same variance across observations. Assumption four requires that there is not a temporal trend over observations. Assumption five requires that the X terms are exogenous or are determined by factors outside the regression model. Assumption six requires that the distribution of disturbances is approximately normal.

The majority of regression assumptions are examined using informal graphical plots. It is stressed that informal plots are simply methods for identifying gross, extreme, or severe violations of regression assumptions. In most cases, moderate violations of regression assumptions have a negligible effect, and so informal methods are adequate and appropriate. When cases are ambiguous, more formal statistical tests should be conducted. Each of the informal and some of the more formal tests are discussed in the following sections.

3.7.1 Linearity

Linearity is checked informally using several plots. These include plots of each of the independent variables on the X axis versus the disturbances (residuals) on the Y axis, and plots of model predicted (fitted) values on the X axis versus disturbances on the Y axis. If the regression model is specified correctly, and relationships between variables are linear, then these plots will produce plots that lack curvilinear trends in the disturbances. Gross, severe, and extreme violations are being sought, so curvilinear trends should be clearly evident or obvious.

Example 3.4

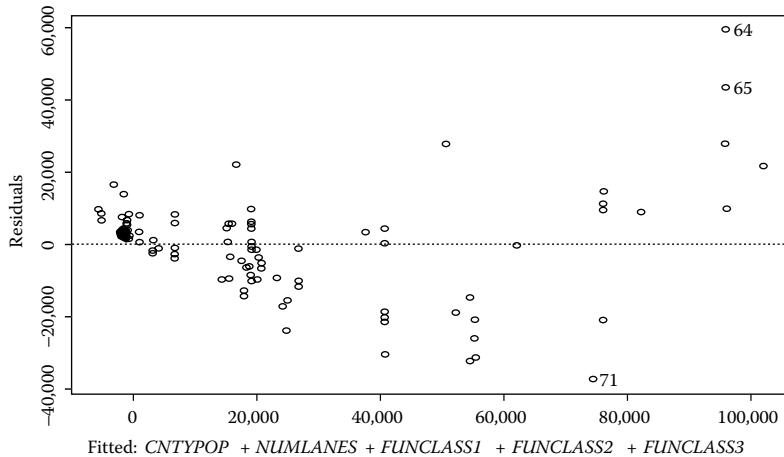
Consider the regression model from Example 3.2. The model depicted is clearly not the most desirable model. Two of the independent variables are statistically insignificant, and linearity may be in question. Figure 3.2 shows the model fitted values versus the model disturbances.

Inspection of the plot shows an evident curvilinear pattern in the disturbances. All disturbances at low predicted *AADT* are positive, the bulk of disturbances between 10,000 and 70,000 *AADT* are negative, and the disturbances after 80,000 *AADT* are positive. This “U-shape” of disturbances is a clear indicator of one or more nonlinear effects in the regression.

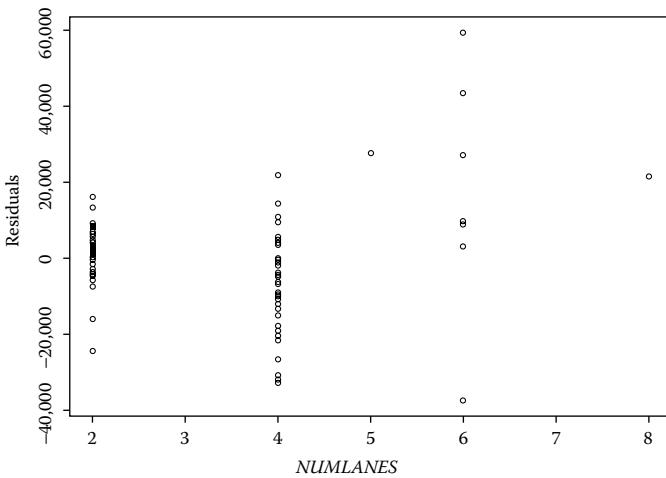
To determine which independent variable(s) is contributing to the nonlinearity, plots of individual independent variables vs. the model disturbances are constructed and examined. After inspection of numerous plots the culprit responsible for the nonlinearity is determined to be the variable *NUMLANES*. Figure 3.3 shows the variable *NUMLANES* versus model disturbances for the model described in Example 3.2.

The nonlinearity between *NUMLANES* and model disturbances suggests that treating the variable *NUMLANES* as a continuous variable may not be the best specification of this variable. Clearly, the variable *NUMLANES* has an effect on *AADT*; however, this effect is probably nonlinear.

Indicator variables for categories of *NUMLANES* are created, and insignificant levels of *FUNCLASS* are removed from the regression. The resulting model

**FIGURE 3.2**

Fitted values versus disturbances (Example 3.3).

**FIGURE 3.3**

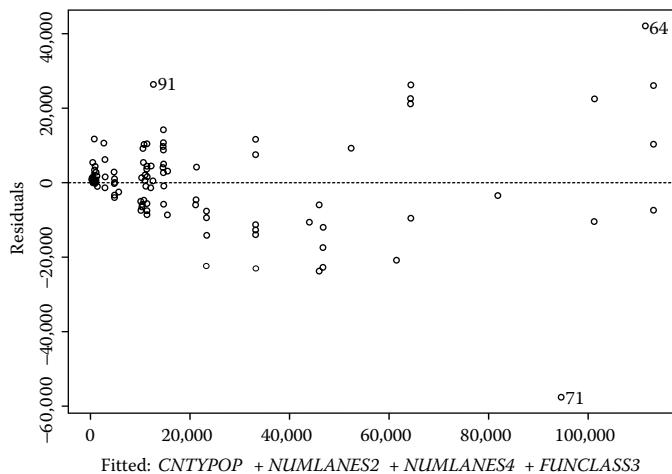
NUMLANES versus disturbances (Example 3.3).

is shown in Table 3.5. The disturbances for the improved model are shown in Figure 3.4. Although there is still a slight U-shaped pattern to the disturbances, the effect is not as pronounced and both positive and negative disturbances are observed along the fitted regression line. Several observations that are outlying with respect to the bulk of the data are noted: observations 71, 91, and 64. A plot of $NUMLANES$ versus the new model disturbances, shown in Figure 3.5, reveals a more acceptable pattern of disturbances, which are distributed around 0 rather randomly and evenly.

TABLE 3.5

Least Squares Estimated Parameters (Example 3.2)

Parameter	Parameter Estimate	Standard Error of Estimate	t-Value	P(> t)
Intercept	58,698.910	5,099.605	11.510	<.0001
CNTYPOP	0.025	0.004	5.859	<.0001
NUMLANES2	-5,8718.141	5,134.858	-11.435	<.0001
NUMLANES4	-48,867.728	4,685.006	-10.431	<.0001
FUNCLASS3	31,349.211	3,689.281	8.497	<.0001

**FIGURE 3.4**

Fitted values versus disturbances (Example 3.5).

Gross departures from the linearity assumption are detected through inspection of disturbance plots. A nonlinear relationship between the response Y and independent variables is revealed through curvilinear trends in the plot. Nonlinear shapes that appear in these plots include "U," inverted "U," "waves," and arcs. Remedial measures to improve nonlinearities through transformation of variable measurement scales are described in Appendix D.

3.7.2 Homoscedastic Disturbances

Constancy of disturbances is called homoscedasticity. When disturbances are not homoscedastic, they are said to be heteroscedastic. This requirement is derived from the variance term in the regression model, which is assumed to be constant over the entire regression. It is possible, for example, to have variance that is an increasing function of one of the independent

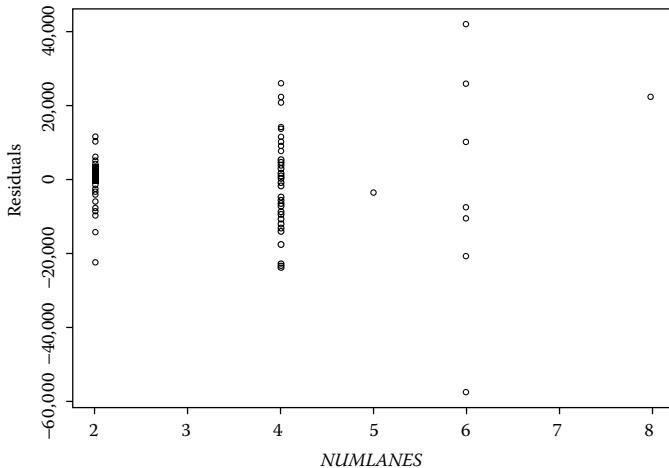


FIGURE 3.5
NUMLANES versus disturbances (Example 3.5).

variables, but this is a violation of the OLS regression assumption and requires special treatment. The consequence of a heteroscedastic regression is reduced precision of beta parameter estimates. That is, regression parameters are less efficient with heteroscedastic compared to homoscedastic disturbances. This finding is intuitive. Since MSE is used to estimate σ^2 , MSE is larger for a heteroscedastic regression. It follows that MSE used in the calculation of the precision of beta parameters will also result in a larger estimate.

Scatter plots are used to assess homoscedasticity. A plot of model fitted values versus disturbances is typically inspected first. If heteroscedasticity is detected, then plots of the disturbances versus independent variables should be conducted to identify the culprit.

Example 3.5

The *AADT* model estimated in Example 3.4 was an improvement over the original model presented in Example 3.2. The apparent nonlinearity is improved, and all model variables are statistically significant and theoretically meaningful. The assumption of homoscedasticity is checked by inspecting the disturbances. The plot in Figure 3.6 shows the plot of disturbances versus model fitted values for the model estimated in Example 3.4. The horizontal lines shown on the plot show a fairly constant band of equidistant disturbances around the regression function. In other words, the disturbances do not become systematically larger or smaller across fitted values of Y, as required by the regression. With the exception of observation numbers 64 and 71, most disturbances fall within these bands. Inspection of the plot reveals lack of a severe, extreme, or significant violation of homoscedasticity.

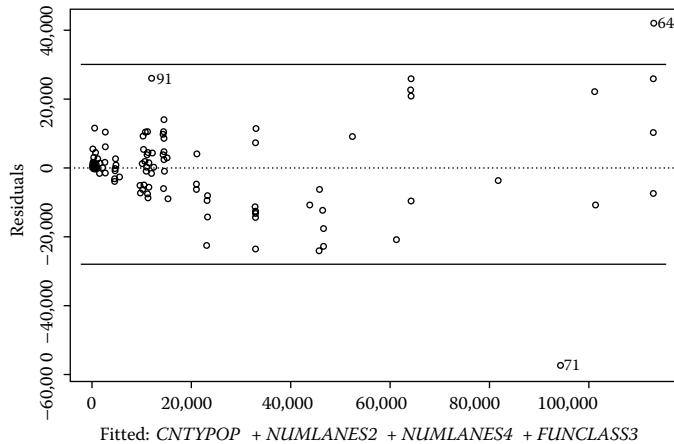


FIGURE 3.6
NUMLANES versus disturbances (Example 3.5).

Often heteroscedasticity is easily detected. In many applications, disturbances that are an increasing function of fitted values of Y are often encountered. This finding simply means that there is greater variability for higher values of Y than for smaller values of Y. Consider a hypothetical model of household level daily trip generation as a function of many sociodemographic independent variables. The variability in trip making for households that make many trips, on average, is much larger than households that make few daily trips. Some high income households, for example, may make extra shopping and recreational trips, whereas similar households may be observed on days when few of these additional nonessential trips are made. This travel behavior results in large variability for those households. Households that depend on transit, in contrast, may not have the financial means or the mobility to exhibit the same variability in day-to-day trip making.

Remedial measures for dealing with heteroscedasticity include transformations on the response variable Y, WLSs, and ridge regression and generalized least squares. Only the first of these, transforming Y, is accomplished within the OLS regression framework. Care must be taken not to improve one situation (heteroscedasticity) at the expense of creating another, such as nonlinearity. Fortunately, fixing heteroscedasticity in many applications also tends to improve nonlinearity. WLS regression is a method used to increase the precision of beta parameter estimates and requires a slight modification to OLS regression. Ridge regression is a technique used to produce biased but efficient estimates of beta parameters. Generalized least squares is presented in Chapter 5.

3.7.3 Uncorrelated Disturbances

Correlated disturbances can result when observations are dependent across individuals, time, or space. For example, traffic volumes recorded every 15 minutes are typically correlated across observations. Traffic volumes recorded at the same time each day would also be correlated. Other examples include observations on people made over time and a survey response to a question posed each year to the same household. Observations can also be correlated across space. For example, grade measurements recorded every 20 feet from a moving vehicle would be correlated.

Correlation of disturbances across time is called serial correlation. The standard plot for detecting serial correlation is a plot of disturbances versus time, or a plot of disturbances versus ordered observations (over space). Approximately normal, homoscedastic, and nonserially correlated disturbances will not reveal any patterns in this plot. Serially correlated disturbances will reveal a trend over time, with peaks and valleys in the disturbances that typically repeat themselves over fixed intervals. A familiar example is the plot of the voltage generated by a person's heart over time. Each beat of the heart results in high voltage output, followed by low voltage readings between heartbeats.

A formal test for autocorrelation is based on the Durbin–Watson statistic. This statistic is provided by most regression software and is calculated from the disturbances of an OLS regression. When no autocorrelation is present the Durbin–Watson statistic will yield a value of 2.0. As the statistic gets farther away from 2.0 one becomes less confident that no autocorrelation is present. Additional details of this test are provided in Greene (2000) and in Chapter 4.

Standard remedial measures for dealing with serial correlation are generalized least squares or time-series analysis, both of which explicitly account for the correlation among disturbances. Chapter 5 presents a discussion of generalized least squares. Chapter 7 presents details of time-series analysis, provides examples of disturbances plots, and describes when time-series analysis is an appropriate substitute for OLS regression.

3.7.4 Exogenous Independent Variables

Assumption number 5 in Table 3.1 is often called the exogeneity assumption, meaning that all independent variables in the regression are exogenous. Exogeneity may be one of the most difficult assumptions in the regression to describe and understand. The phrase "a variable is exogenous" means that the variable is expected to vary "autonomously," or independently of other variables in the model. The value of an exogenous variable is determined by factors outside the model. Alternatively, an endogenous variable is a variable whose variation is caused by other exogenous or endogenous variables in the model (one or more of the independent variables, or the dependent variable).

Exogeneity does not mean that two independent variables cannot covary in the regression; many variables do covary, especially when the regression is estimated on observational data. In the absence of a formal theory between them, correlated variables in the regression are assumed to be merely associated with one another and not causally related. If a causal relationship exists between independent variables in the model, then an alternative model structure is needed, such as a multiequation model (see Chapter 5) or a structural equation model (see Chapter 9). When the “feedback effects” caused by endogeneity are ignored statistical inferences are biased.

As an example of endogeneity, consider the following hypothetical linear regression model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \varepsilon$$

Y = current military expenditures, \$ (millions)

X_1 = number of past military conflicts

X_2 = gross domestic product, \$ (millions)

Recall that Y is to be influenced by variables X_1 and X_2 , and under the exogeneity assumption X_1 and X_2 are influenced by factors outside of the regression. Past military conflicts X_1 presumably would directly influence current military expenditures because a history of military conflicts would encourage a future military preparedness. However, the direction of causality might also be in the reverse direction—being prepared militarily might enable a country to become involved in more conflicts, since, after all, it is prepared. In this example the variable X_1 in this model is said to be endogenous. In the case of endogenous variable X_1 , the variation in Y not explained by X_1 is ε . Under endogeneity, however, ε also contains variation in X_1 not explained by Y —or the feedback of Y on X_1 .

To see the problem caused by endogeneity, consider a case where the OLS regression model and equations are given, respectively, by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad \text{and} \quad Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

Assume that X is endogenous, and is a function of Y ; thus $X = Y\kappa + e^*$. The variable X consists of a systematic component, which is a function of Y and a stochastic component e^* . Clearly, the variable X is correlated with the disturbance term e^* . However, the OLS regression equation does not contain a separate disturbance term e^* , and instead the term e contains both the unexplained variation of X on Y and Y on X . If the effect of endogeneity is

ignored, the OLS regression parameter is estimated as usual (see Equation 3.17), where the variables X and Y are standardized

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

Evaluating the estimator for bias, and using the fact that the covariance of two variables X and ε is given as

$$COV(X, \varepsilon) = E[(X - E[X])(\varepsilon - E[\varepsilon])] = E[X\varepsilon] = \sum_{i=1}^n X_i \varepsilon_i$$

and the variance of standardized X is

$$VAR(X) = \sum_{i=1}^n X_i^2$$

then

$$\begin{aligned} E[\hat{\beta}_i] &= E\left[\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right] = E\left[\frac{\sum_{i=1}^n X_i (X_i \beta + \varepsilon_i)}{\sum_{i=1}^n X_i^2}\right] \\ &= E\left[\frac{\sum_{i=1}^n X_i^2 \beta + \sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2}\right] = E\left[\frac{\sum_{i=1}^n X_i^2 \beta}{\sum_{i=1}^n X_i^2} + \frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2}\right] \\ &= \beta + E\left[\frac{COV(X_i, \varepsilon_i)}{VAR(X)}\right] = \beta + Bias \end{aligned} \quad (3.40)$$

As depicted in Equation 3.40, when the covariance between X and ε is zero the bias term disappears, which is the case under exogeneity. When endogeneity is present, however, the covariance between X and ε is nonzero and the least squares estimate is biased. The direction of bias depends on the covariance between X and ε . Because the variance of X is always positive, a negative covariance will result in a negative bias, whereas a positive covariance will result in positive bias. In rare cases when the variance of X dwarfs the magnitude of the covariance, the bias term is neglected.

3.7.5 Normally Distributed Disturbances

An assumption imposed simply to allow regression parameter inferences to be drawn is that disturbances are approximately normally distributed.

In addition, the disturbances, or residuals, have an expected value of zero. Recall that the assumption of normality led to regression parameters that are approximately t distributed. It is the t distribution that allows probabilistic inferences about the underlying true parameter values. If making inferences is not important, then the assumption of normality is ignored without consequence. For example, if the prediction of future mean responses is the only modeling purpose, then ignoring normality will not hinder the ability to make predictions.

In addition to future predictions, however, interest is often also focused on inferences about the values of beta parameters, and so the normality assumption must be checked. There are nongraphical, graphical, and nonparametric methods for assessing normality. The degrees to which these tools are used to assess normality depend on the degree of satisfaction obtained using more informal methods. If graphical methods suggest serious departures from normality, more sophisticated methods might be used to test the assumption further. The various methods include the following:

1. Summary statistics of the disturbances, including minimum, first and third quartiles, median, and maximum values of the disturbances. Recall that the normal distribution is symmetric and the mean and median should be equivalent and approximately equal to zero. Thus, inspection of the summary statistics for symmetry about zero is useful. Most statistical software packages either provide summary statistics of the disturbances by default or generate the statistics with simple commands.
2. Histograms of the disturbances. A histogram of the disturbances should reveal the familiar bell-shaped normal curve. There should be an approximately equivalent number of observations that fall above and below zero, and they should be distributed such that disturbances above and below zero are mirror images of each other. Most statistical software packages can produce histograms of the disturbances rather easily. Histograms show symmetry but do not provide information about the appropriate allocation of probability in the tails of the normal distribution.
3. Normal probability quantile–quantile (Q–Q) plots of the disturbances. The normal distribution has approximately 67% of the observations within 1 standard deviation above and below the mean (0 in the case of the disturbances), 95% within two standard deviations, and 99% within three standard deviations. Normal Q–Q plots are constructed such that normally distributed disturbances will plot on a perfectly straight line. Departures from normality will reveal departures from the straight-line fit to the Q–Q plot. It is through cumulative experience plotting and examining Q–Q plots that an ability to detect extreme, severe, or significant departures from

normality is developed. Taken together with summary statistics and histograms, Q–Q plots provide sufficient evidence on whether to accept or reject approximate normality of the disturbances.

- Nonparametric methods such as the chi-square goodness-of-fit (GOF) test or the Kolmogorov–Smirnov GOF test. In particular when sample sizes are relatively small, these methods might seem to reveal serious departures from normality. Under such circumstances statistical tests that assess the consistency of the disturbances with normality are performed. In essence, both the chi-square GOF test and the Kolmogorov–Smirnov GOF test provide probabilistic evidence of whether the data were likely to have been drawn from a normal distribution. Readers wishing to pursue these simple tests should refer to Chapter 2, or consult Neter et al. (1996) or Conover (1980).

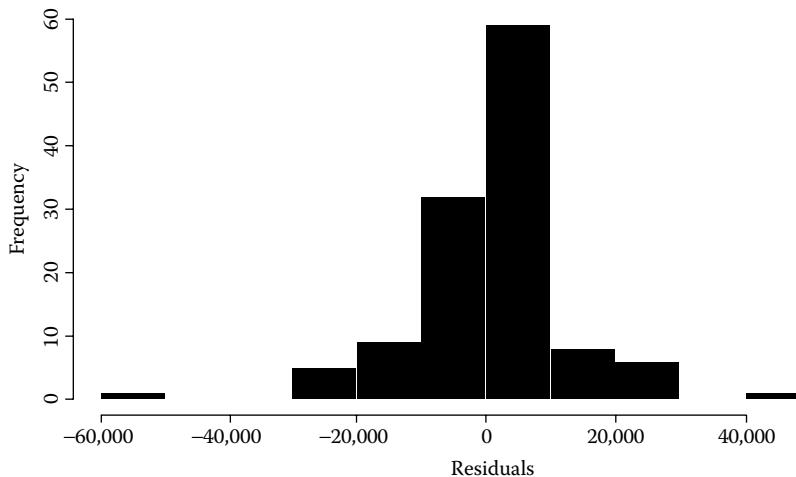
Sample size directly influences the degree to which disturbances exhibit “normal” behavior. Normally distributed small samples can exhibit apparently nonnormal behavior as viewed through diagnostic tools. Disturbances obtained from sample sizes less than 30 should be expected to raise difficulties in assessing the normality assumption. On the other hand, disturbances obtained from large sample sizes (e.g., greater than 100) should reveal behavior consistent with normality if indeed the disturbances are normal.

Example 3.6

Consider the regression model in Example 3.4. A check of the homoscedasticity assumption produced no evidence to reject it. Because the data were collected at different points in space and not over time, serial correlation is not a concern. The normality assumption is now checked. Inspection of standard summary statistics provides an initial glimpse at the normality assumption.

Minimum	1 st Quartile	Median	3 rd Quartile	Maximum
–57,878	–4,658	671.5	3,186	42,204

There are 121 observations in the sample—a relatively large sample. The minimum value is –57,878, a fairly large negative number, whereas the maximum is only 42,204. Symmetry of the disturbances may be in question, or perhaps there are a handful of observations that are skewing the “tails” of the disturbances distribution. Similarly, the 1st and 3rd quartiles are of different magnitudes, almost by 25%, and so symmetry is again in question. The median should be zero, but is less than 1% of the maximum disturbance and so does not appear to raise significant doubt about symmetry. In addition, potential skew appears to be associated with the negative disturbances in contrast to the median, which exhibits positive skew. In summary, the evidence is mixed and inconclusive, but certainly raises some doubts regarding normality.

**FIGURE 3.7**

Histogram of residuals (model in Example 3.4).

A histogram of the disturbances is shown in Figure 3.7. The Y axis shows the number of observations, and the X axis shows bins of disturbances. Disturbances both positive and negative appear to be outlying with respect to the bulk of the data. The peak of the distribution does not appear evenly distributed around zero, the expected result under approximate normality. What cannot be determined is whether the departures represent serious, extreme, and significant departures, or whether the data are consistent with the assumption of approximate normality. In aggregate the results are inconclusive.

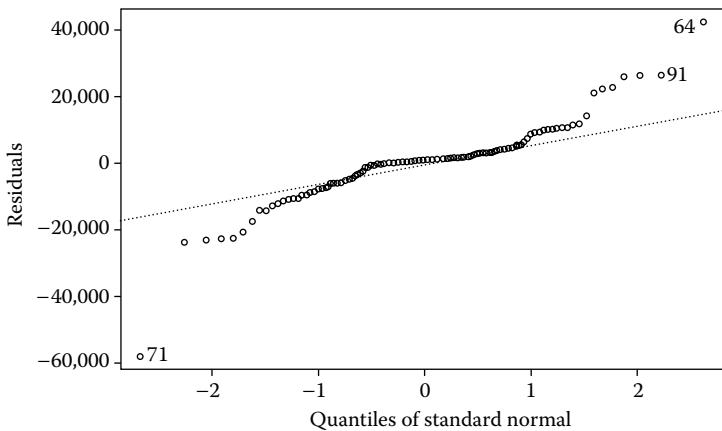
A normal Q–Q plot is shown in Figure 3.8. When disturbances are normally distributed the Q–Q plot reveals disturbances that lie on the dotted straight line shown in the figure.

Inspection of the figure shows that the tails of the disturbances' distribution appear to depart from normality. Several observations, specifically observations 71, 91, and 64, appear to seriously depart from normality, as shown in Figure 3.8.

On the basis of the cumulative evidence from the summary statistics, the histogram, and the normal Q–Q plot, the statistical model can probably be improved. Outlying observations, which should be examined, could be partly responsible for the observed departures from normality.

3.8 Regression Outliers

Identifying influential cases is important because influential observations may have a disproportionate affect on the fit of the regression line. If an influential observation is outlying with respect to the underlying relationship for some known or unknown reason, then this observation serves to

**FIGURE 3.8**

Q–Q plot of residuals (model in Example 3.4).

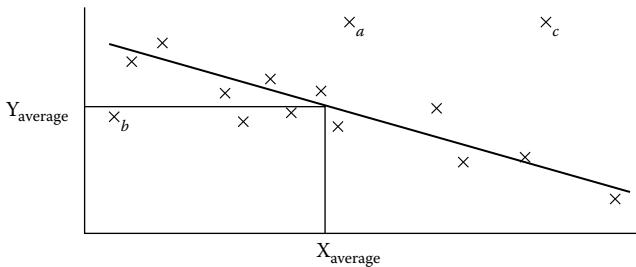
misrepresent the true relationship between variables. As a result, the risk is that an “errant” observation dominates the fitted regression line and thus influences inferences drawn. It is recommended that outlying influential observations be systematically checked to make sure they are legitimate observations.

An observation is influential with respect to its position on the Y axis, its position on the X axis, or both. Recall that the sum of squared disturbances is minimized using OLS. Thus, an observation two times as far from the fitted regression line as another will receive four times the weight in the fitted regression. In general, observations that are outlying with respect to both \bar{X} and Y have the greatest influence on the fit of the regression and need to be scrutinized. Figure 3.9 shows a hypothetical regression line fit to data with three outlying observations. Observation *a* in the figure is relatively distant from \bar{Y} . Similarly, observation *b* is relatively distant from \bar{X} , and observation *c* is relatively distant from both \bar{X} and \bar{Y} .

3.8.1 The Hat Matrix for Identifying Outlying Observations

A matrix called the “hat” matrix is often used to identify outlying observations. Recall that the matrix solution to the least squares regression parameter estimates is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Fitted values of Y are given as $\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. The hat matrix \mathbf{H} is defined such that $\hat{\mathbf{Y}} = \mathbf{H} \mathbf{Y}$. Thus, the $n \times n$ hat matrix is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (3.41)$$

**FIGURE 3.9**

Regression line with three influential observations a , b , and c .

Fitted values of Y_i are expressed as linear combinations of the observed Y_i through the use of the hat matrix. The disturbances e can also be expressed as linear combinations of the observations \mathbf{Y} using the hat matrix

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad (3.42)$$

Because $VAR[\mathbf{Y}] = \sigma^2$, it is straightforward to show that the variance–covariance matrix is expressed in terms of the hat matrix as

$$VAR[e] = \sigma^2(\mathbf{I} - \mathbf{H}) \quad (3.43)$$

Finally, with subscripting i for observations, the variance of disturbances e_i is computed as

$$VAR[e_i] = \sigma^2 [1 - h_{ii}] \quad (3.44)$$

where h_{ii} is the i^{th} element on the main diagonal of the hat matrix. This element is obtained directly from the relation $h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X}) \mathbf{x}_i$, where \mathbf{x}_i corresponds to the \mathbf{H}_i vector derived from the i^{th} observation (row) in \mathbf{X} . The diagonal elements in h_{ii} possess useful properties. In particular, their values are always between 0 and 1, and their sum is equal to p , the number of regression model parameters.

Some interesting observations are noteworthy regarding the h_{ii} contained in the hat matrix

1. The larger is h_{ii} , the more important Y_i is in determining the fitted value \hat{Y}_i .
2. The larger is h_{ii} , the smaller is the variance of the disturbance e_i . This effect is seen from the relation $VAR[e_i] = \sigma^2[1 - h_{ii}]$. In other words, high leverage values have smaller disturbances.

3. As a rule of thumb, a leverage value is thought to be large when any specific leverage value is greater than twice the average of all leverage values. It can easily be seen that the average value is p/n , where p is the number of parameters and n is the number of observations.

3.8.2 Standard Measures for Quantifying Outlier Influence

Perhaps the most common measure used to assess the influence of an observation is Cook's distance D . Cook's distance quantifies the impact of removal of each observation from the fitted regression function on estimated parameters in the regression function. If the effect is large, then D_i is also large. Thus, Cook's distance provides a relative measure of influence for observations in the regression.

In algebraic terms, Cook's distance measure is given by (see Neter et al. 1996)

$$D_i = \frac{\sum_{j=1}^n (Y_j - \hat{Y}_{j(i)})^2}{p(MSE)} = \frac{e_i^2}{p(MSE)} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right] \quad (3.45)$$

The first form in Equation 3.45 considers the sum of squared observed minus fitted values with the i^{th} case deleted from the regression. This computation is burdensome because it requires computing the sum of squares for each of the i observations. The second equivalent form shown in Equation 3.45 makes use of the hat matrix discussed previously and simplifies computation considerably. The diagonals in the hat matrix are computed and used along with the i^{th} disturbance to determine the Cook's distance value across observations. An equivalent form of Cook's distance in matrix form is given as

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T (\hat{Y} - \hat{Y}_{(i)})}{p(MSE)} \quad (3.46)$$

Cook's distance depends on two factors, the magnitudes of the i^{th} disturbance and the leverage value h_{ii} . Larger disturbances, larger leverage values, or both lead to large values of D .

Although Cook's distance is a commonly applied measure in statistical software packages, there are other measures for assessing the influence of observations. All methods use similar principles for assessing the influence of outliers (see Neter et al. 1996).

3.8.3 Removing Influential Data Points from the Regression

So far, methods for quantifying influential observations have been discussed. There are many possible scenarios that could give rise to outliers. The following

describes how these scenarios arise and appropriate actions that should be taken.

1. Misspecification error. A specified model may be inappropriate and fail to account for some important effects, particularly with respect to influential cases. For example, an observation or group of observations could possess a trait (effect) in common that is omitted from the regression model and whose effect, therefore, is not estimated. If the omitted trait is available or subsequently measurable, it should be included in the model. Another misspecification error might be a nonlinearity in the regression—this error could be improved by applying transformations or by using a nonlinear regression technique. In cases where an outlying observation(s) is suspected of possessing (sharing) a trait that cannot be identified, a difficult decision remains. Removing the observation(s) from the regression begs for outside criticism, with the accusation that “the data were fit to the model.” Leaving the observation(s) in the model results in an apparently unsatisfactory model that is open to criticism for “lack of fit.” Perhaps the most defensible course of action in these cases is to leave the outlying observations in the model and provide possible explanations for their being anomalous. If the outlying observations are also influential on the fit of the regression, robust regression techniques are applied to reduce their relative influence, such as absolute deviation regression.
2. Coding error. An influential data point (or points) was recorded incorrectly during data collection. The interim step between observation of a data point and recording of that observation often presents an opportunity for a recording or coding error. For example, an observer writing down radar-detected maximum speeds on a field data-collection sheet might accidentally record a 56 (mph) when the true reading was 65. In documented cases when a coding error is probable, usually through examination of data on the collection forms, erasures on the forms, illegibility, and so on, removal of the data point(s) from the regression is justified.
3. Data-collection error. Influential observations were the result of malfunctioning equipment, human error, or other errors that occurred during data collection. For example, an in-vehicle “black box” that records continuous vehicle speed may not have been calibrated correctly, and thus collects “bad” data during a data-collection period. The outlier resulting from this type of error is often difficult to diagnose, as in some cases there is insufficient information to determine whether an error occurred. It is typically more defensible to leave these types of outliers in the model—as removing them from the regression without demonstrable reason reduces the credibility of the research.

4. Calculation error. Often there is a significant amount of data manipulation that occurs before analysis. Human error can easily creep into data manipulations—especially when these manipulations are not automated, leading to erroneous data. Typically, the best remedy for this situation is not to discard an outlier but instead fix the calculation and rerun the regression with the corrected observations.

Removing outliers from the regression should raise suspicion among those who critique the regression results. In effect, removal of outliers from a regression analysis results in an improved fit of the data to the model. The objective of regression in contrast is to have the model fit the data. Removing outliers without proper justification raises the possibility that data have been manipulated to support a particular hypothesis or model. As a result, complete documentation and justification for removing any and all data from an analysis is good practice.

Example 3.7

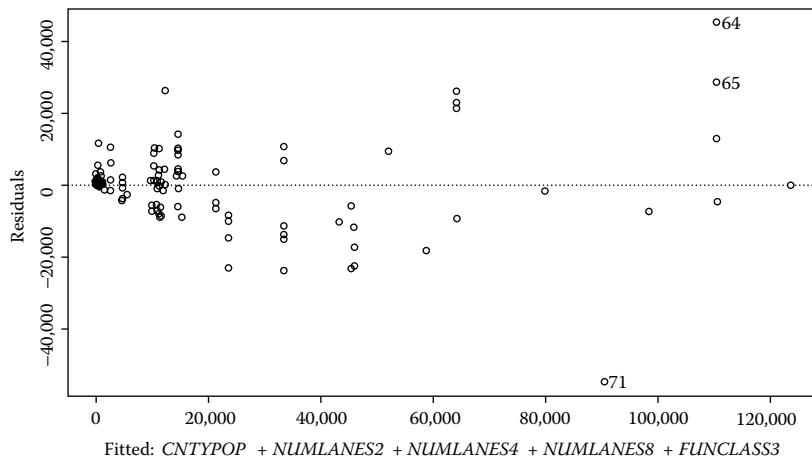
Consider the model estimated in Example 3.4 with a focus on outliers. Two plots are of prime interest—disturbances versus fitted values, and observations versus Cook's distance. Figure 3.10 shows a plot of disturbances versus fitted observations, and Figure 3.11 shows observations versus Cook's distance.

Three outliers or influential cases are identified as having disproportionate influence on the fit of the regression function: observations 58, 64, and 71. Having considerable influence on the fitted regression, these observations may mischaracterize the underlying “true” relationship if they are erroneous. An inquiry into the nature of observations 58, 64, and 71 is worthwhile. Table 3.6 shows summary statistics for the three outlying observations. Observations 58 and 64 are observations with relatively large observed *AADT* and *CNTYPOP* in comparison to the bulk of the data. All facilities were urban interstates.

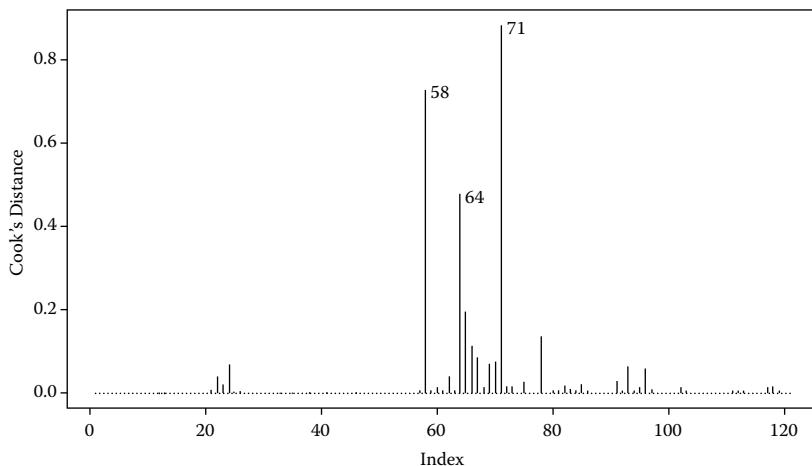
The field data-collection forms are examined to identify whether there were errors in coding, equipment problems, or any other extenuating circumstances. A comment in the field notes regarding observation 64 suggests that there was some question regarding the validity of the *AADT* count, as the counting station had not been calibrated correctly. For observation 71 there is a series of erasures and corrections made to the *AADT* figure, and it is not clear whether the coded figure is correct. Errors or extenuating circumstances for observation 58 cannot be found. As a result observations 64 and 71 are removed from the analysis, leaving observation 58 as a valid observation.

A reestimation of the model after removing observations 64 and 71 and entering the variable *NUMLANES8*, which is now statistically significant, is shown in Table 3.7. The model is similar to the previous models, although the disturbances have been improved compared to the previous model. A histogram of the disturbances of the revised model, shown in Figure 3.12, reveals a more balanced bell-shaped distribution.

Although there are outlying observations with respect to the bulk of observations, they cannot be removed without proper justification. Additional variables

**FIGURE 3.10**

Disturbances versus fitted observation (model in Example 3.4).

**FIGURE 3.11**

Observations versus Cook's distance, D (model in Example 3.4).

that might influence the *AADT* on various facilities—such as characteristics of the driving population, the nature of cross-traffic such as driveways from industrial, commercial, and residential areas (for noninterstates only), and perhaps roadway geometric and operational characteristics—are missing from the set of explanatory variables. However, a fairly reliable predictive tool for forecasting *AADT* given predictor variables in the model has been obtained.

TABLE 3.6

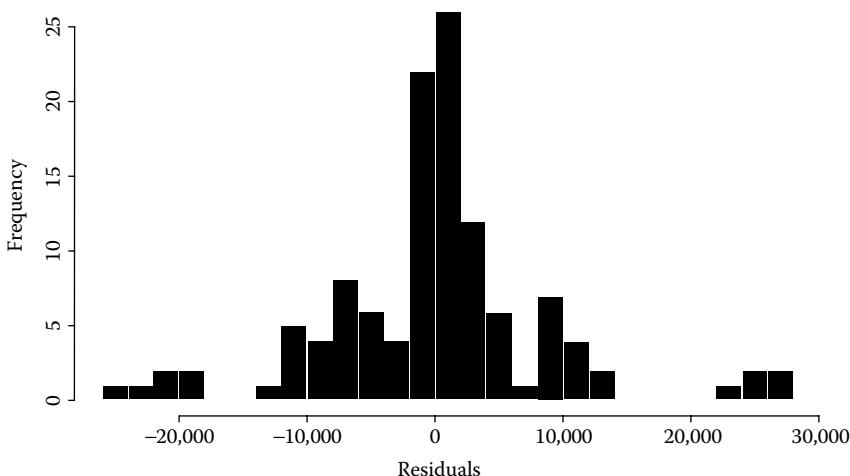
Assessment of Outliers (Example 3.6)

Variable	Mean	Median	Observation		
			58	64	71
AADT	19,440	8,666	123,665	155,547	36,977
CNTYPOP	263,400	113,600	459,784	941,411	194,279
NUMLANES2	3.099	2.000	8	6	6
FUNCLASS3	2.727	2.000	3	3	3

TABLE 3.7

Least Squares Estimated Parameters (Example 3.6)

Parameter	Parameter Estimate	Standard Error of Estimate	t-Value	P(> t)
Intercept	59,771.42	4,569.86	13.08	<.0001
CNTYPOP	0.02	0.00	7.22	<.0001
NUMLANES2	-59,274.88	4,569.13	-12.97	<.0001
NUMLANES4	-48,875.17	4,269.35	-11.45	<.0001
NUMLANES8	2,2261.20	10,024.23	2.22	.0284
FUNCLASS3	31,841.77	2,997.36	10.62	<.0001
R-squared		0.89		
F-statistic	192.1 on 5 and 113 degrees of freedom, the p-value is <.0001			

**FIGURE 3.12**

Histogram of disturbances (Example 3.6).

3.9 Regression Model GOF Measures

Goodness-of-fit statistics are useful for comparing the results across multiple studies, for comparing competing models within a single study, and for providing feedback on the extent of knowledge about the uncertainty involved with the phenomenon of interest. Three measures of model GOF are discussed: R -squared, adjusted R -squared, and the generalized F test.

To develop the R -squared GOF statistic, some basic notions are required. Sum of squares and mean squares are fundamental in both regression and analysis of variance. The sum of square errors (disturbances) is given by

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

the regression sum of squares is given by

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

and the total sum of squares is given by

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

The SSE is the variation of the fitted regression line around the observations. The SSR is the variation of the fitted regression line around \bar{Y} , and SST is the total variation—the variation of each observation around \bar{Y} . It also is straightforward to show algebraically that $SST = SSR + SSE$.

Mean squares are just the sum of squares divided by their degrees of freedom. SST has $n - 1$ degrees of freedom, because 1 degree of freedom is lost in the estimation of \bar{Y} . SSE has $n - p$ degrees of freedom, because p parameters are used to estimate the fitted regression line. Finally, SSR has $p - 1$ degrees of freedom associated with it. As one would expect, the degrees of freedom are additive such that $n - 1 = n - p + p - 1$. The mean squares, then, are $MSE = SSE/(n - p)$ and $MSR = SSR/(p - 1)$. The coefficient of determination, R -squared, is defined as

$$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (3.47)$$

R^2 is the proportionate reduction of total variation accounted for by the independent variables (X). It is commonly interpreted as the proportion of total variance explained by X . When $SSE = 0$, $R^2 = 1$, and all of the variance is

explained by the model. When $\text{SSR} = 0$, $R^2 = 0$, and there is no association between \mathbf{X} and \mathbf{Y} .

Because R^2 can only increase when variables are added to the regression model (SST stays the same, and SSR can only increase even when statistically insignificant variables are added), an adjusted measure, R^2_{adjusted} , is used to account for the degrees of freedom changes as a result of different numbers of model parameters, and allows for a reduction in R^2_{adjusted} as additional, potentially insignificant variables are added. The adjusted measure is considered to be superior for comparing models with different numbers of parameters. The adjusted coefficient of multiple determination is

$$R^2_{\text{adjusted}} = 1 - \frac{\text{SSE}/n-p}{\text{SST}/n-1} = 1 - \left(\frac{n-1}{n-p} \right) \frac{\text{SSE}}{\text{SST}} \quad (3.48)$$

The following guidelines should be applied:

- The R^2 and R^2_{adjusted} measures provide only relevant comparisons with previous models that have been estimated on the phenomenon under investigation. Thus, an R^2_{adjusted} of 0.40 in one study may be considered “good” only if it represents an improvement over similar studies and the model provides new insights into the underlying data-generating process. Thus, it is possible to obtain an improvement in the R^2 or R^2_{adjusted} value without gaining a greater understanding of the phenomenon being studied. It is only the combination of a comparable R^2_{adjusted} value and a contribution to the fundamental understanding of the phenomenon that justifies the claim of improved modeling results.
- The absolute values of R^2 and R^2_{adjusted} measures are not sufficient measures to judge the quality of a model. Thus, an R^2 of 0.20 from a model of a phenomenon with a high proportion of unexplained variation might represent a breakthrough in the current level of understanding, whereas an R^2 of 0.90 of another phenomenon might reveal no new insights or contributions. Thus, it is often better to explain a little of a lot of total variance rather than a lot of a little total variance.
- Relatively large values of R^2 and R^2_{adjusted} are often caused by data artifacts. Small variation in the independent variables can result in inflated values. This effect is particularly troublesome if the model is needed for predictions outside the range of the independent variables. Extreme outliers can also inflate R^2 and R^2_{adjusted} values.
- The R^2 and R^2_{adjusted} assume a linear relation between the response and predictor variables, and can give grossly misleading results if the relation is nonlinear. In some cases R^2 could be relatively large

and suggest a good linear fit, when the true relationships are curvilinear. In other cases R^2 could suggest a poor fit when the relationships are nonlinear. This possible nonlinearity emphasizes the need to plot, examine, and become familiar with data before statistical modeling.

- The R^2 and R^2_{adjusted} values are bound by 0 and 1 only when an intercept term is included in the regression model. When the intercept is forced through zero, the R^2 and R^2_{adjusted} values can exceed the value 1 and more caution needs to be used when interpreting them.

Another measure for assessing model fit is the generalized F test. This general and flexible approach is for testing the statistical difference between competing models. First, a full or unrestricted model is estimated. This full model could contain ten independent variables, for example, or the “best” model specified to date. The full model is fit using the method of least squares and SSE is obtained—the sum of square errors for the full model. For convenience, the sum of square errors for the full model is denoted as

$$\text{SSE}_F = \sum_{i=1}^n (Y_i - \hat{Y}_{Fi})^2$$

where the predicted value of Y is based on the full model.

A reduced model is then estimated, which represents a viable competitor to the full model with fewer variables. This reduced model could contain nine independent variables, or none, leaving only the Y -intercept term B_0 . The sum of squared errors is estimated for the competing or reduced model, where

$$\text{SSE}_R = \sum_{i=1}^n (Y_i - \hat{Y}_{Ri})^2$$

The logic of the F test is to compare the values of SSE_R and SSE_F . Recall from the discussion of R -squared that SSE can only be reduced by adding variables into the model, thus $\text{SSE}_R \geq \text{SSE}_F$. If these two sum of square errors are the same, then the full model has done nothing to improve the fit of the model; there is just as much “lack of fit” between observed and predicted observations as with the reduced model, so the reduced model is superior. Conversely, if SSE_F is considerably smaller than SSE_R , then the additional variables add value to the regression by adding sufficient additional explanatory power. In the generalized F test the null and alternative hypotheses are as follows:

$$H_0: \text{all } \beta_k = 0$$

$$H_a: \text{all } \beta_k \neq 0$$

In this test the null hypothesis is that all of the additional parameters in the full model (compared to the reduced model) β_k are equal to zero.

When the null hypothesis is true (making the F test a conditional probability), the F^* statistic is approximately F distributed, and is given by

$$F^* = \frac{(\text{SSE}_R - \text{SSE}_F)/df_R - df_F}{\text{SSE}_F/df_F} \approx F(1 - \alpha; df_R - df_F, df_F) \quad (3.49)$$

where $df_F = n - p_F$ and $df_R = n - p_R$ (n is the number of observations and p is the number of parameters). To calculate this test statistic, the sum of square errors for the two models is first computed, then the F^* statistic is compared to the F distribution with appropriate numerator and denominator degrees of freedom (see Appendix C, Table C.4 for the F distribution). Specifically,

$$\begin{aligned} \text{If } F^* \leq F(1 - \alpha; df_R - df_F, df_F), & \quad \text{then conclude } H_0 \\ \text{If } F^* \geq F(1 - \alpha; df_R - df_F, df_F), & \quad \text{then conclude } H_a \end{aligned} \quad (3.50)$$

The generalized F test is useful for comparing models of different sizes. When the difference in size between two models is one variable (e.g., a model with six vs. five independent variables), the F test yields an equivalent result to the t test for that variable. Thus, the F test is most useful for comparing models that differ by more than one independent variable.

A modification of this test, the Chow test, is used to test temporal stability of model parameters. For two time periods three models must be estimated: a regression model using the complete data; a regression model using data from period 1; and a regression model using data from period 2. Equation 3.49 is applied with SSE_R computed from the all-data model (sample size n_T) and SSE_F computed as the sum of SSEs from the two time-period models (sample sizes n_1 and n_2 where $n_1 + n_2 = n_T$). For this model, $df_F = n_1 + n_2 - 2p_F$ (where p_F is the number of variables in each of the time-period models, multiplied by 2 because the number of variables is the same for both time-period models) and $df_R = n_T - p_R$ (where p_R is the number of variables in all-data model). The F -statistic again is $F(1 - \alpha; df_R - df_F, df_F)$.

Example 3.8

Continuing with Example 3.7, GOF statistics are compared. Table 3.7 shows GOF statistics for the model. The R -squared value is 0.8947. This model suggests that if the specification is correct (e.g., linear, homoscedastic, etc.), then the collection of independent variables accounts for about 89% of the uncertainty or variation in $AADT$. This result is quite satisfying because previous studies have achieved R -squared values of around 70%. It is only with respect to other models on the same phenomenon that R -squared comparisons are meaningful.

This standard statistic provided by most statistical analysis software is a test of the naïve regression model (the model with the intercept term only) versus the specified model. This assumption is verified by checking the numerator and denominator degrees of freedom. The numerator degrees of freedom is the reduced minus full model degrees of freedom, or $(n_R - p_R) - (n_F - p_F) = (119 - 1) - (119 - 6) = 5$, and the denominator degrees of freedom is the degrees of freedom for the full model, or $(119 - 6) = 113$. The value of the F statistic, or F^* , is 192.1. Although not reported by the statistics package, F^* is much larger than the critical value of F , and reflects the tail of the F distribution, where probability is less than 0.01%. Thus, if the naïve model is correct, then the probability of observing data with all specified model parameters equal to zero is less than 0.01%. This test result provides objective evidence that the specified model is better than a naïve model.

3.10 Multicollinearity in the Regression

Multicollinearity among independent variables is a problem often encountered with observational data. The only way to ensure that multicollinearity is minimized or eliminated is to conduct a carefully designed experiment. Although designed experiments should be conducted when possible, often an experiment is impossible due to the lack of ability to control many variables of interest.

In the presence of multicollinearity or intercorrelation between variables, numerous reasonable and well-intentioned research-oriented questions are difficult to answer (Neter et al. 1996), such as the following. What is the relative importance of various factors (predictors) in the regression? What is the magnitude of the effect of a given predictor variable on the response? Can some predictor variables be dropped from the model without predictive or explanatory loss? Should omitted predictor variables be included in the model?

Multicollinearity enters the regression when independent variables are correlated with each other or when independent variables are correlated with omitted variables that are related to the dependent variable. Perfect multicollinearity between predictor variables represents a singularity problem (see Appendix A). In these unusual cases, one variable provides no additional information over and above its perfectly correlated counterpart. Often perfect multicollinearity is a result of a data coding error, such as coding indicator variables for all levels of a nominal scale variable. Perfect multicollinearity, typically resulting in error messages in statistical analysis packages, means that one independent variable is redundant.

Near or high multicollinearity is a more difficult and common problem, and produces the following effects:

1. The consequence of near multicollinearity is the high standard errors of the estimated parameters of the collinear variables. Near multicollinearity does not prevent least squares from obtaining a best fit to the data, nor does it affect inferences on mean responses or new

observations, as long as inferences are made within the observation limits specified by the joint distribution of the correlated variables. For example, if household income and value of vehicle are strongly positively correlated in a model of total daily vehicle trips, then making an inference for a household with high income and a low-valued vehicle may be difficult.

2. Estimated regression parameters have large sampling variability when predictor variables are highly correlated. As a result, estimated parameters vary widely from one sample to the next, perhaps resulting in counterintuitive signs. In addition, adding or deleting a correlated variable in the regression can change the regression parameters.
3. The standard interpretation of a regression parameter does not apply: one cannot simply adjust the value of an independent variable by one unit to assess its affect on the response, because the value of the correlated independent variable will also change.

The most common remedial measures are now described. The interested reader is directed to other references to explore in greater detail remedial measures for dealing with multicollinearity (e.g., Greene 2000; Myers 1990; Neter et al. 1996).

1. Often pair wise correlation between variables is used to detect multicollinearity. It is possible, however, to have serious multicollinearity between groups of independent variables, for example, X_3 is highly correlated with a linear combination of X_1 and X_2 . The variance inflation factor (VIF) is often used to detect serious multicollinearity.
2. Ridge regression, a method that produces biased but efficient estimators for obtaining parameter estimates, is also used to avoid the inefficient parameter estimates obtained under multicollinearity.
3. Often when variables have serious multicollinearity one of the offending variables is removed from the model. This action is justified when there are theoretical grounds for removing the variable. For example, an independent variable thought to have an associative rather than causal relationship with the response might be omitted. Another justification might be that only one of the variables is collected in practice and used for future predictions. Removing a variable, however, also removes any unique effects this variable may have on the outcome variable.
4. Doing nothing is a common response to multicollinearity. The effects of leaving correlated variables in the model and the limitations involved with interpreting the model should be recognized and documented.

5. An ideal remedy for multicollinearity is to control levels of the suspected multicollinear variables during the study design phase. Although this solution requires the ability to understand the threats of multicollinearity before data collection, it is the optimal statistical solution. For example, a suspected high correlation between the variables household income and value of vehicle might be remedied by stratifying households to ensure that all stratifications of low, medium, and high household income and low, medium, and high value of vehicle are obtained.

Multicollinearity is perhaps the most common negative consequence to the regression resulting from observational data. Good planning in the study design is the best remedy for multicollinearity. Other remedies are performed post hoc, and range from doing nothing to applying ridge regression techniques.

Finally, it is important to keep in mind that highly correlated variables may not necessarily cause estimation problems. It is quite possible that two highly correlated variables produce well-defined parameter estimates (estimates with low standard errors). The problem arises when the standard errors of one or both of the correlated variables are high.

3.11 Regression Model-Building Strategies

The following model-building strategies are frequently used in linear regression analysis.

3.11.1 Stepwise Regression

Stepwise regression is a procedure that relies on a user-selected criterion, such as R -squared, adjusted R -squared, F -ratio, or other GOF measures, to select a “best” model among competing models generated by the procedure. Stepwise regression procedures can either be backward or forward. Backward stepwise regression starts by comparing models with large numbers of independent variables (as many as, say, 30 or 40) and sequentially removing one independent variable at each step. The variable removed is the one that contributes least to the GOF criterion. The procedure iterates until a regression model is obtained in the final step. The user can then compare “best” models of different sizes in the computer printout. As one would expect, forward stepwise begins with a simple regression model and sequentially grows the regression by adding the variable with the largest contribution to the GOF criterion.

There are several problems with stepwise regression. First, the procedure is mechanical, resulting in perhaps many models that are not appealing theoretically but simply produce superior GOF statistics. Second, the number of model fits is limited by the user-defined independent variables. For example, if two-way interactions are not “entered” into the procedure, they will not be considered. Finally, in some programs once a variable is entered into the regression it cannot be removed in subsequently estimated models, despite the real possibility that an overall improvement in GOF is possible only if it is removed.

3.11.2 Best Subsets Regression

Best subsets regression is a procedure that fits all permutations of p -independent variables, and can grow or shrink models similar to stepwise procedures. Best subsets procedures take longer than stepwise methods for a given set of variables. However, computing capability is rarely a problem except for extremely large data sets with many variables. So, in the majority of applications best subsets procedures are preferred to stepwise ones. In all other respects best subsets and stepwise procedures are similar. The same criticisms apply to best subsets—it is a mechanical process that provides little insight in the way of new independent variables and may produce many models that lack theoretical appeal.

3.11.3 Iteratively Specified Tree-Based Regression

Iteratively specified tree-based regression is a procedure that enables the development of statistical models in an iterative way to help identify functional forms of independent variables (see Washington 2000b). Because the procedure is not offered in neatly packaged statistical analysis software, it is more difficult to apply. However, some of the limitations of stepwise and best subset procedures are overcome. The procedure involves iterating back and forth between regression trees and OLS to obtain a theoretically appealing model specification. It is particularly useful when there are many noncontinuous independent variables that could potentially enter the regression model. For the theory of regression trees and transportation applications of this methodology, the reader should consult Breiman et al. (1984), Washington (2000b), Wolf et al. (1998), Washington and Wolf (1997), and Washington et al. (1997).

3.12 Estimating Elasticities

It is extremely useful—particularly in practice—to estimate the responsiveness and sensitivity of a dependent variable Y with respect to changes in one or more independent variables X. Although the magnitudes of the

TABLE 3.8

Elasticity Estimates for Various Functional Forms

Model	Elasticity
Linear	$\frac{\Delta Y_i}{\Delta X_i} \cdot \frac{X_i}{Y_i} = \beta_i \left(\frac{X_i}{Y_i} \right)$
Log-linear	$\frac{\Delta Y_i}{\Delta X_i} \cdot \frac{X_i}{Y_i} = \beta_i (X_i)$
Linear-log	$\frac{\Delta Y_i}{\Delta X_i} \cdot \frac{X_i}{Y_i} = \beta_i \left(\frac{1}{Y_i} \right)$
Double log	$\frac{\Delta Y_i}{\Delta X_i} \cdot \frac{X_i}{Y_i} = \beta_i$

Adapted from McCarthy, P.S. (2001). *Transportation Economics Theory and Practice: A Case Study Approach*. Blackwell, Boston, MA.

regression parameters yield this information in terms of measurement units, it is sometimes more convenient to express the sensitivity in terms of the percentage change in the dependent variable resulting from a 1% change in an independent variable. This measure is the elasticity, and is useful because it is dimensionless unlike an estimated regression parameter, which depends on the units of measurement. The elasticity of the dependent variable Y with respect to an independent variable X is given as

$$e_i = \beta_i \frac{X_i}{Y_i} \approx \frac{\partial Y_i}{\partial X_i} \times \frac{X_i}{Y_i} \quad (3.51)$$

Roughly speaking, a value $e_i = 2.7$ suggests that a 1% increase in X_i results in a 2.7% increase in Y . Table 3.8 summarizes the elasticity estimation for four commonly used functional forms in regression models.

3.13 Censored Dependent Variables—Tobit Model

Occasionally, dependent variables are encountered that are censored in that their measurements are clustered at a lower threshold (left censored), an upper threshold (right censored), or both. For example, one may wish to study vehicle accident rates (number of accidents per vehicle-mile traveled)

on a number of highway segments by observing the number of accidents that occur over some time period. However, some highway segments may not have an observed accident over the time period being considered, in which case there is a clustering of observations at zero in addition to the nonzero observations (those observations where at least one accident was observed over the time period being studied). Note that censored data are the result of observations beyond the censor limit being recorded at the limit (e.g., a car going at 100 mph being recorded as going at 55 mph, the maximum detectable speed of an instrument), whereas truncated data are the result of data beyond the truncation limit being discarded (e.g., the 100 mph car observation is discarded). When encountering censored or truncated data, there are at least three reasons for not simply conducting an analysis on all nonzero observations: (1) it is apparent that by focusing solely on the nonzero observations some potentially useful information is ignored; (2) ignoring some sample elements would affect the degrees of freedom and the t - and F -statistics; and (3) a simple procedure for obtaining efficient and/or asymptotically consistent estimators by confining the analysis to the positive subsample is lacking.

Censored dependent variables are analyzed using a number of techniques, the most popular of which is the Tobit model as introduced by James Tobin (1958). The Tobit model takes the form

$$\begin{aligned} Y_i^* &= \beta X_i + \varepsilon_i, & i = 1, 2, \dots, N \\ Y_i &= Y_i^* & \text{if } Y_i^* > 0 \\ &= 0 & \text{if } Y_i^* \leq 0 \end{aligned} \quad (3.52)$$

where Y_i^* is an implicit, stochastic index (latent variable), which is observed only when positive, which is observed only when positive N is the number of observations, Y_i is the dependent variable, X_i is a vector of explanatory variables, β is a vector of estimable parameters, and ε_i is a normally and independently distributed error term with zero mean and constant variance σ^2 . The corresponding likelihood function for the Tobit model is

$$L = \prod_0 [1 - \Phi(\beta X_i / \sigma)] \prod_1 \sigma^{-1} \phi[(Y_i - \beta X_i) / \sigma] \quad (3.53)$$

where Φ is the standard normal distribution function and ϕ is the standard normal density function.

With this model, the expected value of the dependent variable for all cases (those that are censored and those that are not), $E[Y_i]$, is

$$E[Y_i] = \beta X_i F(z) + \sigma f(z) \quad (3.54)$$

where $z = \beta X_i / \sigma$ is the z -score for an area under the normal curve, $F(z)$ is the cumulative normal distribution function, associated with the proportion of

cases above the zero, $f(z)$ the unit normal density (value of the derivative of the normal curve at a particular point), and σ the standard deviation of the error term. And, the relationship among the expected value of all observations, $E[Y_i]$; the expected value for cases above zero, $E[Y'_i]$; and the probability of being above zero, $F(z)$ is

$$E[Y_i] = F(z)E[Y'_i] \quad (3.55)$$

where Y' denotes observations above zero (not censored).

For the Tobit model the standard assumptions for linear regression listed in Table 3.1 also apply. Careful examination of the model structure reveals that estimation of this model using OLS leads to serious specification errors because in Equation 3.61 the zero observations are not generated by the same process that generates the positive observations. OLS estimation yields biased and inconsistent parameter estimates, because unbiasedness and consistency require that $\varepsilon_i \sim N(0, \sigma^2)$.

Example 3.9

Accident and traffic data are available for a 5-year period (1 January, 1995 to 31 December, 1999) representing vehicular accidents that occurred on 325 interstate highway segments in Indiana. By dividing the total number of accidents on each interstate segment by the total number of vehicles that traveled over the segment during the 5-year period and the length of the segment, the number of accidents per vehicle-mile traveled is determined. However, 128 of the 325 interstate highway segments do not have any reported accidents over the 5-year period. These segments are censored at zero (accidents per vehicle-mile traveled are zero) while the other 197 interstate segments have nonzero accident rates.

The variables available for model estimation are given in Table 3.9 and Tobit model estimation results are provided in Table 3.10. Table 3.10 shows that a wide range of pavement, roadway geometric characteristics and traffic characteristics were found to have a statistically significant effect on the accident rate. In this table, variables with a positive parameter estimate increase the accident rate and those variables with a negative parameter estimate decrease the accident rate. The motivated reader can refer to Anastasopoulos et al. (2008) for a complete and detailed description on these model results, including the calculation and interpretation of elasticities (as discussed in Section 3.11).

3.14 Box–Cox Regression

The Box–Cox transformation is a method for generalizing the linear regression model and for selecting an appropriate specification form for a given

TABLE 3.9

Variables Available to Model Accident Rates

Variable No.	Variable Description
1	Segment identification number
2	Number of segment observations for each interstate
3	Number of single vehicle accidents per 100-million vehicle miles travelled
4	Interstate (64: I-64, 65: I-65, 70: I-70, 74: I-74, and 164: I-164)
5	Average friction in the road segment over the 5-year period (measured at 40 mi/h)
6	Minimum friction reading in the road segment over the 5-year period
7	Maximum friction reading in the road segment over the 5-year period
8	Standard deviation of the friction readings in the road segment over the 5-year period
9	Age of the pavement in 1999
10	Average international roughness index reading on the road segment over the 5-year period
11	Minimum international roughness index reading in the road segment over the 5-year period
12	Maximum international roughness index reading in the road segment over the 5-year period
13	Standard deviation of the international roughness index readings in the road segment over the 5-year period
14	Average rutting (in inches) in the road segment over the 5-year period
15	Minimum rut (in inches) reading in the road segment over the 5-year period
16	Maximum rut (in inches) reading in the road segment over the 5-year period
17	Standard deviation of the rut (in inches) readings in the road segment over the 5-year period
18	Average pavement condition rating (PCR, 0–100 scale) in the road segment over the 5-year period
19	Minimum pavement condition rating (PCR, 0–100 scale) in the road segment over the 5-year period
20	Maximum pavement condition rating (PCR, 0–100 scale) in the road segment over the 5-year period
21	Standard deviation of the pavement condition rating (PCR, 0–100 scale) in the road segment over the 5-year period
22	Average pavement quality index (0–100 scale) in the road segment over the 5-year period
23	Total number of vehicles over the 5 years
24	Segment length (in miles)
25	Total number of ramps in the opposite direction
26	Total number of ramps in the viewing direction
27	Number of lanes
28	Pavement surface type (1 if asphalt, 0 if concrete)
29	Median configuration (1 if depressed, 2 if depressed with bumps, 3 if berms, 4 if flush, 5 if sloped, and 6 if rock wall)
30	Median surface (0 if concrete, 1 if asphalt, 2 if grass, 3 if paved, 4 if grass with trees, 5 if grass with bushes, 6 if trees, and 7 if rock)

continued

TABLE 3.9 (continued)

Variables Available to Model Accident Rates

Variable No.	Variable Description
31	Median width (in feet)
32	Presence of median barrier (1 if present, 0 if absent)
33	Median barrier type (1 if w-beam, 2 if concrete, 3 if cable, 4 if cable, 5 if box-beam, 6 if rock wall)
34	Median barrier location (0 if left, 1 if middle left, 2 if middle, 3 if middle right, 4 if right)
35	Presence of interior shoulder (1 if present, 0 absent)
36	Interior shoulder width (in feet)
37	Interior shoulder surface (1 if asphalt ; 0 if concrete)
38	Interior rumble strips (1 if present, 0 if absent)
39	Outside shoulder width (in feet)
40	Outside shoulder surface (1 if asphalt ; 0 if concrete)
41	Outside rumble strips (1 if present, 0 if absent)
42	Outside barrier type (1 if w-beam, 2 if concrete, 3 if wire rope, 4 if cable, 5 if box-beam, 6 if rock wall)
43	Outside barrier location (1 if less than 15 feet, 2 if greater than 15 feet)
44	Average daily traffic over the 5 years
45	Average daily traffic of trucks over the 5 years
46	Percentage of single unit trucks (average daily)
47	Percentage of combination trucks (average daily)
48	Speed limit of the road segment
49	State maximum speed limit
50	Number of bridges in the road segment
51	Horizontal curve 1 type (1 if inside, 2 if outside)
52	Length of horizontal curve 1 (in feet)
53	Radius of horizontal curve 1 (in feet)
54	Horizontal curve 2 type (1 if inside, 2 if outside)
55	Length of horizontal curve 2 (in feet)
56	Radius of horizontal curve 2 (in feet)
57	Horizontal curve 3 type (1 if inside, 2 if outside)
58	Length of horizontal curve 3 (in feet)
59	Radius of horizontal curve 3 (in feet)
60	Horizontal curve 4 type (1 if inside, 2 if outside)
61	Length of horizontal curve 4 (in feet)
62	Radius of horizontal curve 4 (in feet)
63	Horizontal curve 5 type (1 if inside, 2 if outside)
64	Length of horizontal curve 5 (in feet)
65	Radius of horizontal curve 5 (in feet)
66	Average radius per horizontal curve in the road segment (in feet)
67	Number of horizontal curves in the road segment
68	Length of vertical curve 1 (in feet)

TABLE 3.9 (continued)

Variables Available to Model Accident Rates

Variable No.	Variable Description
69	Vertical curve 1 type (1 if crest, 2 if sag)
70	Curve length divided by the difference in connecting grades (K-value) for vertical curve 1
71	Length of vertical curve 2 in feet
72	Vertical curve 2 type (1 if crest, 2 if sag)
73	Curve length divided by the difference in connecting grades (K-value) for vertical curve 2
74	Length of vertical curve 3 in feet
75	Vertical curve 3 type (1 if crest, 2 if sag)
76	Curve length divided by the difference in connecting grades (K-value) for vertical curve 3
77	Number of vertical curves in the road segment
78	Pavement surface change in the road segment (1 if change, 0 if no change)
79	Changes in vertical profile (1 if change, 0 if no change)
80	Number of bridges per mile
81	Number of horizontal curves per mile
82	Number of vertical curves per mile
83	Number of accidents per 100-million vehicle-miles travelled

TABLE 3.10

Tobit Regression Estimation of Accidents per 100-Million Vehicle-Miles Traveled

Variable	Parameter Estimate	t-Ratio
Constant	173.41	5.91
Interstate indicator variable (1 if I-70 or I-164, 0 otherwise)	-29.52	-3.45
<i>Pavement Characteristics</i>		
High-friction indicator variable (1 if all 5-year friction readings are 40 or higher, 0 otherwise)	-33.52	-3.52
Smooth pavement indicator variable (1 if 5-year international roughness index readings are below 75, 0 otherwise)	-31.26	-3.71
Excellent rutting indicator variable (1 if all 5-year rutting readings are below 0.12 inches, 0 otherwise)	-26.52	-2.92
Good rutting indicator variable (1 if all 5-year average rutting readings are below 0.2 inches, 0 otherwise)	-15.64	-1.88
Average pavement condition rating indicator variable (1 if greater than 95, 0 otherwise)	17.35	2.64
<i>Geometric Characteristics</i>		
Median width indicator variable (1 if greater than 74 feet, 0 otherwise)	-15.47	-2.41
Median barrier presence indicator variable (1 if present, 0 otherwise)	-79.96	-4.80

continued

TABLE 3.10 (continued)

Tobit Regression Estimation of Accidents per 100-Million Vehicle-Miles Traveled

Variable	Parameter Estimate	t-Ratio
Inside shoulder width indicator variable (1 if 5 feet or greater, 0 otherwise)	-33.28	-4.26
Outside shoulder width (in feet)	-6.20	-2.91
Number of bridges (per road section)	-7.10	-1.80
Rumble strips indicator variable (1 if both inside and outside rumble strips are present, 0 otherwise)	-22.75	-2.47
Number of vertical curves per mile	-5.09	-2.73
Ratio of the vertical curve length over the road section length (in tenths)	2.95	2.28
Horizontal curve's degree curvature indicator variable (1 if average degrees per road section is greater than 2.1, 0 otherwise)	-19.42	-2.02
Number of ramps in the driving direction per lane-mile	22.16	3.36
<i>Traffic Characteristics</i>		
Average daily number of passenger cars (in 1,000 vehicles per day)	-0.80	-2.82
Average daily percent of combination trucks	-1.04	-3.41
Number of observations 325.		
Log-likelihood at zero -1,724.20		
Log-likelihood at convergence -1,242.99		
Adapted from Anastasopoulos et al. 2008.		

model (see Appendix D for a general discussion of variable transformations). The procedure involves transforming a variable from X_i to X_i^* such that

$$X_i^* = \frac{X_i^\lambda - 1}{\lambda} \quad (3.56)$$

where λ is an estimated parameter. The basic form of this transformation is

$$\frac{Y_i^\lambda - 1}{\lambda} = \alpha + \beta \left(\frac{X_i^\lambda - 1}{\lambda} \right) + \varepsilon_i \quad (3.57)$$

For a given value of λ , a linear regression is estimated using OLS estimation. Table 3.11 lists some of the common λ parameter values and the corresponding functional specifications used in OLS regression. The equation becomes

$$Y = a + \sum_{k=1}^K \beta_k X_k^{(\lambda)} + \varepsilon \quad (3.58)$$

TABLE 3.11

Common λ Parameter Values and Corresponding Functional Specifications

λ_y, λ_x Values	Functional Form
$\lambda_y = \lambda_x = 1$	Linear regression [Y on X]
$\lambda_y = \lambda_x = 0$	Double log regression [$LN(Y)$ on $LN(X)$]
$\lambda_y = 0, \lambda_x = 1$	Log-linear regression [$LN(Y)$ on X]
$\lambda_y = 1, \lambda_x = 0$	Linear-log regression [{Y on $LN(X)$ }]

The estimated parameter λ need not be restricted to values in the $[0, 1]$ interval, and it is possible to obtain a wide range of values; implying that the Box–Cox procedure is not limited to the set of linear, double log, and semi-logarithmic functional forms shown in Table 3.11. Common values for the parameter λ range between –2 and 2. Further, λ is generally assumed to be the same for all variables in the model because differences become computationally cumbersome. Finally, if λ is not known, the regression model in Equation 3.58 becomes nonlinear in the parameters and nonlinear least squares estimation must be used.

4

Violations of Regression Assumptions

A number of assumptions or requirements must be substantively met for the linear regression model parameters to be best linear unbiased estimators (BLUE). That is, parameters must be unbiased, asymptotically efficient, and consistent. The six main regression model assumptions are as follows: the disturbance terms have zero mean; the disturbance terms are normally distributed; the regressors and disturbance terms are not correlated; the disturbance terms are homoscedastic; the disturbance terms are not serially correlated (nonautocorrelated); and a linear-in-parameters relationship exists between the dependent and independent variable(s).

This chapter extends the practical understanding of the linear regression model development and answers some questions that arise as a result of violating the previously listed assumptions. It answers questions such as the following. Why does it matter that a particular assumption has been violated? How can regression violations be detected? What are the substantive interpretations of violations of the assumptions? What remedial actions should be taken to improve on the model specification?

4.1 Zero Mean of the Disturbances Assumption

The assumption that the disturbance term has a mean of zero is the least restrictive and “damaging” of the violations examined. A violation of this assumption can result from consistent positive or negative errors of measurement in the dependent variable (Y) of a regression model. Two distinct forms of this violation are possible. The first and most commonly encountered violation occurs when $E(\varepsilon_i) = \mu \neq 0$. In this case μ is subtracted from the ε_i to obtain the new disturbances $\varepsilon_i^* = \varepsilon_i - \mu$, which have zero mean. The regression model then becomes

$$Y_i = \beta_0^* + \beta X_i + \varepsilon_i^* \quad (4.1)$$

where $\beta_0^* = \beta_0 + \mu$. It is clear that only β_0^* and β are estimated and not β or μ , which suggests that β and μ cannot be extracted from β_0^* without additional assumptions. Despite this limitation, Equation 4.1 satisfies the remaining

assumptions and as a result ordinary least squares (OLS) yields BLUE estimators for β_0^* and β . In practical terms this means that a nonzero mean for the disturbance terms affects only the intercept term and not the slope (β). If a regression model does not contain an intercept term, however, the β is biased and inconsistent.

The second form of the violation occurs when $E(\varepsilon_i) = \mu_i$, suggesting that the disturbances vary with each i . In this case the regression model of Equation 4.1 could be transformed by adding and subtracting μ_i . However, with this approach, $\beta_{0i}^* = \beta_0 + \mu_i$ would vary with each observation, resulting in more parameters than observations (there are $n\beta_0^*$ terms and one β to be estimated using n observations). This model cannot be estimated unless the data consist of repeated observations such as when analyzing panel data (see Chapter 6 for more details).

4.2 Normality of the Disturbances Assumption

The violation of the assumption that the disturbance terms are approximately normally distributed is commonly the result of the existence of measurement errors in the variables and/or unobserved parameter variations. The hypothesis of normally distributed disturbances is straightforwardly tested using many statistical software packages employing either quantile-quantile plots of the disturbances or χ^2 and/or Kolmogorov-Smirnov tests, which provide inferential statistics on normality.

The property of normality results in OLS estimators that are minimum variance unbiased (MVU) and estimators that are identical to the ones obtained using maximum likelihood estimation (MLE). Furthermore, normality permits the derivation of the distribution of these estimators which, in turn, permits t and F tests of hypotheses to be performed. A violation of the normality assumption therefore results in hypothesis-testing problems and OLS estimates that cannot be shown to be efficient, since the Cramer-Rao bound cannot be determined without a specified distribution. The parameter estimates are consistent; however, further pointing to hypothesis-testing problems. To correct the problem of non-normally distributed disturbances, two alternative approaches are typically used. First, if the sample size used for model estimation is large, then Theil (1978) suggests that the hypothesis of normality is inferred asymptotically for the OLS estimates by relying on the central limit theorem (Theil's proof applies for cases of fixed X in panel data samples, zero mean, and constant variance). Second, alternative functional forms are used for the regression model such as the gamma regression model (Greene 1990b) and the stochastic frontier regression model (Aigner et al. 1977).

4.3 Uncorrelatedness of Regressors and Disturbances Assumption

The assumption that the random regressors are contemporaneously uncorrelated with the disturbances is the most important assumption of all to meet. If this assumption is not met, not only are the parameter estimates biased and inconsistent, but the methods for detecting and dealing with violations of other assumptions will fail. To illustrate, consider the slope estimator for the usual simple linear regression model:

$$\hat{\beta}_{OLS} = \beta + \frac{\sum_i^n X_i \varepsilon_i}{\sum_i^n X_i^2} \quad (4.2)$$

When $\hat{\beta}_{OLS}$ is an unbiased estimator of β , the ratio division on the right-hand side of Equation 4.2 is equal to zero. When there is no correlation between X_i and ε_i , then

$$E\left(\frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2}\right) = \sum_{i=1}^n \left[E\left(\frac{X_i}{\sum_{i=1}^n X_i^2}\right) E(\varepsilon_i) \right] = 0$$

(the first equality is valid because $COV(X, \varepsilon) = 0$ and the second because $E(\varepsilon_i) = 0$). Thus, when $COV(X, \varepsilon) \neq 0$, $\hat{\beta}_{OLS}$ is not an unbiased estimator of β .

A violation of this assumption usually arises in four ways. The first is when the explanatory variables of the regression model include a lagged dependent variable and its disturbances are serially correlated (autocorrelation). For example, consider the model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 Y_{i-1} + u_i \\ u_i &= \varepsilon_i - \rho \varepsilon_{i-1} \end{aligned} \quad (4.3)$$

where the u_i and ε_i are independent and identically distributed according to $\varepsilon_i \sim N(0, \sigma^2)$. The disturbance $u_i = \varepsilon_i - \rho \varepsilon_{i-1}$ is correlated with the regressor Y_{i-1} (this correlation exists because the disturbance has two components $(\varepsilon_i, \varepsilon_{i-1})$ and Y_{i-1} is dependent on one of them (ε_{i-1}) , as shown in Equation 4.3. This model is frequently called a correlated error model, as both u_i and u_{i-1} contain the component ε_{i-1} . When this correlation occurs, the OLS estimator is inconsistent in most cases (possible corrections for this problem are discussed in Section 4.5).

The second source is when the random regressors are nonstationary. The concept of stationarity is discussed in more detail in Chapter 7; however, with a nonstationary random regressor, the assumption that the sample variance of an independent variable X approaches a finite limit is violated.

Under these circumstances, the OLS estimator may no longer be consistent and the usual large sample approximate distribution is no longer appropriate. As a result, confidence intervals and hypothesis tests are no longer valid. To solve the problem, either the process generating the observations on the random regressor needs to be modeled or the regressor needs to be transformed to obtain stationarity (usually by differencing; see Chapters 7 and 8 for a more in-depth discussion on stationarity and differencing).

The third source is when the values of a number of variables are jointly determined (simultaneous equations model). Consider the following system:

$$\begin{aligned} Y_i &= \beta_0 + \beta X_i + \varepsilon_i \\ X_i &= \beta'_0 + \gamma Y_i + \delta Z_i \end{aligned} \quad (4.4)$$

In this model Y_i, X_i are jointly determined by the two equations whereas Z_i is exogenously determined. The OLS estimation of β, γ is inconsistent because X_i and ε_i are correlated (X_i depends on Y_i through the first equation and Y_i depends on ε_i through the second equation and hence X_i depends on ε_i and so are correlated). Modeling systems of simultaneous equations is discussed in Chapter 5.

The fourth source is when one or more of the variables are measured with error (measurement error). Consider the variables X_i and Y_i that are measured with error and X_i^* and Y_i^* the true, yet unobservable, variables. Then,

$$\begin{aligned} X_i &= X_i^* + u_i \\ Y_i &= Y_i^* + v_i \end{aligned} \quad (4.5)$$

where u_i and v_i are random disturbances capturing the measurement errors for the i th observation. For these two disturbance terms the usual assumptions are valid

$$\begin{aligned} E[u_i] &= E[v_i] = 0 \\ VAR(u_i) &= E[u_i^2] = \sigma_u^2, \quad VAR(v_i) = E[v_i^2] = \sigma_v^2 \\ COV(u_i, u_j) &= E[u_i u_j] = 0, \quad COV(v_i, v_j) = E[v_i v_j] = 0, \quad \text{for } i \neq j \\ COV(u_i, v_j) &= E[u_i v_j] = 0, \quad \forall i, j \end{aligned}$$

The problem with measuring the variables with error arises when trying to estimate a model of the following form:

$$Y_i^* = \beta_0 + \beta X_i^* + \varepsilon_i \quad (4.6)$$

When Y_i^* and X_i^* are unobservable, the following model is estimated:

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i, \quad \text{where } \varepsilon_i = v_i - \beta u_i \quad (4.7)$$

As Griffiths et al. (1993) show, $E(\varepsilon_i) = 0$ and $VAR(\varepsilon_i) = \beta_{0v}^2 + \beta^2 \sigma_u^2$, which suggest that ε_i has zero mean and constant variance. However, ε_i and Y_i are correlated resulting in biased and inconsistent OLS estimators. To illustrate, the covariance between ε_i and Y_i is estimated as

$$\begin{aligned} COV(X_i, \varepsilon_i) &= E[(X_i - E[X_i])(\varepsilon_i - E(\varepsilon_i))] \\ &= E[u_i(v_i - u_i\beta)] = E[u_i v_i] - \beta E[u_i^2] = -\beta^2 \sigma_u^2 \neq 0 \end{aligned}$$

Further, Griffiths et al. (1993) show that the inconsistency of the OLS estimator for the slope is

$$\hat{\beta}_{OLS} = \beta \left(1 - \frac{\sigma_u^2}{\sigma_x^2}\right) \quad (4.8)$$

This finding suggests that the inconsistency of the estimator depends on the magnitude of the variance of the measurement error relative to the variance of X_i . Similar to the case when both X_i and Y_i are measured with error, it is possible that only one of the two variables (X_i or Y_i) is measured with error. In the first case, measurement error in X_i , the parameters are once again biased and inconsistent. In the second case, measurement error in Y_i only, the slope parameter will be unbiased and consistent but with an increased error variance. The most common technique for dealing with the problem of measurement error is to use instrumental variables estimation. This technique is commonly applied to estimate systems of simultaneous equations and is discussed in Chapter 5.

4.4 Homoscedasticity of the Disturbances Assumption

The assumption that the disturbances are homoscedastic—constant variance across observations—arises in many practical applications, particularly when cross-sectional data are analyzed (the term heteroscedasticity means “differing variance” and comes from the Greek words “hetero” (different) and “skedasis” (dispersion); as a grammatical note, the correct spelling of this violation is heteroskedasticity, not heteroscedasticity; but, the former is

so well established in the literature that rarely does the term appear with a k). For example, when analyzing household mobility patterns, there is often greater variation in mobility among high-income families than low-income families, possibly due to the greater flexibility in traveling allowed by higher incomes. Figure 4.1 (left) shows that as a variable X (income) increases, the mean level of Y (mobility) also increases, but the variance around the mean value remains the same at all levels of X (homoscedasticity). On the other hand, as Figure 4.1 (right) shows that, although the average level of Y increases as X increases, its variance does not remain the same for all levels of X (heteroscedasticity).

Heteroscedasticity, or unequal variance across observations, does not typically occur in time-series data because changes in the variables are likely to be of similar magnitude over time. This similarity in error magnitude is not always true, however. When heteroscedasticity is observed in time-series data it is analyzed using autoregressive conditionally heteroscedastic (ARCH) models, which are discussed in Chapter 7 (see Engle (1982) and Maddala (1988) for more details on these models).

To formalize, heteroscedasticity implies that the disturbances have a non-constant covariance, that is $E(\varepsilon_i^2) = \sigma_i^2$, $i = 1, 2, \dots, n$. For the linear regression model, $\hat{\beta}_{OLS}$ (given in Equation 4.2) is still unbiased and consistent because these properties do not depend on the assumption of constant variance. However, the variance of $\hat{\beta}_{OLS}$ is now different since

$$VAR(\hat{\beta}_{OLS}) = VAR\left(\frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2}\right) = \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\left(\sum_{i=1}^n X_i^2\right)^2} \quad (4.9)$$

where the second equality is valid because the $VAR(\varepsilon_i)$ is now σ_i^2 . Note that if $\sigma_i^2 = \sigma^2$, this equation reverts back to

$$\frac{\sigma^2}{\sum_{i=1}^n X_i^2}$$

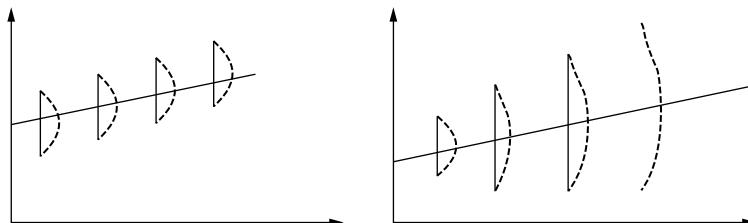


FIGURE 4.1
Homoscedasticity (left); heteroscedasticity (right).

the usual formula for $VAR(\hat{\beta}_{OLS})$ under homoscedasticity. Further, it is straightforward to show that $E(s^2)$ will involve all of the σ_i^2 and not a single common σ^2 (Baltagi 1993). This finding implies that when a software package reports the $VAR(\hat{\beta}_{OLS})$ as

$$\frac{s^2}{\sum_{i=1}^n X_i^2}$$

for a regression problem, it is committing two errors; it is not using the correct formula for the variance (Equation 4.9) and it is using s^2 to estimate a common σ^2 when the σ_i^2 are different. The magnitude of the bias resulting from this error depends on the heteroscedasticity and the regressor. Under heteroscedasticity, OLS estimates lose efficiency and are no longer BLUE. However, they are still unbiased and consistent. In addition, the standard error of the estimated parameters is biased and any inference based on the estimated variances, including the reported t and F statistics is misleading.

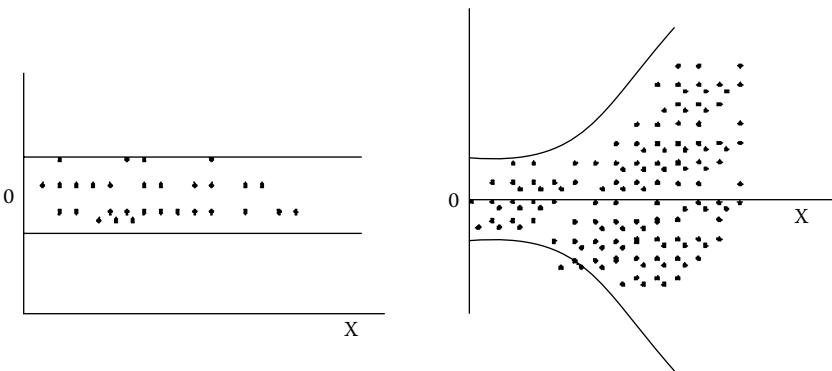
4.4.1 Detecting Heteroscedasticity

When the disturbances in a regression model (ε_i) have constant variance σ^2 (homoscedastic), the residual plot of $\hat{\varepsilon}_i$ versus any one of the explanatory variables X should yield disturbances that appear scattered randomly around the zero line, with no differences in the amount of variation in the residuals regardless of the value of X (Figure 4.2, left). If the residuals are more spread out for larger values of X than for smaller ones, then the assumption of constant variance may be violated (Figure 4.2, right). Conversely, variance could be smaller for high values of X under heteroscedasticity.

The intuitive and rather straightforward graphical approach for detecting heteroscedasticity was formalized into a statistical test by Park (1966). The Park test suggests that if heteroscedasticity is present, the heteroscedastic variance σ_i^2 may be systematically related to one or more of the explanatory variables. Consider, for example, a two-variable regression model of Y on X ; after the model is fitted the following regression is run:

$$LN(\varepsilon_i^2) = \beta_0 + \beta LN(X_i) + u_i \quad (4.10)$$

where u_i is a disturbance term (the ε_i are obtained from the original regression). The Park test involves testing the null hypothesis of homoscedasticity ($\beta = 0$). If a statistically significant relationship exists between $LN(\varepsilon_i^2)$ and $LN(X_i)$, the null hypothesis is rejected and remedial measures need to be taken. If the null hypothesis is not rejected, then the constant (β) from

**FIGURE 4.2**

Homoscedastic disturbances (left); heteroscedastic disturbances (right).

Equation 4.10 is interpreted as giving the value of the homoscedastic variance (σ^2).

Another common test for heteroscedasticity is the Goldfeld–Quandt test (or split sample regressions of Goldfeld and Quandt 1965). This test investigates the null hypothesis of homoscedasticity against the alternative hypothesis that $\sigma_i^2 = cx_i^2$. The Goldfeld–Quandt test proceeds by estimating two OLS regressions. One is estimated using data hypothesized to be associated with low-variance errors and the other is estimated using data from high-variance errors (typically the middle 20% of the data set is excluded). The null hypothesis of homoscedasticity cannot be rejected if the residual variances for the two regressions are approximately equal, whereas, if they are not equal, the null hypothesis is rejected. The Goldfeld–Quandt test statistic is

$$\frac{\text{SSE}_{\text{low variance}}}{\text{SSE}_{\text{high variance}}} \approx F\left(\alpha; \frac{n-d-2k}{2}, \frac{n-d-2k}{2}\right)$$

where SSE is the sum of square error, n is the total sample size, d is the number of observations excluded, and k is the number of independent variables in the model.

There are numerous other tests for investigating heteroscedasticity, including the Breusch and Pagan (1979) test which is an alternative to the Goldfeld–Quandt test when an ordering of the observations according to increasing variance is not simple. A further option is the White (1980) test that does not depend on the normality assumption for the disturbances as do the Goldfeld–Quandt and Breusch–Pagan tests.

4.4.2 Correcting for Heteroscedasticity

When heteroscedasticity is detected, either of two approaches are taken to make the σ_i^2 terms approximately equal: weighted least squares (WLS) or variance stabilizing transformations. The second approach transforms the dependent variable in a way that removes heteroscedasticity and is applicable when the variance of Y_i is a function of its mean (see Appendix D for more information on variable transformations).

The WLS approach to addressing heteroscedasticity in a regression model is a common procedure. Suppose, for example, that $VAR(\varepsilon_i) = \sigma_i^2 = c_i \sigma^2$, where the c_i are known constants. Heteroscedasticity models in which the variance of the disturbances is assumed to be proportional to one of the regressors raised to a power, are extensively reviewed by Geary (1966), Goldfeld and Quandt (1972), Kmenta (1971), Lancaster (1968), Park (1966), and Harvey (1976). Constancy of variance is achieved by dividing both sides of the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \dots, n \quad (4.11)$$

by c_i , yielding

$$Y_i/c_i = \beta_0/c_i + \beta_1 X_{i1}/c_i + \cdots + \beta_k X_{ik}/c_i + \varepsilon_i/c_i, \quad i = 1, \dots, n \quad (4.12)$$

Each $\omega_i = (c_i)^{-2}$ is called a weight and is the result of minimizing the weighted sum of squares,

$$\sum \omega_i (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2 \quad (4.13)$$

When the σ_i^2 , or a proportional quantity, are known, the weights are not difficult to compute. Estimates for the β of Equation 4.12 are obtained by minimizing Equation 4.13 and are called the WLS estimates because each Y and X observation is weighted (divided) by its own (heteroscedastic) standard deviation σ_i .

The variance stabilizing transformations approach has the following basis: for any function $f(Y)$ of Y with a continuous first derivative $f'(Y)$, a finite second derivative $f''(Y)$, and $\mu_i = E(Y_i)$, it follows that

$$f(Y_i) - f(\mu_i) = (Y_i - \mu_i)f'(\mu_i) + \frac{1}{2}(Y_i - \mu_i)^2 f''(\kappa) \quad (4.14)$$

where κ lies between Y_i and μ_i . Squaring both sides of Equation 4.14 and with $(Y_i - \mu_i)^2$ small, yields

$$VAR(f(Y_i)) \approx (f'(\mu_i))^2 \sigma_i^2(\mu_i) \quad (4.15)$$

where $\sigma_i^2(\mu_i)$ is the variance of Y_i with mean μ_i . To find a suitable transformation f of Y_i that would make $VAR(f(Y_i))$ approximately constant, the following equation needs to be solved:

$$f'(\mu_i) = c/\sigma_i(\mu_i) \quad (4.16)$$

where c is a constant. This transformation f is called a variance stabilizing transformation. For example, if $\sigma_i = \mu_i$, then Equation 4.16 yields $f(\mu_i) = LN(\mu_i)$. Frequently used transformations include the square root transformation when the error variance is proportional to an independent variable, and $(1/X_i)$ when the error variance is proportional to X_i^2 .

Example 4.1

A study of the effects of operating subsidies on transit system performance was undertaken for transit systems in the State of Indiana (see Karlaftis and McCarthy 1998, for additional details and a review of the empirical literature). Data were collected (see Table 4.1 for available variables) for 17 transit systems over a 2-year period (1993 and 1994). Although these are—strictly speaking—panel data (they contain time-series data for a cross-section of transit systems), they are of very “short” duration (only 2 years of data exist) and will be analyzed as if they were cross-sectional. A brief overview of the data set suggests the possible existence of heteroscedasticity because systems of vastly different sizes and operating characteristics are pooled together. To determine significant factors that affect performance (performance is measured with the use of performance indicators), a linear regression model was estimated for each of these indicators. Estimation results are shown in Tables 4.2 through 4.4. For each of

TABLE 4.1

Variables Available for the Analysis

Variable Name	Variable Description
SYSTEM	Transit system name
YEAR	Year when data were collected
OE_RVM	Operating expenses (U.S. dollars) per revenue vehicle mile
PAX_RVM	Passengers per revenue vehicle mile
OE_PAX	Operating expenses per passenger in U.S. dollars
POP	Transit system catchment area population
VEH	Number of vehicles available
EMP	Total number of employees
FUEL	Total annual gallons of fuel
C_LOC	Total subsidies from local sources in U.S. dollars
C_STAT	Total subsidies from state sources in U.S. dollars
C_FED	Total subsidies from the federal government in U.S. dollars

TABLE 4.2
Regression of Operating Expenses per Revenue Vehicle Mile on Selected Independent Variables (*t*-Statistics in Parentheses)

Explanatory Variable	OLS	WLS*	WLS†	WLS‡
Constant	1.572 (8.554)	1.579 (11.597)	1.578 (9.728)	1.718 (11.221)
POP	-2.168E-06 (-0.535)	8.476E-07 (0.562)	-9.407E-07 (-0.582)	-4.698E-07 (-0.319)
VEH	-8.762E-04 (-0.025)	0.031 (2.242)	0.012 (1.191)	0.003 (0.466)
EMP	3.103E-03 (0.198)	-0.011 (-1.814)	-0.003 (-0.603)	0.001 (0.308)
FUEL	-1.474E-05 (-3.570)	-1.443E-05 (-9.173)	-1.334E-05 (-7.741)	-1.367E-05 (-8.501)
C_LOC	1.446E-06 (1.075)	1.846E-06 (4.818)	1.350E-06 (3.965)w	1.320E-06 (5.052)
C_STAT	3.897E-06 (1.305)	1.352E-06 (1.390)	2.647E-06 (3.050)	2.961E-06 (4.183)
C_FED	2.998E-06 (1.990)	3.794E-06 (6.777)	3.221E-06 (5.79)	2.938E-06 (6.958)
R ²	0.555	0.868	0.917	0.942

Note: WLS* weighted least squares (weight: FUEL^{1,1}); WLS† weighted least squares (weight: VEH^{1,3}); WLS‡ weighted least squares (weight: EMP^{1,(6)}).

TABLE 4.3

Regression of Passengers per Revenue Vehicle Mile on Selected Independent Variables (*t*-Statistics in Parentheses)

Explanatory Variable	OLS	WLS*	WLS†
Constant	0.736 (7.888)	0.771 (9.901)	0.756 (11.689)
POP	-3.857E-06 (-1.877)	-1.923E-06 (-2.583)	-1.931E-06 (-3.729)
VEH	5.318E-03 (0.299)	0.023 (3.336)	0.025 (13.193)
EMP	-1.578E-03 (-0.198)	-0.009 (-3.162)	-0.010 (-13.026)
FUEL	-2.698E-07 (-0.129)	-7.069E-07 (-1.120)	-4.368E-07 (-1.041)
C_LOC	-3.604E-07 (-0.528)	4.982E-08 (0.341)	5.753E-08 (0.807)
C_STAT	1.164E-06 (0.768)	-2.740E-07 (-0.620)	-4.368E-07 (-2.089)
C_FED	1.276E-06 (1.669)	1.691E-06 (5.906)	1.726E-06 (11.941)
R ²	0.558	0.859	0.982

WLS* weighted least squares (weight: POP^{2.1}); WLS† weighted least squares (weight: VEH^{2.1}).

TABLE 4.4

Regression of Operating Expenses per Passenger on Selected Independent Variables (*t*-Statistics in Parentheses)

Explanatory Variable	OLS	WLS*
Constant	2.362 (12.899)	1.965 (17.144)
POP	5.305E-06 (1.315)	3.532E-06 (3.123)
VEH	-1.480E-02 (-0.425)	-0.007 (-0.677)
EMP	7.339E-03 (0.470)	0.005 (1.018)
FUEL	-1.015E-05 (-2.467)	-1.028E-05 (-10.620)
C_LOC	1.750E-06 (1.307)	1.273E-06 (5.700)
C_STAT	5.460E-07 (0.183)	1.335E-06 (1.982)
C_FED	4.323E-08 (0.029)	7.767E-07 (1.794)
R ²	0.460	0.894

WLS* weighted least squares (weight: POP^{2.1}).

these indicators, both plots of the disturbances and the Goldfeld–Quandt test (at the 5% significance level) indicate the existence of heteroscedasticity. As a result, a series of WLS estimations were undertaken and the results appear in the same tables.

In the absence of a valid hypothesis regarding the variable(s) causing heteroscedasticity, a “wild goose chase” might ensue, as in this example. The most striking result, which applies to all three estimations and for all weights, is that the slope parameters (β) are (more) significant in the WLS estimation than in the OLS estimation. As previously noted, in the presence of heteroscedasticity, OLS estimators of standard errors are biased (and, one cannot foretell the direction of this bias). In this case study the bias is upward, implying

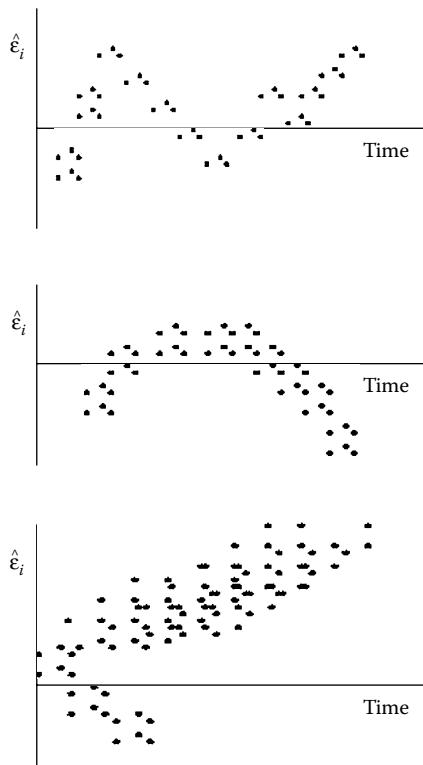
that it overestimates the standard error. One would have to accept the WLS estimates as more trustworthy because they have explicitly accounted for heteroscedasticity.

Besides the statistically related nature of these findings, there are important policy implications from these differences; for example, OLS results suggest the absence of a direct effect of operating subsidies on performance, whereas the WLS results suggest a generally degrading effect of subsidies on performance (with the exception of state-level subsidies on passengers per revenue vehicle miles). As a final point, WLS estimation does not completely alleviate heteroscedasticity; as such, it may be appropriate to use variance stabilizing transformations to fully account for this problem.

4.5 No Serial Correlation in the Disturbances Assumption

The assumption that disturbances corresponding to different observations are uncorrelated is frequently violated, particularly when modeling time-series data. When disturbances from different—usually adjacent time periods—are correlated, then the problem is called disturbance (error term) serial correlation or autocorrelation. Serial correlation frequently occurs in time-series studies mainly as a result of error term correlation over time; for example, consider an analysis of monthly time-series data involving the regression of transit output (vehicle miles) on labor, fuel, and capital inputs (a production function). If there is a labor strike affecting output (vehicle miles) in one month, there may be reason to believe that this disruption may affect output in subsequent month(s), leading to serial correlation in the disturbances. It is important to note that serial correlation may be the result of other problems, besides cyclical phenomena and “shocks,” such as model misspecification (explanatory variables omitted from the model), misspecified dynamics (a static model is estimated when dynamic dependence is present; see Section 7.2 for more details), and *ARCH* effects (error variance is autoregressive; see Section 7.3.1 for more details); these problems, although present in many transportation investigations, are largely ignored in both research and practice. Figure 4.3 shows a variety of patterns that correspond to autocorrelated disturbances.

Similar to serial correlation, spatial correlation may occur when data are taken over contiguous geographical areas sharing common attributes. For example, household incomes of residents on many city blocks are typically not unrelated to household incomes of residents from a neighboring block. Spatial statistics is an area of considerable research interest and potential. Readers interested in this scientific area should refer to Anselin (1988) for an in-depth treatment of the subject.

**FIGURE 4.3**

Patterns of autocorrelated disturbance.

To formalize the discussion on serial correlation, suppose the following regression model is analyzed:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (4.17)$$

where the usual assumption of independence for the disturbance ε_i is violated and instead of being random and independent, the disturbances are dependent as $\varepsilon_i = \rho \varepsilon_{i-1} + u_i$, and where u_i are independent normally distributed with constant variance. When disturbances in one time period are correlated with disturbances in (only) the previous time period, first-order serial correlation is examined. First-order autocorrelation is widely examined in empirical work because of its computational tractability and because corrective procedures are readily available in most statistical software; however, higher-order autocorrelation is prevalent in transportation (time-series) data and phenomena and should not be ignored. The strength of serial correlation depends on the parameter ρ , which is called the autocorrelation coefficient. When $\rho = 0$ the ε_i are independent; when $0 \leq \rho \leq 1$ ($-1 \leq \rho \leq 0$) the

ε_i are positively (negatively) correlated. In the case of positive serial correlation, successive error terms are given by

$$\begin{aligned}\varepsilon_i &= \rho\varepsilon_{i-1} + u_i \\ \varepsilon_{i-1} &= \rho\varepsilon_{i-2} + u_{i-1} \\ &\vdots \\ \varepsilon_{i-n+1} &= \rho\varepsilon_{i-n} + u_{i-n+1}\end{aligned}$$

Combining the first and second equations yields $\varepsilon_i = \rho(\rho\varepsilon_{i-2} + u_{i-2}) + u_i = \rho^2\varepsilon_{i-2} + \rho\varepsilon_{i-1} + u_i$, which implies that the process has “memory,” in that it “remembers” past conditions to some extent. The strength of this memory is reflected in ρ . Continuing the substitution and reversing the order of the terms yields

$$\varepsilon_i = u_i + \rho u_{i-1} + \rho^2 u_{i-2} + \cdots + \rho^n u_{i-n+1}$$

And, since $E(u_i) = E(u_{i-1}) = \cdots = E(u_{i-n+1}) = 0$ and $E(\varepsilon_i) = E(\varepsilon_{i-1}) = \cdots = E(\varepsilon_{i-n+1})$, then $VAR(\varepsilon_i) \neq VAR(u_i)$. Furthermore, $VAR(u_i) = \sigma_u^2$, and it follows that

$$\begin{aligned}\sigma_\varepsilon^2 &= VAR(u_i) + \rho VAR(u_{i-1}) + \rho^2 VAR(u_{i-2}) + \cdots + \rho^n VAR(u_{i-n+1}) \\ &= \sigma_u^2 (1 + \rho^2 + \rho^4 + \cdots + \rho^{2n}) \\ &= \frac{\sigma_u^2}{1 - \rho}\end{aligned}$$

It follows that under serial correlation the estimated disturbances variance $\hat{\sigma}_\varepsilon^2$ is larger than the true variance of the random independent disturbances σ_u^2 by a factor of $1/(1 - \rho^2)$. As such, the OLS estimates lose efficiency and are no longer BLUE. On the other hand, they are unbiased and consistent, similar to the case of heteroscedasticity.

4.5.1 Detecting Serial Correlation

The disturbances in a regression model may not have apparent correlation over time. In such cases a plot of the disturbances over time should appear scattered randomly around the zero line (Figure 4.2, right). If the disturbances show a trend over time, then serial correlation is obvious (Figure 4.3). Serial correlation is also diagnosed using the well-known and widely used Durbin–Watson (DW) test. The DW d statistic is computed as (Durbin and Watson 1951)

$$d = \frac{\sum_{i=1}^n (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^n \hat{\varepsilon}_i^2} \quad (4.18)$$

When the disturbances are independent, d should be approximately equal to 2. When the disturbances are positively (negatively) correlated, d should be less than (more than) 2. Therefore, this test for serial correlation is based on whether d is equal to 2 or not. Unfortunately, the critical values of d depend on the X_i , which vary across data sets. To overcome this inconvenience, Durbin and Watson established upper (d_U) and lower (d_L) bounds for this critical value (Figure 4.4). It is clear that if $d < d_L$ or $d > d_U$, the null hypothesis of no serial correlation is rejected. If $d_L < d < d_U$ the null hypothesis cannot be rejected. But what should be done when d falls in the inconclusive region? Although some procedures have been developed to examine these cases further, they have not been incorporated into standard statistical software. A prudent approach is to consider the inconclusive region as suggestive of serial correlation and take corrective measures. If the results after correction are the same as before correction, then the corrective process was probably not necessary. If the results are different, then the process was necessary.

Note here that the DW test is applicable for regression models with an intercept (constant) term and without a lagged dependent variable as a regressor. An alternative test, typically used in the presence of a lagged dependent variable(s), is Durbin's h statistic (Durbin 1970)

$$h = \left(1 - \frac{d}{2}\right) \sqrt{\frac{T}{1-T[VAR(\hat{\beta})]}} \quad (4.19)$$

where d is the reported DW statistic ($d \approx 2(1-\hat{\rho})$), T is the number of observations, and $VAR(\hat{\beta})$ is the square of the standard error of the parameter of the lagged dependent variable (note that the h test is not valid for $T[VAR(\hat{\beta})] > 1$). Durbin showed that the h statistic is approximately normally distributed with unit variance and, as a result, the test for first-order

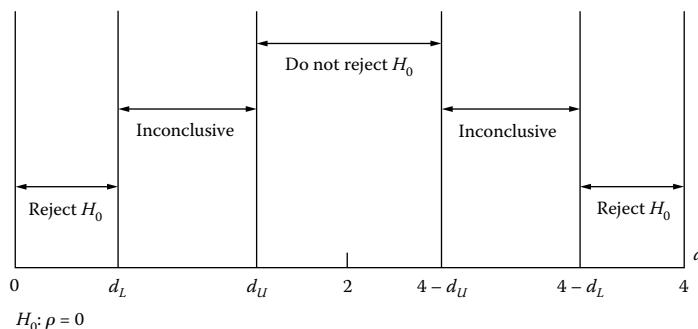


FIGURE 4.4
Durbin–Watson critical values.

serial correlation is conducted directly using the tables of the standard normal distribution (see Appendix C).

4.5.2 Correcting for Serial Correlation

As noted previously, serial correlation of the disturbances may be the result of omitted variables, of correlation over time, and a consequence of the nature of the phenomenon under study. In the first case, it is best to locate and include the missing variable in the model rather than to attempt other corrections. In the second case, the most common approach is to transform the original time-series variables in the regression so that the transformed model has independent disturbances. Consider again the regression model of Equation 4.17, with disturbances that "suffer" from first-order positive serial correlation (most of the discussion regarding autocorrelation and possible corrections concentrates on first-order serial correlation; readers interested in the theoretical aspects of higher-order serial correlation should refer to Shumway and Stoffer 2000),

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ \varepsilon_i &= \rho \varepsilon_{i-1} + u_i \end{aligned} \quad (4.20)$$

To remove serial correlation, Y_i and X_i are transformed into Y_i^* and X_i^* as follows:

$$Y_i^* = Y_i - \rho Y_{i-1} \quad \text{and} \quad X_i^* = X_i - \rho X_{i-1}, \quad i = 1, \dots,$$

The first time-period observations are transformed using $Y_1^* = \sqrt{1 - \rho^2} Y_1$ and $X_1^* = \sqrt{1 - \rho^2} X_1$ (many software packages, depending on the procedure used to estimate ρ , ignore the first observation although recent research has shown this to be inappropriate because it may lead to loss of information and adversely affect the results of the new regression, particularly for short time-series). The new regression model after the transformations is

$$Y_i^* = \beta_0 + \beta_1 X_i^* + u_i$$

where the disturbances u_i are independent. The critical aspect for the success of the previously mentioned approach is the estimation of ρ , which is not known in advance. If $\hat{\rho}$, the estimated autocorrelation parameter is a consistent estimate of ρ , then the corresponding estimates of β_0 and β_1 are asymptotically efficient. Numerous methods for estimating ρ are described in the literature and are readily available in standard statistical software packages; the most common ones are the Cochrane–Orcutt (1949) method, the Hildreth–Lu (1960) search procedure, the Durbin (1960) method, and the Beach–McKinnon (1978) maximum likelihood procedure.

Example 4.2

Monthly bus demand and socioeconomic data for Athens, Greece, were collected for the period between April 1995 and April 2000 to assess the sensitivity of bus travelers to public transport costs (fares), as well as to assess other factors that are hypothesized to influence demand (the variables available for this analysis appear in Table 4.5). Public transport demand cannot be directly estimated, so two different measures were used: monthly ticket and travel card sales; their sum approximates total transit demand. In general, two separate models, one for ticket and one for travel card sales should be estimated, to allow for the possibility of differential price elasticities between these two measures. In this example the latter is examined.

Table 4.6 presents the results using both OLS estimation and estimation with a correction for serial correlation (ρ was estimated using maximum likelihood and the first observation of the series was not discarded). Correction for serial correlation was done because the data reflect a monthly time series and the results of the DW d -test suggest that serial correlation is present. Table 4.6 indicates that before the correction for first-order serial correlation, the t statistics were biased upward (for the estimated parameters for *PRICE*, *NUMAUTOS*, and *INCOMEPERCAP*, for example), whereas others were downward biased (*VMT*, *UNEMPLOYMENT*, *METRO*). Note that the d statistic for this model was much higher after first-order correction (from 1.18 to 1.53).

Finally, because the data collected were monthly observations, besides first-order serial correlation that is readily available in most statistical packages, serial correlation from first to 12th order was investigated. As is shown in Table 4.7, the first- and 12th-order serial correlation parameters are clearly statistically significant and conceptually important; some others, such as the fifth and seventh order, are also statistically significant, but their practical interpretation is not clear. The results for higher-order correction are interesting, particularly when compared to the first-order correction. For example, in many cases such as for *VMT*, *NUMAUTOS*, and *INCOMEPERCAP*, the t -statistics yield values closer to those that resulted in

TABLE 4.5

Variables Available for Public Transport Demand Estimation Model

Variable Abbreviation	Explanation
<i>PRICE</i>	Ticket/travel card price in U.S. dollars
<i>VMT</i>	Total monthly vehicle miles traveled
<i>STRIKES</i>	Total monthly hours of strikes
<i>NUMAUTOS</i>	Number of automobiles
<i>UNEMPLOYMENT</i>	Unemployment (%)
<i>FUELPRICE</i>	Auto fuel price in U.S. dollars
<i>INCOMEPERCAP</i>	Income per capita in U.S. dollars per year
<i>TIMETREND</i>	Time trend
<i>MARCH</i>	Dummy variable for March
<i>AUGUST</i>	Dummy variable for August
<i>DEC</i>	Dummy variable December
<i>METRO</i>	Dummy variable for the introduction of two new Metro lines (0: before, 1: after the introduction)

TABLE 4.6

Regression Model Coefficient Estimates for Travel Card Sales in Athens
(*t*-Statistics in Parentheses)

Independent Variable	No Correction for Autocorrelation	MLE*	MLE†
Constant	-13.633 (-2.97)	-2.371 (-0.37)	-15.023 (-4.21)
PRICE	-0.764 (-6.42)	-0.771 (-4.96)	-0.765 (-9.84)
VMT	0.351 (1.54)	0.382 (2.12)	0.06 (0.35)
NUMAUTOS	2.346 (3.79)	0.509 (0.65)	2.713 (5.54)
UNEMPLOYMENT	0.033 (0.36)	0.165 (1.35)	0.059 (0.70)
FUELPRICE	-0.007 (-0.05)	-0.134 (-0.60)	-0.002 (-0.03)
INCOMEPERCAP	2.105 (3.35)	-0.013 (-0.02)	2.611 (5.51)
MARCH	0.024 (1.43)	0.006 (0.57)	0.021 (1.79)
AUGUST	-0.216 (-10.10)	-0.161 (-12.51)	-0.287 (-11.73)
DEC	0.026 (1.56)	0.004 (0.38)	0.072 (3.83)
METRO	-0.003 (-0.15)	0.042 (1.46)	-0.004 (-0.29)
First-order ρ^{\ddagger}	—	0.715 (7.71)	See Table 4.7
Durbin–Watson statistic	1.189	1.531	2.106
R ²	0.83	0.89	0.95

*MLE estimation of ρ for first-order serial correlation.

†MLE estimation of ρ for first- to 12th-order serial correlation.

‡ ρ ® autocorrelation coefficient.

TABLE 4.7

Estimates of Autoregressive Parameters
(Correction for Serial Correlation)

Lag	Coefficient	<i>t</i> -Statistic
1	0.344	2.65
2	0.17	1.33
3	-0.18	-1.42
4	-0.22	-1.70
5	0.24	1.86
6	0.14	1.12
7	-0.27	-1.95
8	-0.22	-1.68
9	0.09	0.70
10	0.01	0.10
11	-0.24	-1.76
12	0.33	2.54

the absence of first-order correction. As a result, the validity of the results for first-order correction is questioned. However, it must be stressed that for formal policy analyses a more in-depth and careful analysis of the disturbances autoregression parameters should be undertaken, because it is clear that the results are fairly sensitive to the order of the correction for serial correlation.

4.6 Model Specification Errors

In the preceding four sections, violations regarding the disturbance terms in the linear regression model were discussed. It was implicitly assumed that the estimated models represent correctly and capture mathematically the phenomena under investigation; statistically speaking, it was assumed that specification bias or specification errors were avoided. A specification error occurs when the functional form of a model is misspecified and instead of the “correct” model another model is estimated. Specification errors, which may result in severe consequences for the estimated model, usually arise in four ways (readers interested in tests for specification errors should refer to Godfrey 1988).

The first is when a relevant variable is omitted from the specified model (underfitting a model). Excluding an important variable from a model is frequently called omitted variable bias. To illustrate, suppose the correct model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (4.21)$$

but, instead of this model, the following is estimated:

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^* \quad (4.22)$$

If the omitted variable X_{2i} is correlated with included variable X_{1i} , then both β_0^* and β_1^* are biased because

$$\hat{\beta}_1^* = \beta_1 + \beta_2 \frac{COV(X_1, X_2)}{VAR(X_2)} \quad (4.23)$$

Furthermore, since the bias (the second term of Equation 4.23) does not diminish to zero as the sample size becomes larger (approaches infinity), the estimated parameters are inconsistent. However, in cases when X_{2i} is not correlated with the included variable X_{1i} ($COV(X_2, X_1) = 0$), then the estimated parameter β_1^* is unbiased and consistent, but β_0^* remains biased. In both cases, with correlated and uncorrelated variables, the estimated error variance of β_1^* from Equation 4.22 is a biased estimator of the variance of the true estimator β_1 and the usual confidence interval and hypothesis-testing procedures are unreliable (for proofs see Kmenta 1986).

The second source of misspecification error occurs when an irrelevant variable is included in the specified model (overfitting a model). Including an irrelevant variable in a model is frequently called irrelevant variable bias; for example, suppose the “correct” model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (4.24)$$

but instead the following model is estimated:

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \varepsilon_i^* \quad (4.25)$$

The inclusion of the irrelevant variable X_{2i} implies that the restriction $\beta_2^* = 0$ has not been considered. The OLS estimators for this incorrect model are unbiased because $E(\hat{\beta}_1^*) = \beta_1$, as well as consistent. However, the estimated β^* are inefficient (because the information that $\beta_2^* = 0$ is excluded), suggesting that their variances will generally be higher than the variances of the β terms, making OLS estimators LUE (linear unbiased estimators) but not BLUE.

The third source of misspecification error arises when an incorrect functional form for the model is used. Assume, for example, that the correct functional form is the one appearing in Equation 4.21 but, instead, the following is estimated:

$$LN(Y_i) = \beta_0^* + \beta_1^* LN(X_{1i}) + \beta_2^* X_{2i} + \varepsilon_i^* \quad (4.26)$$

The problem with specifying an incorrect functional form is similar to omitted variable bias: the parameter estimates are biased and inconsistent. When developing a statistical model, it is not only important that all theoretically relevant variables be included, but also that the functional form be correctly specified. Box–Cox transformations, discussed in Chapter 3, are a widely used approach for evaluating the appropriateness of the functional form of a model.

The fourth source of misspecification error arises when the independent variables are highly correlated (multicollinearity). In general, in a regression it is hoped that the explanatory variables are highly correlated with the dependent variable but, at the same time, are not highly correlated with other independent variables. The severity of the problem depends on the degree of multicollinearity. Low correlations among the regressors do not result in any serious deterioration in the quality of the estimated model. However, high correlations may result in highly unstable OLS estimates of the regression parameters. In short, in the presence of high to extreme multicollinearity the OLS estimators remain BLUE, but the standard errors of the estimated parameters are disproportionately large. As a result, the t test values for the parameter estimates are generally small and the null hypothesis that the parameters are zero may not be rejected, even though the associated variables are important in explaining the variation in the dependent variable. Finally, under multicollinearity the parameter estimates are usually unstable and, because of the high standard errors, reliable estimates of the regression parameters may be difficult to obtain. As a result, counterintuitive parameter signs are frequent and dropping or adding an independent variable to the model causes large changes in the magnitudes and directions of other independent variables (already present in the model).

5

Simultaneous-Equation Models

Some transportation data are best modeled by a system of interrelated equations. Examples include the interrelation of utilization of individual vehicles (measured in kilometers driven) in multivehicle households, the interrelation between travel time from home to an activity and the duration of the activity, the interrelation between the temperature at a sensor in a pavement and the temperature at adjacent sensors in the pavement, and the interrelation of average vehicle speeds by lane with the vehicle speeds in adjacent lanes. All of these examples produce interrelated systems of equations where the dependent variable in one equation is the independent variable in another.

Interrelated systems of equations create a potentially serious estimation problem if their interrelated structure is not considered. This problem arises because estimation of equation systems by the ordinary least squares (OLS) violates a key OLS assumption in that a correlation between regressors and disturbances is present because not all independent variables are fixed in random samples (one or more of the independent variables is endogenous and OLS estimates are erroneous). All too often, the general issue of endogeneity, such as that resulting from simultaneous-equation models, is ignored in the analysis of transportation data. The consequences of ignoring endogeneity are that erroneous conclusions and inferences will result.

5.1 Overview of the Simultaneous-Equations Problem

To illustrate a simultaneous-equations problem, consider the task of modeling the utilization of vehicles (kilometers driven per year) in households that own two vehicles. This problem represents a classic simultaneous-equations application because the usage of one vehicle is interrelated with the usage of the other (or others). A households' travel allocation is limited and, as a result, as one vehicle is used more, the other is used less (see Mannering 1983 for a detailed study of this problem). To show this problem, consider annual vehicle utilization equations (one for each vehicle) of the following linear form (suppressing the subscripting for each observation n to improve exposition),

$$u_1 = \beta_1 Z_1 + \alpha_1 X + \lambda_1 u_2 + \varepsilon_1 \quad (5.1)$$

$$u_2 = \beta_2 Z_2 + \alpha_2 X + \lambda_2 u_1 + \varepsilon_2 \quad (5.2)$$

where u_1 is the kilometers per year that vehicle 1 is driven, u_2 is the kilometers per year that vehicle 2 is driven, Z_1 and Z_2 are vectors of vehicle attributes (for vehicles 1 and 2, respectively), X is a vector of household characteristics, β , α are vectors of estimable parameters, λ are the estimable scalars, and ε are the disturbance terms.

If Equations 5.1 and 5.2 are estimated separately using OLS, one of the key assumptions giving rise to best linear unbiased estimators (BLUE) is violated because correlation between regressors and disturbances is present since an independent variable, u_2 in Equation 5.1 or u_1 in Equation 5.2 is not fixed in repeated samples. To be fixed in repeated samples, the value of the dependent variable (left-hand side variable) must not influence the value of an independent variable (right-hand side variable). This constraint is not the case in Equations 5.1 and 5.2 since in Equation 5.1, the independent variable u_2 varies as the dependent variable u_1 varies and in Equation 5.2, the independent variable u_1 varies as the dependent variable u_2 varies. Thus, u_2 and u_1 are said to be endogenous variables in Equations 5.1 and 5.2, respectively. As a result, estimating Equations 5.1 and 5.2 separately using OLS results in biased and inconsistent estimates of model parameters.

5.2 Reduced Form and the Identification Problem

To understand the issues involved in estimating simultaneous-equations model parameters, a natural starting point is to consider a reduced form solution. In Equations 5.1 and 5.2 the problem becomes a simple one of solving two equations and two unknowns to arrive at reduced forms. Substituting Equation 5.2 for u_2 in Equation 5.1 gives

$$u_1 = \beta_1 Z_1 + \alpha_1 X + \lambda_1 [\beta_2 Z_2 + \alpha_2 X + \lambda_2 u_1 + \varepsilon_2] + \varepsilon_1 \quad (5.3)$$

rearranging,

$$u_1 = \frac{\beta_1}{1 - \lambda_1 \lambda_2} Z_1 + \frac{\alpha_1 + \lambda_1 \alpha_2}{1 - \lambda_1 \lambda_2} X + \frac{\lambda_1 \beta_2}{1 - \lambda_1 \lambda_2} Z_2 + \frac{\lambda_1 \varepsilon_2 + \varepsilon_1}{1 - \lambda_1 \lambda_2} \quad (5.4)$$

and similarly substituting Equation 5.1 for u_1 in Equation 5.2 gives

$$u_2 = \frac{\boldsymbol{\beta}_2}{1 - \lambda_2 \lambda_1} \mathbf{Z}_2 + \frac{\boldsymbol{\alpha}_2 + \lambda_2 \boldsymbol{\alpha}_1}{1 - \lambda_2 \lambda_1} \mathbf{X} + \frac{\lambda_2 \boldsymbol{\beta}_1}{1 - \lambda_2 \lambda_1} \mathbf{Z}_1 + \frac{\lambda_2 \varepsilon_1 + \varepsilon_2}{1 - \lambda_2 \lambda_1} \quad (5.5)$$

Because the endogenous variables u_1 and u_2 are replaced by their exogenous determinants, Equations 5.4 and 5.5 are estimated using OLS as

$$u_1 = \boldsymbol{a}_1 \mathbf{Z}_1 + \boldsymbol{b}_1 \mathbf{X} + c_1 \mathbf{Z}_2 + \xi_1, \text{ and} \quad (5.6)$$

$$u_2 = \boldsymbol{a}_2 \mathbf{Z}_2 + \boldsymbol{b}_2 \mathbf{X} + c_2 \mathbf{Z}_1 + \xi_2 \quad (5.7)$$

where,

$$\boldsymbol{a}_1 = \frac{\boldsymbol{\beta}_1}{1 - \lambda_1 \lambda_2}; \quad \boldsymbol{b}_1 = \frac{\boldsymbol{\alpha}_1 + \lambda_1 \boldsymbol{\alpha}_2}{1 - \lambda_1 \lambda_2}; \quad c_1 = \frac{\lambda_1 \boldsymbol{\beta}_2}{1 - \lambda_1 \lambda_2}; \quad \xi_1 = \frac{\lambda_1 \varepsilon_2 + \varepsilon_1}{1 - \lambda_1 \lambda_2} \quad (5.8)$$

$$\boldsymbol{a}_2 = \frac{\boldsymbol{\beta}_2}{1 - \lambda_2 \lambda_1}; \quad \boldsymbol{b}_2 = \frac{\boldsymbol{\alpha}_2 + \lambda_2 \boldsymbol{\alpha}_1}{1 - \lambda_2 \lambda_1}; \quad c_2 = \frac{\lambda_2 \boldsymbol{\beta}_1}{1 - \lambda_2 \lambda_1}; \quad \xi_2 = \frac{\lambda_2 \varepsilon_1 + \varepsilon_2}{1 - \lambda_2 \lambda_1} \quad (5.9)$$

Ordinary least squares estimation of the reduced form models (Equations 5.6 and 5.7) is called indirect least squares (ILS) and is discussed later in this chapter. While estimated reduced form models are readily used for forecasting purposes, if inferences are to be drawn from the model system, the underlying parameters need to be determined. Unfortunately, uncovering the underlying parameters (the $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and λ), in reduced form models is problematic because either too little or too much information is often available. For example, note that Equations 5.8 and 5.9 provide two possible solutions for $\boldsymbol{\beta}_1$

$$\boldsymbol{\beta}_1 = \boldsymbol{a}_1 (1 - \lambda_1 \lambda_2) \quad \text{and} \quad \boldsymbol{\beta}_1 = \frac{c_2 (1 - \lambda_2 \lambda_1)}{\lambda_2} \quad (5.10)$$

The previous illustration reveals an important problem with the estimation of simultaneous equations—model identification. In some instances, it may be impossible to determine the underlying parameters. In these cases, the modeling system is said to be unidentified. In cases where exactly one equation solves the underlying parameters, the model system is said to be exactly identified. When more than one equation solves the underlying parameters (as shown in Equation 5.10), the model system is said to be overidentified.

A commonly applied rule to determine if an equation system is identified (exactly or overidentified) is the order condition. The order condition

determines an equation to be identified if the number of all variables excluded from an equation in an equation system is greater than or equal to the number of endogenous variables in the equation system minus one. For example, in Equation 5.1, the number of elements in the vector Z_2 , which is an exogenous vector excluded from the equation, must be greater than or equal to one because there are two endogenous variables in the equation system (u_1 and u_2).

5.3 Simultaneous-Equation Estimation

There are two broad classes of simultaneous-equation techniques—single-equations estimation methods and systems estimation methods. The distinction between the two is that systems methods consider all of the parameter restrictions (caused by overidentification) in the entire equation system and account for possible contemporaneous (cross-equation) correlation of disturbance terms. Contemporaneous disturbance-term correlation is an important consideration in estimation. For example, in the vehicle utilization equation system (Equations 5.1 and 5.2), one would expect ε_1 and ε_2 to be correlated because vehicles operated by the same household will share unobserved effects (common to that household) that influence vehicle utilization. Because system estimation approaches are able to utilize more information (parameter restrictions and contemporaneous correlation), they produce variance–covariance matrices that are at worst equal to, and in most cases smaller than, those produced by single-equation methods (resulting in lower standard errors and higher t -statistics for estimated model parameters).

Single-equation methods include instrumental variables (IV), ILS, two-stage least squares (2SLS), and limited information maximum likelihood (LIML). Systems estimation methods include three-stage least squares (3SLS) and full information maximum likelihood (FIML). These different estimation techniques are discussed below.

5.3.1 Single-Equation Methods

Indirect least squares involve OLS estimation of reduced form equations. In the case of vehicle utilization in two-vehicle households, this estimation requires OLS estimation depicted in Equations 5.6 and 5.7. The estimates of these reduced form parameters are used to determine the underlying model parameters solving Equations like 5.8 and 5.9. Because solving underlying model parameters tends to result in nonlinear equations (such as Equation 5.10), the unbiased estimates of the reduced form parameters (from Equations 5.6 and 5.7) do not produce unbiased estimates of the underlying model parameters. This result is because the ratio of unbiased parameters is also

biased. Moreover, there exists the problem of having multiple estimates of underlying model parameters if the equation system is over identified (as in Equation 5.10).

An IV approach is the most simplistic approach to solving the simultaneous-equations estimation problem. This approach simply replaces the endogenous variables on the right-hand side of the equations in the equation system with an instrumental variable—a variable that is highly correlated with the endogenous variable it replaces and is not correlated to the disturbance term. For example, in Equations 5.1 and 5.2, the IV approach is to replace u_2 and u_1 , respectively, with appropriate IV and estimate Equations 5.1 and 5.2 using OLS. This approach yields consistent parameter estimates. The problem, however, is one of finding suitable instruments, which is difficult to near impossible in many cases.

Two-stage least squares is an extension of IV in that it seeks the best instrument for endogenous variables in the equation system. Stage 1 regresses each endogenous variable on all exogenous variables. Stage 2 uses regression-estimated values from stage 1 as instruments, and estimates each equation using OLS. The resulting parameter estimates are consistent, and studies have shown that most small sample properties of 2SLS are superior to ILS and IV.

Limited information maximum likelihood maximizes the likelihood function of the reduced form models, generally assuming normally distributed error terms. Unlike ILS, the likelihood function is written to account for parameter restrictions (critical for over identified models). This parameterization alleviates the ILS problem of having multiple estimates of underlying model parameters in over identified equations.

In selecting among the single-equation estimation methods, 2SLS and LIML have obvious advantages in cases where the equations are over identified. In cases where the equations are exactly identified and the disturbances are normally distributed, ILS, IV, 2SLS, and LIML produce the same results. In over identified cases, 2SLS and LIML have the same asymptotic variance-covariance matrix so the choice becomes one of minimizing computational costs. A summary of single-equation estimation methods is provided in Table 5.1.

5.3.2 System-Equation Methods

System-equation methods are typically preferred to single-equation methods because they account for restrictions in overidentified equations and contemporaneous (cross-equation) disturbance-term correlation (the correlation of disturbance terms across the equation system). 3SLS is the most popular of the system-equation estimation methods. In 3SLS, stage 1 is to get the 2SLS estimates of the model system. In stage 2, the 2SLS estimates are used to compute residuals from which cross-equation disturbance-term correlations are calculated. In stage 3, generalized least squares (GLS) is used

TABLE 5.1
Summary of Single-Equation Estimation Methods for Simultaneous Equations

Method	Procedure	Resulting Parameter Estimates
Indirect least squares (ILS) Instrumental variables (IV)	Applies ordinary least squares to the reduced form models Uses an instrument (a variable that is highly correlated with the endogenous variable it replaces, but is not correlated to the disturbance term) to estimate individual equations	Consistent but not unbiased Consistent but not unbiased
Two-stage least squares (2SLS)	Approach finds the best instrument for endogenous variables. Stage 1 regresses each endogenous variable on all exogenous variables. Stage 2 uses regression-estimated values from stage 1 as instruments, and estimates equations with ordinary least squares	Consistent but not unbiased. Generally better small sample properties than ILS or IV
Limited information maximum likelihood (LIML)	Uses maximum likelihood to estimate reduced form models. Can incorporate parameter restrictions in over identified equations	Consistent but not unbiased. Has same asymptotic variance–covariance matrix as 2SLS

to compute parameter estimates. Appendix 5A provides an overview of the GLS-estimation procedure.

Because of the additional information considered (contemporaneous correlation of disturbances), 3SLS produces more efficient parameter estimates than single-equation estimation methods. An exception is when there is no contemporaneous disturbance-term correlation. In this case, 2SLS and 3SLS parameter estimates are identical.

Full information maximum likelihood extends LIML by accounting for contemporaneous correlation of disturbances. The assumption typically made for estimation is that the disturbances are multivariate normally distributed. Accounting for contemporaneous error correlation complicates the likelihood function considerably. As a result, FIML is seldom used in simultaneous-equation estimation. And, because under the assumption of multivariate normally distributed disturbances, FIML and 3SLS share the same asymptotic variance/covariance matrix, there is no real incentive to choose FIML over 3SLS in most applications. A summary of system-equation estimation methods is provided in Table 5.2.

Example 5.1

To demonstrate the application of a simultaneous-equations model, consider the problem of studying mean vehicle speeds by lane on a multilane freeway. Due to the natural interaction of traffic in adjacent lanes, a classic simultaneous-equations problem is appropriate because lane mean speeds are determined, in part, by the lane mean speeds in adjacent lanes. This problem was first studied by Shankar and Mannering (1998) and their data and approach are used herein.

For this example, data represent speeds obtained from a six-lane freeway with three lanes in each direction separated by a large median (each direction is considered separately). A summary of the available data is shown in Table 5.3. At the point where the data were gathered, highly variable seasonal weather conditions were present. As a consequence, seasonal factors are expected to play a role. The data were collected over a period of a year, and the mean speeds, by lane, were the mean of the spot speeds gathered over 1-hour periods. The equation system is written as (see Shankar and Mannering 1998)

$$s_R = \boldsymbol{\beta}_R \mathbf{Z}_R + \lambda_R s_C + \varepsilon_R \quad (5.11)$$

$$s_C = \boldsymbol{\beta}_C \mathbf{Z}_C + \lambda_C s_L + \tau_C s_R + \varepsilon_C \quad (5.12)$$

$$s_L = \boldsymbol{\beta}_L \mathbf{Z}_L + \lambda_L s_C + \varepsilon_L \quad (5.13)$$

where s are the mean speeds (over a 1-hour period in kilometers/hr) for the right-most lane (subscript R) relative to the direction of travel (the slow lane), the center lane (subscript C), and the left lane (subscript L), \mathbf{Z} are the vectors of exogenous variables influencing the mean speeds in the corresponding lanes, $\boldsymbol{\beta}$ are the

TABLE 5.2
Summary of System Estimation Methods for Simultaneous Equations

Method	Procedure	Resulting Parameter Estimates
Three-stage least squares (3SLS)	Stage 1 gets 2SLS estimates of the model system. Stage 2 uses the 2SLS estimates to compute residuals to determine cross-equation correlations. Stage 3 uses generalized least squares (GLS) to estimate model parameters	Consistent and more efficient than single-equation estimation methods.
Full information maximum likelihood (FIML)	Similar to LIML but accounts for contemporaneous correlation of disturbances in the likelihood function	Consistent and more efficient than single-equation estimation methods. Has same asymptotic variance-covariance matrix as 3SLS

TABLE 5.3
Lane Mean-Speed Model Variables

Variable No.	Description
1	Mean speed in the right lane in kilometers per hour (gathered over a 1-hour period)
2	Mean speed in the center lane in kilometers per hour (gathered over a 1-hour period)
3	Mean speed in the left lane in kilometers per hour (gathered over a 1-hour period)
4	Traffic flow in right lane (vehicles per hour)
5	Traffic flow in center lane (vehicles per hour)
6	Traffic flow in left lane (vehicles per hour)
7	Proportion of passenger cars (including pick-up trucks and minivans) in the right lane
8	Proportion of passenger cars (including pick-up trucks and minivans) in the center lane
9	Proportion of passenger cars (including pick-up trucks and minivans) in the left lane
10	Month that speed data was collected (1 = January, 2 = February, etc.)
11	Hour in which data was collected (the beginning hour of the 1-hour data collection period)

vectors of estimable parameters, λ and τ are the estimable scalars, and ε are the disturbance terms.

The equation system (Equations 5.11, 5.12, and 5.13) is estimated with 2SLS and 3SLS and the estimation results are presented in Table 5.4. These results show that there are noticeable differences between 2SLS and 3SLS parameter estimates in the equations for mean speeds in the right, center, and left lanes. These differences underscore the importance of accounting for contemporaneous correlation of disturbance terms. In this case, it is clear that the disturbances ε_R , ε_C , ε_L share unobserved factors occurring over the hour during which mean speeds are calculated. These unobserved factors, which are captured in equation disturbances, could include vehicle disablements, short-term driver distractions, weather changes, and so on. One would expect contemporaneous disturbance-term correlation to diminish if more complete data were available to estimate the model. Then, the difference between 2SLS and 3SLS parameter estimates would diminish. The interested reader should see Shankar and Mannering (1998) for a more detailed analysis of these mean-speed data.

TABLE 5.4
Simultaneous-Equation Estimation Results (*t*-Statistics in Parentheses)

Variable Description	Two-Stage Least Squares (2SLS)	Three-Stage Least Squares (3SLS)
<i>Dependent variable: Mean speed in the right lane (km/h)</i>		
Constant	-20.82 (-8.48)	-67.92 (-4.75)
Mean speed in the center lane (km/h)*	1.10 (55.6)	1.49 (12.8)
Winter: 1 if mean speeds were determined from speed data gathered during hours in the months of November, December, January, or February, 0 if not	-0.455 (-1.59)	-2.00 (-4.33)
Spring: 1 if mean speeds were determined from speed data gathered during hours in the months of March, April, or May, 0 if not	0.08 (0.28)	-1.49 (-3.25)
AM peak period: 1 if mean speeds were determined from speed data gathered during the hours from 7:00 AM to 10:00 AM, 0 if not	-1.99 (-6.44)	-0.29 (-1.48)
<i>Dependent variable: Mean speed in the center lane (km/h)</i>		
Constant	23.71 (12.52)	13.58 (6.32)
Mean speed in the right lane (km/h)*	0.23 (3.07)	0.38 (4.84)
Mean speed in the left lane (km/h)*	0.58 (8.77)	0.52 (7.69)
PM peak period: 1 if mean speeds were determined from speed data gathered during the hours from 5:00 PM to 7:00 PM, 0 if not	-1.79 (-5.29)	-0.57 (-5.40)
<i>Dependent variable: Mean speed in the left lane (km/h)</i>		
Constant	-23.40 (-8.43)	-20.56 (-5.19)
Mean speed in the center lane (km/h)*	1.21 (57.7)	1.20 (37.7)
Number of observations	6,362	6,362
R-squared—mean speed right lane equation	0.35	0.71
R-squared—mean speed center lane equation	0.32	0.90
R-squared—mean speed left lane equation	0.34	0.82
3SLS system R-squared	—	0.81

*Endogenous variable.

5.4 Seemingly Unrelated Equations

In some transportation-related studies it is possible to have a series of dependent variables that are considered a group, but do not have direct interaction as simultaneous equations do. An example is the normal driving speed of drivers on interstate highways that have posted speed limits of 55 mi/h, 65 mi/h, and 70 mi/h. Following research reported in Mannerling (2007), consider three equations that model a driver's normal driving speed relative to the speed limit under various speed limits; (1) number of miles per hour normally driven above the speed limit on an interstate with a 70 mi/h speed limit and little traffic (is negative if normally driven below the speed limit under these conditions), (2) number of miles per hour normally driven above the speed limit on an interstate with a 65 mi/h speed limit and little traffic, and (3) number of miles per hour normally driven above the speed limit on an interstate with a 55 mi/h speed limit and little traffic. To estimate a statistical model for each of these three questions, the use of OLS regressions is an obvious choice. Under standard linear regression assumptions, which includes the assumption that the model has all of the information relating to the regression equation and variables, estimated regression parameters for the three equations are unbiased and efficient (see Chapter 4). However, if some information is not taken into account, the properties relating to the unbiasedness and efficiency of estimated parameters cannot be determined. A classic example of potentially missing information is the knowledge that the disturbance term in one of these three regression equations is correlated with the disturbance term in another. This correlation exists for the three equations relating to normal driving speed at different speed limits because the unobserved factors that determine driving speed for each speed limit are likely to be highly correlated.

When studying the number of miles per hour normally driven above the speed limit at various speed limits, the following equation system is written (omitting subscripting for observation number):

$$S_{70} = \beta_{70} Z + \alpha_{70} X + \varepsilon_{70} \quad (5.14)$$

$$S_{65} = \beta_{65} Z + \alpha_{65} X + \varepsilon_{65} \quad (5.15)$$

$$S_{55} = \beta_{55} Z + \alpha_{55} X + \varepsilon_{55} \quad (5.16)$$

where S_{70} , S_{65} , and S_{55} are the number of miles per hour drivers normally drive above the speed limit (with little traffic) for 70, 65, and 55 mi/h speed limits, respectively (these variables can take on positive values if respondents normally drive above the speed limit and negative values if they normally drive below it); Z is the vector of driver and driver-household characteristics, X is the vector of driver preferences and opinions, β , α are the vectors

of estimable parameters, and ε are disturbance terms. Note that Equations 5.14, 5.15, and 5.16 do not directly interact with each other. That is, S_{70} does not directly determine S_{65} , S_{65} does not directly affect S_{55} , and so on, as one would expect in a classic simultaneous-equation system (such as that in Example 5.1). However, because all three responses represented by Equations 5.14, 5.15, and 5.16 are from the same driver, these equations are likely to share unobserved characteristics. In this case, the equations are seemingly unrelated but include a contemporaneous (cross-equation) correlation of error terms. If Equations 5.14, 5.15, and 5.16 are estimated separately by OLS, the parameter estimates are consistent but not efficient. To obtain efficient estimates, the contemporaneous correlation of disturbances ε_1 , ε_2 , and ε_3 must be considered and the common approach for doing so is seemingly unrelated regression estimation (SURE) as developed by Zellner (1962). Estimation of seemingly unrelated equations is accomplished by using GLS as is done in the third stage of 3SLS. Please see Appendix 5A for further details on GLS estimation.

Example 5.2

To demonstrate the application of a simultaneous-equations model, consider the problem discussed above and represented by Equations 5.14, 5.15, and 5.16. To estimate these models, data are available from a sample of licensed drivers who indicated they regularly drove on Indiana freeways (see Mannering 2007 for further details). The sample consists of undergraduate and graduate students at Purdue, primarily from engineering disciplines. A total of 204 drivers provided complete information. Of these 204 respondents, 194 reported normally driving above a 55 mi/h speed limit, 190 reported normally driving above a 65 mi/h speed limit, and 178 reported normally driving above a 70 mi/h speed limit. Only one person reported they normally drove below the 55 mi/h speed limit (nine reported they drive at the 55 mi/h speed limit), three people normally drove below the 65 mi/h speed limit (11 reported they drive at the 65 mi/h speed limit), and six people drove below the 70 mi/h speed limit (20 reported they drive at the 70 mi/h speed limit). Summary statistics for the data are presented in Table 5.5.

The equation system (Equations 5.14 through 5.16) is estimated with SURE and estimation results are presented in Table 5.6. These results show that the variables are of plausible sign and generally significant. If these same models are estimated as separate OLS regressions, the results are quite different (see Mannering 2007 for a further discussion of the model presented in Table 5.6).

5.5 Applications of Simultaneous Equations to Transportation Data

Although there are many transportation problems that lend themselves naturally to simultaneous-equations approaches, there have been surprisingly

TABLE 5.5

Sample Statistics for Data Used in Example 5.2 (Standard Deviation in Parentheses when Appropriate)

Variable Description	Values
Percent believing Indiana's recently raised the speed limits from 65 mi/h to 70 mi/h is: too fast/about right/still too slow	2/72/26
Average normal driving speed on an interstate with a 55 mi/h speed limit and little traffic	65.92 (6.24)
Average normal driving speed on an interstate with a 65 mi/h speed limit and little traffic	74.05 (5.03)
Average normal driving speed on an interstate with a 70 mi/h speed limit and little traffic	77.88 (5.24)
Percent rating the quality of pavements on Indiana interstates as: poor/fair/good/very good/don't know	11/30/44/9/6
Percent rating the quality of pavements on Indiana interstates as: worse than adjacent states/about the same/better than adjacent states/don't know	12/45/16/27
Percent believing the following luxury car brands provide the most prestige: Acura/Audi/BMW/Cadillac/Infiniti/Jaguar/Lexus/Lincoln/Mercedes Benz	1/6/23/4/2/19/10/2/33
Percent: female/male	26/74
Percent: married/single/separated/divorced/other	25/71/0/0/4
Average age	25.00 (6.46)
Percent with highest completed level of education: some high school/high school diploma/technical college degree/college degree/postgraduate degree	1/48/4/29/18
Percent with annual household income as: no income/under \$10,000/\$10,000–\$19,999/\$20,000–\$29,999/\$30,000–\$39,999/\$40,000–\$49,999/\$50,000–\$74,999/\$75,000–\$100,000/over \$100,000	4/1/16/8/6/5/17/16/27
Average number of people living in household	3.52 (1.52)
Average number of children in household that are under age 6	0.18 (0.50)
Average number of children in household that are aged 6–16	0.26 (0.60)
Average number of people in household that work outside the home	1.91 (1.17)
Average number of licensed motor vehicles in household	2.82 (1.41)
Average number of years licensed	7.04 (6.11)

few transportation applications. The most common example is the utilization of individual vehicles in multivehicle households. This problem has been studied using 3SLS by Mannering (1983) and Greene and Hu (1984). Other examples include individuals' travel time to trip-generating activities, such as shopping and the length of time they are involved in that activity (Hamed and Mannering 1993). There are countless applications of simultaneous equations in the general economic literature. The reader is referred to Pindyck and Rubinfeld (1997), Kmenta (1997), Kennedy (1998), and Greene (2007) for additional information and examples.

TABLE 5.6

Seemingly Unrelated Regression Equation (SURE) Estimation Results for the Number of Miles Per Hour Above the Speed Limit Drivers Report as Their Usual Speed (*t*-Statistics in Parentheses)

Variable	Estimated Parameter (55 mi/h speed limit)	Estimated Parameter (65 mi/h speed limit)	Estimated Parameter (70 mi/h speed limit)	Estimated Parameter (70 mi/h speed limit)
Constant	11.11 (12.37)	10.88 (10.27)	10.88 (10.27)	10.86 (8.44)
<i>Driver/household attributes</i>				
Male indicator (1 if driver is male, 0 otherwise)	—	0.470 (1.37)	—	—
Driver age (years)	—	-0.088 (-2.47)	-0.129 (-2.83)	-0.129 (-2.83)
High-income indicator (1 if household's total annual income is \$75,000 or greater, 0 otherwise)	1.958 (2.06)	2.099 (2.71)	2.099 (2.71)	1.255 (1.52)
Low-income indicator (1 if household's total annual income is less than \$30,000, 0 otherwise)	-0.763 (-1.17)	—	—	—
Number of children under the age of 6 years in the household	-1.207 (-1.54)	-0.993 (-1.48)	-1.359 (-1.87)	-1.359 (-1.87)
Late-license indicator (1 if driver was first licensed at age 17 or greater, 0 otherwise)	-3.851 (-4.38)	-2.148 (-3.00)	-2.148 (-3.00)	-1.919 (-2.50)
<i>Driver opinions</i>				
Good pavement indicator (1 if driver believes pavement quality on Indiana interstates is good or very good, 0 otherwise)	1.160 (1.46)	1.067 (1.62)	1.006 (1.43)	1.006 (1.43)
German prestige indicator (1 if driver believes if German-brand vehicles are the most prestigious, 0 otherwise)	1.137 (1.85)	—	—	0.582 (1.35)
Japanese-prestige indicator (1 if driver believes if Japanese-brand vehicles are the most prestigious, 0 otherwise)	-1.348 (-1.25)	-0.814 (-1.35)	—	—
R-squared	0.226	0.200	0.170	0.170
Number of observations	195	—	—	—
Equation system R-squared	0.202	—	—	—

Appendix 5A. A Note on GLS Estimation

Ordinary least squares assumptions are that disturbance terms have equal variances and are not correlated. GLS is used to relax these OLS assumptions. Under OLS assumptions, in matrix notation,

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \mathbf{I} \quad (5A.1)$$

where $E(\cdot)$ denotes expected value, $\boldsymbol{\epsilon}$ is an $n \times 1$ column vector of equation disturbance terms (where n is the total number of observations in the data), $\boldsymbol{\epsilon}^T$ is the $1 \times n$ transpose of $\boldsymbol{\epsilon}$, σ^2 is the disturbance term variance, and \mathbf{I} is the $n \times n$ identity matrix,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & 1 \end{bmatrix} \quad (5A.2)$$

When heteroskedasticity is present, $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is $n \times n$ matrix,

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & . & 0 \\ 0 & \sigma_2^2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \sigma_n^2 \end{bmatrix} \quad (5A.3)$$

For disturbance-term correlation $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \boldsymbol{\Omega}$, where

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho & . & \rho^{N-1} \\ \rho & 1 & . & \rho^{N-2} \\ . & . & . & . \\ \rho^{N-1} & \rho^{N-2} & . & 1 \end{bmatrix} \quad (5A.4)$$

Recall that under OLS assumptions the disturbance terms have equal variances and are not correlated, resulting in parameters being estimated as,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (5.17)$$

where $\hat{\boldsymbol{\beta}}$ is a $p \times 1$ column vector (where p is the number of parameters), \mathbf{X} is a $n \times p$ matrix of data (where n is the number of observations), \mathbf{X}^T is

the transpose of \mathbf{X} , and \mathbf{Y} is an $n \times 1$ column vector. GLS generalizes this expression by using a matrix that considers for correlation among equation error terms ($\boldsymbol{\Omega}$), Equation 4 is rewritten as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{Y} \quad (5.18)$$

The most difficult aspect of GLS estimation is obtaining an estimate of the $\boldsymbol{\Omega}$ matrix. In 3SLS, it is estimated using the initial 2SLS parameter estimates. In SURE, $\boldsymbol{\Omega}$ is estimated from initial OLS estimates of individual equations.

6

Panel Data Analysis

Traditionally, statistical and econometric models have been estimated using cross-sectional or time-series data. In an increasing number of applications, however, there is availability of data based on cross sections of individuals observed over time (or other observational units such as firms, geographic entities, etc.). These data, which combine cross-sectional and time-series characteristics, are panel (or pooled) data, and allow researchers to construct and test realistic behavioral models that cannot be identified using only cross-sectional or time-series data. The modeling of panel data raises new specification issues, however, such as heterogeneity, which, if not explicitly accounted for, may lead to model parameters that are inconsistent and/or meaningless. This chapter focuses on the development of panel data regression models that account for heterogeneity in a variety of ways. Furthermore, this chapter discusses issues related to the possible distortions from both the cross-sectional (heteroscedasticity) and time-series (serial correlation) dimensions to which panel data are vulnerable.

6.1 Issues in Panel Data Analysis

Panel data analyses provide a number of advantages over analyses based solely on cross-sectional or time-series data (Hsiao 1986). First, from a statistical perspective, by increasing the number of observations, panel data have higher degrees of freedom and less collinearity, particularly in comparison with time-series data, thus improving the efficiency of the parameter estimates. Second, panel data allow a researcher to analyze questions that cannot be adequately addressed using either time-series or cross-sectional data. For example, suppose that a cross-section of public transit agencies reveals that, on average, public transit subsidies are associated with 20% increased ridership. A homogeneous population of public transit firms might be construed to imply that a firm's ridership will increase by 20% given transit subsidies. This implication might also apply to a sample of heterogeneous firms. However, an alternative explanation in a sample of heterogeneous public transit firms is that the subsidies have no effect (0% increase) on four-fifths of the firms, and raise ridership by 100% on one-fifth of the firms. Although

these competing hypotheses cannot be tested using a cross-sectional sample (in the absence of a cross-sectional variable that “explains” this difference), it is possible to test between them by identifying the effect of subsidies on a cross-section of time series for the different transit systems. Third, panel data allow researchers to test whether more simplistic specifications are appropriate. With panel data, it is possible to account for cross-sectional heterogeneity by introducing additional parameters into the model. Thus, testing for cross-sectional homogeneity is equivalent to testing the null hypothesis that these additional parameters are equal to zero.

Compared with cross-sectional or time-series data, panel data raise new specification issues that need to be considered during analysis. The most important of these is heterogeneity bias (Hausman and Taylor 1981). Heterogeneity refers to the differences across cross-sectional units that may not be appropriately reflected in the available data—that is, an existing explanatory variable or variables. If heterogeneity across cross-sectional units is not accounted for in a statistical model, estimated parameters are biased because they capture part of the heterogeneity. Indeed, as noted by Greene (2000), cross-sectional heterogeneity should be the central focus of panel data analysis.

A second issue is serial correlation of the disturbance terms, which occurs in time-series studies when the disturbances associated with observations in one time period are dependent on disturbances from prior time periods. For example, positive serial correlation frequently occurs in time-series studies, either because of correlation in the disturbances or because of a high degree of temporal correlation in the cumulative effects of omitted variables (serial correlation is discussed in depth in Section 4.5). Although serial correlation does not affect the unbiasedness or consistency of regression parameter estimates, it does affect their efficiency. In the case of positive serial correlation, the estimates of the standard errors are smaller than the true standard errors. In other words, the regression estimates are unbiased but their associated standard errors are biased downward, which biases the t -statistics upward and increases the likelihood of rejecting a true null hypothesis. A third issue is heteroscedasticity, which refers to the variance of the disturbances not being constant across observations. As discussed in Chapter 4 (Section 4.4), this problem arises in many practical applications, particularly when cross-sectional data are analyzed, and, similar to serial correlation, it affects the efficiency of the estimated parameters.

Notwithstanding the advantages associated with panel data, failure to account for cross-sectional heterogeneity, serial correlation and heteroscedasticity could seriously affect the estimated parameters and lead to inappropriate statistical inferences.

6.2 One-Way Error Component Models

Variable-intercept models across individuals or time (one-way models) and across both individuals and time (two-way models), the simplest and most straightforward models for accounting for cross-sectional heterogeneity in panel data, arise when the null hypothesis of overall homogeneity is rejected. The variable-intercept model assumes that the effects of omitted variables may be individually unimportant but are collectively significant, and thus is considered to be a random variable that is independent of included independent variables. Examples of error-component models in the transportation literature include Dee (1999), Eskelanda and Feyziolub (1997), Loizides and Tsionas (2002), and Chu and Durango-Cohen (2008). Because heterogeneity effects are assumed to be constant for given cross-sectional units or for different cross-sectional units during one time period, they are absorbed by the intercept term as a means to account for individual or time heterogeneity (Hsiao 1986). More formally, a panel data regression is written as

$$Y_{it} = \alpha + \boldsymbol{\beta} \mathbf{X}_{it} + u_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (6.1)$$

where i refers to the cross-sectional units (individuals, counties, states, etc.), t refers to the time periods, α is a scalar, $\boldsymbol{\beta}$ is a $P \times 1$ vector, and \mathbf{X}_{it} is the i^{th} observation on the P^{th} explanatory variable. A one-way error component model for the disturbances, which is the most commonly utilized panel data formulation, is specified as

$$u_{it} = \mu_i + v_{it} \quad (6.2)$$

where μ_i is the unobserved cross-sectional specific effect and v_{it} are random disturbances. For example, in an investigation of the effects of operating subsidies on the performance of transit systems, Y_{it} might be a performance indicator for transit system i for year t , \mathbf{X}_{it} a vector of explanatory variables, including operating subsidies, and the μ_i capture the unobservable transit system-specific effects such as, say, managerial skills of transit system administrators.

When the μ_i are assumed to be fixed parameters to be estimated and the v_{it} are random disturbances that follow the usual regression assumptions, then combining Equations 6.1 and 6.2 yields the following model where inference is conditional on the particular n cross-sectional units that are observed (transit systems), and is thus called a fixed-effects model:

$$Y_{it} = \alpha + \boldsymbol{\beta} \mathbf{X}_{it} + \mu_i + v_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (6.3)$$

on which ordinary least squares (OLS), which provide best linear unbiased estimators (BLUE), are used to obtain α , β , and μ_i . Note that, particularly when n is large, many indicator variables are included in the model and the matrices to be inverted by OLS are of large dimension $(n + P)$. As such, a least squares dummy variable (LSDV) estimator for Equation 6.3 is obtained for β (this estimator is also called the within-group estimator because only the variation within each group is utilized in forming the estimator). Testing for the joint significance of the included fixed effects parameters (the dummy variables), is straightforwardly conducted using the Chow F test

$$F_0 = \frac{(\text{RRSS} - \text{URSS})/(n-1)^{H_0}}{\text{URSS}/(NT - n - P)} \sim F_{n-1, n(T-1)-P} \quad (6.4)$$

where RRSS are the restricted residual sums of squares from OLS on the pooled model and URSS are the unrestricted residual sums of squares from the LSDV regression.

The fixed-effects specification suffers from an obvious shortcoming in that it requires the estimation of many parameters and the associated loss of degrees of freedom. This shortcoming is avoided if the μ_i are considered to be random variables, such that $\mu_i \sim IID(0, \sigma_\mu^2)$, $v_{it} \sim IID(0, \sigma_v^2)$, the μ_i and v_{it} are independent, and the X_{it} are independent of the μ_i and $v_{it} \forall i, t$.

Unlike the fixed-effects model where inferences are conditional on the particular cross-sectional units sampled, an alternative formulation, called the random-effects model, is an appropriate specification if n cross-sectional units are randomly drawn from a large population. Furthermore, it is straightforward to show that a random-effects specification implies a homoscedastic disturbances variance $VAR(u_{it}) = \sigma_\mu^2 + \sigma_v^2$ for all i, t , and serial correlation only for disturbances of the same cross-sectional unit (Hsiao 1986). In general, the following equations apply to random-effects models

$$\begin{aligned} COV(u_{it}, u_{js}) &= \sigma_\mu^2 + \sigma_v^2 && \text{for } i = j, t = s \\ &= \sigma_\mu^2 && \text{for } i = j, t \neq s \\ &= 0 && \text{otherwise} \end{aligned}$$

and

$$\begin{aligned} COR(u_{it}, u_{js}) &= 1 && \text{for } i = j, t = s \\ &= \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_v^2} && \text{for } i = j, t \neq s \\ &= 0 && \text{otherwise} \end{aligned}$$

To obtain a generalized least squares (GLS, see Chapter 5) estimator of the random-effects regression parameters, the disturbance variance–covariance matrix Ω from Equation 6.3 needs to be estimated. This matrix is typically large, and the estimation is successful only with a small n and T (Baltagi 1985). As a result, this estimation is typically accomplished using either one of a variety of modified GLS approaches (see, for example, Nerlove 1971b; Swamy and Arora 1972; Taylor 1980; and Wansbeek and Kapteyn 1982, 1983) or using maximum likelihood estimation (Breusch 1987).

The existence of two fundamentally different model specifications for accounting for heterogeneity in panel data analyses raises the obvious question of which specification—fixed or random effects—is preferred in any particular application. Despite the research interest this issue has attracted, it is not a straightforward question to answer, especially when one considers that the selection of whether to treat the effects as fixed or random can make a significant difference in the estimates of the parameters (see Hausman 1978 for an example). The most important issue when considering these alternative specifications is the context of the analysis. In the fixed-effects model, inferences are conditional on the effects that are in the sample, while in the random-effects model inferences are made unconditionally with respect to the population of effects (Hsiao 1986). In other words, the essential difference between these two modeling specifications is whether the inferences from the estimated model are confined to the effects in the model or whether the inferences are made about a population of effects (from which the effects in the model are a random sample). In the former case the fixed-effects model is appropriate, whereas the latter is suited for the random-effects model.

Although the choice between the two different models is based on the a priori hypotheses regarding the phenomenon under investigation, the fixed-effects model has a considerable virtue in that it does not assume that the individual effects are uncorrelated with the regressors $E(u_{it} | \mathbf{X}_{it}) = 0$, as is assumed by the random-effects model. In fact, the random-effects model may be biased and inconsistent due to omitted variables (Chamberlain 1978; Hausman and Taylor 1981). With the intent of identifying potential correlation between the individual effects and the regressors, Hausman (1978) devised a test to examine the null hypothesis of no correlation between the individual effects and \mathbf{X}_{it} . This test assumes that under the null hypothesis both the LSDV and the GLS are consistent and asymptotically efficient, whereas under the alternative hypothesis the GLS is biased and inconsistent for β , but the LSDV remains unbiased and consistent. As a result, the test is based on the difference $\hat{\mathbf{d}} = \hat{\beta}_{\text{GLS}} - \hat{\beta}_{\text{LSDV}}$ and the assumption that $\text{COV}(\hat{\beta}_{\text{GLS}}, \hat{\mathbf{d}}) = \mathbf{0}$. The test statistic is given by

$$h = \hat{\mathbf{d}}' [\text{VAR}(\hat{\mathbf{d}})]^{-1} \hat{\mathbf{d}}$$

which, under the null hypothesis, is asymptotically distributed as χ^2_P , where P denotes the dimension of β (note that the Hausman test is not valid for heteroscedastic or serially correlated disturbances; an alternative Wald statistic for the Hausman test, not sensitive to heteroscedasticity or serial correlation, was derived by Arellano 1993). Hausman's test is not a tool for conclusively deciding between the fixed- and random-effects specifications. A rejection of the null hypothesis of no correlation suggests the possible inconsistency of the random-effects model and the possible preference for a fixed-effects specification.

6.2.1 Heteroscedasticity and Serial Correlation

As previously noted, the disturbance components of the models in Equations 6.1 and 6.2 are vulnerable to distortions from both the cross-sectional and time-series dimensions, which, as discussed in Chapter 4, amount to problems associated with heteroscedasticity and serial correlation. In both cases the parameter estimates are unbiased and consistent, but are not efficient.

In the case of heteroscedasticity, the homoscedastic error-component model of Equation 6.2 is generalized to have a heteroscedastic μ_i ; that is, $\mu_i \sim (0, w_i^2)$, $i = 1, \dots, n$, but $v_{it} \sim IID(0, \sigma_v^2)$. To correct for heteroscedasticity weighted least squares (WLS) is applied (see Chapter 4). However, the feasibility of the WLS estimates in these cases requires estimates of σ_v^2 and w_i^2 for $i = 1, \dots, n$, which requires large T , small n , and $T \gg n$ (Baltagi 2008). As a result, the application of WLS in panel data is not as straightforward as in OLS regression. To this end, Baltagi and Griffin (1988) derived a GLS approach for estimating heteroscedastic disturbances in panel data and Magnus (1982) developed a maximum likelihood approach that assumes normality.

The disturbances in the model shown in Equation 6.3 are assumed to be correlated over time due to the presence of the same cross-sectional units observed over time. This correlation was shown to be $COR(u_{it}, u_{is}) = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_\epsilon^2)$, for $i = j, t = s$. This assumption is not ideal for the basic error component model because it may result in unbiased and consistent but inefficient parameter estimates. Lillard and Willis (1978) generalized the error-component model to the case of first-order serially correlated disturbances as follows: $v_{it} = \rho v_{i,t-1} + \epsilon_{it}$, where $|\rho| < 1$, $\mu_i \sim IID(0, \sigma_\mu^2)$, and $\epsilon_{it} \sim IID(0, \sigma_\epsilon^2)$. Furthermore, the μ_i are independent from v_{it} . Baltagi and Li (1991) derived the appropriate matrix for transforming the autocorrelated disturbances into random terms (disturbances). When using panel data, the transformation has to be applied for all n cross-sectional units individually, leading to a potentially significant loss of degrees of freedom, especially when n is large, if the first time period observations are discarded and not transformed using $Y_1^* = \sqrt{1-\rho^2} Y_1$ and $X_1^* = \sqrt{1-\rho^2} X_1$ (see Chapter 4 for more details on

this transformation). Note that corrections for second- and fourth-order serial correlation are found in Baltagi (2008).

6.3 Two-Way Error Component Models

The disturbances presented in Equation 6.2 are further generalized to include time-specific effects. As Wallace and Hussain (1969), Nerlove (1971b), and Amemiya (1971) suggest, this generalization is called a two-way error components model, whose disturbances are written as

$$u_{it} = \mu_i + \lambda_t + v_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (6.5)$$

where μ_i is the unobserved cross-sectional specific effect discussed in the previous section, λ_t denotes the unobservable time effects, and v_{it} are random disturbances. Note here that λ_t is the individual invariant and accounts for any time-specific effect that is not included in the regression. For example, it could account for strike year effects that disrupt transit service, and so on.

When the μ_i and λ_t are assumed to be fixed parameters to be estimated and v_{it} are random disturbances that follow the usual regression assumptions, combining Equations 6.1 and 6.5 yields a model where inferences are conditional on the particular n cross-sectional units (transit systems) and are to be made over the specific time period of observation. This model is called a two-way fixed effects error component model and is given as

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \lambda_t + v_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (6.6)$$

where X_{it} are assumed independent of the v_{it} for all i, t . Inference for this two-way fixed-effects model is conditional on the particular n individuals and over the T time periods of observation. Similar to the one-way fixed-effects model, the computational difficulties involved with obtaining the OLS estimates for β are circumvented by applying the within transformation of Wallace and Hussain (1969). Also, similar to the one-way model, testing for the joint significance of the included cross-sectional and time period fixed-effects parameters (the dummy variables) is straightforwardly computed using an F test

$$F_0 = \frac{(\text{RRSS} - \text{URSS}) / ((n+T-2)^{H_0})}{\text{URSS} / ((n-1)(T-1)-P)} \sim F_{(n+T-2), (n-1)(T-1)-P} \quad (6.7)$$

where RRSS are the restricted residual sums of squares from OLS on the pooled model and URSS are the unrestricted residual sums of squares from the regression using the within transformation of Wallace and Hussain (1969). As a next step, the existence of individual effects given time effects can also be tested (Baltagi 1995).

Similar to the one-way error-component model case, if both the μ_i and λ_t are random with $\mu_i \sim IID(0, \sigma_\mu^2)$, $\lambda_t \sim IID(0, \sigma_\lambda^2)$, $v_{it} \sim IID(0, \sigma_v^2)$, the μ_i , λ_t , and v_{it} independent and the X_{it} independent of the μ_i , λ_t , and $v_{it} \forall i, t$, then this formulation is called the two-way random-effects model. Furthermore, it straightforward to show that a random-effects specification implies a homoscedastic disturbance variance, where $VAR(u_{it}) = \sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2$ for all i, t , and

$$\begin{aligned} COV(u_{it}, u_{js}) &= \sigma_\mu^2 && \text{for } i = j, t \neq s \\ &= \sigma_\lambda^2 && \text{for } i \neq j, t = s \\ &= 0 && \text{otherwise} \end{aligned}$$

and

$$\begin{aligned} COR(u_{it}, u_{js}) &= \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2} && \text{for } i = j, t \neq s \\ &= \frac{\sigma_\lambda^2}{\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2} && \text{for } i \neq j, t = s \\ &= 1 && \text{for } i = j, t = s \\ &= 0 && \text{for } i \neq j, t \neq s \end{aligned}$$

Estimation of the two-way random-effects model is typically accomplished using the GLS estimators of Wallace and Hussain (1969) and Amemiya (1971), or by using maximum likelihood estimation (Baltagi and Li 1992). For this model specification, Breusch and Pagan (1979) derived a Lagrange-multiplier test for the null hypothesis $H_0: \sigma_\mu^2 = \sigma_\lambda^2 = 0$; this test is based on the normality of the disturbances and is fairly straightforward to compute (for details, see Breusch and Pagan 1979). Finally, for the two-way error component models, Hausman's (1978) test is still based on the difference between the fixed-effects estimator, including both cross-sectional and time indicator variables and the two-way

random effects GLS estimator. However, Kang (1985) showed the test is not valid for the two-way error component model and proposed a modified Hausman test.

Example 6.1

The effectiveness of safety belt use in reducing motor vehicle-related fatalities has been the subject of much research interest in the past few years (for more details, see Derrig et al. 2002). To investigate the hypothesized relationship between various exogenous factors, including seat belt usage rates and traffic fatalities, Derrig et al. (2002) compiled a panel data set of demographic, socio-economic, political, insurance, and roadway variables for all 50 U.S. states over a 14-year period (1983 through 1996). This data set was subsequently enriched with additional information by R. Noland of Rutgers University, and is used in this analysis (for additional information, see Noland and Karlaftis (2005)). The variables used in the analysis appear in Table 6.1.

In this case study, one potential accident indicator variable, total fatalities per vehicle-miles traveled, is used and is regressed on a number of independent variables. For the purposes of this case study, a number of alternative modeling formulations were investigated: OLS regression (no correction for panel effects); a fixed-effects model (state effects); a two-way fixed-effects model (state and time effects); a random-effects model; a fixed-effects model with correction for first-order serial correlation; and a random-effects model with correction for first-order serial correlation. The parameter estimates and corresponding *t*-statistics appear in Table 6.2. Interestingly, the results and the significance of the parameter estimates in particular show ample variation between the different specifications. For example, the parameter for percent seat belt use (*seatbelt*) is significant at the 99% level for both the fixed- and random-effects specifications but loses much of this significance when incorporating a two-way fixed-effects model or a correction for serial correlation (indicating that without correction for serial correlation the standard error of the parameters was downward biased; that is, the models without correction underestimated the standard error). On the other hand, the parameters for the hospitals per square miles (*HOSPAREA*) variable are significant for both random-effects specifications but are not significant for any of the fixed-effects formulations.

This at times highly debated issue highlights the importance of model selection. As previously discussed, when inferences are confined to the effects in the model, the effects are more appropriately considered to be fixed. When inferences are made about a population of effects from which those in the data are considered to be a random sample, then the effects should be considered random. In these cases, a fixed-effects model is defended on grounds that inferences are confined to the sample. In favor of selecting the fixed-effects rather than the random-effects formulation were the Hausman tests presented in Table 6.3.

TABLE 6.1

Variables Available for Analysis

Variable Abbreviation	Variable Description
<i>STATE</i>	Stawte
<i>YEAR</i>	Year
<i>STATENUM</i>	State ID number
<i>DEATHS</i>	Number of traffic-related deaths
<i>INJURED</i>	Number of traffic-related injuries
<i>PRIMLAW</i>	Primary seat belt law
<i>SECLAW</i>	Secondary seat belt law
<i>TOTVMT</i>	Total VMT
<i>PI92</i>	Per-capita income 1992 (\$)
<i>POP</i>	Total population
<i>HOSPAREA</i>	Hospitals per square mile
<i>ETHANOL</i>	Alcohol consumption total ethanol by volume
<i>ETHPC</i>	Per-capita ethanol consumption
<i>SEATBELT</i>	Percent seat belt use
<i>PERC1524</i>	Percent population 15–24
<i>PERC2544</i>	Percent population 25–44
<i>PERC4564</i>	Percent population 45–64
<i>PERC65P</i>	Percent population 65 plus
<i>PERC75P</i>	Percent population 75 plus
<i>INMILES</i>	Total lane miles (excluding local roads)
<i>PRECIP</i>	Annual precipitation in inches
<i>EDUC</i>	Percent bachelors degrees
<i>PERCRINT</i>	Percentage of vehicle-miles driven in rural interstate highway miles

TABLE 6.2
Alternative Modeling Specifications for Fatalities per VMT (*t*-Statistics in Parentheses)

Independent Variable	OLS: No Correction	Parameter Estimates					
		Model 1 Fixed Effects	Model 2 Fixed Effects	Model 3 Two-Way Fixed Effects	Model 4 Random Effects	Model 5 Fixed Effects w/AR(1)	Model 6 Random Effects w/AR(1)
Constant	1.69 (9.22)	—	—	0.83 (3.83)	—	—	0.28 (1.2)
YEAR	-7.90E-04 (-8.62)	-6.71E-04 (-3.75)	—	-4.10E-04 (-3.87)	-4.50E-04 (-2.21)	-1.30E-04 (-1.15)	
P192	-6.50E-07 (-7.12)	3.30E-07 (1.61)	5.5E-08 (0.23)	-1.90E-07 (-1.33)	1.30E-07 (0.62)	-3.70E-07 (-2.57)	
POP	1E-10 (2.4)	1.1E-09 (2.21)	9.3E-10 (1.91)	1.30E-10 (1.35)	1.50E-09 (2.81)	1.20E-10 (1.26)	
HOSPAREA	-0.18 (-1.88)	-0.35 (-0.61)	-0.55 (-0.95)	-0.64 (-3.22)	-0.19 (-0.35)	-0.54 (-2.79)	
ETHPC	1.62 (3.65)	2.6 (1.72)	1.81 (1.19)	1.43 (1.73)	-1.97 (-1.11)	-3.70E-03 (-0.004)	
SEATBELT	-2.20E-03 (-1.67)	-4.00E-03 (-3.3)	2.37E-03 (-1.87)	-4.10E-03 (-3.590)	-2.50E-03 (-1.95)	-2.30E-03 (-1.91)	
PERC1524	-0.148 (-5.63)	0.084 (3.22)	0.074 (2.85)	0.036 (1.713)	0.13 (4.41)	0.032 (1.33)	
PERC2544	-0.184 (-10.61)	0.097 (2.77)	0.081 (2.26)	1.22E-02 (0.51)	0.16 (4.3)	0.031 (1.24)	
PERC4564	0.081 (5.48)	0.063 (1.66)	0.022 (0.56)	0.037 (1.6)	0.052 (1.25)	0.011 (0.51)	
PERC75P	-0.298 (-11.29)	0.226 (2.29)	0.15 (1.48)	-6.20E-03 (-0.15)	0.31 (2.980)	3.80E-03 (0.090)	
INMILES	-2.4E-09 (-1.11)	-3.6E-08 (-1.47)	-3.6E-08 (-1.49)	-2.8E-09 (-0.55)	-3.3E-08 (-1.35)	-4.4E-09 (-0.88)	
PRECIP	-3.10E-05 (-0.8)	3.30E-04 (2.210)	2.30E-04 (1.61)	2.10E-04 (2.77)	2.40E-04 (1.48)	1.80E-04 (2.46)	
<i>Model Statistics</i>							
N	400	400	400	400	350	350	
R ²	0.650	0.916	0.923	0.650	0.926	0.651	

TABLE 6.3

Fixed-Effects versus Random-Effects Model Tests

Test Statistics					
Hausman Test (2 vs. 4)			Hausman Test (5 vs. 6)		
H	d.f.	p-Value*	H	d.f.	p-Value*
83.76	12	.000	89.04	12	.000

**p* values <.1 favor fixed-effects models (at the 90% level).

6.4 Variable-Parameter Models

The discussion in the two previous sections concentrated on models for which the effects of omitted variables are considered as either one-way effects (individual specific, time specific) or two-way effects (individual and time specific). However, there may be cases in which the cross-sectional units examined possess different unobserved socioeconomic and demographic background factors, resulting in response variables that vary over time and across different cross-sectional units (Hsiao 1986). For example, in Chapter 4, data from small, medium, and large transit systems from Indiana were used to investigate the effects of operating subsidies on the performance of the transit systems (Example 4.1). Karlaftis and McCarthy (1998) originally analyzed these data using an *F* test (described in the previous section) and rejected the null hypothesis of common intercept for the transit systems. However, they also rejected the null hypothesis of common slope parameters for systems of different size groupings (small, medium, and large). This finding implies that operating subsidies have differential effects (different signs and/or magnitudes) depending on the size of the transit systems being examined.

In general, when the underlying hypothesis of the relationship between the variables examined is considered theoretically sound but the data do not support the hypothesis of the parameters being equal, it is possible to allow for variations in parameters across cross-sectional units and/or time to account for interunit and/or interperiod heterogeneity. The first case of a model that accounts for parameters that vary over cross-sectional units but are invariant over time is written as

$$\begin{aligned} Y_{it} &= \sum_{k=1}^p \beta_{ki} X_{kit} + u_{it} \\ &= (\beta_k + \alpha_{ki}) X_{kit} + u_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \end{aligned} \tag{6.8}$$

where β is the common-mean parameter vector and α_{ki} are the individual deviations from the common mean (β). If the characteristics of individual samples are of interest, then the α_{ki} are treated as fixed constants (fixed-parameter model), whereas if the population characteristics are of interest and the cross-sectional units available in the sample have been randomly drawn from the population, then the α_{ki} are treated as random variables having zero means and constant variances and covariances (random-parameter model; for more details, see Swamy 1971).

Similar to the model of Equation 6.8, if the assumption is made that model parameters are needed to account for individual cross-sectional unit heterogeneity and for specific time periods, then the model is rewritten as

$$Y_{it} = \sum_{k=1}^P (\beta_k + \alpha_{ki} + \lambda_{kt}) X_{kit} + u_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (6.9)$$

If the α_{ki} , λ_{kt} , and β_k are treated as fixed constants, then the model is a fixed-parameter model, whereas if the α_{ki} and λ_{kt} are considered random and the β_k fixed, the model corresponds to the mixed analysis of variance model (Harley and Rao 1967; Hsiao 1975). Despite the potential flexibility of the variable-parameter models, they have not been used as extensively in empirical work as the one- and two-way error-component models because of the computational complexities involved in their estimation. Rao (1973) developed a Lagrange-multiplier test for the null hypothesis that regression parameters do not vary across cross-sectional units and/or time to test for the need to develop variable-parameter models. Also, Mundlak (1978) proposed a test for the appropriateness of the random-parameter model (similar to the Hausman test described in the previous section).

6.5 Additional Topics and Extensions

The discussions in this chapter have concentrated on the development of single-equation regression models on a balanced panel data set. Although this development is well established in the literature and most of the models described are readily available in many commercially available statistical software packages, research in the development of models using panel data is evolving in many different directions.

First is the issue of unbalanced panel data models. The discussion to this point has concentrated on “complete” or “balanced” panels, referring to cases where cross-sectional units (individuals) are observed over the entire

sample period. Experience shows that incomplete panels are more common in empirical research (Archilla and Madanat 2000). For example, in collecting transit data over time, it is possible to find that some transit systems have discontinued operating, new systems have been formed, and others have been combined. These are typical scenarios that could lead to an unbalanced or incomplete data set. Both one-way and two-way error-component models for unbalanced panel data are estimable (see Moulton 1986; Wansbeek and Kapteyn 1989).

Second is the issue of error-component models in systems of equations (simultaneous equations; see Chapter 5 for more details). In the simultaneous equations framework, error-component models are viewed as pertaining either to SURE systems (seemingly unrelated regression equations) or to pure simultaneous equations systems. In the former, the application of fixed- and random-effect models is straightforward, and these are easily implemented in commercially available software (see Avery 1977, for model development and Karlaftis et al. 1999, for a transportation application). The same ease of practical application does not apply for simultaneous equations. This specialized topic is described in Baltagi (1995).

Third is the issue of error components in limited and discrete dependent variable models. The available literature in this area is vast. For limited dependent variable models such as the fixed- and random-effects Tobit model, readers are referred to Maddala (1983), Heckman and McCurdy (1980), and Honore (1992). For discrete dependent variable models, readers should refer to Hsiao (1986) for a rigorous treatment of the subject and Greene (2000) for estimation aspects. Detailed estimation aspects of discrete outcome models with panel data are found in Keane (1994), McFadden (1989), and Madanat et al. (1997). Bhat (2000) and Hensher (2001) offer some interesting applications in the transportation area.

7

Background and Exploration in Time Series

Time series models have been the focus of considerable research and development in recent years in many disciplines, including transportation. This interest stems from the insights that are gained when observing and analyzing the behavior of a variable over time; a time series is a sequence of observations arranged by their time of outcome. For example, modeling and forecasting a number of macroscopic traffic variables such as traffic flows, speeds, and occupancies are an indispensable part of most congestion management systems. Forecasting passenger volumes at airport terminals, furthermore, is an essential input to engineering planning and design processes. Transportation variables observed over time frequently result in a time series modeling application, where focus is on predicting the values of a variable based on a series of past values at regular time intervals.

Analyses of time series data concentrate on gaining an improved understanding of the data generating mechanism, modeling the behavior of data over time, and forecasting future outcomes. A fundamental property that sets time series methods apart from other approaches is that time series data are not independently generated and hence procedures that assume independently and identically distributed data are inappropriate. To this end, appropriate methods have been developed for handling such data and remain the focus of this and the following chapter.

When analyzing time series data, time-domain or frequency-domain approaches are often used. The time-domain approach assumes that adjacent points in time are correlated and that future values are related to past and present ones. The frequency-domain approach assumes that time series characteristics relate to periodic or sinusoidal variations that are reflected in the data. The focus of Chapters 7 and 8 is on time-domain approaches due to their predominance in the transportation literature; until recently, relatively few applications in the transportation literature used frequency-domain approaches; but, the use of Fourier transforms and wavelet analyses are becoming increasingly popular (classical transportation related applications of the—Fourier based—frequency-domain approach are Stathopoulos and Karlaftis (2001b), Peeta and Anastassopoulos (2002). Qiao (2005) offers an excellent introduction to wavelet analyses in transportation, Teng and Qi (2003) an interesting application of wavelets in incident detection and Jiang and Adeli (2004, 2005) an application to traffic flow forecasting).

7.1 Exploring a Time Series

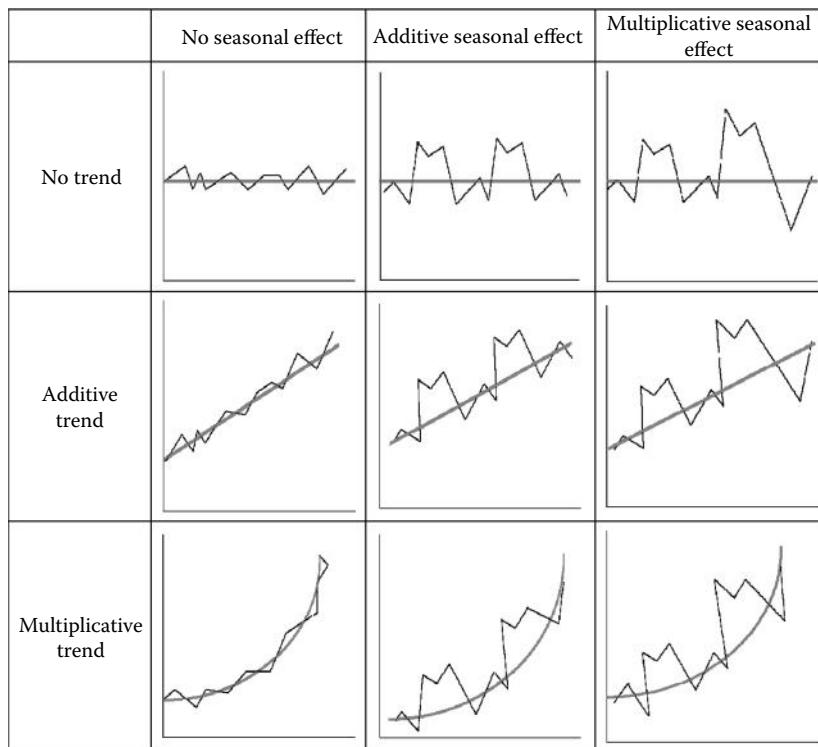
Time series analyses are based upon the assumption that a given time series is the sum of three components, or characteristics, that are frequently encountered in such data. The mathematical representation of the decomposition approach is written as $Y_t = f(T_t, S_t, E_t)$, where Y_t is the (actual) time series value at time t , T_t is the trend component at time t , S_t is the seasonal component at time t , and E_t is the irregular (or random) component at time t (an example of time series decomposition in transportation is the work of Young and Ord 2004). The decomposition of a time series into its characteristic components is based upon the assumption that these “movements” interact in either an additive ($T_t + S_t + E_t$) or a multiplicative ($T_t \times S_t \times E_t$) manner. The literature suggests that an additive model is more appropriate when the magnitude of the seasonal component (S_t) does not vary with the level of the series (T_t), while a multiplicative model is more prevalent when time series have seasonal variation that increases with the level of the series; further, a multiplicative relationship is fit as an additive relationship to the logarithm of the data, for example, $\ln(Y_t) = \ln(T_t) + \ln(S_t) + \ln(E_t)$. A graphical representation of time series components along with their combined effect appear in Figure 7.1.

7.1.1 Trend Component

Trend in a time series is, usually, a gradual change in some property of the series and is frequently referred to as a long-term (or secular) movement that corresponds to the general direction in which a time series graph moves over time. The movement is indicated by a trend curve, frequently represented by a simple trend line. Some of the appropriate methods for estimating trends in a time series are the usual least squares approach, the moving average method, and the semiaverages method (Spiegel and Stephens 1998). Although the notion of a trend is rather simple and straightforward, there are many methods in the time series literature that assume the lack of a trend (stationarity assumption; see Section 7.2.2 for more details), or where a trend actually distorts the statistical relationship of interest, and hence its removal is an important process in time series analyses (see Section 7.1.2).

7.1.2 Seasonal Component

When time-series tend to follow identical or nearly identical patterns within equal time periods, they are recognized as seasonal movements or variations. Seasonal variations range from yearly periods in economic time series to daily periods in traffic data (when, e.g., seasons or periods are considered the peak and off-peak travel times). For example, in Figures 7.2 and 7.3 it is fairly straightforward to identify an annual pattern in monthly international airline passenger totals in the United States; the annual pattern suggests

**FIGURE 7.1**

Time series components and their combined appearance.

that there is an annual repetitive cycle where the yearly low is in November and the yearly high is in July–August. To quantify variations within these periods, say monthly variation within a year, a seasonal index is used. A seasonal index is used to obtain a set of measurements that reflects seasonal variation (e.g., monthly) in the values of a variable (e.g., monthly traffic crashes) for each of the subdivisions of a period. The average seasonal index for the entire period is equal to 100%, and fluctuates above and below this value across subdivisions of the period. A number of methods are used for calculating a seasonal index.

The first is the average percentage method. In this method, data (e.g., monthly traffic crashes) for each subdivision are expressed as a percentage of the average across subdivisions in the period (e.g., yearly traffic crashes divided by 12). When the corresponding subdivisions for all periods are averaged and extreme values are disregarded (outliers are dealt with appropriately), then the resulting values give the seasonal index whose mean is 100%.

The second method for calculating seasonal trend is the percentage trend or ratio-to-trend method. In this approach, data are subdivided into appropriate periods and each subdivision's data are expressed as a percentage of

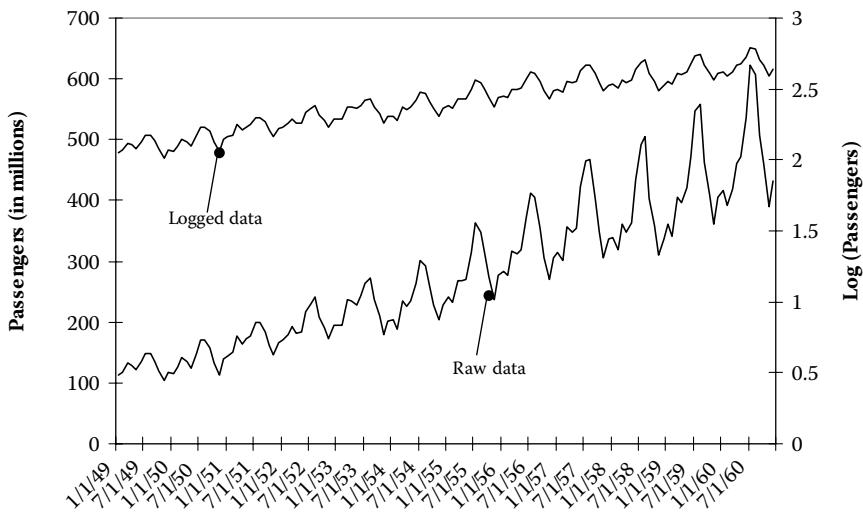


FIGURE 7.2
Monthly airline passenger data.

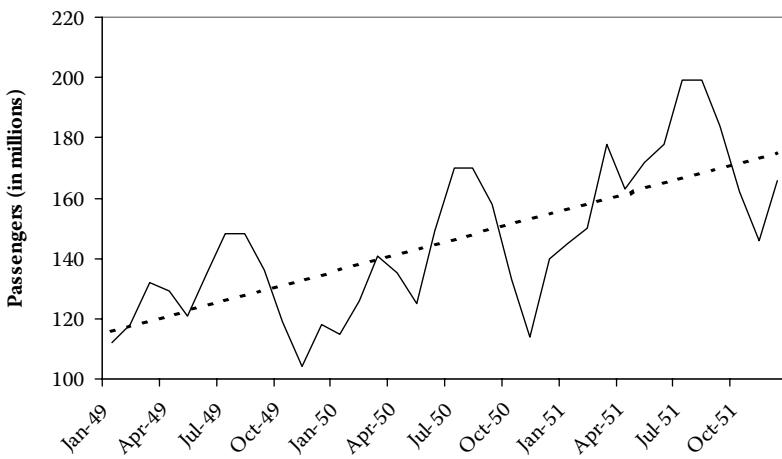


FIGURE 7.3
First 3 years of airline passenger data; actual data, solid line; fitted (linear trend), dashed line.

the series' trend value. The values for all periods are averaged yielding the seasonal index whose mean should be 100%.

The third method is the percentage moving average or ratio-to-moving average method where, for appropriate periods of the data, a period moving average is computed. From the period moving average a two subdivision moving average is computed (called a centered moving average). The required

index is produced by expressing the original data as percentages of the centered moving average. Finally, if the original subdivision data are divided by their corresponding seasonal index number, they are deseasonalized or adjusted for seasonal variation.

7.1.3 Irregular (Random) Component

Irregular or random movements do not follow any ordinary pattern or trend and, as such, are characterized by intense variations of a usually short duration. Estimation of irregularities is done after eliminating trend and seasonal components. Usually, irregularities are of minimum importance to practical problems but they should be handled carefully as they may represent important information regarding the underlying process of the data examined (random movements may also include outliers).

7.1.4 Filtering of Time Series

As previously mentioned, it is at times important to remove the trend components in a time series, either because of the stationarity assumption or the possibility that a trend distorts the “true” underlying relationship between the variables examined. Filtering is a process for assisting in trend and, frequently, intense variation removal.

7.1.5 Curve Fitting

When a time series exhibits a gradual change over time, the trend component is often represented as a function of time. Obviously, the simplest form of this detrending function is the usual least squares line, which indicates a linear trend. Linear regression is used to fit the model

$$Y_t = a + \beta T + \varepsilon_t \quad (7.1)$$

where Y_t is the original series at time t , a is the regression intercept, β is the regression parameter for time T , and ε_t are the residuals; the trend is given as (the fitted) $a + \beta T$. Despite the simplicity and appeal of the linear trend, other forms for the trend component such as the quadratic function, the logistic and Gompertz functions may be more appropriate (see Falk et al. 2006, for a more extensive discussion).

7.1.6 Linear Filters and Simple Moving Averages

Time series data often contain randomness that leads to widely varying predictions. To eliminate the unwanted effect of this randomness, linear

filters are widely used. Given a set of observations $V_1, V_2, V_3, \dots, V_t$, the linear transformation (Falk et al. 2006)

$$V_t^* := \sum_{i=-j}^k a_i V_{t-i}, \quad t = k+1, \dots, n-j \quad (7.2)$$

is a linear filter with weights a_{-j}, \dots, a_k , and where V_t^* is the “output” and V_t is the input of the series. A filter whose weights sum to one, that is, $\sum_{i=-j}^k a_i = 1$ is the well known moving average method.

In general, the simple moving averages method takes a set of recent observed values, finds their average, and uses it to predict a value for an upcoming time period. The term moving average implies that when a new observation becomes available, it is available for forecasting immediately (Makridakis et al. 1989). Given the set of observations $V_1, V_2, V_3, \dots, V_t$, a definition of the moving average of order N is the sequence of arithmetic means given by

$$\frac{V_1 + V_2 + \dots + V_t}{N}, \quad \frac{V_2 + V_3 + \dots + V_{t+1}}{N}, \quad \frac{V_3 + V_4 + \dots + V_{t+2}}{N} \quad (7.3)$$

The higher the number of available observations the “smoother” the series will become because increasing the number of observations tends to minimize the effects of irregular movements that are present in the data. A generalized mathematical representation of the moving averages method is

$$\hat{V}_{t+1} = S_t = \frac{V_t + V_{t-1} + \dots + V_{t-N+1}}{N} = N^{-1} \cdot \sum_{i=t-N+1}^t V_i \quad (7.4)$$

where \hat{V}_{t+1} is the predicted value for the series at time t , S_t is the smoothed value at time t , V_i is the value for the series at time i , i is the time period corresponding to $i = t - N + 1, \dots, t$, and N is the number of values (order) included in the moving average estimation. As inferred from Equation 7.4, the method assigns equal weights and thus equal importance to the last N observations of the time series. The method of moving averages has two major limitations. The first is the requirement for a large number of past observations to provide adequate smoothing. The second is that it assigns equal weights to each observation; this second point contradicts the usual empirical evidence that more recent observations should be closer in value to future ones so they should have an increased importance in the process. To address these two limitations, exponential smoothing methods were developed.

7.1.7 Exponential Smoothing Filters

Exponential smoothing resembles the moving averages method because it is used to smooth past data to eliminate irregularities. Suppose Equation 7.4

is used to compute the moving average value of a series with only its most recently observed value (period t) available. The observed value is approximated in period $t-N$ with a reasonable estimate such as the predicted value from period t (\hat{V}_t). This result implies that

$$\begin{aligned}\hat{V}_{t+1} &= \frac{V_t}{N} - \frac{\hat{V}_t}{N} + \hat{V}_t \\ &= \frac{1}{N} \cdot V_t + \left(1 - \frac{1}{N}\right) \cdot \hat{V}_t\end{aligned}\quad (7.5)$$

Equation 7.5 suggests that the most recently observed value in period t is weighted by $(1/N)$ and the forecast for the same period is weighted by $(1/(1-N))$. Letting $k = 1/N$, Equation 7.5 becomes

$$\hat{V}_{t+1} = k \cdot V_t + (1-k) \cdot \hat{V}_t \quad (7.6)$$

which is the general form of the equation for exponential smoothing. Equation 7.5 accounts for the first limitation of the moving average method because there is no longer a need to store a large number of past data since only the most recent observation is used. The value of \hat{V}_t is expanded such that $\hat{V}_t = kV_{t-1} + (1-k)\hat{V}_{t-1}$. From Equation 7.6 and its expansion the following is obtained:

$$\begin{aligned}\hat{V}_{t+1} &= kV_t + (1-k)(kV_{t-1} + (1-k)\hat{V}_{t-1}) \\ &= kV_t + k(1-k)V_{t-1} + (1-k)^2\hat{V}_{t-1}\end{aligned}\quad (7.7)$$

By recursively substituting \hat{V}_i for each $i = t, t-1, \dots$, the final equation becomes

$$\hat{V}_{t+1} = kV_t + k(1-k)V_{t-1} + k(1-k)^2V_{t-2} + k(1-k)^3V_{t-3} + \dots \quad (7.8)$$

from which it is apparent that the more recently observed values are assigned higher weights. This operation relaxes the limitation of the unweighted average used in the moving averages method by using exponential smoothing of previous points to forecast future values. Exponential smoothing requires only the most recently observed value of the time series and a value for k . Furthermore, with the use of software and experience, this method is effective for smoothing patterns of data. In cases of strong seasonal effects, other methods of smoothing could also be considered. For example, linear exponential smoothing is used when there is a trend in the observed data. In these cases, implementing single exponential smoothing would provide

a random disturbances term for each observation. Using this smoothing technique, a smoothed estimate of the trend is

$$T_t = d(S_t - S_{t-1}) + (1-d)T_{t-1} \quad (7.9)$$

where S_t is the equivalent of the single exponential smoothed value for period t , T_t is the smoothed trend for period t , and d is the smoothing weight. Equation 7.9 operates much like Equation 7.8 in that the most recent trend has a larger weight than others. Combining these two equations gives

$$\hat{V}_t = S_t = kV_t + (1-k)(\hat{V}_{t-1} + T_{t-1}) \quad (7.10)$$

where the additional term T_{t-1} is used to smooth the effects of trend.

Finally, Winters' linear and seasonal exponential smoothing method is similar to linear exponential smoothing but attempts to also account for seasonal movements of past data in addition to trends. The method uses three equations to smooth for randomness, trends, and seasonality, yielding

$$\begin{aligned} \hat{S}_t &= S_t = k \frac{V_t}{SM_{t-L}} + (1-k)(S_{t-1} + T_{t-1}) \\ T_t &= d(S_t - S_{t-1}) + (1-d)T_{t-1} \\ SM_t &= y \frac{V_t}{S_t} + (1-y)SM_{t-L} \end{aligned} \quad (7.11)$$

where S is the smoothed value of the deseasonalized time series, T is the smoothed value for the trend, SM is the smoothed value for the seasonal factor, and L is the length of seasonality (length of seasonality indicates the periods between repeating high or low values; e.g., in monthly data such as the one in Figure 7.2, the length of seasonality would be equal to 12). The third Equation in 7.11 provides an index for seasonality effects; this index is the ratio X_i/S_i , the current value over the smoothed value (see Kitagawa and Gerch 1984). However, irregularity effects are not completely eliminated by accounting for seasonality, and the ratio and past seasonality value SM_{t-L} are weighted accordingly. The second Equation in 7.11 has the same role as in simple linear exponential smoothing. Finally, in the first equation, the first term is divided by SM_{t-L} to eliminate the seasonality effects from X_t . For computational purposes, the value SM_{t-L} is used because without it S_t and SM_t cannot be calculated.

Exponential smoothing, with or without seasonality, has been frequently applied due to the successful application of the progressive decay of the weights of past values. This property is particularly applicable when modeling traffic conditions approaching congestion, since traffic variables typically exhibit extreme peaks, unstable behavior, and rapid fluctuations. In

such cases, the best-fit model must respond quickly and needs to be less influenced by past values, making exponential smoothing a useful traffic engineering smoothing tool in practice (Williams et al. 1998).

Example 7.1

Figure 7.2 presents a classic data set from the time series literature (Box and Jenkins 1976). The data are monthly international airline passenger totals from January, 1949 to December, 1960. Inspection of the plot reveals a trend, a yearly pattern, and an increasing variance. The graph also presents the graph of the log transform used to address the increasing variance (as noted in Chapter 4, a log transformation is a variance stabilizing transformation that can, at times, be effective in cases of increasing variance).

One may be interested in developing both a forecast model for next month's passenger total based on these data and for smoothing the original series. Box and Jenkins developed a model for these data that has become known as the airline model because of its relationship to the airline passenger data. A further illustration of the importance of correcting for increasing variance is seen when comparing the goodness-of-fit of the airline model to the raw data with the goodness-of-fit of the airline model to the logged data (Table 7.1).

In a further investigation, an initial value calculation for trend and seasonal effects in the first 3 years of the airline data is estimated (Figure 7.3 and Table 7.2). Because the seasonal effect appears to increase with level, multiplicative seasonal factors are appropriate. In Figure 7.3, a fitted straight line is shown through the first 3 years of data. The intercept of this line is 114 passengers. This value is used for the initial level. The slope of the line is 1.7 passengers, and is used for the initial trend. Table 7.2 presents the calculation of the initial seasonal factors. The average of the first 3 years of data is 145.5. A ratio of observed passenger total to this average is computed for each of the first 3 years of observations. Finally, the three instances of each month are averaged. It so happens that the calculated initial seasonal factors average to exactly 1.0. In cases when multiplicative seasonal factors do not average to 1.0, they should all be divided by the average to obtain an average of 1.0. Additive seasonal factors are calculated in a similar fashion using the difference between observed and average values instead of the ratio. Additive seasonal factors should sum to zero, which is achieved by subtracting an equal allocation of the sum from each of the seasonal factors.

TABLE 7.1

Characteristic Measures for Raw and Logged Airline Data

Measure	Raw Data	Natural Log Transform
Mean square error	135.49	114.85
Root mean square error	11.64	10.71
Mean absolute percentage error	3.18%	2.93%
Mean absolute error	8.97%	8.18%

TABLE 7.2

First 3 Years of Airline Passenger data

Month	Airline Passengers	Average First 3 Years	Actual to Average Ratio	Seasonal Factors
Jan-49	112	145.5	0.81	0.85
Feb-49	118	145.5	0.81	0.90
Mar-49	132	145.5	0.91	1.03
Apr-49	129	145.5	0.89	0.98
May-49	121	145.5	0.83	0.96
Jun-49	135	145.5	0.93	1.06
Jul-49	148	145.5	1.02	1.18
Aug-49	148	145.5	1.02	1.18
Sep-49	136	145.5	0.93	1.10
Oct-49	119	145.5	0.82	0.95
Nov-49	104	145.5	0.71	0.83
Dec-49	118	145.5	0.81	0.97
Jan-50	115	145.5	0.79	
Feb-50	126	145.5	0.87	
Mar-50	141	145.5	0.97	
Apr-50	135	145.5	0.93	
May-50	125	145.5	0.86	
Jun-50	149	145.5	1.02	
Jul-50	170	145.5	1.17	
Aug-50	170	145.5	1.17	
Sep-50	158	145.5	1.09	
Oct-50	133	145.5	0.91	
Nov-50	114	145.5	0.78	
Dec-50	140	145.5	0.96	
Jan-51	145	145.5	1.00	
Feb-51	150	145.5	1.03	
Mar-51	178	145.5	1.22	
Apr-51	163	145.5	1.12	
May-51	172	145.5	1.18	
Jun-51	178	145.5	1.22	
Jul-51	199	145.5	1.37	
Aug-51	199	145.5	1.37	
Sep-51	184	145.5	1.26	
Oct-51	162	145.5	1.11	
Nov-51	146	145.5	1.00	
Dec-51	166	145.5	1.14	

7.1.8 Difference Filter

An important and widely used type of filtering is differencing. The technique of differencing is simple; the first observation is subtracted from the second, the second from the third, and so on. The final outcome has no trend. Formally, the mathematical operator for this process is the differencing operator D . Assume that B is a backshift operator such that

$$B \cdot X_t = X_{t-1} \quad (7.12)$$

For operator B two properties hold

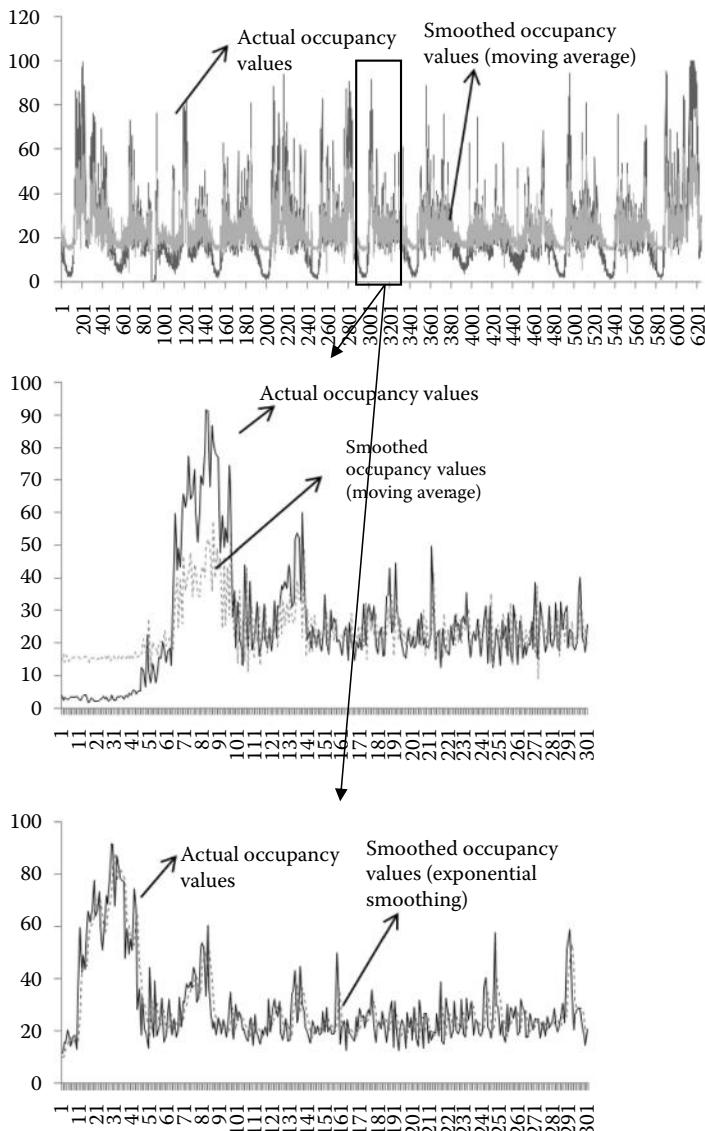
$$\begin{aligned} B^n \cdot X_t &= X_{t-n} \\ B^n \cdot B^m \cdot X_t &= B^{n+m} \cdot X_t = X_{t-n-m} \end{aligned} \quad (7.13)$$

Here, $B \cdot B^{-1} = 1$ and, if c is constant, $B \cdot c = c$. Since $(1-B) \cdot X_t = X_t - X_{t-1}$ for $t \geq 2$, it is straightforward to show that the differencing operator D is $D = 1 - B$. This expression is used for first order differencing and analogous expressions are derived for higher order differencing. First order differencing D is used for removing a linear trend and second order differencing D^2 is used for removing a quadratic trend (Shumway and Stoffer 2000). Further, there may be cases of seasonal variation in time series. In these cases, it may be necessary to obtain a seasonal difference where the differencing operator is denoted as D_S^N and where N is the order of differencing among periods, and S is the order of differencing within a period. For example, for monthly data and a yearly period the differencing operator is $D_{12} \cdot X_t = (1 - B^{12}) \cdot X_t = X_t - X_{t-12}$. As Anderson (1976) discussed, the variance of a time series generally decreases with each additional differencing up to some point where it actually starts to increase; when the variance increases, the series has been overdifferenced.

Example 7.2

Toward meeting the goal of reducing congestion, decreasing travel times and giving drivers the ability to make better route and departure time choices, a dynamic traffic map and advanced traveler information system for the central Athens area has been operating since 1996 (NTUA 1996). As a first attempt to meet driver needs, travel time information is given over the internet updated every 15 minutes. Speeds and travel times along 17 major urban signalized arterials are estimated using volume and occupancy data collected directly from the controllers installed at 144 locations in the downtown urban network (for more information see Stathopoulos and Karlaftis 2001a, 2002, 2003).

The dataset available is sizeable. In an effort to reduce the size, a new dataset was created containing 3 minute volume (number of vehicles per 3 minutes) and occupancy measurements (measured in %) for the first 13 days of May of 2000. This approach yielded a data set containing approximately 6,250 data points to be used in the analysis. Figure 7.4 depicts a time series graph for occupancy along with smoothed values obtained through a 3-period moving average process and

**FIGURE 7.4**

Time series graphs for traffic occupancy (actual and smoothed data).

exponential smoothing. Figure 7.5 is a time series plot for occupancy but after applying the difference filter (first differencing); the differenced series are centered (as expected) around zero with significantly smoothed variation.

Example 7.3

Using the same data as in Example 4.2, but focusing on an analysis of monthly travel card sales, Figure 7.6 depicts the original series along with a 2 period moving average and exponential smoothing values. As was the case for ticket sales, there is a pronounced decrease in sales for the month of August that needs to be accounted for in further analyses.

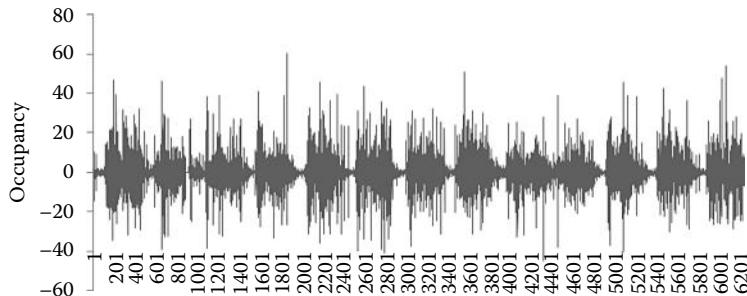


FIGURE 7.5

Time series graphs for traffic occupancy (after first differencing).

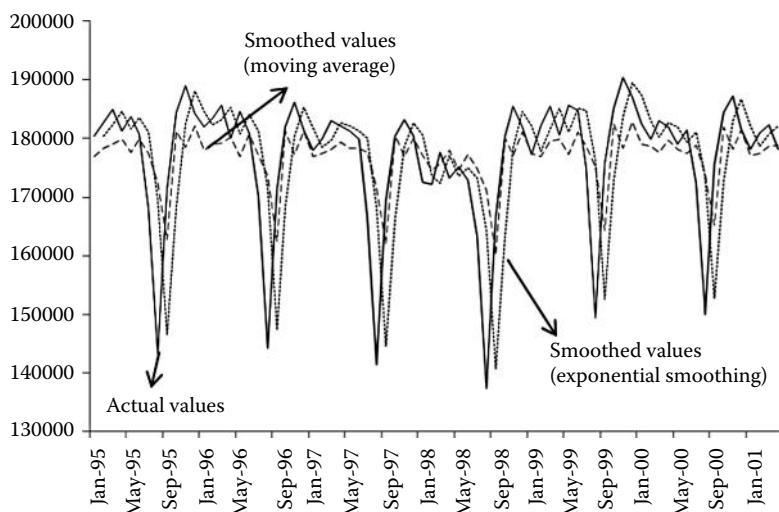


FIGURE 7.6

Time series graphs for travel card sales (actual and smoothed data).

7.2 Basic Concepts: Stationarity and Dependence

The analysis of time series has been largely characterized and is, in turn, based on two fundamental concepts: stationarity and dependence; stationarity refers to the trend and variability in a time series, while dependence refers to the relationship correlation between successive observations in a time series.

7.2.1 Stationarity

Possibly the most important concept in time series analysis is stationarity. Time series models apply to horizontal, or in statistical terms, stationary data only. Thus, before proceeding with modeling a time series, the data must be horizontal (without a trend). More formally, assume the existence of a time series X_t , where t is the observation period. The time series is strictly stationary if the joint probability distributions of $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ and $(X_{t_1+L}, X_{t_2+L}, \dots, X_{t_n+L})$ are the same for all t_1, t_2, \dots, t_n and L (length of seasonality). This implies that the joint distribution of $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ is time invariant, a strong condition that requires verification in practice (Tsay 2002).

A weaker notion of stationarity, called weak stationarity, applies when both the mean of X_t , and the covariance between X_t and X_{t+L} are time invariant (for an arbitrary L). For $n = 1$, the univariate distribution of X_t is the same to that of X_{t+L} . Accordingly, $E(X_{t+L}) = E(X_t)$ and $VAR(X_t) = VAR(X_{t+L})$ implying that the mean μ and variance σ^2 of the time series are constant over time (Shumway and Stoffer 2000). For $n = 2$, the joint probability distributions of (X_{t_1}, X_{t_2}) and (X_{t_1+L}, X_{t_2+L}) are the same and have equal covariances

$$COV(X_{t_1}, X_{t_2}) = COV(X_{t_1+L}, X_{t_2+L}) \quad (7.14)$$

The previous condition depends only upon the lag L . The covariance between X_t and X_{t+L} is called autocovariance (γ_κ) and is given by

$$\gamma_\kappa = COV(X_t, X_{t+L}) = E[(X_t - \mu) \cdot (X_{t+L} - m)] \quad (7.15)$$

and $\gamma_0 = VAR(X_t) = \sigma^2$.

7.2.2 Dependence

A second important concept in time series is the concept of autocorrelation, which refers to the correlation of a time series with its own past and future values. The basic tool for applying advanced methods of time series modeling

is the autocorrelation coefficient ρ_L . An autocorrelation coefficient, similar to the correlation coefficient, describes the association between observations of a particular variable across time periods. For example, when the linear dependence between X_t and X_{t-L} is of interest, the concept of correlation is generalized to autocorrelation, where the correlation coefficient between X_t and X_{t-L} is called the lag- L autocorrelation of X_t . The coefficient is denoted by ρ_L and is a function of L only (under weak stationarity conditions). The lag- L autocorrelation of X_t is written as follows:

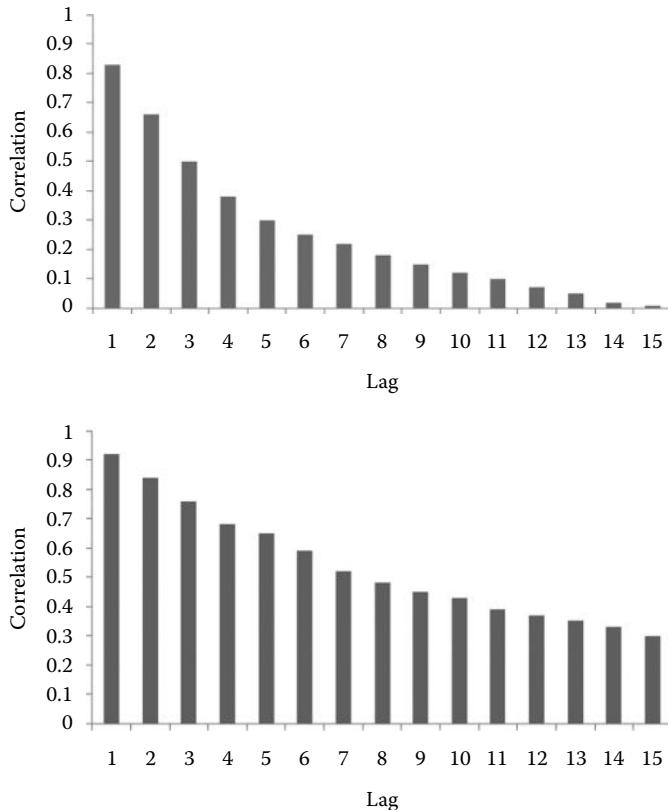
$$\rho_L = \frac{COV(X_t, X_{t-L})}{\sqrt{VAR(X_t)VAR(X_{t-L})}} = \frac{COV(X_t, X_{t-L})}{\sqrt{VAR(X_t)VAR(X_t)}} = \frac{\gamma_L}{\gamma_0} \quad (7.16)$$

When the $VAR(X_t) = VAR(X_{t-L})$ property for a weakly stationary time series is satisfied, it follows that $\rho_0 = 1$, $\rho_L = \rho_{-L}$ and $-1 \leq \rho_L \leq 1$. Finally, a weakly stationary time series is not serially correlated if and only if $\rho_L = 0$ for all $L > 0$ (Tsay 2002).

As discussed, autocorrelation values range from -1 to 1 and provide important insight regarding patterns of the data being examined. If, for example, data are completely random, then autocorrelation values are typically close to 0 . A measure similar to autocorrelation is the partial autocorrelation coefficient (ϕ_{kk}), which has interesting characteristics that are helpful in selecting an appropriate time series model in practice. Partial autocorrelation is defined as the correlation between two data points X_t and X_{t-L} , after removing the effect of the intervening variables $X_{t-1}, \dots, X_{t-L+1}$. Both the autocorrelation and partial autocorrelation coefficients are used extensively in time series modeling. Finally, the function ρ_1, ρ_2, \dots is the autocorrelation function (ACF) of X_t that fully characterizes a linear time series and is used to capture the linear dynamics in data (Figure 7.7 depicts a typical ACF for Stationary (top) and nonstationary (bottom) data).

Example 7.4

Using the data from Examples 7.2 and 7.3, Figures 7.8 through 7.10 depict the ACF and partial autocorrelation function (PACF) graphs for travel card sales, occupancy and volume, respectively. The ACF graph for monthly card sales depicts an interesting pattern since although correlation values for most lags are low, the lag12 correlation is high (approximately 0.70). This effect might be expected, as the data analyzed are monthly and as such, 12th order autocorrelation is expected (this finding along with its potential implications in modeling and policy recommendations are discussed in Example 4.2). In the case of both occupancy and volume, the ACF graphs depict a clearly nonstationary process as the function slowly decays.

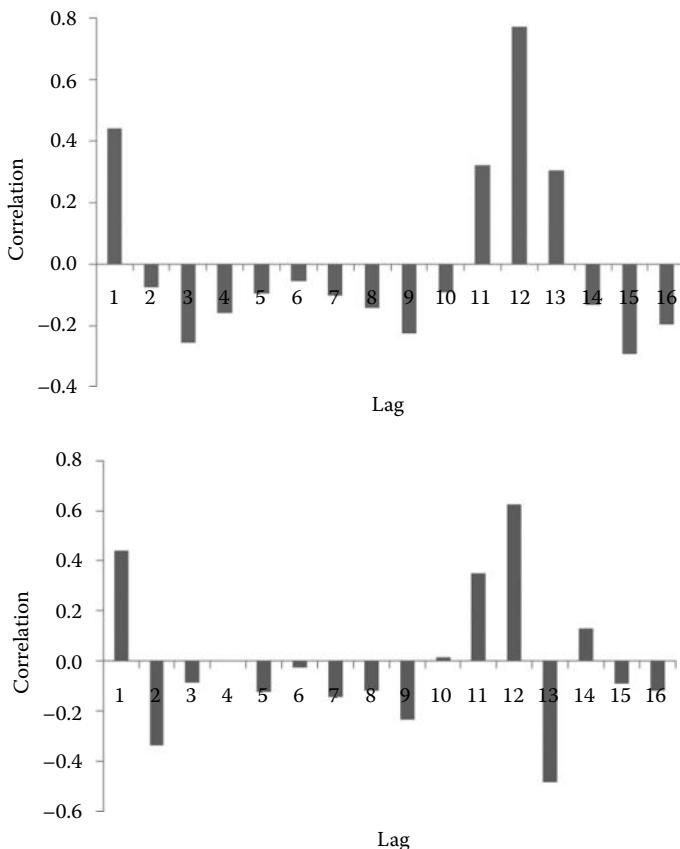
**FIGURE 7.7**

Example ACF graphs for stationary (top) and nonstationary (bottom) data.

7.2.3 Addressing Nonstationarity

As previously mentioned, one of the requirements of time series analyses is that the mean and variance of the series be constant over time (stationarity). Obviously, this assumption is violated when a series exhibits trend and the usual estimation methods do not apply. In general, a stationary time series has the same statistical behavior at each point in time. Models based on nonstationary series usually exhibit large errors and have biased parameter estimates and, as such, obtaining a stationary series is important before analyzing the series.

A first step in investigating stationarity is to use either the ACF for the series (Figure 7.7), or to use a formal statistic to examine whether autocorrelations of X_t are jointly zero. The Ljung–Box (1978) statistic is used for testing the null hypothesis $H_0: \rho_1 = \dots = \rho_m = 0$ against the alternative that $H_1: \rho_i \neq 0$ for some $i \in \{1, \dots, m\}$ as follows (The Ljung–Box statistic is a modification of

**FIGURE 7.8**

ACF (top) and PACF (bottom) graphs for travel card sales.

the well-known Portmanteau test to increase the power of the test in finite samples; Tsay 2002):

$$Q(m) = T(T+2) \sum_{L=1}^m \frac{\rho_L^2}{T-L} \quad (7.17)$$

The $Q(m)$ statistic is asymptotically χ^2 distributed with m degrees of freedom.

7.2.4 Differencing and Unit-Root Testing

Removal of the trend and, frequently, of nonconstant variance is obtained through a variety of filtering approaches (see Section 7.1.2 for more details). However, in transportation research, stationarity is most often obtained through first-order differencing; that is, it is widely believed that

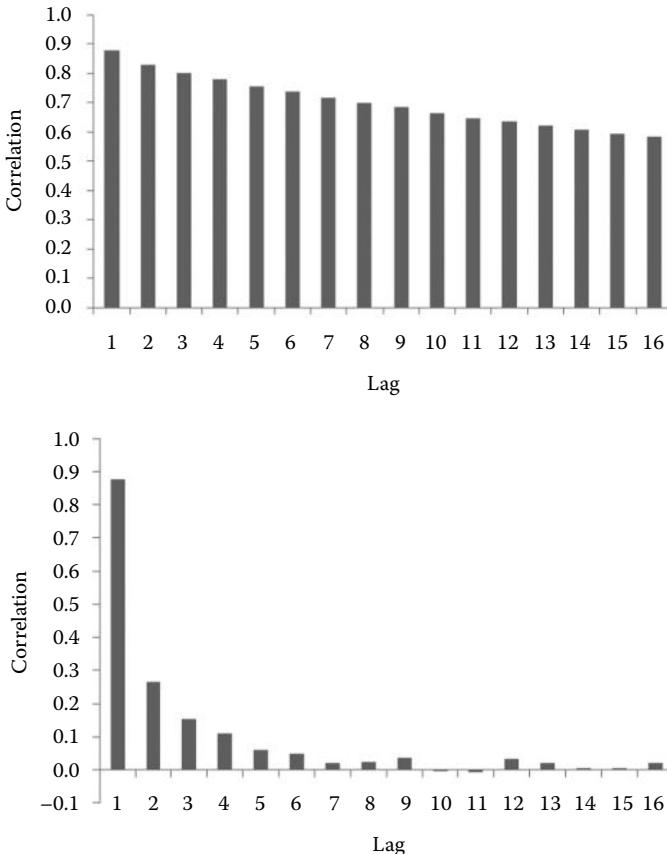


FIGURE 7.9
ACF (top) and PACF (bottom) graphs for occupancy.

a nonstationary series is transformed into a stationary when first-order differences, rather than the original series, are used (first order differencing for X_t is $Z_t = X_t - X_{t-1}$). In this case, the original series X_t are called unit-root nonstationary requiring first-order differencing to become stationary.

Several tests for nonstationarity (unit-root tests) have been devised. Experiments conducted by Granger and Engle (1984) indicate that the most satisfactory of these tests are the Dickey–Fuller tests. In this test, the null hypothesis is that the variable under consideration, say, X_t , is not stationary (i.e., it requires at least one differencing to become stationary; the alternative hypothesis is that the time series is already stationary). To test this hypothesis, the following regression models is estimated:

For a unit root:

$$\Delta X_t = \gamma X_t + \varepsilon_t \quad (7.18)$$

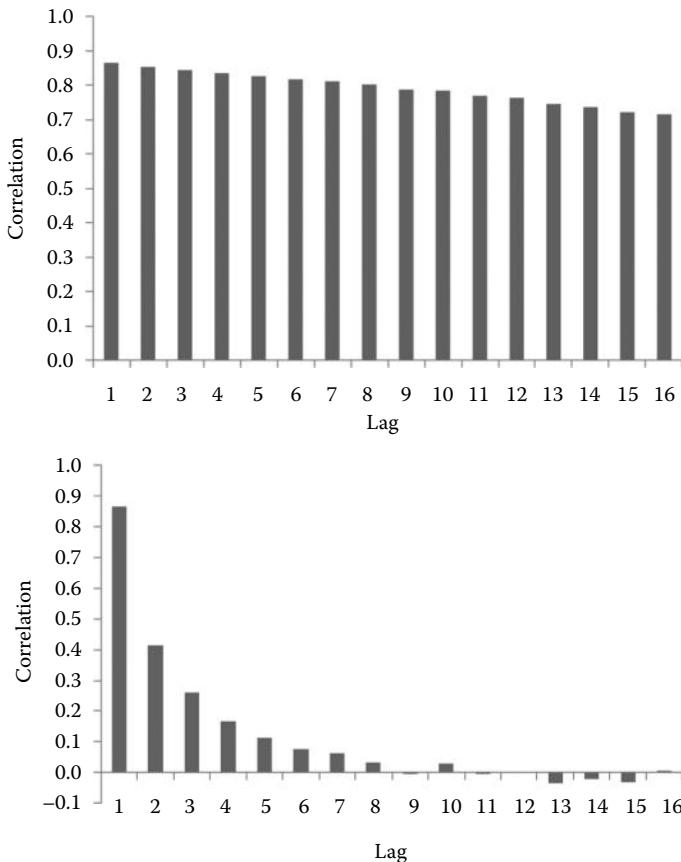


FIGURE 7.10
ACF (top) and PACF (bottom) graphs for volume.

for a unit root with drift:

$$\Delta X_t = a_0 + \gamma X_t + \varepsilon_t \quad (7.19)$$

for a unit root with drift and time trend

$$\Delta X_t = a_0 + \gamma X_t + \delta t + \varepsilon_t \quad (7.20)$$

In each case, testing for a unit root is equivalent to testing whether $\gamma = 0$. Two important points need to be made here. First, the test statistic does not follow the usual t -tests but rather the Dickey–Fuller distribution, which is not explicitly provided but is obtained via Monte-Carlo and bootstrap methods

(Falk et al. 2006). The p -values for the Dickey–Fuller tests are provided by most software packages that offer time series analysis. Second, the tests may not distinguish between true unit-root processes ($\gamma = 0$) and near unit-root processes (γ “close” to zero), and are hence said to have low statistical power.

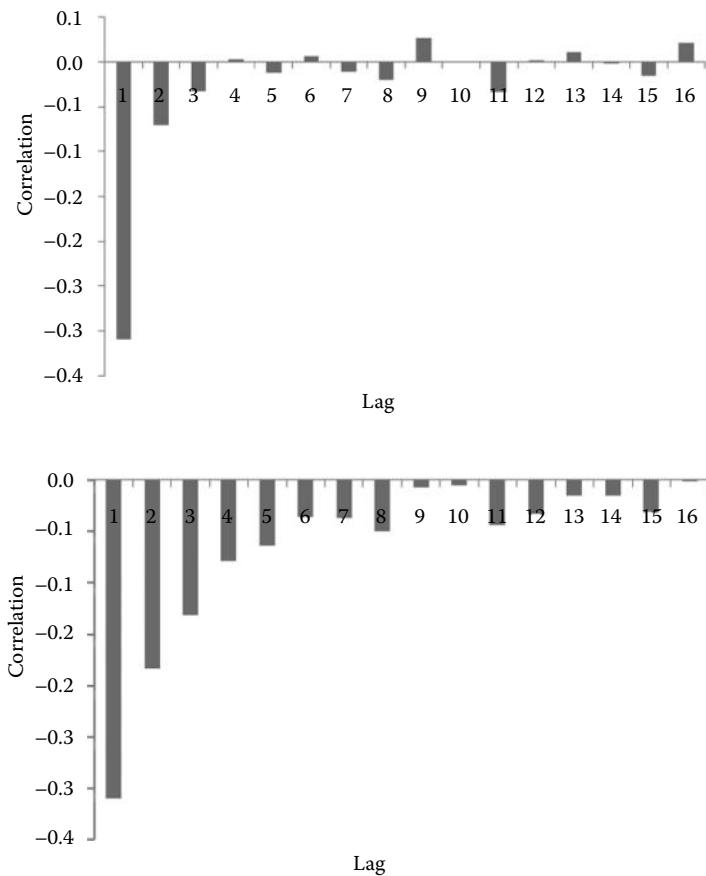
7.2.5 Fractional Integration and Long Memory

This last point, that a process may be near unit root, raises an interesting and widely ignored topic at least in transportation research. Series are most frequently differenced d times to achieve stationarity (the differenced series are said to be integrated of order d ; for $d = 1$ the series are $I(1)$ integrated and contain a unit root). However, as Granger and Joyeux (1980) showed, imposing erroneous differentiation parameters, that is forcing d to be equal to 1 when it should not be so, leads to overdifferentiation of the series and forces an artificial (and erroneous) correlation structure on the estimated models. Further, $I(0)$ and $I(1)$ model structures cannot account for persistence in a time series commonly referred to as long-memory. A series has long memory when significant autocorrelation is present in “widely” separated observations in time.

Both short and long-memory structures are accounted for through fractional integration, where d can take on any value in the $[-1, 1]$ region, including fractional values; obviously, d can still be 0 or 1. In general, data are better modeled through fractional integration since strict $I(0)$ or $I(1)$ processes are avoided and both long-term persistence and short-term correlation in a series are explicitly modeled (Hosking 1981). The differentiation parameter (d) is associated with interesting statistical properties in a time series (Hosking 1981; Odaki 1993); for example, when $d = 1$ the series is a unit-root process; when $d = 0$ the series is stationary; when $0 < d \leq \frac{1}{2}$, the series is fractionally integrated and exhibits long memory, while when $\frac{1}{2} < d < 1$ stationarity in the series cannot be verified (for a transportation application of fractional integration see Karlaftis and Vlahogianni 2009).

Example 7.5

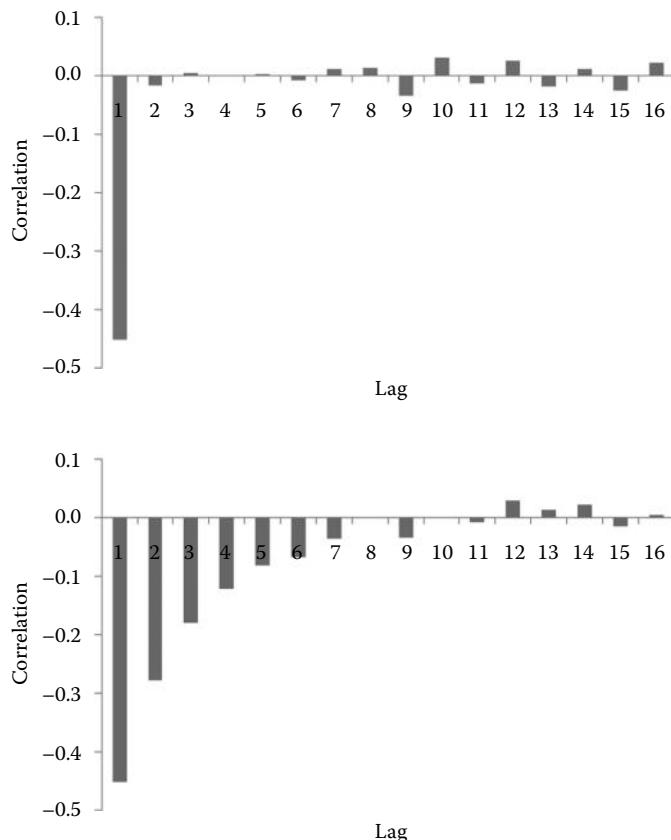
Returning to the traffic and card sales data from the previous examples, Figures 7.11 and 7.12 present the ACF and PACF graphs for volume and occupancy after first-order differencing. The graphs clearly indicate that the variables are now stationary. However, the Dickey–Fuller approach was also used to test for the null hypothesis of the existence of a unit root in volume, occupancy, and travel card sales (Table 7.3 for the results). Some interesting findings emerge; first, despite the “successful”—from a graphical perspective—transformation of the occupancy and volume variables into stationary through first-order differencing (suggesting the existence of a unit root), the tests reject the null hypothesis of a unit root. Second, the tests for the travel cards

**FIGURE 7.11**

ACF (top) and PACF (bottom) graphs for occupancy after first differencing.

sales data do not yield the same result since, if Equation 7.18 was used (unit root), the finding would be that there is a unit root, while if Equations 7.19 and 7.20 were used (unit root with drift and drift and time trend, respectively), the result would suggest rejection of the null hypothesis of a unit root.

As a further step in this investigation, the GPH approach was used to estimate the fractional integration parameter for each of these variables (Geweke and Porter-Hudak 1983). The results show the fractional integration parameter for occupancy and volume to be 0.35 and 0.20 respectively which is a clear indication of long memory (see Karlaftis and Vlahogianni 2009, for a more in-depth discussion of the implications of fractional integration in practice). The fractional integration parameter for the travel card sales is 0.73, indicating that stationarity cannot be verified, further complicating the largely inconclusive findings from the Dickey–Fuller tests.

**FIGURE 7.12**

ACF (top) and PACF (bottom) graphs for volume after first differencing.

TABLE 7.3

Unit-Root Testing and Fractional Integration Parameter Estimation

	Volume	Occupancy	Travel Card Sales
Dickey–Fuller unit-root test (Equation 7.18)	-6.15**	-12.10**	0.30
Dickey–Fuller unit-root test (Equation 7.19)	-21.22**	-20.23**	-5.26**
Dickey–Fuller unit-root test (Equation 7.20)			-5.29**
GPH* estimate of fractional differencing parameter (d)	.202457	.353631	

**Reject null hypothesis of unit root at the 1% significance level.

*Geweke and Porter–Hudak (1983).

7.3 Time Series in Regression

Regression models are often developed using time series data as researchers are interested in examining relationships between variables over time. The use of time series data in regression, despite potential advantages, needs to consider four important aspects: error term serial correlation; dynamic dependence; volatility; and spurious regression, cointegration and causality.

7.3.1 Serial Correlation

Serial correlation, or autocorrelation, is a frequently encountered artifact when using regression to model time series data because the errors associated with observations persist in future observations. Autocorrelation leaves ordinary least squares (OLS) parameter estimates unbiased, but the estimates are no longer efficient. This effect can lead to estimated parameters' standard error being smaller than their true standard errors, leading to larger t-statistics and potentially incorrect conclusions in hypothesis testing. Section 4.5 of the book covers, in detail, methods for detecting and correcting for serial correlation in regression.

7.3.2 Dynamic Dependence

Dynamic dependence is frequently encountered with time series data in regression and occurs when a current outcome depends on past events. To address this dependence, regression models employing time series data frequently include a time lag variable in the regression. For example, assume that a substantial period of time elapses between a change in the price of airline tickets and the effects on passenger demand. If the time between changes in the independent variables and their effects on the dependent variable is sufficiently long, then a lagged explanatory variable may be included in the model. A general form of a regression model with lagged independent variables is given as

$$Y_t = \alpha + \sum_{i=0}^{\infty} \beta_i X_{t-i} + \varepsilon_t \quad (7.21)$$

When the lagged variables have long-term effects, then infinite lag models whose effects fade slowly over time are often used. When changes in the independent variables tend to have short-term effects on the dependent variable, then finite lag models should be considered. An example of an infinite lag model is the geometric lag model where the weights of the lagged variables are all positive and decline geometrically over time. This model

implies that the most recent observations receive the largest weight and that the influence of observations fades uniformly over time. The general form of this model is

$$\begin{aligned} Y_t &= \alpha + \beta(X_t + wX_{t-1} + wX_{t-2}^2 + \dots) + \varepsilon_t \\ &= \alpha + \beta \sum_{s=0}^{\infty} w^s X_{t-s} + \varepsilon_t, \quad 0 < w < 1 \end{aligned} \quad (7.22)$$

The weights of the explanatory variables in the geometric lag model never become zero but decrease in such a way that after a reasonable passage of time the effects become negligible. The form described in Equation 7.22 yields some difficulties regarding the large number of regressors. To simplify the model, if lagging observations are limited to one period ($t-1$), then the model becomes

$$Y_t = \alpha(1-w) + wY_{t-1} + \beta X_t + u_t \quad (7.23)$$

where $u_t = \varepsilon_t - w\varepsilon_{t-1}$.

The adaptive expectations model, another popular specification, assumes that changes in the dependent variable Y_t are related to changes in the expected level of the explanatory independent variable vector X_t . The form of this model is

$$Y_t = \alpha^* + \beta^* X_t^* + \varepsilon_t^* \quad (7.24)$$

where X^* denotes the desired or expected level of X_t and is given by

$$X_t^* = \theta X_t + (1-\theta) X_{t-1}^*; \quad 0 < \theta \leq 1 \quad (7.25)$$

which indicates that the expected level of X_t is a weighted average of its current and past observations.

7.3.3 Volatility

Many transportation applications are concerned with the volatility in a time series; for example, researchers have discussed the volatility of traffic volume around its mean, or the volatility of demand with respect to price increases. With conditional variance (volatility) models, rather than modeling the mean, concern is focused on modeling the variance. In economic time series, variance is usually an expression of risk or volatility and it is frequently observed that large values of a variable X_t have greater fluctuations than small values of X_t . To model such data, Engle (1982) introduced the

autoregressive conditionally heteroscedastic models (ARCH) and Bollerslev (1986) the generalized autoregressive conditionally heteroscedastic models (GARCH). The essential characteristic of these models is that they capture changes in the variability of a time series (in contrast to ARIMA models which assume constant variability, as discussed in Chapter 8). The simplest ARCH(t) model is given as (Shumway and Stoffer 2000)

$$h_t^2 = \text{var}(\varepsilon_t | \psi_t) = a_0 + a_1 \sum_{i=1}^p a_i \varepsilon_{t-i}^2 \quad (7.26)$$

where $\varepsilon_t = Y_t - E[Y_t | X_t]$, the information set $\psi_t = [\varepsilon_{t-j} : j \geq 1]$, $a_0 > 0$, $a_j \geq 0, j = 1, \dots, p$, and where p is the number of lags of the error term to be included in estimating volatility. In the model of Equation 7.26, large positive or negative errors are more likely to be followed by another large error of either sign, and small errors to be followed by small errors. The ARCH(1), with $p = 1$, model is straightforwardly extended to the ARCH(m) form ($p = m$). However, despite their theoretical appeal and practicality, ARCH models assume that shocks, be they negative or positive, have a similar effect on volatility, which may not always be the case. ARCH models also tend to over predict volatility because they are generally slow to respond to large isolated shocks (Tsay 2002). To address some of these shortcomings, Bollerslev (1986) developed the GARCH model as an extension of the ARCH model where

$$h_t^2 = a_0 + \sum_{i=1}^p a_i \varepsilon_{t-i}^2 + \sum_{j=1}^q b_j h_{t-j}^2 \quad (7.27)$$

For example, the first order ($p = q = 1$) GARCH model (GARCH (1,1)) is given as follows

$$h_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 h_{t-1}^2 \quad (7.28)$$

where p is the number of lags of the error term, and q the number of lags of volatility in previous periods, to be included in estimating volatility.

There are other variations of the ARCH model such as the EARCH (Exponential ARCH) model and the IARCH (Integrated ARCH) model (transportation examples with ARCH and GARCH are the papers by Gillen and Martin 2009; Glen and Martin 2004; Harrod 2009; Karlaftis 2010; and Kavussanos et al. 2004). Modeling volatility is also achieved using nonparametric approaches. One of the most commonly used nonparametric approaches for estimating the conditional mean and variance is the Kernel estimator (Tjøstheim 1994). Nonparametric approaches, and particularly Kernel estimators, have found vast applicability in the transportation area (Kharoufeh and Goulias 2002; Smith et al. 1999).

7.3.4 Spurious Regression and Cointegration

Regression models built using time series data are seriously hampered when nonstationary variables are used. As Granger and Newbold (1974) showed, when linear regression models are developed with independent nonstationary variables, the statistical significance of the parameter is overstated. In practice, this indicates that using nonstationary variables in a regression model may lead to spurious results (i.e., although two variables may be unrelated their regression may indicate a statistically significant relationship).

To overcome the possibility of “spurious regression,” statisticians suggest that models are built using first differences of the variables rather than the original series (differencing is discussed in Section 7.1.2). However, many theories (particularly economic theories) are developed upon levels of the variables rather than the differences, and differences do not make full use of these theories. To this end, as Granger (1981) and Granger and Engle (1987) noted, although in the long-run time series may be nonstationary, short-run changes in the series may frequently be stationary. However, concentrating on stationary data leads to loss of valuable long-run information because many long-run associations between variables are meaningful and not spurious.

Murray (1994) discusses an interesting example of the association between two time series. Suppose a drunk is walking quiet randomly (random walk) with his dog; the dog, largely following the drunk, also seems to be on a random walk if viewed in isolation. However, the dog’s path is not aimless and is largely predictable, conditional on the drunk’s position. This phenomenon, by which two processes (variables) may fluctuate to some degree randomly but the distance between them will be relatively constant, was termed by Granger and Engle as “cointegration” and captures the “true” relationship between two nonstationary time series; in essence, two time series are cointegrated when the difference between them is stationary. To formalize the discussion, consider the usual regression model

$$\mathbf{Y}_t = \alpha + \boldsymbol{\beta} \mathbf{X}_t + \varepsilon_t \quad (7.29)$$

where \mathbf{Y}_t and \mathbf{X}_t are $I(1)$ nonstationary time series (they contain a unit root); in general, $\mathbf{Y}_t - \boldsymbol{\beta} \mathbf{X}_t$ is also $I(1)$. However, if ε_t is $I(0)$ (it is stationary), then $\mathbf{Y}_t - \boldsymbol{\beta} \mathbf{X}_t$ is also $I(0)$ suggesting that the linear combination of $\mathbf{Y}_t - \boldsymbol{\beta} \mathbf{X}_t$ has the statistical properties of a stationary variable and variables \mathbf{Y}_t and \mathbf{X}_t are cointegrated. In general, nonstationary variables whose linear combination is stationary are cointegrated, and the vector $[1, -\boldsymbol{\beta}]$ is called the *cointegrating vector* (the concept of cointegration also applies to fractionally integrated and not only to $I(1)$ series). In the case of cointegration, the long-run relationship between \mathbf{Y}_t and \mathbf{X}_t (their joint upward movement over time for example)

is distinguished from their short-run relationship (the deviations of \mathbf{Y}_t and \mathbf{X}_t from their long-run trend). Had differencing been used to make \mathbf{Y}_t and \mathbf{X}_t stationary, however, the long-run relationship between them would have been hidden.

Following the definition and formulation of the concept of cointegration, Granger and Weiss (1983) introduced the Granger representation theorem with far-reaching empirical implications as it estimates long-run relationships between nonstationary variables without differencing. To illustrate, consider a more general form for the regression model of Equation 7.29 where \mathbf{Y}_t and \mathbf{X}_t are cointegrated, ε_t is white noise, and p and q are the lags to be used (chosen so as to “make” ε_t is white noise)

$$\mathbf{Y}_t = \sum_{i=1}^p \gamma_i \mathbf{X}_{t-i} + \sum_{i=1}^q \delta_i \mathbf{Y}_{t-i} + \varepsilon_t \quad (7.30)$$

The Granger representation theorem postulates that the model of Equation 7.30 is regarded as having been generated by an error correction model (ECM) with cointegrating vector $[1, -\boldsymbol{\beta}]$ of the form

$$\Delta \mathbf{Y}_t = a(\mathbf{Y}_{t-1} - \boldsymbol{\beta} \mathbf{X}_{t-1}) + \sum_{i=1}^{p-1} \gamma_i^* \mathbf{X}_{t-i} + \sum_{i=1}^{q-1} \delta_i^* \mathbf{Y}_{t-i} + \varepsilon_t \quad (7.31)$$

Equation 7.30 yields long-run parameters (elasticities if using a double log specification), Equation 7.31 estimates short-run parameters, while parameter a represents the rate of adjustment from short- to long-run equilibrium (called strength of disequilibrium correction). A general strategy for examining and modeling cointegrated series includes the following:

1. Determining if the variables are $I(1)$ (using, e.g., the Dickey–Fuller tests of Equations 7.18 through 7.20).
2. Estimating a cointegrating regression and testing for residual stationarity; if the residuals are $I(0)$, then there is cointegration between the variables.
3. Long-run parameters are obtained by estimating a usual regression (such as Equation 7.30), while short-run parameters and rate of adjustment are obtained by estimating the corresponding ECM model (Equation 7.31 for an example).

There are certain limitations to cointegration analyses. Pretest procedures (step 1 above) may be inconclusive, there may be substantial small-sample bias, and structural breaks in the time series may cause difficulties. In transportation, research using cointegration approaches includes papers by Dargay and Hanly (2002), Romilly (2001), and Crotte et al. (2009).

7.3.5 Causality

An issue closely related to cointegration is that of causality. One would expect that if two variables are cointegrated, then there must be some causality running in, at least, one direction. Causality is a common concern in many statistical investigations. For example, an issue of considerable concern in public transit is in evaluating the effect of subsidies on the performance of a transit system. There may be an issue with reverse causality, which suggests that system performance is not only affected by public transit subsidies but also affects the level of subsidies. If true and left unaccounted for, then endogeneity bias may influence estimated parameters in the resulting statistical model.

To investigate causality (whether evidence supports that changes in one variable cause changes in another) Granger (1969) and later Sims (1972) developed an appropriate procedure. The central idea is based on the premise that while a variable can be predicted from its past values, when a second variable's past values improve on the predictions of the first, then this second variable is said to "Granger-cause" the first. To formalize, consider two time series variables \mathbf{X}_t and \mathbf{Y}_t . Granger-causality tests are conducted by running two sets of regressions on these variables. First, \mathbf{Y}_t is regressed on lagged values of \mathbf{Y}_t and lagged values of \mathbf{X}_t in an unrestricted regression model such that

$$\mathbf{Y}_t = \sum_{i=1}^m a_i \mathbf{Y}_{t-i} + \sum_{i=1}^m \beta_i \mathbf{X}_{t-i} + \varepsilon_t \quad (7.32)$$

where m is the number of lagged variables. Next, \mathbf{Y}_t is regressed only on lagged values of \mathbf{Y}_t in the restricted regression model such that

$$\mathbf{Y}_t = \sum_{i=1}^m a_i \mathbf{Y}_{t-i} + \varepsilon_t^* \quad (7.33)$$

Then, an F -test is conducted on the null hypothesis that $H_0: \beta_1 = \beta_2 = \dots = \beta_m$. To conduct the test, the respective sum of squared residuals from Equations 7.32 and 7.33 are estimated as follows: $SSR_1 = \sum_{t=1}^T \varepsilon_t^2$ and $SSR_0 = \sum_{t=1}^T \varepsilon_t^{*2}$. The test statistic is

$$S = \frac{(SSR_0 - SSR_1)/m}{RSS_1/(T-2m-1)} \sim F_{m, T-2m-1} \quad (7.34)$$

If the test statistic is greater than the critical value of the test at the selected significance level, then the null hypothesis that \mathbf{Y} does not Granger-cause \mathbf{X} is rejected. An asymptotically equivalent test is given by

$$S = \frac{T(\text{SSR}_0 - \text{SSR}_1)}{\text{RSS}_1} \sim \chi^2(m) \quad (7.35)$$

The regression models of Equations 7.32 and 7.33 are then estimated with the dependent and explanatory variables reversed. To ensure objectivity of the results, because the tests are sensitive to the number of lags chosen, tests have to be run using different values for m (number of lags for the regressions). For a discussion on some pitfalls in Granger-causality testing see Lee et al. (2002), for an extension to panel data see Erdil and Yetkiner (2009), while a transportation application of Granger causality can be found in Karlaftis and McCarthy (1998).

Example 7.6

Data from the Athens transit system are used to examine the question of transit performance and its relation to subsidization. This example is similar in aim to Example 4.1 (for further details and an analysis on the issues of subsidization and performance see Iacono 2006; Karlaftis and McCarthy 1998; or Romilly 2001).

TABLE 7.4

Testing for Causality: Does Oe/tvm Help Predict Subsidies?

Variable Description	Dependent Variable: Subsidies	
	Parameter Estimate	t-Statistics
<i>(a) Unrestricted Model</i>		
Intercept	41,763	-0.51
Subsidies lag 1	0.692	3.35
Subsidies lag 2	0.369	0.16
Oe/tvm lag 1	84,094	1.95
Oe/tvm lag 2	-67,174	-1.5
Number of observations	28	
RSS ₁	8.48 × 10 ⁹	
<i>(b) Restricted Model</i>		
Intercept	4,680	0.55
Subsidies lag 1	0.8128	0.0006
Subsidies lag 2	0.2580	0.258
Number of observations	28	
RSS ₀	1.07 × 10 ¹⁰	

TABLE 7.5

Testing for Causality: Does Subsidy Help Predict Oe/tvm? (2 Lags Used)

Variable Description	Dependent Variable: Oe/tvm	
	Parameter Estimate	t-Statistics
<i>(a) Unrestricted Model</i>		
Intercept	0.050	1.37
Oe/tvm lag 1	1.205	5.74
Oe/tvm lag 2	-0.261	-1.19
Subsidies lag 1	1.82×10^{-6}	1.74
Subsidies lag 2	-1.32×10^{-6}	-1.19
Number of observations	28	
RSS ₁	0.21719	
<i>b) Restricted Model</i>		
Intercept	0.057	1.36
Oe/tvm lag 1	1.247	5.96
Oe/tvm lag 2	-0.224	-1.02
Number of observations	28	
RSS ₀	0.24886	

TABLE 7.6

Testing for Causality: Does Subsidy Help Predict Oe/tvm? (3 Lags Used)

Variable Description	Dependent Variable: Oe/tvm	
	Parameter Estimate	t-Statistics
<i>(a) Unrestricted Model</i>		
Intercept	0.061	1.76
Oe/tvm lag 1	1.321	6.82
Oe/tvm lag 2	-1.01	-3.05
Oe/tvm lag 3	0.66	2.91
Subsidies lag 1	2.99×10^{-6}	3.01
Subsidies lag 2	-2.20×10^{-6}	-1.81
Subsidies lag 3	2.77×10^{-7}	0.25
Number of observations	28	
RSS ₁	0.144	
<i>(b) Restricted Model</i>		
Intercept	0.063	1.46
Oe/tvm lag 1	1.305	6.31
Oe/tvm lag 2	-0.657	-2.02
Oe/tvm lag 3	0.398	1.77
Number of observations	28	
RSS ₀	0.21628	

In both transportation and economics, an issue of importance is the effect of subsidies on a system's performance and whether "reverse causality" may be present. Most empirical work assumes that subsidies affect performance in that higher subsidies may lower (or raise) a system's performance. However, potential endogeneity bias may be present if system performance is not only affected by subsidies but also affect subsidy levels. To test for potential bias, annual performance data (operating expenses per total vehicle kilometers (oe/tvm)) and subsidies (in €10³) from 1980 to 2007 were collected and Granger-causality tests conducted.

The results for the models tested appear in Tables 7.4 through 7.6; models are based on Equations 7.32 and 7.33. The two-lag regressions in Table 7.4 check for the hypothesis that performance (oe/tvm) helps predict subsidies. (i.e., whether oe/tvm Granger-cause subsidies). Using Equations 7.34 and 7.35, the test results are 3.09 and 7.54, respectively, which, at the 95% significance level, leads to a rejection of the null hypothesis that performance does not Granger-cause subsidies. From a policy perspective this result suggests that performance Granger-causes subsidies.

The same approach was used to test whether subsidies Granger-cause performance (Table 7.5). Using the same models (with two lags) and the same significance level, the hypothesis that subsidies do not Granger-cause performance could not be rejected. Thus, it does not appear that subsidies Granger-cause performance. In contrast, when three lags were used in the regressions (Table 7.6) the finding suggests that subsidies Granger-cause performance.

This difference in findings, which may have important modeling and policy implications, shows that Granger-causality tests are sensitive to the choice of lag length used for the models.

8

Forecasting in Time Series: Autoregressive Integrated Moving Average (ARIMA) Models and Extensions

Methods like time series decomposition or exponential smoothing discussed in the previous chapter are often used (besides smoothing) to predict future values of variables observed over time. This chapter focuses on the ARIMA family of models that are mathematical models of the persistence or auto-correlation (correlation across values) in a time series. Unlike the use of time series in regression, ARIMA models describe the behavior of a variable in terms of its past values. These models are rather simple and straightforward to develop and are useful for forecasting time series even in the absence of explanatory variables. ARIMA models are widely used in almost all fields of transportation with marked success; examples of such applications are Lau et al. (2009), Masten (2007), and Smith et al. (2002).

8.1 Autoregressive Integrated Moving Average Models

Developed and discussed throughout this chapter, the usual notation for an ARIMA model is ARIMA(p, d, q) where p refers to the models' autoregressive order, q refers to the moving average order, and d refers to the degree of differencing needed to achieve stationarity (see Section 7.2.1 for a discussion on stationarity). When a series is stationary, ARIMA models become ARMA. There are three common types of time series models developed in most research: the autoregressive models (AR), the moving average models (MA), and the autoregressive moving average models (ARMA).

In AR, the current observation in a series is expressed as a linear function of p previous observations, a constant term, and a disturbance term, and is expressed as

$$\begin{aligned} X_t &= k_1 X_{t-1} + k_2 X_{t-2} + \dots + k_p X_{t-p} + \theta_0 + \varepsilon_t \\ &= \theta_0 + \sum_{i=1}^p k_i X_{t-i} + \varepsilon_t \end{aligned} \tag{8.1}$$

where $X_t, X_{t-1}, \dots, X_{t-p}$ are the observations in periods $t, t-1, \dots, t-p$, p is the number of periods (lags) considered in the development of the model k_i are the autoregressive parameters θ_0 is the constant term, and ε_t is the disturbance for period t (this model is written as AR(p)). Provided the process is stationary, then its mean μ is given as

$$E(X_t) = \mu = \frac{\theta_0}{1 - k_1 - k_2 - \dots - k_p} \quad (8.2)$$

and its variance γ_0 as

$$VAR(X_t) = \gamma_0 = \frac{\sigma_\alpha^2}{1 - k_1\rho_1 - k_2\rho_2 - \dots - k_p\rho_p} \quad (8.3)$$

where σ_α^2 is the variance of the disturbances. Finally, the autocorrelation coefficient is expressed by the following recursive equation

$$\rho_\kappa = k_1\rho_{\kappa-1} + k_2\rho_{\kappa-2} + \dots + k_p\rho_{\kappa-p} \quad (8.4)$$

Consider two popular cases of the general model presented in Equation 8.1. The AR(1) has only one term

$$X_t = k_1 X_{t-1} + \theta_0 + \varepsilon_t \quad (8.5)$$

from which it is inferred that, for the process to be stationary $|k_1| < 1$. It is also inferred that the partial autocorrelation coefficients $\phi_{11} = \rho_1$, and $\phi_{\kappa\kappa} = 0$ for $\kappa > 1$ (see Section 7.2.2 for a short description of partial autocorrelation coefficients). The AR(2) has two terms

$$X_t = k_1 X_{t-1} + k_2 X_{t-2} + \theta_0 + \varepsilon_t \quad (8.6)$$

In this case, there are three stationarity constraints: $|k_2| < 1$, $k_1 + k_2 < 1$, $k_2 - k_1 < 1$ and, $\phi_{11} = \rho_1$, $\phi_{22} = (\rho_2 - \rho_1^2)/(1 - \rho_1^2)$, $\phi_{\kappa\kappa} = 0$ for $\kappa > 2$.

In the MA of order q it is assumed that the current observation is the sum of the current and weighted past disturbances as well as a constant term

$$\begin{aligned} X_t &= \theta_0 + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - \dots - \lambda_q \varepsilon_{t-q} \\ &= \theta_0 - \sum_{i=1}^q \lambda_i X_{t-i} + \varepsilon_t \end{aligned} \quad (8.7)$$

where X_t is the observation in period t , q is the number of periods (order), λ_i are the moving average parameters, θ_0 is the constant term, and ε_t is the random disturbances term for periods $t, t-1, \dots, t-q$. The mean, variance, autocorrelation coefficient, and partial autocorrelation coefficients for this series are given as

$$\begin{aligned} E(X_t) &= \mu = \theta_0 \\ VAR(X_t) &= \gamma_0 = \sigma_\alpha^2 \cdot (1 + \lambda_1^2 + \lambda_2^2 + \dots + \lambda_q^2) \\ \rho_\kappa &= \begin{cases} (-\lambda_\kappa + \lambda_1 \lambda_{\kappa+1} + \lambda_2 \lambda_{\kappa+2} + \dots + \lambda_{q-\kappa} \lambda_\kappa) \cdot \left(\frac{\sigma_\alpha^2}{\gamma_0}\right), & \kappa = 1, 2, \dots, q \\ 0 & \text{otherwise} \end{cases} \\ \phi_{\kappa\kappa} &= -\lambda_1^\kappa \left(\frac{1 - \lambda_1^2}{1 - \lambda_1^{2\kappa+2}} \right) \end{aligned} \quad (8.8)$$

A special case of the previous model, MA(2), is given as

$$X_t = \theta_0 + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} \quad (8.9)$$

with parameters that satisfy: $|\lambda_1| < 1, \lambda_1 + \lambda_2 < 1, \lambda_2 - \lambda_1 < 1$.

Furthermore, there are time series models that have both autoregressive and moving average terms. These models are written as ARMA(p, q) and have the following general form

$$\begin{aligned} X_t &= \theta_0 + \varepsilon_t + k_1 X_{t-1} + k_2 X_{t-2} + \dots + k_p X_{t-p} - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - \dots - \lambda_q \varepsilon_{t-q} \\ &= \theta_0 + \sum_{i=1}^p k_i X_{t-i} - \sum_{i=1}^q \lambda_i X_{t-i} + \varepsilon_t \end{aligned} \quad (8.10)$$

where the notation is consistent with that of Equations 8.1 and 8.7. Stationarity must be ensured to implement the model. As an example, consider the ARMA(3,2) model that combines an autoregressive component of order 3 and a moving average component of order 2

$$X_t = \theta_0 + \varepsilon_t + k_1 X_{t-1} + k_2 X_{t-2} + k_3 X_{t-3} - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} \quad (8.11)$$

There are more advanced models that belong to the ARIMA family such as the seasonal ARMA and ARIMA (SARMA and SARIMA) and the general multiplicative model. The SARMA(P, Q) and SARIMA(P, D, Q) models

are extensions of the ARMA and ARIMA models and include seasonality effects of time series. The general multiplicative model is used to account for possible autocorrelation both for seasonal and non seasonal lags. Its order is $(p, q, d) \times (P, Q, D)_S$. A special case of the model is the ARIMA $(0, d, 1) \times (0, D, 1)_S$ —a common moving average model in both the seasonal and non-seasonal parts

$$W_t = \lambda_0 + (1 - \lambda_1 B) \cdot (1 - \Lambda_1 B^S) \varepsilon_t \quad (8.12)$$

where the backward shift operator is B for nonseasonal effects and B^S for seasonal effects (see Section 7.1.8 for a discussion on backward shift operators). In this model, there are nonzero autocorrelations at lags 1, $S - 1$, S , $S + 1$, for which the autocorrelation coefficients are

$$\rho_1 = \frac{-\lambda_1}{1 + \lambda_1^2}, \quad \rho_S = \frac{-\Lambda_1}{1 + \Lambda_1^2}, \quad \rho_{S-1} = \rho_{S+1} = \rho_S \rho_1 \quad (8.13)$$

For more detailed information regarding the seasonal and multiplicative models readers should refer to Box and Jenkins (1976), Abraham and Ledolter (1983), and Williams and Hoel (2003) for a transportation application.

8.2 Box–Jenkins Approach

The Box and Jenkins (1976) approach to estimating models of the general ARIMA family is currently one of the most widely implemented strategies for modeling univariate time series data. Although the four-stage approach (order selection, parameter estimation, diagnostic checking, and forecasting) was originally designed for modeling time series with the ARIMA model, the underlying strategy is applicable to a wide variety of statistical modeling approaches. It provides a logical framework that helps researchers find their way through data uncertainty (Makridakis et al. 1989).

8.2.1 Order Selection

The first step is order selection. The task of identifying the proper model is probably the most difficult one in practice. The values of p and q , the orders of the autoregressive and moving average terms, need to be obtained before applying a model. The selection of p and q is a tedious process because many possible combinations need to be examined. This selection is most frequently

done by examining the autocorrelation and partial autocorrelation functions (ACF and PACF, respectively) of the variable(s) to be modeled. It is important to note that before attempting to specify p and q , the data must be stationary with the values of p and q found by examining the ACF and PACF of stationary data. The general rules that apply in interpreting these two functions appear in Table 8.1. When the ACF tails off exponentially to 0, the model is AR and its order is determined by the number of significant lags in the PACF (Graphical examples of ACF and PACF corresponding to frequently encountered models appear in Figures 8.1 through 8.4.). When the PACF tails off exponentially to 0 the model is a MA, and its order is determined by the number of statistically significant lags in the ACF. When both the ACF and PACF tail-off exponentially to 0 the model is ARMA. Finally, when the data is seasonal the same procedure is used to estimate the P , D , Q parameters.

Besides the graphical approach to order selection, there are a number of information criteria for obtaining the order of (p, q) . The common approach is to select pairs of (p, q) that minimize some function that is based on an estimate $\hat{\sigma}_{p,q}^2$ of the variance of ε_t (Equation 8.10). The most widely used functions include

1. Akaike's information criterion (AIC)

$$AIC(p, q) = \log(\hat{\sigma}_{p,q}^2) + 2 \frac{p+q+1}{n+1} \quad (8.14)$$

2. Bayesian information criterion (BIC)

$$BIC(p, q) = \log(\hat{\sigma}_{p,q}^2) + \frac{(p+q)\log(n+1)}{n+1} \quad (8.15)$$

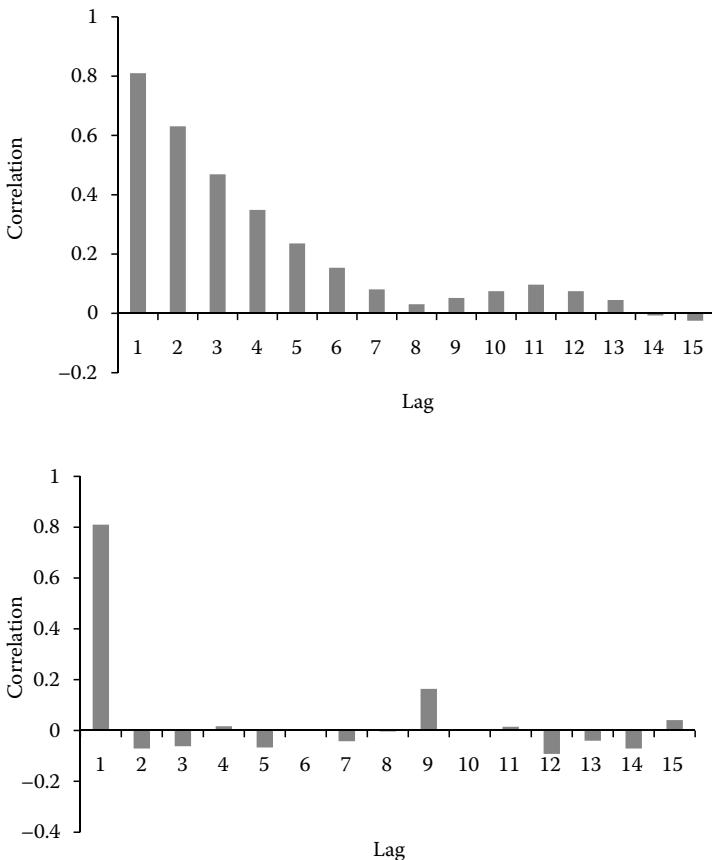
3. Hannan–Quinn criterion

$$HQ(p, q) = \log(\hat{\sigma}_{p,q}^2) + 2 \frac{2(p+q)c \log(\log(n+1))}{n+1}, \quad c > 1 \quad (8.16)$$

TABLE 8.1

ACF and PACF Behavior for ARMA Models

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off exponentially	Cuts off after lag q	Tails off exponentially
PACF	Cuts off after lag p	Tails off exponentially	Tails off exponentially

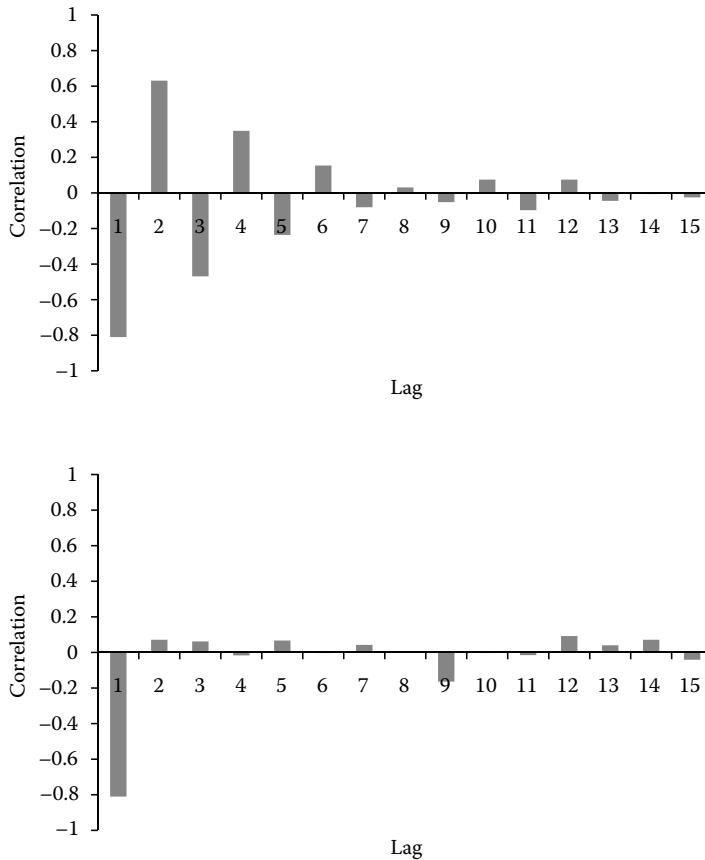
**FIGURE 8.1**

Typical ACF (top) and PACF (bottom) graphs for an AR(1) model with $k_1 = 0.8$.

As Falk et al. (2006) suggest, although the BIC and HQ lead to strongly consistent estimators for model order, AIC does not underestimate the model's order, while the BIC criterion is generally preferred for large samples.

8.2.2 Parameter Estimation

The second step is estimating model parameters. After selecting a model based on its ACF and PACF, its parameters are estimated by maximizing the corresponding likelihood function. However, there are no closed-form expressions for the maximum likelihood estimators for the parameters in the general ARMA(p, q) model. Even for the case of the AR(1) model, the maximum likelihood estimate involves the solution of a cubic equation. Because of the complexity of the MLE procedure for the ARMA (p, q) model, other estimation procedures such as the conditional least squares and unconditional

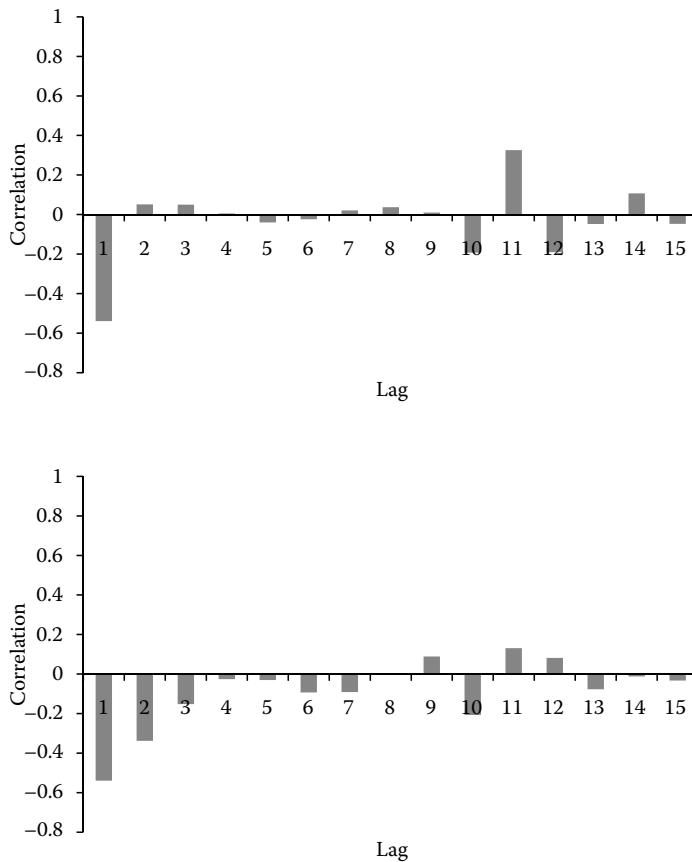
**FIGURE 8.2**

Typical ACF (top) and PACF (bottom) graphs for an AR(1) model with $k_1 = -0.8$.

least squares have been developed and are widely used in practice (for details see Abraham and Ledolter 1983; Box and Jenkins 1976; Cryer 1986; and Falk et al. 2006).

8.2.3 Diagnostic Checking

The third step is the examination of the fitted model's disturbances. These errors may be either random or follow a systematic pattern. If the errors are random, the implication is that the selected model has eliminated the underlying patterns from the data and only random errors remain (similar to an ideal linear regression analysis). If the errors follow a systematic pattern, the underlying patterns in the data were not completely removed and, as such, the disturbances are not random. In a correctly specified model, the residuals are uncorrelated as reflected by the absence of significant values in the

**FIGURE 8.3**

Typical ACF (top) and PACF (bottom) graphs for an MA(1) model with $\lambda_1 = +0.6$.

residual term's ACF. Further, the Ljung–Box $Q(m)$ statistic as described in Section 7.2.3 is used to check the closeness of the residuals to white noise (white noise is a stationary time series with zero autocorrelation). Note, however, that for residuals the $Q(m)$ statistic is asymptotically χ^2 distributed with $m-p-q$ degrees of freedom (the degrees of freedom are modified to account for the p AR and q MA parameters estimated). If the residuals are random (white noise), the model is suitable for forecasting.

8.2.4 Forecasting

After the model is developed and the residuals checked for randomness, the model is suitable for forecasting. Assuming the model is a (stationary)

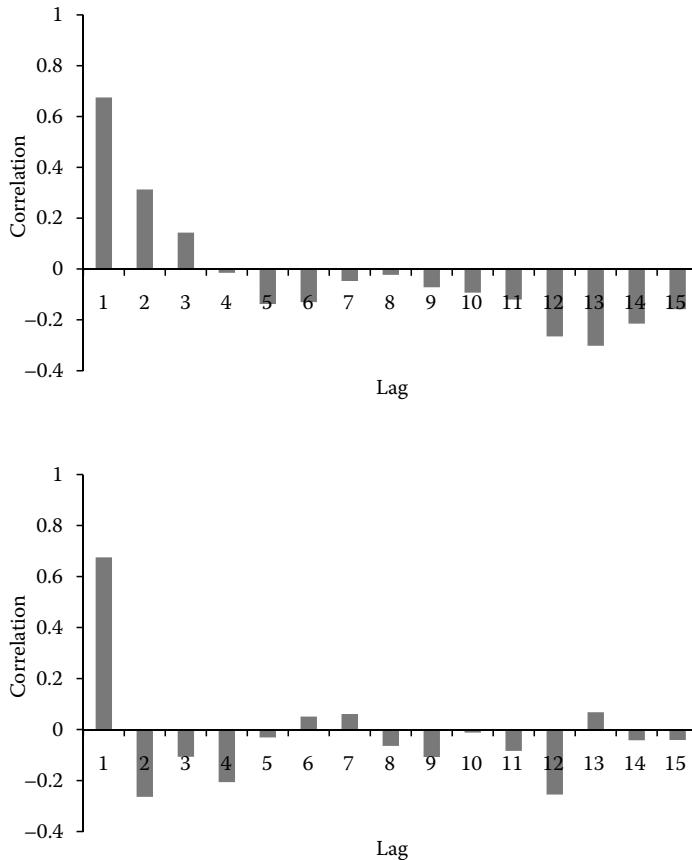


FIGURE 8.4
Typical ACF (top) and PACF (bottom) graphs for an ARMA(1,1) model.

ARMA(p, q) process, the best one-step ahead forecast is given by (Tsay 2002):

$$\hat{X}_{t+1} = E(X_{t+1} | X_t, X_{t-1}, \dots, X_{t+1-p}) = \theta_0 + \sum_{i=1}^p k_i X_{t+1-i} - \sum_{i=1}^q \lambda_i \epsilon_{t+1-i} \quad (8.17)$$

with forecast error $e_{t+1} = X_{t+1} - \hat{X}_{t+1} = \epsilon_{t+1}$ and error variance $VAR(e_{t+1}) = \sigma_\epsilon^2$. Recursively, the h -step ahead forecast is given as

$$\begin{aligned} \hat{X}_{t+h} &= E(X_{t+h} | X_{t+h-1}, X_{t+h-2}, \dots, X_{t+1-p}) \\ &= \theta_0 + \sum_{i=1}^p k_i X_{t+h-i} - \sum_{i=1}^q \lambda_i \epsilon_{t+h-i} \end{aligned} \quad (8.18)$$

with forecast error $e_{t+h} = X_{t+h} - \hat{X}_{t+h}$; note that Equation 8.18 holds for $h - i \leq 0$ (the order of the model is larger than the forecast horizon), while when $h - i > 0$ then $\varepsilon_{t+h-i} = 0$.

Because forecasting is often the primary goal of time series analysis, it is imperative that predictive accuracy be assessed. Usually, accuracy implies how well the model reproduces the already known data (goodness-of-fit) (Makridakis et al. 1989). An accuracy measure is often defined in terms of the forecasting error, which is the difference of the actual and the predicted values of the series (ε_i). Some measures of accuracy that are used with most time series models are shown in Table 8.2. Many of these measures can also be used for other types of models, including linear regression, count models, and so on.

The first four performance measures shown in Table 8.2 are absolute measures and are of limited value when used to compare different time series. The MSE is the most frequently used measure in the literature. However, its value is frequently questioned in evaluating the relative forecasting accuracy between different data sets because of its dependence on the scale of measurement and its sensitivity to outliers (Chatfield 1992; Clemens and Hendry 1993). In transportation modeling, commonly used measures for evaluating the accuracy of

TABLE 8.2

Some Measures of Accuracy Used to Evaluate Time Series Models

Measure	Corresponding Equation
Mean error	$ME = \frac{\sum_{i=1}^n \varepsilon_i}{n}$
Mean absolute deviation	$MAD = \frac{\sum_{i=1}^n \varepsilon_i }{n}$
Sum of squared errors	$SSE = \sum_{i=1}^n \varepsilon_i^2$
Mean squared error	$MSE = \frac{\sum_{i=1}^n \varepsilon_i^2}{n}$
Root mean squared error	$RMSE = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n}}$
Standard deviation of errors	$SDE = \sqrt{\frac{\sum_{i=1}^n \varepsilon_i^2}{n-1}}$
Percentage error	$PE_t = \frac{X_t - F_t}{X_t} \cdot 100\%$
Mean percentage error	$MPE = \frac{\sum_{i=1}^n PE_i}{n}$
Mean absolute percentage error	$MAPE = \frac{\sum_{i=1}^n PE_i }{n}$

the forecasting models are the mean square error (MSE), the mean absolute deviation (MAD), and the mean absolute percent error (MAPE). Much of the time series literature related to transportation problems has demonstrated the need to use absolute percent error as a basis for comparison in order to eliminate the effect of the variability observed in most transportation data sets.

Example 8.1

Continuing with the use of the traffic dataset used in Chapter 7 (Examples 7.2, 7.4, and 7.5), an ARMA forecasting model is developed for both volume and occupancy using the Box–Jenkins approach outlined in Section 8.2.

Step 1: Order Selection

As depicted in Figure 8.5, the ACF for both variables demonstrate considerably slow decay indicating a lack of stationarity (this finding was also verified from the

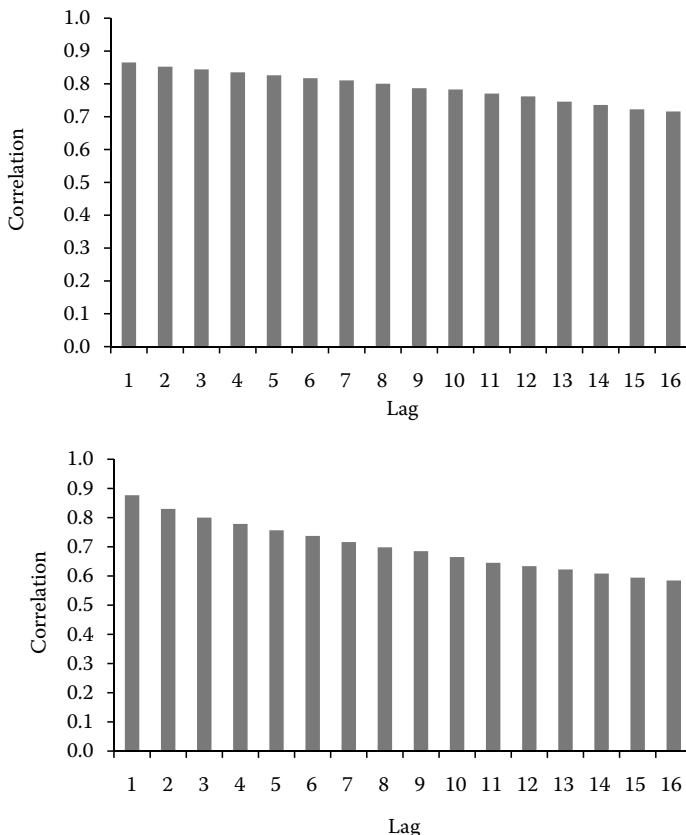


FIGURE 8.5
ACF graphs for volume (top) and occupancy (bottom).

Dickey–Fuller tests for unit roots presented in Example 7.5). Interestingly, using only first-order differencing yields ACF and PACF for both variables that tail off exponentially (Figures 8.6 and 8.7). As previously indicated, series whose ACF and PACF both tail off exponentially lend themselves to a combined ARMA process the characteristics of which indicate an ARMA(1,1) model. However, a number of models were developed and the BIC from Equation 8.15 was also estimated for all the models developed.

Step 2: Parameter Estimation

Parameters were estimated for a variety of models using maximum likelihood estimation and the results for two of the models appear in Table 8.3. All parameters (both for the AR and the MA terms) in both models are statistically significant at the 95% significance level.

Step 3: Diagnostic Checking

In this step, the fitted-model disturbances are checked; for the model to be correctly specified and suitable for forecasting, residuals should be random and not follow a systematic pattern. Figure 8.8 depicts the residual ACF graphs for both volume and occupancy for the ARIMA(1,1) model whose parameters were shown above. For both variables, the ACF indicates a lack of a systematic pattern and the appearance of white noise; this observation was examined further using the Ljung–Box statistic (Section 7.2.3), which, at the 95% significance level, indicates the residuals being close to white noise.

Step 4: Forecasting

The final step is forecasting; Tables 8.4 and 8.5 present the forecasting accuracy results and BIC values from the application of a variety of different approaches to both variables, including exponential smoothing, AR, MA, and ARMA models (all models were estimated on the differenced data). The success of these approaches for one-step ahead forecasting was evaluated with the MAE and MAPE measure; as is apparent, all models yield similar forecasts and BIC values and could be used for forecasting. As a final step in evaluating the forecasting process, Figure 8.9 depicts the actual against the predicted values for the ARIMA(1,1) models. These graphs are useful tools in evaluating the forecasting ability of the developed models and for uncovering outliers and failed forecasts (as are present in the forecasts for volume) as well as possible heteroskedasticity in the forecasts.

8.3 Autoregressive Integrated Moving Average Model Extensions

Some interesting extension to the basic ARIMA models described and developed in Section 8.2 are the random parameter autoregressive (RPA) models,

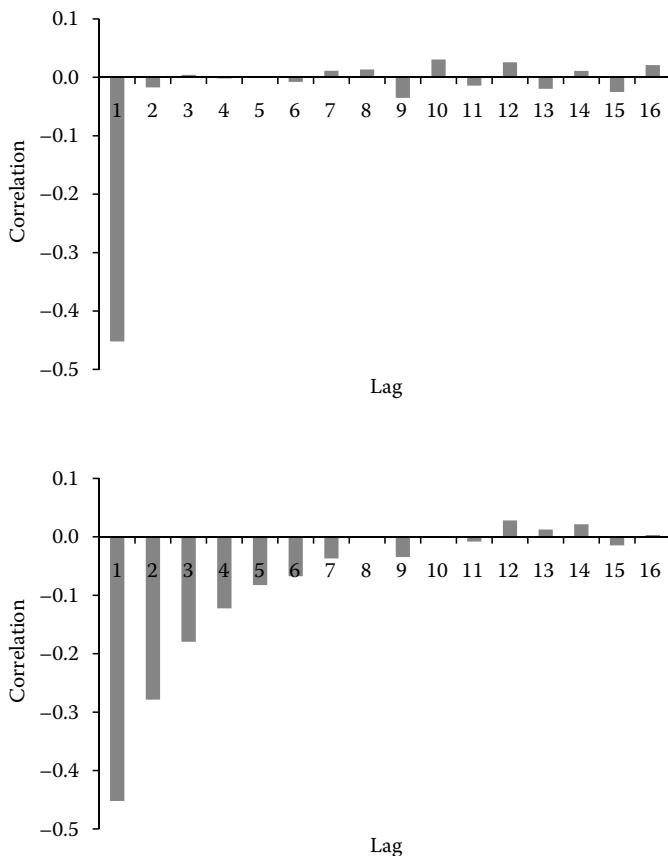
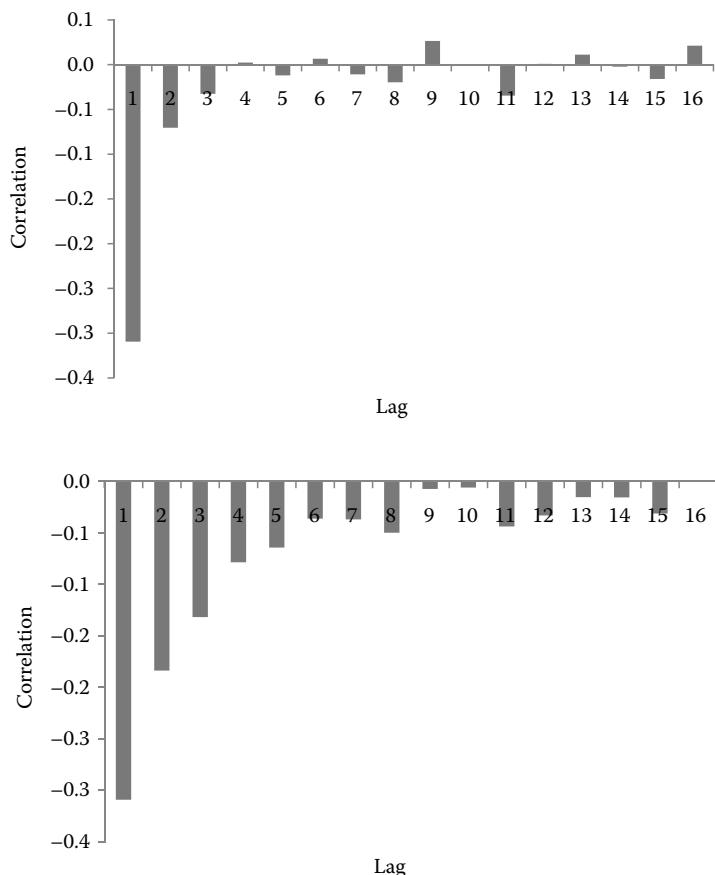


FIGURE 8.6
ACF (top) and PACF (bottom) graphs for volume after first differencing.

the stochastic volatility (SV) models, the autoregressive conditional duration (ACD) models, and the integer-valued ARMA (INARMA) models.

8.3.1 Random Parameter Autoregressive Models

The modeling approaches presented in the previous section assume that the estimated parameters are constant across observations. However, there may be some unobservable factors that affect the observations differentially. The random parameters autoregressive RPA model has been introduced to account for variability across observations with the parameters of the model varying across observations according to some distribution (see Chapter 16).

**FIGURE 8.7**

ACF (top) and PACF (bottom) graphs for occupancy after first differencing.

TABLE 8.3

Parameter Estimates for the ARIMA(1, 1) Models

Volume	Coefficient Estimate	t-Test	p-Value
Constant	0.001	0.017	.987
Volume lag 1	0.193	2.997	.019
MA lag 1	0.690	51.163	.000
<i>Occupancy</i>			
Constant	0.001	0.018	.986
Occupancy lag 1	0.299	12.804	.000
MA lag 1	0.707	40.852	.000

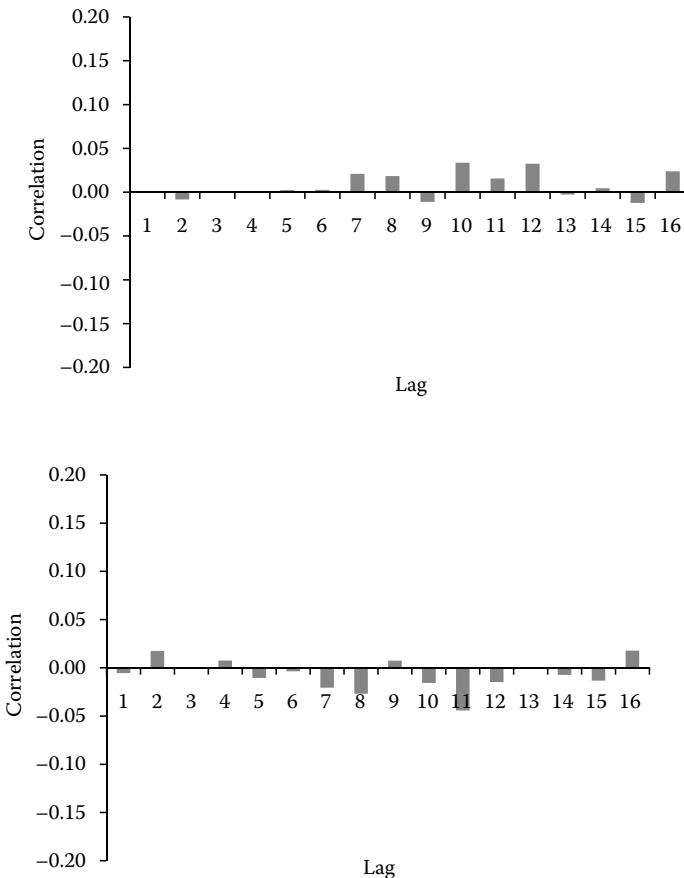


FIGURE 8.8
Residual ACF graphs for volume (top) and occupancy (bottom).

for an extensive discussion of random parameter models). When a time series X_t has an RPA(p) specification, it is written as

$$X_t = \theta_0 + \sum_{i=1}^p (k_i + \varphi_{it}) X_{t-i} + \varepsilon_t \quad (8.19)$$

where p is the order of the model and φ_{it} are the random terms (that are independent of ε_t and X_t) with zero mean and covariance matrix Ω_φ (Nicholls and Quinn 1982).

TABLE 8.4

Results Comparison for Different Time Series Models (Volume)

Volume	Mean Absolute Error	Mean Absolute Percent Error (%)	BIC
Exponential smoothing	11	13.14	5.43
AR(1)	12	12.92	5.57
MA(1)	11	13.14	5.44
ARMA(1, 1)	11	13.16	5.42
AR(2)	12	12.63	5.49
MA(2)	11	13.16	5.43
ARMA(2, 2)	11	13.15	5.43

TABLE 8.5

Results Comparison for Different Time Series Models (Occupancy)

Occupancy	Mean Absolute Error	Mean Absolute Percent Error (%)	BIC
Exponential smoothing	5.5	26.32	4.25
AR(1)	5.7	26.77	4.30
MA(1)	5.5	26.33	4.25
ARMA(1, 1)	5.5	26.97	4.23
AR(2)	5.6	26.38	4.27
MA(2)	5.5	26.68	4.24
ARMA(2, 2)	5.5	26.94	4.24

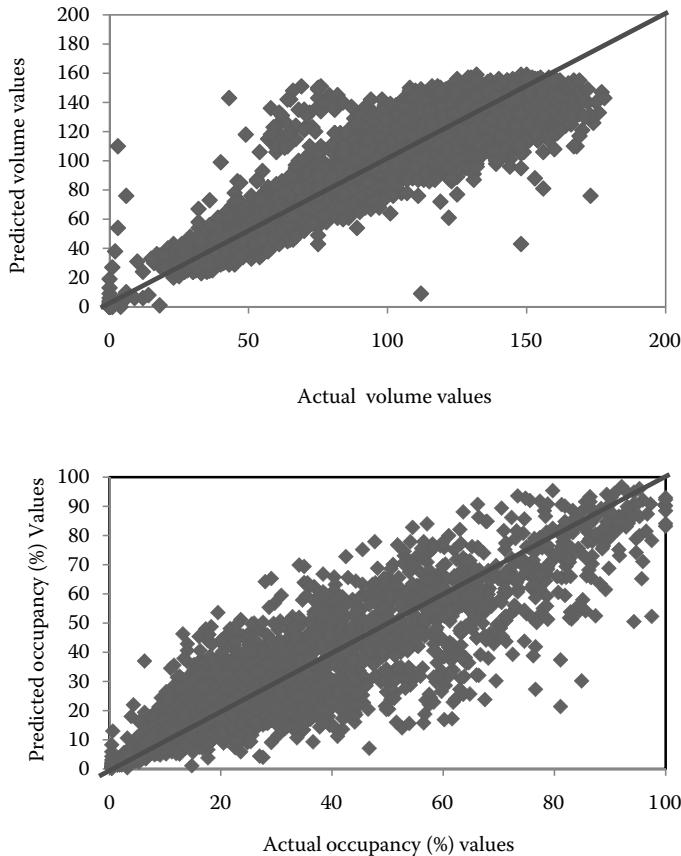
8.3.2 Stochastic Volatility Models

As noted in Chapter 7 (Section 7.3.3), interest frequently centers around the volatility in a time series such as the volatility of traffic volume around its mean. SV models have been developed to combine AR models with the evolution of volatility (as was the case for the ARCH models discussed in Section 7.3.3). For example, consider a time series X_t with a generalized auto regressive conditionally heteroscedastic (GARCH) (1, 1) error structure. Following Shumway and Stoffer (2000), X_t is written as

$$X_t = h_t \varepsilon_t \quad (8.20)$$

with $h_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + b_1 h_{t-1}^2$. Equation 8.20 is rewritten as

$$X_t = \exp\left(\frac{1}{2} \log h_t^2\right) \varepsilon_t \quad (8.21)$$

**FIGURE 8.9**

Actual versus predicted graphs for volume (top) and occupancy (bottom).

Let $h_t^* = \ln h_t^2$ and $X_t^* = \text{LN}(X_t^2)$ and rewrite Equation 8.21 as

$$X_t^* = h_t^* + \text{LN}(\varepsilon_t^2) \quad (8.22)$$

Equation 8.22 is considered an observation equation with the stochastic variance h_t^* as the unobservable state process. Similar to the basic GARCH equation, the volatility follows an autoregression as follows

$$h_t^* = \gamma_0 + \gamma_1 h_{t-1}^* + u_t \quad (8.23)$$

where u_t is white noise. Equations 8.22 and 8.23 jointly form the SV model that could be estimated as a state-space model (see Section 8.4 for more details on state-space models), provided ε_t^2 is log-normally distributed (Shumway

and Stoffer 2000). However, because this condition is rarely observed, SV model estimation is exceedingly complex and usually done using Bayesian methods (see Chapter 17).

8.3.3 Autoregressive Conditional Duration Models

As discussed in Chapter 10 more extensively, duration models are concerned with the time elapsed until the occurrence of an event or the duration of an event. This type of investigation may also present itself in the case of time series data when researchers are interested in the duration of congestion or the time until congestion occurs using time series traffic data (Stathopoulos and Karlaftis 2002).

Using the principles of GARCH models and similar to the approach of SV models discussed above, Engle and Russell (1998) proposed the ACD(r,s) model as follows:

$$\begin{aligned} X_i &= D_i \varepsilon_i \\ D_i &= \alpha + \sum_{j=1}^r \gamma_j X_{i-j} + \sum_{j=1}^s \delta_j D_{i-j} \end{aligned} \quad (8.24)$$

where i is the event whose duration X_i is studied, X_i is a time series, D_i is the expected adjusted duration between the $(i - 1)$ th and the i th events and ε_i is independently and identically distributed nonnegative random variables with $E(\varepsilon_i) = 1$. When ε_i is exponentially distributed, the model is referred to as the exponential ACD model or EACD(r,s), while when ε_i follows the Weibull distribution then the model is referred to as the Weibull ACD model or WACD(r,s) (for estimation and computation of ACD models see Tsay (2002)).

8.3.4 Integer-valued ARMA Models

In many cases in transportation research, count data are available on a time series dimension (a time series of count data is an integer-valued nonnegative sequence of count observations over time, as is frequently encountered in safety investigations). Several models for the analysis of time series of count data are available, but the INARMA class of models, which evolved similarly to the continuous ARMA models, have found applications in many research areas (Brijs et al. 2008; Quddus 2007, 2008; Yannis and Karlaftis 2010). The most commonly encountered form of the INARMA model is the INAR(1) process that is defined as

$$Y_t = \alpha Y_{t-1} + \varepsilon_t, \quad t = 2, \dots, T \quad (8.25)$$

where ε_t is assumed independently and identically distributed Poisson with $E(\varepsilon_t) = \lambda > 0$ and independent of Y_{t-1} . McKenzie (1985) has shown

when $\alpha \in (0, 1)$ and Y_t is discrete self-decomposable, the AR(1) process is stationary (as Van Harn and Steutel (1979) prove, discrete self-decomposable distributions have properties similar to those of their continuous counterparts). This model follows the “usual” AR(1) model in that it explicitly models serial correlation as lags of the endogenous variables, but where the scalar multiplication is replaced by a binomial thinning operator (α). The operator, introduced by Van Harn and Steutel (1977), is defined as $\alpha y = \sum_{i=1}^y u_i$, where u_i is a sequence of binary random variables where each component i either “survives” with probability α (i.e. $u_i = 1$) or does “not survive” with probability $(1 - \alpha)$. First- and second-order moments, properties, and assumptions made by this model is found in Brannas (1994, 1995).

8.4 Multivariate Models

The models discussed in the previous sections are univariate in nature in that they are used to forecast only one variable at a time. However, in the long run, a single variable may not be appropriate because the phenomenon under study may be influenced by other external events. In general, there are three approaches for modeling and forecasting multivariate time series: the transfer function models, the vector AR, and the state–space (Kalman filter) models. Kalman Filters are the most general of all multivariate models and are briefly reviewed here.

Before proceeding, it is important to note that the state–space model and the more widely known Kalman filter model refer to the same basic underlying model. The term state–space refers to the model and the term Kalman filter refers to the estimation of the state. Most times, the process of estimating a state–space model begins by estimating an ARMA equivalent to capture the statistically significant AR and MA dimensions. However, the state–space approach has advantages over the more widely used family of ARIMA models (see Durbin 2000). First, it is based on the analysis of the model structure and as such the various components that make up the series (trend, seasonal variation, cycle, etc.) together with the effects of explanatory variables and interventions are modeled separately before being integrated in the final model. Second, it is flexible in that changes in the structure of the modeled system over time are explicitly modeled. Third, multivariate observations are handled as straightforward extensions of univariate theory unlike ARIMA models. Fourth, it is rather simple to deal with missing observations. Fifth, the Markovian nature of state–space models enables increasingly large models to be handled effectively without disproportionate increases in computational efficiency. And, finally, it is straightforward to

predict future values by projecting the Kalman filter forward into the future and obtaining their estimated values and standard errors.

On the other hand, the mathematical theory of state space is quite complex. To understand it fully on a mathematical level requires a good background in probability theory, matrix theory, multivariate statistics, and the theory of Hilbert space. However, the state-space model is so useful, particularly for multivariate time series processes, that it deserves wider attention in the transportation literature. Readers interested in the finer mathematical details of state-space modeling should refer to Shumway and Stoffer (2000) and Durbin (2000); Whittaker et al. (1997), and Stathopoulos and Karlaftis (2003) offer transportation applications of this approach. Other approaches to multivariate time series analysis are found in Chu and Durango-Cohen (2008), Issarayangyun and Greaves (2007), and Link et al. (2009).

The concept of the state of a system emerged from physical science. In that field, the state consists of a vector that includes all the information about the system that carries over into the future. That is, the state vector consists of a linearly independent set of linear combinations from the past that are correlated with future endogenous variables. To construct such a vector, the simplest and most straightforward way is through the one-step predictor, then the two-step predictor, and so forth, so long as the predictors extracted in this way are linearly independent. The set of predictors extracted in this way are, by construction, a state vector. They contain all the relevant information from the past since all predictors of the future, for any time horizon, are linearly dependent upon them. It is also a minimal set since they are linearly independent. The space spanned by the selected predictors is called the predictor space. The state-space model for an r -variate (multivariate) time series $\{\mathbf{Y}_t, t = 1, \dots, n\}$ is written as

$$\begin{aligned}\mathbf{Y}_t &= \mathbf{M}_t \mathbf{X}_t + \mathbf{D} \mathbf{Z}_t + \mathbf{u}_t \\ \mathbf{X}_t &= \Phi \mathbf{X}_{t-1} + \mathbf{w}_t\end{aligned}\tag{8.26}$$

These two equations are called the *observation equation* and the *state equation*, respectively, where \mathbf{X}_t , a p -vector, is the (observed) state of the process at time t . The evolution of \mathbf{X}_t is governed by the state equation: the state at time t evolves by autoregression on the previous state and is further modified by the innovation term \mathbf{w}_t . The sequence $[\mathbf{w}_t, t = 1, \dots, n]$ consists of independent and identically distributed random p -vectors. The unobserved (latent) data vector, \mathbf{Y}_t , is a possibly time-dependent linear transformation of \mathbf{X}_t with the addition of a regression on the k -vector \mathbf{Z}_t of known exogenous regressors and observational noise modeled by \mathbf{u}_t . The estimation of Equation (8.26) is done by maximum likelihood (Shumway and Stopher 2000).

8.5 Nonlinear Models

Linear time series approaches and particularly ARMA processes have largely dominated the field of time series modeling. However, in many cases both the theory and the data support a nonlinear model structure. There is ample evidence that many classical data sets such as weather (De Groot and Wurtz 1991), financial time series (Engel 1982) and transportation data (Dougherty 1995) arise from nonlinear processes. To a large extent, transportation systems represent highly nonlinear problems (Dougherty 1995). The probable existence of nonstandard phenomena such as nonnormality, asymmetric cycles, bimodality, nonlinear relationships between lagged variables, variation of prediction performance over the series, and sensitivity to initial conditions cannot be adequately modeled by the conventional linear ARMA processes (Tong 1990). A time series X_t is said to be linear if it can be written as follows (Tsay 2002):

$$X_t = \theta_0 + \sum_{i=1}^{\infty} \lambda_i \varepsilon_{t-i} \quad (8.27)$$

where θ_0 is a constant, λ_i are parameters to be estimated with $\lambda_0 = 1$, and the distribution of ε_t is continuous with $E(\varepsilon_t) = 0$ (frequently, the stronger assumption of normality is made for ε_t). Linear models have a number of advantages. First, they have a straightforward mathematical formulation and are easy to estimate. Second, they are based on the normality assumption and, as such, statistical inference is directly applicable in the form of confidence intervals and so on. And third, they have resulted in satisfactory models with "good" predictions for many years.

However, ARMA models are not adaptable and have difficulty in modeling time series with asymmetrically distributed data, outburst and tendency for taking on extreme values. The rationale for concentrating on nonlinear models for univariate time series is improved modeling and better description of the data. In these cases it is expected that a nonlinear model will describe the data better than a simple linear process (Tjøstheim 1994). The construction of a nonlinear model largely consists of modeling the mean and variance of the process given past information on X_t (Tjøstheim 1994), and is accomplished using either parametric or nonparametric approaches.

8.5.1 Testing for Nonlinearity

There are a number of tests available in the literature that can address nonlinearity in a model. Tests are based on the premise that, under linearity, residuals of a correctly specified model should be independent. Any violation of independence in the residuals indicates inadequacy of the developed model, including the linearity assumption (Tsay 2002). The Q-Statistic of squared residuals and the RESET test are two common tests for independence (readers interested in additional tests can refer to Tsay 2002 and Shumway and Stoffer 2000).

The Q -Statistic of squared residuals test is a modification of the Ljung–Box statistic, already presented in Section 7.2.3. It is modified to test the residuals of an ARMA(p,q) model and the test statistic is given as (McLeod and Li 1983)

$$Q(m) = T(T+2) \sum_{i=1}^m \frac{\hat{\rho}_i^2}{T-i} \quad (8.28)$$

where T is the sample size, m is the number of autocorrelations to be used in the test, and $\hat{\rho}_i^2$ is the lag- i ACF of the residuals. The statistic is asymptotically χ^2 distributed with $m-p-q$ degrees of freedom.

The RESET test, originally proposed by Ramsey (1969) and later modified by McLeod and Li (1983), Kennan (1985), and Tsay (1986), is straightforward to use with linear AR models. The test is based on running two AR(p) models

$$X_t = \sum_{t=1}^p k_t X_{t-1} + \varepsilon_t \quad (8.29)$$

$$\hat{\varepsilon}_t = \sum_{t=1}^p v_t X_{t-1} + \sum_{t=1}^p \tau_t \hat{X}_{t-1}^2 \quad (8.30)$$

For both models the sum of squared residuals are then calculated (denoted as SSR_0 for the first model (Equation 8.29) and SSR_1 for the second). The idea behind the test is that, if the first model is correctly specified, then parameters v_t, τ_t in the second model should be zero. This specification is assessed with an F -test as follows:

$$F \frac{(SSR_0 - SSR_1)/\delta}{SSR_1/(T-p-\delta)}, \quad \delta = \frac{p(p+1)}{2} \quad (8.31)$$

which, under linearity, is F distributed with δ and $T-p-\delta-1$ degrees of freedom.

8.5.2 Bilinear Models

A generalization of the usual autoregressive model of Equation 8.1 is a second-order expansion of the ARMA model (Equation 8.10), to provide better fit to the data (Chen and Tsay 1993b; Granger and Andersen 1978; Liu and Brockwell 1988). The bilinear model is written as

$$X_t = \theta_0 + \sum_{i=1}^p k_i X_{t-i} + \sum_{j=1}^q \lambda_j \varepsilon_{t-j} + \sum_{i=1}^r \sum_{j=1}^s v_{ij} X_{t-i} \varepsilon_{t-j} + \varepsilon_t \quad (8.32)$$

with $p, q, r, s \geq 0$.

8.5.3 Threshold Autoregressive Models

Threshold AR are based on the premise that AR(p) models are fit “locally” using the same process and experience from developing “global” AR(p) models. Local models are selected based on a partition of the $\mathbf{X}_{t-1} = (\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-p})$, vectors into J mutually exclusive and exhaustive regions, say R_1, \dots, R_J , where there is evidence that the time series characteristics change “significantly.” The idea is that by fitting separate models based on a threshold for X_t , the models can give an improved fit. The usual threshold AR(p) models (commonly referred to as self-exciting threshold autoregressive, SETAR) are written as

$$X_t = \theta_0^j + \sum_{i=1}^p k_i^j X_{t-i} + \varepsilon_t^j, \quad X_{t-1} \in R_j; \quad j = 1, \dots, J \quad (8.33)$$

where ε_t^j are white noise series with variance σ_j^2 for $j = 1, \dots, J$ (models of similar logic are developed with periodic parameters when the phenomena studies have significant seasonal differences). In traffic forecasting, researchers have identified different thresholds (speed levels for example), where different ARMA models should be developed with different parameters (see Ishak and Al-Deek 2002; Kamarianakis and Prastakos 2007). As an example, suppose researchers want to forecast speeds on urban freeways and notice that there is a marked change in traffic dynamics for speeds above 50 mi/h. The following model could be fit (as two separate ARMA models), assuming that $p = 2$ and X_t is a time series of speed measurements

$$\begin{aligned} X_t &= \theta_0^1 + \sum_{i=1}^2 k_i^1 X_{t-i} + \varepsilon_t^1, \quad X_{t-1} < 50 \text{ mi/h} \\ X_t &= \theta_0^2 + \sum_{i=1}^2 k_i^2 X_{t-i} + \varepsilon_t^2, \quad X_{t-1} \geq 50 \text{ mi/h} \end{aligned} \quad (8.34)$$

SETAR models are sometimes criticized as having a discontinuous conditional mean equation. To overcome this limitation, Terasvirta (1994) developed the smooth transition TAR (STAR(p)) models whose conditional mean function is differentiable. Although an improvement over the SETAR model, STAR models are complicated to interpret.

Another important and popular development in threshold AR is that by which thresholds do not depend on X_t but are given by a probability π_j with $\sum_{j=1}^J \pi_j = 1$ (Tong 1983, introduces the idea of probability “switching”). In practice this probability is frequently given by the exponential distribution and yields the exponential threshold autoregressive (ETRA) models. However, the idea of “switching” between models based on some probability

led Hamilton (1989) and McCulloch and Tsay (1993) to propose the Markov switching autoregressive (MSA) models.

The difference between the MSA and the previous models is that the transition is determined by a Markov chain. Consider the TAR Equation 8.33 in which switching is determined by a transition probability matrix from one threshold to another, rather than by specific values of X_t ; this finding implies that a threshold, say S_t (at time t) is given by a $J \times J$ transition probability matrix $P(S_t = j | S_{t-1} = i)$ for $i, j = 1, \dots, J$. A simple example is the following (Enders 2004):

$$\begin{aligned} X_t &= \theta_0^1 + k_1^1 X_{t-1} + \varepsilon_t^1 \\ X_t &= \theta_0^2 + k_1^2 X_{t-1} + \varepsilon_t^2 \end{aligned} \quad (8.35)$$

where the top model is used in the first “regime” and the bottom in the second. In general, in a two regime situation, the transition probability matrix contains elements p_{11}, p_{12}, p_{21} and p_{22} , where p_{11} is the probability that the system remains in regime 1 (while already in regime 1), and p_{12} is the probability that the system switches from regime 1 to regime 2, with $p_{12} = (1 - p_{11})$. The same applies to elements p_{21} and p_{22} and where the switching process is a first-order Markov process. Following Enders (2004), the probabilities p_{11} and p_{22} (and, obviously, $(1 - p_{11})$ and $(1 - p_{22})$) are conditional probabilities; that is, given that the system is in regime 1, $(1 - p_{11})$ is the probability that it switches into regime 2. The unconditional probability, that is the probability that the system is in a given regime is given as (for regime 1): $p_1 = (1 - p_{22}) / (2 - p_{11} - p_{22})$ (same applies for regime 2). Finally, transition probabilities are estimated along with the AR(p) models. In transportation applications, Malyshkina et al. (2009), Malyshkina and Mannering (2009), and Malyshkina and Mannering (2010) use the MSA principles in a negative binomial count model, multinomial discrete outcome, and zero-inflated count model context, respectively).

8.5.4 Functional Parameter Autoregressive Models

Chen and Tsay (1993a) proposed an adoptive and general formulation for the AR, termed the functional parameter autoregressive (FAR) models, which are written as

$$X_t = \sum_{i=1}^p f_i(Z_{t-1}) X_{t-i} + \varepsilon_t \quad (8.36)$$

where Z_{t-1} is a vector of lagged values of X_t but may also include other exogenous variables at time $t - 1$. To estimate FAR models nonparametric methods

such as Kernel regression, local linear regression, and so on, need to be used to obtain the functional parameters $f_i(\cdot)$.

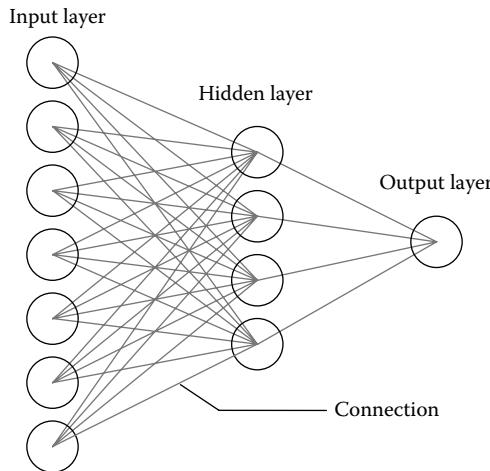
8.5.5 Neural Networks

Neural Networks are an extremely popular class of nonlinear statistical models that has been widely applied to all areas of transportation research (an extensive and up-to-date review of the application of neural networks in transportation research is found in Karlaftis and Vlahogianni 2009). The use of neural networks in transportation engineering emerged as a new alternative to forecasting complex, probably nonlinear, and nonstationary phenomena. Neural networks have met with widespread popularity, at least in part because they are generic, accurate, and convenient mathematical models able to emulate numerical model components easily.

Neural networks are considered to belong to the class of generalized nonlinear nonparametric models inspired by studies of the human brain. Their main advantage is that they can approximate any nonlinear function to an arbitrary degree of accuracy with a suitable Network architecture (Hornik et al. 1989). Neural networks have the inherent propensity for storing empirical knowledge and are used in any of three basic manners (Haykin 1999): (1) as models of biological nervous systems; (2) as real-time adaptive signal processors/controllers; and (3) as data-analytic methods. In transportation research, neural networks have almost exclusively been used as data-analytic methods due to their ability to work with massive amounts of data, modeling flexibility, learning and generalization ability, adaptability, and good predictive ability.

Neural network architecture, much like the human brain, is composed of simple processors called neurons or nodes, and numerous connections between them. A neural network is composed of many processing elements that are usually organized into a sequence of layers with full or partial connections between them (Figure 8.10). Usually, a neural network consists of an input layer, where data are presented to the network, and an output layer that links the response of the network to a given input. Frequently, to capture potential nonlinearity in the data other, intermediate layers called hidden layers, are included in the network architecture. The processing in the neurons is done by an “activation function” that controls the output of each node. Neural networks essentially “train” or “learn” through adaptation of their connection weights, as the hidden neurons organize themselves so that different neurons learn to recognize different features of the total input space.

Neural network training is performed iteratively until the average sum squared error between the computed and the desired output over all the training patterns is minimized. This kind of network derives its outcome from the manner that connection weights are adjusted to reduce output errors during the learning phase (Principe et al. 2000). Output errors are

**FIGURE 8.10**

The topology of a typical artificial neural network.

calculated by comparing the desired output with the actual output. The output is obtained from forward propagation of the input through the network. Next, output errors are propagated back to the hidden and input layers and connection weights in the network are modified to minimize a global error function. For back-propagation neural networks, the error function is usually the generalized delta rule (a variation of the least mean square theory), and a sigmoidal function is used as the activation function (Principe et al. 2000). Validation of the performance of a neural network is done using a separate set of testing data that broadly resembles the training data. Once the training and testing phases are found to be successful, the neural network model can then be put into practical applications. For a more detailed description of neural networks, the interested reader can refer to the classic textbooks by Bishop (1995), Ripley (1996), Haykin (1999) and Principe et al. (2000). A concise taxonomy of neural networks is found in Haykin (1999). The statistical perspective of neural networks both in terms of structure and of learning are extensively studied by Cheng and Titterington (1994), Kuan and White (1994), Ripley (1993, 1994), and Sarle (1994).

Although the goal of both statistical modeling and neural networks is the same, the two have kept each other at arm's length and researchers frequently fail to communicate and even understand each other's work. Frequently, researchers proficient in one of the two approaches argue fervently in support of their chosen method, making the selection of analysis approach one of the hotly debated topics in the literature. At least part of this methodological divide is due to the terminology used by each discipline,

TABLE 8.6

Equivalence of Statistical and Neural Network Terminology

Statistics	Neural Networks
• Independent/predicted variable	• Input/output
• Dependent variables	• Targets or training values
• Residuals	• Errors
• Estimation	• Training, learning, adaptation, or self-organization
• Estimation criterion	• Error function, cost function, Lyapunov function
• Observations	• Patterns or training pairs
• Parameter estimates	• (Synaptic) weights
• Interactions	• Higher order neurons
• Transformations	• Functional links
• Regression and discriminant analysis	• Supervised learning or heteroassociation
• Data reduction	• Unsupervised learning, encoding or autoassociation
• Cluster analysis	• Competitive learning or adaptive vector quantization

Adapted from Sarle 1994 and Karlaftis and Vlahogianni 2010.

TABLE 8.7

Equivalence of Statistical and Neural Network Methodologies

Neural Networks	Statistical Models
Feed-forward neural network with no hidden layer	Generalized linear models • (Multivariate) Linear regression • Logistic regression • Linear discriminant function
• Simple linear perceptron	
• Simple nonlinear perceptron	
• adaline	
Feed-forward neural network with one hidden layer	Projection pursuit regression.
Generalized regression neural network	Kernel regression
Probabilistic neural network	kernel discriminant analysis
Competitive learning networks	k-means clustering
Kohonen self-organizing maps	Discrete approximations to principal curves and surfaces
Hybrid networks (supervised and unsupervised learning)	Principal components regression
Learning vector quantization	Variation of nearest neighbor discriminant analysis
Hebbian learning	Principal component analysis

Adapted from Sarle 1994 and Karlaftis and Vlahogianni 2010.

as terms utilized in neural network modeling are entirely different from those found in statistics. Sarle (1994) made an effort to codify statistical terminology in neural network terms (Table 8.6). This ambiguity in terminology has led to considerable confusion regarding the direct correspondence between the two approaches (see Table 8.7 for some of the modeling analogies).

9

Latent Variable Models

This chapter presents tools for illuminating structure in data in the presence of measurement difficulties, endogeneity, and unobservable or latent variables. Structure in data refers to relationships between variables in the data, including direct or causal relationships, indirect or mediated relationships, associations, and the role of errors of measurement in the models. Measurement difficulties include challenges encountered when an unobservable or latent construct is indirectly measured through exogenous observable variables. Examples of latent variables in transportation include attitudes toward transportation policies or programs such as gas taxes, van-pooling, high-occupancy vehicle (HOV) lanes, and mass transit. Interest might also be centered on the role of education, socioeconomic status, or attitudes on driver decision making—which are also difficult to measure directly. As discussed in previous chapters, endogeneity refers to variables whose values are largely determined by factors or variables included in the statistical model or system.

In many analyses, the initial steps attempt to uncover structure in data that can then be used to formulate and specify statistical models. These situations arise predominately in observational settings—when the analyst does not have control over many of the measured variables, or when the study is exploratory and there are not well-articulated theories regarding the structure in the data. There are several approaches to uncovering data structure. Principal components analysis is widely used as an exploratory method for revealing structure in data. Factor analysis, a close relative of principal components analysis, is a statistical approach for examining the underlying structure in multivariate data. And, structural equation models (SEMs) refer to a formal modeling framework developed specifically for dealing with unobservable or latent variables, endogeneity among variables, and complex underlying data structures encountered in social phenomena often entwined in transportation applications.

9.1 Principal Components Analysis

Principal components analysis has two primary objectives: to reduce a relatively large multivariate data set, and to interpret data (Johnson and

Wichern 1992). These objectives are accomplished by explaining the variance–covariance structure using a few linear combinations of the originally measured variables. Through this process a more parsimonious description of the data is provided—reducing or explaining the variance of many variables with fewer well chosen combinations of variables. If, for example, a large proportion (70%–90%) of the total population variance is attributed to a few uncorrelated principal components, then these components can replace the original variables without much loss of information, and also describe different dimensions in the data. Principal components analysis relies on the correlation matrix of variables—so the method is suitable for variables measured on the interval and ratio scales.

If the original variables are uncorrelated, then principal components analysis accomplishes nothing. Thus, for example, experimental data with randomized treatments is not well suited to principal components analysis. However, observational data, which typically contains a large number of correlated variables, is ideally suited for principal components analysis. If it is found that the variance in 20 or 30 original variables is described adequately with four or five principal components (dimensions), then principal components analysis will have succeeded.

Principal components analysis begins by noting that n observations, each with P variables or measurements upon them, is expressed in an $n \times P$ matrix \mathbf{X} :

$$\mathbf{X}_{n \times P} = \begin{bmatrix} x_{11} & \dots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nP} \end{bmatrix} \quad (9.1)$$

Principal components analysis is not a statistical model, and there is no distinction between dependent and independent variables. If the principal components analysis is useful, there are $K < n$ principal components, with the first principal component

$$Z_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1P}x_p \quad (9.2)$$

that maximizes the variability across individuals, subject to the constraint

$$a_{11}^2 + a_{12}^2 + \dots + a_{1P}^2 = 1 \quad (9.3)$$

Thus, $VAR[Z_1]$ is maximized given the constraint in Equation 9.3, with the constraint imposed simply because the solution is indeterminate otherwise because one could simply increase one or more of the a_{ij} values to increase the variance. A second principal component Z_2 is then sought

that maximizes the variability across individuals subject to the constraints that $a_{21}^2 + a_{22}^2 + \dots + a_{2p}^2 = 1$ and $COR[Z_1, Z_2] = 0$. In keeping, a third principal component is added subject to the same constraint on the a_{ij} values, with the additional constraint that $COR[Z_1, Z_2, Z_3] = 0$. Additional principal components are added up to P , the number of variables in the original data set.

The eigenvalues of the sample variance–covariance matrix \mathbf{X} are the variances of the principal components. The corresponding eigenvector provides the coefficients to satisfy Equation 9.3.

Recall from Appendix E that the symmetric $P \times P$ sample variance–covariance matrix is given as

$$S^2[\mathbf{X}] = \begin{bmatrix} s^2(x_1) & s(x_1, x_2) & \dots & s(x_1, x_p) \\ s(x_2, x_1) & s^2(x_2) & \dots & s(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ s(x_p, x_1) & s(x_p, x_2) & \dots & s^2(x_p) \end{bmatrix} \quad (9.4)$$

The diagonal elements of this matrix represent the estimated variances of random variables 1 through P , while the off-diagonal elements represent the estimated covariances between variables.

The sum of the eigenvalues λ_p of the sample variance–covariance matrix is equal to the sum of the diagonal elements in $S^2[\mathbf{X}]$, or the sum of the variances of the P variables in matrix \mathbf{X} , that is

$$\lambda_1 + \lambda_2 + \dots + \lambda_p = VAR(x_1) + VAR(x_2) + \dots + VAR(x_p) \quad (9.5)$$

Because the sum of the diagonal elements represents the total sample variance, and the sum of the eigenvalues is equal to the trace of $S^2[\mathbf{X}]$, then the variance in the principal components accounts for all of the variation in the original data. There are P eigenvalues, and the proportion of total variance explained by the j^{th} principal component is given by

$$VAR_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, \quad j = 1, 2, \dots, P \quad (9.6)$$

To avoid excessive influence of measurement units, the principal components analysis is carried out on a standardized variance–covariance matrix, or the correlation matrix. The correlation matrix is simply the variance–covariance matrix as obtained by using the standardized variables instead of the original variables, such that $Z_{ij} = X_{ij} - \bar{X}_j / \sigma_j$; $i = 1, 2, \dots, n$; $j = 1, 2, \dots, P$ replace the original X_{ij} 's. Because the correlation matrix is often used, variables used in principal components analysis are restricted to interval and

ratio scales unless corrections are made. Using the correlation matrix, the sum of the diagonal terms, and the sum of eigenvalues, is equal to P (the number of variables).

With the basic statistical mechanics in place, the basic steps in principal components analysis are (see Manly 1986) as follows: standardize all observed variables in the X matrix; calculate the variance–covariance matrix, which is the correlation matrix after standardization; determine the eigenvalues and corresponding eigenvectors of the correlation matrix (the coefficients of the i^{th} principal component are given by the eigenvector, while the variance is given by the eigenvalue); and discard any components that account for a relatively small proportion of the variation in the data.

The intent of principal components analysis is to illuminate underlying commonality, or structure in the data. In exploratory studies relying on observational data, it is often the case that some variables measure the same underlying construct, and this underlying construct is what is being sought in principal components analysis. The goal is to reduce the number of potential explanatory variables and gain insight as to what underlying dimensions have been captured in the data.

Example 9.1

A survey of 281 commuters was conducted in the Seattle metropolitan area. The survey's intent was to gather information on commuters' opinions or attitudes regarding HOV lanes (lanes that are restricted for use by vehicles with two or more occupants). Commuters were asked a series of attitudinal questions regarding HOV lanes and their use, in addition to a number of sociodemographic and commuting behavior questions. A description of the variables is provided in Table 9.1.

It is useful to assess whether a large number of variables, including attitudinal, sociodemographic, and travel behavior, is distilled into a smaller set of variables. To do this assessment, a principal components analysis is performed on the correlation matrix (not the variance–covariance matrix). Figure 9.1 shows a graph of the first 10 principal components. The graph shows that the first principal component represents about 19% of the total variance, the second principal component an additional 10%, the third principal component about 8%, the fourth about 7%, and the remaining principal components about 5% each. Ten principal components account for about 74% of the variance, and six principal components account for about 55% of the variance contained in the 23 variables that were used in the principal components analysis. Thus, there is some evidence that some variables, at least, are explaining similar dimensions of the underlying phenomenon.

Table 9.2 shows the variable coefficients for the 6 principal components. For example, the first principal component is given by

$$\begin{aligned} Z_1 = & -0.380(HOV\text{Pst5}) + 0.396(DAP\text{st5}) - 0.303(CrPP\text{st5}) - 0.109(CrPP\text{st52Mr}) \\ & - 0.161(Bus\text{Pst5}) - 0.325(HOV\text{SavTime}) + 0.364(HOVO\text{pnn}) \\ & - 0.339(GTToHOV2) + 0.117(Gend) \end{aligned}$$

TABLE 9.1

Variables Collected During HOV Lane Survey, Seattle, WA

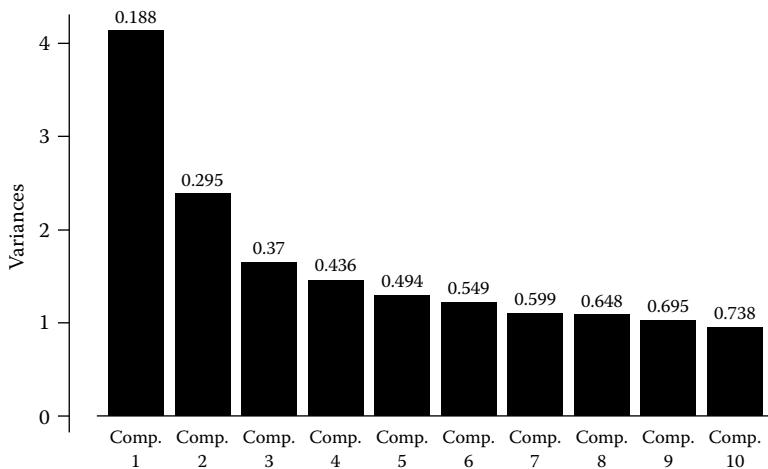
Variable No.	Abbreviation: Variable Description
1	<i>Mode</i> : Usual mode of travel: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if vanpool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
2	<i>HOVUse</i> : Have used HOV lanes: 1 if yes, 0 if no
3	<i>HOVMode</i> : If used HOV lanes, what mode is most often used: 0 in a bus, 1 in two person carpool, 2 in three or more person carpool, 3 in vanpool, 4 alone in vehicle, 5 on motorcycle
4	<i>HOVDecline</i> : Sometimes eligible for HOV lane use but do not use: 1 if yes, 0 if no
5	<i>HOVDecReas</i> : Reason for not using HOV lanes when eligible: 0 if slower than regular lanes, 1 if too much trouble to change lanes, 2 if HOV lanes are not safe, 3 if traffic moves fast enough, 4 if forget to use HOV lanes, 5 if other
6	<i>Mode1Yr</i> : Usual mode of travel one year ago: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if vanpool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
7	<i>Com1Yr</i> : Commuted to work in Seattle a year ago: 1 if yes, 0 if no
8	<i>FlexStar</i> : Have flexible work start times: 1 if yes, 0 if no
9	<i>ChngDepTm</i> : Changed departure times to work in the last year: 1 if yes, 0 if no
10	<i>MinErlyWrk</i> : On average, number of minutes leaving earlier for work relative to last year
11	<i>MinLtWrk</i> : On average, number of minutes leaving later for work relative to last year
12	<i>DepChngReas</i> : If changed departure times to work in the last year, reason why: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
13	<i>ChngRte</i> : Changed route to work in the last year: 1 if yes, 0 if no
14	<i>ChngRteReas</i> : If changed route to work in the last year, reason why: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
15	<i>I90Cm</i> : Usually commute to or from work on Interstate 90: 1 if yes, 0 if no
16	<i>I90Cmt1Yr</i> : Usually commuted to or from work on Interstate 90 last year: 1 if yes, 0 if no
17	<i>HOVPst5</i> : On your past five commutes to work, how often have you used HOV lanes
18	<i>DAPst5</i> : On your past five commutes to work, how often did you drive alone
19	<i>CrPPst5</i> : On your past five commutes to work, how often did you carpool with one other person

continued

TABLE 9.1 (continued)

Variables Collected During HOV Lane Survey, Seattle, WA

Variable No.	Abbreviation: Variable Description
20	<i>CrPPst52Mr</i> : On your past five commutes to work, how often did you carpool with two or more people
21	<i>VnPPst5</i> : On your past five commutes to work, how often did you take a vanpool
22	<i>BusPst5</i> : On your past five commutes to work, how often did you take a bus
23	<i>NonMotPst5</i> : On your past five commutes to work, how often did you bicycle or walk
24	<i>MotPst5</i> : On your past five commutes to work, how often did you take a motorcycle
25	<i>OthPst5</i> : On your past five commutes to work, how often did you take a mode other than those listed in variables 18 through 24
26	<i>ChgRtePst5</i> : On your past five commutes to work, how often have you changed route or departure time
27	<i>HOVSavTime</i> : HOV lanes save all commuters time: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
28	<i>HOVAdUse</i> : Existing HOV lanes are being adequately used: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
29	<i>HOVOpn</i> : HOV lanes should be open to all traffic: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
30	<i>GPToHOV</i> : Converting some regular lanes to HOV lanes is a good idea: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
31	<i>GTTtoHOV2</i> : Converting some regular lanes to HOV lanes is a good idea only if it is done before traffic congestion becomes serious: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
32	<i>Gend</i> : Gender: 1 if male, 0 if female
33	<i>Age</i> : Age in years: 0 if under 21, 1 if 22–30, 2 if 31–40, 3 if 41–50, 4 if 51–64, 5 if 65 or greater
34	<i>HHIncm</i> : Annual household income (U.S. dollars per year): 0 if no income, 1 if 1–9,999, 2 if 10,000–19,999, 3 if 20,000–29,999, 4 if 30,000–39,999, 5 if 40,000–49,999, 6 if 50,000–74,999, 7 if 75,000–100,000, 8 if over 100,000
35	<i>Educ</i> : Highest level of education: 0 if did not finish high school, 1 if high school, 2 if community college or trade school, 3 if college/university, 4 if postcollege graduate degree
36	<i>FamSiz</i> : Number of household members
37	<i>NumAdlt</i> : Number of adults in household (aged 16 or more)
38	<i>NumWrks</i> : Number of household members working outside the home
39	<i>NumCars</i> : Number of licensed motor vehicles in the household
40	<i>ZipWrk</i> : Postal zip code of work place
41	<i>ZipHm</i> : Postal zip code of home
42	<i>HOVCmnt</i> : Type of survey comment left by respondent regarding opinions on HOV lanes: 0 if no comment on HOV lanes, 1 if comment not in favor of HOV lanes, 2 comment positive toward HOV lanes but critical of HOV lane policies, 3 comment positive toward HOV lanes, 4 neutral HOV lane comment

**FIGURE 9.1**

Cumulative proportion of variance explained by 10 principal components: HOV lane survey data.

All of the variables had estimated coefficients (or loadings). However, coefficients less than 0.1 were omitted from Table 9.2 due to their relatively small magnitude. The first principal component loaded strongly on travel behavior variables and HOV attitude variables. In addition, Z_1 increases with decreases in any non-drive alone travel variables (HOV, Car Pool, Bus), increases with decreases in pro-HOV attitudes, and increases for males. By analyzing the principal components in this way, some of the relationships between variables is better understood.

9.2 Factor Analysis

Factor analysis is a close relative of principal components analysis. It was developed early in the twentieth century by Karl Pearson and Charles Spearman with the intent to gain insight into psychometric measurements, specifically the directly unobservable variable intelligence (Johnson and Wichern 1992). The aim of the analysis is to reduce the number of P variables to a smaller set of parsimonious $K < P$ variables. The objective is to describe the covariance among many variables in terms of a few unobservable factors. There is one important difference, however, between principal components and factor analysis. Factor analysis is based on a specific statistical model, whereas principal components analysis is not. As was the case with principal components analysis, factor analysis relies on the correlation matrix, and so factor analysis is suitable for variables measured on interval and ratio scales.

TABLE 9.2
Factor Loadings of Principal Components Analysis: HOV Lane Survey Data. Loadings < 0.10 Shown as Blanks

Variable	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
<i>Travel Behavior Variables</i>						
HOVPst5	-0.380		-0.284	0.236		
DAPst5	0.396	0.274	-0.283		0.128	
CrPPst5	-0.303	-0.223	0.240	0.282	0.221	
CrPPst52Mr	-0.109		0.167	0.196	-0.107	
VnPPst5		-0.146				
BusPst5	-0.161	-0.227	0.112	-0.514		-0.395
NonMotPst5					0.471	
MotPst5		0.104		0.381	-0.418	
ChgRtePst5			0.525		-0.302	
<i>HOV Attitude Variables</i>						
HOVSavTime	-0.325		0.301	-0.140		
HOVAdUse	-0.321	0.227		-0.133		
HOVOpn	0.364	-0.216	0.210			
GPToHOV	-0.339	0.230	-0.115			
GTToHOV2	-0.260	0.245	-0.153			
<i>Sociodemographic Variables</i>						
Gend	0.117	0.388	0.180		-0.199	
Age		0.268	0.341	-0.363	-0.270	
HHIncm	0.304	0.131	0.489		0.101	
Educ	0.188	0.247	0.443		0.247	
FamSiz	0.429	-0.122				
NumAdlt	0.516	-0.188	-0.128	-0.133		
NumWkhs	0.451	-0.242	-0.137			
NumCars	0.372	-0.106		0.107	-0.268	

Just as for other statistical models, there should be a theoretical rationale for conducting a factor analysis (Pedhazur and Pedhazur Schmelkin 1991). One should not simply “feed” all variables into a factor analysis with the intention to uncover real dimensions in the data. There should be a theoretically motivated reason to suspect that some variables are measuring the same underlying phenomenon, with a subsequent examination of whether the data support this expected underlying measurement model or process.

The factor analysis model is formulated by expressing the X_i 's as linear functions, such that,

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \dots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \dots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots && \ddots && \vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \dots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \quad (9.7)$$

where, in matrix notation the factor analysis model is given as

$$(\mathbf{X} - \boldsymbol{\mu})_{p \times 1} = \mathbf{L}_{p \times m} \mathbf{F}_{m \times 1} + \boldsymbol{\varepsilon}_{p \times 1} \quad (9.8)$$

where F 's are factors, ℓ_{ij} 's are the factor loadings. The ε 's are associated only with the X_i 's, and the p random errors and m factor loadings are unobservable or latent. With p equations and $p + m$ unknowns, the unknowns cannot be directly solved without additional information. To solve for the unknown factor loadings and errors, restrictions are imposed. The types of restrictions determine the type of factor analysis model. The factor rotation method used determines the type of factor analysis model, orthogonal or oblique. Factor loadings that are either close to one or close to zero are sought. A factor loading close to one suggests that a variable X_i is largely influenced by F_j . In contrast, a factor loading close to zero suggests that a variable X_i is not substantively influenced by F_j . A collection of factor loadings that is as diverse as possible is sought, lending itself to easy interpretation.

The orthogonal factor analysis model satisfies the following conditions:

$\mathbf{F}, \boldsymbol{\varepsilon}$ are independent

$$E[\mathbf{F}] = \mathbf{0}$$

$$COV[\mathbf{F}] = \mathbf{I} \quad (9.9)$$

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

$$COV[\boldsymbol{\varepsilon}] = \boldsymbol{\psi}, \text{ where } \boldsymbol{\psi} \text{ is a diagonal matrix}$$

Varimax rotation, which maximizes the sum of the variances of the factor loadings, is a common method for conducting an orthogonal rotation, although there are many other methods.

The oblique factor analysis model relaxes the restriction of uncorrelated factor loadings, resulting in factors that are nonorthogonal. Oblique factor analysis is conducted with the intent to achieve more interpretable structure. Specifically, computational strategies have been developed to rotate factors so as to best represent clusters of variables, without the constraint of orthogonality. However, the oblique factors produced by such rotations are often not easily interpreted, sometimes resulting in factors with less-than-obvious meaning (i.e., with many cross-loadings).

Interpretation of factor analysis is straightforward. Variables that have high factor loadings are thought to be highly influential in describing the factor, whereas variables with low factor loadings are less influential in describing the factor. Inspection of the variables with high factor loadings on a specific factor is used to uncover structure or commonality among the variables. One must then determine the underlying constructs that are common to variables that load highly on specific factors.

Example 9.2

Continuing from the previous example, a factor analysis on continuous variables is conducted to determine which variables might be explaining similar underlying phenomenon. The same set of variables used for the principal components analysis is again used in an exploratory factor analysis with orthogonal varimax rotation. Because six principal components explained roughly 55% of the data variance, the number of factors estimated was six.

Table 9.3 shows the factor loadings resulting from this factor analysis. As was done in the principal components analysis, factor loadings less than 0.1 are blanks in the table. Table 9.3, in general, is more sparse than Table 9.2, simply because factors are orthogonal to one another (correlations between factors are zero), and because a factor set solution was sought that maximizes the variances of the factor loadings. Inspection of Table 9.3 shows that the equation for X_1 is given as $HOVPst5 = 0.306F_1 + 0.658F_2 + 0.323F_3 + 0.147F_4 + 0.147F_5$, where factors one through five are unobserved or latent. As in the principal components analysis, travel behavior variables and HOV attitudinal variables seem to load heavily on factor 1. That many of the factors include travel behavior variables suggest that many dimensions of travel behavior exist. There appear to be two factors that reflect dimensions in the data related to attitudes toward HOV lanes. Sociodemographic variables do not seem to load on any particular factor, and so probably represent unique dimensions in the data.

9.3 Structural Equation Modeling

Structural equation models represent a natural extension of a measurement model, and represents a mature statistical modeling framework. The SEM is

TABLE 9.3

Factor Loadings of Factor Analysis: HOV Lane Survey Data. Loadings < 0.10 Shown as Blanks

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
<i>Travel Behavior Variables</i>						
<i>HOVPst5</i>	0.306	0.658	0.323	0.147		0.147
<i>DAPst5</i>	-0.278	-0.846	-0.358	-0.206		-0.167
<i>CrPPst5</i>	0.221	0.930	-0.217	-0.166	-0.109	
<i>CrPPst52Mr</i>	0.112			0.987		
<i>VnPPst5</i>						0.675
<i>BusPst5</i>	0.113	0.108	0.983			
<i>NonMotPst5</i>						0.298
<i>MotPst5</i>					0.992	
<i>ChgRtePst5</i>	-0.125		-0.113		0.158	
<i>HOV Attitude Variables</i>						
<i>HOVSavTime</i>	0.734	0.142				
<i>HOVAdUse</i>	0.642	0.110				0.135
<i>HOVOpen</i>	-0.814	-0.117				
<i>GPToHOV</i>	0.681	0.156				
<i>GTToHOV2</i>	0.515					
<i>Sociodemographic Variables</i>						
<i>Age</i>		-0.128				
<i>HHIncm</i>			-0.167			
<i>Educ</i>						
<i>FamSiz</i>						-0.129
<i>NumWrks</i>	0.105					
<i>NumCars</i>						

a tool developed largely by clinical sociologists and psychologists. It is designed to deal with several difficult modeling challenges, including cases in which some variables of interest to a researcher are unobservable or latent and are measured using one or more exogenous variables, endogeneity among variables, and complex underlying social phenomena. For instance, in a study of transit ridership, one might want to determine the impact of a latent variable such as “attitude towards transit” on transit ridership. Because the attitude toward transit itself is not directly observable, one might ask a question or series of questions with the intent to indirectly measure this variable. These observed variables—answers to attitudinal questions—are poorly measured variables meant to reflect the latent variable attitude toward transit. In addition, experiences riding transit might influence attitudes toward transit, and attitudes toward transit might influence their experiences—an endogenous relationship where each variable could influence the other.

When measurement errors in independent variables are incorporated into a regression equation (via a poorly measured variable), the variances of the measurement errors in the regressors are transmitted to the model error, thereby inflating the model error variance. The estimated model variance is thus larger than if no measurement errors are present. This outcome would have deleterious effects on standard errors of coefficient estimates, and goodness-of-fit (GOF) criteria, including the standard F -ratio and R -squared measures. If parameters are estimated using ordinary least squares then parameter estimates are biased and are a function of the measurement error variances. The SEM framework resolves potential problems by explicitly incorporating measurement errors into the modeling framework. In addition, the SEM model can accommodate a latent variable as a dependent variable, something that cannot be done in the traditional regression analysis.

9.3.1 Basic Concepts in Structural Equation Modeling

SEM's have two components, a measurement model and a structural model. The measurement model is concerned with how well various measured exogenous variables measure latent variables. A classical factor analysis is a measurement model, and determines how well various variables load on a number of factors or latent variables. The measurement models within a SEM incorporate estimates of measurement errors of exogenous variables and their intended latent variable. The structural model is concerned with how the model variables are related to one another. SEMs allow for direct, indirect, and associative relationships to be explicitly modeled, unlike ordinary regression techniques which implicitly model associations. It is the structural component of SEMs that enables substantive conclusions to be made about the relationship between latent variables, and the mechanisms underlying a process or phenomenon. The structural component of SEMs is similar to a system of simultaneous equations discussed previously in Chapter 5. Because of the ability of SEMs to specify complex underlying relationships, SEMs lend themselves to graphical representations and these graphical representations have become the standard means for presenting and communicating information about SEMs.

Like factor and principal components analyses, SEMs rely on information contained in the variance–covariance matrix. Similar to other statistical models, the SEM requires the specification of relationships between observed and unobserved variables. Observed variables are measured, whereas unobserved variables are latent variables—similar to factors in a factor analysis—which represent underlying unobserved constructs. Unobserved variables also include error terms that reflect the portion of the latent variable not explained by their observed counterparts. In a SEM, there is a risk that the number of model parameters sought will exceed the number of model equations needed to solve them. Thus, there is a need to distinguish between

fixed and free parameters—fixed parameters being set by the analyst and free parameters being estimated from the data. The collection of fixed and free parameters specified by the analyst will imply a variance–covariance structure in the data, which is compared to the observed variance–covariance matrix to assess model fit.

There are three types of relationships that are modeled in the SEM. An association is a casual (not causal) relationship between two independent variables, and is depicted as a double headed arrow between variables. A direct relationship is where the independent variable influences the dependent variable, and is shown with a directional arrow, where the direction of the arrow is assumed to coincide with the direction of influence from the exogenous to the endogenous variable. An indirect relationship is when an independent variable influences a dependent variable indirectly through a third independent variable. For example, variable A has a direct effect on variable B , which has a direct effect on variable C : so variable A has an indirect effect on variable C . Note that in this framework a variable may serve as both an endogenous variable in one relationship, and an exogenous variable in another.

Figure 9.2 shows a graphical representation of two different linear regression models with two independent variables, as is often depicted in the SEM nomenclature. The independent variables X_1 and X_2 , shown in rectangles, are measured exogenous variables, have direct effects on variable Y_1 , and are correlated with each other. The model in the bottom of the figure reflects a fundamentally different relationship among variables. First, variables X_3 and X_4 directly influence Y_2 . Variable X_4 is also directly influenced by variable X_3 . The SEM model shown in the top of the figure implies a different variance–covariance matrix than the model shown in the bottom of the figure. The models also show that although the independent variables have direct affects on the dependent variable, they do not fully explain the variability in Y , as reflected by the error terms, depicted as ellipses in the figure. The additional

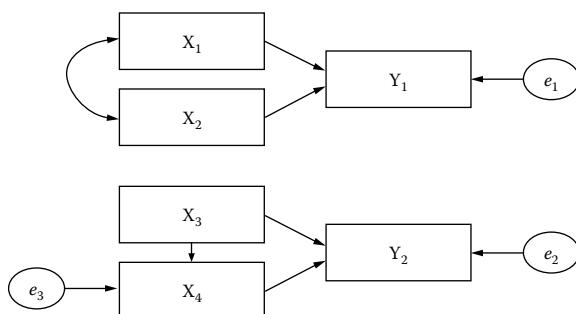


FIGURE 9.2

SEMs depicting standard linear regression model with two variables: top model depicts no correlation between independent variables; bottom model depicts model with correlation between independent variables.

error term, error 3, is that portion of variable X_4 not fully explained by variable X_3 . Latent variables, if entered into these models, would also be depicted as ellipses in the graphical representation of the SEM.

An obvious issue of concern is how these two different SEMs depicted in Figure 9.2 imply different variance–covariance matrices. The model depicted in the top of Figure 9.2 represents a linear regression model with two independent variables that covary, such that $Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error}_1$. The model depicted in the bottom of the figure represents two simultaneous regressions, $Y_2 = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \text{error}_2$ and $X_4 = \beta_0 + \beta_5 X_3 + \text{error}_3$. In this second SEM model, the variable X_4 serves as both an exogenous and an endogenous variable. The collective set of constraints implied by these two SEMs determines the model implied variance–covariance structure. The original correlation matrix is completely reproduced if all effects, direct, indirect, and correlated, are accounted for in a model. This saturated model is uninteresting simply because there is no parsimony achieved by such a model. Without compromising the statistical validity of the model, a natural goal is to simplify an underlying complex data generating process with a relatively simple model. How the path is drawn in the development of SEMs determines the presumed variance–covariance matrix.

Example 9.3

To examine simple path models, consider the relationships between an attitudinal variable in the HOV survey and two behavioral variables. Table 9.4 shows the results of two SEMs. These overly simple models illustrate the two SEM models depicted in Figure 9.2, with the exception that no covariation between independent variables is allowed in Model I. Model I is a linear regression model with two variables, *CRPPst5* and *HOVPst5*. The second SEM consists of two simultaneous regressions, as shown in the figure. The constraint of zero covariance between *CRPPst5* and *HOVPst5* in Model I does not agree well with the observed covariance of 1.51 between these two variables. These two competing models imply different variance–covariance matrices, and the agreement of these two model-implied covariance matrices with the observed variance–covariance matrix is used to assess model fit.

Example 9.3 illustrates how SEMs imply sets of regression equations, which in turn have implications on the variance–covariance matrix. It is the discrepancy between the observed and the implied variance–covariance matrix that forms the basis of SEMs. When latent variables are brought into the SEM framework the true intention of SEMs is realized. The SEM is able to cope with both endogenous and exogenous variables, as well as observed and unobserved variables, expressed as embedded linear relationships reflecting complex underlying relationships. To see a couple of detailed examples of how complex SEMs are estimated and how they have been applied in the construction and transit industries, see Molenaar et al. (2000 and 2009) and Karlaftis et al. (2002), respectively.

TABLE 9.4

Implied Variance–Covariance Matrices for Two Simple SEM's: HOV Survey Data

Sample Variance–Covariance Matrix	HOVPst5	CRPPst5	HOVSavTi
HOVPst5	3.29		
CRPPst5	1.51	2.32	
HOVSavTi	0.74	0.67	1.97
<i>Implied Variance–Covariance Matrix: SEM Model I</i>			
$\text{HOVSavTi} = 1.52 + 0.201(\text{CRPPst5}) + 0.123(\text{HOVPst5})$			
HOVPst5	3.28		
CRPPst5	0.00	2.30	
HOVSavTi	0.43	0.46	1.89
<i>Implied Variance–Covariance Matrix: SEM Model II</i>			
$\text{HOVSavTi} = 1.58 + 0.224(\text{HOVPst5})$			
$\text{HOVPst5} = 0.55 + 0.650(\text{CRPPst5})$			
HOVPst5	2.292		
CRPPst5	1.490	3.281	
HOVSavTi	0.333	0.734	1.959

9.3.2 Fundamentals of Structural Equation Modeling

The focus here is to provide a general framework of SEMs, to demonstrate how the parameters are estimated, and to illustrate how results are interpreted and used. The interested reader can consult Hoyle (1995) and Arminger et al. (1995) for a presentation and discussion of the full gamut of alternative SEM specifications.

Structural equation models, similar to other statistical models, are used to evaluate theories or hypotheses using empirical data. The empirical data are contained in a $P \times P$ variance–covariance matrix S , which is an unstructured estimator of the population variance–covariance matrix Σ . A SEM is then hypothesized to be a function of Q unknown structural parameters (in parameter vector θ), which in turn will generate a model-implied variance–covariance matrix $\Sigma(\theta)$. All variables in the model, whether observed or latent, are classified as either independent (endogenous) or dependent (exogenous). A dependent variable in a SEM diagram is a variable that has a one-way arrow pointing to it. The set of these variables is collected into a vector η , while independent variables are collected in the vector ξ , such that (following Bentler and Weeks 1980)

$$\eta = \beta\eta + \gamma\xi + \varepsilon \quad (9.10)$$

where β and γ are estimated vectors of coefficients that contain regression coefficients for the dependent and independent variables, respectively,

and $\boldsymbol{\epsilon}$ is a vector of regression errors. The exogenous factor covariance matrix is represented as $\boldsymbol{\Phi} = COV[\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^T]$, and the error covariance matrix as $\boldsymbol{\psi} = COV[\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^T]$.

The variance–covariance matrix for the model in Equation 9.10 is

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{G}(\mathbf{I} - \boldsymbol{\beta})^{-1} \boldsymbol{\gamma} \boldsymbol{\Phi} \boldsymbol{\gamma}^T (\mathbf{I} - \boldsymbol{\beta})^{-1} \mathbf{G}^T \quad (9.11)$$

where \mathbf{G} is a selection matrix containing either 0 or 1 to select the observed variables from all the dependent variables in $\boldsymbol{\eta}$. There are P^2 elements or simultaneous equations in Equation 9.11, one for each element in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Some of the P^2 equations are redundant, however, leaving $P^* = P(P-1)/2$ independent equations. These P^* independent equations are used to solve for unknown parameters $\boldsymbol{\theta}$, which consist of the vector $\boldsymbol{\beta}$, the vector $\boldsymbol{\gamma}$, and $\boldsymbol{\Phi}$. The estimated model-implied variance–covariance matrix is then given as $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$.

Model identification in SEM can present serious challenges. There are Q unknown model parameters (comprising $\boldsymbol{\theta}$), which must be solved using P^* simultaneous independent equations. There are two necessary and sufficient conditions for SEM identification. The first is that the number of simultaneous equations must be equal to or greater than the number of unknown model parameters, such that $Q \leq P^*$. The second is that each and every free model parameter must be identified, which often is difficult (Hoyle 1995).

Once the SEM has been specified, and identification conditions are met, solutions for the parameters are obtained. Parameters are estimated using a discrepancy function criterion, where the differences between the sample variance–covariance matrix and the model-implied variance–covariance matrix are minimized. The discrepancy function is

$$F = F(\mathbf{S}, \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})) \quad (9.12)$$

Different estimation methods in SEM have varying distributional assumptions, and in turn require different discrepancy functions. For example, maximum likelihood (MLE) estimated model parameters, which requires that specific distributional and variable assumptions are met, are obtained using the discrepancy function

$$F_{MLE} = LN |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + TRACE[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{S}] - LN |\mathbf{S}| - p \quad (9.13)$$

For detailed discussions on other discrepancy functions and corresponding estimation methods, including MLE, generalized least squares (GLS), asymptotically distribution-free (ADF), scale-free least squares (SLS), unweighted least squares (ULS), and Browne's method, see Arbuckle and Wotheke (1995), Hoyle (1995), or Arminger et al. (1995).

A useful feature of discrepancy functions is that they are useful for testing the null hypothesis that $H_0: \boldsymbol{\Sigma}(\theta) = \boldsymbol{\Sigma}$, where

$$\chi^2 = F(n-1) \approx \chi^2(\alpha, P^* - Q) \quad (9.14)$$

This equation shows—given that the model is correct, variables are approximately multivariate normally distributed, and the sample size is sufficiently large—that the product of the minimized discrepancy function and sample size minus one is asymptotically chi-square distributed with degrees of freedom equal to $P^* - Q$. Also, it is straightforward to show that SEM parameter estimates are asymptotically unbiased, consistent, and asymptotically efficient (Hoyle 1995).

Equation 9.14 needs to be applied with care. Its unsuitability as a criterion for model assessment and selection was pointed out early in SEM theory development because the test statistic is largely a function of sample size (Bentler and Bonett 1980; Gullikson and Tukey 1958; Joreskog 1969). Thus, the χ^2 best serves the analyst in the selection of the best from competing models estimated on the same data, and whose absolute value should be evaluated with respect to sample size on which the statistic is estimated.

9.3.3 Nonideal Conditions in the Structural Equation Model

Recall that ideal conditions in SEM include multivariate normality of independent variables, the correct model functional form, independent and dependent variables measured on the interval or ratio scale, and a sufficiently large sample size. A large number of studies have been conducted to assess the impact of continuous yet nonnormal variables on SEMs (see, for instance, Browne 1984; Chou et al. 1991; Finch et al. 1994; Hu et al. 1992; and Kline 1998). Nonnormality can arise from poorly distributed continuous variables or coarsely categorized continuous variables. Nonnormality is detected in a number of ways, including box plots, histograms, normal probability plots, and by inspection of multivariate kurtosis. Numerous studies have arrived at similar conclusions regarding the impact of nonnormality in SEMs. The χ^2 test statistic becomes inflated as the data become more nonnormal. In addition, the GLS and MLE methods of parameter estimation produce inflated χ^2 test statistics with small sample sizes, even if multivariate normality is satisfied. In addition, model GOF indices are underestimated under nonnormality and nonnormality leads to moderate to severe underestimation of standard errors of parameter estimates.

There are several remedies for dealing with nonnormality. The asymptotically distribution-free estimator (ADF) is a GLS estimation approach that does not rely on multivariate normality (Browne 1984). The ADF estimator produces asymptotically unbiased estimates of the χ^2 test statistic, parameter estimates, and standard errors. The scaled χ^2 test statistic, developed by

Satorra and Bentler (see Satorra 1990), corrects or rescales the χ^2 test statistic so that it approximates the referenced χ^2 distribution.

Bootstrapping is a third method for dealing with nonnormal samples. Bootstrapping is based on the principle that the obtained random sample is a fair representation of the population distribution, and by resampling from this sample, estimates of parameters and their standard errors obtained are reliable estimates of the true population parameters. Efron and Tibshirani (1986) have demonstrated that in many studies the sampling distribution is reasonably approximated by data obtained from a single sample. Details of the bootstrap approach to SEM is provided in Bollen and Stine (1992).

Nominal and ordinal scale variables also cause problems in SEMs—resulting in biased estimates of χ^2 test statistics and estimated parameters and their standard errors. One approach, developed by Muthén (1984), consists of a continuous/categorical variable methodology (CVM) weighted least squares estimator and discrepancy function, which results in unbiased, consistent, and efficient parameter estimates when variables are measured on nominal and ordinal scales. However, this estimator requires large sample sizes (at least 500–1,000 cases), and is difficult to estimate for overly complex models (Hoyle 1995). Other approaches include variable reexpressions (Cattell and Burdsal 1975), variable transformations (Daniel and Wood 1980; Emerson and Stoto 1983), and alternating conditional expectations and Box–Cox transformations (de Veaux 1990).

Interactions and nonlinear effects arise frequently in the modeling of real data. In SEM, interactions and nonlinear effects present challenges above and beyond those encountered in simple linear regression. There are two general approaches to handling these problems; the indicant product approach, and the multisample approach. The indicant product approach is only well developed for multiplicative cases, and requires a centering transformation. The multisample approach is more flexible, avoids some multicollinearity and distributional problems associated with the product indicant approach, and is suitable under the widest range of conditions (Rigdon et al. 1998). Most currently available SEM software packages can accommodate the multisample approach. The reader is encouraged to consult Schumacker and Marcoulides (1998) for a complete treatment of this topic as it relates to SEMs.

9.3.4 Model Goodness-of-Fit Measures

Model GOF measures are an important part of any statistical model assessment. GOF measures in SEMs are an unsettled topic, primarily as a result of lack of consensus on which GOF measures serve as “best” measures of model fit to empirical data (Arbuckle and Wothke 1995). For detailed discussions of these debates and a multitude of SEM GOF methods see Mulaik et al. (1989), MacCallum (1990), Steiger (1990), Bollen and Long (1993), and Arbuckle and Wothke (1995).

Several important concepts are routinely applied throughout SEM GOF tests that enable the assessment of statistical models. A saturated model is

a model that is perfectly fit to the data—the variance–covariance structure is completely unconstrained and represents an unappealing model. It is the most general model possible, and is used as a standard of comparison to the estimated model. Because the saturated model is as complex as the original data, it does not summarize the data into succinct and useful relationships. In contrast, the independence model is constrained such that no relationships exist in the data and all variables in the model are independent of each other. This model presents the “worst case” model. The saturated and independence models are typically viewed as two extremes within which the best model lies.

There are a large number of GOF criteria available for assessing the fit of SEMs. Several important and widely used GOF measures are described, and the reader is encouraged to examine other GOF measures in the references provided.

The first class of GOF indices includes measures of parsimony. Models with few parameters are preferred to models with many parameters, providing that the important underlying model assumptions are not violated. This modeling philosophy is borne by a general desire to explain complex phenomena with as simple a model as possible. Three simple measures of parsimony are the number of model parameters Q , the degrees of freedom of the model being tested $df = P^* - Q$, and the parsimony ratio

$$PR = \frac{d}{d_i} \quad (9.15)$$

where d is the degrees of freedom of the estimated model and d_i is the degrees of freedom of the independence model. The PR represents the number of parameter constraints of the estimated model as a fraction of the number of constraints in the independence model (a higher PR is preferred).

There are several GOF indices based on the discrepancy function F shown in Equation 9.12. As stated previously, the χ^2 test statistic, derived from the discrepancy function, needs to be treated with care because it is dependent largely on sample sizes—small samples tending to accept (fail to reject) the null hypothesis, and large samples tending to reject the null hypothesis.

The X^2 statistic is the minimum value of the discrepancy function F times its degrees of freedom (see Equation 9.14). The p -value is the probability of obtaining a discrepancy function as large as or larger than the one obtained by random chance if the model is correct, distributional assumptions are correct and the sample size is sufficiently large. The statistic $X^2/(\text{model degrees of freedom})$ has been suggested as a useful fit measure. Rules of thumb have suggested that this measure (except under ULS and SLS estimation) should be close to 1 for correct models. In general, researchers have recommended this statistic lie somewhere less than 5, with values close to 1 being preferred (Byrne 1989; Carmines and McIver 1981; Marsh and Hocevar 1985).

Another class of fit measures is based on the population discrepancy. These measures rely on the notion of a population discrepancy function (as opposed to the sample discrepancy function) to estimate GOF measures, including the noncentrality parameter (NCP), the root mean square error of approximation (RMSEA), and PCLOSE, the *p*-value associated with a hypothesis test of $\text{RMSEA} \leq 0.05$. For details on these measures the reader should consult Steiger et al. (1985) and Browne and Cudeck (1993).

Information theoretic measures are designed primarily for use with MLE methods, and are meant to provide a measure of the amount of information contained in a given model. There are many measures used to assess fit in this class. The Akaike information criterion (Akaike 1987) is given as

$$AIC = 2Q - 2LL(\boldsymbol{\theta}) \quad (9.16)$$

where Q is the number of parameters and $LL(\boldsymbol{\theta})$ is the log-likelihood at convergence. Lower values of AIC are preferred to higher values because higher values of $-2LL(\boldsymbol{\theta})$ correspond to greater lack of fit. In the AIC criterion a penalty is imposed on models with larger numbers of parameters, similar to the adjusted *R*-square measure in regression. The Browne-Cudeck (1989) criterion is similar to AIC, except it imposes a slightly greater penalty for model complexity than does AIC. It is also the only GOF measure in this class of measures designed specifically for the analysis of moment structures (Arbuckle and Wothke 1995).

Yet another class of GOF measures is designed to compare the fitted model to baseline models. The normed fit index, developed by Bentler and Bonett (1980), is given as

$$NFI = 1 - \frac{X^2}{X_b^2} \quad (9.17)$$

where X_b^2 is the minimum discrepancy of the baseline model comparison, usually the saturated or independence model. NFI will take on values between 0 and 1. Values close to one indicate a close fit to the saturated model (as typically measured to the terribly fitting independence model). Rules of thumb for this measure suggest that models with a NFI less than 0.9 can be substantially improved (Bentler and Bonnet 1980). Other GOF measures in this category include the relative fit index (RFI), the incremental fit index (IFI), the Tucker-Lewis coefficient, and the comparative fit index (CFI), discussion on which is found in Bollen (1986), Bentler (1990), and Arbuckle and Wothke (1995).

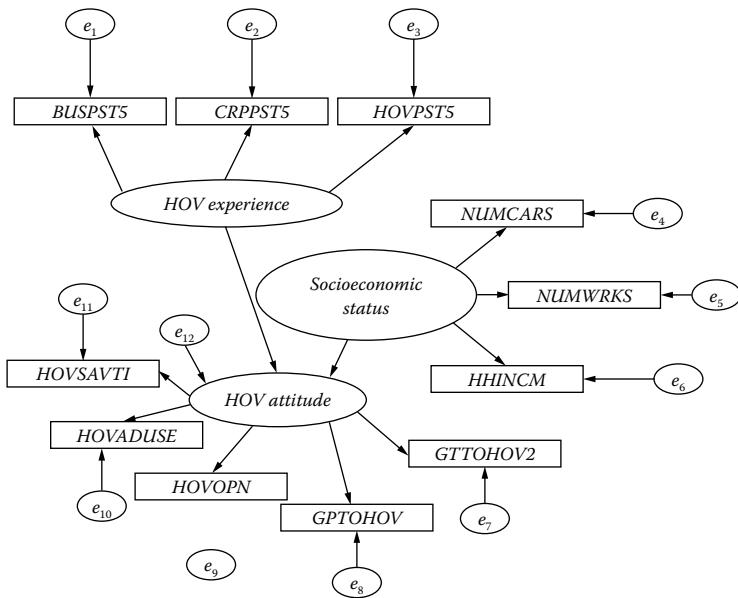
9.3.5 Guidelines for Structural Equation Modeling

Similar to other statistical modeling methods, practical guidelines are useful in the estimation of SEMs. To review, there are a couple of reasons for using a SEM. First, SEMs handle measurement problems well. They are ideal when exogenous variables measure an underlying unobservable or latent construct or constructs. Second, SEMs provide a way to check the entire structure of data assumptions, not just whether the dependent variable predictions fit observations well. Third, SEMs cope with endogeneity among variables well. When numerous endogenous variables are present in a model, the SEM framework can account for the feedback present in these relationships.

The need for the analyst to have a well-articulated theory regarding the data generating process before structural equation modeling is compounded. The complexity of variable relationships accommodated in the SEM framework translates to a significant increase in the potential for data mining. One must be careful to prespecify direct, indirect, and associative relationships between variables that correspond with theory and expectation. In addition, an understanding of the measurement model or challenge in the data is needed along with knowing which underlying latent constructs are being sought, and which variables reflect these constructs. In transportation applications, latent constructs include characteristics like attitudes in favor or against a policy or program, educational background, level of satisfaction with a program or service, and socioeconomic status to name but a few. Finally, assessment of a SEM should take into account many criteria, including theoretical appeal of the model specification, overall χ^2 GOF between observed and implied variance–covariance matrices, individual variable coefficients and their standard errors, and GOF indices. As always, the analyst should rely on consensus building over time to establish a model as a reliable depiction of the underlying reality.

Example 9.4

The results from the principal components and factor analyses are used to develop the SEM specification. In the SEM presented here, three important latent variables are thought to define the structure in the HOV survey data. The first is *HOV attitude*, which reflects the respondent's attitude toward HOV lanes and their use, positive values reflecting support for HOV lanes and negative values reflecting distaste for HOV lanes. The second is *socioeconomic status*, which reflects the relative level of personal wealth combined with relative social status. The third latent variable thought to be important is *HOV experience*, which is thought to measure how much experience using HOV lanes the respondent has had. It is expected that *HOV experience* will influence someone's *HOV attitude*, that is, someone who has experienced the benefits from utilizing an HOV facility will have a greater attitude toward HOV facilities (presuming that they are effective). It might also be that *HOV attitude* influences *HOV experience*, that is, respondents with a better attitude are more likely to use HOV facilities. It is also suspected that

**FIGURE 9.3**

SEM model of HOV survey data: three latent variables.

respondents with higher *socioeconomic* status will in general have better attitudes toward HOV facilities.

As shown in Figure 9.3 (with error terms represented in ellipses), the responses to questions in the HOV survey were determined by underlying latent variables (identified in part by the factor and principal components analyses). For example, *HOV experience* is influenced positively by a respondent who regularly engages in HOV use, either through carpooling, bus, or some other HOV qualifying use. The relationship between latent variables and the measured variables is shown in the figure. Table 9.5 shows the estimation results for the SEM depicted in Figure 9.3. The initial expectation belief that *HOV experience* and *HOV attitude* were mutually endogenous was not supported in the modeling effort. Instead, *HOV experience* directly influenced *HOV attitude*. As seen in the table, there is negligible support for an effect of *socioeconomic status* on *HOV attitude*; however, it was the best specification of all specifications tested, and overall model GOF indices are reasonable. All other estimated model parameters were statistically significant at well beyond the 90% level of confidence.

As shown in Table 9.5, the SEM reflects 12 simultaneous regression equations. Constraints were imposed to ensure identifiability, for example, the mean and variance of the variable socioeconomic status were set to 0 and 1, respectively. Some of the data were missing as a result of incomplete surveys—MLE estimates were used to impute missing values (for details see Arbuckle and Wothke 1995). Because of computational constraints, other estimation methods are not available when there are missing data—limiting the number of estimation methods that were available.

TABLE 9.5

Maximum Likelihood Estimation Results for SEM of HOV Survey Data

Estimated Parameters	Estimate	Standard Error	Z-Value
<i>Regression Parameters</i>			
<i>HOV Attitude</i> \leftarrow Socioeconomic Status	-0.001	0.080	-0.017
<i>HOV Attitude</i> \leftarrow HOV Experience	0.507	0.078	6.509
<i>CRPPST5</i> \leftarrow HOV Experience	0.848	0.085	10.001
<i>HOVPST5</i> \leftarrow HOV Experience	1.724	0.079	21.908
<i>BUSPST5</i> \leftarrow HOV Experience	0.386	0.061	6.324
<i>HOVSAVTI</i> \leftarrow HOV Attitude	0.905	0.068	13.333
<i>HOVADUSE</i> \leftarrow HOV Attitude	0.661	0.056	11.875
<i>HOVOPN</i> \leftarrow HOV Attitude	-1.164	0.075	-15.505
<i>GTTOHOV2</i> \leftarrow HOV Attitude	0.542	0.065	8.283
<i>NUMCARS</i> \leftarrow Socioeconomic Status	0.832	0.185	4.497
<i>NUMWRKS</i> \leftarrow Socioeconomic Status	0.374	0.091	4.114
<i>HHINCM</i> \leftarrow Socioeconomic Status	0.431	0.126	3.410
<i>GPTOHOV</i> \leftarrow HOV Attitude	0.847	0.067	12.690
<i>Intercepts</i>			
<i>HOVPST5</i>	0.973	0.105	9.234
<i>CRPPST5</i>	0.643	0.089	7.259
<i>BUSPST5</i>	0.290	0.061	4.785
<i>HOVADUSE</i>	1.268	20.053	0.063
<i>HOVOPN</i>	1.733	0.089	19.458
<i>GPTOHOV</i>	1.757	0.077	22.914
<i>HOVSAVTI</i>	1.787	0.079	22.742
<i>GTTOHOV2</i>	1.892	0.071	26.675
<i>NUMCARS</i>	2.418	0.061	39.524
<i>HHINCM</i>	6.220	0.088	70.537
<i>NUMWRKS</i>	1.849	0.046	40.250
<i>Goodness-of-Fit Measures</i>			
Degreews of freedom = $P^* - Q$			77 - 34 = 43
Chi-square			129.66
Parsimony ratio			0.652
Chi-square/degrees of freedom			3.02
Akaike's Information Criterion: Saturated/fitted/independence			154/198/4503
Browne-Cudeck Criterion: Saturated/fitted/independence			150/200/4504
Normed Fit Index			0.971

GOF statistics for the SEM are shown at the bottom of the table. The GOF statistics are in general encouraging. The chi-square statistics divided by model degrees of freedom is around 3, an acceptable statistic. Akaike's information criterion close to the value for the saturated model. Finally, the normed fit index is about 0.971, a value fairly close to 1.

Given the significance of the SEM variables, the theoretical appeal of the structural portion of the model, and the collective GOF indices, the SEM is considered to be a good approximation of the underlying data generating process. Some improvements are possible, especially the role of the variable socioeconomic status.

10

Duration Models

In many instances, one encounters the need to study the elapsed time until the occurrence of an event or the duration of an event. Data such as these are referred to as duration data, and are encountered often in the field of transportation research. Examples include the time until a vehicle accident occurs, the time between vehicle purchases, the time devoted to an activity (shopping, recreational, etc.), and the time until the adoption of new transportation technologies. Duration data are usually continuous and can, in most cases, be modeled using least-squares regression. The use of estimation techniques that are based on hazard functions, however, can often provide additional insights into the underlying duration problem (Hensher and Mannering 1994; Kiefer 1988).

10.1 Hazard-Based Duration Models

Hazard-based duration modeling has been extensively applied in a number of fields, including biostatistics (Fleming and Harrington 1991; Kalbfleisch and Prentice 1980) and economics (Kiefer 1988). In the transportation field, the application of such models has been relatively limited. One example of duration modeling is research on trip-generating activities such as the length of time spent shopping and engaging in recreational/social visits and the length of time spent staying at home between trips (see Bhat 1996a, 1996b; Ettema and Timmermans 1997; Hamed and Mannering 1993; Kharoufeh and Goulias 2002).

To study duration data, hazard-based models are applied to study the conditional probability of a time duration ending at some time t , given that the duration has continued until time t . For example, consider the duration of time until a vehicle accident occurs as beginning when a person first becomes licensed to drive (see Mannering 1993). Hazard-based duration models can account for the possibility that the likelihood of a driver becoming involved in an accident may change over time. As time passes, driving experience is gained and there may be changes in drivers' confidence, skill, and risk-taking behavior. Thus, one would expect the probability of an accident to increase or decrease, although it may also be possible that the probability of an accident would remain constant over time. Probabilities that change as time passes are ideally suited to hazard-function analyses.

Developing hazard-based duration models begins with the cumulative distribution function

$$F(t) = P(T < t) \quad (10.1)$$

where P denotes probability, T is a random time variable, and t is some specified time. Using the example of time until an accident, Equation 10.1 gives the probability of having an accident before some transpired time t . The density function corresponding to this distribution function (the first derivative of the cumulative distribution with respect to time) is

$$f(t) = \frac{dF(t)}{dt} \quad (10.2)$$

and the hazard function is

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (10.3)$$

where $h(t)$ is the conditional probability that an event will occur (e.g., an accident, death, or end of a shopping trip) between time t and $t + dt$, given that the event has not occurred up to time t . In words, $h(t)$ gives the rate at which event durations are ending at time t (such as the duration in an accident-free state that would end with the occurrence of an accident), given that the event duration has not ended up to time t . The cumulative hazard $H(t)$ is the integrated hazard function, and provides the cumulative rate at which events are ending up to or before time t .

The survivor function, which provides the probability that a duration is greater than or equal to some specified time t is also frequently used in hazard analyses for interpretation of results. The survivor function is

$$S(t) = P(T \geq t) \quad (10.4)$$

If one of these functions is known—the density, cumulative distribution, survivor, hazard, or integrated hazard—any of the others are readily obtained. The relationships between these functions are as follows

$$\begin{aligned} S(t) &= 1 - F(t) = 1 - \int_0^t f(t) dt = \text{EXP}[-H(t)] \\ f(t) &= \frac{d}{dt} F(t) = h(t) \text{EXP}[-H(t)] = -\frac{d}{dt} S(t) \\ H(t) &= \int_0^t h(t) dt = -\text{LN}[S(t)] \\ h(t) &= \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \frac{d}{dt} H(t) \end{aligned} \quad (10.5)$$

Graphically, hazard, density, cumulative distribution, and survivor functions are illustrated in Figure 10.1.

The slope of the hazard function (the first derivative with respect to time) has important implications. It captures dependence of the probability of a duration ending on the length of the duration (duration dependence). Consider a driver's probability of having an accident and the length of time without having an accident. Figure 10.2 presents four possible hazard functions for this case. The first hazard function $h_1(t)$ has $dh_1(t)/dt < 0$ for all t . This hazard is monotonically decreasing in duration, implying that the longer drivers go without having an accident, the less likely they are to have one soon. The second hazard function is nonmonotonic and has $dh_2(t)/dt > 0$ and $dh_2(t)/dt < 0$ depending on the length of duration t . In this case the accident probabilities increase or decrease in duration. The third hazard function has $dh_3(t)/dt > 0$ for all t and is monotonically increasing in duration. This implies that the longer drivers go without having an accident the more likely they are to have an accident soon. Finally, the fourth hazard function has $dh_4(t)/dt = 0$, which means that accident probabilities are independent of duration and no duration dependence exists.

In addition to duration dependence, hazard-based duration models account for the effect of covariates on probabilities. For example, for vehicle accident probabilities, gender, age, and alcohol consumption habits would all be expected to influence accident probabilities. Proportional hazards and accelerated lifetime models are two alternate methods that have been popular in accounting for the influence of covariates.

The proportional-hazards approach assumes that the covariates, which are factors that affect accident probabilities, act multiplicatively on some

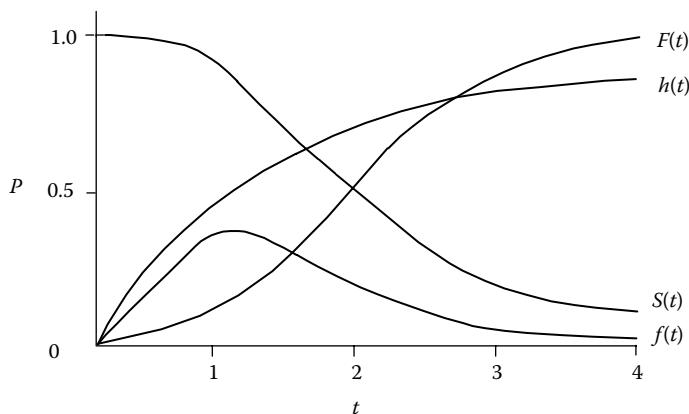


FIGURE 10.1

Illustration of hazard ($h(t)$), density ($f(t)$), cumulative distribution ($F(t)$), and survivor functions ($S(t)$).

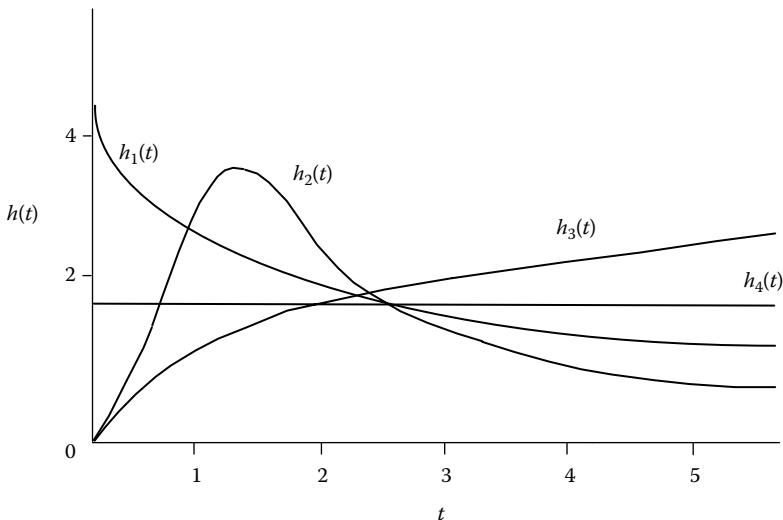
**FIGURE 10.2**

Illustration of four alternate hazard functions.

underlying hazard function. This underlying (or baseline) hazard function is denoted by $h_o(t)$, and is the hazard function assuming that all elements of a covariate vector \mathbf{X} are zero. For simplicity, covariates are assumed to influence the baseline hazard with the function $EXP(\boldsymbol{\beta}\mathbf{X})$, where $\boldsymbol{\beta}$ is a vector of estimable parameters. Thus the hazard rate with covariates is

$$h(t|\mathbf{X}) = h_o(t)EXP(\boldsymbol{\beta}\mathbf{X}) \quad (10.6)$$

This proportional-hazards approach is illustrated in Figure 10.3.

The second approach for incorporating covariates in hazard-based models is to assume that the covariates rescale (accelerate) time directly in a baseline survivor function, which is the survivor function when all covariates are zero. This accelerated lifetime method again assumes covariates influence the process with the function $EXP(\boldsymbol{\beta}\mathbf{X})$. The accelerated lifetime model is written as

$$S(t|\mathbf{X}) = S_o[t EXP(\boldsymbol{\beta}\mathbf{X})] \quad (10.7)$$

which leads to the conditional hazard function

$$h(t|\mathbf{X}) = h_o[t EXP(\boldsymbol{\beta}\mathbf{X})]EXP(\boldsymbol{\beta}\mathbf{X}) \quad (10.8)$$

Accelerated lifetime models have, along with proportional-hazards models, enjoyed widespread use and are estimable by standard maximum likelihood methods (see Kalbfleisch and Prentice 1980).

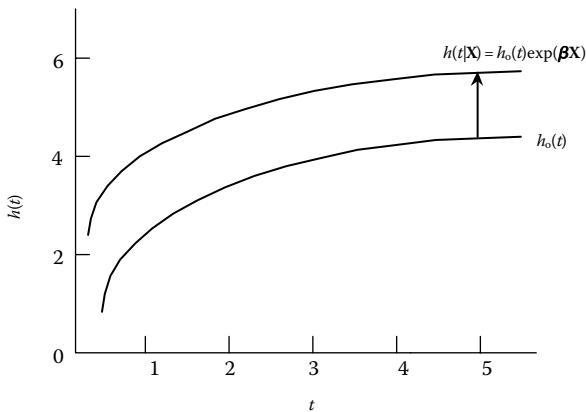
**FIGURE 10.3**

Illustration of the proportional-hazards model.

10.2 Characteristics of Duration Data

Duration data are often left or right censored. For example, consider the time of driving a vehicle until a driver's first accident. Suppose data are only available for reported accidents over a specified time period beginning at time a in Figure 10.4 and ending at time b . Observation 1 is not observed, since it does not fall within the time period of observation. Observation 2 is left and right censored because it is not known when driving began and the first accident is not observed in the a to b time interval. Observation 3 is complete with both start and ending times in the observed period. Observations 4 and 6 are left censored and observation 5 is right censored.

Hazard-based models can readily account for right-censored data (see Kalbfleisch and Prentice 1980). With left-censored data the problem becomes one of determining the distribution of duration start times so that they are exploited to determine the contribution of the left-censored data to the model's likelihood function. Left-censored data creates a far more difficult problem because of the additional complexity added to the likelihood function. For additional details regarding censoring problems in duration modeling, the interested reader should refer to Kalbfleisch and Prentice (2002), Heckman and Singer (1984), and Fleming and Harrington (1991).

Another challenge may arise when a number of observations end their durations at the same time. This is referred to as the problem of tied data. Tied data can arise when data collection is not precise enough to identify exact duration-ending times. When duration exits are grouped at specific times, the likelihood function for proportional and accelerated lifetime

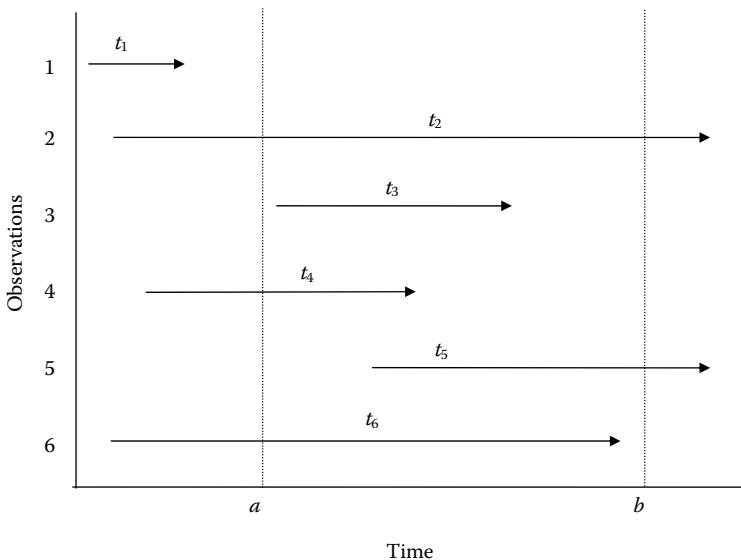


FIGURE 10.4
Illustration of duration data.

models becomes increasingly complex. Kalbfleisch and Prentice (2002) and Fleming and Harrington (1991) provide an excellent discussion on the problem of tied data.

10.3 Nonparametric Models

Although applications in the transportation field are rare, there are numerous nonparametric (distribution free) survival analysis applications in fields such as medicine and epidemiology. Because of the predominant use of semi-parametric and parametric methods in the field of transportation, only the fundamentals of nonparametric methods are covered here. A lack of transportation applications of nonparametric methods in no way implies limited utility of these methods. Nonparametric methods afford the ability to model survival or duration data without relying on specific or well-behaved statistical distributions.

There are two popular approaches for generating survival functions for nonparametric methods, the product-limit (PL) method developed by Kaplan and Meier (1958), and life tables. The PL estimate is based on individual survival times, whereas the life table method groups survival times into intervals.

The basic method for calculating survival probabilities using the PL begins by specifying the probability of surviving r years (without event A occurring) as the conditional probability of surviving r years given survival for $r-1$ years times the probability of surviving $r-1$ years (or months, days, minutes, etc.). In notation, the probability of surviving k or more years is given by

$$\hat{S}(k) = (p_k | p_{k-1}) \dots (p_4 | p_2)(p_3 | p_2)(p_2 | p_1)(p_1) \quad (10.9)$$

where $(p_k | p_{k-1})$ is the proportion of observed subjects surviving to period k , given survival to period $k-1$, and so on. This PL estimator produces a survival step-function that is based purely on the observed data. The Kaplan–Meier method provides useful estimates of survival probabilities and a graphical presentation of the survival distribution. It is the most widely applied nonparametric method in survival analysis (Lee 1992).

A few observations about this nonparametric are worth mentioning. Kaplan–Meier estimates are limited. If the largest (survival) observation is right-censored, the PL estimate is undefined beyond this observation. If the largest observation is not right censored, then the PL estimate at that time equals zero, which is technically correct because no survival time longer than this was observed. In addition, the median survival time cannot be estimated if more than 50% of the observations are censored and the largest observation is censored.

The PL method assumes that censoring is independent of survival times. If this is false, the PL method is inappropriate. For example, a study of drivers' crash probabilities, where drivers remove themselves from the study due to the burden of participating would violate the independence assumption.

Life table analysis, an alternative method that is similar to nonparametric survival analysis, is described in detail by Lee (1992).

10.4 Semiparametric Models

Both semiparametric and fully parametric hazard-based models have been widely cited in the literature. Fully parametric models assume a distribution of duration times (e.g., Weibull, exponential, etc.) and also have a parametric assumption on the functional form of the covariates' influence on the hazard function (usually $\text{EXP}(\beta X)$ as discussed previously). Semiparametric models, in contrast, are more general in that they do not assume a duration-time distribution, although they do retain the parametric assumption of the covariate influence.

A nonparametric approach for modeling the hazard function is convenient when little or no knowledge of the functional form of the hazard is available. Such an approach was developed by Cox (1972) and is based on

the proportional-hazards approach. The Cox proportional-hazards model is semiparametric because $\text{EXP}(\boldsymbol{\beta}\mathbf{X})$ is still used as the functional form of the covariate influence. The model is based on the ratio of hazards—so that the probability of an observation i exiting a duration at time t_i , given that at least one observation exits at time t_i , is given as

$$\frac{\text{EXP}(\boldsymbol{\beta}\mathbf{X}_i)}{\sum_{j \in R_i} \text{EXP}(\boldsymbol{\beta}\mathbf{X}_j)} \quad (10.10)$$

where R_i denotes the set of observations j , with durations greater than or equal to t_i . This model is readily estimated using standard maximum likelihood methods. If only one observation completes its duration at each time (no tied data), and no observations are censored, the partial log-likelihood is

$$LL = \sum_{i=1}^I \left[\boldsymbol{\beta}\mathbf{X}_i - \sum_{j \in R_i} \text{EXP}(\boldsymbol{\beta}\mathbf{X}_j) \right] \quad (10.11)$$

If no observations are censored and tied data are present with more than one observation exiting at time t_i , the partial log-likelihood is the sum of individual likelihoods of the n_i observations that exit at time t_i

$$LL = \sum_{i=1}^I \left[\boldsymbol{\beta} \sum_{j \in t_i} \mathbf{X}_j - ni \sum_{j \in R_i} \text{EXP}(\boldsymbol{\beta}\mathbf{X}_j) \right] \quad (10.12)$$

Example 10.1

A survey of 204 Seattle-area commuters was conducted to examine the duration of time that commuters delay their work-to-home trips in an effort to avoid peak-period traffic congestion. Of the 204 commuters surveyed, 96 indicated that they sometimes delayed their work-to-home trip to avoid traffic congestion (see Mannering and Hamed 1990, for additional details). These commuters provided their average time delay—thus each commuter has a completed delay duration so that neither left nor right censoring is present in the data. Observed average commuter delays ranged from 4 minutes to 4 hours with a sample average of 51.3 minutes and a standard deviation of 37.5 minutes.

To determine significant factors that affect the duration of commuters' delay, a Cox proportional hazards is estimated using the variables shown in Table 10.1. Model estimation results are presented in Table 10.2. Table 10.2 shows that the multiplicative effect of the male indicator variable increases the hazard (see Equation 10.7) and thus shortens the expected commuting delay. The finding that

TABLE 10.1

Variables Available for Work-To-Home Commute Delay Model

Variable No.	Variable Description
1	Minutes delayed to avoid congestion
2	Primary activity performed while delaying: 1 if perform additional work, 2 if engage in nonwork activities, or 3 if do both
3	Number of times delayed in the past week to avoid congestion
4	Mode of transportation used on work-to-home commute: 1 if by single occupancy vehicle, 2 if by carpool, 3 if by vanpool, 4 if by bus, 5 if by other
5	Primary route to work in Seattle area: 1 if Interstate 90, 2 if Interstate 5, 3 if State Route 520, 4 if Interstate 405, 5 if other
6	In the respondent's opinion, is the home-to-work trip traffic congested: 1 if yes, 0 if no
7	Commuter age in years: 1 if under 25, 2 if 26–30, 3 if 31–35, 4 if 36–40, 5 if 41–45, 6 if 46–50, 7 if over 50
8	Respondent's gender 1 if female, 0 if male
9	Number of cars in household
10	Number of children in household
11	Annual household income (US dollars per year): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if 50,000–59,999, 6 if over 60,000
12	Respondent has flexible work hours? 1 if yes, 0 if no
13	Distance from work to home (in kilometers)
14	Respondent faces level of service D or worse on work-to-home commute? 1 if yes, 0 if no
15	Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time
16	Population of work zone
17	Retail employment in work zone
18	Service employment in work zone
19	Size of work zone (in hectares)

TABLE 10.2

Cox Proportional Hazard Model Estimates of the Duration of Commuter Work-To-Home Delay to Avoid Congestion

Variable Description	Estimated Parameter	t-Statistic
Male indicator (1 if male commuter, 0 if not)	0.246	1.07
Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time	-1.070	-2.59
Distance from home to work in kilometers	-0.028	-1.73
Resident population of work zone	-0.1761E-04	-1.58
Number of observations	96	
Log-likelihood at zero	-361.56	
Log-likelihood at convergence	-325.54	

men have shorter delays is not highly significant though, with a t -statistic of 1.07. The ratio of actual travel time, at the time of expected departure, to free-flow (uncongested) travel time decreases the hazard and increases delay. This variable captures the impact of high levels of congestion on increasing overall work-to-home commute departure delay. The greater was the distance from home to work the greater was the departure delay—perhaps capturing the time needed for congestion to dissipate. Finally, the greater the resident population in the area (zone) of work, the greater the length of delay. This variable likely reflects the availability of activity opportunities during the delay.

Note from Table 10.2 that a constant is not estimated. It is not possible to estimate a constant because the Cox proportional-hazards model is homogeneous of degree zero in \mathbf{X} (see Equation 10.10). Thus, any variable that does not vary across observations cannot be estimated, including the constant term.

10.5 Fully Parametric Models

With fully parametric models, a variety of distributional alternatives for the hazard function have been used with regularity in the literature. These include gamma, exponential, Weibull, log-logistic, and Gompertz distributions, among others. Table 10.3 shows the names and corresponding hazard functions (with their associated distribution parameters) for a variety of parametric duration models. The choice of any one of these alternatives is

TABLE 10.3

Some Commonly used Hazard Functions for Parametric Duration Models

Name	Hazard Function $h(t)$
Compound exponential	$h(t) = \frac{P}{t + (P / \lambda_0)}$
Exponential	$h(t) = \lambda$
Exponential with gamma heterogeneity	$h(t) = \frac{\lambda}{1 + \theta\lambda t}$
Gompertz	$h(t) = (P)EXP^{\lambda t}$
Gompertz-Makeham	$h(t) = \lambda_0 + \lambda_1 EXP^{\lambda_2 t}$
Log-logistic	$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1 + (\lambda t)^P}$
Weibull	$h(t) = (\lambda P)(\lambda t)^{P-1}$
Weibull with gamma heterogeneity	$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1 + \theta(\lambda t)^P}$

justified on theoretical grounds or statistical evaluation. The choice of a specific distribution has important implications relating not only to the shape of the underlying hazard, but also to the efficiency and potential biasedness of the estimated parameters.

This chapter looks closely at three distributions: exponential, Weibull, and log-logistic. For additional distributions used in hazard-based analysis the reader is referred to Kalbfleisch and Prentice (2002) and Cox and Oakes (1984).

The simplest distribution to apply and interpret is the exponential. With parameter $\lambda > 0$, the exponential density function is

$$f(t) = \lambda EXP(-\lambda t) \quad (10.13)$$

with hazard,

$$h(t) = \lambda \quad (10.14)$$

Equation 10.14 implies that this distribution's hazard is constant, as illustrated by $h_4(t)$ in Figure 10.2. This means that the probability of a duration ending is independent of time and there is no duration dependence. In the case of the time until a vehicle accident occurs, the probability of having an accident stays the same regardless of how long one has gone without having an accident. This is a potentially restrictive assumption because it does not allow for duration dependence.

The Weibull distribution is a more generalized form of the exponential. It allows for positive duration dependence (hazard is monotonic increasing in duration and the probability of the duration ending increases over time), negative duration dependence (hazard is monotonic decreasing in duration and the probability of the duration ending decreases over time), or no duration dependence (hazard is constant in duration and the probability of the duration ending is unchanged over time). With parameters $\lambda > 0$ and $P > 0$, the Weibull distribution has the density function,

$$f(t) = \lambda P(\lambda t)^{P-1} EXP[-(\lambda t)^P] \quad (10.15)$$

with hazard

$$h(t) = (\lambda P)(\lambda t)^{P-1} \quad (10.16)$$

As indicated in Equation 10.16, if the Weibull parameter P is greater than one, the hazard is monotone increasing in duration (see $h_3(t)$ in Figure 10.2); if P is less than one, it is monotone decreasing in duration (see $h_1(t)$ in Figure 10.2); and if P equals one, the hazard is constant in duration and reduces to the exponential distribution's hazard with $h(t) = \lambda$ (see $h_4(t)$ in Figure 10.2). Because the Weibull distribution is a more generalized form of the exponential distribution, it provides a more flexible means of capturing duration dependence.

However, it is still limited because it requires the hazard to be monotonic over time. In many applications, a nonmonotonic hazard is theoretically justified.

The log-logistic distribution allows for nonmonotonic hazard functions and is often used as an approximation of the more computationally cumbersome lognormal distribution. The log-logistic with parameters $\lambda > 0$ and $P > 0$ has the density function

$$f(t) = \lambda P(\lambda t)^{P-1}[1 + (\lambda t)^P]^{-2} \quad (10.17)$$

and hazard function

$$h(t) = \frac{(\lambda P)(\lambda t)^{P-1}}{1 + (\lambda t)^P} \quad (10.18)$$

The log-logistic's hazard is identical to the Weibull's except for the denominator. Equation 10.18 indicates that if $P < 1$, then the hazard is monotone decreasing in duration (see $h_1(t)$ in Figure 10.2); if $P = 1$, then the hazard is monotone decreasing in duration from parameter λ ; and if $P > 1$, then the hazard increases in duration from zero to an inflection point, $t_i = (P - 1)^{1/P}/\lambda$, and decreases toward zero thereafter (see $h_2(t)$ in Figure 10.2).

Example 10.2

Using the same data as in Example 10.1, the work-to-home departure delay is examined using exponential, Weibull, and log-logistic proportional-hazards models. Model estimation results are presented in Table 10.4. Note that these estimated fully parametric models include a constant term. Unlike the Cox proportional-hazards model, such terms are estimable.

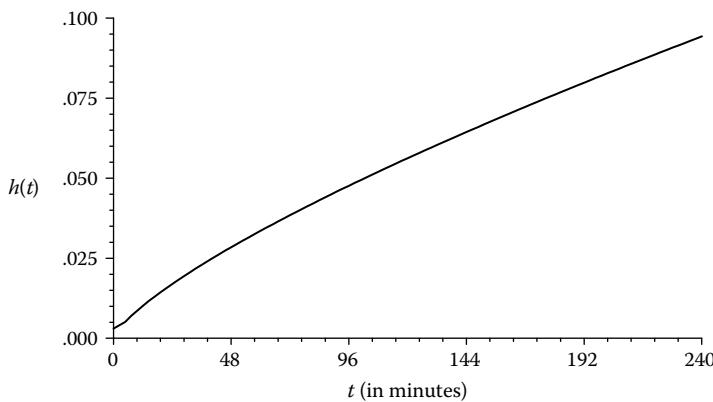
The signs of the parameters are identical to those found earlier in the Cox model. Of particular interest in these models are the duration effects. The exponential model assumes that the hazard function is constant, and thus the probability of a departure delay ending is independent of the time spent delaying. This assumption is tested by looking at the Weibull model estimation results. Recall the exponential model implicitly constrains $P = 1.0$. So if the Weibull estimate of P is significantly different from 1.0, the exponential model will not be valid and the hazard function is not constant over time. Table 10.4 shows the Weibull model parameter P is positive (indicating a monotonically increasing hazard) and significantly different from 0. To compare exponential and Weibull models, the appropriate test of whether P is significantly different from 1 is given as

$$t = \frac{\beta - 1}{S(\beta)} = \frac{1.745 - 1}{0.175} = 4.26 \quad (10.19)$$

TABLE 10.4

Hazard Model Parameter Estimates of the Duration of Commuter Work-To-Home Delay to Avoid Congestion (*t*-Statistics in Parentheses)

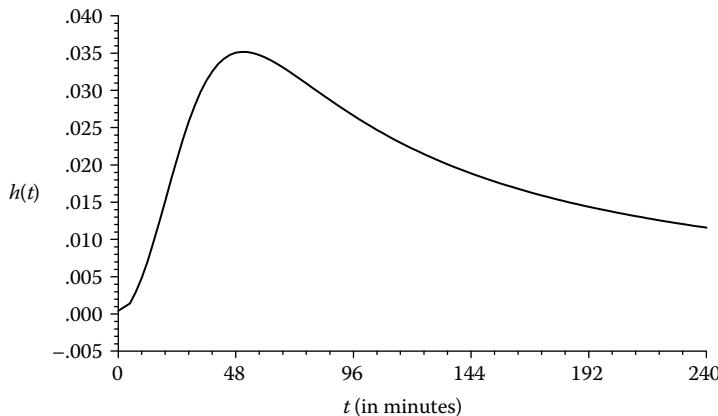
Variable Description	Exponential	Weibull	Log-Logistic
Constant	1.79 (1.02)	1.74 (2.46)	1.66 (3.07)
Male indicator (1 if male commuter, 0 if not)	0.173 (0.46)	0.161 (1.03)	0.138 (0.953)
Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time	-0.825 (-1.04)	-0.907 (-3.00)	-0.752 (-3.02)
Distance from home to work in kilometers	-0.035 (-0.65)	-0.032 (-1.55)	-0.041 (-2.69)
Resident population of work zone	-0.130E-05 (-0.57)	-0.137E-04 (-1.68)	-0.133E-05 (-1.90)
<i>P</i> (distribution parameter)	–	1.745 (9.97)	2.80 (10.17)
λ	0.020 (5.67)	0.018 (14.88)	0.024 (15.67)
Number of observations	96	96	96
Log-likelihood at convergence	-113.75	-93.80	-91.88

**FIGURE 10.5**

Estimated hazard function for the Weibull model in Example 10.2.

This *t*-statistic shows that *P* is significantly different from 1.0 and the Weibull model is preferred over the exponential. The resulting hazard function for the Weibull distribution is illustrated in Figure 10.5.

The log-logistic distribution relaxes the monotonicity of the hazard function. Table 10.4 shows that the parameter *P* is greater than one indicating that the hazard increases in duration from zero to an inflection $t_i = (P - 1)^{1/P}/\lambda$ and decreases

**FIGURE 10.6**

Estimated hazard function for the log-logistic model in Example 10.2.

toward zero thereafter. The value t_i is computed to be 51.23 minutes, the time after which the hazard starts to decrease. The log-logistic hazard for this problem is shown in Figure 10.6.

10.6 Comparisons of Nonparametric, Semiparametric, and Fully Parametric Models

The choice among nonparametric, semiparametric, and fully parametric methods for estimating survival or duration models is often complicated. When there is little information about the underlying distribution due to the small size of the sample or the lack of a theory that would suggest a specific distribution, a nonparametric approach may be appropriate. Nonparametric methods are a preferred choice when the underlying distribution is not known, while parametric methods are more suitable when underlying distributions are known or are theoretically justified (Lee 1992).

Semiparametric models may also be a good choice when little is known about the underlying hazard distribution. However, there are two drawbacks. First, duration effects are difficult to quantify. This drawback is problematic if a primary area of interest is how probabilities of duration exits change with respect to duration. Second, a potential loss in efficiency may result. It is straightforward to show that in data where censoring exists and the underlying survival distribution is known, the Cox semiparametric proportional-hazards model does not produce efficient parameter estimates. However, a number of studies have shown that the loss in efficiency is typically small (Efron 1977; Oaks 1977).

Comparing various hazard distributional assumptions for fully parametric models can also be difficult. Determining the relative difference between a Weibull and exponential model could be approximated by the significance of the Weibull's P parameter, which represents the difference between the two distributions (see Equations 10.15 and 10.17). In Example 10.2, it is noted that the Weibull model is superior to the exponential because the Weibull-estimated value of P is significantly different from 1.0. A likelihood ratio test comparing the two models is conducted using the log-likelihoods at convergence. The test statistic is

$$X^2 = -2(LL(\boldsymbol{\beta}_e) - LL(\boldsymbol{\beta}_w)) \quad (10.20)$$

where $LL(\boldsymbol{\beta}_e)$ is the log-likelihood at convergence for the exponential distribution and $LL(\boldsymbol{\beta}_w)$ is the log-likelihood at convergence for the Weibull distribution. This X^2 statistic is χ^2 distributed with one degree of freedom (representing the additional parameter estimated, P). In the case of the models discussed in Example 10.2, this outcome results in a X^2 test statistic of 39.9 $[-2(-113.75 - (-93.80))]$. With one degree of freedom, a confidence level of over 99.99% is obtained. This finding indicates that in less than 1 in 100 samples would the Weibull model by chance alone provide superior fit compared to the exponential model. This comparison is made directly because exponential and Weibull models are derivatives. This property is sometimes referred to as nesting since the exponential is simply a special case of the Weibull (with P constrained to be equal to one).

The difference between the Weibull and log-logistic models and other distributions is more difficult to test because the models may not be nested. One possible comparison for distributional models that are not nested is to compare likelihood ratio statistics (see Nam and Manning 2000). In cases like these, the likelihood ratio statistic is

$$X^2 = -2(LL(\mathbf{0}) - LL(\boldsymbol{\beta}_e)) \quad (10.21)$$

where $LL(\mathbf{0})$ is the initial log-likelihood (with all parameters equal to zero) and $LL(\boldsymbol{\beta}_e)$ is the log-likelihood at convergence. This X^2 statistic is χ^2 distributed with the degrees of freedom equal to the number of estimated parameters included in the model. One could select the distribution that provided the highest level of significance for this statistic to determine the best-fit distribution.

An informal method suggested by Cox and Oakes (1984) is to use plots of the survival and hazard distributions obtained using nonparametric methods to guide selection of a parametric distribution. With this approach, a visual inspection of the shapes and characteristics (inflection points, slopes) of the survival and hazard curves is used to provide insight into the selection of an appropriate parametric distribution.

10.7 Heterogeneity

While formulating proportional-hazard models an implicit assumption made is that the survival function (see Equation 10.6) is homogenous across observations. Thus, all of the variation in durations is assumed to be captured by the covariate vector X . A problem arises when some unobserved factors, not included in X , influence durations. This phenomenon is referred to as unobserved heterogeneity and may result in a major specification error that can lead one to draw erroneous inferences on the shape of the hazard function, in addition to producing inconsistent parameter estimates (Gourieroux et al. 1984; Heckman and Singer 1984).

In fully parametric models the most common approach to account for heterogeneity is to introduce a heterogeneity term designed to capture unobserved effects across the population and to work with the resulting conditional survival function. With a heterogeneity term, w , having a distribution over the population $g(w)$, along with a conditional survival function $S(t|w)$ the unconditional survival function becomes

$$S(t) = \int S(t|w)g(w) dw \quad (10.22)$$

To see how this heterogeneity term is applied, consider a Weibull distribution with gamma heterogeneity (Hui 1990). Without loss of generality, w is assumed to be gamma distributed with mean 1 and variance = $1/k$. So

$$g(w) = \frac{k^k}{\Gamma(k)} e^{-kw} w^{k-1} \quad (10.23)$$

With the Weibull distribution and $S(t) = f(t)/h(t)$, Equations 10.17 and 10.18 give

$$S(t|w) = e^{-(\omega\lambda t)^p} \quad (10.24)$$

The unconditional survival distribution can then be written as (with $\theta = 1/k$)

$$S(t) = \int_0^\infty S(t|w)g(w)dw = [1 + \theta(\lambda t)^p]^{-1/\theta} \quad (10.25)$$

resulting in the hazard function

$$h(t) = \lambda P(\lambda t)^{p-1}[S(t)]^\theta \quad (10.26)$$

Note that if $\theta = 0$, heterogeneity is not present because the hazard reduces to Equation 10.18 and the variance of the heterogeneity term w is zero.

Example 10.3

Using the data used in Example 10.1, a Weibull model is estimated with gamma heterogeneity and compared with the results of the Weibull model reported in Table 10.4. Model estimation results are presented in Table 10.5. While there is some variation in the parameter values and corresponding *t*-statistics, of particular concern is the estimation of θ (with its corresponding *t*-statistic of 1.58). The correct comparison of the two models is the likelihood ratio test

$$X^2 = -2(LL(\boldsymbol{\beta}_w) - LL(\boldsymbol{\beta}_{wh})) \quad (10.27)$$

where $LL(\boldsymbol{\beta}_w)$ is the log-likelihood at convergence for the Weibull model and $LL(\boldsymbol{\beta}_{wh})$ is the log-likelihood at convergence for the Weibull model with gamma heterogeneity. This test results in an X^2 statistic of 3.84 [$-2(-93.80 - (-91.88))$]. With one degree of freedom, a confidence level of 95% is obtained, suggesting that heterogeneity is present in the underlying Weibull survival process (assuming the Weibull specification is correct).

The selection of a heterogeneity distribution should not be taken lightly. The consequences of incorrectly specifying $g(w)$ is potentially severe and can result in inconsistent estimates as demonstrated both theoretically and empirically by Heckman and Singer (1984). Thus, while the gamma distribution has been a popular approach in accounting for heterogeneity, some caution must be exercised when interpreting estimation results based on

TABLE 10.5

Weibull versus Weibull with Gamma Heterogeneity Model Parameter Estimates of the Duration of Commuter Work-To-Home Delay to Avoid Congestion (*t*-Statistics in Parentheses)

Variable Description	Weibull	Weibull with Gamma Heterogeneity
Constant	1.74 (2.46)	1.95 (3.08)
Male indicator (1 if male commuter, 0 if not)	0.161 (1.03)	0.216 (1.58)
Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time	-0.907 (-3.00)	-0.751 (-2.70)
Distance from home to work in kilometers	-0.032 (-1.55)	-0.034 (-1.92)
Resident population of work zone	-0.137E-04 (-1.68)	-0.120E-05 (-1.57)
P (distribution parameter)	1.745 (9.97)	2.31 (6.18)
λ	0.018 (14.88)	0.024 (15.67)
θ	—	0.473 (1.58)
Number of observations	96	96
Log-likelihood at convergence	-93.80	-91.88

the selection of a specific parametric form for heterogeneity. Fortunately, for choosing among distributions, it has been shown that if the correct distribution is used for the underlying hazard function, parameter estimates are not highly sensitive to alternate distributional assumptions of heterogeneity (Kiefer 1988). Also, Meyer (1990) and Han and Hausman (1990) have shown that if the baseline hazard is nonparametric (a Cox proportional-hazards model), the choice of heterogeneity distribution may be less important with the estimation results being less sensitive to the distribution chosen. In other work, Bhat (1996a) has explored the use of nonparametric heterogeneity corrections in semiparametric and parametric models. The reader is referred to his work for further information on heterogeneity-related issues.

10.8 State Dependence

In duration modeling, state dependence refers to a number of processes that seek to establish a relationship between past duration experiences and current durations. Conceptually, the understanding of how past experience affects current behavior is a key component in modeling that captures important habitual behavior. State dependence is often classified into three types. Type I state dependence is duration dependence. As discussed previously, duration dependence is the conditional probability of a duration ending soon, given that it has lasted to some known time. Type I state dependence is captured in the shape of the hazard function (see Figure 10.2).

Type II state dependence is occurrence dependence. This dependence is reflected by the effect of previous durations on a current duration. For example, suppose interest is focused on modeling the duration of a traveler's time spent shopping. The number of times the traveler previously shopped during the same day may have an influence on current shopping duration. This dependence is accounted in the model by including the number of previous shopping durations as a variable in the covariate vector.

Type III state dependence, lagged duration dependence, accounts for the effect that lengths of previous durations have on a current duration. Returning to the example of the time spent shopping, an individual may have developed habitual behavior with respect to the time spent shopping that would make the length of previous shopping durations an excellent predictor of current shopping duration. This dependence is accounted for in a model's covariate vector by including the length of previous durations as a variable.

When including variables that account for Type II or Type III state duration dependence, the findings are easily misinterpreted. Misinterpretation arises because unobserved heterogeneity is often captured in the parameter estimates of the state dependence variables. To illustrate, suppose income is

an important variable for determining shopping duration. If income is omitted from the model, it becomes part of the unobserved heterogeneity. Because it is correlated with occurrence and lagged duration dependence, the parameter estimate of these variables is capturing both true state dependence and spurious state dependence (unobserved heterogeneity). Heterogeneity corrections (as in Equation 10.24) in the presence of occurrence and lagged duration variables do not necessarily help untangle true state dependence from heterogeneity (see Heckman and Borjas 1980). One simple solution is to instrument state variables by regressing them against exogenous covariates and using the regression-predicted values as variables in the duration model.

10.9 Time-Varying Covariates

Covariates that change over a duration are problematic. For example, suppose the duration between an individual's vehicular accidents is of particular interest. During the time between accidents it is possible that an individual may change the type of vehicle driven. This in turn could affect the probability of having a vehicular accident. If the covariate vector X changes over the duration being studied, parameter estimates may be biased. Time-varying covariates are difficult to account for, but are incorporated in hazard models by constructing the covariate vector as a function of time. The hazard and likelihood functions are then appropriately rewritten. The likelihood function becomes more complex, but estimation is typically simplified because time-varying covariates usually make only a few discrete changes over the duration being studied. However, the interpretation of findings and, specifically duration effects, is much more difficult. The interested reader should see Peterson (1976), Cox and Oakes (1984), and Greene (2000) for additional information on the problems and solutions associated with time-varying covariates.

10.10 Discrete-Time Hazard Models

An alternative to standard continuous-time hazard models is to use a discrete-time approach. In discrete-time models, time is segmented into uniform discrete categories and exit probabilities in each time period are estimated using a logistic regression or other discrete outcome modeling approach (see Chapters 12 and 13). For example, for developing a model of commuters' delay in work-to-home departure to avoid traffic congestion, one could categorize

the observed continuous values into 30-minute time intervals. Then, a logistic regression could be estimated to determine the exit probabilities (probability of leaving work for home) in each of these time periods. This discrete approach allows for a general form of the hazard function (and duration effects) that can change from one time interval to the next as shown in Figure 10.7. However, the hazard is implicitly assumed to be constant during each of the discrete-time intervals.

Discrete hazard models have the obvious drawback of inefficient parameter estimators because of information lost in discretizing continuous data. However, when the discrete-time intervals are short, and the probability of exiting in any discrete-time interval is small, studies have shown that discrete hazard models produce results similar to those of continuous-time hazard models (Abbot 1985; Green and Symons 1983; Ingram and Kleinman 1989).

Although discrete-time hazard models are inefficient, they provide at least two advantages over their continuous-time counterparts. First, time-varying covariates are easily handled by incorporating new, time-varying values in the discrete-time intervals (there is still the assumption that they are constant over the discrete-time interval). Second, tied data (groups of observations exiting a duration at the same time), which, as discussed previously, are problematic when using continuous data approaches, are not a problem with discrete hazard models. This lack of difficulty is because grouped data fall within a single discrete-time interval and the subsequent lack of information

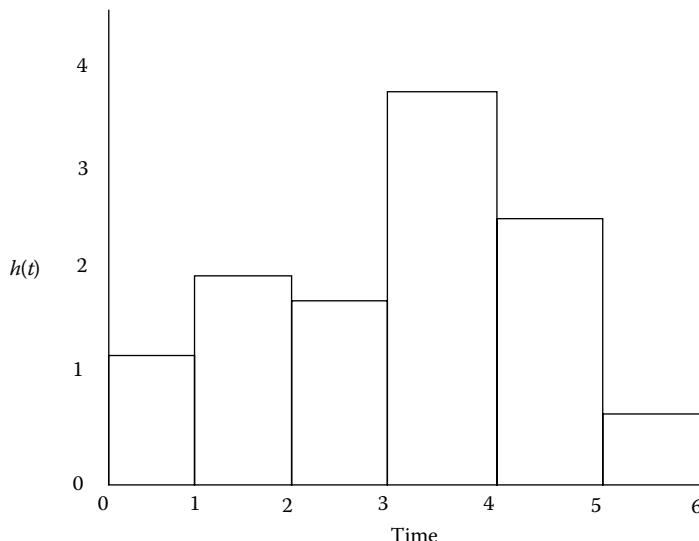


FIGURE 10.7

Illustration of hazard function using discrete data analysis.

on “exact” exit times does not complicate model estimation. Discrete-time hazard models continue to be a viable alternative to continuous-time approaches because problems associated with tied data and time-varying covariates are sometimes more serious than the efficiency losses encountered when discretizing the data.

Advances continue to be made by a number of researchers in the field of discrete-time hazard models. For example, Han and Hausman (1990) developed a generalized discrete-time hazard approach with gamma heterogeneity that demonstrates the continuing promise of the approach.

10.11 Competing Risk Models

The normal assumption for hazard model analyses is that durations end as a result of a single event. For example, in the commuters’ delay example discussed previously, the delay duration ends when the commuter leaves for home. However, multiple outcomes that produce different durations can exist. For example, if the commuter delay problem is generalized to consider any delay in the trip from work to avoid traffic congestion (not just the work-to-home trip option), then multiple duration-ending outcomes would exist. These are referred to as “competing risks” and could include work-to-home, work-to-shopping, and work-to-recreation trips as options that would end commuters’ delay at work.

The most restrictive approach to competing risk models is to assume independence among the possible risk outcomes. If this assumption is made, separate hazard models for each possible outcome are often estimated (see Gilbert 1992; Katz 1986). This approach is limiting because it ignores the interdependence among risks that are often present and of critical interest in the analysis of duration data. Accounting for interdependence among competing risks is not easily done due to the resulting complexity in the likelihood function. However, Diamond and Hausman (1984) have developed a model based on strict parametric assumptions about the interdependence of risks. In subsequent work, Han and Hausman (1990) extend this approach by providing a more flexible form of parametric interdependence that also permits one to statistically test for independence among competing risks. More recent work has dealt with a number of aspects relating to competing risk in the context of transportation-related problems. This research includes the work of Bhat (1996), Hensher (1998), Yamamoto et al. (1999), and Yamamoto et al. (2004).

Part III

Count and Discrete Dependent Variable Models

11

Count Data Models

Count data consist of nonnegative integer values and are encountered frequently in the modeling of transportation-related phenomenon. Examples of count data variables in transportation include the number of driver route changes per week, the number of trip departure changes per week, drivers' frequency of use of Intelligent Transportation System technologies over some time period, number of vehicles waiting in a queue, and the number of accidents observed on road segments per year. A common mistake is to model count data as continuous data by applying standard least squares regression. This approach is not strictly correct because regression models yield predicted values that are nonintegers and can also predict values that are negative, both of which are inconsistent with count data. These limitations make standard regression analysis inappropriate for modeling count data without modifying dependent variable.

Count data are properly modeled with a number of methods, the most popular of which are Poisson and negative binomial regression models. Poisson regression is the more popular of the two, and is applied to a wide range of transportation count data. The Poisson distribution approximates rare-event count data, such as accident occurrence, failures in manufacturing or processing, and number of vehicles waiting in a queue. One requirement of the Poisson distribution is that the mean of the count process equals its variance. When the variance is significantly larger than the mean, the data are said to be overdispersed. There are numerous reasons for overdispersion, some of which are discussed later in this chapter. In many cases, overdispersed count data are successfully modeled using negative binomial model.

11.1 Poisson Regression Model

To help illustrate the principle elements of a Poisson regression model, consider the number of accidents occurring per year at various intersections in a city. In a Poisson regression model, the probability of intersection i having y_i accidents per year (where y_i is a nonnegative integer) is given by

$$P(y_i) = \frac{\text{EXP}(-\lambda_i)\lambda_i^{y_i}}{y_i!} \quad (11.1)$$

where $P(y_i)$ is the probability of intersection i having y_i accidents per year and λ_i is the Poisson parameter for intersection i , which is equal to intersection i 's expected number of accidents per year, $E[y_i]$. Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of events per period) as a function of explanatory variables. For the intersection–accident example, explanatory variables might include intersections' geometric conditions, signalization, pavement types, visibility, and so on. The most common relationship between explanatory variables and the Poisson parameter is the log-linear model,

$$\lambda_i = \text{EXP}(\boldsymbol{\beta}\mathbf{X}_i) \text{ or, equivalently } \text{LN}(\lambda_i) = \boldsymbol{\beta}\mathbf{X}_i \quad (11.2)$$

where \mathbf{X}_i is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters. In this formulation, the expected number of events per period is given by $E[y_i] = \lambda_i = \text{EXP}(\boldsymbol{\beta}\mathbf{X}_i)$. This model is estimable by standard maximum likelihood methods, with the likelihood function given as

$$L(\boldsymbol{\beta}) = \prod_i \frac{\text{EXP}[-\text{EXP}(\boldsymbol{\beta}\mathbf{X}_i)][\text{EXP}(\boldsymbol{\beta}\mathbf{X}_i)]^{y_i}}{y_i!} \quad (11.3)$$

The log of the likelihood function is simpler to manipulate and more appropriate for estimation, and is given as

$$\text{LL}(\boldsymbol{\beta}) = \sum_{i=1}^n [-\text{EXP}(\boldsymbol{\beta}\mathbf{X}_i) + y_i \boldsymbol{\beta}\mathbf{X}_i - \text{LN}(y_i!)] \quad (11.4)$$

As with most statistical models, the estimated parameters are used to make inferences about the unknown population characteristics thought to impact the count process. Maximum likelihood estimates produce Poisson parameters that are consistent, asymptotically normal and asymptotically efficient.

11.2 Interpretation of Variables in the Poisson Regression Model

To provide some insight into the implications of parameter estimation results, elasticities are computed to determine the effects of the independent variables. Elasticities provide an estimate of the impact of a variable on the expected frequency and are interpreted as the effect of a 1% change in the variable on the expected frequency λ_i . For example, an elasticity of -1.32 is interpreted as a 1% increase in the variable reduces the expected

frequency by 1.32%. Elasticities are the correct way of evaluating the relative impact of each variable in the model. Elasticity of frequency λ_i is defined as

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik} \quad (11.5)$$

where E represents the elasticity, x_{ik} is the value of the k^{th} independent variable for observation i , β_k is the estimated parameter for the k^{th} independent variable, and λ_i is the expected frequency for observation i . Note that elasticities are computed for each observation i . It is common to report a single elasticity as the average elasticity over all i .

The elasticity in Equation 11.5 is only appropriate for continuous variables such as highway lane width, distance from outside shoulder edge to roadside features, and vertical curve length. It is not valid for noncontinuous variables such as indicator variables that take on values of zero or one. For indicator variables, a pseudoelasticity is computed to estimate an approximate elasticity of the variables. The pseudoelasticity gives the incremental change in frequency caused by changes in the indicator variables. The pseudoelasticity, for indicator variables, is computed as

$$E_{x_{ik}}^{\lambda_i} = \frac{\text{EXP}(\beta_k) - 1}{\text{EXP}(\beta_k)} \quad (11.6)$$

Another way to interpret the effect of specific variables is to calculate marginal effects. Unlike elasticities, which measure the effect that a 1% change in a variable x has on the dependent variable, marginal effects reflect the effect of a “one unit” change in x on the dependent variable (in the Poisson regression the dependent variable is λ_i). Marginal effects are computed as

$$ME_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\partial x_{ik}} = \beta_k \text{EXP}(\beta_k \mathbf{X}_i) \quad (11.7)$$

Marginal effects are often more easily interpreted than elasticities, particularly with regard to the effect of variables that indicator variables (taking on values of zero or one) and some integer variables. For example, if one were considering the effect that the number of children in the family has on the number of shopping trips made per month, marginal effects would give the effect of an additional child (which would make sense) whereas elasticities would give the effect of a 1% increase in the child variable (which would need additional interpretation). However, because elasticities and marginal effects both determine the impact of specific variables, it is common to report one or the other—but not both because they are redundant.

Finally, note that the Poisson probabilities for observation i are calculated using the following recursive formulas:

$$\begin{aligned} P_{0,i} &= \text{EXP}(-\lambda_i) \\ P_{j,i} &= \left(\frac{\lambda_i}{j} \right) P_{j-1,i}, \quad j = 1, 2, 3, \dots; \quad i = 1, 2, \dots, n \end{aligned} \quad (11.8)$$

where $P_{0,i}$ is the probability that observation i experiences 0 events in the specified observation period.

11.3 Poisson Regression Model Goodness-of-Fit Measures

There are numerous goodness-of-fit statistics used to assess the overall fit of the Poisson regression model. As mentioned in previous chapters, when selecting among alternate models, goodness-of-fit statistics should be considered along with model plausibility and agreement with expectations.

The likelihood ratio test is a common test used to assess two competing models. It provides evidence in support of one model, usually a full or complete model, over another competing model that is restricted by having a reduced number of model parameters. The likelihood ratio test statistic is

$$X^2 = -2[LL(\boldsymbol{\beta}_R) - LL(\boldsymbol{\beta}_U)] \quad (11.9)$$

where $LL(\boldsymbol{\beta}_R)$ is the log-likelihood at convergence of the “restricted” model (sometimes considered to have all parameters in $\boldsymbol{\beta}$ equal to 0, or just to include the constant term, to test overall fit of the model) and $LL(\boldsymbol{\beta}_U)$ is the log-likelihood at convergence of the unrestricted model. The X^2 statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of parameters in the restricted and unrestricted model (the difference in the number of parameters in the $\boldsymbol{\beta}_R$ and the $\boldsymbol{\beta}_U$ parameter vectors).

The sum of model deviances G^2 is equal to zero for a model with perfect fit. Note, however, that because observed y_i is an integer, while the predicted expected value $\hat{\lambda}_i$ is continuous, a G^2 equal to zero is a theoretical lower bound. This statistic is given as

$$G^2 = 2 \sum_{i=1}^n y_i \ln \left(\frac{y_i}{\hat{\lambda}_i} \right) \quad (11.10)$$

An equivalent measure to R^2 in ordinary least squares linear regression is not available for a Poisson regression model due to the nonlinearity of the conditional mean ($E[y|X]$) and heteroscedasticity in the regression. A similar statistic is based on standardized residuals

$$R_p^2 = 1 - \frac{\sum_{i=1}^n \left[(y_i - \hat{\lambda}_i) / \sqrt{\hat{\lambda}_i} \right]^2}{\sum_{i=1}^n \left[(y_i - \bar{y}) / \sqrt{\bar{y}} \right]^2} \quad (11.11)$$

Where the numerator is similar to an error or residual sum of squares and the denominator is similar to a total sum of squares.

Another measure of overall model fit is the ρ^2 statistic (also sometimes referred to as the McFadden ρ^2). The ρ^2 statistic is

$$\rho^2 = 1 - \frac{LL(\boldsymbol{\beta})}{LL(\mathbf{0})} \quad (11.12)$$

where $LL(\boldsymbol{\beta})$ is the log-likelihood at convergence with parameter vector $\boldsymbol{\beta}$ and $LL(\mathbf{0})$ is the initial log-likelihood (with all parameters set to zero). A perfect model has a likelihood function equal to one (all selected alternative outcomes are predicted by the model with probability one, and the product of these across the observations is also one) and the log-likelihood is zero yielding a ρ^2 of one (see Equation 11.12). Thus the ρ^2 statistic is between zero and one, while the closer it is to one the more variance the estimated model is explaining. Additional goodness-of-fit measures for the Poisson regression model are found in Cameron and Windmeijer (1993), Gurmu and Trivedi (1994), and Greene (1995b).

Example 11.1

Accident data from California (1993–1998) and Michigan (1993–1997) were collected (Vogt 1999 and Vogt and Bared 1998). The data represent a culled data set from the original studies, which included data from four states across numerous time periods and over five different intersection types. A reduced set of explanatory variables is used for injury accidents on 3-legged stop controlled intersections with two lanes on the minor and four lanes on the major road. The accident data are thought to be approximately Poisson or negative binomial distributed, as suggested by previous studies on the subject (Harwood et al. 2000; Miaou 1994; Miaou and Lum 1993; Milton and Mannering 1998; Poch and Mannering 1996; and Shankar et al. 1995). The variables in the study are summarized in Table 11.1.

Table 11.2 shows the parameter estimates of a Poisson regression estimated on the accident data. This model contains a constant and four variables; two average annual daily traffic (AADT) variables, median width, and number of driveways. The mainline

TABLE 11.1

Summary of Variables in California and Michigan Accident Data

Variable Description	Maximum/ Minimum Values	Mean of Observations	Standard Deviation of Observations
Indicator variable for state: (0 = California; 1 = Michigan)	1/0	0.29	0.45
Count of injury accidents over observation period	13/0	2.62	3.36
Average annual daily traffic on the major road	33,058/2,367	12870	6798
Average annual daily traffic on the minor road	3,001/15	596	679
Median width on the major road in feet	36/0	3.74	6.06
Number of driveways within 250 feet of intersection center	15/0	3.10	3.90

TABLE 11.2

Poisson Regression of Injury Accident Data

Variable Description	Estimated Parameter	t-Statistic
Constant	-0.826	-3.57
Average annual daily traffic on major road	0.00000812	6.90
Average annual daily traffic on minor road	0.000550	7.38
Median width in feet	-0.0600	-2.73
Number of driveways with 250 feet of intersection	0.0748	4.54
Number of observations	84	
Restricted log-likelihood (constant term only)	-246.18	
Log-likelihood at convergence	-169.26	
Chi-squared (associated p-value)	153.85 (<.0000001)	
R _p -Squared	0.4792	
G ²	176.5	

AADT appears to have a smaller influence than the minor road AADT, contrary to what is expected. Also, as median width increases, accidents decrease. Finally, the number of driveways close to the intersection increases the number of intersection injury accidents. The signs of the estimated parameters are in line with expectation.

The mathematical expression for this Poisson regression model is as follows,

$$\begin{aligned}
 E[y_i] &= \lambda_i = \text{EXP}(\beta \mathbf{x}_i) \\
 &= \text{EXP}(-0.83 + 0.000008(\text{AADT1}_i) + 0.0005(\text{AADT2}_i) \\
 &\quad - 0.06(\text{MEDIAN}_i) + 0.07(\text{DRIVE}_i))
 \end{aligned}$$

where $AADT_1$ is the AADT on the major road for intersection i , $AADT_2$ is the AADT on the minor road for intersection i , $MEDIAN_i$ is the median width on the major road in feet for intersection i , and $DRIVE_i$ is number of driveways within 250 feet of intersection center for intersection i . The model parameters are additive in the exponent, or multiplicative on the expected value of y_i . As in a linear regression model, standard errors of the estimated parameters are provided, along with approximate t values, and p -values associated with the null hypothesis of zero effect. In this case, the results show all estimated model parameters to be statistically significant beyond the 0.01 level of significance.

Example 11.2

Inspection of the output shown in Table 11.2 reveals numerous properties of the fitted model. The value of the log-likelihood function for the fitted model is -169.25 , whereas the restricted log-likelihood is -246.18 . The restricted or reduced log-likelihood is associated with a model with the constant term only. A likelihood ratio test comparing the fitted model and the reduced model results in $X^2 = 153.85$, which is sufficient to reject the fit of the reduced model. Thus it is extremely unlikely (p -value less than .0000001) that randomness alone would produce the observed decrease in the log-likelihood function.

Note that G^2 is 186.48, which is only relevant in comparison to other competing models. The R_p^2 is 0.48, which again serves as a comparison to competing models. A model with higher X^2 , lower G^2 , and a higher R_p^2 is sought in addition to a more appealing set of individual predictor variables, model specification, and agreement with expectation and theory.

Table 11.3 shows the average elasticities computed for the variables in the model shown in Table 11.2. These elasticities are obtained by enumerating through the sample (applying Equation 11.5 to all observations) and computing the average of these elasticities.

Poisson regression is a powerful analysis tool but, as with all statistical methods, it is used inappropriately if its limitations are not fully understood. There are three common analysis errors (Lee and Mannering 2002). The first is the failure to recognize that data are truncated. The second is the failure to recognize that the mean and variance are not equal, as required by the Poisson distribution. The third is the failure to recognize that the data contain a preponderance of zeros (zero accidents). These limitations and their remedies are now discussed.

TABLE 11.3

Average Elasticities for the Poisson Regression Model Shown in Table 11.2

Variable Description	Elasticity
Average annual daily traffic on major road	1.045
Average annual daily traffic on minor road	0.327
Median width in feet	-0.228
Number of driveways with 250 feet of intersection	0.232

11.4 Truncated Poisson Regression Model

Truncation of data can occur in the routine collection of transportation data. For example, if the number of times per week a commuter changes their departure time from work (on their work-to-home commute) in response to traffic congestion reports during weekdays, the data are right truncated at 5, which is the maximum number of departure-time changes in any given week. Estimating a Poisson regression model without accounting for this truncation results in biased estimates of the parameter vector β , and erroneous inferences are drawn. Fortunately, the Poisson model is adapted easily to account for such truncation. The right-truncated Poisson model is written as (see Johnson and Kotz 1970):

$$P(y_i) = \frac{[\lambda_i^{y_i} / y_i!]}{\left[\sum_{m_i=0}^r (\lambda_i^{m_i} / m_i!) \right]} \quad (11.13)$$

where $P(y_i)$ is the probability of commuter i making y_i departure changes per week, λ_i is the Poisson parameter for commuter i , m_i is the number of departure changes per week, and r is the right truncation (in this case, 5 times per week).

Example 11.3

Consider the sample of Seattle commuters used in Example 10.1. Although data were obtained from 204 commuters from the travel survey conducted in the Seattle metropolitan area (see Mannering and Hamed 1990a, only 96 of these commuters indicated that they ever delayed their work-to-home trip departure time. For these 96 commuters, a model of the number of times they changed their departure time on their work-to-home trip to avoid traffic congestion in the last week is developed. There are likely to be a significant number of zeroes in the data because, even though everyone in the 96-commuter sample indicated they sometimes delay, some of the respondents may not have delayed during the last week of the survey. These data are nonnegative integers and are thus well suited to the Poisson regression approach. However, because there are only 5 work days in the work week, the data are truncated.

The variables available for model estimation are shown in Table 11.4. In terms of summary statistics for the dependent variable, the mean number of departure-time changes per week is 1.833 with a variance of 1.877 (standard deviation of 1.37). These numbers are similar and will most certainly satisfy the Poisson assumption of equal mean and variance (see additional detail in the next section).

Model estimation results are given in Table 11.5 and corresponding average marginal effects are given in Table 11.6. The results show that commuters whose work-to-home routes included State Route 520 and Interstate 5 were less likely to delay their departure times. The marginal effects show that State Route 520 users have an average mean departure-time changes per week that is 0.874 lower

TABLE 11.4

Variables Available to Model the Number of Departure-Time Changes Per Week

Variable No.	Variable Description
1	Household number
2	Do you ever delay work-to-home departure to avoid traffic congestion? 1 if yes, 0 if no
3	If sometimes delay, on average how many minutes do you delay?
4	If sometimes delay, do you 1 if perform additional work, 2 if engage in nonwork activities, or 3 do both?
5	If sometimes delay, how many times have you delayed in the past week?
6	Mode of transportation used work-to-home: 1 if car, 2 if carpool, 3 if vanpool, 4 if bus, 5 if other.
7	Primary route (work-to-home): 1 if I-90, 2 if I-5, 3 if SR-520, 4 if I-405, 5 if other
8	Do you generally encounter traffic congestion on your work-to-home trip? 1 if yes, 2 if no
9	Commuter age in years: 1 if less than 25, 2 if 26–30, 3 if 31–35, 4 if 26–40, 5 if 41–45, 6 if 46–50, 7 if greater than 50
10	Gender: 1 if male, 0 if female
11	Number of cars in household
12	Number of children in household
13	Annual household income (U.S. dollars): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if 50,000–59,999, 6 if greater than 60,000
14	Do you have flexible work hours? 1 if yes, 0 if no
15	Distance from work to home (in kilometers)
16	Commuter faces level-of-service D or worse? 1 if yes, 0 if no
17	Ratio of actual travel time to free-flow travel time
18	Population of work zone
19	Retail employment in work zone
20	Service employment in work zone
21	Size of work zone (in acres)

and Interstate 5 users are 0.582 lower. It was also found that the more cars in the household, the fewer the departure-time changes, with each additional car resulting in an average drop of a 0.191 in the mean departure-time changes per week. This result could reflect the possibility that households with a greater number of cars have less scheduling flexibility. Marginal effects show that commuters with flexible work hours had an average reduction of 0.613 in the mean departure-time changes per week. This finding is likely because commuters with flexible work hours adjust their arrival at work to avoid the afternoon traffic and thus are less likely to delay their work-to-home departure. It was also found that as the distance from work to home increased the mean departure-time changes per week decreased (by an average of 0.22 per kilometer). Finally, commuters that used car and commuters that were less than 31 years old were both found to have higher mean departure-time changes per week (by an average of 0.430 and 0.448, respectively as indicated in Table 11.6).

TABLE 11.5

Truncated Poisson Regression of the Number of Departure-Time Changes Per Week

Variable Description	Estimated Parameter	t-Statistic
Constant	-0.931	-2.37
State Route 520 indicator (1 if primary work-to-home route includes SR-520, 0 otherwise)	-0.567	-2.18
Interstate 5 indicator (1 if primary work-to-home route includes I-5, 0 otherwise)	-0.377	-1.99
Number of cars in household	-0.124	-1.47
Flexible work hours indicator (1 if commuter has flexible work hours, 0 otherwise)	-0.397	-2.22
Distance from work to home (in kilometers)	-0.0142	-1.20
Car indicator (1 if commuter uses a car, 0 otherwise)	0.279	1.42
Young commuter indicator (1 if the commuter is less than 31 years old, 0 otherwise)	0.291	1.70
Number of observations	96	
Restricted Log-likelihood (constant term only)	-160.56	
Log-likelihood at convergence	-149.02	
Chi-squared and associated p-value	23.09	<.00164

TABLE 11.6

Average Marginal Effects for the Truncated Poisson Regression Model Shown in Table 11.5

Variable Description	Marginal Effect
State Route 520 indicator (1 if primary work-to-home route includes SR-520, 0 otherwise)	-0.874
Interstate 5 indicator (1 if primary work-to-home route includes I-5, 0 otherwise)	-0.582
Number of cars in household	-0.191
Flexible work hours indicator (1 if commuter has flexible work hours, 0 otherwise)	-0.613
Distance from work to home (in kilometers)	-0.022
Car indicator (1 if commuter uses a car, 0 otherwise)	0.430
Young commuter indicator (1 if the commuter is less than 31 years old, 0 otherwise)	0.448

11.5 Negative Binomial Regression Model

A common analysis error is a result of failing to satisfy the property of the Poisson distribution that restricts the mean and variance to be equal, when $E[y_i] = \text{VAR}[y_i]$. If this equality does not hold, the data are said to be under

dispersed ($E[y_i] > VAR[y_i]$) or overdispersed ($E[y_i] < VAR[y_i]$), and the parameter vector is biased if corrective measures are not taken. Overdispersion can arise for a variety of reasons, depending on the phenomenon under investigation (for additional discussion and example see Karlaftis and Tarko 1998). The primary reason in many studies is that variables influencing the Poisson rate across observations have been omitted from the regression.

The negative binomial model is derived by rewriting Equation 11.2 such that, for each observation i

$$\lambda_i = EXP(\beta \mathbf{X}_i + \varepsilon_i) \quad (11.14)$$

where $EXP(\varepsilon_i)$ is a Gamma-distributed disturbance term with mean 1 and variance α . The addition of this term allows the variance to differ from the mean as shown below

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \quad (11.15)$$

The Poisson regression model is regarded as a limiting model of the negative binomial regression model as α approaches zero, which means that the selection between these two models is dependent upon the value of α . The parameter α is often referred to as the overdispersion parameter. The negative binomial distribution has the form

$$P(y_i) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i} \quad (11.16)$$

where $\Gamma(\cdot)$ is a gamma function. This formulation results in the likelihood function

$$L(\lambda_i) = \prod_i \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i} \right)^{y_i} \quad (11.17)$$

When the data are overdispersed, the estimated variance term is larger than one would expect under a true Poisson process. As overdispersion gets larger, so does the estimated variance, and consequently all of the standard errors of parameter estimates become inflated.

A test for overdispersion is provided by Cameron and Trivedi (1990) based on the assumption that under the Poisson model, $(y_i - E[y_i])^2 - E[y_i]$ has mean zero, where $E[y_i]$ is the predicted count $\hat{\lambda}_i$. Thus, null and alternative hypotheses are generated such that

$$H_0: VAR[y_i] = E[y_i]$$

$$H_A: VAR[y_i] = E[y_i] + \alpha g(E[y_i]).$$

where $g(E[y_i])$ is a function of the predicted counts that is most often given values of $g(E[y_i]) = E[y_i]$ or $g(E[y_i]) = E[y_i]^2$. To conduct this test, a simple linear regression is estimated where Z_i is regressed on W_i , where,

$$\begin{aligned} Z_i &= \frac{(y_i - E(y_i))^2 - y_i}{E(y_i)\sqrt{2}} \\ W_i &= \frac{g(E(y_i))}{\sqrt{2}} \end{aligned} \quad (11.18)$$

After running the regression ($Z_i = bW_i$) with $g(E[y_i]) = E[y_i]$ and $g(E[y_i]) = E[y_i]^2$, if b is statistically significant in both cases, then H_0 is rejected for the particular function g .

Example 11.5

Consider the truncated Poisson regression of Example 11.3. Recall that, for this example, the mean number of departure-time changes per week is 1.833 with a variance of 1.877 (standard deviation of 1.37). Because the mean and variance are numerically close the Poisson regression model is likely to be suitable. Supporting this assertion, when a truncated negative binomial model is estimated using the data described in Table 11.4, the resulting estimate of α (see Equation 11.15) is 0.00000898 with a standard error of 0.1093, which yields a t -statistic of 0.000082. Since this value of α is not significantly different from zero the Poisson model is appropriate (the differences in the log-likelihoods at convergence between the Poisson and negative binomial models in this case is also virtually zero, again reflecting the negative binomial formulation is not statistically justified).

Example 11.6

Consider again the Poisson regression model on injury accidents estimated previously. Even though the model presented in Example 11.2 appears to fit the data fairly well, overdispersion might be expected. Using the regression test proposed by Cameron and Trivedi, with $g(E[y_i]) = E[y_i]$, $b = 1.28$ with a t -statistic = 2.75, and with $g(E[y_i]) = E[y_i]^2$, $b = 0.28$ with a t -statistic = 2.07. Based on the model output, both cases are statistically significant at the 5% level of significance, or 95% level of confidence. This result suggests that random sampling does not satisfactorily explain the magnitude of the overdispersion parameter, and a Poisson model is rejected in favor of a negative binomial model.

TABLE 11.7

Negative Binomial Regression of Injury Accident Data

Variable Description	Estimated Parameter	t-Statistic
Constant	-0.931	-2.37
Average annual daily traffic on major road	0.0000900	3.47
Average annual daily traffic on minor road	0.000610	3.09
Median width in feet	-0.0670	-1.99
Number of driveways with 250 feet of intersection	0.0632	2.24
Overdispersion parameter α	0.516	3.09
Number of observations	84	
Restricted Log-likelihood (constant term only)	-169.26	
Log-likelihood at convergence	-153.28	
Chi-squared and associated p-value	31.95	<.0000001

Example 11.7

Having evidence that overdispersion is present, a negative binomial model is estimated using the accident data. The results of the estimated negative binomial regression model are shown in Table 11.7 in Example 11.1.

As with the Poisson regression model, the signs of the estimated parameters are expected and are significant. In addition, the overdispersion parameter is statistically significant, confirming that the variance is larger than the mean. The restricted log-likelihood test suggests that the fitted model is better than a model with only the constant term.

11.6 Zero-Inflated Poisson and Negative Binomial Regression Models

There are certain phenomena where an observation of zero events during a given time period can arise from two qualitatively different conditions. One condition may result from simply failing to observe an event during the observation period. Another qualitatively different condition may result from an inability to ever experience an event. Consider the following example. A transportation survey asks how many times you have taken mass transit to work during the past week. An observed zero could arise in two distinct ways. First, last week you may have opted to take the vanpool instead of mass transit. Alternatively, you may never take transit, as a result of other commitments on the way to and from your place of employment. Thus two states are present, one being a normal count-process state and the other being a zero-count state.

At times what constitutes a zero-count state may be less clear. Consider vehicle accidents occurring per year on 1-kilometer sections of highway. For straight sections of roadway with wide lanes, low traffic volumes, and no roadside objects, the likelihood of a vehicle accident occurring may be extremely small, but still present because an extreme human error could cause an accident. These sections are considered to be in a zero-accident state because the likelihood of an accident is so small (perhaps the expectation is that a reported accident occurs once in a 100-year period). Thus, the zero-count state may refer to situations where the likelihood of an event occurring is extremely rare in comparison to the normal-count state where event occurrence is inevitable and follows some known count process (see Lambert 1992). Two aspects of this nonqualitative distinction of the zero state are noteworthy. First, there is a preponderance of zeroes in the data—more than is expected under a Poisson process. Second, a sampling unit is not required to be in the zero or near zero state into perpetuity, and can move from the zero or near zero state to the normal-count state with positive probability. Thus, the zero or near zero state reflects one of negligible probability compared to the normal state. Regardless, the interpretation of such models must be conducted with extreme care and their application carefully considered (see Lord et al. 2004).

Data obtained from two-state regimes (normal-count and zero-count states) often suffer from overdispersion if considered as part of a single, normal-count state because the number of zeroes is inflated by the zero-count state. Example applications in transportation are found in Miaou (1994), Shankar et al. (1997), and Malyshkina and Mannering (2009). It is common not to know if the observation is in the zero state—so the statistical analysis process must uncover the separation of the two states as part of the model estimation process. Models that account for this dual-state system are referred to as zero-inflated models (see Greene 2007; Lambert 1992; and Mullahey 1986).

To address phenomena with zero-inflated counting processes, the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models have been developed. The ZIP model assumes that the events, $Y = (y_1, y_2, \dots, y_n)$, are independent and the model is

$$\begin{aligned} y_i = 0 &\quad \text{with probability } p_i + (1 - p_i) \text{EXP}(-\lambda_i) \\ y_i = y &\quad \text{with probability } \frac{(1 - p_i) \text{EXP}(-\lambda_i) \lambda_i^y}{y!} \end{aligned} \quad (11.19)$$

where p_i is the probability of being in the zero state and y is the number of events per period. Maximum likelihood estimates are used to estimate the parameters of a ZIP regression model and confidence intervals are constructed by likelihood ratio tests.

The ZINB regression model follows a similar formulation with events, $\mathbf{Y} = (y_1, y_2, \dots, y_n)$, being independent and

$$\begin{aligned} y_i = 0 & \text{ with probability } p_i + (1-p_i) \left[\frac{1/\alpha}{(1/\alpha) + \lambda_i} \right]^{1/\alpha} \\ y_i = y & \text{ with probability } (1-p_i) \left[\frac{\Gamma((1/\alpha)+y)}{\Gamma(1/\alpha)y!} u_i^{1/\alpha} (1-u_i)^y \right], \quad y = 1, 2, 3, \dots \end{aligned} \tag{11.20}$$

where $u_i = (1/\alpha)/[(1/\alpha) + \lambda_i]$. Maximum likelihood methods are again used to estimate the parameters of a ZINB regression model.

Zero-inflated models imply that the underlying data-generating process has a splitting regime that provides for two types of zeros. The splitting process is assumed to follow a logit (logistic) or probit (normal) probability process, or other probability processes. A point to remember is that there must be underlying justification to believe the splitting process exists (resulting in two distinct states) before fitting this type of statistical model. There should be a basis for believing that part of the process is in a zero-count state.

To test the appropriateness of using a zero-inflated model rather than a traditional model, Vuong (1989) proposed a test statistic for nonnested models that is well suited for situations where the distributions (Poisson or negative binomial) are specified. The statistic is calculated as (for each observation i)

$$m_i = LN \left(\frac{f_1(y_i | \mathbf{x}_i)}{f_2(y_i | \mathbf{x}_i)} \right) \tag{11.21}$$

where $f_1(y_i | \mathbf{X}_i)$ is the probability density function of model 1, and $f_2(y_i | \mathbf{X}_i)$ is the probability density function of model 2. Using this approach, Vuongs' statistic for testing the nonnested hypothesis of model 1 versus model 2 is (Greene 2000; Shankar et al. 1997)

$$V = \frac{\sqrt{n} \left[(1/n) \sum_{i=1}^n m_i \right]}{\sqrt{(1/n) \sum_{i=1}^n (m_i - \bar{m})^2}} = \frac{\sqrt{n} (\bar{m})}{S_m} \tag{11.22}$$

where \bar{m} is the mean $((1/n) \sum_{i=1}^n m_i)$, S_m is standard deviation, and n is a sample size. Vuongs' value is asymptotically standard normal distributed (to be compared to z -values), and if $|V|$ is less than $V_{critical}$ (1.96 for a 95% confidence

level), the test does not support the selection of one model over another. Large positive values of V greater than $V_{critical}$ favor model 1 over model 2, whereas large negative values support model 2. For example, if comparing negative binomial alternatives, one would let $f_1(\cdot)$ be the density function of the ZINB and $f_2(\cdot)$ be the density function of the negative binomial model. In this case, assuming a 95% critical confidence level, if $V > 1.96$ the statistic favors the ZINB whereas a value of $V < -1.96$ favors the negative binomial. Values in between would mean that the test was inconclusive. The Vuong test can also be applied in an identical manner to test zero-inflated Poisson and Poisson models.

Because overdispersion will almost always include excess zeros, it is not always easy to determine whether excess zeros arise from true overdispersion, unobserved heterogeneity, or from an underlying splitting regime. The uncertain origin of zeroes could lead one to erroneously choose a negative binomial model when the correct model may be a ZIP. For example, recall from Equation 11.13 that the simple negative binomial gives

$$VAR[y_i] = E[y_i] [1 + \alpha E[y_i]] \quad (11.23)$$

For the ZIP model it is straightforward to show that

$$VAR[y_i] = E[y_i] + \left[1 + \frac{p_i}{1-p_i} E[y_i] \right] \quad (11.24)$$

Thus the term $p_i/(1-p_i)$ could be erroneously interpreted as α . To provide some guidance in unraveling this problem, consider the Vuong statistic comparing ZINB and negative binomial (with $f_1(\cdot)$ being the density function of the ZINB and $f_2(\cdot)$ being the density function of the negative binomial model for Equation 11.20). Shankar et al. (1997) provide model-selection guidelines for this case based on possible values of Vuong-test and overdispersion parameter (α) t -statistics. These guidelines, for the

TABLE 11.8

Decision Guidelines for Model Selection (Using the 95% Confidence Level) among Negative Binomial (NB), POISSON, Zero-Inflated POISSON (ZIP) and Zero-Inflated Negative Binomial (ZINB) Models Using the Vuong Statistic and the Overdispersion Parameter α

		<i>t</i> -Statistic of the NB Overdispersion Parameter α	
		< 1.96	> 1.96
Vuong statistic for ZINB($f_1(\cdot)$) and NB($f_2(\cdot)$) comparison		< -1.96	ZIP or Poisson as alternative to NB
		> 1.96	ZINB

95% confidence level, are presented in Table 11.8. Again, it must be stressed that great care must be taken in the selection of models. If there is not a compelling reason to suspect that a two states might be present, the use of a zero-inflated model may be simply capturing model misspecification that could result from factors such as unobserved effects (heterogeneity) in the data.

Example 11.6

Consider the sample of Seattle commuters used in Example 11.3. It is known that 108 of the 204 commuters in the Seattle metropolitan area sample (see Mannerling and Hamed 1990a) indicated that they never change their work-to-home trip departure time in response to congestion. And, of the 96 of these commuters that indicated that they do sometimes delay their trips, several of these did not change their normal departure time in the last week in response to congestion. When using the entire 204 commuters, a ZIP model is appropriate because 108 people are in the “zero-state,” while the remaining 96 follow a Poisson process with regard to the number of departure-time changes made in the last week (see Example 11.3).

Zero-inflated Poisson estimation results (with right truncation due to the maximum number of departure-time changes being 5) are presented in Table 11.9. The results show that commuters in the count state whose work-to-home routes included State Route 520 and Interstate 5 were less likely to delay their departure times (as was the case in Example 11.3). The results also show that commuters with household incomes less than 30,000 U.S. dollars per year were more likely to change their departure times and that those commuters with flexible work hours were less likely to change their departure times (again this is likely because commuters with flexible work hours adjust their arrival at work to avoid the afternoon traffic and thus are less likely to delay their work-to-home departure).

Turning to the factors that make commuters more or less likely to be in the zero state, the table shows that those commuters using a car and those facing level-of-service D or worse (extremely congested traffic conditions) are less likely to be in the zero state. Both of these make sense because they would both be associated with commuters that were more likely to change to avoid congestion and thus less likely to be in the zero state. Finally, the higher the ratio of actual travel time to free-flow travel time (the travel time when there is little traffic on the system), the lower the likelihood of being in the zero state. This variable again reflects the relative degree of congestion on the work-to-home trip and, as congestion increases, a commuter is less likely to be in the zero state.

Finally, the Young statistic for testing the ZIP versus the normal-count Poisson is computed as 8.95. The absolute value of the t-statistic for the negative binomial overdispersion parameter is less than 1.96 (see Example 11.5), while the 8.95 test values (which are greater than 1.96) in the lower quadrant of Table 11.8 suggest that the ZIP is the correct model.

TABLE 11.9

Zero-Inflated Truncated Poisson Regression of the Number of Departure-Time Changes Per Week: Logistic Distribution Splitting Model

Variable Description	Estimated Parameter	t-Statistic
<i>Poisson Departure-Time Change Count State</i>		
Constant	0.814	4.22
State Route 520 indicator (1 if primary work-to-home route includes SR-520, 0 otherwise)	-0.602	-2.52
Interstate 5 indicator (1 if primary work-to-home route includes I-5, 0 otherwise)	-0.295	-1.36
Low-income indicator (1 if commuter's household income is less than 30,000 U.S. dollars per year, 0 otherwise)	0.281	1.37
Flexible work hours indicator (1 if commuter has flexible work hours, 0 otherwise)	-0.263	-1.31
<i>Zero Departure-Time Change State</i>		
Constant	11.051	5.34
Car indicator (1 if commuter uses a car, 0 otherwise)	-1.111	-1.65
Level-of-service D indicator (1 if commuter faces level-of-service D or worse during work-to-home commute, 0 otherwise)	-2.941	-2.45
Ratio of actual travel time to free-flow travel time on work-to-home commute	-4.753	-3.29
Number of observations	204	
Log-likelihood at convergence (Poisson)	-288.62	
Log-likelihood at convergence (zero-inflated Poisson)	-187.18	
Vuong statistic for testing zero-inflated Poisson versus the normal-count Poisson model	8.95	

11.7 Random-Effects Count Models

In some cases, there may be reason to expect correlation among observations. This correlation could arise from spatial considerations (data from the same geographic region may share unobserved effects), temporal considerations (such as in panel data—where data collected in the same time period could share unobserved effects), or a combination of the two. To account for such correlation, random effects (where the common unobserved effects are assumed to be distributed over the spatial/temporal units according to some distribution and shared unobserved effects are assumed to be uncorrelated with independent variables) and fixed effects (where common unobserved

effects are accounted for by indicator variables and shared unobserved effects are assumed to be correlated with independent variables) models are considered. In the context of count models, Hausman et al. (1984) examined random effects and fixed effects negative binomial models for panel data (which has temporal considerations) in their study of and development patents.

For the purposes of demonstration, consider the random-effects approach that assumes that the shared unobserved effects are distributed across the population of temporal or spatial groups according to some predetermined distribution (an application of random effects applied to ordered probability models is provided in Section 14.2 of Chapter 14). To consider random effects in a count data model, Equation 11.2 is rewritten as

$$LN(\lambda_{ij}) = \boldsymbol{\beta} \mathbf{X}_{ij} + \eta_j \text{ or, equivalently } \lambda_{ij} = EXP(\boldsymbol{\beta} \mathbf{X}_{ij}) EXP(\eta_j) \quad (11.25)$$

where λ_{ij} is the expected number of events for observation i belonging to group j (e.g., a spatial or temporal group expected to share unobserved effects), \mathbf{X}_{ij} is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of estimable parameters, and η_j is a random effect for observation group j . This specification yields the Poisson model (see Equation 11.1) as

$$P(y_{ij} | \mathbf{X}_{ij}, \eta_j) = \frac{EXP[-EXP(\boldsymbol{\beta} \mathbf{X}_{ij}) EXP(\eta_j)] [EXP(\boldsymbol{\beta} \mathbf{X}_{ij}) EXP(\eta_j)]^{y_{ij}}}{y_{ij}!} \quad (11.26)$$

The most common model is derived by assuming η_j are assumed to be randomly distributed across groups such that $EXP(\eta_j)$ is Gamma-distributed with mean one and variance α (see Hausman et al. (1984) and note similarities with negative binomial model in Equation 10.13). Recall that the Poisson model restricts the mean and variance to be equal which in this case is $E[y_{ij}] = VAR[y_{ij}]$. However, with the random effects as in Equation 11.26 the Poisson variance to mean ration is now $1 + \lambda_{ij}/(1/\alpha)$.

Using the same approach as above, the random-effects negative binomial model can also be readily derived (Hausman et al. 1984). Transportation applications of fixed and random-effects Poisson and negative binomial models have been quite limited. However, two studies are noteworthy in this regard. Johansson (1996) studied the effect of a lowered speed limit on the number of accidents on roadways in Sweden and Shankar et al. (1998) compared standard negative binomial and random-effects negative binomial models in a study of accidents caused by median crossovers in Washington State. The reader is referred to these sources and Hausman et al. (1984) for additional information on random and fixed effects count models, as well as the more general random parameters count model described in Chapter 16 of this book.

12

Logistic Regression

The regression models described in Chapter 3 are derived with the assumption that the dependent variable is continuous—an interval or ratio scale variable. There are numerous cases when the dependent variable is discrete—count data, a series of qualitative rankings, and categorical outcomes are examples. For these cases, a wide variety of techniques is applied and is the subject of later chapters. This chapter addresses the modeling and analysis of data when the outcome variable is binary: whether a pavement has deteriorated beyond a certain threshold; whether a person walks to work; and whether fog is present (or not) at the time of a motor vehicle crash. When the intent is to model binary outcomes as a function of predictor variables, logistic regression is often an appropriate method. Logistic regression has been applied to model a wide variety of transportation data: White and Washington (2001) modeled safety restraint use as a function of law enforcement and other factors; Haider and Chatti (2009) examined fatigue cracking in flexible pavements; Lachapelle and Frank (2009) studied whether transit and car trips were associated with meeting recommended daily levels of physical activity; and Wong et al. (2009) examined risk factors associated with airport risk and accidents.

12.1 Principles of Logistic Regression

The goal of logistic regression, much like linear regression, is to identify a well fitting, defensible model that describes the relationship between a binary dependent variable and a set of independent or explanatory variables. As usual, an often unstated assumption is that the independent variables directly or indirectly influence the outcome, and that the independent variables are used to either explain or predict outcomes—depending upon the particular study objectives.

Before presenting the logistic regression modeling framework, it is worthwhile to present some basic principles useful in the development and interpretation of these models; odds, odds ratios, and the log of odds ratios, or log(odds).

Odds are related to probability but are conceptually and numerically different. Suppose that the probability that a crash at an intersection involves

a rear-end collision is $P = .1$. It follows that the probability that a crash does not involve a rear end collision is $1 - P = .9$.

Odds describe likelihoods of events. Odds are related to probability such that $O = P/(1 - P)$, so in this case $O = 0.1/0.9$. One would say, then, that the (conditional) odds that a crash is a rear-end collision 1 in 9 (1:9 or 1/9), or equivalently that the odds that a crash is not a rear-end collision is 9–1 (9:1 in 9/1). Equal probability binomial outcomes, such as flipping a fair coin, have $P = .5$ and odds of 1 to 1 (for every head outcome there is a tail outcome, on average). Odds approaching 0 correspond with low-probability events, whereas odds approaching infinity correspond with high-probability events. As a result, odds between 0 and 1 represent probabilities between 0 and .5, while odds between 1 and ∞ represent probabilities between .5 and 1. Obtaining P from O is straightforward, as $P = O/(O + 1)$; thus for a fair coin $P = (1/1)/((1/1) + 1) = 1/2 = 0.5$, or for rear-end crashes as described previously, $P = (1/9)/((1/9) + 1) = 1/9 = 0.1$.

Odds ratios are useful for comparing the likelihood of two events. Continuing from the previous example, one might ask which crash is more likely, a rear-end or sideswipe (suppose the odds of a sideswipe crash are 1 in 4)? Computing the odds ratio gives, $[1/9]/[1/4] = 4/9$ suggesting that the odds of a rear-end crash compared to a sideswipe crash are 4 in 9 (i.e., 4 rear-end crashes corresponds with the occurrence of 9 sideswipe crashes on average).

For analytical convenience the odds ratio is often scaled using a natural logarithmic transform, which gives the log of the odds ratio (*LN* odds ratio). The *LN* odds ratio is convenient analytically with probabilities between 0 and .5 corresponding to log odds ratios between $-\infty$ and 0, while *LN* odds ratios between 0 and ∞ correspond to probabilities between .5 and 1. The *LN* odds ratio scale is symmetric about zero just as probabilities are symmetric around .5 (1 unit above zero is comparable to one unit below zero, etc.) This convenient symmetry does not exist when using the odds ratio scale. As a result, for most modeling applications the *LN* odds ratio scale is often used.

12.2 Logistic Regression Model

For logistic regression, the dependent is the population proportion or probability (P) that the resulting outcome indicates the presence of a condition—usually denoted using a binary indicator variable coded as 1 or 0. Thus in a dataset one might code gender of a female respondent as 1, and 0 for males.

In developing the logistic regression equation, the LN of the odds represents a logit transformation, where the logit is a function of covariates such that

$$Y_i = \log it(P_i) = LN\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_K X_{K,i} \quad (12.1)$$

and where β_0 is the model constant and the β_1, \dots, β_K are the unknown parameters corresponding with the explanatory variables (X_k , $k = 1, \dots, K$ the set of independent variables). In Equation 12.1, the unknown binomial probabilities are a function of explanatory variables (which may include both continuous and discrete variables).

The unknown parameters in Equation 12.1 are typically estimated using maximum likelihood methods, which are discussed elsewhere in the text (see Section 3.2.2). These parameters, once estimated, are used to estimate the probability that the outcome takes the value 1 as a function of covariates using

$$P_i = \frac{EXP[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_K X_{K,i}]}{1 + EXP[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_K X_{K,i}]} \quad (12.2)$$

where all terms are as defined previously.

The logistic regression equation is approximately linear in the middle ranges and logarithmic at extreme values. A simple transformation of Equation 12.1 yields (with Greek betas replaced by their estimates)

$$\left(\frac{P_i}{1-P_i}\right) = EXP[\hat{\beta}_0 + \hat{\beta}_i X_i] = EXP[\hat{\beta}_0] EXP[\hat{\beta}_i X_i] \quad (12.3)$$

which shows that when the value of an explanatory variable increases by one unit, and all other variables are held constant, the probability ratio becomes

$$\begin{aligned} \left(\frac{P_i}{1-P_i}\right)^* &= EXP[\hat{\beta}_0] EXP[\hat{\beta}_i (X_i + 1)] \\ &= EXP[\hat{\beta}_0] EXP[\hat{\beta}_i X_i] EXP[\hat{\beta}_i] \\ &= \left(\frac{P_i}{1-P_i}\right) EXP[\hat{\beta}_i] \end{aligned} \quad (12.4)$$

Thus, an increase in the independent variable X_i by one unit (all other factors held constant, which is typically only possible when multicollinearity does not exist), the odds ($P_i/(1-P_i)$) increase by the factor $EXP[\hat{\beta}_i]$. The factor $EXP[\hat{\beta}_i]$ is the odds ratio and indicates the relative amount by which the odds of an outcome increases (odds ratio >1) or decreases (odds ratio <1) when the value of the corresponding independent variable increases by 1 unit.

Example 12.1

The Hoosier Service Patrol Program in Northwest Indiana was initiated in 1991 to patrol a 16-mile stretch of Interstate 80-94 known as the Borman Expressway and an 8-mile portion of Interstate 65. The patrol would look for, respond, and provide services to remediate incidents (vehicle crashes and disablements) along the corridors (for more detail see Karlaftis et al. 1999). After an assist was completed, the patrol person would record a variety of information on the driver and incident characteristics as shown in Table 12.1. Variables thought to be associated with the likelihood of a secondary incident (a crash or vehicle disablement that occurs in the queues created by the first incident, for example) were collected. Clearance times (the time it takes to clear the road of vehicles and debris) are generally positively associated with secondary incident likelihoods simply due to increased exposure. The type of vehicle is also important because specific types of vehicles (large trucks, passenger cars, etc.) may be more difficult to remove, may result in more closed lanes, and so on. Road position may help to indicate the severity of an incident, the stopping times required to avoid secondary incidents, and so on. Season of the year reflects aspects such as traction and visibility as well as probability of inclement weather conditions. Finally, weekday versus weekend travel may reflect the familiarity of drivers with ambient road conditions, attitudes and aggressiveness, and possibly vehicle mix.

A logistic regression model estimated on these data is shown in Table 12.2. Recall that the model is constructed to predict or explain the probability that a primary incident results in a secondary incident. Three explanatory variables were found to be statistically significant, the clearance time, season of winter, and whether the incident involved a van.

The probabilities of an incident leading to a secondary incident are estimated using Equation 12.2. For example, let us examine incidents not involving a van, occurring in summer, and taking 15, 30, and 60 minutes to clear (note that the indicator variables for van and winter are zero and so the corresponding coefficients are dropped from the model):

$$P_{15} = \frac{EXP[-1.224 + 0.027(15)]}{1 + EXP[-1.224 + 0.027(15)]} = 0.305$$

$$P_{30} = \frac{EXP[-1.224 + 0.027(30)]}{1 + EXP[-1.224 + 0.027(30)]} = 0.397$$

$$P_{60} = \frac{EXP[-1.224 + 0.027(60)]}{1 + EXP[-1.224 + 0.027(60)]} = 0.597$$

The probability is about 30% that an incident cleared in 15 minutes will result in a secondary incident, based on this logistic regression model. As the positive coefficient suggests, an increase in clearance time will increase the probability that a secondary incident occurs—with 30 and 60 minute clearance times resulting in about 40% and 60% probability of secondary incidents respectively.

TABLE 12.1

Hoosier Service Patrol Variables

Variable No.	Variable Description
1	Indicator variable for primary incident (0 = not linked to secondary incident; 1 = linked to secondary incident)
2	Clearance time of incident in minutes
3	Indicator variables for season of year (fall, winter, spring, summer)
4	Indicator variables for vehicle type (car, truck, semi, van, bus)
5	Road position at time of incident (left lane, right lane, right shoulder, and ramp)
6	Indicator variable for weekend travel
7	Indicator variable for AM or PM peak period travel

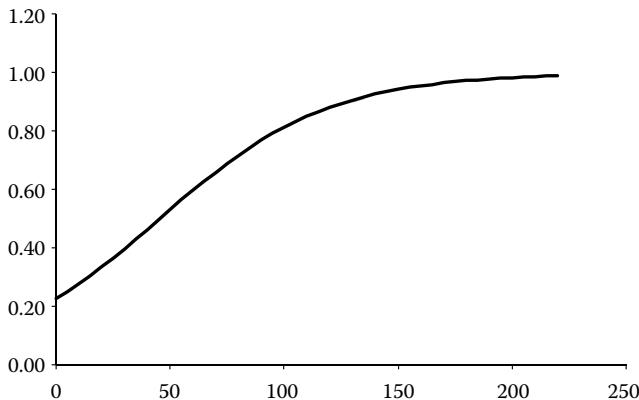
TABLE 12.2

Logistic Regression Model of Hoosier Service Patrol Data: Coefficients

Variable	Parameter Estimate	t-statistic	P > z	95% Coefficient Interval
Constant	-1.22	-8.23	.000	-1.516, -0.933
Clearance time of incident in minutes	0.027	6.35	.000	0.0187, 0.0354
Winter indicator variable (1 if incident occurred in the winter, 0 otherwise)	-0.417	-2.23	.026	-0.783, -0.051
Van indicator (1 if vehicle is a van, 0 otherwise)	-1.848	-2.48	.013	-3.309, -0.386
<i>Odds Ratio</i>				
Clearance time of incident in minutes	1.027412			1.019, 1.036
Winter indicator variable (1 if incident occurred in the winter, 0 otherwise)	0.6590597			0.457, 0.951
Van indicator (1 if vehicle is a van, 0 otherwise)	0.1575847			0.0365, 0.6795
Number of observations	741			
Log likelihood	-446.64			
Likelihood ratio chi-square (3 d.f.)	63.30			
Probability > chi-square	<.0001			
Pseudo R-squared	0.0662			

The probabilities are useful, but what are the odds ratios for the variables? As shown in Table 12.2, for each additional minute of clearance time, there is almost a 3% increase in the likelihood that a incident will involve a secondary incident. So, for example, a clearance time of 60 minutes increases the likelihood by almost 100% of a secondary incident compared to a clearance time of 30 minutes. An incident during winter decreases the likelihood of a secondary incident by about 35%, while an incident involving a van decreases a secondary incident by about 84%.

Logistic regression is an extremely useful tool for analyzing binary outcome data as a function of explanatory variables. Of course the logistic regression

**FIGURE 12.1**

Predicted probabilities of secondary incident: Hoosier logistic regression.

model is complicated rather quickly with random effects, panel data, and other anomalies. For example, suppose that the Hoosier Service Patrol Program staff became more efficient and applied more advanced methods over time to mitigate the impact of incidents—and thus months since beginning of service was thought to be important—requiring a panel approach to estimate the effect of time. Or, suppose it was believed that random effects influence the probability of secondary incidents as a result of unmeasured factors at the scene of each incident (e.g., interference by police, ambulance, and by fire is random across incidents). As one might expect there are methods for coping with such necessary model enhancements.

Finally, it is worth commenting on the form of the logistic distribution and model probabilities obtained from logistic regression models. Figure 12.1 shows a plot of the predicted probabilities of a secondary incident (from the Hoosier Service Patrol Program) as a function of clearance time. By the nature of the logit transformation, model predictions are asymptotic to 0 and 1. In this particular model, an incident that is cleared immediately upon arrival of the service patrol still presents a 20% probability of a secondary incident. This logit or logistic form is an artifact that is extremely useful because actual probabilities are constrained between 0 and 1 (exclusive). For example, a linear regression on probabilities would not constrain them properly.

There are other binary model forms other than the logistic that may be used to fit binary outcomes as a function of explanatory variables. The probit regression model is often an alternative to the logistic regression (using the logit transformation), and relies on the normal instead of the logistic distribution. Readers wishing greater depth on logistic regression models than provided here should consult Hilbe (2009) and Hosmer and Lemeshow (2000).

13

Discrete Outcome Models

Discrete or nominal scale data often play a dominant role in transportation because many interesting policy-sensitive analyses deal with such data. Examples of discrete data in transportation include the mode of travel (automobile, bus, rail transit), the type or class of vehicle owned, and the type of a vehicular accident (run-off-road, rear-end, head-on, etc.). From a conceptual perspective, such data are classified as those involving a behavioral choice (choice of mode or type of vehicle to own) or those simply describing discrete outcomes of a physical event (type of vehicle accident). The methodological approach used to statistically model these conceptual perspectives is often identical. However, the underlying theory used to derive these models is often quite different. Discrete models of behavioral choices are derived from economic theory, often leading to additional insights in the analysis of model estimation results, whereas models of physical phenomena are derived from simple probabilistic theory.

Some discrete data encountered in transportation applications are ordered. An example is telecommuting-frequency data that have outcomes of never, sometimes, and frequently. In contrast to data that are not ordered, ordinal discrete data possess additional information on the ordering of responses that is often used to improve the efficiency of the model's parameter estimates. For example, using the information that "never" is less than "sometimes" which is less than "frequently" can increase the efficiency of parameter estimates. Guidelines and considerations for choosing between ordered and unordered discrete modeling approaches are presented later in this chapter.

13.1 Models of Discrete Data

In formulating a statistical model for unordered discrete outcomes, it is common to start with a linear function of covariates that influences specific discrete outcomes. For example, in the event of a vehicular accident, possible discrete crash outcomes are rear-end, sideswipe, run-off-road, head-on, turning, and other. Let T_{in} be a linear function that determines discrete outcome i for observation n such that,

$$T_{in} = \boldsymbol{\beta}_i \mathbf{X}_{in} \quad (13.1)$$

where β_i is a vector of estimable parameters for discrete outcome i and X_{in} is a vector of the observable characteristics (covariates) that determine discrete outcomes for observation n . To arrive at a statistically estimable probabilistic model, a disturbance term ε_{in} is added, giving

$$T_{in} = \beta_i X_{in} + \varepsilon_{in} \quad (13.2)$$

The addition of the disturbance term is supported on a number of grounds such as the possibility that (1) variables have been omitted from Equation 13.1 (some important data may not be available), (2) the functional form of Equation 13.1 may be incorrectly specified (it may not be linear), (3) proxy variables may be used (variables that approximate missing variables in the database), and (4) variations in β_i that are not accounted for (β_i may vary across observations). As is shown later in this chapter, the existence of some of these possibilities can have adverse effects on the estimation of β_i , so caution must be exercised in the interpretation of ε_{in} .

To derive an estimable model of discrete outcomes with I denoting all possible outcomes for observation n , and $P_n(i)$ being the probability of observation n having discrete outcome i ($i \in I$)

$$P_n(i) = P(T_{in} \geq T_{In}) \quad \forall I \neq i \quad (13.3)$$

By substituting Equation 13.2 into Equation 13.3, the latter is written as

$$P_n(i) = P(\beta_i X_{in} + \varepsilon_{in} \geq \beta_I X_{In} + \varepsilon_{In}) \quad \forall I \neq i \quad (13.4)$$

or

$$P_n(i) = P(\beta_i X_n - \beta_I X_n \geq \varepsilon_{In} - \varepsilon_{in}) \quad \forall I \neq i \quad (13.5)$$

Estimable models are developed by assuming a distribution of the random disturbance term, ε . Two popular modeling approaches are probit and logit.

13.2 Binary and Multinomial Probit Models

A distinction is often drawn between binary models (models that consider two discrete outcomes) and multinomial models (models that consider 3 or more discrete outcomes) because the derivation between the two can vary significantly. Probit models arise when the disturbance term ε_{In} in Equation 13.5

is assumed to be normally distributed. In the binary case (two outcomes, denoted 1 or 2) Equation 13.5 is written as

$$P_n(1) = P(\boldsymbol{\beta}_1 \mathbf{X}_{1n} - \boldsymbol{\beta}_2 \mathbf{X}_{2n} \geq \varepsilon_{2n} - \varepsilon_{1n}) \quad (13.6)$$

This equation estimates the probability of outcome 1 occurring for observation n , where ε_{1n} and ε_{2n} are normally distributed with mean = 0, variances σ_1^2 and σ_2^2 , respectively, and the covariance is σ_{12} . An attractive feature of normally distributed variates is that the addition or subtraction of two normal variates also produces a normally distributed variate. In this case $\varepsilon_{2n} - \varepsilon_{1n}$ is normally distributed with mean zero and variance $\sigma_1^2 + \sigma_2^2 - \sigma_{12}^2$. The resulting cumulative normal function is

$$P_n(1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\boldsymbol{\beta}_1 \mathbf{X}_{1n} - \boldsymbol{\beta}_2 \mathbf{X}_{2n})/\sigma} \text{EXP}\left[-\frac{1}{2}w^2\right] dw \quad (13.7)$$

If $\Phi(\cdot)$ is the standardized cumulative normal distribution, then

$$P_n(1) = \Phi\left(\frac{\boldsymbol{\beta}_1 \mathbf{X}_{1n} - \boldsymbol{\beta}_2 \mathbf{X}_{2n}}{\sigma}\right) \quad (13.8)$$

where $\sigma = (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})^{0.5}$. The term $1/\sigma$ in Equations 13.7 and 13.8 is a scaling of the function determining the discrete outcome and can be set to any positive value, although $\sigma = 1$ is typically used. An example of the general shape of this probability function is shown in Figure 13.1.

The parameter vector $(\boldsymbol{\beta})$ is readily estimated using standard maximum likelihood methods. If δ_{in} is defined as being equal to 1 if the observed

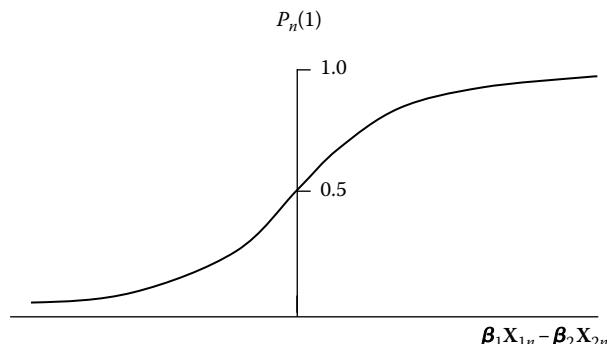


FIGURE 13.1

General shape of probit outcome probabilities.

discrete outcome for observation n is i and zero otherwise, the likelihood function is

$$L = \prod_{n=1}^N \prod_{i=1}^I P(i)^{\delta_{in}} \quad (13.9)$$

where N is the total number of observations. In the binary case with $i = 1$ or 2, the log-likelihood from Equation 13.9 is (again the log likelihood is used in estimation without loss of generality because the log transformation does not affect the ordering)

$$LL = \sum_{n=1}^N \left(\delta_{1n} \ln \Phi \left(\frac{\beta_1 X_{1n} - \beta_2 X_{2n}}{\sigma} \right) + (1 - \delta_{1n}) \ln \Phi \left(\frac{\beta_1 X_{1n} - \beta_2 X_{2n}}{\sigma} \right) \right) \quad (13.10)$$

The derivation of the multinomial probit model is provided in numerous sources, most notably Daganzo (1979). The problem with the multinomial probit is that the outcome probabilities are not closed form and estimation of the likelihood functions requires numerical integration. The difficulties of extending the probit formulation to more than two discrete outcomes have lead researchers to consider other disturbance term distributions.

13.3 Multinomial Logit Model

From a model estimation perspective, a desirable property of an assumed distribution of disturbances (ϵ 's in Equation 13.5) is that the maximums of randomly drawn values from the distribution have the same distribution as the values from which they were drawn. The normal distribution does not possess this property (the maximums of randomly drawn values from the normal distribution are not normally distributed). A disturbance term distribution with such a property greatly simplifies model estimation because Equation 13.10 could be applied to the multinomial case by replacing $\beta_2 X_{2n}$ with the highest value (maximum) of all other $\beta_i X_{in} \neq 1$. Distributions of the maximums of randomly drawn values from some underlying distribution are referred to as extreme value distributions (Gumbel 1958). Extreme value distributions are categorized into three families: Type 1, Type 2, and Type 3 (see Johnson and Kotz 1970). The most common extreme value distribution is the Type 1 distribution (sometimes referred to as the Gumbel distribution). It has the desirable property that maximums of randomly drawn values from the extreme value

Type 1 distribution are also extreme value Type 1 distributed. The probability density function of the extreme value Type 1 distribution is,

$$F(\varepsilon) = \eta \text{EXP}(-\eta(\varepsilon - \omega)) \text{EXP}(-\text{EXP}(-\eta(\varepsilon - \omega))) \quad (13.11)$$

with corresponding density function

$$F(\varepsilon) = \text{EXP}(-\text{EXP}(-\eta(\varepsilon - \omega))) \quad (13.12)$$

where η is a positive scale parameter, ω is a location parameter (mode), and the mean is $\omega + 0.5772/\eta$. An illustration of the extreme value Type 1 distribution with location parameter equal to zero and various scale parameter values is presented in Figure 13.2.

To derive an estimable model based on the extreme value Type 1 distribution, following Daniel McFadden's original derivations (McFadden 1978, 1981), a revised version of Equation 13.4 yields

$$P_n(i) = P(\beta_i X_{in} + \varepsilon_{in} \geq \max_{\forall l \neq i} (\beta_l X_{ln} + \varepsilon_{ln})) \quad (13.13)$$

For the extreme value Type 1 distribution, if all ε_{ln} are independently and identically (same variances) distributed random variates with modes ω_{ln}

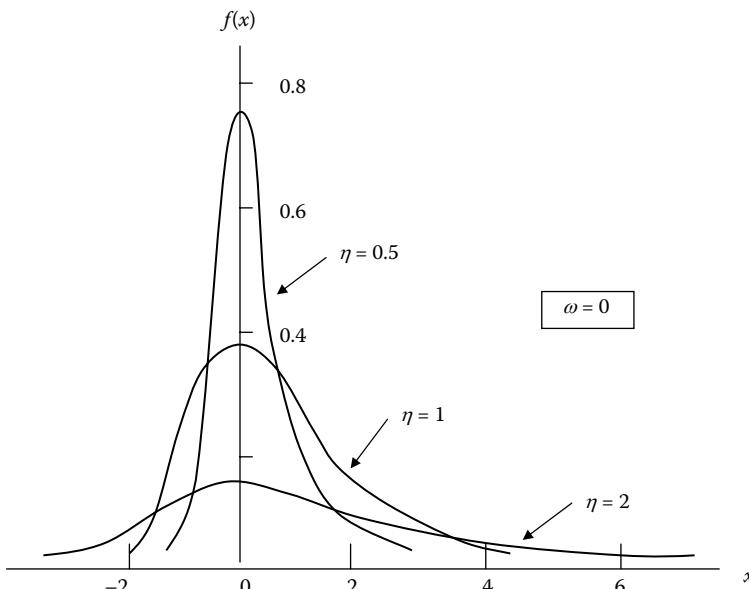


FIGURE 13.2

Illustration of an extreme value Type I distribution.

and a common scale parameter η (which implies equal variances), then the maximum of $\beta_i X_{in} + \varepsilon_{in}$ is extreme value Type 1 distributed with mode

$$\frac{1}{\eta} \ln \sum_{\forall I \neq i} \text{EXP}(\eta \beta_i X_{in}) \quad (13.14)$$

and scale parameter η (see Gumbel 1958). If ε'_n is a disturbance term associated with the maximum of all possible discrete outcomes $\neq i$ (see Equation 13.13) with mode equal to zero and scale parameter η , and $\beta' X'_n$ is the parameter and covariate product associated with the maximum of all possible discrete outcomes $\neq i$, then it is shown that

$$\beta' X'_n = \frac{1}{\eta} \ln \sum_{\forall I \neq i} \text{EXP}(\eta \beta_i X_{in}) \quad (13.15)$$

This result arises because for extreme value Type 1 distributed variates, ε , the addition of a positive scalar constant say, a , changes the mode from ω to $\omega + a$ without affecting the scale parameter η (see Johnson and Kotz 1970). So, if ε'_n has mode equal to zero and scale parameter η , adding the scalar shown in Equation 13.14 gives an extreme value distributed variate with mode ($\beta' X'_n$) equal to Equation 13.14 (as shown in Equation 13.15) and scale parameter η .

Using these results, Equation 13.13 is written as

$$P_n(i) = P(\beta_i X_{in} + \varepsilon_{in} \geq \beta' X'_n + \varepsilon'_n) \quad (13.16)$$

or

$$P_n(i) = P(\beta' X'_n + \varepsilon'_n - \beta_i X_{in} - \varepsilon_{in} \leq 0) \quad (13.17)$$

And, because the difference between two independently distributed extreme value Type 1 variates with common scale parameter η is logistic distributed

$$P_n(i) = \frac{1}{1 + \text{EXP}[\eta(\beta' X'_n - \beta_i X_{in})]} \quad (13.18)$$

rearranging terms

$$P_n(i) = \frac{\text{EXP}[\eta(\beta_i X_{in})]}{\text{EXP}[\eta(\beta_i X_{in})] + \text{EXP}[\eta(\beta' X'_n)]} \quad (13.19)$$

Substituting with Equation 13.15 and setting $\eta = 1$ (there is no loss of generality, Johnson and Kotz 1970) the equation becomes

$$P_n(i) = \frac{\text{EXP}[\beta_i X_{in}]}{\text{EXP}[\beta_i X_{in}] + \text{EXP}\left[LN \sum_{\forall I \neq i} \text{EXP}(\beta_I X_{In})\right]} \quad (13.20)$$

or

$$P_n(i) = \frac{\text{EXP}[\beta_i X_{in}]}{\sum_{\forall I} \text{EXP}(\beta_I X_{In})} \quad (13.21)$$

which is the standard multinomial logit (MNL) formulation. For estimation of the parameter vectors (β 's) by maximum likelihood, the log-likelihood function is

$$LL = \sum_{n=1}^N \left(\sum_{i=1}^I \delta_{in} \left[\beta_i X_{in} - LN \sum_{\forall I} \text{EXP}(\beta_I X_{In}) \right] \right) \quad (13.22)$$

where I is the total number of outcomes, δ is as defined in Equation 13.9, and all other variables are as defined previously.

When applying the MNL model it is important to realize that the choice of the extreme value Type 1 distribution is made on the basis of computational convenience, although this distribution is similar to the normal distribution (see Figure 13.3 for a two-outcome model and compare to Figure 13.1). Other restrictive assumptions made in the derivation of the model are the independence of disturbance terms and their common variance (homoscedasticity). These assumptions have estimation consequences that are discussed later in this chapter.

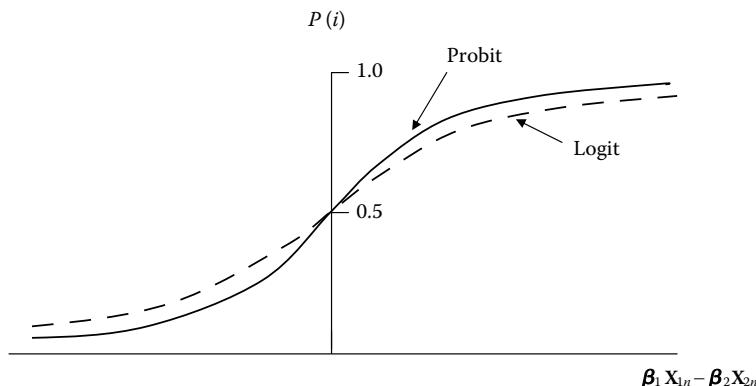


FIGURE 13.3

Comparison of binary logit and probit outcome probabilities.

13.4 Discrete Data and Utility Theory

In many applications it is useful to tie the statistical model to an underlying theoretical construct. An example of this approach in the economic and transportation literature is the integration of discrete outcome models and utility theory. Traditional approaches from microeconomic theory have decision makers choosing among a set of alternatives such that their utility (satisfaction) is maximized subject to the prices of the alternatives and an income constraint (see Nicholson 1978). When dealing with decision makers' choices among discrete alternatives, discrete outcome models are often referred to as discrete choice models, but it is important to remember, in a more general sense, that any discrete outcome can, whether a consumer choice or not, be modeled using a discrete outcome model.

Because utility theory consists of decision makers selecting a utility-maximizing alternative based on prices of alternatives and an income constraint, any purchase affects the remaining income and thus all purchases are interrelated. This constraint creates an empirical problem because theoretically one cannot isolate specific choice situations. For example, when considering a traveler's choice of mode (e.g., car, bus, and rail), one would also have to consider the choice of breakfast cereal because the purchase of any good will affect remaining income. This difficulty can make the empirical analysis of choice data virtually intractable. However, the economic literature provides a number of options that require placing restrictions on the utility function. To illustrate these, a utility function is defined that is determined by the consumption of m goods (y_1, y_2, \dots, y_m) such that

$$u = f(y_1, y_2, \dots, y_m) \quad (13.23)$$

As an extremely restrictive case it is assumed that the consumption of one good is independent of the consumption of all other goods. The utility function is then written as

$$u = f_1(y_1) + f_2(y_2) + \dots + f_m(y_m) \quad (13.24)$$

This additive utility function, in nearly all applications, is unrealistically restrictive. For example, the application of such an assumption implies that the acquisition of two types of breakfast cereal are independent although it is clear that the purchase of one will affect the purchase of the other. A more realistic restriction on the utility function is to separate decisions into groups and to assume that consumption of goods within the groups is independent of those goods in other groups. For example, if utility is a function of the

consumption of breakfast cereal and a word-processing software package, these two choices are separated and modeled in isolation. This separability is an important construct applied economic theory (Phlips 1974). It is this property that permits the focus on specific choice groups such as the choices of travel mode to work.

The previous discussion views utility in the traditional sense in that utility is maximized subject to an income constraint and this maximization produces a demand for goods y_1, y_2, \dots, y_m . When applying discrete outcome models, the utility function is typically written with prices and incomes as arguments. The utility function written in this way is said to be indirect, and the relationship between this indirect utility and the resulting demand equation for some good m is given by Roy's identity (Dubin and McFadden 1984; Mannerling and Winston 1985; Phlips 1974)

$$y_m^0 = -\frac{\partial V/\partial p_m}{\partial V/\partial Inc} \quad (13.25)$$

where V is the indirect utility, p_m is the price of good m , Inc is the decision maker's income, and y_m^0 is the utility-maximizing demand for good m . As is shown later, the application of this equation proves to be critical in model development and the economic interpretation of model findings.

Applying the utility framework within discrete outcome models is straightforward. Using the notation above, T in Equations 13.1, 13.2, and 13.3 becomes the utility determining the choice (as opposed to a function determining the outcome). But in addition to the restrictions typically used in applying utility theory to discrete outcome models (separability of utility functions and the use of an indirect utility function with prices and income as arguments in the function), behavioral implications are noteworthy. The derivations of discrete outcome models provided herein imply that the model is compensatory. That is, changes in factors that determine the function T in Equation 13.1 for each discrete outcome do not matter as long as the total value of the function remains the same. This phenomenon is potentially problematic in some utility-maximizing choice situations. For example, consider the choice of a vehicle purchase where price, fuel efficiency, and seating capacity determine the indirect utility (variables in the X vector in Equation 13.1). Decision makers with families of five may extract little utility from a two-seat sports car. Yet, the model assumes that the decision-maker is compensated by lower prices or increased fuel efficiency. Some argue that a more realistic framework is one in which there are thresholds beyond which the decision maker does not accept an alternative. Models that use thresholds and relax the compensatory assumption are referred to as noncompensatory models. Due to their inherent complexity, noncompensatory models have been rarely used in practice (Tversky 1972).

13.5 Properties and Estimation of MNL Models

The objective of MNL model estimation is to estimate a function that determines outcome probabilities. As an example, for choice models based on utility theory this function is the indirect utility. To illustrate, consider a commuter's choice of route from home to work where the choices are to take an arterial, a two-lane road, or a freeway. Using the form of the MNL model given in Equation 13.21, the choice probabilities for the three routes are for each commuter n (n subscripting omitted)

$$P(a) = \frac{e^{V_a}}{e^{V_a} + e^{V_t} + e^{V_f}}, \quad P(t) = \frac{e^{V_t}}{e^{V_a} + e^{V_t} + e^{V_f}}, \quad P(f) = \frac{e^{V_f}}{e^{V_a} + e^{V_t} + e^{V_f}} \quad (13.26)$$

where $P(a)$, $P(t)$, and $P(f)$, are the probabilities that commuter n selects the arterial, two-lane road, and freeway, respectively, and V_a , V_t and V_f are corresponding indirect utility functions. Broadly speaking, the variables defining these functions are classified into two groups—those that vary across outcome alternatives and those that do not. In route choice, distance and number of traffic signals are examples of variables that vary across outcome alternatives. Commuter income and other commuter-specific characteristics (number of children, number of vehicles, and age of commuting vehicle) are variables that do not vary across outcome alternatives. The distinction between these two sets of variables is important, because the MNL model is derived using the difference in utilities (see Equation 13.17). Because of this differencing, estimable parameters relating to variables that do not vary across outcome alternatives can, at most, be estimated in $I - 1$ of the functions determining the discrete outcome (I is the total number of discrete outcomes). The parameter of at least one of the discrete outcomes must be normalized to zero to make parameter estimation possible (this constraint is illustrated in a forthcoming example).

Given these two variables types, the utility functions for Equation 13.26 are defined as

$$\begin{aligned} V_a &= \beta_{1a} + \boldsymbol{\beta}_{2a} \mathbf{X}_a + \boldsymbol{\beta}_{3a} \mathbf{Z} \\ V_t &= \beta_{1t} + \boldsymbol{\beta}_{2t} \mathbf{X}_t + \boldsymbol{\beta}_{3t} \mathbf{Z} \\ V_f &= \beta_{1f} + \boldsymbol{\beta}_{2f} \mathbf{X}_f + \boldsymbol{\beta}_{3f} \mathbf{Z} \end{aligned} \quad (13.27)$$

where \mathbf{X}_a , \mathbf{X}_t , and \mathbf{X}_f are vectors of variables that vary across arterial, two-lane, and freeway choice outcomes, respectively, as experienced by commuter n , \mathbf{Z} is a vector of characteristics specific to commuter n , β_1 are constant terms, $\boldsymbol{\beta}_2$ are vectors of estimable parameters corresponding to outcome-specific variables in \mathbf{X} vectors, and $\boldsymbol{\beta}_3$ are vectors corresponding to variables that do not vary across outcome alternatives. Note that the constant terms are effectively

the same as variables that do not vary across alternate outcomes and at most are estimated for $I - 1$ of the outcomes.

Example 13.1

A survey of 151 commuters was conducted. Information was collected on their route selection on their morning trip from home to work. All commuters departed from the same origin (a large residential complex in suburban State College, Pennsylvania) and went to work in the downtown area of State College. Distance was measured precisely from parking lot of origin to parking lot of destination so there is a variance in distances among commuters even though they departed and arrived in the same general areas. Commuters had a choice of three alternate routes; a four-lane arterial (speed limit = 60 km/h, two lanes each direction), a two-lane highway (speed limit = 60 km/h, one lane each direction), and a limited access four-lane freeway (speed limit = 90 km/h, two lanes each direction). Each of these three routes shared some common portions for access and egress because, for example, the same road to the downtown area is used by both freeway and two-lane road alternatives since the freeway exits on to the same city street as the two-lane road.

To analyze commuters' choice of route, a MNL model is estimated given the variables shown in Table 13.1. Model estimation results are presented in Table 13.2. Based on this table, the estimated utility functions are (see Equation 13.27):

$$\begin{aligned} V_a &= -0.942(DISTA) \\ V_t &= 1.65 - 1.135(DISTT) + 0.128(VEHAGE) \\ V_f &= -3.20 - 0.694(DISTF) + 0.233(VEHAGE) + 0.764(MALE) \end{aligned} \quad (13.28)$$

The interpretation of the parameters is straightforward. First, consider the constant terms. Note that V_a does not have a constant because constants are estimated as variables that do not vary across alternate outcomes and therefore appear in at most $I - 1$ functions (constants are estimated as $\beta\mathbf{X}$ with \mathbf{X} being a vector of 1s, and this vector does not vary across alternate outcomes). The lack of constant in the arterial function establishes it as a 0 baseline. Thus, all else being equal, the two-lane road is more likely to be selected (with its positive constant) relative to the arterial and the freeway is less likely to be selected relative to the arterial (with its negative constant). And, all else being equal, the freeway is less likely to be selected than the two-lane road. Omitting a constant from the arterial function is arbitrary and a constant could have been omitted from one of the other functions instead and the same convergence of the likelihood function would have resulted. The key here is that all of the constants are relative and the omission of one constant over another simply rescales the remaining constants. For example, if the constant is omitted from the two-lane function (so it becomes zero) the estimated constant for the arterial is -1.65 and the estimated constant for the freeway is -4.85 (thus preserving the same magnitude of differences).

The distance from home to work along the various routes was also included in the model. Distance varies across routes so this variable is included in all utility functions. The negative sign indicates that increasing distance decreases the

TABLE 13.1

Variables Available for Home to Work Route-Choice Model Estimation

Variable No.	Variable Description
1	Route chosen: 1 if arterial, 2 if two-lane road, 3 if freeway
2	Traffic flow rate of arterial at time of departure (vehicles per hour)
3	Traffic flow rate of two-lane road at time of departure (vehicles per hour)
4	Traffic flow rate of freeway at time of departure (vehicles per hour)
5	Number of traffic signals on the arterial
6	Number of traffic signals on the two-lane road
7	Number of traffic signals on the freeway
8	Distance on the arterial in kilometers
9	Distance on the two-lane road in kilometers
10	Distance on the freeway in kilometers
11	Seat belts: 1 if wearing, 0 if not
12	Number of passengers in vehicle
13	Commuter age in years: 1 if less than 23, 2 if 24–29, 3 if 30–39, 4 if 40–49, 5 if 50 and over 50
14	Gender: 1 if male, 0 if female
15	Marital status: 1 if single, 0 if married
16	Number of children in household (aged 16 or less)
17	Annual household income (US dollars per year): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if more than 50,000
18	Age of the vehicle used on the trip in years

TABLE 13.2

Multinomial Logit Estimation Results for the Choice of Route to Work

Variable Description	Variable Mnemonic	Estimated Parameter	t Statistic
Two-lane road constant		1.65	1.02
Freeway constant		-3.20	-1.16
<i>Variables that vary across alternate outcomes</i>			
Distance on the arterial in kilometers		-0.942	-3.99
Distance on the two-lane road in kilometers	DISTT	-1.135	-5.75
Distance on the freeway in kilometers	DISTF	-0.694	-2.24
<i>Variables that do not vary across alternate outcomes</i>			
Male indicator (1 if male commuter, 0 if not defined for the freeway utility function)	MALE	0.766	1.19
Vehicle age in years (defined for the two-lane road utility function)	VEHAGE	0.128	1.87
Vehicle age in years (defined for the freeway utility function)	VEHAGE	0.233	2.75
Number of observations		151	
Log-likelihood at zero		-165.89	
Log-likelihood at convergence		-92.51	

likelihood of a route being selected. The estimation of separate parameters for the three distances indicates that distance is not valued equally on the three routes. In this case, distance on the two-lane road is most onerous, followed by distance on the arterial and distance on the freeway.

The positive parameter on male indicator variable indicates that men are more likely to take the freeway. The vehicle age variables indicate that the older the commuting vehicle the more likely the commuter will drive on the freeway, followed by two lane. This variable does not vary across alternate outcomes so is captured in at most $I - 1$ functions (in this case two). Age of the vehicle on the arterial is implicitly set to zero and the same relativity logic discussed above for the constants applies.

13.5.1 Statistical Evaluation

The statistical significance of individual parameters in a MNL model is approximated using a one-tailed t test. To determine if the estimated parameter is significantly different from zero, the test statistic t^* , which is approximately t distributed, is

$$t^* = \frac{\beta - 0}{S.E.(\beta)} \quad (13.29)$$

where $S.E.(\beta)$ is the standard error of the parameter. Note that because the MNL is derived from an extreme value distribution and not a normal distribution, the use of t -statistics is not strictly correct although in practice it is a reliable approximation of the true significance.

A more general and appropriate test is the likelihood ratio test. This test is used for a wide variety of reasons such as assessing the significance of individual parameters, evaluating overall significance of the model (similar to the F test for ordinary least squares regression), examining the appropriateness of estimating separate parameters for the same variable in different outcome functions (such as the separate distance parameters estimated in Example 13.1), and for examining the transferability of results over time and space. The likelihood ratio test statistic is

$$\chi^2 = -2[LL(\beta_R) - LL(\beta_U)] \quad (13.30)$$

where $LL(\beta_R)$ is the log-likelihood at convergence of the “restricted” model and $LL(\beta_U)$ is the log-likelihood at convergence of the “unrestricted” model. This χ^2 statistic is χ^2 distributed with degrees of freedom equal to the difference in the numbers of parameters between the restricted and unrestricted model (the difference in the number of parameters in the β_R and the β_U parameter vectors). By comparing the improvement in likelihoods as individual variables are added, this test is the correct way to test for the significance of individual variables because, as mentioned above, the more commonly used t -statistic is based on the assumption of normality which is not strictly true for the MNL model.

Example 13.2

For the model estimated in Example 13.1, it is interesting to test if the effects of distance are different across routes (as estimated) or whether there is one effect across all routes. To test this effect, the model shown in Table 13.2 needs to be reestimated while constraining the distance parameters to be the same across all three functions (see Equation 13.28). Keeping all other variables in the model, the single distance parameter is estimated as -1.147 (t^* of -5.75), and the log-likelihood at convergence of this restricted model (estimating 1 instead of 3 distance parameters) is -94.31. With the unrestricted model's log-likelihood at convergence of at -92.51 (as shown in Table 13.2) the application of Equation 13.29 yields an χ^2 value of 3.6. The degrees of freedom is 2 (since 2 fewer parameters are estimated in the restricted model). An χ^2 value of 3.6 and 2 degrees of freedom yields a level of confidence of 0.835. Thus using a 90% level of confidence the model of a generic distance effect cannot be rejected (the model with individual effects is not preferred to the model with a generic effect). The interpretation is that sampling variability alone can explain the observed difference in log-likelihoods, and the model with individual distance effects is rejected.

Finally, a common measure of overall model fit is the ρ^2 statistic (it is similar to R^2 in regression models in terms of purpose, and is occasionally referred to as the McFadden ρ^2 statistic). The ρ^2 statistic is

$$\rho^2 = 1 - \frac{LL(\boldsymbol{\beta})}{LL(\mathbf{0})} \quad (13.31)$$

where $LL(\boldsymbol{\beta})$ is the log-likelihood at convergence with parameter vector $\boldsymbol{\beta}$ and $LL(\mathbf{0})$ is the initial log-likelihood (with all parameters set to zero). The perfect model has a likelihood function equal to one (all selected alternative outcomes are predicted by the model with probability one, and the product of these across the observations are also equal to one) and the log-likelihood is zero giving a ρ^2 of one (see Equation 13.31). Thus the ρ^2 statistic lies between zero and one, and a statistic close to one suggests that the model is predicting the outcomes with near certainty. For Example 13.1, the ρ^2 is 0.442 [$1 - (-92.51/-165.89)$].

As is the case with R^2 in regression analysis, the disadvantage of the ρ^2 statistic is that it will always improve as additional parameters are estimated even though the additional parameters may be statistically insignificant. To account for the estimation of potentially insignificant parameters a corrected ρ^2 is estimated as

$$\text{corrected } \rho^2 = 1 - \frac{LL(\boldsymbol{\beta}) - K}{LL(\mathbf{0})} \quad (13.32)$$

where K is the number of parameters estimated in the model (the number of parameters in the vector β). For Example 13.1, the corrected ρ^2 is 0.394 [$1 - ((-92.51 - 8) / -165.89)$].

13.5.2 Interpretation of Findings

Two other techniques are used to assess individual parameter estimates; elasticities and marginal rates of substitution. Elasticity computations measure the magnitude of the impact of specific variables on the outcome probabilities. Elasticity is computed from the partial derivative for each observation n (n subscripting omitted):

$$E_{x_{ki}}^{P(i)} = \frac{\partial P(i)}{\partial x_{ki}} \times \frac{x_{ki}}{P(i)} \quad (13.33)$$

where $P(i)$ is the probability of outcome i and x_{ki} is the value of variable k for outcome i . It is readily shown by taking the partial derivative of the MNL model that Equation 13.33 becomes

$$E_{x_{ik}}^{P(i)} = [1 - P(i)] \beta_{ki} x_{ki} \quad (13.34)$$

Elasticity values are interpreted as the percent effect that a 1% change in x_{ki} has on the outcome probability $P(i)$. If the computed elasticity value is less than one, the variable x_{ki} is said to be inelastic and a 1% change in x_{ki} will have less than a 1% change in outcome i 's selection probability. If the computed elasticity is greater than one it is said to be elastic and a 1% change in x_{ki} will have more than a 1% change in outcome i 's selection probability.

Example 13.3

To determine the elasticity's for distances on the three routes based on the model estimated in Example 13.1 (assuming this model is tenable), Equation 13.34 is applied over all observations N (the entire 151 commuters). The average elasticity for distance (averaged over N commuters) on the arterial is -6.48 , suggesting that for the average commuter a 1% increase in distance on the arterial decreases the probability of the arterial being selected by 6.48%. Among the 151 commuters in the sample, elasticity's range from a high of -13.47 to a low of -1.03 . The computed average elasticity's for distances on the two-lane road and the freeway are -3.07 and -6.60 , respectively. These findings show that distance on the two-lane road has the least effect on the selection probabilities.

There are several points to keep in mind when using elasticities. First, the values are point elasticity's and as such are valid only for small changes x_{ik} and considerable error may be introduced when an elasticity is used to estimate the probability change caused by a doubling of x_{ki} . Second, elasticities are not applicable to indicator variables (those variables that take on values of 0 or 1 such as the male indicator in Table 13.2). Some measure of the sensitivity of indicator variables is made by computing a pseudoelasticity. The equation is

$$E_{x_{ki}}^{P(i)} = \frac{\text{EXP}[\Delta(\beta_i x_i)] \sum_{\forall l} \text{EXP}(\beta_{kl} x_{kl})}{\text{EXP}[\Delta(\beta_i x_i)] \sum_{\forall I_n} \text{EXP}(\beta_{kl} x_{kl}) + \sum_{\forall I \neq I_n} \text{EXP}(\beta_{kl} x_{kl})} - 1 \quad (13.35)$$

where I_n is the set of alternate outcomes with x_k in the function determining the outcome and I is the set of all possible outcomes. See Ulfarsson and Mannering (2004) for a discussion and application of this equation.

The elasticities described previously are referred to as direct elasticities because they capture the effect that a change in a variable determining the likelihood of alternate outcome i has on the probability of outcome i being selected. It may also be of interest to determine the effect of a variable influencing the probability of outcome j may have on the probability of outcome i . For example, referring to the route-choice problem (Example 13.1), interest may be centered on knowing how a change in the distance on the arterial affects the probability of the two-lane road being chosen. Known as a cross-elasticity, its value is computed using the equation

$$E_{x_{kj}}^{P(i)} = -P(j) \beta_{kj} x_{kj} \quad (13.36)$$

Note that this equation implies that there is one cross-elasticity for all i ($i \neq j$). This finding suggest that an increase in distance on the arterial results in an equal increase in the likelihood of the two-lane and freeway alternatives being chosen. This property of uniform cross-elasticities is an artifact of the error-term independence assumed in deriving the MNL model. This assumption is not always realistic and is discussed further when independence of irrelevant alternatives (IIA) properties and nested logit models are addressed later in this chapter. When computed using the sample of 151 commuters, the average cross-elasticity for distance on the arterial is 1.63 indicating that a 1% increase in distance on the arterial increases the probability of the two-lane road being selected by 1.63% and the probability of the freeway being accepted by 1.63%.

Because logit models are compensatory (as discussed in Section 13.4), marginal rates of substitution are computed to determine the relative magnitude of any two parameters estimated in the model. In MNL models, this rate is computed simply as the ratio of parameters for any two variables in question (in this case a and b)

$$MRS(i)_{ba} = \frac{\beta_{ia}}{\beta_{ib}} \quad (13.37)$$

Example 13.4

One might want to know the marginal rate of substitution between distance and vehicle age on the two-lane road based on the model estimated in Example 13.1. The estimated parameters are -0.942 for distance and 0.128 for vehicle age. The marginal rate of substitution between distance and vehicle age is -0.136 km/vehicle-year ($0.128/-0.942$). This estimate means that each year a vehicle ages (which increases the probability of the two-lane route being selected) distance is increased 0.136 km on average while the same route-choice probability is maintained.

13.5.3 Specification Errors

A number of restrictive assumptions are made when deriving the MNL model and when these are not met potentially serious specification errors may result. In addition, as with any statistical model, there are consequences in omitting key variables and including irrelevant variables in the model estimation. Arguably the most commonly overlooked and misunderstood limitation of the MNL model is what is referred to as the IIA property. Recall that a critical assumption in the derivation of the MNL model is that the disturbances (ϵ in Equation 13.5) are independently and identically distributed (IID). When this assumption does not hold, a major specification error results. This problem arises when only some of the functions, which determine possible outcomes, share unobserved elements (that show up in the disturbances). The “some of the functions” is a critical distinction because if all outcomes shared the same unobserved effects, the problem would self correct because in the differencing of outcome functions (see Equation 13.17) common unobserved effects would cancel out. Because the common elements cancel in the differencing, a logit model with only two outcomes can never have an IIA violation.

To illustrate the IIA problem, it is easily shown that for MNL models the ratio of any two-outcome probabilities is independent of the functions determining any other outcome since (with common denominators as shown in Equation 13.21)

$$\frac{P(1)}{P(2)} = \frac{\text{EXP}[\boldsymbol{\beta}_1 \mathbf{X}_1]}{\text{EXP}[\boldsymbol{\beta}_2 \mathbf{X}_2]} \quad (13.38)$$

for each observation n (n subscripting omitted). To illustrate the problem that can arise as a result of this property, consider the estimation of a model of choice of travel mode to work where the alternatives are to take a personal vehicle, a red transit bus, or a blue transit bus. The red and blue transit buses clearly share unobserved effects that will appear in their disturbance terms and they will have exactly the same functions ($\beta_{rb}X_{rb} = \beta_{bb}X_{bb}$) if the only difference in their observable characteristics is their color. For illustrative purposes, assume that, for a sample commuter, all three modes have the same value of $\beta_i X_i$'s (the red bus and blue bus will, and assume that costs, time, and other factors that determine the likelihood of the personal vehicle being chosen works out to the same value as the buses). The predicted probabilities yield each mode with a 33% chance of being selected. This outcome is unrealistic since the correct answer is a 50% chance of taking a personal vehicle and a 50% chance of taking a bus (both red and blue bus combined) and not 33.33% and 66.67%, respectively, as the MNL would predict. The consequences of an IIA violation are incorrect probability estimates.

In most applications the IIA violation is more subtle than in the previous example. There are a number statistical tests that are conducted to test for IIA violations. One of the more common of these tests was developed by Small and Hsiao (1985). The Small-Hsiao IIA test is easy to conduct. The procedure is to first split the data randomly into two samples (N^A and N^B) containing the same number of observations. Two separate models are then estimated producing parameter estimates β^A and β^B . A weighted average of these parameters is obtained from

$$\beta^{AB} = \left(1/\sqrt{2}\right)\beta^A + \left(1-1/\sqrt{2}\right)\beta^B \quad (13.39)$$

Then, a restricted set of outcomes D is created as a subsample from the full set of outcomes. The sample N^B is then reduced to include only those observations in which the observed outcome lies in D . Two models are estimated with the reduced sample ($N^{B'}$) using D as if it were the entire outcome set (B' in superscripting denotes the sample reduced to observations with outcomes in D). One model is estimated by constraining the parameter vector to be equal to β^{AB} as computed above. The second model estimates the unconstrained parameter vector $\beta^{B'}$. The resulting log-likelihoods are used to evaluate the suitability of the model structure by creating a chi-squared statistic with the number of degrees of freedom equal to the number of parameters in β^{AB} (also the same number as in $\beta^{B'}$). This statistic is

$$\chi^2 = -2[LL^{B'}(\beta^{AB}) - LL^{B'}(\beta^{B'})] \quad (13.40)$$

The test is then repeated by interchanging the roles of the N^A and N^B subsamples (reducing the N^A sample to observations were the observed

outcomes lie in D and proceed). Using the same notation, Equation 13.39 becomes

$$\boldsymbol{\beta}^{BA} = (1/\sqrt{2})\boldsymbol{\beta}^B + (1 - 1/\sqrt{2})\boldsymbol{\beta}^A \quad (13.41)$$

and the chi-squared statistic is

$$X^2 = -2[LL^{A'}(\boldsymbol{\beta}^{BA}) - LL^{A'}(\boldsymbol{\beta}^A)] \quad (13.42)$$

Example 13.5

It might be plausible that the model estimated in Example 13.1 may have an IIA violation because the two-lane road and the arterial may share unobserved elements (because they are lower level roads relative to the freeway (no access control, lower design speeds, etc.). A Small-Hsiao test is used to evaluate this suspicion. The sample is randomly split to give 75 observations in N^A and 76 observations in N^B . Two models (one for the A and B subsamples) are estimated with the eight parameters as shown in Table 13.2. Based on the estimation results, $\boldsymbol{\beta}^{AB}$ is computed as shown in Equation 13.39. The N^B sample is then reduced to include only those observations in which the arterial or two-lane road were chosen. The models required to compute Equation 13.40 are estimated (only 4 parameters are estimated because the parameters specific to the freeway cannot be estimated because freeway users are no longer in the sample). The estimation results give $LL^B(\boldsymbol{\beta}^{AB}) = -25.83$ and $LL^B(\boldsymbol{\beta}^B) = -24.53$. Applying Equation 13.40 yields a X^2 of 2.6 with 4 degrees of freedom. This statistic suggests that the MNL structure is not rejected at any reasonable confidence level (confidence level is 37.3%). Following the same procedure reversing the N^A and N^B subsamples and applying Equations 13.41 and 13.42 gives a X^2 of 0.18 with 4 degrees of freedom which means the MNL structure again cannot be rejected at any reasonable confidence level (confidence level is 0.4%). Based on this test, the MNL model structure cannot be refuted and IIA violations resulting from shared unobserved effects from the arterial and two-lane road are not statistically significant.

Aside from IIA violations, there are a number of other possible specification errors. These include the following.

Omitted Variables. The omission of a relevant explanatory variables results in inconsistent estimates of logit model parameters and choice probabilities if any of the following hold: (1) the omitted variable is correlated with other variables included in the model, (2) the mean values of the omitted variable vary across alternate outcomes and outcome-specific constants are not included in the model, or (3) the omitted variable is correlated across alternate outcomes or has a different variance in different outcomes. Because one or more of these conditions are likely to hold, omitting relevant variables is a serious specification problem.

Presence of an Irrelevant Variable. Estimates of parameter and choice probabilities remain consistent in the presence of an irrelevant variable but the standard errors of the parameter estimates will increase (loss of efficiency).

Disturbances that are not IID. Dependence among a subset of possible outcomes causes the IIA problem resulting in inconsistent parameter estimates and outcome probabilities. Having disturbances with different variances (not identically distributed) also results in inconsistent parameter estimates and outcome probabilities. Such a case could arise in a model of choice of mode of work where “comfort” was a major part of the disturbance. In this case, the wide variety of vehicles may make the variance of the disturbance term bigger for the vehicle-mode option than the disturbance terms for the other competing modes such as bus and train.

Random Parameter Variations. Standard MNL estimation assumes that the estimated parameters are the same for all observations. Violations of this assumption can occur if there is some compelling reason (to believe) that parameters vary across observations in a way that cannot be accounted for in the model. For example, suppose that a price parameter, the cost of taking the mode, is estimated in a model of choice of mode to work. The parameter estimate inherently assumes that price is equally onerous across the population. Realistically, price is less important to wealthy people. In these cases, estimating the parameter associated with a transformed variable (such as price divided income) instead of simply price is an attempt to account for this variation. In other applications, the parameter variation may be essentially random (such as the effect of patience on the parameter associated with mode travel time) in that available data cannot uncover the relationship causing the variation in the parameter. Such random parameter estimates give inconsistent estimates of parameters and outcome probabilities. The mixed logit model described in Chapter 16 addresses this modeling challenge.

Correlation between Explanatory Variables and Disturbances and Endogenous Variables. If a correlation exists between X and ε then parameter estimates are inconsistent. An example of such correlation is a model of mode-to-work choice where distance is an explanatory variable and comfort is unobserved. If it is the practice of the transit agency to put new comfortable buses on longer routes a correlation will exist. An example of an endogenous variable problem is a discrete model of the severity of vehicle accidents in icy conditions (with discrete outcome categories such as property damage only, injury, and fatality) and the presence of an ice-warning sign (an X variable). Because ice-warning signs are likely to be posted at high-severity locations, their presence is correlated with severity and thus is endogenous (see Carson and Mannering 2001 for this example).

Erroneous Data. If erroneous data are used, parameter and outcome probabilities are incorrectly estimated (also erroneous).

State Dependence and Heterogeneity. A potential estimation problem can arise in discrete outcome models if information on previous outcomes is used to determine current outcome probabilities. To illustrate, consider the prior example of a commuter choosing among three routes on a morning commute to work. On day 1, the driver is observed taking the freeway. In modeling the commuter's choice on day 2, it is tempting to use information observed from the previous day as an independent variable in the X vector. Conceptually, this approach may make sense because it could be capturing important habitual behavior. Such habitual behavior is called state dependence (Heckman 1981). Unfortunately, the inclusion of such a state variable may also capture residual heterogeneity, which would lead one to observe spurious state dependence. To illustrate, suppose that the disturbance term includes unobserved characteristics (unobserved heterogeneity) that are commonly present among drivers selecting specific routes. If a variable indicating previous route selection is included, it is unclear whether the estimated parameter of this variable is capturing true habitual behavior (state dependence) or is simply picking up some mean commonality in drivers' disturbance terms (unobserved heterogeneity). This distinction is important because the presence or absence of habitual behavior could lead one to draw significantly different behavioral conclusions. Isolating true state dependence from unobserved heterogeneity is not an easy task (see Heckman 1981). As a result, extreme caution must be used when interpreting the parameters of variables that are based on previous outcomes.

One general comment about endogeneity and discrete outcome models is noteworthy. This comment relates to the effect of a single observation on the right-hand-side (explanatory) variable. Consider the choice of a used vehicle (make, model, and vintage) in which price is an explanatory variable. Naturally, because there are a limited number of specific used vehicles, such as a 2006 Mazda Miata, an increase in selection probabilities will increase the price so that supply–demand equilibrium is met, ensuring that the demand for the 2006 Mazda Miata does not exceed the supply of 2006 Miatas. Thus one could argue that price is endogenous (as it would surely be in an aggregate regression model of vehicle demand). However, because the effect of any single observation on vehicle price is infinitesimal, many have contended that price is exogenous for estimation purposes. Still, if one were to forecast used-vehicle market shares using the model (using the summation of individual outcome probabilities as a basis), vehicle prices would need to be forecasted internally as a function of total vehicle demand. As a result prices must be treated as endogenous for forecasting. See Berkovec (1985) and Mannerling and Winston (1987b) for examples of such forecasting strategies.

Authors of recent work have argued that price variables continue to present problems in individual discrete outcome models. This view is held because prices tend to be higher for products that have attributes that are observed by the consumer and not by the analyst. An example is unobserved aspects

of vehicle quality (such as fit and finish) that result in a higher price. This missing variable sets up a correlation between the prices variables and the disturbances that causes an obvious specification problem (see above).

13.5.4 Data Sampling

There are two general types of sampling strategies for collecting data to estimate discrete outcome models, random and stratified random sampling. All standard MNL model derivations assume that the data used to estimate the model are drawn randomly from a population of observations. Complications may arise in the estimation when sampling is not random. Stratified sampling refers to a host of nonrandom sampling alternatives. The idea of stratified sampling is that some known population of observations is partitioned into subgroups (strata) and random sampling is conducted in each of these subgroups. This type of sampling is particularly useful when one wants to gain information on a specific group that is a small percentage of the total population of observations (such as transit riders in the choice of transportation mode or households with incomes exceeding \$200,000 per year). Note that random sampling is a special case of stratified sampling in that the number of observations chosen from each strata is in exact proportion to the size of the strata in the population of observations.

There are four special cases of stratified sampling—exogenous sampling, outcome-based sampling, enriched sampling, and double sampling.

Exogenous sampling refers to sampling that in which selection of the strata is based on values of the X (right-hand side) variables. For example, in the previous route-choice example (Example 13.1), to ensure that users of older vehicles are well represented, one may have chosen to sample 75 commuters with vehicles of age 10 years or less and 76 commuters commuting in vehicles 10 years or greater. In such cases, Manski and Lerman (1977) have shown that standard maximum likelihood estimation (treating the sample as though it was a random sample) is appropriate.

Outcome-based sampling may be used to get a sufficient representation of a specific outcome or may be an artifact of the data-gathering process. For example, the percentage of individuals choosing public transit as a mode from home to work in most U.S. cities is very small (on the order of 5% or less). To get a sufficient representation of transit riders, one may survey transit users at bus stops and thus overrepresent transit users in the overall mode-choice sample. If the proportions of outcomes in the sample are not equal to the proportions of outcomes in the overall population, an estimation correction must be made. The correction is straightforward providing that a full set of outcome-specific constants is specified in the model (since constants do not vary across alternate outcomes $I - 1$ constants must be specified, where I is the set of outcomes). Under these conditions, standard MNL estimation correctly estimates all parameters except for the outcome-specific

constants. To correct the constant estimates, each constant must have the following subtracted from it

$$LN\left(\frac{SF_i}{PF_i}\right) \quad (13.43)$$

where SF_i is the fraction of observations having outcome i in the sample and PF_i is the fraction of observations having outcome i in the total population.

Example 13.6

In Example 13.1 a random sample was used and the population fractions were: 0.219 choosing the arterial, 0.682 choosing the two-lane road, and 0.099 choosing the freeway. Suppose the estimation results shown in Table 13.2 were obtained by equal sampling of the three alternatives ($SF_a = 0.333$, $SF_t = 0.333$, $SF_f = 0.333$). The correct constant estimations would then have to be calculated. Using the estimated constants in Table 13.2 and applying Equation 13.43, the correct constant estimate is $0 - LN(0.333/0.219)$ or -0.419 for the arterial, $1.65 - LN(0.333/0.682)$ or 0.933 for the two-lane road, and $-3.20 - LN(0.333/0.099)$ or -4.41 for the freeway.

Enriched sampling and double sampling are the last two general types of stratified sampling. Enriched sampling is the merging of a random (or random stratified) sample with a sample of another type. For example, a random sample of route choices may be merged with a sample of commuters observed taking one of the routes such as the freeway. Some types of enriched sampling problems reduce to the same correction used for the outcome-based samples. Others may result in estimation complications (see Cosslett 1981).

Finally, double sampling usually refers to the process where information from a random sample is used to direct the gathering of additional data often targeted at oversampling underrepresented components of the population. Estimation of MNL models with double sampling complicates the likelihood function. The reader is referred to Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), and Ben-Akiva and Lerman (1985) for details on estimation alternatives in sampling and additional sampling strategies.

13.5.5 Forecasting and Aggregation Bias

When using logit models for forecasting aggregate impacts there is a potential for forecasting bias that arises from the nonlinearity of the model. To illustrate this problem, consider what happens in a simple case when a population

consists of two observations, a and b in a model with one independent variable x . To predict the population probabilities, one approach is to use the average x in the model to come up with a probability of outcome i , over the population, of $P(i|x_{avg})$. However, Figure 13.4 shows that this approach is incorrect due to the nonlinearity of outcome probabilities with respect to x . In the two-observation case, the correct population probability is the average of observation as and bs individual probability as shown by $P_{ab}(i|x_{ab})$ in Figure 13.4. Figure 13.4 shows the bias that can result by using x_{avg} instead of averaging probabilities over the entire population. Given this potential bias, the forecasting problem becomes one of minimizing this bias (often referred to as aggregation bias), since one rarely has the entire population needed to forecast outcome shares in the population.

Theoretically, all bias in estimating outcome shares in a population are eliminated by

$$S_i = \int_X g_i(\mathbf{X}) h(\mathbf{X}) d\mathbf{X} \quad (13.44)$$

where S_i is the population share of outcome i , $g_i(\mathbf{X})$ is the functional representation of the model and $h(\mathbf{X})$ is the distribution of model variables over the population. The problem in applying this equation is that $h(\mathbf{X})$ is not completely known. However, there are four common approximations to Equation 13.44 that are used in practice; sample enumeration, density functions, distribution moments, and classification.

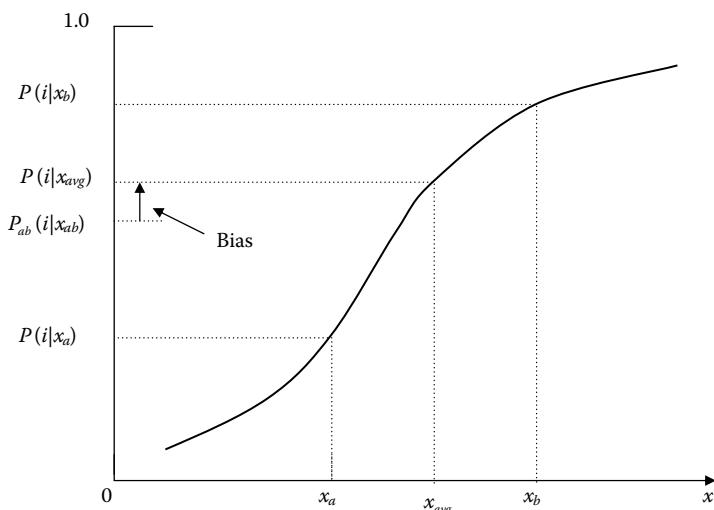


FIGURE 13.4
Example of population aggregation bias.

- **Sample enumeration.** This procedure involves using the same sample that was used to estimate the model to predict the population probabilities. Outcome probabilities for all observations in the sample are computed and these probabilities are averaged to approximate the true population outcome probabilities for each alternative. Applying this approach with the route-choice data for the model in Example 13.1, average route-choice probabilities are .218, .684, and .098 for the arterial, two-lane road, and freeway, respectively. To illustrate the possible bias, if average values of X are used, population route-choice probabilities are estimated to be .152, .794, and .053 for the arterial, two-lane road, and freeway, respectively, which are far from the observed shares of 0.219, 0.682, and 0.099, respectively. Sample enumeration represents an excellent approximation to Equation 13.44 but is restrictive when transferring results to other populations.
- **Density functions.** Constructing density function approximations of x_s can approximate Equation 13.44. The advantage of this approach is that it has great flexibility in applying results to different populations. The disadvantages are that the functions are often difficult to construct on theoretical or empirical grounds and it is difficult to capture covariances among x_s .
- **Distribution moments.** This approach attempts to describe $h(X)$ by considering moments and cross-moments to represent the spread and shape of variable distributions and their interactions in the population. The major limitation of this approach is the gathering of theoretical and/or empirical information to support the representation of moments and cross-moments.
- **Classification.** The classification approach attempts to categorize the population into nearly homogeneous groups and use averages of x_s from these groups. This approach is easy to apply but the assumption of homogeneity among population groups is often dubious and can introduce considerable error.

For additional information on forecasting and possible aggregation bias, the interested reader should see Koppelman (1975) for a detailed exploration of the topic and Mannering and Harrington (1981) for an example of an application of alternate bias-reduction techniques.

13.5.6 Transferability

A concern with all models is whether their estimated parameters are transferable spatially (among regions or cities) or temporally (over time). From a spatial perspective, transferability is desirable because it means that parameters of models estimated in other places are used, thus saving the cost of additional data collection and estimation. Temporal transferability ensures that forecasts

made with the model have some validity in that the estimated parameters are stable over time. When testing spatial and temporal transferability, likelihood ratio tests are applied. Suppose the transferability of parameters between two regions a and b is tested (or equivalently between two time periods). To test for transferability the following likelihood ratio test:

$$X^2 = -2[LL(\boldsymbol{\beta}_T) - LL(\boldsymbol{\beta}_a) - LL(\boldsymbol{\beta}_b)] \quad (13.45)$$

is applied, where $LL(\boldsymbol{\beta}_T)$ is the log-likelihood at convergence of the model estimated with the data from both regions (or both time periods), $LL(\boldsymbol{\beta}_a)$ is the log-likelihood at convergence of the model using region a data, and $LL(\boldsymbol{\beta}_b)$ is the log-likelihood at convergence of the model using region b data. In this test the same variables are used in all three models (total model, region a model, and region b model). This X^2 statistic is χ^2 distributed with degrees of freedom equal to the summation of the number of estimated parameters in all regional models (a and b in this case but additional regions can be added to this test) minus the number of estimated parameters in the overall model. The resulting X^2 statistic provides the confidence level that the null hypothesis (that the parameters are the same) is rejected. Alternatively, one could conduct the following test

$$X^2 = -2[LL(\boldsymbol{\beta}_{ba}) - LL(\boldsymbol{\beta}_a)] \quad (13.46)$$

where $LL(\boldsymbol{\beta}_{ba})$ is the log-likelihood at convergence of a model using the converged parameters from region b (using only region b 's data) on region a 's data (restricting the parameters to be region b 's estimated parameters), and $LL(\boldsymbol{\beta}_a)$ is the log-likelihood at convergence of the model using region a data. This test can also be reversed using $LL(\boldsymbol{\beta}_{ab})$ and $LL(\boldsymbol{\beta}_b)$. The statistic is χ^2 distributed with the degrees of freedom equal to the number of estimated parameters in $\boldsymbol{\beta}_{ba}$ and the resulting X^2 statistic provides the probability that the models have different parameters. The combination of these two tests yields a good assessment of the model's transferability.

A key criterion in demonstrating the spatial or temporal transferability of models is that they be well specified. Omitted variables and other specification errors may lead one to erroneously reject transferability.

13.6 Nested Logit Model (Generalized Extreme Value Models)

As noted previously, a restrictive property of the MNL model (MNL) is the IIA. Recall that this property arises from the model's derivation, which assumes independence of disturbance terms (ε_{in} 's), among alternate

outcomes. To overcome the IIA limitation in simple MNL models, McFadden (1981) developed a class of models known as generalized extreme value (GEV) models, which includes the MNL model and extensions. McFadden's derivation of this class of models, based on the assumption of GEV disturbance terms, allows any IIA problem to be readily addressed. The nested logit model is one of the more commonly used models in this class. The idea behind a nested logit model is to group alternate outcomes suspected of sharing unobserved effects into nests (this sharing sets up the disturbance term correlation that violates the derivation assumption). Because the outcome probabilities are determined by differences in the functions determining these probabilities (both observed and unobserved, see Equation 13.17), shared unobserved effects will cancel out in each nest providing that all alternatives in the nest share the same unobserved effects. This canceling out will not occur if a nest (group of alternatives) contains some alternative outcomes that share unobserved effects and others that do not (resulting in an IIA violation in the nest).

As an example of a nested structure, consider the route-choice problem described in Example 13.1. Suppose it is suspected that the arterial and two-lane road share unobserved elements (being lower level roads relative to the freeway with no access control, lower design speeds). When developing a nested structure to deal with the suspected disturbance term correlation, a structure shown visually in Figure 13.5 is used. By grouping the arterial and two-lane road in the same nest their shared unobserved elements cancel.

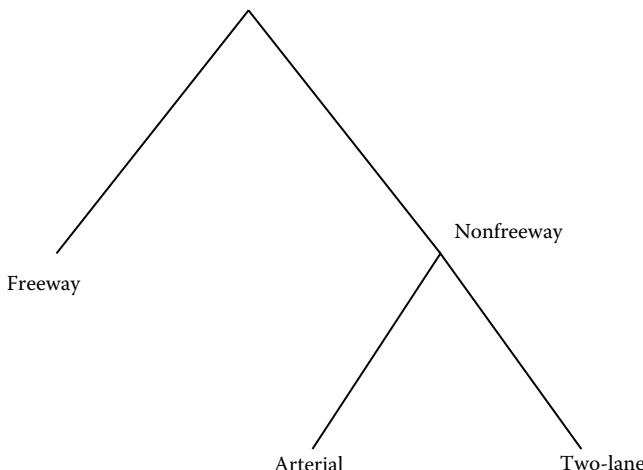


FIGURE 13.5

Nested structure of route choice with shared disturbances between arterial and two-lane road.

Mathematically, McFadden (1981) has shown the GEV disturbance assumption leads to the following model structure for observation n choosing outcome i :

$$P_n(i) = \frac{EXP[\boldsymbol{\beta}_i \mathbf{X}_{in} + \phi_i LS_{in}]}{\sum_{\forall I} EXP[\boldsymbol{\beta}_I \mathbf{X}_{in} + \phi_I LS_{in}]} \quad (13.47)$$

$$P_n(j|i) = \frac{EXP[\boldsymbol{\beta}_{ji} \mathbf{X}_{jn}]}{\sum_{\forall J} EXP[\boldsymbol{\beta}_{ji} \mathbf{X}_{jn}]} \quad (13.48)$$

$$LS_{in} = LN \left[\sum_{\forall J} EXP(\boldsymbol{\beta}_{ji} \mathbf{X}_{jn}) \right] \quad (13.49)$$

where $P_n(i)$ is the unconditional probability of observation n having discrete outcome i , \mathbf{X} are vectors of measurable characteristics that determine the probability of discrete outcomes, $\boldsymbol{\beta}$ are vectors of estimable parameters, $P_n(j|i)$ is the probability of observation n having discrete outcome j conditioned on the outcome being in outcome category i (e.g., for the nested structure shown in Figure 13.5 the outcome category i is nonfreeway and $P_n(j|i)$ is the binary logit model of the choice between the arterial and two-lane road), J is the conditional set of outcomes (conditioned on i), I is the unconditional set of outcome categories (the upper two branches of Figure 13.5), LS_{in} is the inclusive value (logsum), and ϕ_i is an estimable parameter. Note that this equation system implies that the unconditional probability of having outcome j is,

$$P_n(j) = P_n(i) \times P_n(j|i) \quad (13.50)$$

In the past, estimation of a nested model was typically done in a sequential fashion, although most modern software packages provide for a simultaneous estimation of all nests using full-information maximum likelihood. In the sequential method, the procedure first estimates the conditional model (Equation 13.48) using only the observations in the sample that are observed having discrete outcomes J . In the example illustrated in Figure 13.5, this approach results in a binary model of commuters observed taking the arterial or the freeway. Once these estimation results are obtained, the logsum is calculated (the denominator of one or more of the conditional models—see Equations 13.48 and 13.49) for all observations, both those selecting J and those not (for all commuters in our example case). Finally, these computed logsums (in our example there is just one logsum) are used as independent variables in the functions as shown in Equation 13.47. Note that not all unconditional outcomes need to have a logsum in their respective functions (the example shown in Figure 13.5 would only have a logsum present in the function for the nonfreeway choice).

Caution needs to be exercised when using the sequential estimation procedure described above because it has been shown that sequential estimation of a nested logit model results in variance–covariance matrices that are too small

and thus t -statistics are inflated (typically by about 10%–15%). This problem is resolved by estimating the entire model at once using full-information maximum likelihood. See Green (2008) for details on this approach and the forthcoming Example 13.7.

It is important to note that the interpretation of the estimated parameter associated with the inclusive values (ϕ_i) has the following important elements:

- ϕ_i must be greater than 0 and less than 1 in magnitude to be consistent with the nested logit derivation (see McFadden 1981).
- If $\phi_i = 1$, the assumed shared unobserved effects in the nest are not significant and the nested model reduces to a simple MNL (see Equations 13.47, 13.48, and 13.49 with $\phi_i = 1$). For the example route-choice model, estimating the nest shown in Figure 13.5 yields a ϕ_i for the nonfreeway outcome of the upper nest that is not significantly different than 1, indicating that the simple MNL model is the appropriate structure and that the arterial and two-lane road do not share significant unobserved effects (this finding is also corroborated by the IIA test conducted in Example 13.5). When estimating the nested structure shown in Figure 13.5, ϕ_i for the nonfreeway choice is estimated as 0.91 with a standard error of 0.348. For testing whether the parameter estimate is significantly different from 1, the t test is (again, this test is an approximation due to the nonnormality in the logit model)

$$t^* = \frac{\beta - 1}{\text{S.E.}(\beta)} = \frac{0.91 - 1}{0.348} = -0.2586 \quad (13.51)$$

- A one-tailed t test gives a confidence level of only 60%. It is important to exercise caution when using estimation packages that calculate t -statistics because they typically report values where β is compared with zero as shown in Equation 13.29 (which would have been 2.62 in the case above). In the previous example the 0.91 parameter estimate is significantly different from zero but not significantly different from one.
- If ϕ_i is less than zero then factors increasing the likelihood of an outcome being chosen in the lower nest will decrease the likelihood of the nest being chosen. For example, in the route-choice case, if distance on the arterial is reduced, the logsum on the nonfreeway nest increases but the negative ϕ_i decreases the likelihood of a nonfreeway route being chosen. This finding does not make sense because an improvement in a nonfreeway route should increase the likelihood of it being chosen.
- If ϕ_i is equal to zero then changes in nest outcome probabilities will not affect the probability of nest selection and the correct model is recursive (separated).

It is important to note that the nested structure is often estimated with more than the two levels shown in Figure 13.5, depending on the number of alternate outcomes and the hypothesized disturbance correlations. In addition, the nesting does not imply a decision tree or an ordering of how decisions are made—the proposed nesting structure conveys nothing about a hierarchical decision-making process. The nesting is purely an empirical method for eliminating IIA violations.

Example 13.7

To illustrate the estimation of a nested logit model using full-information maximum likelihood, consider an example of a model of motorcyclists' injury severity, as described in Savolainen and Mannering (2007). The data consist of 2,273 single-vehicle motorcycle accidents in the state of Indiana. There are four possible severity outcomes: no-injury (property damage only and possible injury); nonincapacitating injury; incapacitating injury; fatality. The variables available for model estimation are shown in Table 13.3.

TABLE 13.3

Variables Available for Motorcycle Injury-Severity Model Estimation

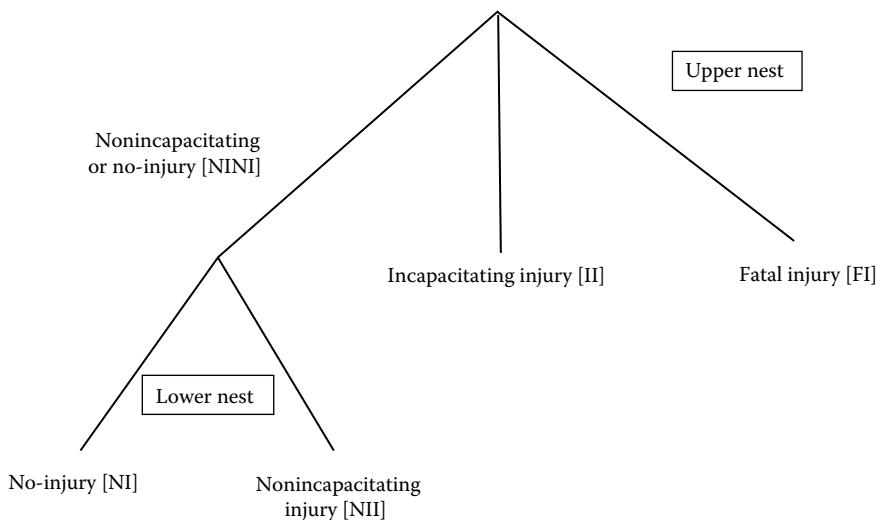
Variable No.	Variable Description
1	Crash identifier (1 or 0)
2	1 if no injury, 0 otherwise
3	1 if nonincapacitating injury, 0 otherwise
4	1 if incapacitating injury, 0 otherwise
5	1 if fatal injury, 0 otherwise
6	Master record number for the crash
7	Motorcycle age in years
8	1 if motorcycle was a Harley-Davidson, 0 otherwise
9	1 if motorcycle was a Honda, 0 otherwise
10	1 if motorcycle was a Yamaha, 0 otherwise
11	1 if motorcycle was a Suzuki, 0 otherwise
12	1 if motorcycle was a Kawasaki, 0 otherwise
13	1 if motorcycle was a Sportbike, 0 otherwise
14	1 if passenger on motorcycle, 0 otherwise
15	1 if collision with animal, 0 otherwise
16	1 if collision with guardrail, 0 otherwise
17	1 if collision with tree, 0 otherwise
18	1 if collision with pedestrian or bicycle, 0 otherwise
19	1 if collision with curb, 0 otherwise
20	1 if collision with pole, 0 otherwise
21	1 if collision with parked vehicle, 0 otherwise
22	1 if collision with wall, 0 otherwise
23	1 if collision with culvert, 0 otherwise
24	1 if collision with ditch, 0 otherwise

TABLE 13.3 (continued)

Variables Available for Motorcycle Injury-Severity Model Estimation

Variable No.	Variable Description
25	1 if collision with embankment, 0 otherwise
26	1 if traffic control devices were present, 0 otherwise
27	Motorcyclist age in years
28	1 if female motorcyclist, 0 otherwise
29	1 if motorcyclist ejected, 0 otherwise
30	1 if motorcyclist pinned, 0 otherwise
31	1 if motorcyclist wearing helmet, 0 otherwise
32	Month of accident:1= January....12 = December
33	Year of crash
34	1 if on local roads, 0 otherwise
35	1 if on state or U.S. roads, 0 otherwise
36	1 if on interstate, 0 otherwise
37	1 if in rural area, 0 otherwise
38	1 if crash at dawn or dusk, 0 otherwise
39	1 if crash in dark, 0 otherwise
40	1 if roadway lighted, 0 otherwise
41	1 if pavement wet, 0 otherwise
42	1 if at intersection, 0 for otherwise
43	1 if on horizontal curve, 0 for otherwise
44	1 if on grade, 0 otherwise
45	1 if on crest vertical curve, 0 otherwise
46	1 if vehicle failure identified, 0 otherwise
47	1 if crash is run-off-the-road, 0 for otherwise
48	1 if alcohol involved (motorcyclist drinking), 0 for otherwise
49	1 if speeding involved (motorcyclist speeding), 0 otherwise
50	1 if no license, 0 otherwise
51	1 if probationary operators license, 0 otherwise
52	1 if operator license with motorcycle endorsement, 0 otherwise
53	1 if learner license, 0 otherwise
54	1 if learner permit, 0 otherwise
55	1 if operator license, 0 otherwise
56	1 if commercial driver's license, 0 otherwise
57	1 if chauffeur's license, 0 otherwise
58	Number of months since motorcyclist's Basic Rider Course was completed
59	1 if motorcyclist passed Basic Rider Course, 0 otherwise
60	Posted speed limit in miles per hour

It is speculated that the less severe accident injuries (no injury and nonincapacitating injuries) may share unobserved effects resulting in a nesting structure as shown in Figure 13.6. As shown in Savolainen and Mannering (2007), estimation of this nested structure produces the parameter estimates given in Table 13.4.

**FIGURE 13.6**

Nested logit structure of the single-vehicle motorcycle crash-injury severity model.

TABLE 13.4

Full-Information Maximum Likelihood Estimation Results for Single-Vehicle Motorcyclist Crash-Injury Severity

Variable Description*	Estimated Parameter	t-Statistic
<i>Lower nest</i>		
Constant [NI]	-1.368	-11.94
Wet pavement indicator (1 if wet pavement, 0 if not) [NI]	0.880	2.86
Helmet indicator (1 if helmet used, 0 if not) [NI]	0.529	4.50
Intersection indicator (1 if accident at intersection, 0 if not) [NI]	0.380	2.82
Motorcycle age indicator (1 if motorcycle less than 5 years old, 0 if not) [NI]	0.278	2.38
Speeding indicator (1 if motorcyclist speeding, 0 if not) [NII]	0.674	3.49
Alcohol indicator (1 if motorcyclist drinking, 0 if not) [NII]	1.445	3.33
Passenger indicator (1 if motorcycle had a passenger, 0 if not) [NII]	0.658	3.80
Female indicator (1 if motorcyclist was female, 0 if not) [NII]	0.773	3.50
<i>Upper nest</i>		
Constant [II]	-2.749	-10.90
Constant [FI]	-5.702	-15.10
Speeding indicator (1 if motorcyclist speeding, 0 if not) [FI]	1.301	3.96
April indicator (1 if accident occurred in April, 0 if not) [FI]	0.959	2.61
July indicator (1 if accident occurred in July, 0 if not) [FI]	0.909	2.98

TABLE 13.4 (continued)

Full-Information Maximum Likelihood Estimation Results for Single-Vehicle Motorcyclist Crash-Injury Severity

Variable Description*	Estimated Parameter	t-Statistic
Basic Riding Course indicator (1 if motorcyclist took the basic riding course more than 2 years ago, 0 if not) [FI]	1.262	2.32
Run-off-road indicator (1 if motorcycle ran off the road, 0 if not) [FI]	1.231	3.95
Tree indicator (1 if motorcycle ran into a tree, 0 if not) [FI]	1.487	3.16
Pole indicator (1 if motorcycle ran into a pole, 0 if not) [FI]	1.388	3.13
Darkness indicator (1 if accident occurred in darkness, 0 if not) [FI]	0.700	2.55
Curb indicator (1 if motorcycle ran into a curb, 0 if not) [NINI]	-0.645	-2.63
Pole indicator (1 if motorcycle ran into a pole, 0 if not) [NINI]	-0.820	-2.53
Culvert indicator (1 if motorcycle ran into a culvert, 0 if not) [NINI]	-1.425	-3.17
Horizontal indicator (1 if accident occurred on a horizontal curve, 0 if not) [NINI]	-0.324	-2.62
Motorcyclist age in years [NINI]	-0.021	-4.38
Alcohol indicator (1 if motorcyclist drinking, 0 if not) [NINI]	-1.032	-2.49
Speeding indicator (1 if motorcyclist speeding, 0 if not) [NINI]	-0.676	-2.96
Helmet indicator (1 if helmet used, 0 if not) [NINI]	0.618	4.55
Darkness indicator (1 if accident occurred in darkness, 0 if not) [NINI]	-0.347	-2.34
Guardrail indicator (1 if motorcycle ran into a guardrail, 0 if not) [NINI]	-0.738	-2.82
Tree indicator (1 if motorcycle ran into a tree, 0 if not) [NINI]	-1.462	-3.59
Speed limit indicator (1 if accident occurred on a road with speed limit over 50 mph, 0 if not) [NINI]	-0.423	-3.30
Inclusive value (logsum) [NINI]	0.421	1.70
Number of Observations	2273	
Log-Likelihood at zero	-3481.50	
Log-Likelihood at convergence	-1967.66	

* Parameter defined for: [NI] No-injury; [NII] Nonincapacitating injury; [II] Incapacitating injury; [FI] Fatal injury; [NINI] Nonincapacitating or no injury.

Note that in this table, the inclusive value (logsum) parameter in the nonincapacitating or no-injury branch of the upper nest [NINI] is estimated as 0.421 with a *t*-statistic of 1.70. This reported *t*-statistic is computed to determine if the parameter is significantly different from zero. Thus the associate standard error is 0.247 (0.421 divided by 1.70). When comparing to see if the estimated parameter is significantly different from one, as is required to justify the nested structure as opposed to a simple MNL structure (Equation 13.51 is applied as)

$$t^* = \frac{\beta - 1}{S.E.(\beta)} = \frac{0.421 - 1}{0.247} = -2.34$$

Using a one-tailed t test gives a confidence level of over 99%, providing convincing evidence that the estimated parameter is less than 1.0. Thus the nested logit model structure is preferred relative to the simple MNL structure indicating that there is correlation between the nonincapacitating outcome and no-injury outcome disturbance terms.

13.7 Special Properties of Logit Models

Two special properties of logit models are noteworthy: (1) subsampling of alternate outcomes for model estimation and (2) using the logsum to estimate consumer welfare effects. There are many discrete outcome situations where the number of alternate outcomes is large. For example, to estimate which make, model, and vintage of vehicle someone may choose to own, a choice model might have thousands of alternative outcomes such as a 2010 Honda Accord, 2009 Honda Accord, 2010 Toyota Camry, 2009 Toyota Camry, and so on. Fortunately, the independently identically distributed extreme value distribution used in the derivation of the MNL model permits consistent estimation of model parameters using a subsample of the available outcome set (McFadden 1978). The estimation procedure is one of reducing the set of outcomes available to each observation to a manageable size. In so doing, one must include the outcome observed for each observation in the estimation sample and supplement this set with additional outcome possibilities that are selected randomly from the complete outcome set (a different set of randomly chosen outcomes is generated for each observation). For example, suppose there are just two consumer observations of individuals choosing over 3,500 different makes, models, and vintages of vehicles. Observation 1 is observed owning a 2004 Mazda Miata and observation 2 is observed owning a 2009 Dodge Caravan. If 10 outcomes are available for estimation, observation 1's outcome set would include the 2004 Mazda Miata (with associated attributes being the \mathbf{X} vector) in addition to 9 other make, model, and vintage options drawn randomly from the 3,500 available vehicles. Observation 2's outcome set would include a 2009 Dodge Caravan (with associated attributes being the \mathbf{X} vector) and 9 different other make, model, and vintage options drawn randomly from the 3,500 available vehicles. This subsampling of alternate outcomes is legitimate for estimation of the β but when one estimates the probabilities and logsums the entire set of outcome alternatives must be used (enumeration through the entire set of outcomes, not just the subsample, must be conducted to get estimates of outcome probabilities). See Mannering and Winston (1985) for an example of this sampling approach using a nested logit modeling structure of vehicle choice.

When logit models are coupled with the theory of utility maximization, the denominator is used to compute important welfare effects. The basis for this calculation is the concept of compensating variation (CV), which is the (hypothetical) amount of money that individuals will pay (or borrow) to make them as well off after a change in \mathbf{X} as they were prior to the change in \mathbf{X} . Small and Rosen (1981) show that the CV for each observation n is

$$CV = -\left(\frac{1}{\lambda}\right) \left[-LN \sum_{\forall I} EXP(\beta_I X_{In}) \right]_{\beta_I X_{In}^o}^{\beta_I X_{In}^f} \quad (13.52)$$

where λ is the marginal utility of income, \mathbf{X}^o refers to initial values of \mathbf{X} , \mathbf{X}^f refers to final values of \mathbf{X} (after the change), and all other terms are as defined previously. In most applications the marginal utility of income is equal in magnitude but opposite in sign to the cost parameter associated with alternate outcomes—the cost parameter estimated in the discrete outcome models. The procedure is to apply Equation 13.52 to the sample (by enumerating all observations) and then expand the results to the population (using the proportion of the sample in the overall population as a basis for this expansion). CV is a powerful tool and has been used to estimate the value consumers place on automobile safety devices (Winston and Mannering 1984), optimal traffic signal timing (Mannering et al. 1990) and the value of various transportation modes in urban areas (Niemeier 1997).

14

Ordered Probability Models

In many transportation applications discrete data are ordered. Examples include when survey respondents are asked for quantitative ratings (e.g., on a scale from 1 to 10 rate the following), ordered opinions (e.g., do you disagree, are neutral, or you agree), or categorical frequency data (e.g., property damage-only crash, injury crash, and fatal crash). While these response data are discrete, as were the data used in the discrete outcome models discussed in Chapter 13, application of the standard or nested multinomial discrete models presented previously do not account for the ordinal nature of the discrete data and thus all information reflected by the ordering is lost. Amemiya (1985) showed that if an unordered model (such as the multinomial logit model [MNL]) is used to model ordered data, the model parameter estimates remain consistent but there is a loss of efficiency. Ordered probability models have been developed to address the problem of ordered discrete data. Due to the restrictions placed on how variables are believed to affect ordered discrete outcome probabilities, not all ordinal data are best modeled using ordered probability models. There are cases where unordered probability models are preferred, even though the data are ordered (these cases are discussed in more detail in Section 14.3).

14.1 Models for Ordered Discrete Data

Ordered probability models (both probit and logit) have been in widespread use since the mid-1970s (see McKelvey and Zavonia 1975). Ordered probability models are derived by defining an unobserved variable z that is used as a basis for modeling the ordinal ranking of data. This unobserved variable is typically specified as a linear function for each observation (n subscripting omitted), such that

$$z = \beta \mathbf{X} + \varepsilon \quad (14.1)$$

where \mathbf{X} is a vector of variables determining the discrete ordering for observation n , β is a vector of estimable parameters, and ε is a random disturbance.

Using this equation, observed ordinal data, y , for each observation are defined as

$$\begin{aligned} y &= 1 && \text{if } z \leq \mu_0 \\ y &= 2 && \text{if } \mu_0 < z \leq \mu_1 \\ y &= 3 && \text{if } \mu_1 < z \leq \mu_2 \\ y &= \dots \\ y &= I && \text{if } z \geq \mu_{I-1} \end{aligned} \quad (14.2)$$

where μ are estimable parameters (referred to as thresholds) that define y , which corresponds to integer ordering, and I is the highest integer ordered response. Note that during estimation, nonnumerical ordered responses such as "never," "sometimes," and "frequently" are converted to integers (e.g., 1, 2, and 3) without loss of generality.

The μ are parameters that are estimated jointly with the model parameters β . The estimation problem then becomes one of determining the probability of I specific ordered responses for each observation n . This determination is accomplished by making an assumption on the distribution of ε . If ε are assumed to be normally distributed across observations with mean = 0 and variance = 1, an ordered probit model results with the ordered selection probabilities being

$$\begin{aligned} P(y = 1) &= \Phi(-\beta X) \\ P(y = 2) &= \Phi(\mu_1 - \beta X) - \Phi(-\beta X) \\ P(y = 3) &= \Phi(\mu_2 - \beta X) - \Phi(\mu_1 - \beta X) \\ P(y = I) &= 1 - \Phi(\mu_{I-1} - \beta X) \end{aligned} \quad (14.3)$$

where $\Phi(\cdot)$ is the cumulative normal distribution

$$\Phi(\mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left[-\frac{1}{2}w^2\right] dw \quad (14.4)$$

Note that in Equation 11.56, threshold μ_0 is set equal to zero without loss of generality (this implies that one need only estimate $I - 2$ thresholds). Figure 14.1 provides an example with five possible ordered outcomes. For estimation, Equation 14.3 is written as

$$P(y = i) = \Phi(\mu_i - \beta X) - \Phi(\mu_{i+1} - \beta X) \quad (14.5)$$

where μ_i and μ_{i+1} represent the upper and lower thresholds for outcome i . The likelihood function is (over the population of N observations)

$$L(y | \beta_1, \dots, \beta_k, \mu_2, \dots, \mu_{I-1}) = \prod_{n=1}^N \prod_{i=1}^I [\Phi(\mu_i - \beta X_n) - \Phi(\mu_{i+1} - \beta X_n)]^{\delta_{in}} \quad (14.6)$$

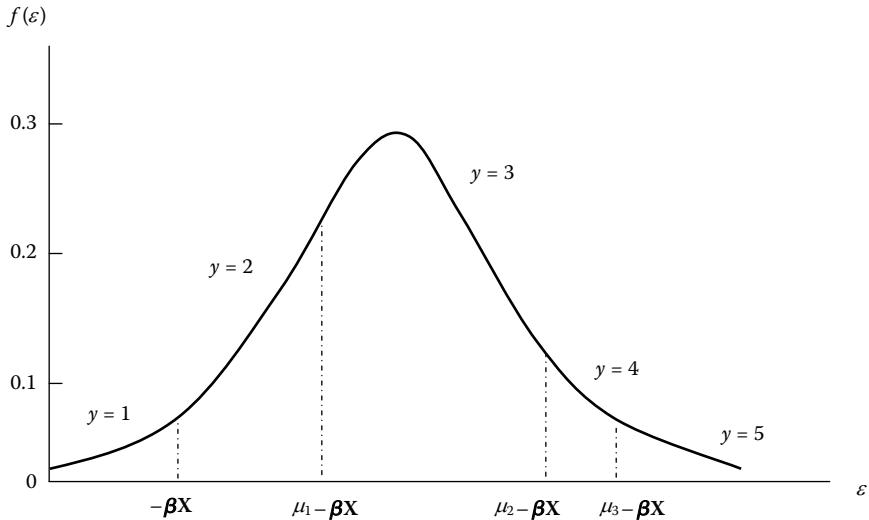
**FIGURE 14.1**

Illustration of an ordered probability model with $\mu_0 = 0$.

where δ_{in} is equal to 1 if the observed discrete outcome for observation n is i , and zero otherwise. This equation leads to a log-likelihood of

$$LL = \sum_{n=1}^N \sum_{i=1}^I \delta_{in} \ln [\Phi(\mu_i - \boldsymbol{\beta} \mathbf{X}_n) - \Phi(\mu_{i+1} - \boldsymbol{\beta} \mathbf{X}_n)] \quad (14.7)$$

Maximizing this log-likelihood function is subject to the constraint $0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_{I-1}$. If the assumption is made that e in Equation 14.1 is logistically distributed across observations with mean = 0 and variance = 1, an ordered logit model results and the derivation proceeds the same as for the ordered probit model. Because the ordered probit model is not afflicted with the estimation difficulties encountered in estimating the multinomial probit model for unordered discrete data, the ordered probit is usually chosen over the ordered logit because of the underlying assumption of normality.

In terms of evaluating the effect of individual estimated parameters in ordered probability models, Figure 14.2 shows that a positive value of β_k implies that an increase in x_k will unambiguously increase the probability that the highest ordered discrete category results ($y = 5$ in Figure 14.2) and unambiguously decreases the probability that the lowest ordered discrete category results ($y = 1$ in Figure 14.2).

A practical difficulty with ordered probability models is associated with the interpretation of intermediate categories, ($y = 2$, $y = 3$, and $y = 4$ in Figure 14.2). Depending on the location of the thresholds, it is not

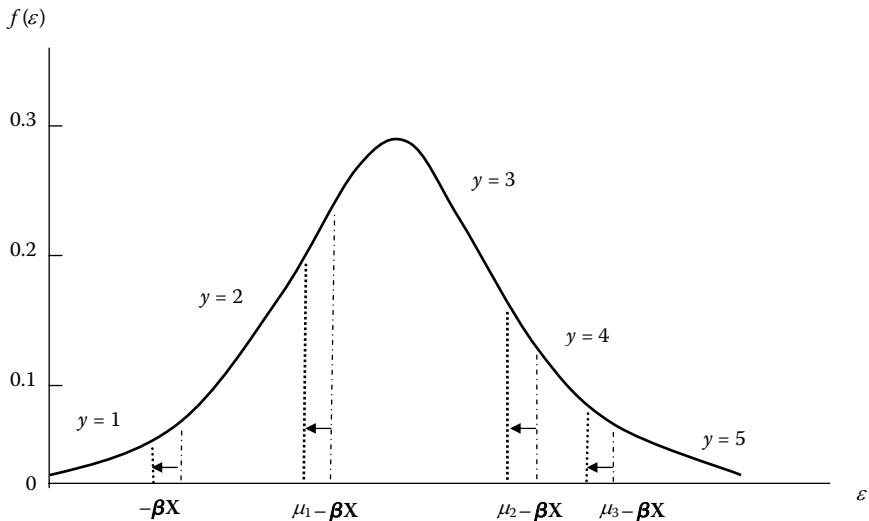
**FIGURE 14.2**

Illustration of an ordered probability models with an increase in βX (with $\mu_0 = 0$).

necessarily clear what effect a positive or negative β_k has on the probabilities of these “interior” categories. This difficulty arises because the areas between the shifted thresholds may yield increasing or decreasing probabilities after shifts to the left or right (see Figure 14.2). For example, suppose the five response categories in Figure 14.2 were; disagree strongly, disagree, neutral, agree, agree strongly. It is tempting to interpret a positive β_k as implying that an increase x_k will increase the likelihood of agreeing but this conclusion is incorrect because of the ambiguity in the interior category probabilities. The correct interpretation is that an increase in x_k increases the likelihood of agreeing strongly and decreases the likelihood of disagreeing strongly.

To obtain a sense of the direction of the effects on the interior categories, marginal effects are computed for each category. For indicator variables, the effects are computed as the difference in the estimated probabilities with the indicator variable changing from zero to one, while all other variables are equal to their means. For continuous variables, the effects are computed from the partial derivatives

$$\frac{\partial P(y=1)}{\partial \mathbf{X}} = -\phi(-\boldsymbol{\beta} \mathbf{X}) \boldsymbol{\beta}'$$

$$\frac{\partial P(y=2)}{\partial \mathbf{X}} = [\phi(\mu_0 - \boldsymbol{\beta} \mathbf{X}) - \phi(\mu_1 - \boldsymbol{\beta} \mathbf{X})] \boldsymbol{\beta}'$$

$$\begin{aligned}
 \frac{\partial P(y = 3)}{\partial \mathbf{X}} &= [\phi(\mu_1 - \boldsymbol{\beta}\mathbf{X}) - \phi(\mu_2 - \boldsymbol{\beta}\mathbf{X})] \boldsymbol{\beta}' \\
 \frac{\partial P(y = \dots)}{\partial \mathbf{X}} &= \dots \\
 \frac{\partial P(y = I)}{\partial \mathbf{X}} &= -\phi(\mu_{I-2} - \boldsymbol{\beta}\mathbf{X}) \boldsymbol{\beta}' \tag{14.8}
 \end{aligned}$$

where $P(y = j)$ is the probability of response category j , $\phi(\cdot)$ is the standard normal density, and all other terms are as previously defined.

The marginal effects for each response category are interpreted as a change in the outcome probability of each threshold category $P(y = j)$ given a unit change in a continuous variable x . In the context of user-perceptions of roadway roughness, a large marginal effect (in absolute value terms) indicates the parameter has a relatively large effect on the users' roughness rankings, while a relatively small marginal effect indicates a relatively minimal effect on users' roughness rankings. A positive marginal effect for a specific roughness ranking indicates an increase in probability for that ranking, while a negative value would correspond to a decrease in probability for that ranking.

Example 14.1

A survey of 281 commuters was conducted in the Seattle metropolitan area. The survey's intent was to gather information on commuters' opinions of high-occupancy vehicle (HOV) lanes (lanes that are restricted for use by vehicles with 2 or more occupants). The variables obtained from this study are provided in Table 14.1.

Among other questions, commuters were asked whether they agreed with the statement "HOV lanes should be open to all vehicles, regardless of vehicle occupancy level" (variable number 29 in Table 14.1). The question provided ordered responses of; strongly disagree, disagree, neutral, agree, agree strongly, and the observed percentage frequency of response in these five categories was 32.74, 21.71, 8.54, 12.10, and 24.91, respectively. To understand the factors determining commuter opinions, an ordered probit model of this survey question is appropriate.

Ordered probit estimation results are presented in Table 14.2. The results show that commuters who normally drive alone and are 50 years old or older are more likely to agree strongly, and less likely to disagree strongly that HOV lanes should be open to all traffic. The findings also show that as income increases and the more frequently a commuter is observed varying from normal route and departure times (presumably in an effort to avoid traffic congestion, the more likely the commuter is to agree strongly (and less likely to disagree strongly). Finally, commuters with flexible work hours were also likely to agree strongly (more likely to disagree strongly). Statistical assessment of this model is the same as that for other maximum likelihood models (t -statistics, $\rho^2 \chi^2$, likelihood ratio tests, etc.). The estimated

TABLE 14.1

Variables Available to Study High-Occupancy Vehicle (HOV) Lanes Opinions

Variable No.	Variable Description
1	Usual mode of travel: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if vanpool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
2	Have used HOV lanes: 1 if yes, 0 if no
3	If used HOV lanes, what mode is most often used: 0 in a bus, 1 in two person carpool, 2 in three or more person carpool, 3 in vanpool, 4 alone in vehicle, 5 on motorcycle
4	Sometimes eligible for HOV lane use but do not use: 1 if yes, 0 if no
5	Reason for not using HOV lanes when eligible: 0 if slower than regular lanes, 1 if too much trouble to change lanes, 2 if HOV lanes are not safe, 3 if traffic moves fast enough, 4 if forgot to use HOV lanes, 5 if other
6	Usual mode of travel one year ago: 0 if drive alone, 1 if two person carpool, 2 if three or more person carpool, 3 if vanpool, 4 if bus, 5 if bicycle or walk, 6 if motorcycle, 7 if other
7	Commuted to work in Seattle a year ago: 1 if yes, 0 if no
8	Have flexible work start times: 1 if yes, 0 if no
9	Changed departure times to work in the last year: 1 if yes, 0 if no
10	On average, number of minutes leaving earlier for work relative to last year
11	On average, number of minutes leaving later for work relative to last year
12	If changed departure times to work in the last year, reason why: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
13	Changed route to work in the last year: 1 if yes, 0 if no
14	If changed route to work in the last year, reason why: 0 if change in travel mode, 1 if increasing traffic congestion, 2 if change in work start time, 3 if presence of HOV lanes, 4 if change in residence, 5 if change in lifestyle, 6 if other
15	Usually commute to or from work on Interstate 90: 1 if yes, 0 if no
16	Usually commuted to or from work on Interstate 90 last year: 1 if yes, 0 if no
17	On your past five commutes to work, how often have you used HOV lanes
18	On your past five commutes to work, how often did you drive alone
19	On your past five commutes to work, how often did you carpool with one other person
20	On your past five commutes to work, how often did you carpool with two or more people
21	On your past five commutes to work, how often did you take a vanpool
22	On your past five commutes to work, how often did you take a bus
23	On your past five commutes to work, how often did you bicycle or walk
24	On your past five commutes to work, how often did you take a motorcycle

TABLE 14.1 (continued)

Variables Available to Study High-Occupancy Vehicle (HOV) Lanes Opinions

Variable No.	Variable Description
25	On your past five commutes to work, how often did you take a mode other than those listed in variables 18 through 24
26	On your past five commutes to work, how often have you changed route or departure time
27	HOV lanes save all commuters time: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
28	Existing HOV lanes are being adequately used: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
29	HOV lanes should be open to all traffic: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
30	Converting some regular lanes to HOV lanes is a good idea: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
31	Converting some regular lanes to HOV lanes is a good idea only if it is done before traffic congestion becomes serious: 0 if strongly disagree, 1 if disagree, 2 if neutral, 3 if agree, 4 if agree strongly
32	Gender: 1 if male, 0 if female
33	Age in years: 0 if under 21, 1 if 22–30, 2 if 31–40, 3 if 41–50, 4 if 51–64, 5 if 65 or greater
34	Annual household income (US. dollars per year): 0 if no income, 1 if 1–9,999, 2 if 10,000–19,999, 3 if 20,000–29,999, 4 if 30,000–39,999, 5 if 40,000–49,999, 6 if 50,000–74,999, 7 if 75,000–100,000, 8 if over 100,000
35	Highest level of education: 0 if did not finish high school, 1 if high school, 2 if community college or trade school, 3 if college/university, 4 if postcollege graduate degree
36	Number of household members
37	Number of adults in household (aged 16 or more)
38	Number of household members working outside the home
39	Number of licensed motor vehicles in the household
40	Postal zip code of work place
41	Postal zip code of home
42	Type of survey comment left by respondent regarding opinions on HOV lanes: 0 if no comment on HOV lanes, 1 if comment not in favor of HOV lanes, 2 comment positive toward HOV lanes but critical of HOV lane policies, 3 comment positive toward HOV lanes, 4 neutral HOV lane comment

threshold values do not have any direct interpretation, but are need to compute outcome probabilities and marginal effects.

The marginal effects for the model are presented in Table 14.3. Each commuter will have their own marginal effects, which depend on the values of \mathbf{X} and the values presented in Table 14.3 are the marginal effects are averaged over the population and give the change in category probabilities resulting from a one unit

TABLE 14.2

Ordered Probit Estimation Results for the Question “High-Occupancy Vehicle (HOV) Lanes Should be Open to All Vehicles, Regardless of Vehicle Occupancy Level” (Disagree Strongly, Disagree, Neutral, Agree, Agree Strongly)

Independent Variable	Estimated Parameter	t-Statistic
Constant	0.707	-3.92
Drive alone indicator (1 if commuter’s usual mode is drive alone, 0 otherwise)	1.144	7.02
Flexible work-time indicator (1 if commuter has flexible work hours, 0 other wise)	0.283	2.06
High-income indicator (1 if commuter’s household income is greater than \$100,000 per year, 0 otherwise)	0.365	1.59
Older age indicator (1 if commuter is 50 years old or older, 0 otherwise)	0.249	1.43
Number of times in the past 5 commutes the commuter changed from normal routes or departure time	0.081	1.59
Threshold 1	0.644	8.61
Threshold 2	0.890	10.55
Threshold 3	1.27	13.10
Number of observations	281	
Log-likelihood at zero	-424.05	
Log-likelihood at convergence	-395.65	
ρ^2	0.067	

change in the independent variable. Indicator variables provide the change in category probabilities following a change in the variable from zero to one. When looking at Table 14.2 and using the signs of the parameter estimates, statements can only be made as to what will happen at the extreme (highest and lowest) categories. For example, a commuter that normally drives alone is more likely to agree strongly and less likely to disagree strongly. The marginal effects shown in Table 14.3 provide more information on the magnitude of the change as well as what happens with the “interior” probability categories. For commuters that drive alone (relative to those that do not), their probability of disagreeing strongly is (on average) 0.4050 lower and disagreeing 0.0456 lower. Their probability of being neutral is 0.0299 higher, agreeing 0.841 higher, and agreeing strongly 0.3366 higher.

14.2 Ordered Probability Models with Random Effects

As in estimating other types of models, databases sometimes contain observations with obvious correlations. For example, correlation can occur in

TABLE 14.3

Computed Marginal Average Effects for Parameter Estimates Shown in Table 14.2 ($y = 1$ [disagree strongly], $y = 2$ [disagree], $y = 3$ [neutral], $y = 4$ [agree], $y = 5$ [agree strongly])

Variable	Marginal Effects				
	[$y = 1$]	[$y = 2$]	[$y = 3$]	[$y = 4$]	[$y = 5$]
Drive alone indicator (1 if commuter's usual mode is drive alone, 0 otherwise)	-0.4050	-0.0456	0.0299	0.0841	0.3366
Flexible work-time indicator (1 if commuter has flexible work hours, 0 otherwise)	-0.1002	-0.0113	0.0074	0.0208	0.0833
High-income indicator (1 if commuter's household income is greater than \$100,000 per year, 0 otherwise)	-0.1291	-0.0145	0.0095	0.0268	0.1073
Older age indicator (1 if commuter is 50 years old or older, 0 otherwise)	-0.0883	-0.0099	0.0065	0.0183	0.0734
Number of times in the past 5 commutes the commuter changed from normal routes or departure time	-0.0285	-0.0032	0.0021	0.0059	0.0237

time-series data when several observations of an individual are made at multiple points in time. Or in cross-sectional data when a single individual is observed repeatedly and is in the dataset multiple times. In both cases, correlation of model disturbances is a concern because the unobserved characteristics of each individual will likely be correlated. Such disturbance correlation will violate disturbance independence assumptions and result in erroneous parameter estimates.

This problem is addressed with a random effects model, which allows for an individual-specific disturbance term (in addition to an overall disturbance term) to account for random disturbances specific to each individual. By rewriting Equation 14.1, the traditional disturbance term are considered as two: the traditional disturbance term unique to each observation, ε_{ic} , and an individual-specific random effect disturbance term, φ_i , (assumed to be normally distributed with mean 0 and variance σ^2)

$$z_{ic} = \beta \mathbf{X}_{ic} + \varepsilon_{ic} + \varphi_i \quad (14.9)$$

where subscripting i denotes individuals, subscripting c indexes the group of observations generated by each individual, and all other terms are as previously defined. While one may expect that a correlation exists between observations generated from the same individual, this may not always be the case. To test for correlation, the Hausman test statistic σ is typically estimated as part of the random effects model and determines the significance of the random effects formulation relative to the standard ordered logit model (Hausman 1978).

Example 14.2

A survey of 56 subjects was conducted on freeways in the Seattle area. Each subject drove a vehicle over 40 freeway segments (thus each subject can generate as many as 40 observations if there is no missing data). As they drove over the test segments, they were asked: "How would you rank the roughness of the road on a scale from one to five—with one being the smoothest (or the best) and five being the roughest (or the worst)?" Data were collected on the type of vehicle being used (minivan, pickup, etc.), in-vehicle-cabin noise (A-scale decibels, dBA), vehicle speed (km/h), socioeconomic information, the pavement's international roughness index (IRI) measurement, age of the roadway surface, information on patching, and the pavement structural condition (PSC), which is calculated separately for flexible and rigid pavements based on the amount and severity of various distresses and its values range from 100 (excellent pavement condition) to zero (completely deteriorated pavement). The variables available for model estimation are shown in Table 14.4. A subset of these data (using 31 subject and 37 freeway segments) was used previously in a research paper by Shafizadeh and Mannerling (2006).

The estimation results are shown in Table 14.5. With regard to the importance of the random effects, note that the Random effect (Hausman test) parameter (σ) is significantly different from zero with a parameter estimate of 0.753 and a t -statistic 14.14. This high level of significance clearly shows that estimating the model with random effects is valid and that ignoring the correlation in disturbances among

TABLE 14.4

Variables Available to Study Pavement Roughness Opinions

Variable No.	Variable Description
1	Individual number
2	Roadway segment number
3	Indicator for number of observations from respondent
4	Roughness ranking: 1 if very smooth; 5 if very rough
5	Sedan: 1 if yes, 0 if no
6	Sport utility vehicle: 1 if yes, 0 if no
7	Pickup: 1 if yes, 0 if no
8	Minivan: 1 if yes, 0 if no

TABLE 14.4 (continued)

Variables Available to Study Pavement Roughness Opinions

Variable No.	Variable Description
9	Noise dBA reading
10	Speed in miles per hour
11	Highway level of service (LOS): 1 if LOS A, 2 if LOS B, 3 if LOS C, 4 if LOS D, 5 if LOSE, 6 if LOS F
12	User regularly uses Interstate 5: 1 if yes, 0 if no
13	User regularly uses Interstate 90: 1 if yes, 0 if no
14	User regularly uses Interstate 405: 1 if yes, 0 if no
15	User regularly uses State Route 520: 1 if yes, 0 if no
16	Female: 1 if yes, 0 if no
17	Married: 1 if yes, 0 if no
18	Age in years: 0 if Less than 21; 1 if 21 to 25; 2 if 26 to 30; 3 if 31 to 35; 4 if 36 to 40; 5 if 41 to 45; 6 if 46 to 50; 7 if 51 to 55; 8 if 56 to 60; 9 if 61 to 65; 10 if 66 to 70; 11 if Over 70
19	Annual household income (U.S. dollars): 0 if no income; 1 if under \$15,000; 2 if \$15,000–\$24,999; 3 if \$25,000–\$34,999; 4 if \$35,000–\$44,999; 5 if \$45,000–\$54,999; 6 if \$55,000–\$64,999; 7 if \$65,000–\$74,999; 8 if \$75,000–\$84,999; 9 if \$85,000–\$99,999; 10 if \$100,000–\$150,000; 11 if over \$150,000
20	Education: 1 if some high school; 2 if high school diploma; 3 if technical college degree (AA); 4 if college degree (BS or BA) 5 if postgraduate degree
21	Vehicle type normally driven: (miscoded, do not use)
22	Number of household vehicles
23	Household size
24	Number of household infants
25	Number of household children
26	Number of household workers
27	International roughness index (IRI) in m/km
28	Roadway surface age
29	Visible wear: 1 if yes, 0 if no
30	Visible joints: 1 if yes, 0 if no
31	Visible patching: 1 if yes, 0 if no
32	Bridge in segment: 1 if yes, 0 if no
33	Surface type: 1 if concrete, 0 if asphalt
34	Rut depth in mm
35	Pavement structural condition index (PSC)
36	Segment length in miles
37	Number of lanes
38	Cracking present: 1 if yes, 0 if no
39	Scaling present: 1 if yes, 0 if no
40	Faulting present: 1 if yes, 0 if no
41	Spalling present: 1 if yes, 0 if no
42	International roughness index (IRI) change from last segment (m/km)
43	Noise change from last segment (dBA)

individual participant pavement ratings would result in serious model specification error leading to inconsistent parameter estimates.

Focusing on individual parameter estimates, the results presented in Table 14.5 show that female participants and those participants over 55 years old were more likely to think the pavement was very smooth and less likely to think the pavement was very rough. This finding suggests that these participants may have a higher tolerance or a different definition of what constitutes a rough pavement. As the IRI increased (indicating a rougher pavement), the age of the pavement surface increased, and when patching of the pavement surface was visible, the pavement was more likely to be rated very rough and less likely to be rated very smooth. These findings are consistent with expectations. Also, as the higher the PSC index (indicating a better PSC) the lower the likelihood that the pavement would be rated very rough and the higher the likelihood that it would be rated very smooth.

TABLE 14.5

Random Effects Ordered Probit Model of User-Perceived Roughness Rankings
(Dependent Variable Responses are Integers between 1 [Very Smooth] and 5
[Very Rough])

Variable Description	Estimated Parameter	t-Statistic
Constant	1.910	5.08
Gender indicator (1 if participant was female, 0 if male)	-0.568	-6.15
Older age indicator (1 if participant was over age 55, 0 otherwise)	-0.599	-5.86
IRI measurement (m/km) of roadway segment	0.797	15.81
Age of roadway segment surface (years)	0.018	5.68
Patch indicator (1 if the segment appeared to have patch work, 0 otherwise)	0.176	2.15
Pavement structural condition (PSC) index of roadway	-0.022	-6.07
Noise (dBA) inside test vehicle during evaluation if age less than 35 years old	0.005	5.47
Noise increase indicator (1 if the noise inside test vehicle during evaluation increases by 3 dBA or more between two adjacent test segments, 0 otherwise)	0.721	5.74
Threshold 1	1.630	34.75
Threshold 2	2.982	70.53
Threshold 3	4.403	86.77
Random effect (Hausman test) parameter, σ	0.753	14.14
Number of observations	2179	
Initial log-likelihood	-3187.27	
Log-likelihood at convergence	-2331.57	
ρ^2	0.268	

For those participants that were less than 35 years of age, higher noise levels inside the vehicle made them more likely to rate the pavement as very rough and less likely to rate it as very smooth, suggesting that younger participants may be more sensitive to noise levels. Finally, if the noise increase from one roadway segment to the next increased by more than 3 dBA, participants were more likely to rate the pavement as very rough and less likely to rate it as very smooth, indicating that the relative increase in sound is an important factor.

The average marginal effects for this model are presented in Table 14.6. As discussed in Example 14.1, marginal effects provide information on the magnitude of the change in all probability categories: a one unit increase in the roadway's IRI (1 m/km) results in an average 0.0069 increase in the probability that participants would rate the pavement as very rough (5 on the 1–5 scale), a 0.1209 increase in the probability that it would be rated 4 on the 1–5 scale, a 0.1209 increase in the probability that it would be rated 3 on the 1–5 scale, a 0.2339 decrease in the probability that it would be rated 2 on the 1–5 scale, and a 0.0841 decrease in the probability that it would be rated as very smooth (1 on the 1–5 scale).

TABLE 14.6

Computed Average Marginal Effects for the Parameter Estimates Shown in Table 14.4 ($y = 1$ [Very Smooth] to $y = 5$ [Very Rough])

Variable Description	Marginal Effects				
	[$y = 1$]	[$y = 2$]	[$y = 3$]	[$y = 4$]	[$y = 5$]
Gender indicator (1 if participant was female, 0 if male)	0.0600	0.1668	-0.1357	-0.0862	-0.0049
Older age indicator (1 if participant was over age 55, 0 otherwise)	0.0632	0.1758	-0.1430	-0.0908	-0.0052
IRI measurement (m/km) of roadway segment	-0.0841	-0.2339	0.1902	0.1209	0.0069
Age of roadway segment surface (years)	-0.0019	-0.0054	0.0044	0.0028	0.0001
Patch indicator (1 if the segment appeared to have patch work, 0 otherwise)	-0.0185	-0.0515	0.0419	0.0266	0.0015
Pavement structural condition (PSC)	0.0023	0.0064	-0.0052	-0.0033	-0.0002
Noise (dBA) inside test vehicle during evaluation if age less than 35 years old	-0.0006	-0.0015	0.0012	0.0008	0.0001
Noise increase indicator (1 if the noise inside test vehicle during evaluation increases by 3 dBA or more between two adjacent test segments, 0 otherwise)	-0.0761	-0.2116	0.1722	0.1094	0.0061

14.3 Limitations of Ordered Probability Models

As mentioned at the beginning of this chapter not all ordinal data are best modeled using ordered probability models since there are numerous occasions when an unordered probability model provides a superior fit to ordered data. Such cases arise because ordered probability models place a restriction on how variables in \mathbf{X} affect outcome probabilities. For example, suppose one is modeling the severity of vehicle accidents with outcomes of property damage only, injury, and fatality. These data are clearly ordered. Suppose that one of the key factors in determining the level of injury is whether an airbag was deployed. The air bag deployment indicator variable in an ordered model would either increase the probability of a fatality (and decrease the probability of property damage only) or decrease the probability of fatality (and increase the probability of property damage only). But the reality may be that the deployment of an airbag reduces the probability of a fatality but increases the probability of minor injuries (from air bag deployment). In an unordered discrete-modeling framework, this would result in the parameter for the air bag deployment variable having a negative value in the severity function for the fatality outcome and also a negative value for the property damage-only outcome (with the injury outcome having an increased probability in the presence of an air bag deployment). In this situation an ordered probability model is not appropriate because it does not have the flexibility to explicitly capture interior category probabilities.

When considering the choice between an ordered probability model and a traditional MNL, the traditional MNL model may have many advantages if the assumptions for the model hold (specifically, the independence of irrelevant alternatives [IIA] property). In addition to the ability to account for a more general influence of variables (see above), it also has the ability to handle outcome-based sampling without severe consequences on parameter estimation. That is, if the MNL IIA property holds a nonrandom sample of outcomes will still give correct parameter estimates for parameters except the alternative specific constant (see Section 13.5.4). In the case of ordered models in the presence of outcome-based samples, all parameters are erroneously estimated. An example of this type of sample is police-reported injury severity in vehicle accidents. Because lower-severity accidents (those involving property damage only) are less likely to be reported to police, most accident databases contain a significant underrepresentation of property damage-only accidents. Estimating a severity outcome model (where choices may be property damage only, possible injury, evident injury, disabling injury, and fatality) using an MNL model will result in correct parameter estimates for all variables except the constant terms. Estimating this model with an ordered probit or logit model will result in erroneous estimates of all model parameters. Yamamoto et al. (2008) discuss this point in detail.

In summary, one must be cautious in the selection of ordered and unordered models. A trade off is inherently being made between recognizing the ordering of responses and losing the flexibility in specification offered by unordered outcome models. In fact, there have been a number of journal articles that have addressed this point (Eluru et al. 2008; and Yamamoto et al. 2008).

15

Discrete/Continuous Models

Statistical approaches to study continuous data (ordinary least-squares regression) and discrete data (logit and probit models) have been used for decades in the analysis of transportation data. However, researchers have also identified a class of transportation-related problems that involve interrelated discrete and continuous data. Examples include consumers' choice of the type of vehicle to own (discrete) and the number of kilometers to drive it (continuous), choice of route (discrete) and driving speed (continuous), and choice of trip-generating activity (discrete) and duration in the activity (continuous).

Interrelated discrete/continuous data are easily overlooked and sometimes difficult to identify. The transportation literature and the literature in many other fields are strewn with well-known examples of data that have been erroneously modeled as continuous data, when the data are really part of an interrelated discrete/continuous process. The consequences of this oversight are serious—biased estimation results that could significantly alter the inferences and conclusions drawn from the data.

15.1 Overview of the Discrete/Continuous Modeling Problem

The problem of interrelated discrete/continuous data can best be viewed as a problem of selectivity, with observed data being an outcome of a selection process that results in a nonrandom sample of data in observed discrete categories. To better illustrate this concept, consider the route-choice problem presented previously in Chapter 13 (Example 13.1). In this example, commuters had a choice of three alternate routes for their morning commute (from home to work): a four-lane arterial (speed limit = 60 km/h, two lanes each direction), a two-lane highway (speed limit = 60 km/h, one lane each direction), and a limited-access four-lane freeway (speed limit = 90 km/h, two lanes each direction). If one were interested in studying average travel speed between the work and home, a classic discrete/continuous problem arises because the observed travel-speed data arise from a self-selection process. As such, commuters observed on each of the three routes represent a nonrandom sample because commuters' choice of route is influenced

by their propensity to drive faster. Commuters that tend to drive faster are more likely to choose the freeway, while commuters that tend to drive slower are more likely to choose the arterial. From a model estimation perspective, travel speeds are only observed on the chosen route. Thus, it is not known how fast observed arterial users would have driven had they taken the freeway, or how fast freeway users would have driven had they taken the arterial. If faster drivers are indeed more likely to choose the freeway and slower drivers more likely to choose the arterial, the speed behavior of arterial drivers on the freeway is likely to be different from the speed behavior of drivers observed on the freeway. If this interrelationship between route (discrete) and travel speed (continuous) is ignored, selectivity bias will result in the statistical models estimated to predict travel speed.

The selectivity-bias problem can best be illustrated by a simple example. Consider the estimation of a regression model of average travel speed on the freeway route from work to home

$$s_{fn} = \beta_f X_n + \xi_{fn} \quad (15.1)$$

where s_{fn} is the average speed of commuter n on the freeway, X_n is the vector of commuter n characteristics influencing average travel speed, β_f is the vector of estimable parameters, and ξ_{fn} is the unobserved characteristics influencing travel speed. The selectivity bias that would result in estimating this equation with observed data is shown in Figure 15.1. In this figure, commuter data indicated by a "+" represents the data of observed freeway users and commuter data indicated by a "-" represents the unobserved speeds of nonfreeway users had they chosen to

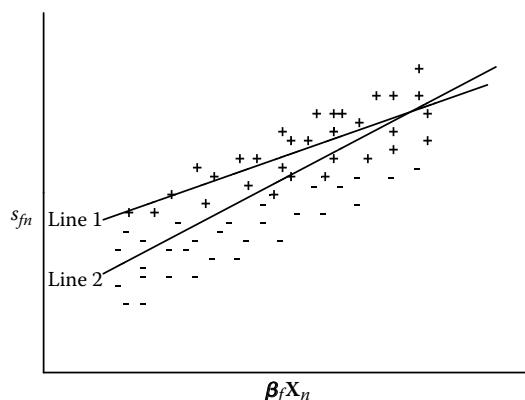


FIGURE 15.1

An illustration of self-selected bias in discrete-continuous data.

drive on the freeway. Because freeway users are a self-selected group (for whatever reasons) of faster drivers (with faster drivers being more likely to choose the freeway), they are underrepresented at lower values of $\beta_f X_n$ and over represented at higher values of $\beta_f X_n$. If Equation 15.1 is estimated using only observed data (observed freeway users only), the resulting estimates are biased as indicated by “line 1” in Figure 15.1. The true equation of freeway speed with all data (observed and unobserved) is given by “line 2.”

From an average speed-forecasting perspective, the problem with “line 1” is that any policy causing a shift in route choice will result in biased forecasts of average speed because of the selectivity bias present in equation estimation. For example, if the arterial was closed for reconstruction, the arterial users moving to the freeway would have a different β reflecting their speeds compared to the β of the originally observed freeway users. This problem also exists if a single equation is estimated for the average speeds on all three routes because each route contains a self-selected sample of commuters that is based on their speed preferences.

15.2 Econometric Corrections: Instrumental Variables and Expected Value Method

Perhaps the most simplistic method of addressing misspecification induced by discrete/continuous processes is to use an instrumental variables approach. To illustrate, consider Example 11.1 with commuters selecting from three possible routes (freeway, four-lane arterial, and two-lane road), and a revision of Equation 15.1 such that the average travel speed to work on commuter n 's chosen route is

$$s_n = \beta X_n + \alpha Z_n + \xi_n \quad (15.2)$$

where s_n is the average speed of commuter n on commuter n 's chosen route, X_n is the vector of commuter n characteristics influencing average travel speed that are not a function of the route chosen (such as income, driver age), Z_n is the vector of characteristics commuter n faces that influence average travel speed that are a function of the route chosen (such as number of traffic signals and travel distance), β and α are the corresponding vectors of estimable parameters, and ξ_n is the unobserved characteristics influencing travel speed. Direct estimation of Equation 15.2 results in biased and inefficient parameter estimates because Z_n is endogenous due to the discrete/continuous interrelation between travel speed and route choice. This endogeneity arises because as a commuter's choice of speed increases, elements

in the Z_n vector will change since the likelihood of selecting specific route is interrelated with speed preferences. As was the case in simultaneous equations models that involve all continuous variables (see Chapter 5), one could replace the elements of Z_n with estimates derived from regressing Z_n against all exogenous variables. The procedure consists of estimating regression equations for all elements of Z_n and to use regression predicted values \hat{Z}_n to estimate Equation 15.2 such that

$$s_n = \beta X_n + \alpha \hat{Z}_n + \xi_n \quad (15.3)$$

An alternative to the instrumental variables approach is to interact the endogenous variables directly with the corresponding discrete-outcome model (the route-choice model in this case). The most common approach is to replace the elements in the endogenous variable vector Z_n with their expected values. Thus, each element of Z_n , z_{jn} , is replaced by (with subscripting j denoting elements of Z_n)

$$E(z_{jn}) = \sum_{\forall i} P_n(i) z_{jn} \quad (15.4)$$

where $P_n(i)$ is the predicted probability of commuter n selecting discrete outcome i , as determined by a discrete-outcome model. Equation 15.2 then becomes

$$s_n = \beta X_n + \alpha \sum_{\forall j} \sum_{\forall i} P_n(i) z_{jn} + \xi_n \quad (15.5)$$

Example 15.1

To demonstrate the expected value selectivity-bias correction method, consider the data used previously in Example 13.1. Recall that these data are from a survey of 151 commuters departing from the same origin (a large residential complex) and going to work in the same downtown area. Distance is measured precisely from parking lot of origin to parking lot of destination so there is a variance in distances among commuters even though they depart and arrive in the same general areas. Commuters are observed choosing one of three alternate routes; a four-lane arterial, a two-lane highway, and a limited-access four-lane freeway. To estimate a model of average commuting speed, the variables shown in Table 15.1 are available.

For the equation correction, to estimate the expected value of endogenous variables that result because of the discrete/continuous interaction, the route-choice probabilities are determined from the logit model estimated in

TABLE 15.1

Variables Available for Home to Work Average Speed Model Estimation

Variable No.	Variable Description
1	Average commuting speed in km/h
2	Route chosen: 1 if arterial, 2 if rural road, 3 if freeway
3	Traffic flow rate at time of departure in vehicles per hour
4	Number of traffic signals on the selected route
5	Distance along the selected route in kilometers
6	Seat belts: 1 if wearing, 0 if not
7	Number of passengers in vehicle
8	Commuter age in years: 1 if less than 23, 2 if 24–29, 3 if 30–39, 4 if 40–49, 5 if 50 and over 50
9	Gender: 1 if male, 0 if female
10	Marital status: 1 if single, 0 if married
11	Number of children (aged 16 or less)
12	Annual household income (U.S. dollars per year): 1 if less than 20,000, 2 if 20,000–29,999, 3 if 30,000–39,999, 4 if 40,000–49,999, 5 if more than 50,000
13	Age of vehicle driven in years
14	Manufacturer origin of vehicle driven: 1 if domestic, 0 if foreign
15	Operating cost of trip based on vehicle fuel efficiency (in U.S. dollars)

Example 13.1. Recall those logit model parameter estimates gave the following utility functions:

$$\begin{aligned} V_a &= -0.942(DISTA) \\ V_t &= 1.65 - 1.135(DISTT) + 0.128(VEHAGE) \\ V_f &= -3.20 - 0.694(DISTF) + 0.233(VEHAGE) + 0.764(MALE) \end{aligned} \quad (15.6)$$

Where *DISTA*, *DISTT*, and *DISTF* are the distances of individual travelers on the arterial, two-lane highway, and freeway route, respectively; *VEHAGE* is the age of the commuter's vehicle in years and *MALE* is an indicator variable that is equal to one if the commuter is a male and zero if female.

These give estimated route probabilities:

$$P(a) = \frac{e^{V_a}}{e^{V_a} + e^{V_t} + e^{V_f}}, \quad P(t) = \frac{e^{V_t}}{e^{V_a} + e^{V_t} + e^{V_f}}, \quad P(f) = \frac{e^{V_f}}{e^{V_a} + e^{V_t} + e^{V_f}} \quad (15.7)$$

TABLE 15.2

Corrected and Uncorrected Regression Models of Average Commute Speed
(*t*-Statistics in Parentheses)

Variable Description	Uncorrected Parameter Estimates	Corrected Parameter Estimates
Constant	14.54 (2.97)	15.37 (2.14)
Distance on chosen route in kilometers	4.08 (7.71)	4.09 (2.48)
Number of signals on chosen route	-0.48 (-1.78)	-1.30 (-0.81)
Number of children	3.10 (2.16)	2.85 (1.75)
Single indicator (1 if commuter is not married, 0 if married)	3.15 (1.52)	3.98 (1.70)
Vehicle age in years	-0.51 (-2.13)	-0.16 (-0.57)
Number of observations	151	151
R-squared	0.32	0.13
Corrected R-squared	0.30	0.10

Table 15.2 gives estimation results for two models. The “uncorrected” model is estimated ignoring discrete/continuous interactions. The “corrected” model uses Equations 15.6 and 15.7 as input to estimate Equation 15.5.

Both models show that commuters tend to have higher average commute speeds if they have longer trips, more children, and are males, and lower average commute speeds as the number of traffic signals they encounter increases and the older their commute vehicle. While the two models show consistency in the direction that the included variables have on average commute speeds, the magnitude of these variables and their statistical significance vary widely in some cases. Specifically, in the corrected model, the number of traffic signals and vehicle age become statistically insignificant. These shifting results underscore the need to consider the discrete/continuous nature of the problem and the role that commuter selectivity is playing in the model. For example, if the uncorrected model was used, erroneous inferences could have been drawn concerning the number of traffic signals and vehicle age.

15.3 Econometric Corrections: Selectivity-Bias Correction Term

Another popular approach to resolve selectivity-bias problems and arrive at unbiased parameter estimates is to develop an expression for a selectivity-bias correction term (see Heckman 1976, 1978, 1979). In the context of the example problem, this approach is implemented by noting that average travel speed s for commuter n from home to work is written as (refer to Equation 15.1)

$$E(s_n | i) = \beta_i \mathbf{X}_n + E(\xi_n | i) \quad (15.8)$$

where $E(s_n|i)$ is the average commute speed of commuter n conditional on the chosen route i , \mathbf{X}_n is the vector of commuter n characteristics influencing average travel speed, $\boldsymbol{\beta}_i$ is the vector of estimable parameters, and $E(\xi_n|i)$ is the conditional unobserved characteristics. Note that variables specific to i (\mathbf{Z}_n in Equation 15.2) are omitted from the right-hand side of this equation. Application of this equation provides bias corrected and consistent estimates of $\boldsymbol{\beta}_i$ because the conditional expectation of ξ_n , $E(\xi_n|i)$ accounts for the non-random observed commute speeds that are selectively biased by commuters' self-selected choice of route.

Equation 15.2 provides bias-corrected estimates of $\boldsymbol{\beta}_i$ because the selectivity bias caused by the nonrandom nature of the observed sample is taken into account in the conditional expectation of ξ_n , $E(\xi_n|i)$. The problem then becomes one of deriving a closed-form representation of $E(\xi_n|i)$ that is used for equation estimation. Such a derivation was developed by Hay (1980) and Durbin and McFadden (1984) on the assumption that the discrete-outcome portion of the problem is represented by a multinomial logit model.

To illustrate this approach (suppressing subscripting n), let γ denote a vector of discrete-outcome disturbance terms ($\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_J$), where J is the total number of discrete outcomes. The conditional expectation ξ (conditioned on discrete outcome k) is written as

$$E(\xi|i) = \left(\frac{1}{P_i} \right) \int_{\gamma \neq i} E(\xi|\gamma) \prod_{j=1}^J f(\varepsilon_j) d\gamma \quad (15.9)$$

where P_i is the probability of discrete outcome i . If assumed that γ is generalized extreme value distributed with σ^2 being the unconditional variance of ξ and ρ_i being the correlation of ξ and the resulting discrete-outcome logistic error terms (resulting from the differencing of $\varepsilon_i - \varepsilon_j$), Hay (1980) and Durbin and McFadden (1984) have shown that Equation 15.9 is

$$E(\xi_n|i) = (-1)^{J+1} (\sigma 6\rho_i / \pi^2) \left[(1/J) \sum_{j \neq i}^J \left[(P_j \ln(P_j)) / (1 - P_j) \right] + \ln(P_i) \right] \quad (15.10)$$

Using Equation 15.10, selectivity bias in discrete/continuous models is corrected by undertaking the following three steps:

1. Estimate a multinomial logit model to predict the probabilities of discrete outcomes i for each observation.
2. Use the logit-predicted outcome probabilities to compute the portion of Equation 15.10 in large brackets ([]) for each observation.

3. Use the values computed in step 2 to estimate Equation 15.8 using standard least-squares regression methods. The term $(-1)^{J+1} \sigma 6\rho_i / \pi^2$ becomes a single estimable parameter.

Thus Equation 15.8 is estimated for each route k as

$$s_{in} = \beta_i X_n + \alpha_i \lambda_n + \eta_n \quad (15.11)$$

where $\alpha_i = (-1)^{J+1} \sigma 6\rho_i / \pi^2$, $\lambda_n = [(1/J) \sum_{j=1}^J [(P_j \ln(P_j)) / (1 - P_j) + \ln(P_j)]]$, and η_n is the disturbance term. It is common practice to not include ρ_i in the summation over discrete outcomes J , thus imposing an equality restriction of the correlations of across ξ_n and $\varepsilon_l - \varepsilon_j$. This restriction is relaxed by moving ρ_i within the summation making it necessary to estimate a total of $J - 1$ selectivity-bias parameters (α) for each continuous equation corresponding to discrete outcome i . However, it has been shown empirically by Hay (1980) and Manner (1986a) that this restriction on ρ_i is reasonable.

Example 15.2

The selectivity-bias correction term approach is demonstrated by estimating an average commute speed model for the two-lane highway users using the data used previously in Example 15.1 and the discrete model estimated in Example 11.1. Also, for comparison purposes, it is interesting to estimate a model that excludes the selectivity-bias correction term.

Model estimation results are presented in Table 15.3. The table shows that 103 of the 151 commuters chose the two-lane highway as their commute route. For both selectivity corrected and uncorrected models, commuters that were men, had more passengers in their vehicle, wore seat belts, were older, and were not married tended to have higher commute speeds. While those commuters with large annual incomes tended to drive slower.

For the selectivity-corrected model, the selectivity-bias parameter is significantly different from zero. This finding provides a strong indication that sample selectivity is present. The null hypothesis of no selectivity bias is rejected with over 99% confidence as indicated by the t statistic. The implication is that the model is seriously misspecified when the selectivity bias term is omitted. This bias is also evident upon inspection of the uncorrected parameter estimates, which show noticeable differences when compared to the parameters in the selectivity-corrected model—particularly the constant parameter. These findings show that if selectivity bias is ignored, erroneous inferences and conclusions are drawn from the estimation results.

15.4 Discrete/Continuous Model Structures

In developing discrete/continuous equation systems, one can provide a construct to link the discrete and continuous components. The most obvious of

TABLE 15.3

Corrected and Uncorrected Regression Models of Average Commute Speed
(*t*-Statistics in Parentheses)

Variable Description	Parameter Estimates with Selectivity Correction	Uncorrected Parameter Estimates
Constant	77.04 (5.25)	26.12 (5.03)
Number of passenger in vehicle	4.31 (1.92)	5.07 (2.14)
Seat belt indicator (1 if commuter wearing seatbelts, 0 if not)	3.72 (1.60)	4.29 (1.74)
Gender indicator (1 if male, 0 if female)	3.52 (1.53)	3.19 (1.30)
Driver age in years	0.23 (1.96)	0.12 (1.57)
Annual household income in U.S. dollars	-0.000175 (-2.01)	-0.00009 (-1.12)
Single indicator (1 if commuter is not married, 0 if married)	4.18 (1.72)	6.03 (2.40)
Selectivity-bias correction α_k	12.55 (3.60)	—
Number of observations	103	103
R-squared	0.22	0.11
Corrected R-squared	0.16	0.05

these constructs is a reduced form approach. A common way to implement a reduced form is to start with the discrete model. Let T_{in} be a linear function that determines discrete outcome i for observation n and let y_{in} be the corresponding continuous variable in the discrete/continuous modeling system. Then write,

$$T_{in} = \boldsymbol{\beta}_i \mathbf{X}_{in} + \boldsymbol{\theta}_i y_{in} + \varepsilon_{in} \quad (15.12)$$

where $\boldsymbol{\beta}_i$ is a vector of estimable parameters for discrete outcome i , \mathbf{X}_{in} is a vector of the observable characteristics (covariates) that determines discrete outcomes for observation n , and ε_{in} is a disturbance term. Let the corresponding continuous equation be the linear function,

$$y_{in} = \boldsymbol{\theta}_i \mathbf{W}_{in} + v_{in} \quad (15.13)$$

where $\boldsymbol{\theta}_i$ is a vector of estimable parameters for the continuous variable observed for discrete outcome i , \mathbf{W}_{in} is a vector of the observable characteristics (covariates) that determine y_{in} , and v_{in} is a disturbance term. Equation 15.13 is estimated using ordinary least squares with appropriate selectivity-bias correction (such as adding a selectivity-bias correction term). For estimation of the discrete-outcome portion of the discrete/continuous process, note that y_{in} is endogenous in Equation 15.12 because y_{in} changes with changing

T_{in} due to the interrelated discrete/continuous structure. To avoid estimation problems, a reduced form model is obtained by substituting Equation 15.13 into Equation 15.12 giving,

$$T_{in} = \beta_i X_{in} + \phi_i \theta_i W_{in} + \phi_i v_{in} + \varepsilon_{in} \quad (15.14)$$

With Equation 15.14, a discrete-outcome model is derived readily. For example, if ε_{in} s are assumed to be generalized extreme value distributed, a multinomial logit model results with the probability of observation n having outcome i as (see Chapter 13)

$$P_n(i) = \frac{EXP(\beta_i X_{in} + \phi_i \theta_i W_{in})}{\sum_{\forall l} EXP(\beta_l X_{ln} + \phi_l \theta_l W_{ln})} \quad (15.15)$$

Note that because the term $\phi_i v_{in}$ does not vary across outcomes i , it cancels out and does not enter the logit model structure. Examples of applications of reduced form models are given in Hay (1980) and Mannering (1986a).

Another common approach to link discrete/continuous model structures is one that is based on utility theory and ensures economic consistency. If the discrete-outcome model is based on utility maximization (see Chapter 13), one can derive a model structure by noting that indirect utility (used as the basis to arrive at outcome probabilities) is related to commodity demand by Roy's identity (as discussed in Chapter 13)

$$y_{in}^0 = -\frac{\partial V_{in}/\partial p_{in}}{\partial V_{in}/\partial Inc_n} \quad (15.16)$$

where V_{in} is the indirect utility of discrete alternative i to consumer n , p_{in} is the consumer's unit price of consuming i , Inc_n is the decision maker's income, and y_{in}^0 is the utility maximizing demand for i . The discrete/continuous link is made by either specifying the indirect utility function V_{in} or the commodity demand y .

As an example, consider consumers' choice of vehicle make, model and vintage, and the annual usage of the vehicle (kilometers driven per year). This problem reflects a classic discrete/continuous application because consumers select a vehicle type, defined by make (Ford), model (Mustang) and vintage (2008) with some expectation of how much it is driven. Aggregate data support the relationship between vehicle type and usage. For example, it is well known that newer model vehicles are driven more, on average, than older vehicles. From a modeling perspective, consumers that own older vehicles are likely to be a self-selected group of people with lower levels of usage while those consumers owning newer vehicles may be people with

higher levels of usage. Because older-vehicle owners are not observed driving newer vehicles and newer-vehicle owners are not observed driving older vehicles, a selectivity-bias problem arises.

To develop an economically consistent model structure, a vehicle utilization equation is specified as (alternatively an indirect utility function could be specified first),

$$y_{in} = \beta_i X_{in} + \alpha_i Z_{in} + \kappa(Inc_n - \pi r_n p_{in}) + v_{in} \quad (15.17)$$

where y_{in} is the annual utilization (e.g., kilometers per year) of vehicle i , X_{in} is the vector of consumer characteristics that determine vehicle utilization, Z_{in} is the vector of vehicle characteristics that determine vehicle utilization, Inc_n is the consumer n 's annual income, r_n is the expected annual vehicle utilization for consumer n , p_{in} is the consumer's unit price of utilization (e.g., dollars per kilometer driven), β_i and α_i are the estimable vectors, κ and π are the estimable parameters, and v_i is the disturbance term. The expected utilization, r_{in} , is needed to capture the income effect of consumption ($Inc_n - \pi r_n p_{in}$) but is determined exogenously by regressing it against exogenous variables and using predicted values in Equation 15.17 (see Mannerling 1986b for a complete discussion).

With Equation 15.17, we can apply Roy's identity as a partial differential equation (see Equation 15.16)

$$\frac{\partial V_{in}}{\partial Inc_n} y_{in}^0 + \frac{\partial V_{in}}{\partial p_{in}} = 0 \quad (15.18)$$

and solve for V_{in} giving

$$V_{in} = \left[\beta_i X_{in} + \alpha_i Z_{in} + \kappa(Inc_n - \pi r_n p_{in}) + v_{in} \right] e^{-\kappa p_{in}} + \varepsilon_{in} \quad (15.19)$$

where ε_{in} is a disturbance term added for discrete-outcome model estimation. If ε_{in} are assumed to be generalized extreme value distributed, a logit model for discrete outcomes results. In estimating Equation 15.17, corrections need to be applied to the vector of vehicle characteristics Z_{in} such as instrumental variables or expected value methods to account for the endogeneity of vehicle characteristics. The major drawback with the economically consistent approach is that a nonlinear form of either the indirect utility function or continuous equation results. For example, from the previous model, the linear utilization equation produced a nonlinear indirect utility function. This nonlinearity complicates estimation of the discrete outcome model. In choosing between a reduced form and the economically consistent approaches, many applications require a trade off between ease of estimation and compromising theory.

15.5 Transportation Application of Discrete/ Continuous Model Structures

For applications of discrete/continuous modeling to a variety of transportation problems the reader is referred to Damm and Lerman (1981); Mannering and Winston (1985); Train (1986); Mannering (1986a, 1986b); Hensher and Milthorpe (1987); Mannering and Hensher (1987); Mannering et al. (1990); Hamed and Mannering (1993); Bhat (2005); Bhat (2008); and Pinjari and Bhat (2010).

Part IV

Other Statistical Methods

16

Random-Parameter Models

The modeling approaches presented in previous chapters treat parameters as constant across observations. That is, the effect of any individual explanatory variable (in the X vector) is the same for each observation or individual. However, this fixed-parameter assumption may be incorrect. For example, consider the effect of fuel price on the number of miles an individual drives. One might expect that individuals with high incomes are less sensitive to changes in fuel prices, an effect that is accommodated in a fixed-parameter model by dividing fuel price by the individual's income as the independent variable (the division by income reflects the less responsive changes in fuel price of higher income people). However, there may be other unobserved factors that affect an individual's sensitivity to fuel price such as an inherent love of driving or unobserved travel related factors such as child day-care responsibilities, poor fuel economy of vehicle, and so on. To account for the influences of unobserved heterogeneity, random-parameter models are derived by assuming that the estimated parameters vary across observations, usually according to some prespecified distribution. When a parameter is found to vary significantly across observations model estimation is considerably more complex because a unique parameter for each observation is estimated for the variable in question. The motivation for such models is to account for unobserved heterogeneity across observations or individuals and to facilitate important new insights regarding the underlying data generating process.

The following sections of this chapter present a number of random-parameter models that address different types of data, including discrete, count, and continuous data. Examples are provided along with interpretations of random-parameter models.

16.1 Random-Parameter Multinomial Logit Model (Mixed Logit Model)

While Chapter 13 reveals that the multinomial logit (MNL) model is suitable for many applications, it also describes a number of weaknesses that can result in erroneous parameters estimates if underlying assumptions are not satisfied. The random-parameters logit addresses several weaknesses of the

standard MNL model by allowing parameter values to vary across observations. The assumption made in the derivation and application of the standard MNL is that parameters are fixed across all observations. When this assumption is incorrect, inconsistent estimates of parameters and outcome probabilities result (see the discussion of this in Section 13.5.3).

To account for the possibility that the parameters may vary across observations, a random-parameters or mixed logit model is appropriate. Following the work presented in McFadden and Train (2000) and Train (2003) to develop the mixed logit modeling approach, consider a function determining discrete outcome probabilities as shown in Chapter 13

$$T_{in} = \boldsymbol{\beta}_i \mathbf{X}_{in} + \varepsilon_{in} \quad (16.1)$$

where $\boldsymbol{\beta}_i$ is a vector of estimable parameters for discrete outcome i and \mathbf{X}_{in} is a vector of the observable characteristics (covariates) that determine discrete outcomes for observation n , and ε_{in} is a disturbance term. As shown in Chapter 13, the assumption that the disturbances are extreme value Type I distributed gives the standard MNL form as

$$P_n(i) = \frac{\text{EXP}[\boldsymbol{\beta}_i \mathbf{X}_{in}]}{\sum_{\forall l} \text{EXP}(\boldsymbol{\beta}_l \mathbf{X}_{ln})} \quad (16.2)$$

where $P_n(i)$ is the probability of observation n having discrete outcome i ($i \in I$ with I denoting all possible outcomes for observation n). Now define a mixed model (a model with a mixing distribution) whose outcome probabilities are defined as $P_n^m(i)$ with

$$P_n^m(i) = \int_{\mathbf{X}} P_n(i) f(\boldsymbol{\beta} | \boldsymbol{\varphi}) d\boldsymbol{\beta} \quad (16.3)$$

where $f(\boldsymbol{\beta} | \boldsymbol{\varphi})$ is the density function of $\boldsymbol{\beta}$ with $\boldsymbol{\varphi}$ referring to a vector of parameters of that density function (mean and variance), and all other terms are as previously defined. Substituting Equation 16.2 into Equation 16.3 gives the mixed logit model,

$$P_n^m(i) = \int_{\mathbf{X}} \frac{\text{EXP}[\boldsymbol{\beta}_i \mathbf{X}_{in}]}{\sum_l \text{EXP}[\boldsymbol{\beta}_l \mathbf{X}_{ln}]} f(\boldsymbol{\beta} | \boldsymbol{\varphi}) d\boldsymbol{\beta} \quad (16.4)$$

In words, Equation 16.4 shows that the mixed logit probabilities $P_n^m(i)$ are the weighted average of the standard MNL probabilities $P_n(i)$ with the weights determined by the density function $f(\boldsymbol{\beta} | \boldsymbol{\varphi})$. Note that in the simplified case where $f(\boldsymbol{\beta} | \boldsymbol{\varphi}) = 1$, the model reduces to the standard MNL. With Equation 16.4, for model estimation, $\boldsymbol{\beta}$ can now account for observation-specific variations

of the effect of \mathbf{X} on outcome probabilities, with the density function $f(\boldsymbol{\beta}|\boldsymbol{\varphi})$ used to determine $\boldsymbol{\beta}$. Mixed logit probabilities are thus a weighted average for different values of $\boldsymbol{\beta}$ across observations where some elements of the parameter vector $\boldsymbol{\beta}$ are fixed parameters and some are random. Many studies have used a continuous form of the density function $f(\boldsymbol{\beta}|\boldsymbol{\varphi})$ in model estimation (such as a normal distribution), although a wide variety of density functions are feasible.

Mixed logit models do not suffer from the independence of irrelevant alternatives problem because the ratio of any two outcome probabilities (see Equation 13.38 in Chapter 13) is no longer independent of any other outcomes' probabilities (see Train 2003 for a complete discussion of this point).

Estimation of the mixed logit model shown in Equation 16.4 by maximum likelihood is undertaken using simulation approaches due to the difficulty in computing the probabilities. Recall from Chapter 13 that the log-likelihood function is written as (see Equation 13.22 for comparison),

$$LL = \sum_{n=1}^N \sum_{i=1}^I \delta_n \ln [P_n^m(i)] \quad (16.5)$$

where N is the total number of observations, I is the total number of outcomes, δ_n is defined as being equal to 1 if the observed discrete outcome for observation n is i and zero otherwise δ is as defined in Equation 13.9, and all other variables are as defined previously. The mixed logit probabilities $P_n^m(i)$ are approximated by drawing values of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\boldsymbol{\varphi})$ given values of $\boldsymbol{\varphi}$ and using these drawn values to estimate the simple logit probability $P_n(i) = \text{EXP}[\boldsymbol{\beta}_i \mathbf{X}_{in}] / \Sigma_i \text{EXP}[\boldsymbol{\beta}_i \mathbf{X}_{in}]$. This procedure is repeated across many samples and the computed logit probabilities are summed and averaged to obtain the simulated probability $\hat{P}_n^m(i)$ in Equation 16.5 used to compute the likelihood function. Because probabilities are simulated, the likelihood function in Equation 16.5 is referred to as the simulated likelihood function.

There has been considerable research on how best to draw values of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\boldsymbol{\varphi})$ so that accurate approximations of the probabilities are obtained with as few draws as possible. Sampling values randomly from the distribution is an obvious starting point, but this approach may not be as efficient as other sampling techniques. Train (2003) reviews a wide variety of sampling techniques and strategies that are often used in this regard. However, the most popular alternative to random draws are Halton sequences (or Halton draws), which are based on a technique developed by Halton (1960) to generate a systematic nonrandom sequence of numbers. Halton draws(samples) have been shown to be significantly more efficient than purely random draws, arriving at accurate probability approximations with far fewer draws (see Bhat 2003 and Train 1999). Details of the evolution of simulation-based maximum likelihood methods for estimating mixed logit models are provided in numerous references, including McFadden and Ruud (1994), Geweke et al. (1994), Boersch-Supan and Hajivassiliou (1993), Stern (1997), and Brownstone and Train (1999).

Example 16.1

A random-parameters logit model is estimated to determine the proportion of police-reported accidents classified as property-damage only, possible injury, and evident injury. Accident data from multilane-divided highways in Washington State are used to estimate this model. These accident data were gathered for specific roadway segments, the lengths of which are defined by the highway's median treatments (safety barriers, cables, or landform barriers). Thus, a roadway segment's beginning point was identified where a previous run of a barrier terminated (or began) and ended where the next run of a barrier was encountered (or the current run ended). The data consist of 275 roadway segments of varying lengths with a mean segment length of roughly 2.4 miles with a standard deviation of about 2.7 miles. Police-reported accidents that occurred on each roadway segment in the 1990 calendar year were used to determine the accident injury-severity proportions. The variables available for model estimation are shown in Table 16.1 (please see Milton et al. 2008 for a further description of these data).

A mixed logit model is estimated to determine the proportions of accidents that result in no injuries, possible injury (accidents where the most severely injured person is determined to be possibly injured by the officer at the accident scene) and evident injury (accidents where the most severely injured person has injuries that are evident to the officer at the accident scene, including nonincapacitating injuries, incapacitating injuries, and fatalities). To estimate possible mixing distributions (see Equation 16.4), 200 Halton draws were used as recommended by Bhat (2003) and others. Model-estimation results are presented in Table 16.2 (due to missing data, only 258 of the original 275 roadway segments are used for model estimation).

Table 16.2 shows eight statistically significant model parameters, including constants. Four of these parameters vary across the roadway-segment population according to a normal distribution (less well-fitting distributions considered were the log-normal, uniform, and triangular). The other four parameters are fixed across the roadway-segment population because the estimated standard deviations of their parameter distributions were statistically insignificant from zero (these fixed-parameter estimates included the constants for the no-injury and possible-injury functions).

Turning to specific estimation results, annual daily travel per lane yields a positive fixed parameter in the no-injury function, indicating that higher vehicle volumes per lane increase the probability that accidents will involve no injury. This outcome is likely to be the result of congestion, with higher vehicle volumes resulting in lower speeds and less severe accidents.

The minimum radius on the roadway segment, defined for the no-injury function, produced a normally distributed parameter with a mean of -0.918E-04 and a standard deviation of 1.11E-04. This result suggests that for 79.1% of roadway segments, a higher minimum radius (segments with their tightest curves being less sharp) results in a lower probability of a no-injury accident and a higher probability of a more severe accident. For some roadway segments, drivers may be compensating for sharper curves by driving more cautiously, thus reducing their injury probabilities. For the remaining 20.9% of the roadway segments, the parameter is positive, indicating that a higher minimum radius results in an increased likelihood of noninjury accidents (and a reduced likelihood of more severe accidents).

The percentage of trucks in the traffic, defined for the possible-injury function, produced a normally distributed negative parameter with a mean of -0.164 and

TABLE 16.1

Variables Available to Model Accident Severity Proportions

Variable No.	Variable Description
1	Segment identification number
2	Injury frequency (dependent variable—property damage only, possible injury, injury)
3	Route number
4	Segment length in miles
5	Number of lanes in increasing milepost direction
6	Number of lanes in decreasing milepost direction
7	Total combined width of all lanes in feet
8	Minimum median shoulder in feet
9	Maximum median shoulder in feet
10	Speed limit in mi/h
11	Indicates urban area (1 if yes, 0 if no)
12	Functional class (1 if local, 2 if collector, 3 if arterial, 4 if principal arterial, 5 if interstate)
13	Average Annual Daily Traffic
14	Daily percentage of single-unit trucks
15	Daily percentage of tractor and trailer trucks
16	Daily percentage of tractor and two-trailer trucks
17	Percent of daily traffic in the peak hour
18	Number of grade breaks in the segment (defined by the presence of a vertical curve or a vertical point of inflection for grade changes that would result in vertical curves below minimum curve lengths)
19	Minimum grade in the segment in percent
20	Maximum grade in the segment in percent
21	Maximum grade difference in the segment in percent
22	Tangent length in the segment in miles
23	Number of horizontal curves in the segment
24	Minimum radius curve on roadway segment (in feet)
25	Segment access control (0 if none, 1 if partial, 2 if full)
26	Median width (1 if less than 30 ft; 2 if 30–40 ft; 3 if 40–50 ft; 4 if 50–60 ft; 5 if higher than 60 ft)
27	Friction value (0–100 with 100 being high)
28	Average daily travel per lane (in vehicles)
29	Segment slope (0 if flat, 1 if slight, 2 if medium, 3 if high)
30	Indicates number of interchanges in the segment
31	Average precipitation per month in inches
32	Average snowfall per month in inches

a standard deviation of 0.132, suggesting that for 89.3% of roadway segments higher truck percentages result in decreased likelihood of a possible-injury accidents and increased likelihood of other injury types. However, for 10.7% of the roadway segments, increasing truck percentages increase the probability of

TABLE 16.2

Mixed Logit Estimation Results for Accident Severity Proportions (All Random Parameters are Normally Distributed)

Variable Description	Estimated Parameter	t-Statistic
<i>No injury</i>		
Constant	3.30	3.54
Average annual daily travel per lane (vehicles per lane)	0.150E-04	1.37
Minimum radius curve on the roadway segment (in feet) (Standard deviation of parameter distribution)	-0. 918E-04 (1.11E-04)	-1.57 (3.44)
<i>Possible injury</i>		
Constant	3.46	3.24
Percentage of trucks (all truck types) in the traffic (Standard deviation of parameter distribution)	-0.164 (0.132)	-3.01 (2.65)
<i>Evident Injury</i>		
Friction value (0–100 with 100 being high)	0.0591	3.01
Number of interchanges per mile of segment length (Standard deviation of parameter distribution)	-2.22 (4.07)	-2.75 (3.18)
Number of grade breaks per mile of segment length—defined by the presence of a vertical curve or a vertical point of inflection for grade changes that would result in vertical curves below minimum curve lengths divided by segment length (Standard deviation of parameter distribution)	-0.297 (0.458)	-2.25 (2.45)
Number of observations	258	
Log-likelihood at zero	-5,116.24	
Log-likelihood at convergence	-4,457.39	

possible-injury accidents. Clearly, the effect of truck percentages can vary significantly depending on the specific roadway segment.

For the evident-injury function, the roadway friction parameter did not vary across roadway segments. The parameter is positive, indicating that higher friction values increase the likelihood of severe accidents. This result is intuitive because higher friction can result in a greater force transfer among vehicles involved in crashes. Thus, one might expect higher friction to decrease the likelihood of an accident (because stopping distances would be shorter, etc.) but to possibly increase the severity due to the higher transfer of forces between vehicles and objects hit.

The interchange density (number of interchanges per mile) on the roadway segment, defined for the evident-injury function, produced a normally distributed parameter with a mean of -2.22 and a standard deviation of 4.07. This finding suggests that for 70.7% of roadway segments, higher interchange density decreases the probability of an evident-injury accident (thus increasing the likelihood of lower injury-severity accidents). This outcome might reflect the effect that high

interchange density has on lowering vehicle speeds. However, for 29.3% of the roadway segments, increasing interchange density increases the likelihood of evident injury. For some roadway segments, possible speed reductions associated with high interchange densities are not sufficient to overcome the increase crash risk introduced by increased vehicle conflict paths associated with interchanges.

Finally, the density of grade breaks on the roadway segment (see Table 16.1 for a detailed description of this variable), defined for the evident-injury function, produced a normally distributed random parameter with a mean of -0.297 and a standard deviation of 0.458. This result indicates that for 74.2% of roadway segments, higher grade-break density decreases the probability of an evident-injury accident (thus increasing the likelihood of lower injury-severity accidents) and for 25.8% of the roadway segments higher densities increase the probability of an evident-injury accident. High grade-break density lowers speeds and thus accident severities on some segments, while on others this lowering of speed is not sufficient to overcome the added danger that such roadway elevation changes may pose.

16.2 Random-Parameter Count Models

As is the case with the logit model, random parameters are introduced into count-data models in a similar fashion. Consider the basic Poisson model presented previously in Equation 11.1

$$P(y_n) = \frac{\text{EXP}(-\lambda_n)\lambda_n^{y_n}}{y_n!} \quad (16.6)$$

where y_n is nonnegative integer count, $P(y_n)$ is the probability of observation n having y_n counts per some time period (e.g., 1 year) and λ_n is the Poisson parameter for observation n and is equal to observation n 's expected number of counts per year, $E[y_n]$. Recall from Chapter 11 that a Poisson regression is estimated by setting

$$\lambda_n = \text{EXP}(\boldsymbol{\beta}\mathbf{X}_n) \quad (16.7)$$

where \mathbf{X}_n is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters. Also recall that negative binomial model is derived by assuming

$$\lambda_n = \text{EXP}(\boldsymbol{\beta}\mathbf{X}_n + \varepsilon_n) \quad (16.8)$$

where $\text{EXP}(\varepsilon_n)$ is a Gamma-distributed error term with mean 1 and variance α .

To allow the parameters to vary across observations to account for possible unobserved heterogeneity (unobserved factors that may vary across observations) with random parameters, Greene (2007) developed estimation

procedures for incorporating random parameters in count-data models. To allow for such random parameters in count-data models, estimable parameters are written as

$$\boldsymbol{\beta}_n = \boldsymbol{\beta} + \omega_n \quad (16.9)$$

where ω_n is a randomly distributed term (e.g., a normally distributed term with mean zero and variance σ^2). With this equation, the Poisson parameter becomes $\lambda_n | \omega_n = \text{EXP}(\boldsymbol{\beta}_n \mathbf{X}_n)$ in the Poisson model and $\lambda_n | \omega_n = \text{EXP}(\boldsymbol{\beta}_n \mathbf{X}_n + \varepsilon_n)$ in the negative binomial with the corresponding probabilities for Poisson or negative binomial now $P(y_i | \omega_i)$. With this random-parameters version of the model, the log likelihood is written as,

$$LL = \sum_{\forall n} \ln \int_{\omega_n} g(\omega_n) P(y_n | \omega_n) d\omega_n \quad (16.10)$$

where $g()$ is the probability density function of the ω_i . Because probability estimations are computationally cumbersome much like the case for the mixed logit, a simulation-based maximum likelihood method is again used (with Halton draws again being an efficient alternative to random draws).

Interestingly, this random-parameters formulation is equivalent to the random effects model presented previously in Section 11.6 of Chapter 11 with the constant term is the only random parameter. The interested reader should consult Greene (2007) for further details on random-parameters count models and Anastasopoulos and Mannerling (2009) for an application with transportation data.

Example 16.2

A random-parameters count model is estimated to determine the total number of accidents that occurred on 275 roadway segments described previously in Example 16.1 (variables available for model estimation are shown in Table 16.1, except for variable 2 in Table 16.1, which is now the total number of accidents of all severity types). Seventeen of the 275 segments have zero accidents, and the mean number of annual accidents (all severity types) on the 275 roadway segments is 16.87 and the standard deviation is 21.39 (variance is 457.5). The difference between the mean and the variance suggests that the negative binomial will likely be needed to account for the overdispersion reflected in the data.

Negative binomial model-estimation results shown in Table 16.3 reveal that eight parameters are statistically significant (including constants) and that four of these parameters are fixed across the roadway segments (standard deviations of their parameter distributions were not statistically different from zero) and four parameters vary across the roadway-segment population according to a normal distribution (other distributions considered were the log-normal, uniform, and triangular, however, the normal distribution yielded superior statistical fit). The table also shows that the negative binomial dispersion parameter is significantly different

TABLE 16.3

Random-Parameters Negative Binomial Model for Annual Accident Frequencies
(All Random Parameters are Normally Distributed)

Variable Description	Estimated Parameter	t-Statistic	Marginal Effect
Constant	2.62	8.65	
100 million vehicle miles traveled per year (computed as annual average daily traffic multiplied by 365 and by segment length) (Standard deviation of parameter distribution)	2.35 (0.940)	22.98 (15.69)	25.18
Low precipitation indicator (1 if annual precipitation was less than 18 inches, 0 otherwise)	-0.451	-6.31	-4.84
Snowfall indicator (1 if annual snowfall was greater than 12 inches, 0 otherwise) (Standard deviation of parameter distribution)	-0.211 (0.607)	-3.62 (13.42)	-2.26
Friction value (0–100 with 100 being high)	-0.0143	-2.47	-0.153
Number of interchanges per mile of segment length (Standard deviation of parameter distribution)	0.214 (0.510)	4.25 (13.41)	2.29
Number of grade breaks in the roadway segment—defined by the presence of a vertical curve or a vertical point of inflection for grade changes that would result in vertical curves below minimum curve lengths	0.0545	7.81	0.584
Number of horizontal curves per mile of segment length (Standard deviation of parameter distribution)	-0.159 (0.129)	-4.94 (7.72)	-1.70
Negative binomial dispersion parameter	10.08	6.28	
Number of observations		275	
Log-likelihood at zero		-1,3541.12	
Log-likelihood at convergence		-951.30	

from zero, indicating that the application of the standard Poisson model is not appropriate for these data. Finally, the log-likelihood values shown in Table 16.3 show strong overall model fit with a large increase in the likelihood function at convergence versus at zero.

Examination of results reveals that as vehicle miles of travel increase on the roadway segment (100 million vehicle miles traveled per year computed as annual average daily traffic multiplied by 365 and by the segment length) the mean accident rate also increases. The estimated parameter is normally distributed across the roadway segments and positive for almost all segments (on only two segments was this parameter slightly negative). The average value for this variable across all segments is 0.258 million vehicle miles traveled per year. This relatively low average value explains why the marginal effect was so high, with a 1 unit increase in

the 100-million-vehicle-miles-traveled variable (which would be several times the mean value) increasing the mean number of accidents per year by an average of 25.18 across all segments.

Roadway segments receiving less than 18 inches of precipitation per year were found to have fewer accidents. This variable is fixed across the roadway-segment population with roadway segments with less than 18 inches of annual precipitation having an average of 4.84 fewer accidents per year (as indicated by the marginal effect). In contrast, the effect of annual snowfalls greater than 12 inches is normally distributed and negative (reducing accident frequency) for 63.6% of the roadway segments and positive (increasing accident frequency) for 37.4% of the roadway segments. It is likely that this variable is capturing variation across Washington State's multilane highways regarding snow removal practices, responses of local drivers to snow (slowing down, etc.), and the familiarity of drivers with snow events across the state (some regions of Washington State rarely get snow while many others regularly experience snow).

The friction variable produced a negative fixed parameter that reduced the frequency of accidents by an average of 0.153 accidents for each one point improvement in the friction value (as indicated by the marginal effect). Note that, when combined with the findings for severity shown in Table 16.2, increasing friction decreases the likelihood of an accident but increases the severity once an accident has occurred.

The number of grade breaks in the roadway segment (which averaged 3.87 per segment) has a positive fixed parameter, where each additional grade break resulted in a 0.584 increase in the annual number of accidents (on average). This finding likely reflects challenges motorists have with elevation differences over the segments such as restricted sight distances.

Finally, the number of horizontal curves per mile in the roadway segment has a normally distributed parameter that is negative for 89.11% of the roadway segments and positive for 10.89% of the roadway segments. For most roadway segments increasing horizontal curve density decreases accident frequency, perhaps because drivers compensate for the more complex roadway geometrics by slowing down and becoming increasingly alert. However, at some roadway segments this possible compensation effect is not sufficient to overcome the more challenging geometrics—and the number of accidents increases with increasing horizontal curve density.

16.3 Random-Parameter Duration Models

Following the same procedure used for count models, random parameters are introduced into duration models (see Chapter 10 for the details of these models). That is, instead of having the explanatory variables act as $\text{EXP}(\beta \mathbf{X}_n)$ as shown in Equations 10.6 and 10.7, a randomly distributed term (ω_n) is introduced as in Equation 16.9 and explanatory variables now act on the hazard as $\text{EXP}(\beta_n \mathbf{X}_n)$, where β now varies across n observations. As with the two random-parameter models presented previously, a simulation-based maximum

likelihood method is again used (with Halton draws again being an efficient alternative to random draws).

Example 16.3

A random-parameters duration model is estimated using the 96 Seattle-area commuters' work-to-home delay data, as shown in Table 10.1 and used in Examples 10.1, 10.2, and 10.3. The results of a random-parameters Weibull duration model are shown in Table 16.4. The table indicates that three of the five previously estimated fixed parameters are random (see Table 10.5 for previous estimates with fixed parameters). When comparing the Weibull fixed-parameters model in Table 10.5 to the random-parameters model in Table 16.4, the random-parameters model appears to be significantly improved, with a likelihood at convergence of -88.63 compared to -93.80 when the parameters are fixed. This finding is significant because a likelihood ratio test comparing these two model values produces an χ^2 statistic of 10.34 (with three degrees of freedom—based on the three estimated standard deviation parameters that distinguish the random-parameters case from the fixed-parameter case), suggesting that the null hypothesis that the fixed-parameters Weibull is correct is rejected at a confidence level of 98.4%.

Turning to specific estimation results shown in Table 16.4, the male-indicator variable is normally distributed across the population and found to decrease delay for 88.1% of the commuters and decrease it for 11.9%. The ratio of actual travel time to free-flow travel time is normally distributed and negative for all but two of the commuters. Finally, the resident population of the work is normally distributed across commuters and negative for all but one commuter.

TABLE 16.4

Random-Parameter Weibull Model-Estimation Results of the Duration of Commuter Work-to-Home Delay (in Minutes) to Avoid Congestion (All Random Parameters are Normally Distributed)

Variable Description	Estimated Parameter	t-Statistic
Constant	-1.95	-12.37
Male indicator (1 if male commuter, 0 if not) (Standard deviation of parameter distribution)	0.164 (0.139)	3.80 (3.22)
Ratio of actual travel time at time of expected departure to free-flow (noncongested) travel time (Standard deviation of parameter distribution)	-0.718 (0.342)	-9.75 (8.18)
Distance from home to work in kilometers	-0.035	-7.46
Resident population of work zone (Standard deviation of parameter distribution)	-1.191E-05 (0.526E-05)	-5.53 (5.56)
P (distribution parameter)	5.25	12.78
Number of observations	96	
Log-likelihood at convergence	-88.63	

17

Bayesian Models

Bayesian models are simply the integration of Bayes' theorem with classical statistical models. Any classical statistical model can be estimated using a Bayesian equivalent. Bayes' theorem, in words, says that subjective prior probabilities may play a role in the estimation of statistical models. The Bayesian statistician uses prior information on parameter values in addition to the current data (or likelihood) at hand to obtain a posterior estimate of parameter values.

Bayesian models are becoming popular within the transportation profession, not primarily due to the benefits of Bayes' theorem, but because of the accessibility of Bayesian models through the use of Markov Chain Monte Carlo (MCMC) methods. MCMC is a sampling-based approach to estimation that is well suited for Bayesian models. In addition, MCMC enables the estimation of complex functional forms—forms that are often difficult to estimate using maximum likelihood methods. As a result, many Bayes' model applications in transportation involve noninformative (also known as diffuse, vague, or ignorance) prior information and complex model functional forms. Because of the accessibility and flexibility offered by MCMC estimation methods for estimating Bayesian statistical models, the transportation profession is seeing an increasing number of applications. As a result, general guidance on these methods is needed.

This chapter cannot describe MCMC Bayesian statistical models in great detail due to their wide variety and complexity. Instead, the necessary basic building blocks and examples are provided, with the intent to develop a basic understanding of MCMC Bayesian models and their estimation. The interested reader can find detailed discussions of Bayesian methods and models in the considerable body of literature that exists on this subject (see, e.g., Congdon [2001]; and Congdon [2003]; Carlin and Louis [1996]; Gelman et al. [1995]; Robert and Casella [1999], among many other references).

17.1 Bayes' Theorem

Bayesian statistical methods originate from an alternative philosophical viewpoint, requiring the analyst to reframe problems in terms of Bayesian logic. This logic combines “subjective” or prior knowledge, typically in the

form of statistical distributions, with “objective” current information (data), to derive meaningful “posterior” distributions.

With parameter vector β , in Bayes’ theorem current information (or likelihood) provided by data Y , a realization of $Y \sim f(Y|\beta)$, is combined with prior information in the form of a prior distribution of unknown parameter values with density $P(\beta)$, which through Bayes’ theorem results in the posterior distribution $P(\beta|Y)$. Bayes’ theorem is given as

$$\pi(\beta|Y) = \frac{f(Y|\beta)P(\beta)}{\int f(Y|\beta)P(\beta)d\beta} \quad (17.1)$$

where $m(x) = \int f(Y|\beta)P(\beta)d\beta$ is the marginal density of Y , or the probability of the data for the model with parameters β . For interesting Bayesian models, the marginal density of Y presents a difficult integration problem, so a similar expression is often used instead, where

$$\pi(\beta|Y) \approx f(Y|\beta)P(\beta) \quad (17.2)$$

In this equation the posterior is proportional to the product of the prior and the likelihood. The relative weights of the likelihood and prior are determined by the variances of the distributions, with smaller variance resulting in greater weight in the determination of the posterior. For example, a normal prior of $P(\beta_1) \sim N(\mu = 10, \sigma = 10)$ would have less influence on the likelihood than would a normal prior of $P(\beta_1) \sim N(\mu = 10, \sigma = 2)$.

The Bayesian theorem asserts that useful information about probabilities is obtained about specific observable events through subjective, expert evaluation or insight. For example, in the simple linear regression model $\hat{y} = \beta_0 + \beta_1 x$, an analyst may know something about the expected value of the unknown model parameter β_1 from prior research, from fundamental knowledge, or from prior data. It could be that $\beta_1 \sim N(\mu, \sigma)$, where μ and σ are known parameters of a normal distribution. Large values of σ might indicate a high degree of uncertainty in the value of β_1 while smaller values may indicate greater confidence. Alternatively, the analyst might believe that $\beta_1 \sim U(l, u)$, where l and u are the lower and upper bounds of a continuous uniform distribution, respectively.

While the majority of this chapter focuses on the mechanics of Bayesian statistical models, it is important to know the arguments for and against Bayesian models. There is a long history of healthy debate between classical and Bayesian statisticians. A common criticism of Bayesian models is the selection of subjective priors to influence parameter values. Often the source of information that guides the selection of priors is the cumulative body of past research. As an example, suppose that ten studies on a phenomenon under investigation show the set of values of β_1 in a linear regression model to be {5.5, 8.1, 9.0, 6.3, 5.3, 6.5, 3.9}. How should this information be expressed in a prior? One could fit these data to a normal distribution and use the

parameters obtained as priors. If there was belief that these values should not dominate the model parameters, then the variance could be increased to reflect greater or less belief in these values. However, analysts must defend the choice of priors, and often it is a challenge to argue convincingly for one prior over another. As mentioned previously, a common approach is to assume that nothing informative is known about priors, allowing “ignorance,” “diffuse,” or “noninformative” priors to be used and thus removing the burden to defend subjective priors. As an example, one might assume that a parameter has a mean of zero and standard deviation 1,000—a relatively flat, noninformative prior. However, the assigning of ignorance priors, as in this example, often results in a model with parameters that could have been obtained via maximum likelihood (or other classical) methods. One would question the motive for using Bayes’ theorem under such circumstances.

The assigning of subjective priors is touted as an asset by Bayesian statisticians. The Bayesian philosophy asserts that almost always something is known or expected about model parameters before estimation, and that knowledge is accumulated in the natural course of discovery. This viewpoint is in contrast to starting with a review of prior research regarding plausible parameter values and effects, only to discard this prior information in favor of letting only the current data influence parameter values. Bayesians argue that knowledge is accumulated by incorporating knowledge from the past into statistical models using Bayes’ theorem, resulting in a continual updating of knowledge.

Bayesian inference also claims interpretive advantage over classical methods, because posterior credible intervals reflect the probability of the null hypothesis being true conditioned on the observed data. In contrast, classical confidence intervals on parameters provide the probability of observing data conditioned on specific parameter values. This data versus parameter conditioned inference space is an outcome of Bayes’ theorem.

Nonetheless, an analyst wishing to apply a Bayesian model is well advised to have a clear motive for so doing. A defense of priors is necessary, and that the use of noninformative priors may raise questions as to why maximum likelihood (or other classical parameter estimation) methods were not used instead. As is discussed in the next section, a practical motive for using Bayesian models might be to estimate complex models that are difficult using classical estimation approaches. As is shown, the estimation, interpretation, and assessment of MCMC Bayesian models is fundamentally different than that of classically estimated statistical models, requiring an enhanced set of statistical tools.

17.2 MCMC Sampling-Based Estimation

As stated previously, MCMC methods provide a convenient mechanism for the estimation of Bayesian statistical models. A major analytical drawback to

Bayesian methods before MCMC was the difficulty of solving complex integration problems and, specifically, the integration of Bayes' risk

$$\iint L(\boldsymbol{\delta}, \boldsymbol{\beta}) f(\mathbf{Y} | \boldsymbol{\beta}) P(\boldsymbol{\beta}) d\mathbf{Y} d\boldsymbol{\beta} \quad (17.3)$$

where $L(\boldsymbol{\delta}, \boldsymbol{\beta})$ is the loss function for parameter vector $\boldsymbol{\beta}$ for estimator $\boldsymbol{\delta}$ and represents the loss incurred from estimating $\boldsymbol{\beta}$ using $\boldsymbol{\delta}$ and also involves a double integration over x and $\boldsymbol{\beta}$. For example, a loss function might be minimum squared error loss (Robert and Casella 1999).

Until recently this and similar integration problems have limited prior distributions to be conjugate, or within the same "family" of distributions. This limitation caused models to be too simple or uninteresting to receive widespread use. Fortunately, MCMC methods and the Gibbs sampler (to be discussed shortly) have drastically improved the ability to estimate complex Bayesian models (Congdon 2001, 2003).

Markov Chain Monte Carlo methods have matured to become part of the standard set of techniques routinely used by statisticians. They are particularly useful for solving classical statistical problems of estimation, prediction, statistical tests, data imputation, and variable selection. The MCMC approach has overcome the difficult problems that arise with models of censored data, missing data, nonlinear explanatory variables, and latent variable models (Robert and Casella 1999), to name just a few applications.

Markov Chain Monte Carlo is a simulation process that enables repeated sampling from known distributions in finite state space that results in a Markov Chain, or series of numbers. An example of a Markov Chain is the random sample of normal variates that is obtained from a random number generator. A Markov Chain operates in discrete time intervals to produce a sequence of evolving random variables, with the probability of transition (evolution) dependent upon its current state. Chains are generated from a transition kernel (which is a conditional probability density function). The resulting chains have desirable properties. Specifically, a stationary probability distribution exists as a result of construction of the Markov chains, and convergence to the limiting or stationary distribution occurs with almost certainty (Robert and Casella 1999) under the right conditions.

The Gibbs sampler is an efficient MCMC method ideally suited for solving Bayesian modeling problems (Robert and Casella 1999). To illustrate, suppose that for $p > 1$ (where p is the number of conditional univariate densities) a random variable vector $\mathbf{Y} = (y_1, y_2, \dots, y_p)$ exists, where the y_i 's are either uni- or multidimensional. In addition, suppose the corresponding univariate conditional densities f_1, f_2, \dots, f_p are simulated (samples are drawn from densities, including the Normal, Lognormal, Binomial, Poisson, Gamma, Beta, etc.).

In other words, it is possible, given reasonable starting values, to simulate the following:

$$y_i | y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_p \sim f_i(y_i | y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_p) \quad (17.4)$$

Then the associated Gibbs sampling algorithm (or Gibbs Sampler) is used to facilitate the transition from $Y(t)$ to $Y(t+1)$, the first and second draws from the Gibbs sampler

1. $y_1(t+1) \sim f_1(y_1 | y_2(t), y_3(t), \dots, y_p(t))$
2. $y_2(t+1) \sim f_2(y_2 | y_1(t+1), y_3(t), \dots, y_p(t))$
- .
- p. $y_p(t+1) \sim f_p(y_p | y_1(t+1), y_2(t+1), \dots, y_{p-1}(t+1)).$ (17.5)

The densities f_1, f_2, \dots, f_p are full conditionals, and are the only densities used for simulation in the Gibbs sampler. For higher dimensional problems the simulations are reduced to univariate simulations. The Gibbs sampler is a special case of the Metropolis–Hastings algorithm and has desirable convergence properties, satisfying the minimum conditions for ergodicity, irreducibility, and Harris recurrence (for additional details regarding the Gibbs sampler and properties of Markov Chains, consult Meyn and Tweedie 1993; Norris 1997; Reznick 1994; Robert and Casella 1999).

Repeated sampling using the Gibbs sampler generates an autocorrelated sequence of numbers, which when subjected to regularity conditions (ergodicity, etc.) eventually “forgets” the starting values used to initialize the chain and converges to a stationary sampling distribution of the posterior density (Congdon 2001). Statistics on the posterior density calculated post-convergence reveal information about the parameters of interest, specifically posterior credible intervals of the model coefficients and disturbances.

Unlike classical statistical methods, Bayesian estimation via MCMC yields the entire density of the parameters of interest, which is obtained from a combination of the likelihood and prior. In contrast, maximum likelihood estimation (using, e.g., Newton–Raphson algorithms) is focused on finding an optimal estimate, resulting in an absolute maximum value (or something close to it) of the likelihood function, and relies on large sample asymptotic properties to make inferences about the distribution of the parameter.

An extensive treatment of MCMC methods is not the intent of this chapter. Interested readers may consult a variety of references for additional information, including Meyn and Tweedie (1993); Norris (1997); Reznick (1994); and Robert and Casella (1999).

Example 17.1

Recall the average annual daily traffic (AADT) model estimated in Chapter 3 (see Example 3.1), where the objective was to model AADT on a particular road section as a function of local population and road characteristics. Suppose an analyst would like to estimate an MCMC Bayesian version of the linear regression model, with knowledge of reliable previously-estimated values for the population effect.

The first task is to specify the Bayesian version of the linear regression model, and then compare and contrast the statistical output of the Bayesian versus the classically estimated models. To begin, suppose the analyst believes that a reasonable functional form for the model is:

$$\text{Std(AADT)} = \beta_0 + \beta_1 \text{CNTYPOP} + \beta_2 \text{FUNCLASS3}$$

where Std(AADT) is the standardized version of AADT ($\text{AADT} - \text{mean(AADT)}/\text{sd(AADT)}$), CNTYPOP (in 10,000's) is the population of the county in which the road segment resides, and FUNCLASS3 is an indicator variable for urban interstates (see Table 3.1). In this specification of the model, parameters indicate a change in the standardized units of the response. That is, a parameter of 1.0 for FUNCLASS3 would indicate that urban interstates are associated with a 1 standard deviation increase in AADT above the mean. The classical estimation results for this model are shown in Table 17.1. The table shows the parameter estimates, standard errors, t values, p values, and 95% confidence intervals. For every 10,000 in county population, there is a 0.008 increase in the number of standard deviations of AADT. Each lane is associated with a 0.31 standard deviation increase, while urban interstates are associated with an increase of 1.16 standard deviations in AADT. The 95% confidence intervals are estimated based on the assumed normality of the estimated parameters, which are necessary in classical estimation. The correct interpretation of this confidence interval is that given model parameter values of zero (the null hypothesis), the probability of observing parameter values within the corresponding confidence interval would occur in 95 out of 100 samples. The square root of the mean squared error ($\text{MSE}^{-1/2}$) is approximately 0.448.

The estimation results for the Bayesian model are shown in Table 17.2. The initial Bayesian model is specified with noninformative priors on all model parameters

TABLE 17.1
Classical Estimation Results for AADT Model

Variable	Parameter Estimate	Standard Error	t -Statistic	$P(> t)$	95% Confidence Interval
Intercept	-1.367159	0.1160665	-11.78	<0.001	-1.597023, 1.137296
CNTYPOP10K	0.0087893	0.001394	6.31	<0.001	0.0060286, 0.01155
NUMLANES	0.3106909	0.0398291	7.80	<0.001	0.2318113, 0.3895704
FUNCLASS3	1.161202	0.1431521	8.11	<0.001	0.8776968, 1.444707
$\text{MSE}^{-1/2}$	0.44776				

TABLE 17.2

Bayesian Estimation Results for AADT Model: Uninformative Priors*

(Model I) STDADT	Parameter Estimate	Standard Error	Markov Chain	2.5%	Median	97.5%
Intercept	-1.37	0.12	0.00	-1.60	-1.37	-1.13
CNTYPOP10K	0.01	0.00	1.424E-5	0.01	0.01	0.01
NUMLANES	0.31	0.04	3.993E-4	0.23	0.31	0.39
FUNCLASS3	1.16	0.14	0.00	0.88	1.16	1.45
MSE ^{-1/2}	0.45	0.03	2.64E-4	0.40	0.45	0.52

*Convergence sample 20,000, postconvergence sample 10,000.

(e.g., $\beta's \sim N(\mu = 0, \sigma^2 = 100,000)$). Thus the priors have virtually no influence on the model parameters, which are dominated by the data. Because the MCMC approach requires a sampling-based specification, the linear regression model is specified such that

$$\beta_0 + \beta_1 \text{CNTYPOP} + \beta_2 \text{FUNCLASS3} \sim N(\mu = \text{Std(AADT)}, \sigma^2 = \text{MSE})$$

where the linear function is approximately normally distributed with mean Std(AADT) and variance equal to the MSE. In this specification MSE is assigned an uninformative prior. The parameters in the Bayesian linear regression model (as previously described) are arbitrarily close to those obtained using classical techniques (i.e., least squares or maximum likelihood parameter estimation) when noninformative priors are assigned.

The Bayesian model output is different than those normally associated with classical estimation methods, however, and reflects the sampling-based approach to parameter estimation. First, the standard errors of estimates are a combination of the variance from the data and the variance across simulations (MC Error). On average, Bayesian standard errors are larger than the standard errors obtained from classical methods (they are less efficient when noninformative priors are used), although the differences are often quite small. The posterior distribution, or credible interval, is entirely captured, and thus it is easy to report any percentile one wishes to examine. Table 17.2 shows the 2.5, 50 (median), and 97.5 percentiles corresponding to the 95% posterior credible interval (in contrast to a confidence interval). The correct interpretation of the posterior credible interval is that the parameter of interest lies within the interval with 95% probability. Two things are noteworthy regarding sigma. First, it is a sampled quantity and thus is stochastic, unlike the calculated value obtained from classical estimation techniques. Second, it is slightly larger than the likelihood-based value shown in Table 17.1, due to the variation introduced from using a sampling-based approach. Finally, 20,000 Markov chains were constructed to ensure model convergence, while parameter estimates were obtained from an additional 10,000 Markov chains. The formation of Markov Chains is discussed later in this chapter.

Suppose that five research studies have revealed the effect of population (CNTYPOP10K) on roadway traffic to be {0.010, 0.015, 0.012, 0.008, 0.021}. A decision has to be made as to how to include this information in the form of a prior. The decision on how to integrate this information into the model influences the estimation results and, as such, should be carefully considered and defended. Suppose, for example, it is believed that past research reflects a plausible range of outcomes, and thus assigns $\beta_1 \sim U(l = 0.008, u = 0.021)$ —a continuous uniform prior. Alternatively, one might believe that the observed values are themselves drawn from a normal, calculating an average and standard deviation of past values such that $\beta_1 \sim N(\mu = 0.0132, \sigma = 0.00507)$. Finally, another option might be to compute a weighted average and standard deviation, convinced that the sample sizes of the corresponding studies (20, 20, 50, 15, 125) reflect the importance that should be placed on the estimates, such that $\beta_1 \sim N(\mu = 0.0167, \sigma = 0.00458)$.

The effects on estimation results of these three different priors are shown in Table 17.3. A number of interesting effects are evident in the table. First, the Bayesian priors affect the parameter estimates for population (CNTYPOP10K), with differing effects depending on the prior selected. In general, the greater the precision of a prior (relative to the precision in the data), the greater the influence on the parameter estimate as specified in Bayes' theorem (Equation 17.1). Also, the distribution of a prior will influence the posterior differently, as well.

TABLE 17.3

Bayesian Estimation Results for AADT Model: Various Informative Priors*

Variable	Parameter Estimate	Standard Error	Markov Chain	2.5%	Median	97.5%
<i>Model II: Uniform Prior: $\beta_1 \sim U(lower = 0.008, upper = 0.021)$</i>						
Intercept	-1.367	0.1176	0.00112	-1.595	-1.366	-1.137
CNTYPOP10K	0.009455	0.001005	1.544E-5	0.008076	0.009286	0.01175
NUMLANES	0.3056	0.03984	4.007E-4	0.2268	0.306	0.383
FUNCLASS3	1.148	0.1452	0.001423	0.8583	1.147	1.43
MSE ^{-1/2}	0.4516	0.02961	3.21E-4	0.3978	0.4502	0.5148
<i>Model III: Normal Prior: $\beta_1 \sim N(\mu = 0.0132, \sigma = 0.00507)$</i>						
Intercept	-1.368	0.118	0.00114	-1.598	-1.368	-1.134
CNTYPOP10K	0.009091	0.001346	1.374E-5	0.006489	0.009082	0.01174
NUMLANES	0.3086	0.04013	3.999E-4	0.229	0.3091	0.3857
FUNCLASS3	1.155	0.144	0.001368	0.8743	1.156	1.441
MSE ^{-1/2}	0.4526	0.02993	2.627E-4	0.3979	0.451	0.5165
<i>Model IV: Normal Prior: $\beta_1 \sim N(\mu = 0.0167, \sigma = 0.00458)$</i>						
Intercept	-1.368	0.1178	0.001137	-1.598	-1.368	-1.134
CNTYPOP10K	0.009457	0.001335	1.36E-5	0.006877	0.009445	0.01209
NUMLANES	0.3059	0.04003	3.992E-4	0.2263	0.3065	0.3828
FUNCLASS3	1.148	0.1436	0.001365	0.8673	1.148	1.433
MSE ^{-1/2}	0.4515	0.02992	2.62E-4	0.3969	0.45	0.5152

*Convergence sample 20,000, postconvergence sample 10,000.

Parameters of other variables are also affected. This finding is an artifact of the data and not of Bayes's theorem. Because variables are correlated to varying degrees, their parameters are not independent. As a result, a prior assigned to a parameter through Bayes' theorem will also influence the parameter estimates of correlated variables. In all cases, the parameter estimate for the population has increased compared to the value obtained using classical methods and Bayes' with a noninformative prior. This observed parameter increase reflects the subjective belief that the estimate should account for the effect of the current information (data) as well as information known from the past (prior studies). Also, the goodness-of-fit as measured by sigma is reasonably close in all models, suggesting that the priors assigned in this example have not significantly skewed goodness-of-fit. The similarity in sigma values is obtained in part because the priors are not extreme (compared to the values obtained from the data), and because parameter estimates are not independent.

17.3 Flexibility of Bayesian Statistical Models via MCMC Sampling-Based Estimation

Example 17.1 showed that Bayes' theorem is used to assign priors to parameters in a linear regression model. There are a great number of potential applications of Bayes' theorem and the assignment of priors in statistical models (see Congdon 2001, 2003). Every statistical model developed and discussed in this text, for example, is estimable using a Bayesian equivalent, where a wide range of priors on parameters are possible. Bayesian methods are quite flexible and are used, for example, to assign priors to ratios of parameters, sums of parameters, order of parameters, hyperparameters (such as the gamma distribution in a negative binomial model), and missing data values. Priors can also include lower and/or upper truncated distributions. It is this ability to influence parameters using prior information that distinguishes Bayesian from classical statistical models.

As stated previously, there may be other motivations for estimating a Bayesian statistical model using MCMC methods, other than the explicit use of Bayes' theorem and the assignment of informative priors. A common motivation is the Bayesian model's ability to estimate parameters associated with complex model forms and functions that are difficult to estimate using classical approaches. The accessibility of Bayesian models via MCMC sampling-based estimation for handling complex statistical models does not mean Bayesian models are superior to classical estimation techniques. Rather, they may simply be more convenient for a particular model specification.

Example 17.2

Recall the Bayesian models estimated in Example 17.1. Suppose, instead of applying informative priors to the regression parameters, the following model is considered

TABLE 17.4

Bayesian Estimation Results for AADT Model: Nonlinear in Parameters Regression Function*

(Model V) STDADT	Parameter Estimate	Standard Error	Markov Chain Error	2.5%	Median	97.5%
Intercept	-2.231	0.1321	0.004302	-2.486	-2.233	-1.964
CNTYPOP10K	0.01235	0.001818	8.056E-5	0.008492	0.01245	0.01571
CNTYPOP10K ^β	-0.1721	0.0933	0.005944	-0.406	-0.1612	-0.01829
NUMLANES	0.3304	0.04024	7.458E-4	0.251	0.3306	0.4093
FUNCLASS3	1.17	0.1415	0.001354	0.8933	1.168	1.452
MSE ^{-1/2}	0.4443	0.02979	3.675E-4	0.3905	0.4428	0.507

*Convergence sample 20,000, postconvergence sample 10,000.

$$\text{LOG(AADT)} = \beta_0 + \beta_1 \text{CNTYPOP} + \text{CNTYPOP}\beta^2 + \beta_3 \text{NUMLANES} + \beta_4 \text{FUNCLASS3}$$

This nonlinear equation is quite easily handled within the MCMC framework because the parameter β_2 is not being calculated directly but instead is sampled conditionally using Equation 17.5. The results of this MCMC Bayesian linear regression model uninformative priors (β 's $\sim N(\mu = 0, \sigma^2 = 100,000)$) are shown in Table 17.4.

The linear regression model with the additional nonlinear parameter for population shows a slight improvement in fit with a reduction in $\text{MSE}^{-1/2}$. A wide range of model specifications could be considered in this context, including latent variable effects, other nonlinear transformations (see Appendix D), truncation, and random parameters models.

17.4 Convergence and Identifiability Issues with MCMC Bayesian Models

So far this chapter has focused on the estimation of Bayesian models using MCMC methods and their interpretation. However, as one would expect, there are issues concerning the assessment of fit of various models and model convergence. Model convergence assessment (the ability to determine when the Markov Chain has converged to the desired, stable, posterior distribution) is discussed first, followed by a discussion of various Bayesian goodness-of-fit criterion.

Model convergence associated with MCMC Bayesian statistical models is analogous to maximizing likelihood functions in classical methods. In the MCMC framework, it is important to ensure that the posterior distributions of all sampled parameters have achieved stationarity. Because statistics are computed on the posterior distributions obtained from the Markov Chains

(see footnote on Table 17.4 for example), it is important to sample from those posteriors only when the chains have converged. Otherwise the sampled posteriors are incorrect, leading to excessive variance, bimodal distributions, and other problems. It is generally true that complex Bayesian models will converge slower than simple ones, while model identifiability problems will also contribute to poor or slow convergence. Uninformative priors can also contribute to slow or poor convergence of MCMC models (especially with complex models) as can correlation among model parameters. Centering or standardizing variables can often improve convergence properties of Bayesian MCMC models (Congdon 2003).

In practice, running multiple simultaneous chains with different starting values of parameters assists in diagnosing both poor identifiability and slow convergence of models. For example, one might simultaneously estimate a single model using two or three chains and examine the convergence properties of several model parameters across the chains. Often, poorly identified models are improved by assigning more informative priors, improved starting values, estimating fewer model parameters, and by centering or standardization of variables.

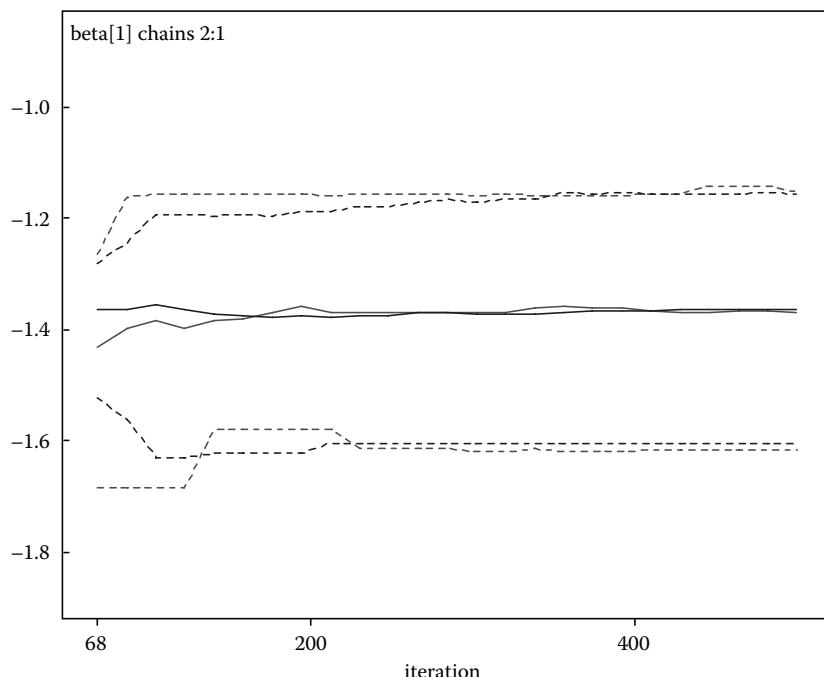
There are other methods for assessing model convergence, such as the Gelman-Rubin statistic as modified in Brooks and Gelman (1998). The basic idea is to monitor chains from divergent starting values and assess convergence by comparing within- and between-chain variation. When the between-variation becomes sufficiently small then the chains have converged. Once chains have converged, postconvergence statistics based on the posterior should reliably reflect the posterior distributions. In practice, it is important to ensure convergence of Bayesian MCMC models before characterizing the posteriors.

Example 17.3

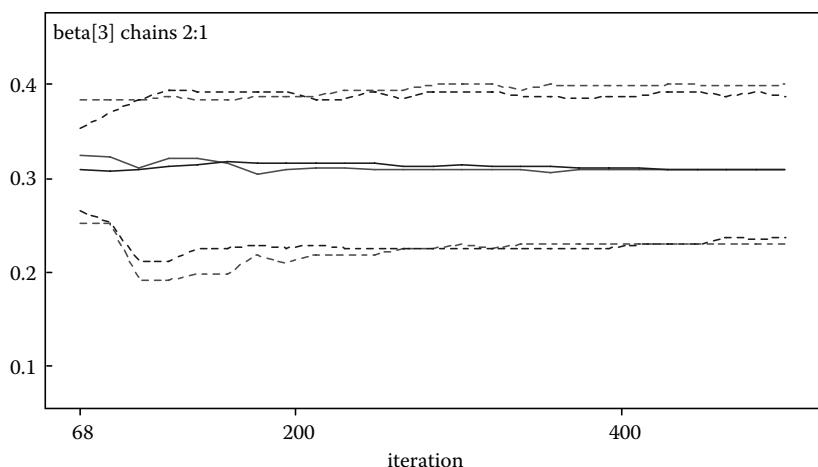
Recall the Bayesian linear regression model of AADT in Table 17.2 and discussed in Example 17.1. Suppose there is interest in examining the convergence properties of this model. Using the 2.5% and 97.5% posterior density values as a rough guide, starting or initial values for the regression parameters of two simultaneous chains are set such that

$$\begin{aligned} \text{Initial Parameter Values Chain 1} &= \{-1.6, 0.006, 0.2, 0.8\} \quad \text{and} \\ \text{Initial Parameter Values Chain 2} &= \{-1.2, 0.02, 0.4, 1.5\}. \end{aligned}$$

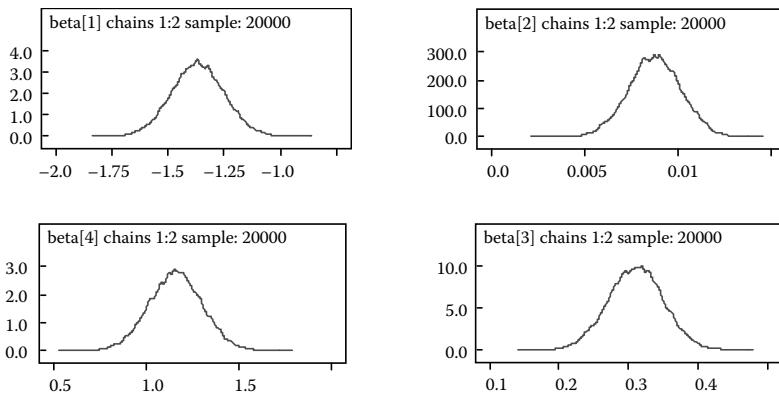
The means are monitored and 95% credible intervals for the model parameters from the simultaneous chains sampled, using these two sets of initial values. Figures 17.1 and 17.2 show the statistics for β_1 and β_2 over about 500 samples (iterations) produced by the Markov Chain and Gibbs sampler. It appears that for β_1 , the two chains (shown in red and blue) become stable around the 400th iteration, while for β_2 converged around 425th (the total number of iterations

**FIGURE 17.1**

Sampled values of β_1 for two chains: means and 95% credible intervals.

**FIGURE 17.2**

Sampled values of β_3 for two chains: means and 95% credible intervals.

**FIGURE 17.3**

Posterior densities of β 's: Bayesian MCMC linear regression model.

was 500). Model parameters β_2 and β_4 (not shown) converged around the 325th and 200th iterations, respectively. Thus, sampling 20,000 chains before drawing sample statistics for this model ensures that statistics are based on the stable posterior distribution. Figure 17.3 shows sampled posterior distributions of the β 's for this model. As is shown, the parameters appear somewhat normally distributed and also are clearly stochastic. Because the priors in this particular model were uninformative (i.e., they do not influence the shape of the posterior distribution), the normal distribution is a result of the Central Limit theorem and not of an influential prior.

17.5 Goodness-of-Fit, Sensitivity Analysis, and Model Selection Criterion Using MCMC Bayesian Models

Of course the appropriate error structure, explanatory variables, and functional form of a statistical model is not always known a priori. Model selection, is often aided by tools to help quantify goodness-of-fit and to make useful comparisons across models. The discussion here on goodness-of-fit measures focuses on three commonly applied methods: the deviance, the scaled deviance, Akaike's information criterion and the Bayes' information criterion.

Suppose L denotes the data likelihood $L = P(\mathbf{X} | \boldsymbol{\theta})$ and D denotes the deviance of a model involving p parameters. The deviance is typically defined as minus twice the log-likelihood, or

$$D = -2LN(L) \quad (17.6)$$

While the deviance is often useful, it lacks an absolute scale, and so a scaled deviance, D' , is often used, where

$$D' = -2\text{LOG}\left(\frac{L}{L_s}\right) \quad (17.7)$$

and L_s is the saturated likelihood obtained by an exact fit of predicted to observed data.

To encourage model parsimony—capturing complexity as simply as possible—a penalized measure (akin to the adjusted R -square measure in linear regression) is obtained by applying Akaike's information criterion (Akaike 1973) to obtain the deviance information criterion (Spiegelhalter et al. 2002), expressed as

$$\text{DIC} = D + 2p \quad \text{or} \quad \text{DIC} = D' + 2p \quad (17.8)$$

where the model with the lowest AIC is preferred, all else being equal, and p is the number of model parameters as defined previously. In MCMC estimation, the effective number of parameters is estimated as the difference between the posterior mean of the deviance and the deviance of the posterior means (see WinBUGS 2007). When AIC is applied to compare models, reduced model deviance is offset by increased model parameters penalizing more complex models.

An additional penalized goodness-of-fit measure, the Bayesian information criterion (BIC), is an asymptotic approximation to the Bayesian posterior probability of a model, the probability that the model is most likely given the data (Schwarz 1978). Although there are different forms of the BIC, the most common version (Congdon 2003) is given as

$$\text{BIC} = D + p\text{LN}(n) \quad (17.9)$$

where n is the sample size and again, where the model with the lowest BIC is preferred, all else being equal. The BIC penalizes complex models more heavily than the AIC.

Despite the utility of these goodness-of-fit measures, complications in MCMC estimation do arise (Congdon 2003). First, DIC and BIC are sampled parameters and thus result in probability densities during MCMC estimation. Thus, one might compare average values of BIC and AIC to across models, as opposed to point estimates in classical estimation procedures. Second, DIC and BIC are typically sensitive to multimodal posteriors, often rendering them biased. Finally, BIC and AIC are ill defined for random effects models with many parameters and therefore should be avoided.

The majority of other standard goodness-of-fit measures are suitable for examining Bayesian models. For example, one can easily calculate R -square, adjusted R -square, mean absolute deviations, MSEs, and similar measures

TABLE 17.5

Goodness-of-Fit Statistics for MCMC Bayesian Linear Regression Models I through V (Posterior Means)

Goodness-of-Fit Statistic	Mean absolute error	$MSE^{-1/2}$	DIC	BIC
Model I: $\beta's \sim N(\mu = 0, \sigma^2 = 100,000)$	0.302	0.4526	155.072	174.248
Model II: $\beta_i \sim U(lower = 0.008, upper = 0.021)$	0.305	0.4516	154.244	171.401
Model III: $\beta_i \sim N(\mu = 0.0132, \sigma = 0.00507)$	0.303	0.4526	154.974	173.868
Model IV: $\beta_i \sim N(\mu = 0.0167, \sigma = 0.00458)$	0.305	0.4515	155.112	173.958
Model V: Nonlinear parameter: CNTYPOP β^2	0.299	0.4443	151.657	173.160

All parameters obtained with convergence sample 20,000, postconvergence sample 10,000.

(penalized and nonpenalized). The main difference is that in MCMC estimation of Bayes' models, goodness-of-fit measures are reflected in densities and not in point estimates. Thus, an R -square produced via classical methods will produce a single estimate, whereas a Bayesian model via MCMC will yield a posterior credible interval on R -square.

Bayesian statistical models, by their nature and derivation, are sensitive to assumptions about priors. This sensitivity is in contrast to classically estimated models, whose parameters are completely determined by the data at hand. One must be careful to examine the sensitivity of model results to subjectively specified priors, to defend the choice of priors and also understand the impact of those choices on model estimation results. As one might expect, the use of priors based on well-accepted prior knowledge or research is easier to defend than priors that may appear arbitrary. One must also remember that the relative precision of the data and the prior will determine how the degree of influence priors will have on model parameters. Thus a sensitivity analysis might include adjustments to the precision estimates of priors as well as the means.

Example 17.4

Suppose one wishes to compare the Bayesian linear regression models of the preceding examples—ranging from Model I (standard linear regression model with uninformative priors) to Model V (nonlinear in parameters regression function) using the various goodness-of-fit statistics. Table 17.5 shows the five models across four goodness-of-fit measures.

Inspection of the goodness-of-fit statistics suggests that Model V(the model with a nonlinear parameter for county population) is superior by all measures except the BIC, by which it is the second best measure. If Model V is appealing from a practical standpoint, that is, the collection of variables is appealing, the functional form is plausible, and the magnitude of effects are defensible, then it may be the preferred model among the five candidates.

Appendix A

Statistical Fundamentals

This appendix provides a brief review of statistical fundamentals that are used in chapters in this book. Because this appendix serves solely as a review of fundamentals, readers wanting greater depth should consult other references that provide more detailed coverage of the topics. For methods from an engineering perspective, see Ash (1993), Johnson (1994), Vardeman (1994), Kotegoda and Rosso (1997), Rosenkrantz (1997), Ayyub (1998), Devore and Farnum (1999), Haldar and Mahadevan (2000), and Vardeman and Jobe (1994). For methods presented from statistics, economics, social sciences, and biology perspectives, see Larson and Marx (1986), Pedhazur and Pedhazur-Schmelkin (1991), Glenberg (1996), Freund and Wilson (1997), Steel et al. (1997), Smith (1998), and Greene (2000).

A.1 Matrix Algebra Review

A review of matrix algebra is undertaken in this section because matrix algebra is a critical and unavoidable tool in multivariate statistical methods. Computations are easier when using matrix algebra than when using simple algebra. Before standard matrix notation is introduced, summation operators, which are useful when working with matrices, are presented. Some common summation operators include

$$\begin{aligned}\sum_{i=1}^N X_i &= X_1 + X_2 + \cdots + X_N \\ \sum_{i=1}^N C &= NC \\ \sum_{i=1}^N (X_i + Y_i) &= \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i \\ \sum_{i=1}^n (X_i C + K) &= C \sum_{i=1}^n X_i + NK\end{aligned}\tag{A.1}$$

An $n \times p$ matrix is expressed as

$$\begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \quad (\text{A.2})$$

where $i =$ the row index and $j =$ the column index of matrix \mathbf{X} . In many statistical modeling applications, the rows of matrix \mathbf{X} correspond to observations within the data set and the columns of matrix \mathbf{X} correspond to independent variables measured on each observation. Therefore a matrix $\mathbf{X}_{n \times p}$ contains observations on n individuals, each with p measured or observed attributes.

A square matrix is a matrix that has an equal number of rows and columns, or when $n = p$. A matrix containing only one column is a column vector, or simply a vector. A row vector is a matrix containing only one row. Matrices are said to be equal only if they have the same dimension and if all corresponding elements in the matrices are equal.

The transpose of a matrix \mathbf{A} is obtained by transposing the columns and rows in matrix \mathbf{A} , and is denoted \mathbf{A}' or \mathbf{A}^T , such that

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \quad (\text{A.3})$$

Matrix addition and subtraction requires that the matrices added or subtracted have the same dimension. The sum or difference of two matrices is the corresponding sum or difference of the individual elements in the two matrices, such that

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} & \dots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{np} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{np} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1p} + b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \dots & a_{np} + b_{np} \end{bmatrix} \quad (\text{A.4})$$

A.1.1 Matrix Multiplication

Matrix multiplication can be done in two ways: a matrix can be multiplied by a scalar or by another matrix. To multiply a matrix by a scalar, each element in the matrix is multiplied by the scalar to obtain the scalar–matrix product. The product of two matrices is found by computing the cross-products of the rows of matrix \mathbf{A} with the columns of matrix \mathbf{B} and summing across cross-products. In matrix multiplication, order of multiplication matters; therefore,

a matrix is premultiplied or postmultiplied by another matrix. In computing products of matrices, matrix dimensions are critical. To obtain the product \mathbf{AB} , the number of columns in matrix \mathbf{A} must equal the number of rows in matrix \mathbf{B} . The size of a resultant matrix \mathbf{C} , which results when matrix \mathbf{A} is postmultiplied by \mathbf{B} , will be a matrix \mathbf{C} that will contain rows equal to the number of rows in matrix \mathbf{A} , and columns equal to the number of columns in matrix \mathbf{B} .

Example A.1

Consider the product of two matrices \mathbf{A} and \mathbf{B} as follows:

$$\mathbf{A}_{(n \times m)} = \begin{bmatrix} 4 & 6 & 2 \\ -5 & 1 & 5 \end{bmatrix}, \text{ and } \mathbf{B}_{(m \times p)} = \begin{bmatrix} 5 & -1 \\ 3 & 3 \end{bmatrix}$$

Premultiplying matrix \mathbf{A} by matrix \mathbf{B} gives

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

Calculating the matrix entries for this matrix product results in

$$\begin{aligned} C_{11} &= (5)(4) + (-1)(-5) = 15; & C_{12} &= (5)(6) + (-1)(1) = 29 \\ C_{13} &= (5)(2) + (-1)(5) = 5; & C_{21} &= (3)(4) + (3)(-5) = -3 \\ C_{22} &= (3)(6) + (3)(1) = 21; & C_{23} &= (3)(2) + (3)(5) = 21 \end{aligned}$$

Thus, the resultant matrix $\mathbf{C}_{2 \times 3}$ is given as

$$\mathbf{C}_{(2 \times 3)} = \begin{bmatrix} 15 & 29 & 5 \\ -3 & 21 & 21 \end{bmatrix}$$

Note that the matrix product \mathbf{AB} is not defined.

When computing products of matrices, the product \mathbf{AB} generally does not equal the product \mathbf{BA} . In fact, it is possible to have the matrix product \mathbf{AB} defined and the product \mathbf{BA} not defined, as in Example A.1. If two matrices \mathbf{A} and \mathbf{B} can be multiplied, then they are said to be conformable. Calculation of matrix products for dimensions up to 3×3 is fairly easy by hand, but as the size of matrices increases, the matrix product computation becomes cumbersome. Computers are ideal for multiplying matrices.

A.1.2 Linear Dependence and Rank of a Matrix

Column vectors are linearly dependent if one vector can be expressed as a linear combination of the other. If this is not the case, then the vectors are linearly independent.

Example A.2

A matrix \mathbf{A} is given as

$$\mathbf{A}_{(3 \times 4)} = \begin{bmatrix} 10 & 4 & 5 & 1 \\ 1 & 4 & 5 & 6/4 \\ 5/3 & 1 & 5/4 & 1 \end{bmatrix}$$

Note that the second column is a linear combination of the third column, such that

$$A_{i,2} = \frac{4}{5} A_{i,3}$$

Stated more formally, scalar multiples of the columns of matrix $\mathbf{A}_{n \times m}$ can be sought such that the sum of scalar and column vectors $\lambda_1 \mathbf{C}_1 + \lambda_2 \mathbf{C}_2 + \dots + \lambda_m \mathbf{C}_m = 0$. If a combination of scalars $\lambda_1, \lambda_2, \dots, \lambda_m$ is found such that the linear combination sums to zero, then at least two of the rows are said to be linearly dependent. If the only set of scalars for which this holds is the set of all 0, then the matrix is said to be linearly independent. Linear or near-linear dependencies can arise in statistical modeling and cause difficulties in data analysis. A linear dependency between two columns suggests that the two columns contain the same information, differing only by a constant.

The rank of a matrix is defined to be the maximum number of linearly independent columns in the matrix. In the previous example, the rank of the matrix is 3. A matrix is said to have full rank if none of the column vectors is linearly dependent. In statistical modeling applications, a full rank matrix of independent variables suggests that each variable is measuring different information, or a different dimension in the data.

A.1.3 Matrix Inversion (Division)

Matrix inversion is the equivalent to division in algebra. In (nonmatrix) algebra, if a number is multiplied by its inverse or reciprocal, the result is always 1, such that $(12)(1/12) = 1$. In matrix algebra, the inverse of a matrix \mathbf{A} is another matrix and is denoted by \mathbf{A}^{-1} . The inverse property holds such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, where \mathbf{I} is the identity matrix, which is a diagonal matrix

whose elements on the main diagonal are all ones and all other elements are zeros, such that

$$\begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \quad (\text{A.5})$$

A matrix inverse is defined only for square matrices. Some matrices do not have an inverse, but if a matrix does have an inverse, the inverse is unique. Also, an inverse of a square $r \times r$ matrix exists only if the matrix has full rank. Such a matrix is said to be nonsingular. A $r \times r$ matrix with rank less than r is said to be singular, and does not have an inverse.

In ordinary algebra the problem $10x = 30$ is solved by dividing both sides of the equation by 10 to obtain $x = 3$. To illustrate the equivalent operation using matrix algebra, consider the following example.

Example A.3

The following three equations have three unknowns, X , Y , and Z :

$$6X + 8Y + 10Z = 40$$

$$10X - 9Y + 24Z = 26$$

$$14X - 10Y - 6Z = 12$$

These equations can be rewritten in matrix form as $\mathbf{AB} = \mathbf{C}$, where

$$\mathbf{A} = \begin{bmatrix} 6 & 8 & 10 \\ 10 & -9 & 24 \\ 14 & -10 & -6 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 40 \\ 26 \\ 12 \end{bmatrix}$$

Then, to solve for the unknown quantities in vector \mathbf{B} , the following matrix operations are performed

$$\mathbf{AB} = \mathbf{C}$$

$$\mathbf{A}^{-1}\mathbf{AB} = \mathbf{A}^{-1}\mathbf{C} \quad (\text{premultiply both sides by } \mathbf{A}^{-1}).$$

$$\mathbf{IB} = \mathbf{A}^{-1}\mathbf{C} \quad (\text{a matrix multiplied by its inverse is I}).$$

$$\mathbf{B} = \mathbf{A}^{-1}\mathbf{C} \quad (\text{a matrix multiplied by I is itself}).$$

One need only find the inverse of matrix \mathbf{A} to solve for the unknown quantities in matrix \mathbf{B} .

Computing matrix inversions for matrices larger than size 3×3 is a formidable computation, and justifies the extensive use of computers for these types of problems. In statistical modeling applications, statistical analysis software is used to perform these operations. The general formula for computing the individual ij^{th} element of matrix \mathbf{A}^{-1} , the inverse of matrix \mathbf{A} is given as

$$a^{ij} = \frac{|C_{ij}|}{|\mathbf{A}|} \quad (\text{A.6})$$

where C_{ij} is the ji^{th} cofactor of \mathbf{A} , and $|\mathbf{A}|$ is the determinant of \mathbf{A} (Greene 2000). The interested reader can examine cofactors, determinants, and matrix inversion computations in more detail by consulting any number of references on matrix algebra (such as Greene 2000 and Smith 1998).

A.1.4 Eigenvalues and Eigenvectors

Eigenvalues are the characteristic roots of a matrix, whereas eigenvectors are the characteristic vectors. Characteristic roots or eigenvalues of a matrix have many applications. A useful set of results for analyzing a square matrix \mathbf{A} arises from the solutions to the following sets of equations:

$$\mathbf{AE} = \boldsymbol{\lambda}\mathbf{E} \quad (\text{A.7})$$

where \mathbf{A} is a square matrix, $\boldsymbol{\lambda}$ is a vector of eigenvalues, and \mathbf{E} contains the matrix eigenvectors. The eigenvalues that solve Equation A.7 can be obtained by performing matrix manipulations, constraining the solutions such that $\mathbf{E}^T\mathbf{E} = 1$ (to remove the indeterminacy), and then solving

$$|\mathbf{A} - \boldsymbol{\lambda}\mathbf{I}| = 0 \quad (\text{A.8})$$

The solutions to these equations are nonzero only if the matrix $|\mathbf{A} - \boldsymbol{\lambda}\mathbf{I}|$ is singular or has a zero determinant. The eigenvalues of a symmetric matrix will always be real, and fortunately most matrices that are solved for eigenvalues and eigenvectors in statistical modeling endeavors are symmetric.

Example A.4

To find the characteristic roots of the matrix \mathbf{X} , where

$$\mathbf{X} = \begin{bmatrix} 4 & 2 \\ 1 & 5 \end{bmatrix}$$

Equation A.8 is applied, so

$$|\mathbf{X} - \lambda \mathbf{I}| = \begin{vmatrix} 4 - \lambda & 2 \\ 1 & 5 - \lambda \end{vmatrix} = (4 - \lambda)(5 - \lambda) - (2)(1) = \lambda^2 - 9\lambda + 18$$

The solutions or eigenvalues of this n^{th} order polynomial are 6 and 3.

With the vector of eigenvalues determined, the original formulation shown in Equation A.7 can be used to find the eigenvectors. Equation A.7 can be manipulated to obtain

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{E} = \mathbf{0} \quad (\text{A.9})$$

Equation A.9 can then be used to find the eigenvectors of matrix \mathbf{A} . Recall, however, that the constraint $\mathbf{E}^T \mathbf{E} = \mathbf{1}$ must be imposed to result in a unique solution.

Example A.5

Continuing with the previous example, the eigenvectors of the matrix \mathbf{X} are obtained by applying Equation A.9

$$(\mathbf{X} - \lambda \mathbf{I})\mathbf{E} = \mathbf{0} = \left(\begin{bmatrix} 4 - \lambda & 2 \\ 1 & 5 - \lambda \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Substituting the values of 6 and 3 for λ , respectively, results in the following solutions:

$$\begin{aligned} \text{substituting } \lambda = 6 : -2E_1 + 2E_2 = 0 &\quad \text{and} \quad E_1 - E_2 = 0 \\ \text{substituting } \lambda = 3 : E_1 + 2E_2 = 0 &\quad \text{and} \quad E_1 - 2E_2 = 0 \end{aligned}$$

There are an infinite number of solutions to these equations because the constraint $\mathbf{E}^T \mathbf{E} = \mathbf{1}$ has not been met. When both constraints are met, the eigenvectors can be shown to be

$$\mathbf{E}_{\lambda=6} = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \mathbf{E}_{\lambda=3} = \pm \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix}$$

A.1.5 Useful Matrices and Properties of Matrices

There are numerous useful matrices that arise in statistical modeling. A symmetric matrix is symmetric if $\mathbf{A} = \mathbf{A}^T$; thus a symmetric matrix must also be square. A diagonal matrix is a square matrix whose off-diagonal

elements are all zeros. There are two types of diagonal matrices, the identity matrix and the scalar matrix. The identity matrix is denoted by \mathbf{I} . It is a diagonal matrix whose elements on the main diagonal are all ones, as shown in Equation A.5. Premultiplying or postmultiplying any $r \times r$ matrix \mathbf{A} by a $r \times r$ identity matrix \mathbf{I} results in \mathbf{A} , such that $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$. The scalar matrix is a diagonal matrix whose main-diagonal elements are a constant. Multiplying an $r \times r$ matrix \mathbf{A} by the $r \times r$ scalar matrix $\lambda\mathbf{I}$ is equivalent to multiplying \mathbf{A} by the scalar λ . A column vector with all elements equal to 1 is denoted by \mathbf{I} . A square matrix with all elements equal to 1 is denoted by \mathbf{J} .

Some useful theorems (shown without proof) in matrix algebra, which enable calculations necessary in statistical modeling applications, include the following:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

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

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

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$$

$$\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$$

$$(\mathbf{A}^T)^T = \mathbf{A}$$

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$$

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

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

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

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

A.1.6 Matrix Algebra and Random Variables

Matrix algebra is extremely efficient for dealing with random variables, random vectors, and random matrices. In particular, matrix algebra is extremely useful for manipulating matrices used in the development of statistical models and methods. This section introduces some of the basic matrices used in the statistical modeling of random phenomena. Other matrices are shown in the text as necessary.

A common starting point in many statistical modeling applications is a matrix of random variables of size $n \times p$

$$\mathbf{X}_{n \times p} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \quad (\text{A.10})$$

where the p columns in matrix \mathbf{X} represent variables (e.g., gender, weight, number of lanes, etc.), and the n rows in matrix \mathbf{X} represent observations across sampling units (e.g., individuals, autos, road sections, etc.).

A mean matrix $\bar{\mathbf{X}}$, representing the means across individuals, is given as

$$E[\mathbf{X}] = \bar{\mathbf{X}} = \begin{bmatrix} E[x_{11}] & \dots & E[x_{1p}] \\ \vdots & \ddots & \vdots \\ E[x_{n1}] & \dots & E[x_{np}] \end{bmatrix} \quad (\text{A.11})$$

where the means of the elements in the mean matrix $\bar{\mathbf{X}}$ are calculated using

$$E[x_{ij}] = \frac{\sum_{i=1}^n x_{ij}}{n}; \quad j = 1, 2, \dots, p \quad (\text{A.12})$$

A deviations matrix is obtained by subtracting matrix $E[\mathbf{X}]$ from \mathbf{X} . Building on the fact that $VAR[\mathbf{X}_i] = E[\mathbf{X}_i - E[\mathbf{X}_i]]$, in matrix form (with matrix sizes shown) the variance–covariance matrix of matrix \mathbf{X} is obtained using

$$VAR[\mathbf{X}]_{p \times p} = E \left\{ \left[\mathbf{X}_{n \times p} - E[\mathbf{X}]_{n \times p} \right]_{p \times n}^T \left[\mathbf{X}_{n \times p} - E[\mathbf{X}]_{n \times p} \right]_{p \times n} \right\} \quad (\text{A.13})$$

A correlation matrix is obtained by computing the variance–covariance matrix using standardized variables, where

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}; \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, p$$

replace the original x_{ij} terms.

A.2 Probability, Conditional Probability, and Statistical Independence

The probability of an event or outcome is the proportion of times the event occurs in a long-run sequence or number of trials (Johnson 1994). In notation, probability is defined as

$$P(A) = \frac{\text{count}(A)}{\text{trials}}; \quad \text{trials} \rightarrow \infty \quad (\text{A.14})$$

where $\text{count}(A)$ is the number of times event A occurs, and trials is the number of times the experiment was repeated or the recorded number of observations where event A could have occurred, typically denoted as n . Note that $P(A)$ converges in probability as the number of trials approaches infinity. Thus, one view of probability supports the notion that many trials need to be observed before obtaining a reliable estimate of probability. In contrast, a Bayesian statistician can generate estimates of probability using subjective information (see Chapter 2 for additional details).

Conditional probability is defined as the probability of one event, say, event A , given that event B has already occurred. In notation, conditional probability is given as

$$P(A | B) = \frac{P(AB)}{P(B)} \quad (\text{A.15})$$

where $P(AB)$ is the joint probability of events A and B occurring together—their joint probability. Conditional probability is a cornerstone of statistical modeling. For example, many statistical models explain or predict the probability of an outcome, Y , given the independent variables \mathbf{X} , such that the model provides $P(Y|\mathbf{X})$.

Statistical hypothesis tests are conditional probabilities. A classical or frequentist statistical hypothesis test, presented in general form, is given as

$$P(\text{data} | \text{true null hypothesis}) = \frac{P(\text{data}) P(\text{true null hypothesis})}{P(\text{true null hypothesis})} \quad (\text{A.16})$$

Often the right-hand side of the denominator in Equation A.16 is unknown and is sought. Thus, the classical hypothesis test does not provide the probability of the null hypothesis being true, but instead the conditional probability of observing the data given a true null hypothesis. Despite this fact, the conditional probability does provide objective evidence regarding the plausibility of the null hypothesis.

Bayes' theorem is also based on the notion of conditional probability and is derived in the following manner:

$$\begin{aligned} P(A|B) &= \frac{P(AB)}{P(B)}; \text{ and} \\ P(B|A) &= \frac{P(AB)}{P(A)}; \text{ so, by substitution} \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)}; \text{ which is Bayes' theorem} \end{aligned} \quad (\text{A.17})$$

Bayes' theorem is used to circumvent practical and philosophical problems of the classical hypothesis test result.

Statistical independence is a useful concept in the development of some statistical methods. It describes a condition where the probability of occurrence of one event A is entirely unaffected by the probability of occurrence of another event B . So if events A and B are independent, then

$$P(AB) = P(A)P(B) \quad (\text{A.18})$$

By using the conditional probability formula in Equation A.15, it can be seen that for statistically independent events

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A) \quad (\text{A.19})$$

Equation A.19 reveals that if events A and B are statistically independent, then the probability that event A occurs given that event B has already occurred is the probability of event A . This is intuitive because event B does not influence event A .

A.3 Estimating Parameters in Statistical Models—Least Squares and Maximum Likelihood

In the text of the book, many pages are devoted to estimating parameters in statistical models. Parameters are an integral part of statistical models. Often, beginning with a theory of a process or phenomenon that is known to generate data, a model is postulated that may involve any number of variables. The general model may be

$$\mathbf{Y} = f(\boldsymbol{\theta}; \mathbf{X}) + \boldsymbol{\varepsilon} \quad (\text{A.20})$$

where \mathbf{Y} is a vector of outcomes, $\boldsymbol{\theta}$ is a vector of estimated model parameters, \mathbf{X} is a matrix of variables across n observations thought to impact or influence \mathbf{Y} , and $\boldsymbol{\varepsilon}$ is a vector of disturbances—that portion of \mathbf{Y} not accounted for by the function of $\boldsymbol{\theta}$ and the \mathbf{X} . The majority of the text of this book discusses various ways of organizing functions f that relate $\boldsymbol{\theta}$ and the matrix \mathbf{X} to the outcome vector \mathbf{Y} . For example, the ordinary least squares regression model relates the \mathbf{X} terms to \mathbf{Y} through the expression $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\theta}$ is the vector of estimated parameters or the $\boldsymbol{\beta}$. The parameters in this or any statistical model are estimated from observational or experimental data.

There are numerous methods for estimating parameters in statistical models. Parameter estimation methods include ordinary or unweighted least squares, weighted least squares, maximum likelihood, and method of moments, to name but a few. These methods are discussed in detail, where appropriate, in the development of statistical models in the text. In this section two common parameter estimation methods are briefly presented, ordinary least squares and maximum likelihood, mainly to familiarize the reader with these concepts.

Ordinary least squares estimation involves minimizing the squared differences between observed and predicted observations, such that

$$Q = \text{MIN} \left(Y_i - \hat{Y}_i \right)^2 \quad (\text{A.21})$$

where \hat{Y}_i is the value of Y predicted by the statistical model for the i^{th} trial or observation. The predicted value of Y is a function of the \mathbf{X} terms and the collection of estimated parameters $\boldsymbol{\theta}$. In least squares estimation, the parameter estimates are obtained by solving Equation A.21. Note that no assumptions about statistical distributions are required to obtain parameter estimates using least squares estimation.

Maximum likelihood methods also provide estimates of parameters in statistical models; however, the method for obtaining them is fundamentally different. The principle behind maximum likelihood is that different populations generate different samples, and so any particular sample is more likely to come from some populations rather than others. For example, if a random sample of y_1, y_2, \dots, y_n was drawn, there is some parameter $\boldsymbol{\theta}$ (for simplicity assumed to be the sample mean) that is most likely to generate the sample. Figure A.1 shows two different statistical distributions A and B , which represent two different assumed sample means. In the figure, the sample mean θ_A associated with distribution A is much more likely to generate the sample of y than the sample mean θ_B associated with distribution B . Maximum likelihood estimation seeks the parameter or set of parameters that are most likely to have generated the observed data (y) among all possible θ . Unlike least

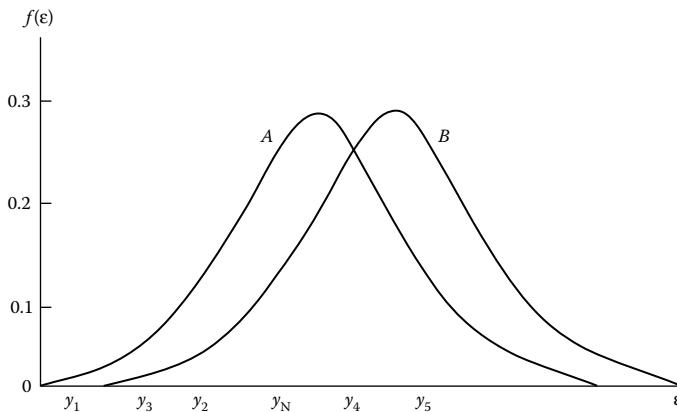
**FIGURE A.1**

Illustration of maximum likelihood estimation.

squares estimation, a particular probability distribution must be specified to obtain maximum likelihood estimates.

Different parameter estimation methods and their associated properties are discussed in the text of the book when appropriate. Desirable properties of estimators in general terms are discussed in Chapter 3.

A.4 Useful Probability Distributions

Probability distributions are key to statistical and econometric analysis. By identifying the properties of underlying data-generating processes, one can easily answer probability-related questions, such as the probability of the occurrence of an event. Fortunately, well-understood probability distributions approximate a large number of commonly studied phenomena. For example, arrival times of vehicles at the back of a queue and accident occurrence on a section of roadway are often approximated well by a Poisson distribution. Many samples, such as vehicle speeds on a segment of freeway, are well approximated by normal distributions.

There are generally two types of probability distributions—continuous and discrete. Continuous distributions arise from variables that can take on any value within a range of values, and are measured on the ratio or interval scale. Discrete distributions arise from ordinal data, or count data—data that are strictly continuous but can only take on integer values. Two useful properties concerning probability distributions are that the probability of any particular outcome lies between 0 and 1, and the sum of probabilities over all possible outcomes is 1.

In this section four commonly used statistical distributions are discussed: the standard normal Z distribution, the t distribution, the χ^2 distribution, and the F distribution. Tables of probabilities for these distributions are provided in Appendix C. Other probability distributions are discussed in relation to specific modeling applications when necessary. In addition, numerous probability distribution functions are described in Appendix B.

A.4.1 The Z Distribution

The Z distribution can be derived from the central limit theorem. If the mean \bar{X} is calculated on a sample of n observations drawn from a distribution with mean μ and known finite variance σ^2 , then the sampling distribution of the test statistic Z change to Z^* is approximately standard normal distributed, regardless of the characteristics of the population distribution. For example, the population could be normal, Poisson, binomial, or beta distributed. The standard normal distribution is characterized by a bell-shaped curve, with mean zero and variance equal to one. A random variable Z^* is computed as follows:

$$Z^* = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \approx Z_\alpha \quad (\text{A.22})$$

where α denotes a confidence level.

The statistic Z^* is a random variable whose distribution approaches that of the standard normal distribution as n approaches infinity. This is an extremely useful result. It says that regardless of the distribution of the population from which samples are drawn, a test statistic computed using Equation A.22 will approach the standard normal distribution as the sample becomes large. In practice, sample sizes of 20 or more will result in a virtually normally distributed sampling distribution of the mean (Glenberg 1996).

The Z distribution is also extremely useful for examining normal distributions with any mean and variance. By using the standard normal transformation,

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (\text{A.23})$$

original variables X_i obtained from any normal distribution can be standardized to new variables Z_i that are standard normal distributed. The distribution functions for both the normal distribution and the standard normal distribution are provided in Appendix B. Percentiles of the standard normal distribution Z are provided in Appendix C.

A.4.2 The t Distribution

The t distribution is used for conducting many of the statistical tests described throughout this book. Recall that the Z^* statistic forms the basis for confidence interval estimation and hypothesis testing when σ^2 is known. When σ^2 is unknown, however, it is natural to replace it with its unbiased estimator s^2 . When this is done, a test statistic t^* is obtained

$$t^* = \frac{\bar{X} - \mu}{s / \sqrt{n}} \approx t_\alpha (v = n - 1) \quad (\text{A.24})$$

where t^* is approximately t distributed with $n - 1$ degrees of freedom.

This probability distribution was developed by W. S. Gossett, an employee of the Guinness Brewery, who in 1919 published his work under the pseudonym "Student." Gossett called the statistic t and, since then, its distribution has been called Student's t distribution. The t distribution is similar to the standard normal. Figure A.2 depicts both a t distribution and a standard normal distribution. Like the standard normal distribution, t is symmetric around zero. It is mound shaped, whereas the normal distribution is bell shaped. The extent to which the t distribution is more spread out than the normal distribution is determined by the degrees of freedom of the distribution. As the sample size becomes large, the t distribution approaches the standard normal Z distribution. It is important to note also that the t distribution requires that the population from which samples are drawn is normal.

The distribution functions for t distribution are provided in Appendix B. Percentiles of the t distribution are provided in Appendix C.

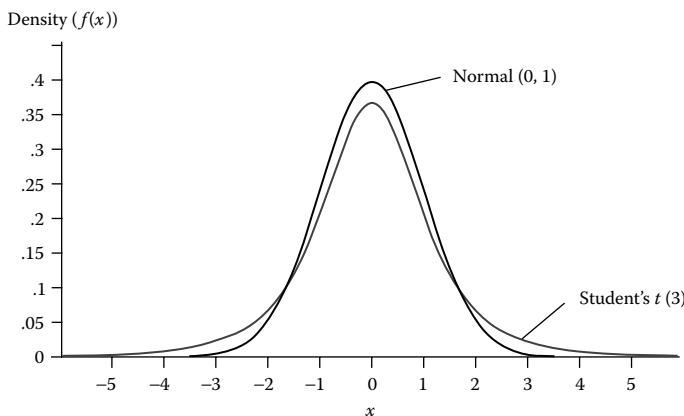


FIGURE A.2

The standard normal Z and the $t_{(3)}$ probability density functions.

A.4.3 The χ^2 Distribution

The χ^2 distribution is extremely useful because it arises in numerous common situations. Statistical theory shows that the square of a standard normal variable Z is χ^2 distributed with 1 degree of freedom. Also, let Z_1, Z_2, \dots, Z_k be k independent standard normal random variables. If each of these random variables is squared, their sum will follow a χ^2 distribution with k degrees of freedom, such that

$$X^2 = \sum_{i=1}^n Z_i^2 \approx \chi^2_{(k)} \quad (\text{A.25})$$

The X^2 statistic shows that the χ^2 distribution arises from the sum of independent squared normal random variables. As a sum of squares, the χ^2 random variable cannot be negative and, as a result, is bounded on the low end by zero. The χ^2 distribution is therefore skewed to the right. Figure A.3 shows several χ^2 distributions with different degrees of freedom. Note that as degrees of freedom increase, the χ^2 distribution approximates the normal distribution. In fact, as the degrees of freedom increase, the χ^2 distribution approaches a normal distribution with mean equal to the degrees of freedom and variance equal to two times the degrees of freedom.

Another useful application of the χ^2 distribution is to assess whether the variance of the sample is the same as the variance in the population. If s^2 is

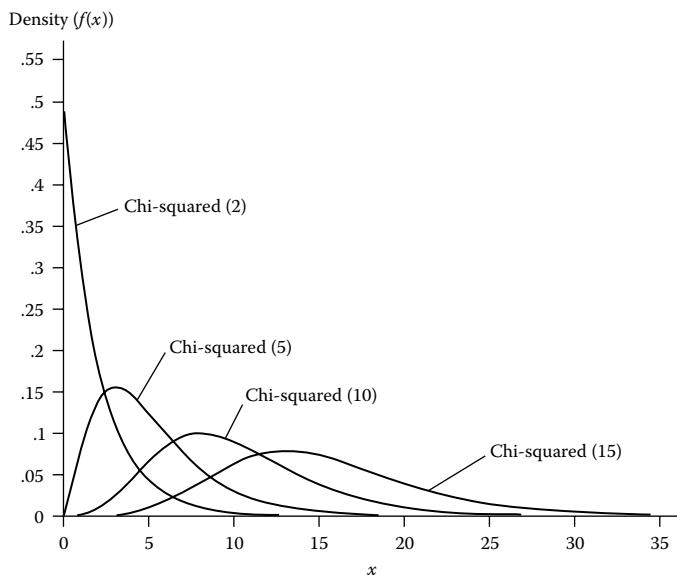


FIGURE A.3

The χ^2 density function with various degrees of freedom.

the estimated variance of a random sample of size n drawn from a normal population having variance σ^2 then the sampling distribution of the test statistic X^2 is approximately χ^2 distributed with $v = n - 1$ degrees of freedom, such that

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \approx \chi^2_\alpha (v = n-1) \quad (\text{A.26})$$

Finally, the χ^2 distribution can be used to compare two distributions. If frequencies of events are observed in certain categories or frequency bins under both observed and expected distributions, then the χ^2 distribution can be used to test whether the observed frequencies are equivalent to the expected frequencies. The test statistic is given as

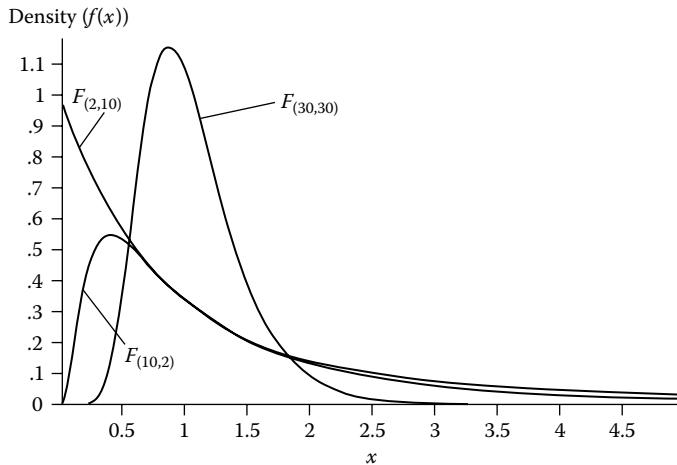
$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_\alpha [I-1, J-1] \quad (\text{A.27})$$

where I and J are the number of rows and columns in a two-way contingency table. The test statistic can easily be made to accommodate multiway tables, where there might be, for example, I rows, J columns, and K elements associated with each ij^{th} bin. In this case the degrees of freedom in Equation A.27 becomes $(I-1)(J-1)(K-1)$. The expected frequency in Equation A.27 may be the result of a model of statistical independence, a frequency distribution based on an assumed statistical distribution like the normal distribution, or an empirical distribution (where two empirical distributions are compared). Caution needs to be applied when using the test statistic shown in Equation A.27, because small expected frequencies can compromise the reliability of the test statistic. In such cases, exact methods should be applied, as described in Mehta and Patel (1983).

The distribution functions for the χ^2 distribution are provided in Appendix B. Percentiles of the χ^2 distribution are provided in Appendix C.

A.4.4 The F Distribution

Another statistical distribution that is extremely useful in statistics, and the final distribution discussed in this section, is the F distribution. This distribution is named after the English statistician Sir Ronald A. Fisher, who discovered it in 1924. The F distribution is approximated by the ratio of two independent χ^2 random variables, each of which is divided by its own degrees of freedom. For example, let χ^2_1 be a χ^2 random variable with 1 degree of freedom, and χ^2_2 be a χ^2 random variable with 2 degrees of freedom. The ratio of

**FIGURE A.4**

The F density function for different degrees of freedom combinations.

these two random variables is F distributed with $\chi_1^2 = 1$ and $\chi_2^2 = 2$ degrees of freedom, respectively. That is,

$$F^* = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2} \approx F_\alpha(\nu_1, \nu_2) \quad (\text{A.28})$$

A useful application of the F distribution is to test the ratio of variances of two samples drawn from a normal population, or to test whether two samples were drawn from a single population with variance σ^2 . If s_1^2 and s_2^2 are the estimated variances of independent random samples of size n_1 and n_2 , respectively, then the sampling distribution of the test statistic F^* is approximately F distributed with ν_1 (numerator) and ν_2 (denominator) degrees of freedom, such that

$$F^* = \frac{s_1^2}{s_2^2} \approx F_\alpha(\nu_1 = n_1 - 1, \nu_2 = n_2 - 1) \quad (\text{A.29})$$

The F distributions, shown in Figure A.4 with different degrees of freedom, are asymmetric—a quality “inherited” from the χ^2 distributions; in addition their shape resembles that of the χ^2 distributions. Note also that $F_{(2,10)} \neq F_{(10,2)}$.

The distribution functions for the F distribution are provided in Appendix B. Percentiles of the F distribution are provided in Appendix C.

Appendix B

Glossary of Terms

A

Abscissa: The horizontal or x -coordinate of a data point as graphed on a Cartesian coordinate system. The y -coordinate is called the ordinate.

Accelerated lifetime model: An approach used in hazard-based analysis of duration data that assumes covariates rescale time directly in a baseline survivor function, which is the survivor function when all covariates are zero.

Accuracy: Degree to which some estimate matches the true state of nature. Typically, accuracy is a measure of the long-run average of predictions compared to the true, often unknown average (see also precision, bias, efficiency, and consistency).

Aggregate: The value of a single variable that summarizes, adds, or represents the mean of a group or collection of data.

Aggregation: Compounding of primary data, usually for the purpose of expressing them in summary or aggregate form.

Aggregation bias: A problem that arises when forecasting aggregate (population) impacts using disaggregate nonlinear models, such as the logit model.

Alpha level: The probability selected by the analyst that reflects the degree of acceptable risk for rejecting the null hypothesis when, in fact, the null hypothesis is true. The degree of risk is not interpretable for an individual event or outcome; instead, it represents the long-run probability of making a Type I error (see also beta and Type II error).

Alternative hypothesis: The hypothesis that one accepts when the null hypothesis (the hypothesis under test) is rejected. It is usually denoted by H_A or H_1 (see also null hypothesis).

Analysis of variance (ANOVA): Analysis of the total variability of a set of data (measured by their total sum of squares) into components that are attributed to different sources of variation. The sources of variation are apportioned according to variation as a result of random fluctuations and those as a result of systematic differences across groups.

Approximation error: In general, an error due to approximation from making a rough calculation, estimate, or guess. When performing numerical calculations, approximations result from rounding errors, for example, $\pi \approx 22/7 \approx 3.1417$.

Arithmetic mean: The result of summing all measurements from a population or sample and dividing by the number of population or sample members. The arithmetic mean is also called the average, which is a measure of central tendency.

Attitude survey: Surveys individually designed to measure reliable and valid information regarding respondent's attitudes and/or preferences to assist in making critical decisions and to focus resources where they are most needed.

Autocorrelation: The temporal association across a time series of observations; also referred to as serial correlation.

Autoregressive (AR) process: Stationary time series that are characterized by a linear relationship between adjacent observations. The order of the process p defines the number of past observations upon which current observations depend.

Autoregressive integrated moving average (ARIMA) process: Time series that are made stationary by differencing and whose differenced observations are linearly dependent on past observations and past innovations.

Autoregressive moving average (ARMA) process: Stationary time series whose observations are linearly dependent on past observations and past innovations.

Average: The arithmetic mean of a set of observations. The average is a measure of central tendency of a scattering of observations, as are also the median and mode.

B

Backshift operator: Model convention that defines stepping backward in a time-indexed data series (see also time series).

Bar chart or diagram: A graph of observed data using a sequence of rectangles, whose widths are fixed and whose heights are proportional to the number of observations, proportion of total observations, or probability of occurrence.

Bayes' theorem: A theorem, developed by Bayes, that relates the conditional probability of occurrence of an event to the probabilities of other events. Bayes' theorem is given as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')},$$

where $P(A|B)$ is the probability of event A given that event B has occurred, $P(A)$ is the probability of event A , and $P(A')$ is the probability of event A not occurring. Bayes' theorem is used to overcome some of the interpretive and philosophical shortcomings of frequentist or classical statistical methods, and to incorporate subjective information for obtaining parameter estimates.

Bayesian statistical philosophy: Bayesian statisticians base statistical inference on a number of philosophical underpinnings that differ in principle from frequentist or classical statistical thought. First, Bayesians believe that research results should reflect updates of prior research results. In other words, prior knowledge should be incorporated formally into current research to obtain the best “posterior” or resultant knowledge. Second, Bayesians believe that much is gained from insightful prior, subjective information regarding to the likelihood of certain types of events. Third, Bayesians use Bayes' theorem to translate probabilistic statements into degrees of belief, instead of a classical confidence interval interpretation (see also frequentist statistical philosophy).

Before–after data: A study where data are collected before and following an event, treatment, or action. The event, treatment, or action applied between the two periods is thought to affect the data under investigation. The purpose of this type of study is to show a relationship between the data and the event, treatment, or action. In experimental research all other factors are either randomized or controlled (see also panel data, time series data, and cross-sectional data).

Bernoulli distribution: Another name for binomial distribution.

Bernoulli trial: An experiment where there is a fixed probability p of “success,” and a fixed probability $1 - p$ of “failure.” In a Bernoulli process, the events are independent across trials.

Best fit: See goodness-of-fit.

Beta distribution: A distribution, which is closely related to the F distribution, used extensively in analysis of variance. In Bayesian inference, the beta distribution is sometimes used as the prior distribution of a parameter of interest. The beta distribution, when used to describe the distribution of lambda of a mixture of Poisson distributions, results in the negative binomial distribution. The beta density is given by

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx}$$

where α and β are shape parameters, $\alpha/(\alpha + \beta)$ is the mean, and variance is $\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)]$.

Beta error: The probability assigned (or implied) by the analyst that reflects the degree of acceptable risk for accepting the null hypothesis when,

in fact, the null hypothesis is false. The degree of risk is not interpretable for an individual event or outcome; instead, it represents the long-run probability of making a Type II error (see also Type I error and alpha).

Beta parameters: The parameters in statistical models, often represented by the Greek letter β . Beta parameters are meant to represent fixed parameters of the population.

Bias: When estimating population parameters, an estimator is biased if its expected value does not equal the parameter it is intended to estimate. In sampling, a bias is a systematic error introduced by selecting items nonrandomly from a population that is assumed to be random. A survey question may result in biased responses if it is poorly phrased (see also unbiased).

Binomial Distribution: The distribution of the number of successes in n trials when the probability of success (and failure) remains constant from trial to trial and the trials are independent. This discrete distribution is also known as Bernoulli distribution or process, and is given by

$$P(Y = x, n; p) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

where x is the number of “successes” out of n trials, p is the probability of success, and ! represents the factorial operator, such that $3! = 3 \times 2 \times 1$ (see also negative binomial).

Bivariate distribution: A distribution resulting from two random variables, for example, the joint distribution of vehicle model year (MY) and annual mileage (AM) driven. If MY and AM are independent variables, then the bivariate distribution is approximately uniform across cells. Because older vehicles are driven less per year on average than newer vehicles, these two variables are dependent; that is, annual mileage is dependent on model year. A contingency table, or cross-classification analysis, is useful for testing statistical independence among two or more variables.

C

Categorical variable: Also called a nominal scale variable, a variable that has no particular ordering, and intervals between these variables are without meaning. A categorical variable is a discrete variable. Examples of categorical variables include vehicle manufacturer and gender. For a categorical variable to be defined appropriately, it must consist of mutually exclusive and collectively exhaustive categories.

Causation or causality: When event A causes event B , there exists a material, plausible, underlying reason or explanation relating event B with event A . Causality is not proved with statistics—statistics and statistical models instead provide probabilistic evidence supporting or refuting a causal relation. The statistical evidence is a necessary but not sufficient condition to demonstrate causality. Causality is more defensible in light of statistical evidence obtained from designed experiments, whereas data obtained from observational and quasi-experiments too often yield relationships where causation cannot be discerned from association (see also post hoc theorizing and correlation).

Censored distribution: A statistical distribution that occurs when a response above or below a certain threshold value is fixed at that threshold. For example, for a stadium that holds 50,000 seats, the number of tickets sold cannot exceed 50,000, even though there might be a demand for more than 50,000 seats.

Central limit theorem: If x_{ave} is calculated on a sample drawn from a distribution with mean μ and known finite variance σ^2 , then the sampling distribution of the test statistic Z is approximately standard normal distributed, regardless of the characteristics of the parent distribution (i.e., normal, Poisson, binomial, etc.). Thus, a random variable Z^* computed as follows is approximately standard normal (mean = 0, variance = 1) distributed:

$$Z^* = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx Z_\alpha$$

The distribution of random variable Z approaches that of the standard normal distribution as n approaches infinity.

Central tendency: The tendency of quantitative data to cluster around some variate value.

Chebyshev's inequality: A useful theorem, which states that for a probability distribution with mean μ and standard deviation σ , the probability that an observation drawn from the distribution differs from μ by more than $k\sigma$ is less than $1/k^2$ or, stated mathematically,

$$P\{|x - \mu| > k\sigma\} < \frac{1}{k^2}$$

Chi-square distribution: A distribution that is of great importance for inferences concerning population variances or standard deviations, for comparing statistical distributions, and for comparing estimated statistical models. It arises in connection with the sampling distribution

of the sample variance for random samples from normal populations. The chi-squared density or distribution function is given by

$$f(x) = x^{n/2-1} \text{EXP}(-x/2) / [2^{n/2} r(n/2)]; \quad x > 0$$

where $n/2$ and 2 are shape and scale parameters, respectively, and $\underline{(z)}$ is the gamma function. The chi-square distribution is a special case of the gamma distribution (with $n/2$ and 2 for shape and scale parameters).

Class: Observations grouped according to convenient divisions of the variate range, usually to simplify subsequent analysis (the upper and lower limits of a class are called class boundaries, the interval between them, the class interval, and the frequency falling into the class is the class frequency).

Classical statistical philosophy: See frequentist and Bayesian statistical philosophy.

Cluster sampling: A type of sampling whereby observations are selected at random from several clusters instead of at random from the entire population. It is intended that the heterogeneity in the phenomenon of interest is reflected within the clusters; that is, members in the clusters are not homogenous with respect to the response variable. Cluster sampling is less satisfactory from a statistical standpoint but often is more economical and/or practical.

Coefficient: See parameter.

Coefficient of determination: Employed in ordinary least squares regression and denoted R^2 , the proportion of total variance in the data taken up or “explained” by the independent variables in the regression model. It is the ratio or proportion of explained variance to total variance and is bounded by 0 and 1 for models with intercept terms. Because adding explanatory variables to a model cannot reduce the value of R^2 , adjusted R^2 is often used to compare models with different numbers of explanatory variables. The value of R^2 from a model cannot be evaluated as “good” or “bad” in singularity; it can only be judged relative to other models that have been estimated on similar phenomenon. Thus an R^2 of 30% for some phenomenon may be extremely informative, whereas for another phenomenon it might be quite uninformative.

Collectively exhaustive: When a categorical variable is collectively exhaustive it represents all possible categories into which a random variable may fall.

Compensating variation: Used in utility theory, the amount of money that would have to be paid (or taken away from) individuals to render them as well off after a change in a specific variable as they were before a change in that variable.

Conditional probability: The long-run likelihood that an event will occur given that a specific condition has already occurred, for example, the probability that it will rain today, given that it rained yesterday. The standard notation for conditional probability is $P(A|B)$, which corresponds to the probability that event A occurs given the event B has already occurred. It can also be shown that

$$P(A|B) = \frac{P(AB)}{P(B)}$$

where $P(AB)$ is the probability that both event A and B will occur.

Confidence coefficient: The measure of probability α (see alpha error) associated with a confidence interval that the interval will include the true population parameter of interest.

Confidence interval or region: A calculated range of values known to contain the true parameter of interest over the average of repeated trials with specific certainty (probability). The correct interpretation of a confidence interval is as follows: If the analyst were to draw samples repeatedly at the same levels of the independent variables and compute the test statistic (mean, regression, slope, etc.), then the true population parameter would lie in the $(1 - \alpha)\%$ confidence interval α times out of 100, conditional on a true null hypothesis.

Confounded variables: In general, a variable in a statistical model that is correlated with a variable that is not included in a model; sometimes called an omitted variable problem. When variables in a model are correlated with variables excluded from a model, then the estimates of parameters in the model are biased. The direction of bias depends on the correlation between the confounded variables. In addition, the estimate of model error is also biased.

Consistency: An estimate of a population parameter, such as the population mean or variance, obtained from a sample of observations is said to be consistent if the estimate approaches the value of the true population parameter as the sample size approaches infinity. Stated more formally

$$P(|\hat{\beta} - \beta| < \varepsilon) \rightarrow 1 \text{ as } n \rightarrow \infty \text{ for all } \varepsilon > 0$$

where $\hat{\beta}$ are estimated parameters, β are true population parameters, n is sample size, and ε is a positive small difference between estimated and true population parameters.

Contemporaneous correlation: Correlation among disturbance terms of different statistical model equations within a system of equations. This correlation is typically caused by shared unobserved effects of variates appearing in multiple equations.

Contingency coefficient: A measure of the strength of the association between two variables, usually qualitative, on the basis of data tallied into a contingency table. The statistic is never negative and has a maximum value less than 1, depending on the number of rows and columns in the contingency table.

Contingency table: Cross-classification or contingency table analysis is a statistical technique that relies on properties of the multinomial distribution, statistical independence, and the chi-square distribution to determine the strength of association (or lack thereof) between two or more factors or variables.

Continuous variable: A variable that is measured either on the interval or ratio scale. A continuous variable can theoretically take on an infinite number of values within an interval. Examples of continuous variables include measurements in distance, time, and mass. A special case of a continuous variable is a data set consisting of counts. Counts consist of nonnegative integer values (see also discrete, nominal, and ordinal data).

Control group: A comparison group of experimental units that do not receive a treatment and are used to provide a comparison to the experimental group with respect to the phenomenon of interest.

Correlation: The relationship (association or dependence) between two or more qualitative or quantitative variables. When two variables are correlated they are said to be statistically dependent; when they are uncorrelated they are said to be statistically independent. For continuous variables Pearson's product moment correlation coefficient is often used, whereas for rank or ordinal data Kendall's coefficient of rank correlation is used.

Correlation coefficient: A measure of the interdependence between two variates or variables. For interval or ratio scale variables it is a fraction, which lies between -1 for perfect negative correlation and 1 for perfect positive correlation.

Correlation matrix: For a set of variables X_1, \dots, X_n , with correlation between X_i and X_j denoted by r_{ij} , this is a square symmetric matrix with values r_{ij} .

Count data: Ratio scale data that are nonnegative integers.

Covariance: The expected value of the product of the deviations of two random variables from their respective means; that is, $E[(X_1 - \mu_1)(X_2 - \mu_2)] = \sum_{i=1-n} [(X_{1i} - \mu_1)(X_{2i} - \mu_2)]/n$. Correlation and covariance are related statistics in that the correlation is the standardized form of the covariance; that is, covariance is a measure of the association in original units, whereas the correlation is the measure of association in standardized units.

Covariates: Term often used to refer to independent variables in the model—the variables in the right-hand side of a statistical model.

Cross-sectional data: Data collected on some variable, at the same point or during a particular period of time, from different geographical

regions, organizations, households, and so on (see also panel data, time series data, and before–after data).

Cumulative distribution function: A function related to a distribution's density function that is written as $F(x) = P[X < x]$, where P denotes probability, X is a random time variable, and x is some specified value of the random variable. The density function corresponding to the distribution function is the first derivative of the cumulative distribution with respect to x , $f(x) = dF(x)/dx$.

Cumulative frequency: The frequency with which an observed variable takes on a value equal to or less than a specified value. Cumulative frequencies are often depicted in bar charts or histograms, known as cumulative frequency distributions.

D

Data: Information or measurements obtained from a survey, experiment, investigation, or observational study. Data are stored in a database, usually in electronic form. Data are discrete (measured on the nominal or ordinal scale) or continuous (measured on the interval or ratio scale).

Data mining: Process undertaken with the intent to uncover previously undiscovered relationships in data. Unfortunately, data mining leads to a higher likelihood of illusory correlation, omitted variable bias, and post hoc theorizing, all of which threaten high-quality research and scientific investigations. Data mining should be undertaken with great caution, and conclusions drawn from it should be made with sufficient caveats and pessimism. The logical role of data mining in research is for generating research ideas that deserve follow-on detailed research investigations.

Deduction: A type of logic or thinking process used in statistical hypothesis testing. Deductive logic stipulates that if event A causes or leads to event B , and event B is not observed to occur, then event A also did not occur. In probabilistic terms, if the null hypothesis is true (A), then the sample data should be observed with high probability (B). If, instead, the sample data are observed with low probability, then the null hypothesis is rejected as implausible (see also induction).

Degrees of freedom: The number of free variables in a set of observations used to estimate statistical parameters. For example, the estimation of the population standard deviation computed on a sample of observations requires an estimate of the population mean, which consumes 1 degree of freedom to estimate—thus the sample standard deviation has $n - 1$ degrees of freedom remaining. The degrees

of freedom associated with the error around a linear regression function has $n - P$ degrees of freedom, as P degrees of freedom have been used to estimate the parameters in the regression model.

Density function: Often used interchangeably with probability density function, the function that defines a probability distribution in that integration of this function gives probability values.

Dependent variables: If a function is given by $Y = f(X_1, \dots, X_n)$, it is customary to refer to X_1, \dots, X_n as independent or explanatory variables, and Y as the dependent or response variable. The majority of statistical investigations in transportation aim to predict or explain values (or expected values) of dependent variables given known or observed values of independent variables.

Descriptive statistics: Statistics used to display, describe, graph, or depict data. Descriptive statistics do not generally include modeling of data, but are used extensively to check assumptions of statistical models.

Deterministic model or process: This model, as opposed to a stochastic model, is one that contains effectively no or negligible random elements and for which, therefore, the future course of the system is completely determined by its position, velocities, and so on. An example of a deterministic model is given by force = mass \times acceleration.

Discrepancy function: Used primarily in structural equation models, this function measures the differences between the observed variance-covariance matrix and the variance-covariance matrix implied by a postulated structural equation model. All else being equal, a small value of the discrepancy function is preferred to a larger value, indicating greater lack of fit to the data (see also structural equations and goodness-of-fit statistics).

Discrete/continuous models: A class of models that has a system of both discrete and continuous interrelated dependent variables.

Discrete variable: A variable (or outcome) that is measured on the nominal or ordinal scale. Examples of discrete variables include mode of travel (with discrete possibilities being car, bus, train) and accident severity (with discrete possibilities being property damage only, injury, fatality).

Dispersion: The degree of scatter or concentration of observations around the center or middle. Dispersion is usually measured as a deviation about some central value such as the mean, standard or absolute deviation, or by an order statistic such as deciles, quintiles, and quartiles.

Distribution: The set of frequencies or probabilities assigned to various outcomes of a particular event or trial. Densities (derived from continuous data) and distributions (derived from discrete data) are often used interchangeably.

Distribution function: See cumulative distribution function.

Disturbances: Also referred to as errors, residuals, and model noise, these are random variables added to a statistical model equation to account for unobserved effects. The choice of their statistical distribution typically plays a key role in parametric model derivation, estimation, and inference.

Double sampling: The process by which information from a random sample is used to direct the gathering of additional data often targeted at oversampling underrepresented components of the population.

Dummy variables: See indicator variables.

E

Econometrics: The development and application of statistical and/or mathematical principles and techniques for solving economic problems.

Efficiency: A statistical estimator or estimate is said to be efficient if it has small variance. In most cases, a statistical estimate is preferred if it is more efficient than alternative estimates. It can be shown that the Cramer–Rao bound represents the best possible efficiency (lowest variance) for an unbiased estimator. That is, if an unbiased estimator is shown to be equivalent to the Cramer–Rao bound, then there are no other unbiased estimators that are more efficient. It is possible in some cases to find a more efficient estimate of a population parameter that is biased.

Elasticity: Commonly used to determine the relative importance of a variable in terms of its influence on a dependent variable. It is generally interpreted as the percent change in the dependent variable induced by a 1% change in the independent variable.

Empirical: Derived from experimentation or observation rather than underlying theory.

Endogenous variable: In a statistical model, a variable whose value is determined by influences or variables within the statistical model. An assumption of statistical modeling is that explanatory variables are exogenous. When explanatory variables are endogenous, parameter estimation problems arise (see also exogenous variable, instrumental variables, simultaneous equations, and structural equations).

Enriched sampling: Refers to merging a random sample with a sample from a targeted population. For example, a random sample of commuters' mode choices may be merged with a sample of commuters observed taking the bus.

Error: In statistics the difference between an observed value and its "expected" value as predicted or explained by a model. In addition, errors occur in data collection, sometimes resulting in outlying observations.

Finally, Type I and Type II errors refer to specific interpretive errors made when analyzing the results of hypothesis tests.

Error mean square: In analysis of variance and regression, the residual or error sum of squares divided by the degrees of freedom. Often called the mean square error (MSE), the error mean square provides an estimate of the residual or error variance of the population from which the sample was drawn.

Error of observation: An error arising from imperfections in the method of observing a quantity, whether due to instrumental or to human factors.

Error rate: In hypothesis testing, the unconditional probability of making an error; that is, erroneously accepting or rejecting a statistical hypothesis. Note that the probabilities of Type I and Type II errors, α and β , are conditional probabilities. The first is subject to the condition that the null hypothesis is true, and the second is subject to the condition that the null hypothesis is false.

Error term: See disturbance.

Error variance: The variance of the random or unexplainable component of a model; the term is used mainly in the presence of other sources of variation, as, for example, in regression analysis or in analysis of variance.

Exogenous variable: In a statistical model, a term referring to a variable whose value is determined by influences outside of the statistical model. An assumption of statistical modeling is that explanatory variables are exogenous. When explanatory variables are endogenous, parameter estimation problems arise (see also endogenous variable, instrumental variables, simultaneous equations, and structural equations).

Expectation: The expected or mean value of a random variable or function of that variable such as the mean or variance.

Experiment: A set of measurements carried out under specific and controlled conditions to discover, verify, or illustrate a theory, hypothesis, or relationship. Experiment is the cornerstone of statistical theory and is the only accepted method for suggesting causal relationships between variables. Experimental hypotheses are not proved using statistics; however, they are disproved. Elements of an experiment generally include a control group, randomization, and repeat observations.

Experimental data: Data obtained by conducting experiments under controlled conditions. Quasi-experimental data are obtained when some factors are controlled as in an experiment, but some factors are not. Observational data are obtained when no exogenous factors other than the treatment are controlled or manipulated. Analysis and modeling based on quasi-experimental and observation data are subject to illusory correlation and confounding of variables.

Experimental design: Plan or blueprint for conducting an experiment.

Experimental error: Any error in an experiment whether due to stochastic variation or bias (not including mistakes in design or avoidable imperfections in technique).

Exponential: A variable raised to a power of x . The function $F(x) = a^x$ is an exponential function.

Exponential distribution: A continuous distribution that is typically used to model life cycles, growth, survival, hazard rate, or decay of materials or events. Its probability density function is given by

$$f(x) = (\lambda) \text{EXP}(-\lambda x)$$

where λ is single parameter of the exponential distribution.

Exponential smoothing: Time series regression in which recent observations are given more weight by way of exponentially decaying regression parameters.

F

F distribution: A distribution of fundamental importance in analysis of variance and regression. It arises naturally as a ratio of estimated variances and sums of squares. The F density function is given as

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{n/2} m^{m/2}}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} \frac{x^{n/2-1}}{(m+nx)^{(n+m)/2}}$$

where $\Gamma(z)$ is the gamma function and n and m are the numerator and denominator degrees of freedom for the F distribution, respectively (see gamma distribution for gamma function).

F ratio: The ratio of two independent unbiased estimates of variance of a normal distribution, which has widespread application in the analysis of variance.

F test: A computed statistic that under an appropriate null hypothesis has an approximate F distribution. It is used routinely to test full and restricted statistical models.

Factor analysis: An analytical technique used to reduce the number of P variables to a smaller set of parsimonious $K < P$ variables, so that the covariance among many variables is described in terms of a few unobserved factors. An important difference between principal components and

factor analysis is that factor analysis is based on a specific statistical model, whereas principal components analysis is not (see also principal components analysis and structural equations).

Factorial experiment: An experiment designed to examine the effect of one or more factors, with each factor applied to produce orthogonal or independent effects on the response.

Fixed effects model: A model appropriate when unobserved heterogeneity of sampled units is thought to represent the effects of a particular sampled unit (see random effects model).

Full factorial experiment: An experiment investigating all the possible treatment combinations that may be formed from the factors under investigation.

Full information maximum likelihood (FIML): A likelihood function written such that all possible information and restrictions related to model estimation are included. For example, in simultaneous equations estimation, the FIML approach accounts for overidentification of model parameters and cross-equation correlation of disturbances.

Frequency: The number of occurrences of a given type of event.

Frequentist statistical philosophy: Also called classical statistical philosophy, this philosophy represents the majority of currently applied statistical techniques in practice, although Bayesian philosophy is gaining increased acceptance. In simple terms, classical statisticians assert that probabilities are obtained through long-run repeated observations of events. Frequentist statistical philosophy results in hypothesis tests that provide an estimate of the probability of observing the sample data conditional on a true null hypothesis, whereas Bayesian statistical philosophy results in an estimate of the probability of the null hypothesis being true conditional on the observed sample data (see also Bayesian statistical philosophy).

G

Gamma distribution: A distribution that includes as special cases the chi-square distribution and the exponential distribution. It has many important applications; in Bayesian inference, for example, it is sometimes used as the a priori distribution for the parameter (mean) of a Poisson distribution. It is also used to describe unobserved heterogeneity in some models. The gamma density is given by

$$f(x) = \frac{x^{\alpha-1} \text{EXP}\left(-\frac{x}{\beta}\right)}{[\Gamma(\alpha)\beta^\alpha]} \quad x > 0$$

where α is the shape parameter, β is a scale parameter, and $\Gamma(z)$ is the gamma function, which is given by the formula

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

When the location parameter is equal to 0 and the scale parameter is equal to 1, the gamma distribution reduces to the standard gamma distribution, with mean γ and standard deviation $\sqrt{\gamma}$.

Gantt chart: A bar chart showing actual performance or output expressed as a percentage of a quota or planned performance per unit of time.

Gaussian distribution: Another name for the normal distribution.

Generalized extreme value (GEV): A generalized distribution used to derive a family of models that includes the multinomial and nested logit models.

Generalized least squares (GLS): An estimation procedure that generalizes least squares estimation by accounting for possible heterogeneity and correlation of disturbance terms.

Goodness of-fit (GOF) statistics: A class of statistics used to assess the fit of a model to observed data. There are numerous GOF measures, including the coefficient of determination R^2 , the F test, the chi-square test for frequency data, and numerous other measures. GOF statistics may be used to measure the fit of a statistical model to estimation data, or data used for validation. GOF measures are not statistical tests, like F -ratio and likelihood ratio tests to compare models.

Gumbel distribution: The density for a continuous variable that approximates many decay, survival, and growth processes, as well provides appealing analytical qualities in analytical problems. The Gumbel distribution is given by

$$F(\varepsilon) = EXP\left[EXP(-\eta[\varepsilon - \omega])\right]$$

where η is a positive scale parameter, ω is a location parameter (mode), and $\omega + 0.5772/\eta$ is the mean.

H

Hazard function: Function used to describe the conditional probability that an event will occur between time t and $t + dt$, given that the

event has not occurred up to time t . It is written as $h(t) = f(t)/[1 - F(t)]$, where $F(t)$ is the cumulative distribution function and $f(t)$ is the density function. In words, the hazard function $h(t)$ gives the rate at which event durations are ending at time t (such as the duration in an accident-free state that would end with the occurrence of an accident), given that the event duration has not ended up to time t .

Heterogeneity: A term used to describe samples or individuals from different populations, which differ with respect to the phenomenon of interest. If the populations are not identical, they are said to be heterogeneous, and by extension the sample data are also said to be heterogeneous (see also unobserved heterogeneity).

Heteroscedasticity: In regression analysis, the property that the conditional distributions of the response variable Y for fixed values of the independent variables do not all have constant variance. Nonconstant variance in a regression model results in inflated estimates of model mean square error. Standard remedies include transformations of the response, and/or employing a generalized linear model.

Histogram: A univariate frequency diagram in which rectangles proportional in area to the class frequencies are erected on sections of the horizontal axis, the width of each section representing the corresponding class interval of the variate.

Holt-Winters smoothing: A seasonal time series modeling approach in which the original series is decomposed into its level, trend, and seasonal components with each of the components modeled with exponentially smoothed regression.

Homogeneity: Term used in statistics to describe samples or individuals from populations that are similar with respect to the phenomenon of interest. If the populations are similar, they are said to be homogeneous, and by extension the sample data are also said to be homogeneous.

Homoscedasticity: In regression analysis, the property that the conditional distributions of Y for fixed values of the independent variable all have the same variance (see also heteroscedasticity).

Hypothesis: In statistics, a statement concerning the value of parameters or form of a probability distribution for a designated population or populations. More generally, a statistical hypothesis is a formal statement about the underlying mechanisms that have generated some observed data.

Hypothesis testing: Term used to refer to testing whether observed data support a stated position or hypothesis. Support of a research hypothesis suggests that the data would have been unlikely if the hypothesis were indeed false (see also Type I and Type II errors).

I

Identification problem: A problem encountered in the estimation of statistical models (in particular simultaneous and structural equations models) that creates difficulty in uncovering underlying equation parameters from a reduced form model.

Illusory correlation: Also called spurious correlation, an omitted variable problem, similar to confounding. Illusory correlation is used to describe the situation where Y and X_1 are correlated, yet the relation is illusory, because X_1 is actually correlated with X_2 , which is the true "cause" of the changes in Y .

Independence of irrelevant alternatives (IIA): A property of multinomial logit models that is an outgrowth of the model derivation, which assumes disturbance terms are independent. If the outcome disturbances are correlated, the model will give erroneous outcome probabilities when forecasting.

Independent events: In probability theory, two events are said to be statistically independent if, and only if, the probability that they will both occur equals the product of the probabilities that each one, individually, will occur. Independent events are not correlated, whereas dependent events are (see also dependence and correlation).

Independent variables: Two variables are said to be statistically independent if, and only if, the probability that they will both occur equals the product of the probabilities that each one, individually, will occur. Independent events are not correlated, whereas dependent events are (see also dependence and correlation).

Indicator variables: Variables used to quantify the effect of a qualitative or discrete variable in a statistical model. Also called dummy variables, indicator variables typically take on values of zero or one. Indicator variables are coded from ordinal or nominal variables. For a nominal variable with n levels, $n - 1$ indicator variables are coded for use in a statistical model. For example, the nominal variable vehicle type: truck, van, or auto, the analyst would code $X_1 = 1$ for truck, 0 otherwise, and $X_2 = 1$ for van, 0 otherwise. When both X_1 and X_2 are coded as zero, the respondent was an auto.

Indirect least squares: A method used in simultaneous equations estimation that applies ordinary least squares to the reduced form model.

Indirect utility: Utility that has prices and incomes as arguments. It is based on optimal consumption and is tied to underlying demand equations by Roy's identity.

Induction: The type of logic or thinking process, where inferences are drawn about an entire class or group based on observations on a few of its members (see also deduction).

Inference: The process of inferring qualities, characteristics, or relationships observed in the sample onto the population. Statistical inference relies on both deductive and inductive logic systems.

Information criteria: Model performance measures that balance decrease in model error with increase in model complexity. In general, information criteria are based on the Gaussian likelihood of the model estimates with a penalty for the number of model parameters.

Innovation series: The zero-mean uncorrelated stochastic component of a time series of observations that remains after all deterministic and correlated elements have been appropriately modeled. The innovations are also referred to as the series noise.

Instrumental variables: A common approach used in model estimation to handle endogenous variables—those variables that are correlated with the disturbance term causing a violation in model assumptions. The ideal instrument is a variable that is highly correlated with the endogenous variable it replaces but is not correlated with the disturbance term.

Interaction: Two variables X_1 and X_2 are said to interact if the value of X_1 influences the value of X_2 positively or negatively. An interaction is a synergy between two or more variables and reflects the fact that their combined effect on a response not only depends on the level of the individual variables, but their combined levels as well.

Interval estimate: The estimation of a population parameter by specifying a range of values bounded by an upper and a lower limit, within which the true value is asserted to lie.

Interval scale: A measurement scale that has equal differences between pairs of points anywhere on the scale, but the zero point is arbitrary; thus, ratio comparisons cannot be made. An example interval scale is the temperature in degrees Celsius. Each interval on the scale represents a single degree; however, 50°C is not twice as hot as 25°C (see also nominal, ordinal, and ratio scales).

J

Joint probability: The joint density function of two random variables or bivariate density.

K

Kendall's coefficient of rank correlation: Denoted as τ_{ij} , where i and j refer to two variables, a coefficient that reflects the degree of linear

association between two ordinal variables, and is bounded between +1 for perfect positive correlation and -1 for perfect negative correlation. The formula for τ_{ij} is given by

$$\tau_{ij} = \frac{S}{\frac{1}{2}n(n-1)}$$

where S is the sum of scores (see nonparametric statistic reference for computing scores) and n is sample size.

L

Latent variable: A variable that is not directly observable. Examples include intelligence, education, and satisfaction. Typically, latent variables or constructs are measured inexactly with many variables. For example, intelligence may be partially captured with IQ score, GPA (grade point average) from college, and number of books read per year.

Least squares estimation: Also called ordinary least squares, a technique of estimating statistical parameters from sample data whereby parameters are determined by minimizing the squared differences between model predictions and observed values of the response. The method may be regarded as possessing an empirical justification in that the process of minimization gives an optimum fit of observation to theoretical models; in restricted cases, such as normal theory linear models, estimated parameters have optimum statistical properties of unbiasedness and efficiency.

Level of significance: The probability of rejecting a null hypothesis, when it is in fact true. It is also known as α , or the probability of committing a Type I error.

Likelihood function: The probability or probability density of obtaining a given set of sample values, from a certain population, when this probability or probability density is regarded as a function of the parameter of the population and not as a function of the sample data (see also maximum likelihood method).

Likelihood ratio test: A test frequently used to test for restrictions in models estimated by maximum likelihood. The test is used for a wide variety of reasons, including the testing of the significance of individual parameters and the overall significance of the model. The likelihood ratio test statistic is $-2[LL(\boldsymbol{\beta}_R) - LL(\boldsymbol{\beta}_L)]$, where $LL(\boldsymbol{\beta}_R)$ is the log-likelihood at convergence of the "restricted" model and $LL(\boldsymbol{\beta}_L)$ is

the log-likelihood at convergence of the “unrestricted” model. This statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of parameters in the restricted and unrestricted models, that is, the number of parameters in the β_R and the β_u parameter vectors.

Limited information maximum likelihood (LIML): A likelihood function written such that some possible information on the model is excluded. For example, in simultaneous equations estimation, the LIML approach does not account for possible cross-equation correlation of disturbances.

Linear correlation: A somewhat ambiguous expression used to denote either Pearson’s product moment correlation, in cases where the corresponding variables are continuous, or a correlation coefficient on ordinal data such as Kendall’s rank correlation coefficient. There are other linear correlation coefficients in addition to the two listed here.

Linear model: A mathematical model in which the equations relating the random variables and parameters are linear in parameters. Although the functional form of the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2$ includes the nonlinear term $\beta_2 X_1^2$, the model itself is linear, because the statistical parameters are linearly related to the random variables.

Logistic distribution: A distribution used for various growth models and in logistic regression models. The general logistic distribution is given by

$$f(x) = \frac{\left(\frac{1}{\beta}\right) EXP[(x - \alpha)/\beta]}{\{1 + EXP[(x - \alpha)/\beta]\}^2}$$

where α is the location parameter and β is the scale parameter. The mean of the general logistic distribution is α , and the variance is $\beta^2\pi^2/3$.

Logit model: A discrete outcome model derived, in multinomial form (three or more outcomes), by assuming the disturbances are Weibull distributed (Gumbel extreme value type I).

Log-logistic distribution: A distribution used in a number of applications, including duration modeling. In duration modeling the log-logistic distribution allows for nonmonotonic hazard functions and is often used as an approximation of the more computationally cumbersome lognormal distribution. The log-logistic with parameters $\lambda > 0$ and $P > 0$ has the density function

$$f(x) = \frac{\lambda P(\lambda x)^{P-1}}{\left[1 + (\lambda x)^P\right]^2}$$

Lognormal distribution: A distribution where random variable x has the log-normal distribution if $LN(x)$ is normally distributed with mean μ and variance σ^2 . The density for the lognormal distribution is given by

$$f(x) = EXP\{-[LN(x) - \mu]^2 / (2\sigma^2)\} / [x(2\pi)^{1/2} \sigma]; \quad x > 0$$

where μ and σ are location and scale parameters of the lognormal distribution. The mean and variance of the lognormal distribution are $EXP(\mu + \sigma^2/2)$ and $EXP[2(\mu + \sigma^2)] - EXP[2\mu + \sigma^2]$, respectively.

M

Marginal effect: An economic term used to measure the effect that a unit change in an independent variable has on the response variable of interest. Marginal effects are indicators of how influential or important a variable is in a particular data-generating process (see also significance).

Marginal rate of substitution (MRS): An economic term used to measure the trade off consumers are willing to make between attributes of a good. For example, if one were to use a logit model in the analysis of discrete outcome data, the MRS is determined from the ratio of parameter estimates.

Maximum likelihood method: A method of parameter estimation in which a parameter is estimated by the value of the parameter that maximizes the likelihood function. In other words, the maximum likelihood estimator is the value of θ that maximizes the probability of the observed sample. The method can also be used for the simultaneous estimation of several parameters, such as regression parameters. Estimates obtained using this method are called maximum likelihood estimates.

Mean: That value of a variate such that the sum of deviations from it is zero; thus, it is the sum of a set of values divided by their number.

Mean square error: For unbiased estimators, an estimate of the population variance usually denoted as MSE. For biased estimators, the mean squared deviation of an estimator from the true value is equal to the variance plus the squared bias. The square root of the mean square error is referred to as the root mean square error.

Median: The middle-most number in an ordered series of numbers. It is a measure of central tendency and is often a more robust measure of central tendency; that is, the median is less sensitive to outliers than is the sample mean.

Mixed multinomial logit models: A class of multinomial logit models that has a mixing distribution introduced to account for random variations in parameters across the population.

Mode: The most common or most probable value observed in a set of observations or sample.

Model: A formal expression of a theory or causal or associative relationship between variables, which is regarded by the analyst as having generated the observed data. A statistical model is always a simplified expression of a more complex process; thus, the analyst should anticipate some degree of approximation a priori. A statistical model that can explain the greatest amount of underlying complexity with the simplest model form is preferred to a more complex model.

Moving average (MA) processes: Stationary time series that are characterized by a linear relationship between observations and past innovations. The order of the process q defines the number of past innovations on which the current observation depends.

Multicollinearity: A term that describes the state when two variables are correlated with each other. In statistical models, multicollinearity causes problems with the efficiency of parameter estimates. It also raises some philosophical issues because it becomes difficult to determine which variables (both, either, or neither) are causal and which are the result of illusory correlation.

Multinomial logit model (MNL): A multinomial (three or more outcome) discrete model derived by assuming the disturbances are Weibull distributed (Gumbel extreme value type I).

Multiple linear regression: A linear regression involving two or more independent variables. Simple linear regression, which is merely used to illustrate the basic properties of regression models, contains one explanatory variable and is rarely if ever used in practice.

Mutually exclusive events: In probability theory, two events are said to be mutually exclusive if and only if they are represented by disjoint subsets of the sample space, namely, by subsets that have no elements or events in common. By definition, the probability of mutually exclusive events A and B occurring is zero.

N

Negative binomial distribution: A discrete distribution characterized by the count of observations that remain unchanged in a binomial process. It also arises as a combination of gamma distributed heterogeneity of Poisson means. The probability distribution of the negative

binomial is given as

$$P(Y_k = n) = C(n-1, k-1)p^k(1-p)^{n-k}; \quad \text{for } n = k, k+1, k+2, \dots,$$

where k is the number of successes, p is the probability of success, and $C(a,b)$ is the number of combinations of b objects taken from a objects, defined as

$$C(a,b) = \frac{a!}{(a-b)!b!}$$

The mean and variance of the negative binomial distribution are k/p and $k(1-p)/p^2$. The negative binomial probability distribution provides the probability of observing k successes in n Bernoulli trials (see also binomial and Bernoulli).

Nested logit model: A modification of the multinomial logit model, a model that is derived from a generalized extreme value distribution and that can eliminate common logit model problems related to the independence of irrelevant alternatives property. The approach of a nested logit model is to group alternative outcomes suspected of sharing unobserved effects into nests of outcome possibilities.

Noise: A convenient term for a series of random disturbances or deviation from the actual distribution. Statistical noise is a synonym for error term, disturbance, or random fluctuation.

Nominal scale: A variable measured on a nominal scale is the same as a categorical or discrete variable. The nominal scale lacks order and does not possess even intervals between levels of the variable. An example of a nominal scale variable is vehicle type, where levels of response include truck, van, and auto (see also ordinal, interval, and ratio scales of measurement).

Nonlinear relation: A relation where a scatter plot between two variables X_1 and X_2 will not produce a straightline trend. In many cases a linear trend is observed between two variables by transforming the scale of one or both variables. For example, a scatter plot of $LN(X_1)$ and X_2 might produce a linear trend. In this case, the variables are said to be nonlinearly related in their original scales, but linear in transformed scale of X_1 .

Nonparametric statistics: Statistical methods that do not rely on statistical distributions with estimable parameters. In general, nonparametric statistical methods are best suited for ordinal scale variables or, better, for dealing with small samples, and for cases when parametric methods assumptions are suspect (see also parametric statistics).

Nonrandom sample: A sample selected by a nonrandom method. For example, a scheme whereby units are self-selected would yield a non-random sample, where units that prefer to participate do so. Some aspects of nonrandom sampling can be overcome, however.

Normal distribution: A continuous distribution that was first studied in connection with errors of measurement and, thus, is referred to as the “normal curve of errors.” The normal distribution forms the cornerstone of a substantial portion of statistical theory. Also called the Gaussian distribution, the normal distribution has the two parameters μ and σ , when $\mu = 0$ and $\sigma = 1$ the normal distribution is transformed into the standard normal distribution. The normal distribution is characterized by its symmetric shape and bell-shaped appearance. The normal distribution probability density is given by

$$\Phi(z) = \frac{\text{EXP}[-(x-\mu)^2/2\sigma^2]}{(2\pi)^{1/2}}$$

Null hypothesis: In general, a term relating to the particular research hypothesis tested, as distinct from the alternative hypothesis, which is accepted if the research hypothesis is rejected. Contrary to intuition, the null hypothesis is often a research hypothesis that the analyst would prefer to reject in favor of the alternative hypothesis, but this is not always the case. Erroneous rejection of the null hypothesis is known as a Type I error, whereas erroneous acceptance of the null hypothesis is known as a Type II error.

O

Observational data: Nonexperimental data. Because there is no control of potential confounding variables in a study based on observational data, the support for conclusions based on observational data must be strongly supported by logic, underlying material explanations, identification of potential omitted variables and their expected biases, and caveats identifying the limitations of the study.

Omitted variable bias: Variables affecting the dependent variable that are omitted from a statistical model are problematic. Irrelevant omitted variables cause no bias in parameter estimates. Important variables that are uncorrelated with included variables will also cause no bias in parameter estimates, but the estimate of σ^2 is biased high. Omitted variables that are correlated with an included variable X_1 will produce biased parameter estimates. The sign of the bias depends on the product of the covariance of the omitted variable and X_1 and β_1 , the biased parameter. For example, if the covariance is negative and β_1 is negative, then the parameter is biased positive. In addition, σ^2 is also biased.

One-tailed test: Also known as a one-sided test, a test of a statistical hypothesis in which the region of rejection consists of either the right-hand tail

or the left-hand tail of the sampling distribution of the test statistic. Philosophically, a one-tailed test represents the analyst's a priori belief that a certain population parameter is either negative or positive.

Ordered probability models: A general class of models used to model discrete outcomes that are ordered (e.g., from low to high). Ordered probit and ordered logit are the most common of these models. They are based on the assumption of normal and Weibull distributed disturbances, respectively.

Ordinal scale: A scale of measurement occurs when a variable can take on ordered values, but there is not an even interval between levels of the variable. Examples of ordinal variables include measuring the desirability of different models of vehicles (Mercedes SLK, Ford Mustang, Toyota Camry), where the response is highly desirable, desirable, and least desirable (see also nominal, interval, and ratio scales of measurement).

Ordinary differencing: Creating a transformed series by subtracting the immediately adjacent observations.

Ordinary least squares (OLS): See least squares estimation.

Orthogonal: A condition of two variables where the linear correlation coefficient between them is zero. In observational data, the correlation between two variables is almost always nonzero. Orthogonality is an extremely desirable property among independent variables in statistical models, and typically arises by careful design of data collection through experimentation or cleverly crafted stated preference surveys.

Outliers: Observations that are identified as such because they "appear" to lie outside a large number of apparently similar observations or experimental units according to a specified model. In some cases, outliers are traced to errors in data collecting, recording, or calculation and are corrected or appropriately discarded. However, outliers sometimes arise without a plausible explanation. In these cases, it is usually the analyst's omission of an important variable that differentiates the outlier from the remaining, otherwise similar observations, or a misspecification of the statistical model that fails to capture the correct underlying relationships. Outliers of this latter kind should not be discarded from the "other" data unless they are modeled separately, and their exclusion justified.

P

p-Value: The smallest level of significance α that leads to rejection of the null hypothesis.

Panel data: Data that are obtained from experimental units across various points in time, but not necessarily continuously over time like time series data (see also before–after data, time series data, and cross-sectional data).

Parameter: Sometimes used interchangeably with coefficient, an unknown quantity that varies over a certain set of inputs. In statistical modeling, it usually occurs in expressions defining frequency or probability distributions in terms of their relevant parameters (such as mean and variance of normal distribution), or in statistical models describing the estimated effect of a variable or variables on a response. Of utmost importance is the notion that statistical parameters are merely estimates, computed from the sample data, which are meant to provide insight regarding what the true population parameter value is, although the true population parameter always remains unknown to the analyst.

Parametric statistics: Methods of statistical analysis that utilize parameterized statistical distributions. These methods typically require that a host of assumptions be satisfied. For example, a standard t test for comparing population means assumes that distributions are approximately normally distributed and have equal variances (see also nonparametric statistics).

Pearson's product moment correlation coefficient: Denoted as r_{ij} , where i and j refer to two variables, a coefficient that reflects the degree of linear association between two continuous (ratio or interval scale) variables and that is bounded between +1 for perfect positive correlation and -1 for perfect negative correlation. The formula for r_{ij} is given by

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

where s_{ij} is the covariance between variables i and j , and s_i and s_j are the standard deviations of variables i and j , respectively.

Point estimate: A single estimated value of a parameter, or an estimate without a measure of sampling variability.

Poisson distribution: A discrete distribution that is often referred to as the distribution of rare events. It is typically used to describe the probability of occurrence of an event over time, space, or length. In general, the Poisson distribution is appropriate when the following conditions hold: the probability of “success” in any given trial is relatively small; the number of trials is large; and the trials are independent. The probability distribution function for the Poisson distribution is given as

$$P(X = x) = P(x; \lambda) = \frac{\lambda^x \text{EXP}(-\lambda)}{x!}, \text{ for } x = 1, 2, 3, \dots, \infty$$

where x is the number of occurrences per interval, and λ is the mean number of occurrences per interval.

Population: In statistical usage, a term that is applied to any finite or infinite collection of individuals. It is important to distinguish between the population, for which statistical parameters are fixed and unknown at any given instant in time, and the sample of the population, from which estimates of the population parameters are computed. Population statistics are generally unknown because the analyst can rarely afford to measure all members of a population, so a random sample is drawn.

Post hoc theorizing: Theorizing that is likely to occur when the analyst attempts to explain analysis results after the fact (post hoc). In this less than ideal approach to scientific discovery, the analyst develops hypotheses to explain the data, instead of the reverse (collecting data to nullify a well-articulated hypothesis). The number of post hoc theories that are developed to “fit” the data is limited only by the imagination of a group of scientists. With an abundance of competing hypotheses, and little forethought as to which hypothesis is afforded more credence, there is little in the way of statistical justification to prefer one hypothesis to another. Especially when data are observational, there is little evidence to eliminate the prospect of illusory correlation.

Power: The probability that a statistical test of some hypothesis rejects the alternative hypothesis when the alternative is false. The power is greatest when the probability of a Type II error is least. Power is $1 - \beta$, whereas level of confidence is $1 - \alpha$.

Precision: Degree of scatter to which repeated estimates agree with the true state of nature. Typically, precision is a measure of the variation of predictions around the true and often unknown average (see also accuracy, bias, efficiency, and consistency).

Prediction interval: A calculated range of values known to contain some future observation over the average of repeated trials with specific certainty (probability). The correct interpretation of a prediction interval is as follows. If the analyst were to repeatedly draw samples at the same levels of the independent variables and compute the test statistic (mean, regression slope, etc.), then a future observation will lie in the $(1 - \alpha)\%$ prediction interval α times out of 100. The prediction interval differs from the confidence interval in that the confidence interval provides certainty bounds around a mean, whereas the prediction interval provides certainty bounds around an observation.

Principal components analysis: An analytical tool used to explain the variance–covariance structure of a relatively large multivariate data set using a few linear combinations of the originally measured variables. It is used to reveal structure in data and enable the identification

of underlying dimensions in the data (see also factor analysis and structural equations).

Probability density functions: Synonymous with probability distributions (sometimes referred to simply as density functions); knowing the probability that a random variable takes on certain values, judgments are made regarding how likely or unlikely were the observed values. In general, observing an unlikely outcome tends to support the notion that chance was not acting alone. By posing alternative hypotheses to explain the generation of data, an analyst can conduct hypothesis tests to determine which of the two competing hypotheses best supports the observed data.

Probit model: A discrete outcome model derived from assuming that the disturbances are multivariate normally distributed.

Proportional hazards model: An approach used in hazard-based analysis of duration data that assumes that the covariates act multiplicatively on some underlying (or base) hazard function.

R

Random effects model: When unobserved heterogeneity of sampled units is thought to represent a representative random sample of effects from a larger population of interest, then these effects are thought to be random. In random effects models, interest is not centered on the effects of sampled units.

Random error: A deviation of an observed value from a true value, which occurs as though chosen at random from a probability distribution of such errors.

Randomization: Process used in the design of experiments. When certain factors cannot be controlled and omitted variable bias has the potential to occur, randomization is used to assign subjects to treatment and control groups randomly, such that any systematically omitted variable bias is distributed evenly among the two groups. Randomization should not be confused with random sampling, which serves to provide a representative sample.

Random sampling: A sample strategy whereby population members have equal probability of being recruited into the sample. Often called simple random sampling, it provides the greatest assurance that the sample is representative of the population of interest.

Random selection: Synonymous with random sampling, a sample selected from a finite population is said to be random if every possible sample has equal probability of selection. This applies to sampling without replacement. A random sample with replacement is still

considered random as long as the population is sufficiently large that the replaced experimental unit has small probability of being recruited into the sample again.

Random variable: A variable whose exact value is not known prior to measurement.

Range: The largest minus the smallest of a set of variate values.

Ratio scale: A variable measured on a ratio scale has order, possesses even intervals between levels of the variable, and has an absolute zero. An example of a ratio scale variable is height, where levels of response include 0.0 and 2000.0 inches (see also discrete, continuous, nominal, ordinal, and interval).

Raw data: Data that have not been subjected to any sort of mathematical manipulation or statistical treatment such as grouping, coding, censoring, or transformation.

Reduced form: An equation derived from the combination of two or more equations. Reduced form models are used in a number of statistical applications, including simultaneous equations modeling and the analysis of interrelated discrete/continuous data.

Regression: A statistical method for investigating the interdependence of variables.

Repeatability: Degree of agreement between successive runs of an experiment.

Replication: The execution of an experiment or survey more than once to increase the precision and to obtain a closer estimation of the sampling error.

Representative sample: A sample that is representative of a population (it is a moot point whether the sample is chosen at random or selected to be "typical" of certain characteristics; therefore, it is better to use the term for samples, which turn out to be representative, however chosen, rather than apply it to a sample chosen with the objective of being representative). Random samples are, by definition, representative samples.

Reproducibility: An experiment or survey is said to be reproducible if, on repetition or replication under similar conditions, it gives the same results.

Residual: The difference between the observed value and the fitted value in a statistical model. Residual is synonymous with error, disturbance, and statistical noise.

Residual method: In time series analysis, a classical method of estimating cyclical components by first eliminating the trend, seasonal variations, and irregular variations, thus leaving the cyclical relatives as residuals.

Robustness: A method of statistical inference is said to be robust if it remains relatively unaffected when all of its underlying assumptions are not met.

Roy's identity: An identity relating an indirect utility function with its underlying demand function.

S

Sample: A part or subset of a population, which is obtained through a recruitment or selection process, usually with the objective of obtaining an improved understanding of the parent population.

Sample size: The number of sampling units which have been sampled from the larger population of ultimate interest.

Sampling error: Also called sampling variability; that part of the difference between an unknown population value and its statistical estimate, which arises from chance fluctuations in data values as a natural consequence of random sampling.

Scatter diagram: Also known as a scatter plot, a graph showing two corresponding scores for an individual as a point; the result is a swarm of points.

Seasonal cycle length: The length of the characteristic recurrent pattern in seasonal time series, given in terms of number of discrete observation intervals.

Seasonal differencing: Creating a transformed series by subtracting observations that are separated in time by one seasonal cycle.

Seasonality: The time series characteristic defined by a recurrent pattern of constant length in terms of discrete observation intervals.

Seemingly unrelated equations: A system of equations that are not directly interrelated, but have correlated disturbances because of shared unobserved effects across disturbance terms.

Selectivity bias: The bias that results from using a sample that is populated by an outcome of some nonrandom process. For example, if a model were developed to study the accident rates of sports car owners, great care would have to be taken in interpreting the estimation results. This is because sports car owners are likely to be a self-selected sample of faster drivers willing to take more risks. Thus in interpreting the results, it would be difficult to untangle the role that having a faster, high-performance car has on accident rates from the fact that riskier drivers are attracted to such cars.

Self-selection in survey collection: Phenomenon that can occur when survey respondents are allowed to deny participation in a survey. The belief is that respondents who are opposed or who are apathetic about the objectives of the survey are less likely to participate, and their removal from the sample will bias the results of the survey. Self-selection can also occur because respondents who are either strongly opposed or strongly supportive of a survey's objectives respond to the survey.

A classic example is television news polls that solicit call-in responses from listeners—the results from which are practically useless for learning how the population at large feels about an issue.

Serial correlation: The temporal association between observations of a series of observations ordered across time. Also referred to as autocorrelation.

Significance: An effect is statistically significant if the value of the statistic used to test it lies outside acceptable limits, that is, if the hypothesis that the effect is not present is rejected. Statistical significance does not imply practical significance, which is determined by also evaluating the marginal effects.

Simultaneous equations: A system of interrelated equations with dependent variables determined simultaneously.

Skewness: The lack of symmetry in a probability distribution. In a skewed distribution the mean and median are not coincident.

Smoothing: The process of removing fluctuations in an ordered series so that the result is “smooth” in the sense that the first differences are regular and higher-order differences are small. Although smoothing is carried out using freehand methods, it is usual to make use of moving averages or the fitting of curves by least squares procedures. The philosophical grounds for smoothing stem from the notion that measurements are made with error, such that artificial “bumps” are observed in the data, whereas the data really should represent a smooth or continuous process. When these “lumpy” data are smoothed appropriately, the data are thought to better reflect the true process that generated the data. An example is the speed–time trace of a vehicle, where speed is measured in integer miles per hour. Accelerations of the vehicle computed from differences in successive speeds are overestimated due to the lumpy nature of measuring speed. Thus an appropriate smoothing process on the speed data results in data that more closely resembles the underlying data-generating process. Of course, the technical difficulty with smoothing lies in selecting the appropriate smoothing process, as the real data are never typically observed.

Spurious correlation: See illusory correlation.

Standard deviation: The sample standard deviation s_x is the square root of the sample variance and is given by the formula

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

where n is the sample size and \bar{x} is the sample mean. The sample standard deviation shown here is a biased estimator, even though the sample variance is unbiased, and the bias becomes larger as the sample size becomes smaller.

Standard error: The square root of the variance of the sampling distribution of a statistic.

Standard error of estimate: The standard deviation of the observed values about a regression line.

Standard error of the mean: Standard deviation of the means of several samples drawn at random from a large population.

Standard normal transformation: Fortunately, the analyst can transform any normal distributed variable into a standard normal distributed variable by making use of a simple transformation. Given a normally distributed variable X , Z is defined such that

$$Z_i = \frac{x_i - \mu_x}{\sigma_x}$$

The new variable Z is normally distributed with $\mu = 0$ and $\sigma = 1$ and is a standard normal variable.

Standard scores: Scores expressed in terms of standard deviations from the mean. Standard scores are obtained using the standard normal transformation (see also normal distribution and Z value).

State dependence: Often used to justify the inclusion of lagged variables in a model that represent previous conditions. For example, if one were modeling an individual's choice of home-to-work mode on Tuesday, using their mode choice on Monday could represent the "mode state." Great caution must be exercised when using and interpreting the results of such state variables because they may be capturing spurious state dependence, which means the parameter is picking up the effects of unobserved factors and not true state effects.

Statistic: A summary value calculated from a sample of observations.

Statistical independence: In probability theory, two events are said to be statistically independent if, and only if, the probability that they both occur is equal to the product of their individual probabilities. That is, one event does not depend on another for its occurrence or nonoccurrence. In statistical notation, statistical independence is given by

$$P(AB) = P(A)P(B)$$

where $P(AB)$ is the probability that both event A and B occur.

Statistical inference: Also called inductive statistics, a form of reasoning from sample data to population parameters; that is, any generalization, prediction, estimate, or decision based on a sample and made about the population. There are two schools of thought in statistical inference, classical or frequentist statistics and Bayesian inference (see also induction and deduction).

Statistics: The branch of mathematics that deals with all aspects of the science of decision making and analysis of data in the face of uncertainty.

Stochastic: An adjective that implies that a process or data-generating mechanism involves a random component or components. A statistical model consists of stochastic and deterministic components.

Stratification: The division of a population into parts, known as strata.

Stratified random sampling: A method of sampling from a population whereby the population is divided into subsamples, known as strata, especially for the purpose of drawing a sample, and then assigned proportions of the sample are sampled from each stratum. The process of stratification is undertaken to reduce the variability of stratification statistics. Strata are generally selected such that interstrata variability is maximized, and intrastrata variability is small. When stratified sampling is performed as desired, estimates of strata statistics are more precise than the same estimates computed on a simple random sample.

Structural equations: Equations that capture the relationships among a system of variables, often a system consisting of latent variables. Structural equations are typically used to model an underlying behavioral theory.

Survivor function: A function that gives the probability that an event duration is greater than or equal to some specified time t . It is frequently used in hazard analyses of duration data and is written $S(t) = P[T \geq t]$, where P denotes probability, T is a random time variable, and t is some specified value of the random variable.

Systematic error: An error that is in some sense biased, having a distribution with a mean that is not zero (as opposed to a random error).

T

t distribution: Distribution of values obtained from the sampling distribution of the mean when the variance is unknown. It is used extensively in hypothesis testing—specifically tests about particular values of population means. The probabilities of the t distribution approach the standard normal distribution probabilities rapidly as n exceeds 30. The density of the t distribution is given as

$$f(x) = C(n) \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

where $C(n)$ is the normalizing constant given by

$$C(n) = \Gamma[(n+1)/2] / [(n\pi)^{1/2} \Gamma(n/2)]$$

The mean of the t distribution is zero and the variance is $n/(n - 2)$ for $n > 2$.

Three-stage least squares (3SLS): A method used in simultaneous equation estimation. Stage 1 obtains two-stage least squares (2SLS) estimates of the model system. Stage 2 uses the 2SLS estimates to compute residuals to determine cross-equation correlations. Stage 3 uses generalized least squares (GLS) to estimate model parameters.

Tied data: A problem encountered in the analysis of duration data in which a number of observations end their durations at the same time. When duration exits are grouped at specific times, the likelihood function for proportional and accelerated lifetime hazard models becomes increasingly complex.

Time series data: A set of ordered observations on a quantitative characteristic of an individual or collective phenomenon taken at different points of time. Although it is not a requirement, it is common for these points to be equidistant in time (see also cross-sectional and before-after data).

Transferability: A concern with all models is whether their estimated parameters are transferable spatially (among regions or cities) or temporally (over time). From a spatial perspective, transferability is desirable because it means that parameters of models estimated in other places are used, thus saving the cost of additional data collection and estimation. Temporal transferability ensures that forecasts made with the model have some validity in that the estimated parameters are stable over time (see also validation).

Transformation: The change in the scale of a variable. Transformations are performed to simplify calculations, to meet specific statistical modeling assumptions, to linearize an otherwise nonlinear relation with another variable, to impose practical limitations on a variable, and to change the characteristic shape of a probability distribution of the variable in its original scale.

Truncated distribution: A statistical distribution that occurs when a response above or below a certain threshold value is discarded. For example, assume that certain instrumentation can read measurements only within a certain range—data obtained from this instrument will result in a truncated distribution, as measurements outside the range are discarded. If measurements are recorded at the extreme range of the measurement device, but assigned a maximum or minimum value, then the distribution is censored.

Two-stage least squares (2SLS): A method used in simultaneous equation estimation. Stage 1 regresses each endogenous variable on all exogenous variables. Stage 2 uses regression-estimated values from stage 1 as instruments, and ordinary least squares is used to obtain statistical models.

Two-tailed test: A statistical test of significance in which the direction of an anticipated effect is not known a priori.

Type I error: If, as the result of a test statistic computed on sample data, a statistical hypothesis is rejected when it should be accepted, that is, when it is true, then a Type I error is committed. The α , or level of significance, is preselected by the analyst to determine the Type I error rate. The level of confidence of a particular test is given by $1 - \alpha$.

Type II error: If, as the result of a test statistic computed on sample data, a statistical hypothesis is accepted when it is false, i.e., when it should have been rejected, then a Type II error is committed. The error rate is preselected by the analyst to determine the Type II error rate, and the power of a particular statistical test is given by $1 - \beta$.

U

Unbiased estimator: An estimator whose expected value (the mean of the sampling distribution) equals the parameter it is supposed to estimate. In general, unbiased estimators are preferred to biased estimators of population parameters. There are rare cases, however, when biased estimators are preferred because they result in estimators with smaller standard errors.

Uniform distribution: Distributions that are appropriate for cases when the probability of obtaining an outcome within a range of outcomes is constant. It is expressed in terms of either discrete or continuous data. The probability density function for the discrete uniform distribution is given by

$$P(X = x) = U(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } \alpha < x < \beta \\ 0 & \text{elsewhere} \end{cases}$$

where x is the value of random variable, α is the lowermost value of the interval for x , and β is the uppermost value of the interval for x .

Unobserved heterogeneity: A problem that arises in statistical modeling when some unobserved factors (not included in the model) systematically vary across the population. Ignoring unobserved

heterogeneity can result in model specification errors that can lead one to draw erroneous inferences on model results.

Utility maximization: A basic premise of economic theory whereby consumers facing choices are assumed to maximize their personal utility subject to cost and income constraints. Utility theory may be used in the derivation and interpretation of discrete outcome models (but is not necessary).

V

Validation: A term used to describe the important activity of defending a statistical model. The only way to validate the generalizability or transferability of an estimated model is to make forecasts or backcasts with a model and compare them to data that were not used to estimate the model. This exercise is called external validation. The importance of this step of model building cannot be overstated, but it remains perhaps the least-practiced step of model building, because it is expensive and time-consuming, and because some modelers and practitioners confuse goodness-of-fit statistics computed on the sample data with the same computed on validation data.

Validity: Degree to which some procedure is founded on logic (internal or formal validity) or corresponds to nature (external or empirical validity).

Variability: A statistical term used to describe and quantify the spread or dispersion of data around its center, usually the mean. Knowledge of data variability is essential for conducting statistical tests and for fully understanding data. Thus, it is often desirable to obtain measures of both central tendency and spread. In fact, it may be misleading to consider only measures of central tendency when describing data.

Variable: A quantity that may take any one of a specified set of values.

Variance: A generic term used to describe the spread of data around the center of the data. The center typically is a mean, a median, or a regression function. The variance is the square of standard deviation.

Variate: A quantity that may take any of the values of a specified set with a specified relative frequency or probability; also known as a random variable.

W

Weibull distribution: A distribution commonly used in discrete outcome modeling and duration modeling. With parameters $\lambda > 0$ and $P > 0$, the Weibull distribution has the density function

$$f(x) = \lambda P (\lambda x)^{P-1} \text{EXP}\left[-(\lambda x)^P\right]$$

Weight: A numerical coefficient attached to an observation, frequently by multiplication, so that it will assume a desired degree of importance in a function of all the observations of the set.

Weighted average: An average of quantities to which have been attached a series of weights to make allowance for their relative importance. Weights are commonly used to accommodate nonrandom samples in statistical models, such as stratified samples, and are used to remediate heterogeneity with weighted least squares.

White noise: For time series analysis, noise that is defined as a series whose elements are uncorrelated and normally distributed with mean zero and constant variance. The residuals from properly specified and estimated time series models should be white noise.

Z

Z-value: A measure of a variable that quantifies its relative position within a standard normal distribution. It is the number of standard deviations from 0. As rules of thumb, Z-values of 1, 2, and 3 standard deviations from zero contain approximately 67%, 95%, and 99% of all observations in the distribution.

Appendix C

Statistical Tables

TABLE C.1

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
-4.00	0.0001	0.0000	1.0000	-3.60	0.0006	0.0002	0.9998
-3.99	0.0001	0.0000	1.0000	-3.59	0.0006	0.0002	0.9998
-3.98	0.0001	0.0000	1.0000	-3.58	0.0007	0.0002	0.9998
-3.97	0.0001	0.0000	1.0000	-3.57	0.0007	0.0002	0.9998
-3.96	0.0002	0.0000	1.0000	-3.56	0.0007	0.0002	0.9998
-3.95	0.0002	0.0000	1.0000	-3.55	0.0007	0.0002	0.9998
-3.94	0.0002	0.0000	1.0000	-3.54	0.0008	0.0002	0.9998
-3.93	0.0002	0.0000	1.0000	-3.53	0.0008	0.0002	0.9998
-3.92	0.0002	0.0000	1.0000	-3.52	0.0008	0.0002	0.9998
-3.91	0.0002	0.0001	1.0000	-3.51	0.0008	0.0002	0.9998
-3.90	0.0002	0.0001	1.0000	-3.50	0.0009	0.0002	0.9998
-3.89	0.0002	0.0001	1.0000	-3.49	0.0009	0.0002	0.9998
-3.88	0.0002	0.0001	1.0000	-3.48	0.0009	0.0003	0.9998
-3.87	0.0002	0.0001	1.0000	-3.47	0.0010	0.0003	0.9997
-3.86	0.0002	0.0001	0.9999	-3.46	0.0010	0.0003	0.9997
-3.85	0.0002	0.0001	0.9999	-3.45	0.0010	0.0003	0.9997
-3.84	0.0003	0.0001	0.9999	-3.44	0.0011	0.0003	0.9997
-3.83	0.0003	0.0001	0.9999	-3.43	0.0011	0.0003	0.9997
-3.82	0.0003	0.0001	0.9999	-3.42	0.0011	0.0003	0.9997
-3.81	0.0003	0.0001	0.9999	-3.41	0.0012	0.0003	0.9997
-3.80	0.0003	0.0001	0.9999	-3.40	0.0012	0.0003	0.9997
-3.79	0.0003	0.0001	0.9999	-3.39	0.0013	0.0003	0.9997
-3.78	0.0003	0.0001	0.9999	-3.38	0.0013	0.0004	0.9996
-3.77	0.0003	0.0001	0.9999	-3.37	0.0014	0.0004	0.9996
-3.76	0.0003	0.0001	0.9999	-3.36	0.0014	0.0004	0.9996
-3.75	0.0003	0.0001	0.9999	-3.35	0.0015	0.0004	0.9996
-3.74	0.0004	0.0001	0.9999	-3.34	0.0015	0.0004	0.9996
-3.73	0.0004	0.0001	0.9999	-3.33	0.0016	0.0004	0.9996
-3.72	0.0004	0.0001	0.9999	-3.32	0.0016	0.0004	0.9996
-3.71	0.0004	0.0001	0.9999	-3.31	0.0017	0.0005	0.9995
-3.70	0.0004	0.0001	0.9999	-3.30	0.0017	0.0005	0.9995
-3.69	0.0004	0.0001	0.9999	-3.29	0.0018	0.0005	0.9995
-3.68	0.0005	0.0001	0.9999	-3.28	0.0018	0.0005	0.9995
-3.67	0.0005	0.0001	0.9999	-3.27	0.0019	0.0005	0.9995
-3.66	0.0005	0.0001	0.9999	-3.26	0.0020	0.0006	0.9994
-3.65	0.0005	0.0001	0.9999	-3.25	0.0020	0.0006	0.9994
-3.64	0.0005	0.0001	0.9999	-3.24	0.0021	0.0006	0.9994
-3.63	0.0006	0.0001	0.9999	-3.23	0.0022	0.0006	0.9994
-3.62	0.0006	0.0001	0.9999	-3.22	0.0022	0.0006	0.9994
-3.61	0.0006	0.0001	0.9999	-3.21	0.0023	0.0007	0.9993
-3.60	0.0006	0.0002	0.9998	-3.20	0.0024	0.0007	0.9993

continued

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
-3.20	0.0024	0.0007	0.9993	-2.70	0.0104	0.0035	0.9965
-3.19	0.0025	0.0007	0.9993	-2.69	0.0107	0.0036	0.9964
-3.18	0.0025	0.0007	0.9993	-2.68	0.0110	0.0037	0.9963
-3.17	0.0026	0.0008	0.9992	-2.67	0.0113	0.0038	0.9962
-3.16	0.0027	0.0008	0.9992	-2.66	0.0116	0.0039	0.9961
-3.15	0.0028	0.0008	0.9992	-2.65	0.0119	0.0040	0.9960
-3.14	0.0029	0.0008	0.9992	-2.64	0.0122	0.0042	0.9959
-3.13	0.0030	0.0009	0.9991	-2.63	0.0126	0.0043	0.9957
-3.12	0.0031	0.0009	0.9991	-2.62	0.0129	0.0044	0.9956
-3.11	0.0032	0.0009	0.9991	-2.61	0.0132	0.0045	0.9955
-3.10	0.0033	0.0010	0.9990	-2.60	0.0136	0.0047	0.9953
-3.09	0.0034	0.0010	0.9990	-2.59	0.0139	0.0048	0.9952
-3.08	0.0035	0.0010	0.9990	-2.58	0.0143	0.0049	0.9951
-3.07	0.0036	0.0011	0.9989	-2.57	0.0147	0.0051	0.9949
-3.06	0.0037	0.0011	0.9989	-2.56	0.0151	0.0052	0.9948
-3.05	0.0038	0.0011	0.9989	-2.55	0.0155	0.0054	0.9946
-3.04	0.0039	0.0012	0.9988	-2.54	0.0158	0.0055	0.9945
-3.03	0.0040	0.0012	0.9988	-2.53	0.0163	0.0057	0.9943
-3.02	0.0042	0.0013	0.9987	-2.52	0.0167	0.0059	0.9941
-3.01	0.0043	0.0013	0.9987	-2.51	0.0171	0.0060	0.9940
-3.00	0.0044	0.0014	0.9987	-2.50	0.0175	0.0062	0.9938
-2.99	0.0046	0.0014	0.9986	-2.49	0.0180	0.0064	0.9936
-2.98	0.0047	0.0014	0.9986	-2.48	0.0184	0.0066	0.9934
-2.97	0.0049	0.0015	0.9985	-2.47	0.0189	0.0068	0.9932
-2.96	0.0050	0.0015	0.9985	-2.46	0.0194	0.0069	0.9930
-2.95	0.0051	0.0016	0.9984	-2.45	0.0198	0.0071	0.9929
-2.94	0.0053	0.0016	0.9984	-2.44	0.0203	0.0073	0.9927
-2.93	0.0054	0.0017	0.9983	-2.43	0.0208	0.0076	0.9925
-2.92	0.0056	0.0018	0.9982	-2.42	0.0213	0.0078	0.9922
-2.91	0.0058	0.0018	0.9982	-2.41	0.0219	0.0080	0.9920
-2.90	0.0060	0.0019	0.9981	-2.40	0.0224	0.0082	0.9918
-2.89	0.0061	0.0019	0.9981	-2.39	0.0229	0.0084	0.9916
-2.88	0.0063	0.0020	0.9980	-2.38	0.0235	0.0087	0.9913
-2.87	0.0065	0.0021	0.9980	-2.37	0.0241	0.0089	0.9911
-2.86	0.0067	0.0021	0.9979	-2.36	0.0246	0.0091	0.9909
-2.85	0.0069	0.0022	0.9978	-2.35	0.0252	0.0094	0.9906
-2.84	0.0071	0.0023	0.9977	-2.34	0.0258	0.0096	0.9904
-2.83	0.0073	0.0023	0.9977	-2.33	0.0264	0.0099	0.9901
-2.82	0.0075	0.0024	0.9976	-2.32	0.0271	0.0102	0.9898
-2.81	0.0077	0.0025	0.9975	-2.31	0.0277	0.0104	0.9896
-2.80	0.0079	0.0026	0.9974	-2.30	0.0283	0.0107	0.9893
-2.79	0.0081	0.0026	0.9974	-2.29	0.0290	0.0110	0.9890
-2.78	0.0084	0.0027	0.9973	-2.28	0.0296	0.0113	0.9887
-2.77	0.0086	0.0028	0.9972	-2.27	0.0303	0.0116	0.9884
-2.76	0.0089	0.0029	0.9971	-2.26	0.0310	0.0119	0.9881
-2.75	0.0091	0.0030	0.9970	-2.25	0.0317	0.0122	0.9878
-2.74	0.0094	0.0031	0.9969	-2.24	0.0325	0.0126	0.9875
-2.73	0.0096	0.0032	0.9968	-2.23	0.0332	0.0129	0.9871
-2.72	0.0099	0.0033	0.9967	-2.22	0.0339	0.0132	0.9868
-2.71	0.0101	0.0034	0.9966	-2.21	0.0347	0.0135	0.9865
-2.70	0.0104	0.0035	0.9965	-2.20	0.0355	0.0139	0.9861

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
-2.20	0.0355	0.0139	0.9861	-1.70	0.0940	0.0446	0.9554
-2.19	0.0363	0.0143	0.9857	-1.69	0.0957	0.0455	0.9545
-2.18	0.0371	0.0146	0.9854	-1.68	0.0973	0.0465	0.9535
-2.17	0.0379	0.0150	0.9850	-1.67	0.0989	0.0475	0.9525
-2.16	0.0387	0.0154	0.9846	-1.66	0.1006	0.0485	0.9515
-2.15	0.0396	0.0158	0.9842	-1.65	0.1023	0.0495	0.9505
-2.14	0.0404	0.0162	0.9838	-1.64	0.1040	0.0505	0.9495
-2.13	0.0413	0.0166	0.9834	-1.63	0.1057	0.0515	0.9485
-2.12	0.0422	0.0170	0.9830	-1.62	0.1074	0.0526	0.9474
-2.11	0.0431	0.0174	0.9826	-1.61	0.1091	0.0537	0.9463
-2.10	0.0440	0.0179	0.9821	-1.60	0.1109	0.0548	0.9452
-2.09	0.0449	0.0183	0.9817	-1.59	0.1127	0.0559	0.9441
-2.08	0.0459	0.0188	0.9812	-1.58	0.1145	0.0570	0.9429
-2.07	0.0468	0.0192	0.9808	-1.57	0.1163	0.0582	0.9418
-2.06	0.0478	0.0197	0.9803	-1.56	0.1182	0.0594	0.9406
-2.05	0.0488	0.0202	0.9798	-1.55	0.1200	0.0606	0.9394
-2.04	0.0498	0.0207	0.9793	-1.54	0.1219	0.0618	0.9382
-2.03	0.0508	0.0212	0.9788	-1.53	0.1238	0.0630	0.9370
-2.02	0.0519	0.0217	0.9783	-1.52	0.1257	0.0643	0.9357
-2.01	0.0529	0.0222	0.9778	-1.51	0.1276	0.0655	0.9345
-2.00	0.0540	0.0227	0.9772	-1.50	0.1295	0.0668	0.9332
-1.99	0.0551	0.0233	0.9767	-1.49	0.1315	0.0681	0.9319
-1.98	0.0562	0.0238	0.9761	-1.48	0.1334	0.0694	0.9306
-1.97	0.0573	0.0244	0.9756	-1.47	0.1354	0.0708	0.9292
-1.96	0.0584	0.0250	0.9750	-1.46	0.1374	0.0722	0.9278
-1.95	0.0596	0.0256	0.9744	-1.45	0.1394	0.0735	0.9265
-1.94	0.0608	0.0262	0.9738	-1.44	0.1415	0.0749	0.9251
-1.93	0.0619	0.0268	0.9732	-1.43	0.1435	0.0764	0.9236
-1.92	0.0632	0.0274	0.9726	-1.42	0.1456	0.0778	0.9222
-1.91	0.0644	0.0281	0.9719	-1.41	0.1476	0.0793	0.9207
-1.90	0.0656	0.0287	0.9713	-1.40	0.1497	0.0808	0.9192
-1.89	0.0669	0.0294	0.9706	-1.39	0.1518	0.0823	0.9177
-1.88	0.0681	0.0301	0.9699	-1.38	0.1540	0.0838	0.9162
-1.87	0.0694	0.0307	0.9693	-1.37	0.1561	0.0853	0.9147
-1.86	0.0707	0.0314	0.9686	-1.36	0.1582	0.0869	0.9131
-1.85	0.0721	0.0322	0.9678	-1.35	0.1604	0.0885	0.9115
-1.84	0.0734	0.0329	0.9671	-1.34	0.1626	0.0901	0.9099
-1.83	0.0748	0.0336	0.9664	-1.33	0.1647	0.0918	0.9082
-1.82	0.0761	0.0344	0.9656	-1.32	0.1669	0.0934	0.9066
-1.81	0.0775	0.0352	0.9648	-1.31	0.1691	0.0951	0.9049
-1.80	0.0790	0.0359	0.9641	-1.30	0.1714	0.0968	0.9032
-1.79	0.0804	0.0367	0.9633	-1.29	0.1736	0.0985	0.9015
-1.78	0.0818	0.0375	0.9625	-1.28	0.1759	0.1003	0.8997
-1.77	0.0833	0.0384	0.9616	-1.27	0.1781	0.1020	0.8980
-1.76	0.0848	0.0392	0.9608	-1.26	0.1804	0.1038	0.8962
-1.75	0.0863	0.0401	0.9599	-1.25	0.1827	0.1056	0.8943
-1.74	0.0878	0.0409	0.9591	-1.24	0.1849	0.1075	0.8925
-1.73	0.0893	0.0418	0.9582	-1.23	0.1872	0.1094	0.8907
-1.72	0.0909	0.0427	0.9573	-1.22	0.1895	0.1112	0.8888
-1.71	0.0925	0.0436	0.9564	-1.21	0.1919	0.1131	0.8869
-1.70	0.0940	0.0446	0.9554	-1.20	0.1942	0.1151	0.8849

continued

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
-1.20	0.1942	0.1151	0.8849	-0.70	0.3123	0.2420	0.7580
-1.19	0.1965	0.1170	0.8830	-0.69	0.3144	0.2451	0.7549
-1.18	0.1989	0.1190	0.8810	-0.68	0.3166	0.2482	0.7518
-1.17	0.2012	0.1210	0.8790	-0.67	0.3187	0.2514	0.7486
-1.16	0.2036	0.1230	0.8770	-0.66	0.3209	0.2546	0.7454
-1.15	0.2059	0.1251	0.8749	-0.65	0.3230	0.2579	0.7421
-1.14	0.2083	0.1271	0.8729	-0.64	0.3251	0.2611	0.7389
-1.13	0.2107	0.1292	0.8708	-0.63	0.3271	0.2643	0.7357
-1.12	0.2131	0.1314	0.8686	-0.62	0.3292	0.2676	0.7324
-1.11	0.2155	0.1335	0.8665	-0.61	0.3312	0.2709	0.7291
-1.10	0.2178	0.1357	0.8643	-0.60	0.3332	0.2742	0.7258
-1.09	0.2203	0.1379	0.8621	-0.59	0.3352	0.2776	0.7224
-1.08	0.2226	0.1401	0.8599	-0.58	0.3372	0.2810	0.7190
-1.07	0.2251	0.1423	0.8577	-0.57	0.3391	0.2843	0.7157
-1.06	0.2275	0.1446	0.8554	-0.56	0.3411	0.2877	0.7123
-1.05	0.2299	0.1469	0.8531	-0.55	0.3429	0.2912	0.7088
-1.04	0.2323	0.1492	0.8508	-0.54	0.3448	0.2946	0.7054
-1.03	0.2347	0.1515	0.8485	-0.53	0.3467	0.2981	0.7019
-1.02	0.2371	0.1539	0.8461	-0.52	0.3485	0.3015	0.6985
-1.01	0.2396	0.1563	0.8438	-0.51	0.3503	0.3050	0.6950
-1.00	0.2420	0.1587	0.8413	-0.50	0.3521	0.3085	0.6915
-0.99	0.2444	0.1611	0.8389	-0.49	0.3538	0.3121	0.6879
-0.98	0.2468	0.1635	0.8365	-0.48	0.3555	0.3156	0.6844
-0.97	0.2492	0.1660	0.8340	-0.47	0.3572	0.3192	0.6808
-0.96	0.2516	0.1685	0.8315	-0.46	0.3589	0.3228	0.6772
-0.95	0.2541	0.1711	0.8289	-0.45	0.3605	0.3264	0.6736
-0.94	0.2565	0.1736	0.8264	-0.44	0.3621	0.3300	0.6700
-0.93	0.2589	0.1762	0.8238	-0.43	0.3637	0.3336	0.6664
-0.92	0.2613	0.1788	0.8212	-0.42	0.3653	0.3372	0.6628
-0.91	0.2637	0.1814	0.8186	-0.41	0.3668	0.3409	0.6591
-0.90	0.2661	0.1841	0.8159	-0.40	0.3683	0.3446	0.6554
-0.89	0.2685	0.1867	0.8133	-0.39	0.3697	0.3483	0.6517
-0.88	0.2709	0.1894	0.8106	-0.38	0.3711	0.3520	0.6480
-0.87	0.2732	0.1921	0.8078	-0.37	0.3725	0.3557	0.6443
-0.86	0.2756	0.1949	0.8051	-0.36	0.3739	0.3594	0.6406
-0.85	0.2780	0.1977	0.8023	-0.35	0.3752	0.3632	0.6368
-0.84	0.2803	0.2004	0.7995	-0.34	0.3765	0.3669	0.6331
-0.83	0.2827	0.2033	0.7967	-0.33	0.3778	0.3707	0.6293
-0.82	0.2850	0.2061	0.7939	-0.32	0.3790	0.3745	0.6255
-0.81	0.2874	0.2090	0.7910	-0.31	0.3802	0.3783	0.6217
-0.80	0.2897	0.2119	0.7881	-0.30	0.3814	0.3821	0.6179
-0.79	0.2920	0.2148	0.7852	-0.29	0.3825	0.3859	0.6141
-0.78	0.2943	0.2177	0.7823	-0.28	0.3836	0.3897	0.6103
-0.77	0.2966	0.2207	0.7793	-0.27	0.3847	0.3936	0.6064
-0.76	0.2989	0.2236	0.7764	-0.26	0.3857	0.3974	0.6026
-0.75	0.3011	0.2266	0.7734	-0.25	0.3867	0.4013	0.5987
-0.74	0.3034	0.2296	0.7703	-0.24	0.3896	0.4052	0.5948
-0.73	0.3056	0.2327	0.7673	-0.23	0.3885	0.4091	0.5909
-0.72	0.3079	0.2358	0.7642	-0.22	0.3894	0.4129	0.5871
-0.71	0.3101	0.2389	0.7611	-0.21	0.3902	0.4168	0.5832
-0.70	0.3123	0.2420	0.7580	-0.20	0.3910	0.4207	0.5793

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
-0.20	0.3910	0.4207	0.5793	0.30	0.3814	0.6179	0.3821
-0.19	0.3918	0.4247	0.5754	0.31	0.3802	0.6217	0.3783
-0.18	0.3925	0.4286	0.5714	0.32	0.3790	0.6255	0.3745
-0.17	0.3932	0.4325	0.5675	0.33	0.3778	0.6293	0.3707
-0.16	0.3939	0.4364	0.5636	0.34	0.3765	0.6331	0.3669
-0.15	0.3945	0.4404	0.5596	0.35	0.3752	0.6368	0.3632
-0.14	0.3951	0.4443	0.5557	0.36	0.3739	0.6406	0.3594
-0.13	0.3956	0.4483	0.5517	0.37	0.3725	0.6443	0.3557
-0.12	0.3961	0.4522	0.5478	0.38	0.3711	0.6480	0.3520
-0.11	0.3965	0.4562	0.5438	0.39	0.3697	0.6517	0.3483
-0.10	0.3970	0.4602	0.5398	0.40	0.3683	0.6554	0.3446
-0.09	0.3973	0.4641	0.5359	0.41	0.3668	0.6591	0.3409
-0.08	0.3977	0.4681	0.5319	0.42	0.3653	0.6628	0.3372
-0.07	0.3980	0.4721	0.5279	0.43	0.3637	0.6664	0.3336
-0.06	0.3982	0.4761	0.5239	0.44	0.3621	0.6700	0.3300
-0.05	0.3984	0.4801	0.5199	0.45	0.3605	0.6736	0.3264
-0.04	0.3986	0.4840	0.5160	0.46	0.3589	0.6772	0.3228
-0.03	0.3988	0.4880	0.5120	0.47	0.3572	0.6808	0.3192
-0.02	0.3989	0.4920	0.5080	0.48	0.3555	0.6844	0.3156
-0.01	0.3989	0.4960	0.5040	0.49	0.3538	0.6879	0.3121
0.00	0.3989	0.5000	0.5000	0.50	0.3521	0.6915	0.3085
0.01	0.3989	0.5040	0.4960	0.51	0.3503	0.6950	0.3050
0.02	0.3989	0.5080	0.4920	0.52	0.3485	0.6985	0.3015
0.03	0.3988	0.5120	0.4880	0.53	0.3467	0.7019	0.2981
0.04	0.3986	0.5160	0.4840	0.54	0.3448	0.7054	0.2946
0.05	0.3984	0.5199	0.4801	0.55	0.3429	0.7088	0.2912
0.06	0.3982	0.5239	0.4761	0.56	0.3411	0.7123	0.2877
0.07	0.3980	0.5279	0.4721	0.57	0.3391	0.7157	0.2843
0.08	0.3977	0.5319	0.4681	0.58	0.3372	0.7190	0.2810
0.09	0.3973	0.5359	0.4641	0.59	0.3352	0.7224	0.2776
0.10	0.3970	0.5398	0.4602	0.60	0.3332	0.7258	0.2742
0.11	0.3965	0.5438	0.4562	0.61	0.3312	0.7291	0.2709
0.12	0.3961	0.5478	0.4522	0.62	0.3292	0.7324	0.2676
0.13	0.3956	0.5517	0.4483	0.63	0.3271	0.7357	0.2643
0.14	0.3951	0.5557	0.4443	0.64	0.3251	0.7389	0.2611
0.15	0.3945	0.5596	0.4404	0.65	0.3230	0.7421	0.2579
0.16	0.3939	0.5636	0.4364	0.66	0.3209	0.7454	0.2546
0.17	0.3932	0.5675	0.4325	0.67	0.3187	0.7486	0.2514
0.18	0.3925	0.5714	0.4286	0.68	0.3166	0.7518	0.2482
0.19	0.3918	0.5754	0.4247	0.69	0.3144	0.7549	0.2451
0.20	0.3910	0.5793	0.4207	0.70	0.3123	0.7580	0.2420
0.21	0.3902	0.5832	0.4168	0.71	0.3101	0.7611	0.2389
0.22	0.3894	0.5871	0.4129	0.72	0.3079	0.7642	0.2358
0.23	0.3885	0.5909	0.4091	0.73	0.3056	0.7673	0.2327
0.24	0.3876	0.5948	0.4052	0.74	0.3034	0.7703	0.2296
0.25	0.3867	0.5987	0.4013	0.75	0.3011	0.7734	0.2266
0.26	0.3857	0.6026	0.3974	0.76	0.2989	0.7764	0.2236
0.27	0.3847	0.6064	0.3936	0.77	0.2966	0.7793	0.2207
0.28	0.3836	0.6103	0.3897	0.78	0.2943	0.7823	0.2177
0.29	0.3825	0.6141	0.3859	0.79	0.2920	0.7852	0.2148
0.30	0.3814	0.6179	0.3821	0.80	0.2897	0.7881	0.2119

continued

TABLE C.1 (continued)

Normal Distribution

<i>z</i>	<i>f(z)</i>	<i>F(z)</i>	$1 - F(z)$	<i>z</i>	<i>f(z)</i>	<i>F(z)</i>	$1 - F(z)$
0.80	0.2897	0.7881	0.2119	1.30	0.1714	0.9032	0.0968
0.81	0.2874	0.7910	0.2090	1.31	0.1691	0.9049	0.0951
0.82	0.2850	0.7939	0.2061	1.32	0.1669	0.9066	0.0934
0.83	0.2827	0.7967	0.2033	1.33	0.1647	0.9082	0.0918
0.84	0.2803	0.7995	0.2004	1.34	0.1626	0.9099	0.0901
0.85	0.2780	0.8023	0.1977	1.35	0.1604	0.9115	0.0885
0.86	0.2756	0.8051	0.1949	1.36	0.1582	0.9131	0.0869
0.87	0.2732	0.8078	0.1921	1.37	0.1561	0.9147	0.0853
0.88	0.2709	0.8106	0.1894	1.38	0.1540	0.9162	0.0838
0.89	0.2685	0.8133	0.1867	1.39	0.1518	0.9177	0.0823
0.90	0.2661	0.8159	0.1841	1.40	0.1497	0.9192	0.0808
0.91	0.2637	0.8186	0.1814	1.41	0.1476	0.9207	0.0793
0.92	0.2613	0.8212	0.1788	1.42	0.1456	0.9222	0.0778
0.93	0.2589	0.8238	0.1762	1.43	0.1435	0.9236	0.0764
0.94	0.2565	0.8264	0.1736	1.44	0.1415	0.9251	0.0749
0.95	0.2541	0.8289	0.1711	1.45	0.1394	0.9265	0.0735
0.96	0.2516	0.8315	0.1685	1.46	0.1374	0.9278	0.0722
0.97	0.2492	0.8340	0.1660	1.47	0.1354	0.9292	0.0708
0.98	0.2468	0.8365	0.1635	1.48	0.1334	0.9306	0.0694
0.99	0.2444	0.8389	0.1611	1.49	0.1315	0.9319	0.0681
1.00	0.2420	0.8413	0.1587	1.50	0.1295	0.9332	0.0668
1.01	0.2396	0.8438	0.1563	1.51	0.1276	0.9345	0.0655
1.02	0.2371	0.8461	0.1539	1.52	0.1257	0.9357	0.0643
1.03	0.2347	0.8485	0.1515	1.53	0.1238	0.9370	0.0630
1.04	0.2323	0.8508	0.1492	1.54	0.1219	0.9382	0.0618
1.05	0.2299	0.8531	0.1469	1.55	0.1200	0.9394	0.0606
1.06	0.2275	0.8554	0.1446	1.56	0.1182	0.9406	0.0594
1.07	0.2251	0.8577	0.1423	1.57	0.1163	0.9418	0.0582
1.08	0.2226	0.8599	0.1401	1.58	0.1145	0.9429	0.0570
1.09	0.2203	0.8621	0.1379	1.59	0.1127	0.9441	0.0559
1.10	0.2178	0.8643	0.1357	1.60	0.1109	0.9452	0.0548
1.11	0.2155	0.8665	0.1335	1.61	0.1091	0.9463	0.0537
1.12	0.2131	0.8686	0.1314	1.62	0.1074	0.9474	0.0526
1.13	0.2107	0.8708	0.1292	1.63	0.1057	0.9485	0.0515
1.14	0.2083	0.8729	0.1271	1.64	0.1040	0.9495	0.0505
1.15	0.2059	0.8749	0.1251	1.65	0.1023	0.9505	0.0495
1.16	0.2036	0.8770	0.1230	1.66	0.1006	0.9515	0.0485
1.17	0.2012	0.8790	0.1210	1.67	0.0989	0.9525	0.0475
1.18	0.1989	0.8810	0.1190	1.68	0.0973	0.9535	0.0465
1.19	0.1965	0.8830	0.1170	1.69	0.0957	0.9545	0.0455
1.20	0.1942	0.8849	0.1151	1.70	0.0940	0.9554	0.0446
1.21	0.1919	0.8869	0.1131	1.71	0.0925	0.9564	0.0436
1.22	0.1895	0.8888	0.1112	1.72	0.0909	0.9573	0.0427
1.23	0.1872	0.8907	0.1094	1.73	0.0893	0.9582	0.0418
1.24	0.1849	0.8925	0.1075	1.74	0.0878	0.9591	0.0409
1.25	0.1827	0.8943	0.1056	1.75	0.0863	0.9599	0.0401
1.26	0.1804	0.8962	0.1038	1.76	0.0848	0.9608	0.0392
1.27	0.1781	0.8980	0.1020	1.77	0.0833	0.9616	0.0384
1.28	0.1759	0.8997	0.1003	1.78	0.0818	0.9625	0.0375
1.29	0.1736	0.9015	0.0985	1.79	0.0804	0.9633	0.0367
1.30	0.1714	0.9032	0.0968	1.80	0.0790	0.9641	0.0359

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
1.80	0.0790	0.9641	0.0359	2.30	0.0283	0.9893	0.0107
1.81	0.0775	0.9648	0.0352	2.31	0.0277	0.9896	0.0104
1.82	0.0761	0.9656	0.0344	2.32	0.0271	0.9898	0.0102
1.83	0.0748	0.9664	0.0336	2.33	0.0264	0.9901	0.0099
1.84	0.0734	0.9671	0.0329	2.34	0.0258	0.9904	0.0096
1.85	0.0721	0.9678	0.0322	2.35	0.0252	0.9906	0.0094
1.86	0.0707	0.9686	0.0314	2.36	0.0246	0.9909	0.0091
1.87	0.0694	0.9693	0.0307	2.37	0.0241	0.9911	0.0089
1.88	0.0681	0.9699	0.0301	2.38	0.0235	0.9913	0.0087
1.89	0.0669	0.9706	0.0294	2.39	0.0229	0.9916	0.0084
1.90	0.0656	0.9713	0.0287	2.40	0.0224	0.9918	0.0082
1.91	0.0644	0.9719	0.0281	2.41	0.0219	0.9920	0.0080
1.92	0.0632	0.9726	0.0274	2.42	0.0213	0.9922	0.0078
1.93	0.0619	0.9732	0.0268	2.43	0.0208	0.9925	0.0076
1.94	0.0608	0.9738	0.0262	2.44	0.0203	0.9927	0.0073
1.95	0.0596	0.9744	0.0256	2.45	0.0198	0.9929	0.0071
1.96	0.0584	0.9750	0.0250	2.46	0.0194	0.9930	0.0069
1.97	0.0573	0.9756	0.0244	2.47	0.0189	0.9932	0.0068
1.98	0.0562	0.9761	0.0238	2.48	0.0184	0.9934	0.0066
1.99	0.0551	0.9767	0.0233	2.49	0.0180	0.9936	0.0064
2.00	0.0540	0.9772	0.0227	2.50	0.0175	0.9938	0.0062
2.01	0.0529	0.9778	0.0222	2.51	0.0171	0.9940	0.0060
2.02	0.0519	0.9783	0.0217	2.52	0.0167	0.9941	0.0059
2.03	0.0508	0.9788	0.0212	2.53	0.0163	0.9943	0.0057
2.04	0.0498	0.9793	0.0207	2.54	0.0158	0.9945	0.0055
2.05	0.0488	0.9798	0.0202	2.55	0.0155	0.9946	0.0054
2.06	0.0478	0.9803	0.0197	2.56	0.0151	0.9948	0.0052
2.07	0.0468	0.9808	0.0192	2.57	0.0147	0.9949	0.0051
2.08	0.0459	0.9812	0.0188	2.58	0.0143	0.9951	0.0049
2.09	0.0449	0.9817	0.0183	2.59	0.0139	0.9952	0.0048
2.10	0.0440	0.9821	0.0179	2.60	0.0136	0.9953	0.0047
2.11	0.0431	0.9826	0.0174	2.61	0.0132	0.9955	0.0045
2.12	0.0422	0.9830	0.0170	2.62	0.0129	0.9956	0.0044
2.13	0.0413	0.9834	0.0166	2.63	0.0126	0.9957	0.0043
2.14	0.0404	0.9838	0.0162	2.64	0.0122	0.9959	0.0042
2.15	0.0396	0.9842	0.0158	2.65	0.0119	0.9960	0.0040
2.16	0.0387	0.9846	0.0154	2.66	0.0116	0.9961	0.0039
2.17	0.0379	0.9850	0.0150	2.67	0.0113	0.9962	0.0038
2.18	0.0371	0.9854	0.0146	2.68	0.0110	0.9963	0.0037
2.19	0.0363	0.9857	0.0143	2.69	0.0107	0.9964	0.0036
2.20	0.0355	0.9861	0.0139	2.70	0.0104	0.9965	0.0035
2.21	0.0347	0.9865	0.0135	2.71	0.0101	0.9966	0.0034
2.22	0.0339	0.9868	0.0132	2.72	0.0099	0.9967	0.0033
2.23	0.0332	0.9871	0.0129	2.73	0.0096	0.9968	0.0032
2.24	0.0325	0.9875	0.0126	2.74	0.0094	0.9969	0.0031
2.25	0.0317	0.9878	0.0122	2.75	0.0091	0.9970	0.0030
2.26	0.0310	0.9881	0.0119	2.76	0.0089	0.9971	0.0029
2.27	0.0303	0.9884	0.0116	2.77	0.0086	0.9972	0.0028
2.28	0.0296	0.9887	0.0113	2.78	0.0084	0.9973	0.0027
2.29	0.0290	0.9890	0.0110	2.79	0.0081	0.9974	0.0026
2.30	0.0283	0.9893	0.0107	2.80	0.0079	0.9974	0.0026

continued

TABLE C.1 (continued)

Normal Distribution

<i>z</i>	<i>f(z)</i>	<i>F(z)</i>	$1 - F(z)$	<i>z</i>	<i>f(z)</i>	<i>F(z)</i>	$1 - F(z)$
2.80	0.0079	0.9974	0.0026	3.30	0.0017	0.9995	0.0005
2.81	0.0077	0.9975	0.0025	3.31	0.0017	0.9995	0.0005
2.82	0.0075	0.9976	0.0024	3.32	0.0016	0.9996	0.0004
2.83	0.0073	0.9977	0.0023	3.33	0.0016	0.9996	0.0004
2.84	0.0071	0.9977	0.0023	3.34	0.0015	0.9996	0.0004
2.85	0.0069	0.9978	0.0022	3.35	0.0015	0.9996	0.0004
2.86	0.0067	0.9979	0.0021	3.36	0.0014	0.9996	0.0004
2.87	0.0065	0.9980	0.0021	3.37	0.0014	0.9996	0.0004
2.88	0.0063	0.9980	0.0020	3.38	0.0013	0.9996	0.0004
2.89	0.0061	0.9981	0.0019	3.39	0.0013	0.9997	0.0003
2.90	0.0060	0.9981	0.0019	3.40	0.0012	0.9997	0.0003
2.91	0.0058	0.9982	0.0018	3.41	0.0012	0.9997	0.0003
2.92	0.0056	0.9982	0.0018	3.42	0.0011	0.9997	0.0003
2.93	0.0054	0.9983	0.0017	3.43	0.0011	0.9997	0.0003
2.94	0.0053	0.9984	0.0016	3.44	0.0011	0.9997	0.0003
2.95	0.0051	0.9984	0.0016	3.45	0.0010	0.9997	0.0003
2.96	0.0050	0.9985	0.0015	3.46	0.0010	0.9997	0.0003
2.97	0.0049	0.9985	0.0015	3.47	0.0010	0.9997	0.0003
2.98	0.0047	0.9986	0.0014	3.48	0.0009	0.9998	0.0003
2.99	0.0046	0.9986	0.0014	3.49	0.0009	0.9998	0.0002
3.00	0.0044	0.9987	0.0014	3.50	0.0009	0.9998	0.0002
3.01	0.0043	0.9987	0.0013	3.51	0.0008	0.9998	0.0002
3.02	0.0042	0.9987	0.0013	3.52	0.0008	0.9998	0.0002
3.03	0.0040	0.9988	0.0012	3.53	0.0008	0.9998	0.0002
3.04	0.0039	0.9988	0.0012	3.54	0.0008	0.9998	0.0002
3.05	0.0038	0.9989	0.0011	3.55	0.0007	0.9998	0.0002
3.06	0.0037	0.9989	0.0011	3.56	0.0007	0.9998	0.0002
3.07	0.0036	0.9989	0.0011	3.57	0.0007	0.9998	0.0002
3.08	0.0035	0.9990	0.0010	3.58	0.0007	0.9998	0.0002
3.09	0.0034	0.9990	0.0010	3.59	0.0006	0.9998	0.0002
3.10	0.0033	0.9990	0.0010	3.60	0.0006	0.9998	0.0002
3.11	0.0032	0.9991	0.0009	3.61	0.0006	0.9999	0.0001
3.12	0.0031	0.9991	0.0009	3.62	0.0006	0.9999	0.0001
3.13	0.0030	0.9991	0.0009	3.63	0.0006	0.9999	0.0001
3.14	0.0029	0.9992	0.0008	3.64	0.0005	0.9999	0.0001
3.15	0.0028	0.9992	0.0008	3.65	0.0005	0.9999	0.0001
3.16	0.0027	0.9992	0.0008	3.66	0.0005	0.9999	0.0001
3.17	0.0026	0.9992	0.0008	3.67	0.0005	0.9999	0.0001
3.18	0.0025	0.9993	0.0007	3.68	0.0005	0.9999	0.0001
3.19	0.0025	0.9993	0.0007	3.69	0.0004	0.9999	0.0001
3.20	0.0024	0.9993	0.0007	3.70	0.0004	0.9999	0.0001
3.21	0.0023	0.9993	0.0007	3.71	0.0004	0.9999	0.0001
3.22	0.0022	0.9994	0.0006	3.72	0.0004	0.9999	0.0001
3.23	0.0022	0.9994	0.0006	3.73	0.0004	0.9999	0.0001
3.24	0.0021	0.9994	0.0006	3.74	0.0004	0.9999	0.0001
3.25	0.0020	0.9994	0.0006	3.75	0.0003	0.9999	0.0001
3.26	0.0020	0.9994	0.0006	3.76	0.0003	0.9999	0.0001
3.27	0.0019	0.9995	0.0005	3.77	0.0003	0.9999	0.0001
3.28	0.0018	0.9995	0.0005	3.78	0.0003	0.9999	0.0001
3.29	0.0018	0.9995	0.0005	3.79	0.0003	0.9999	0.0001
3.30	0.0017	0.9995	0.0005	3.80	0.0003	0.9999	0.0001

TABLE C.1 (continued)

Normal Distribution

z	$f(z)$	$F(z)$	$1 - F(z)$	z	$f(z)$	$F(z)$	$1 - F(z)$
3.80	0.0003	0.9999	0.0001	3.90	0.0002	1.0000	0.0001
3.81	0.0003	0.9999	0.0001	3.91	0.0002	1.0000	0.0001
3.82	0.0003	0.9999	0.0001	3.92	0.0002	1.0000	0.0000
3.83	0.0003	0.9999	0.0001	3.93	0.0002	1.0000	0.0000
3.84	0.0003	0.9999	0.0001	3.94	0.0002	1.0000	0.0000
3.85	0.0002	0.9999	0.0001	3.95	0.0002	1.0000	0.0000
3.86	0.0002	0.9999	0.0001	3.96	0.0002	1.0000	0.0000
3.87	0.0002	1.0000	0.0001	3.97	0.0001	1.0000	0.0000
3.88	0.0002	1.0000	0.0001	3.98	0.0001	1.0000	0.0000
3.89	0.0002	1.0000	0.0001	3.99	0.0001	1.0000	0.0000
3.90	0.0002	1.0000	0.0001	4.00	0.0001	1.0000	0.0000

TABLE C.2Critical Values for the t Distribution

v	α						
	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
45	1.301	1.679	2.014	2.412	2.690	3.281	3.520
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
100	0.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.091	3.291

TABLE C.3Critical Values for the Chi-Square Distribution $\chi^2_{\alpha, v}$

v	α							
	.9999	.9995	.999	.995	.99	.975	.95	.90
1	.07157	.06393	.05157	.04393	.0002	.0010	.0039	.0158
2	.0002	.0010	.0020	.0100	.0201	.0506	.1026	.2107
3	.0052	.0153	.0243	.0717	.1148	.2158	.3518	.5844
4	.0284	.0639	.0908	.2070	.2971	.4844	.7107	1.0636
5	.0822	.1581	.2102	.4117	.5543	.8312	1.1455	1.6103
6	.1724	.2994	.3811	.6757	.8721	1.2373	1.6354	2.2041
7	.3000	.4849	.5985	.9893	1.2390	1.6899	2.1673	2.8331
8	.4636	.7104	.8571	1.3444	1.6465	2.1797	2.7326	3.4895
9	.6608	.9717	1.1519	1.7349	2.0879	2.7004	3.3251	4.1682
10	.889	1.2650	1.4787	2.1559	2.5582	3.2470	3.9403	4.8652
11	1.1453	1.5868	1.8339	2.6032	3.0535	3.8157	4.5748	5.5778
12	1.4275	1.9344	2.2142	3.0738	3.5706	4.4038	5.2260	6.3038
13	1.7333	2.3051	2.6172	3.5650	4.1069	5.0088	5.8919	7.0415
14	2.0608	2.6967	3.0407	4.0747	4.6604	5.6287	6.5706	7.7895
15	2.4082	3.1075	3.4827	4.6009	5.2293	6.2621	7.2609	8.5468
16	2.7739	3.5358	3.9416	5.1422	5.8122	6.9077	7.9616	9.3122
17	3.1567	3.9802	4.4161	5.6972	6.4078	7.5642	8.6718	10.0852
18	3.5552	4.4394	4.9048	6.2648	7.0149	8.2307	9.3905	10.8649
19	3.9683	4.9123	5.4068	6.8440	7.627	8.9065	10.1170	11.6509
20	4.3952	5.3981	5.9210	7.4338	8.2604	9.5908	10.8508	12.4426
21	4.8348	5.8957	6.4467	8.0337	8.8972	10.2829	11.5913	13.2396
22	5.2865	6.4045	6.9830	8.6427	9.5425	10.9823	12.3380	14.0415
23	5.7494	6.9237	7.5292	9.2604	10.1957	11.6886	13.0905	14.8480
24	6.2230	7.4527	8.0849	9.8862	10.8564	12.4012	13.8484	15.6587
25	6.7066	7.9910	8.6493	10.5197	11.5240	13.1197	14.6114	16.4734
26	7.1998	8.5379	9.2221	11.1602	12.1981	13.8439	15.3792	17.2919
27	7.7019	9.0932	9.8028	11.8076	12.8785	14.5734	16.1514	18.1139
28	8.2126	9.6563	10.3909	12.4613	13.5647	15.3079	16.9279	18.9392
29	8.7315	10.2268	10.9861	13.1211	14.2565	16.0471	17.7084	19.7677
30	9.2581	10.8044	11.5880	13.7867	14.9535	16.7908	18.4927	20.5992
31	9.7921	11.3887	12.1963	14.4578	15.6555	17.5387	19.2806	21.4336
32	10.3331	11.9794	12.8107	15.1340	16.3622	18.2908	20.0719	22.2706
33	10.8810	12.5763	13.4309	15.8153	17.0735	19.0467	20.8665	23.1102
34	11.4352	13.1791	14.0567	16.5013	17.7891	19.8063	21.6643	23.9523
35	11.9957	13.7875	14.6878	17.1918	18.5089	20.5694	22.4650	24.7967

continued

TABLE C.3 (continued)Critical Values for the Chi-Square Distribution $\chi^2_{\alpha,\nu}$

v	α							
	.9999	.9995	.999	.995	.99	.975	.95	.90
36	12.5622	14.4012	15.3241	17.8867	19.2327	21.3359	23.2686	25.6433
37	13.1343	15.0202	15.9653	18.5858	19.9602	22.1056	24.0749	26.4921
38	13.7120	15.6441	16.6112	19.2889	20.6914	22.8785	24.8839	27.3430
39	14.2950	16.2729	17.2616	19.9959	21.4262	23.6543	25.6954	28.1958
40	14.8831	16.9062	17.9164	20.7065	22.1643	24.4330	26.5093	29.0505
41	15.48	17.54	18.58	21.42	22.91	25.21	27.33	29.91
42	16.07	18.19	19.24	22.14	23.65	26.00	28.14	30.77
43	16.68	18.83	19.91	22.86	24.40	26.79	28.96	31.63
44	17.28	19.48	20.58	23.58	25.15	27.57	29.79	32.49
45	17.89	20.14	21.25	24.31	25.90	28.37	30.61	33.35
46	18.51	20.79	21.93	25.04	26.66	29.16	31.44	34.22
47	19.13	21.46	22.61	25.77	27.42	29.96	32.27	35.08
48	19.75	22.12	23.29	26.51	28.18	30.75	33.10	35.95
49	20.38	22.79	23.98	27.25	28.94	31.55	33.93	36.82
50	21.01	23.46	24.67	27.99	29.71	32.36	34.76	37.69
60	27.50	30.34	31.74	35.53	37.48	40.48	43.19	46.46
70	34.26	37.47	39.04	43.28	45.44	48.76	51.74	55.33
80	41.24	44.79	46.52	51.17	53.54	57.15	60.39	64.28
90	48.41	52.28	54.16	59.20	61.75	65.65	69.13	73.29
100	55.72	59.90	61.92	67.33	70.06	74.22	77.93	82.36
200	134.02	140.66	143.84	152.24	156.43	162.73	168.28	174.84
300	217.33	225.89	229.96	240.66	245.97	253.91	260.88	269.07
400	303.26	313.43	318.26	330.90	337.16	346.48	354.64	364.21
500	390.85	402.45	407.95	422.30	429.39	439.94	449.15	459.93
600	479.64	492.52	498.62	514.53	522.37	534.02	544.18	556.06
700	569.32	583.39	590.05	607.38	615.91	628.58	639.61	652.50
800	659.72	674.89	682.07	700.73	709.90	723.51	735.36	749.19
900	750.70	766.91	774.57	794.47	804.25	818.76	831.37	846.07
1,000	842.17	859.36	867.48	888.56	898.91	914.26	927.59	943.13
1,500	1,304.80	1,326.30	1,336.42	1,362.67	1,375.53	1,394.56	1,411.06	1,430.25
2,000	1,773.30	1,798.42	1,810.24	1,840.85	1,855.82	1,877.95	1,897.12	1,919.39
2,500	2,245.54	2,273.86	2,287.17	2,321.62	2,338.45	2,363.31	2,384.84	2,409.82
3,000	2,720.44	2,751.65	2,766.32	2,804.23	2,822.75	2,850.08	2,873.74	2,901.17
3,500	3,197.36	3,231.23	3,247.14	3,288.25	3,308.31	3,337.93	3,363.53	3,393.22
4,000	3,675.88	3,712.22	3,729.29	3,773.37	3,794.87	3,826.60	3,854.03	3,885.81

TABLE C.3 (continued)Critical Values for the Chi-Square Distribution $\chi^2_{\alpha,v}$

v	α								
	.9999	.9995	.999	.995	.99	.975	.95	.90	
4,500	4,155.71	4,194.37	4,212.52	4,259.39	4,282.25	4,315.96	4,345.10	4,378.86	
5,000	4,636.62	4,677.48	4,696.67	4,746.17	4,770.31	4,805.90	4,836.66	4,872.28	
5,500	5,118.47	5,161.42	5,181.58	5,233.60	5,258.96	5,296.34	5,328.63	5,366.03	
6,000	5,601.13	5,646.08	5,667.17	5,721.59	5,748.11	5,787.20	5,820.96	5,860.05	
6,500	6,084.50	6,131.36	6,153.35	6,210.07	6,237.70	6,278.43	6,313.60	6,354.32	
7,000	6,568.49	6,617.20	6,640.05	6,698.98	6,727.69	6,769.99	6,806.52	6,848.80	
7,500	7,053.05	7,103.53	7,127.22	7,188.28	7,218.03	7,261.85	7,299.69	7,343.48	
8,000	7,538.11	7,590.32	7,614.81	7,677.94	7,708.68	7,753.98	7,793.08	7,838.33	
8,500	8,023.63	8,077.51	8,102.78	8,167.91	8,199.63	8,246.35	8,286.68	8,333.34	
9,000	8,509.57	8,565.07	8,591.09	8,658.17	8,690.83	8,738.94	8,780.46	8,828.50	
9,500	8,995.90	9,052.97	9,079.73	9,148.70	9,182.28	9,231.74	9,274.42	9,323.78	
1,000	9,482.59	9,541.19	9,568.67	9,639.48	9,673.95	9,724.72	9,768.53	9,819.19	

Critical Values for the Chi-Square Distribution $\chi^2_{\alpha,v}$

v	α								
	.10	.05	.025	.01	.005	.001	.0005	.0001	
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276	12.1157	15.1367	
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155	15.2018	18.4207	
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662	17.7300	21.1075	
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668	19.9974	23.5127	
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150	22.1053	25.7448	
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577	24.1028	27.8563	
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219	26.0178	29.8775	
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245	27.8680	31.8276	
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772	29.6658	33.7199	
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883	31.4198	35.5640	
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641	33.1366	37.3670	
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095	34.8213	39.1344	
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282	36.4778	40.8707	
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233	38.1094	42.5793	
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973	39.7188	44.2632	
16	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524	41.3081	45.9249	
17	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902	42.8792	47.5664	
18	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124	44.4338	49.1894	
19	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202	45.9731	50.7955	
20	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147	47.4985	52.3860	

continued

TABLE C.3 (continued)Critical Values for the Chi-Square Distribution $\chi^2_{\alpha,v}$

v	α							
	.10	.05	.025	.01	.005	.001	.0005	.0001
21	29.6151	32.6706	25.4789	38.9322	41.4011	46.7970	49.0108	53.9620
22	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679	50.5111	55.5246
23	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282	52.0002	57.0746
24	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786	53.4788	58.6130
25	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197	54.9475	60.1403
26	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520	56.4069	61.6573
27	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760	57.8576	63.1645
28	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923	59.3000	64.6624
29	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012	60.7346	66.1517
30	40.2560	43.7730	46.79792	50.8922	53.6720	59.7031	62.1619	67.6326
31	41.4217	44.9853	48.2319	52.1914	55.0027	61.0983	63.5820	69.1057
32	42.5847	46.1943	49.4804	53.4858	56.3281	62.4872	64.9955	70.5712
33	43.7452	47.3999	50.7251	54.7755	57.6484	63.8701	66.4025	72.0296
34	44.9032	48.6024	51.9660	56.0609	58.9639	65.2472	67.8035	73.4812
35	46.0588	49.8018	53.2033	57.3421	60.2748	66.6188	69.1986	74.9262
36	47.2122	50.9985	54.4373	58.6192	61.5812	67.9852	70.5881	73.3650
37	48.3634	52.1923	55.6680	59.8925	62.8833	69.3465	71.9722	77.7977
38	49.5126	53.3835	56.8955	61.1621	64.1814	70.7029	73.3512	79.2247
39	50.6598	54.5722	58.1201	62.4281	65.4756	72.0547	74.7253	80.6462
40	51.8051	55.7585	59.3417	63.6907	66.7660	73.4020	76.0946	82.0623
41	52.95	56.94	60.56	64.95	68.05	74.74	77.46	83.47
42	54.09	58.12	61.78	66.21	69.34	76.08	78.82	84.88
43	55.23	59.30	62.99	67.46	70.62	77.42	80.18	86.28
44	56.37	60.48	64.20	68.71	71.89	78.75	81.53	87.68
45	57.51	61.66	65.41	69.96	73.17	80.08	82.88	89.07
46	58.64	62.83	66.62	71.20	74.44	81.40	84.22	90.46
47	59.77	64.00	67.82	72.44	75.70	82.72	85.56	91.84
48	60.91	65.17	69.02	73.68	76.97	84.04	86.90	93.22
49	62.04	66.34	70.22	74.92	78.23	85.35	88.23	94.60
50	63.17	67.50	71.42	76.15	79.49	86.66	89.56	95.97
60	74.40	79.08	83.30	88.38	91.95	99.61	102.69	109.50
70	85.53	90.53	95.02	100.43	104.21	112.32	115.58	122.75
80	96.58	101.88	106.63	112.33	116.32	124.84	128.26	135.78
90	107.57	113.15	118.14	124.12	128.30	137.21	140.78	148.63
100	118.50	124.34	129.56	135.81	140.17	149.45	153.17	161.32

TABLE C.3 (continued)Critical Values for the Chi-Square Distribution $\chi^2_{\alpha,v}$

v	α							
	.10	.05	.025	.01	.005	.001	.0005	.0001
200	226.02	233.99	241.06	249.45	255.26	267.54	272.42	283.06
300	331.79	341.40	349.87	359.91	366.84	381.43	387.20	399.76
400	436.65	447.63	457.31	468.72	476.61	493.13	499.67	513.84
500	540.93	553.13	563.85	576.49	585.21	603.45	610.65	626.24
600	644.80	658.09	669.77	583.52	692.98	712.77	720.58	737.46
700	748.36	762.66	775.21	789.97	800.13	821.35	829.71	847.78
800	851.67	866.91	880.28	895.98	906.79	929.33	938.21	957.38
900	954.78	970.90	985.03	1,001.63	1,013.04	1,036.83	1,046.19	1,066.40
1,000	1,057.72	1,074.68	1,089.53	1,106.97	1,118.95	1,143.92	1,153.74	1,174.93
1,500	1,570.61	1,609.23	1,630.35	1,644.84	1,674.97	1,686.81	1,712.30	
2,000	2,081.47	2,105.15	2,125.84	2,150.07	2,166.16	2,214.68	2,243.81	
2,500	2,591.04	2,617.43	2,640.47	2,667.43	2,685.89	2,724.22	2,739.25	2,771.57
3,000	3,099.69	3,128.54	3,153.70	3,183.13	3,203.28	3,245.08	3,261.45	3,296.66
3,500	3,607.64	3,638.75	3,665.87	3,697.57	3,719.26	3,764.26	3,781.87	3,819.74
4,000	4,115.05	4,148.25	4,177.19	4,211.01	4,234.14	4,282.11	4,300.88	4,341.22
4,500	4,622.00	4,657.17	4,689.78	4,723.63	4,748.1	4,798.87	4,818.73	4,861.40
5,000	5,128.58	5,165.61	5,197.88	5,235.57	5,261.34	5,314.73	5,335.62	5,380.48
5,500	5,634.83	5,673.64	5,707.45	5,746.93	5,773.91	5,829.81	5,851.68	5,898.63
6,000	6,140.81	6,181.31	6,216.59	6,257.78	6,285.92	6,344.23	6,367.02	6,415.98
6,500	6,646.54	6,688.67	6,725.36	6,768.18	6,797.45	6,858.05	6,881.74	6,932.61
7,000	7,152.06	7,195.75	7,233.79	7,278.19	7,308.53	7,371.35	7,395.90	7,448.62
7,500	7,657.38	7,702.58	7,741.93	7,787.86	7,819.23	7,884.18	7,909.57	7,964.06
8,000	8,162.53	8,209.19	8,249.81	8,297.20	8,329.58	8,396.59	8,422.78	8,479.00
8,500	8,667.52	8,715.59	8,757.44	8,806.26	8,839.60	8,908.62	8,935.59	8,993.48
9,000	9,172.36	9,221.81	9,264.85	9,315.05	9,349.34	9,420.30	9,448.03	9,507.53
9,500	9,677.07	9,727.86	9,772.05	9,823.60	9,858.81	9,931.67	9,960.13	1,0021.21
10,000	1,0181.66	1,0233.75	1,0279.07	1,0331.93	1,0368.03	1,0442.73	1,0471.91	1,0534.52

TABLE C.4
Critical Values for the F Distribution

For given values of v_1 and v_2 , this table contains values of $F_{0.1;v_1,v_2}$; defined by $\text{Prob}[F \geq F_{0.1;v_1,v_2}] = \alpha = 0.1$

		v_1												
		1	2	3	4	5	6	7	8	9	10	50	100	∞
v_2		1	2	3	4	5	6	7	8	9	10	50	100	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	62.69	63.01	63.33	
2	28.53	9.00	9.16	9.24	9.33	9.35	9.37	9.38	9.39	9.47	9.48	9.49		
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.15	5.14	5.13	
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.80	3.78	3.76	
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.15	3.13	3.10	
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.77	2.75	2.72	
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.52	2.50	2.47	
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.35	2.32	2.29	
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.22	2.19	2.16	
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.12	2.09	2.06	
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.04	2.01	1.97	
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	1.87	1.94	1.90	
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	1.92	1.88	1.85	
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	1.87	1.83	1.80	
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.83	1.79	1.76	
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.79	1.76	1.72	
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.76	1.73	1.69	
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.74	1.70	1.66	
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.71	1.67	1.63	
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.69	1.65	1.61	
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.61	1.56	1.52	
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.44	1.39	1.34	
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.35	1.29	1.20	
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.71	1.67	1.63	1.60	1.24	1.17		

For given values of v_1 and v_2 , this table contains values of $F_{0.05, v_1, v_2}$ defined by $\text{Prob}[F \geq F_{0.05, v_1, v_2}] = \alpha = 0.05$

v_2	v_1										50	100	∞
	1	2	3	4	5	6	7	8	9	10	50	100	∞
1	161.4	199.5	215.7	224.6	230.2	236.8	238.9	240.5	241.9	251.8	253.0	254.3	
2	18.51	19.00	19.16	19.25	19.30	19.35	19.37	19.40	19.48	19.49	19.50	19.50	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.81	8.79	8.58	8.55	8.53	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.70	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.44	4.41	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.75	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.32	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.02	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.80	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.64	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.51	2.46	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.40	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.31	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.24	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.18	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.12	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.08	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.04	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.00	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	1.97	1.91	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	1.84	1.78	1.71
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.60	1.52	1.45
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.48	1.39	1.28
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.35	1.25	1.00

continued

TABLE C.4 (continued)Critical Values for the F DistributionFor given values of v_1 and v_2 , this table contains values of $F \geq F_{0.025, v_1, v_2}$ defined by $\text{Prob}[F \geq F_{0.025, v_1, v_2}] = \alpha = 0.025$

v_2	v_1										50	100	∞
	1	2	3	4	5	6	7	8	9	10	50	100	
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	1008	1013	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.01	13.96	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.38	8.32	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.14	6.08	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	4.98	4.92	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.28	4.21	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	3.81	3.74	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.47	3.40	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.22	3.15	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.03	2.96	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	2.87	2.80	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	2.74	2.67	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.64	2.56	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.55	2.47	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.47	2.40	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.41	2.33	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.35	2.27	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.30	2.22	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.25	2.17	2.09
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.08	2.00	1.91
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	1.75	1.66	1.54
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.59	1.48	1.37
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.43	1.27	1.00

For given values of v_1 and v_2 , this table contains values of $F_{0.01, v_1, v_2}$; defined by $\text{Prob}[F \geq F_{0.01, v_1, v_2}] = \alpha = 0.01$

v_2	v_1									10	50	100	∞
	1	2	3	4	5	6	7	8	9	10	50	100	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,981	6,022	6,056	6,303	6,334	6,336
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.35	26.24	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	13.69	13.58	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.24	9.13	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.09	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	5.86	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.07	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.52	4.41	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.12	4.01	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	3.81	3.71	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	3.57	3.47	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.38	3.27	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.22	3.11	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.08	2.98	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	2.97	2.86	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	2.87	2.76	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	2.78	2.68	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	2.71	2.60	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	2.64	2.54	2.42
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.40	2.29	2.17
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	1.95	1.82	1.70
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	1.74	1.60	1.45
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	1.53	1.32	1.00

continued

TABLE C.4 (continued)Critical Values for the F DistributionFor given values of v_1 and v_2 , this table contains values of $F_{0.005, v_1, v_2}$ defined by $\text{Prob}[F \geq F_{0.005, v_1, v_2}] = \alpha = 0.005$

v_2	v_1	1	2	3	4	5	6	7	8	9	10	50	100	∞
1	16.211	20,000	21.615	22.500	23.056	23.437	23.715	23.925	24.091	24.224	25.211	25.337	25.465	
2	198.5	199.0	199.2	199.3	199.3	199.3	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	42.21	42.02	41.83	
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	19.67	19.50	19.32	
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	12.45	12.30	12.14	
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	9.17	9.03	8.88	
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	7.35	7.22	7.08	
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	6.22	6.09	5.95	
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	5.45	5.32	5.19	
10	12.83	9.43	8.08	7.34	6.97	6.54	6.30	6.12	5.97	5.85	4.90	4.77	4.64	
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	4.49	4.36	4.23	
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.17	4.04	3.90	
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	3.91	3.78	3.65	
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	3.70	3.57	3.44	
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	3.52	3.39	3.26	
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	3.37	3.25	3.11	
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.25	3.12	2.98	
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.14	3.01	2.87	
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.04	2.91	2.78	
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	2.96	2.83	2.69	
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	2.65	2.52	2.38	
50	8.63	5.90	4.83	4.23	3.85	3.58	3.38	3.22	3.09	2.99	2.10	1.95	1.81	
100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	1.84	1.68	1.51	
∞	7.88	5.30	4.28	3.762	3.35	3.09	2.90	2.74	2.62	2.52	1.60	1.36		

For given values of v_1 and v_2 , this table contains values of $F_{0.001, v_1, v_2}$ defined by $\text{Prob}[F \geq F_{0.001, v_1, v_2}] = \alpha = 0.001$

v_2	v_1										50	100	∞
	1	2	3	4	5	6	7	8	9	10	50	100	∞
2	998.5	999.0	999.2	999.3	999.4	999.4	999.4	999.4	999.4	999.4	999.5	999.5	999.5
3	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2	124.7	124.1	123.5
4	74.14	61.25	45.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	44.88	44.47	44.05
5	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	24.44	24.12	23.79
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	16.31	16.03	15.75
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	12.20	11.95	11.70
8	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	9.80	9.57	9.33
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	8.26	8.04	7.81
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	7.19	6.98	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	6.42	6.21	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	5.83	5.63	5.42
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	5.37	5.17	4.97
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	5.00	4.81	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	4.70	4.51	4.31
16	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	4.45	4.26	4.06
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	4.24	4.05	3.85
18	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	4.06	3.87	3.67
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	3.90	3.71	3.51
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	3.77	3.58	3.38
25	13.88	9.22	7.45	5.89	5.46	5.15	4.91	4.71	4.56	3.28	3.09	2.89	
50	12.22	7.96	5.46	4.90	4.51	4.22	4.00	3.82	3.67	2.44	2.25	2.06	
100	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	2.08	1.87	1.65
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	1.75	1.45	1.00

TABLE C.5Durbin-Watson Statistic Significance Points of d_L and d_U : 5%

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
	d_L	d_u								
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Source: Durbin, J. and Watson, G. S., *Biometrika*, 38, 159–178, 1951. By permission of Oxford University Press.

TABLE C.6Durbin-Watson Statistic Significance Points of d_L and d_U : 2.5%

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
	d_L	d_u								
15	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
16	0.98	1.24	0.86	1.40	0.75	1.59	0.64	1.80	0.53	2.03
17	1.01	1.25	0.90	1.40	0.79	1.58	0.68	1.77	0.57	1.98
18	1.03	1.26	0.93	1.40	0.82	1.56	0.72	1.74	0.62	1.93
19	1.06	1.28	0.96	1.41	0.86	1.55	0.76	1.72	0.66	1.90
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
21	1.10	1.30	1.01	1.41	0.92	1.54	0.83	1.69	0.73	1.84
22	1.12	1.31	1.04	1.42	0.95	1.54	0.86	1.68	0.77	1.82
23	1.14	1.32	1.06	1.42	0.97	1.54	0.89	1.67	0.80	1.80
24	1.16	1.33	1.08	1.43	1.00	1.54	0.91	1.66	0.83	1.79
25	1.18	1.34	1.10	1.43	1.02	1.54	0.94	1.65	0.86	1.77
26	1.19	1.35	1.12	1.44	1.04	1.54	0.96	1.65	0.88	1.76
27	1.21	1.36	1.13	1.44	1.06	1.54	0.99	1.64	0.91	1.75
28	1.22	1.37	1.15	1.45	1.08	1.54	1.01	1.64	0.93	1.74
29	1.24	1.38	1.17	1.45	1.10	1.54	1.03	1.63	0.96	1.73
30	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	0.98	1.73
31	1.26	1.39	1.20	1.47	1.13	1.55	1.07	1.63	1.00	1.72
32	1.27	1.40	1.21	1.47	1.15	1.55	1.08	1.63	1.02	1.71
33	1.28	1.41	1.22	1.48	1.16	1.55	1.10	1.63	1.04	1.71
34	1.29	1.41	1.24	1.48	1.17	1.55	1.12	1.63	1.06	1.70
35	1.30	1.42	1.25	1.48	1.19	1.55	1.13	1.63	1.07	1.70
36	1.31	1.43	1.26	1.49	1.20	1.56	1.15	1.63	1.09	1.70
37	1.32	1.43	1.27	1.49	1.21	1.56	1.16	1.62	1.10	1.70
38	1.33	1.44	1.28	1.50	1.23	1.56	1.17	1.62	1.12	1.70
39	1.34	1.44	1.29	1.50	1.24	1.56	1.19	1.63	1.13	1.69
40	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
45	1.39	1.48	1.34	1.53	1.30	1.58	1.25	1.63	1.21	1.69
50	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
55	1.45	1.52	1.41	1.56	1.37	1.60	1.33	1.64	1.30	1.69
60	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
65	1.49	1.55	1.46	1.59	1.43	1.62	1.40	1.66	1.36	1.69
70	1.51	1.57	1.48	1.60	1.45	1.63	1.42	1.66	1.39	1.70
75	1.53	1.58	1.50	1.61	1.47	1.64	1.45	1.67	1.42	1.70
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
85	1.56	1.60	1.53	1.63	1.51	1.65	1.49	1.68	1.46	1.71
90	1.57	1.61	1.55	1.64	1.53	1.66	1.50	1.69	1.48	1.71
95	1.58	1.62	1.56	1.65	1.54	1.67	1.52	1.69	1.50	1.71
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72

Source: Durbin, J. and Watson, G. S., *Biometrika*, 38, 159–178, 1951. By permission of Oxford University Press.

TABLE C.7Durbin–Watson Statistic Significance Points of d_L and d_U : 1%

n	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
	d_L	d_U								
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Source: Durbin, J. and Watson, G. S., *Biometrika*, 38, 159–178, 1951. By permission of Oxford University Press.

Appendix D

Variable Transformations

D.1 Purpose of Variable Transformations

Transformations are used to present or translate data to different scales. In modeling applications, transformations are often used to improve the compatibility of data with assumptions underlying a modeling process, to linearize a relationship between two variables whose relationship is nonlinear, or to modify the range of values of a variable. Both dependent and independent variables can be transformed.

Although transformations may result in an improvement of specific modeling assumptions, such as linearity or homoscedasticity, they can result in the violation of other assumptions. Thus, transformations must be used in an iterative fashion, with close monitoring of other modeling assumptions as transformations are made.

A difficulty arises when transforming the dependent variable Y because the variable is expressed in a form that may not be of interest to the investigation, such as the log of Y , the square root of Y , or the inverse of Y . When comparing statistical models, the comparisons should always be made on the original untransformed scale of Y . These comparisons extend to goodness-of-fit statistics and model validation exercises. An example in transportation of the necessity to compare models in original units is described in Fomunung et al. (2000).

Transformations not only reflect assumptions about the underlying relation between variables, but also the underlying disturbance structure of the model. For example, exponential transformations imply a multiplicative disturbance structure of the underlying model, and not an additive disturbance structure that is assumed in linear regression. For example, when the underlying function $Y = \alpha \text{EXP}(\beta X) + \varepsilon$ is suspected, a log transformation gives $\text{LN}(Y) = \text{LN}(\alpha \text{EXP}(\beta X) + \varepsilon) = \text{LN}[(\alpha \text{EXP}(\beta X))(1 + \varepsilon/\alpha \text{EXP}(\beta X))] = \text{LN}(\alpha) + \beta X + \text{LN}(1 + \varepsilon/\alpha \text{EXP}(\beta X))$. Although the model is linear, the disturbance term is not the one specified in ordinary least squares regression because it is a function of X , α , and β , and thus is multiplicative. It is important that disturbance terms always be checked after transformations to make sure they are still compatible with

modeling assumptions. Guidelines for applying transformations include the following:

1. Transformations on a dependent variable will change the distribution of disturbance terms in a model. Thus, incompatibility of model disturbances with an assumed distribution can sometimes be remedied with transformations of the dependent variable.
2. Nonlinearities between the dependent variable and an independent variable are often linearized by transforming the independent variable. Transformations of an independent variable do not change the distribution of disturbance terms.
3. When a relationship between a dependent and independent variable requires extensive transformations to meet linearity and disturbance distribution requirements, often there are alternative methods for estimating the parameters of the relation, such as nonlinear and generalized least squares regression.
4. Confidence intervals computed on transformed variables need to be computed by transforming back to the original units of interest.

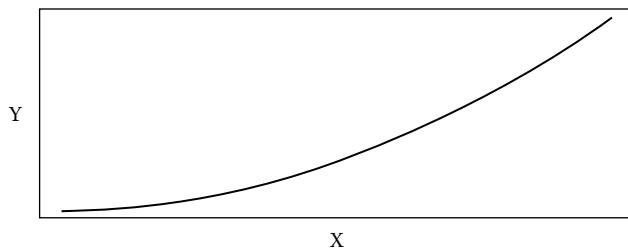
D.2 Commonly Used Variable Transformations

Taken in the context of modeling the relationship between a dependent variable Y and independent variable X , there are several motivations for transforming a variable or variables. It should be noted that most transformations attempt to make the relation between Y and X linear because linear relationships are easier to model. Transformations done to Y and X in their originally measured units are merely done for convenience of the modeler, and not because of an underlying problem with the data.

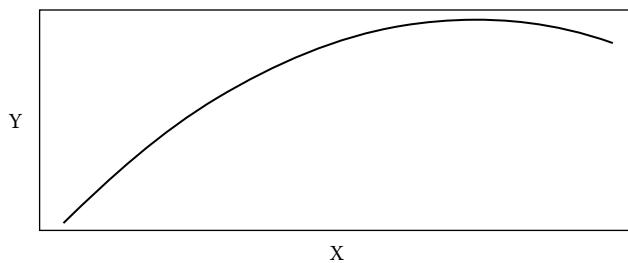
The following figures show common transformations used to linearize a relationship between two random variables, X and Y . Provided are plots of the relationships between X and Y in their untransformed states, and then some examples of transformations on X , Y , or both that are used to linearize the relation.

D.2.1 Parabolic Transformations

Parabolic transformations are used to linearize a nonlinear or curvilinear relation. The parabolic transformation is used when the true relation between Y and X is given as $Y = \alpha + \beta X + \gamma X^2$. The transformation is done by simply adding a squared or quadratic term to the right-hand side of the equation, which is really more than a mere transformation.

**FIGURE D.1**

Parabolic relation (functional form: $Y = \alpha + \beta X + \gamma X^2$, where $\alpha > 0, \beta > 0, \gamma > 0$).

**FIGURE D.2**

Parabolic relation (functional form: $Y = \alpha + \beta X + \gamma X^2$, where $\alpha > 0, \beta > 0, \gamma < 0$).

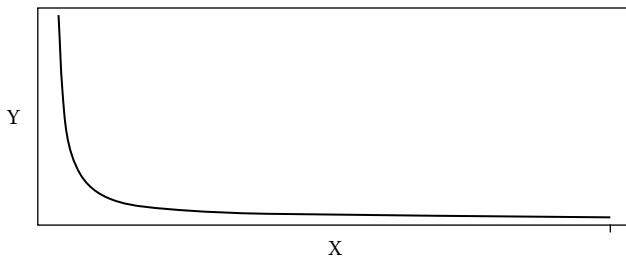
The nature of the relationship depends on the values of α , β , and γ . Figure D.1 shows an example of a relation between Y and X when all parameters are positive, and Figure D.2 shows an example of the relation between Y and X when α and β are positive and γ is negative.

D.2.2 Hyperbolic Transformations

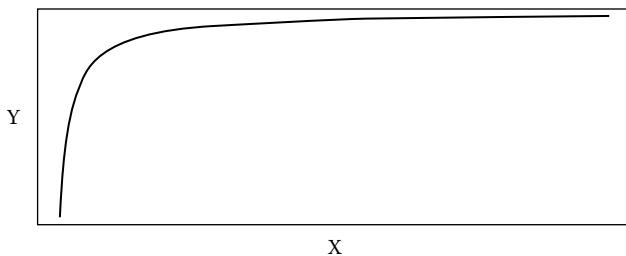
Hyperbolic transformations are used to linearize a variety of curvilinear shapes. A hyperbolic transformation is used when the true relation between Y and X is given as $Y = X/(\alpha + \beta X)$, as shown in Figures D.3 and D.4. By transforming both Y and X using the inverse transformation, one can generate a linear relationship such that $1/Y = \beta_0 + \beta_1(1/X) + \text{disturbance}$. In this transformation, $\alpha = \beta_1$ and $\beta = \beta_0$. Figure D.3 shows an example of a relation between Y and X when α is positive, whereas Figure D.4 shows an example of the relation between Y and X when α is negative.

D.2.3 Exponential Functions

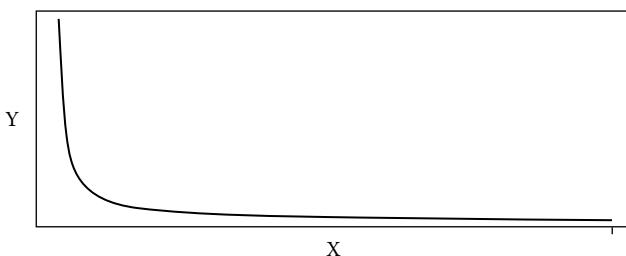
The natural log transformation is used to correct heterogeneous variance in some cases, and when the data exhibit curvature between Y and X of a certain

**FIGURE D.3**

Hyperbolic relation (functional form: $Y = X/(\alpha + \beta X)$, where $\alpha > 0$).

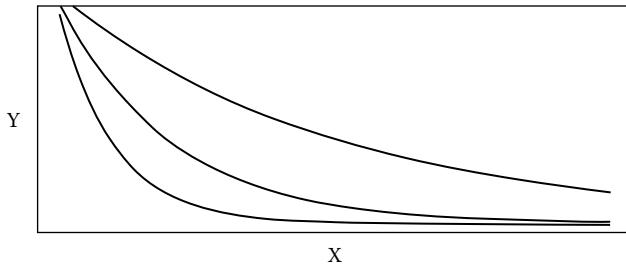
**FIGURE D.4**

Hyperbolic relation (functional form: $Y = X/(\alpha + \beta X)$, where $\alpha < 0$).

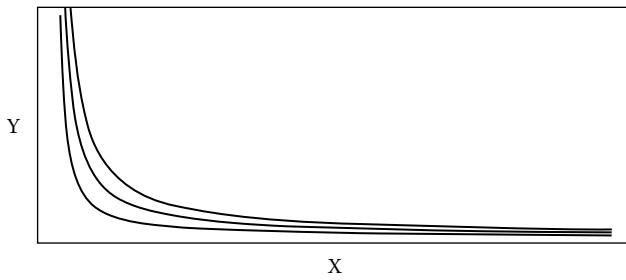
**FIGURE D.5**

Exponential relation (functional form: $Y = \alpha EXP(\beta X)$, where $\beta > 0$).

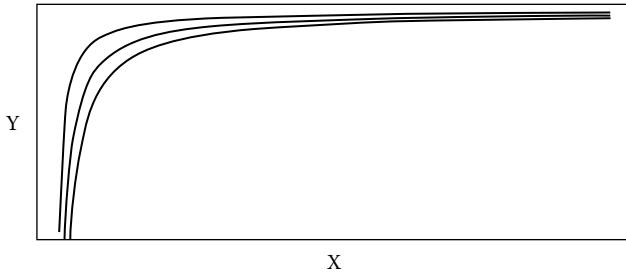
type. Figures D.5 and D.6 show the nature of the relationship between Y and X for data that are linearized using the log transformation. The nature of the underlying relation is $Y = \alpha EXP(\beta X)$, where α and β are parameters of the relation. To get this relation in linear model form, one transforms both sides of the equation to obtain $LN(Y) = LN(\alpha EXP(\beta X)) = LN(\alpha) + LN(EXP(\beta X)) = LN(\alpha) + \beta X = \beta_0 + \beta_1 X$. In linearized form $\beta_0 = LN(\alpha)$ and $\beta_1 = \beta$. Figure D.5 shows examples of the relation between Y and X for $\beta > 0$, and Figure D.6 shows examples of the relation between Y and X for $\beta < 0$.

**FIGURE D.6**

Exponential functions (functional form: $Y = \alpha \text{EXP}(\beta X)$, where $\beta < 0$).

**FIGURE D.7**

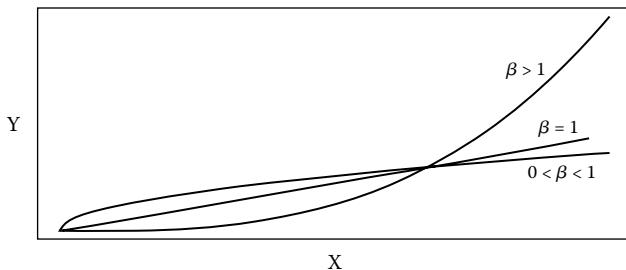
Inverse exponential functions (functional form: $Y = \alpha \text{EXP}(\beta/X)$, where $\beta > 0$).

**FIGURE D.8**

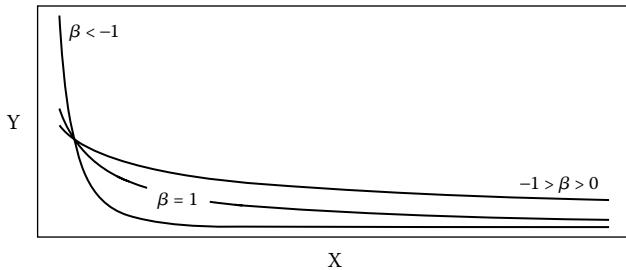
Inverse exponential functions (functional form: $Y = \alpha \text{EXP}(\beta/X)$, where $\beta < 0$).

D.2.4 Inverse Exponential Functions

Sometimes the exponential portion of the mechanism is proportional to the inverse of X instead of untransformed X . The underlying relationship, which is fundamentally different from the exponential relations shown in Figures D.5 and D.6, is given by $Y = \alpha \text{EXP}(\beta/X)$. By taking the log of both sides of this relationship one obtains the linear model form of the relation,

**FIGURE D.9**

Power functions (functional form: $Y = \alpha X^\beta$, where $\beta > 0$).

**FIGURE D.10**

Power functions (functional form: $Y = \alpha X^\beta$, where $\beta < 0$).

$LN(Y) = LN(\alpha) + \beta/X = \beta_0 + 1/\beta_1 X$. Figure D.7 shows examples of the inverse exponential relation when β is positive, and Figure D.8 shows examples when β is negative.

D.2.5 Power Functions

Power transformations are needed when the underlying structure is of the form $Y = \alpha X^\beta$ and transformations on both variables are needed to linearize the function. The linear form of the power function is $LN(Y) = LN(\alpha X^\beta) = LN(\alpha) + \beta LN(X) = \beta_0 + \beta_1 LN(X)$. The shape of the power function depends on the sign and magnitude of β . Figure D.9 depicts examples of power functions with β greater than zero, and Figure D.10 depicts examples of power functions with β less than zero.

References

- Abbot, R. (1985). Logistic regression in survival analysis. *American Journal of Epidemiology* 121, 465–471.
- Abraham, B. and Ledolter, T. (1983). *Statistical Methods for Forecasting*. John Wiley & Sons, NY.
- Aczel, A. (1993). *Complete Business Statistics*. Irwin, Homewood, IL.
- Aigner, D., Lovell, K., Schmidt, P. (1977). Formulation and estimation of stochastic frontier production models. *Journal of Econometrics* 6, 21–37.
- Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. 2nd International Symposium of Information Theory. Tsahkadsor, 1971, 267–281.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* 52, 317–332.
- Amemiya, T. (1971). The estimation of the variances in a variance-components model. *International Economic Review* 12, 1–13.
- Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, MA.
- Anastasopoulos, P. and Mannering, F. (2009). A note on modeling vehicle-accident frequencies with random-parameters count models. *Accident Analysis and Prevention* 41, 153–159.
- Anastasopoulos, P., Tarko, A., Mannering, F. (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis and Prevention* 40, 768–775.
- Anderson, O. (1976). *Time Series Analysis and Forecasting*. London: Butterworths.
- Ansari, A. and Bradley, R. (1960). Rank sum tests for dispersion. *Annals of Mathematical Statistics* 31, 1174–1189.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Boston, MA.
- Arbuckle, J. and Wotheke, W. (1995). *AMOS Users' Guide*. Version 4.0. SmallWaters Corporation, Chicago, IL.
- Arbuthnott, I. (1910). An argument for Divine Providence taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions* 27, 186–190.
- Archilla, A. and Madanat, S. (2000). Development of a pavement rutting model from experimental data. *ASCE Journal of Transportation Engineering* 126, 291–299.
- Arellano, M. (1993). On the testing of correlated effects with panel data. *Journal of Econometrics* 59, 87–97.
- Arminger, G., Clogg, C., Sobel, M. (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum Press, NY.
- Ash, C. (1993). *The Probability Tutoring Book*. IEEE Press, NY.
- Avery, R. (1977). Error components and seemingly unrelated regressions. *Econometrica* 45, 199–209.
- Ayyub, B. (1998). *Uncertainty Modeling and Analysis in Civil Engineering*. CRC Press, Boca Raton, FL.
- Baltagi, B. (1985). *Econometric Analysis of Panel Data*. John Wiley & Sons, NY.

- Baltagi, B. (1998). *Econometrics*. Springer, NY.
- Baltagi, B. (2008). *Econometric Analysis of Panel Data*, 4th Edition, Wiley.
- Baltagi, B. and Griffin, J. (1988). A generalized error component model with heteroscedastic disturbances. *International Economic Review* 29, 745–753.
- Baltagi, B. and Li, Q. (1991). A transformation that will circumvent the problem of autocorrelation in an error component model. *Journal of Econometrics* 48, 385–393.
- Baltagi, B. and Li, Q. (1992). A monotonic property for iterative GLS in the two-way random effects model. *Journal of Econometrics* 53, 45–51.
- Bartels, R. (1982). The rank version of von Neumann's ratio test for randomness. *Journal of the American Statistical Association* 77, 40–46.
- Beach, C. and McKinnon, J. (1978). A maximum likelihood procedure for regression with autocorrelated errors. *Econometrics* 46, 51–58.
- Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis*. MIT Press, Cambridge, MA.
- Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin* 107, 238–246.
- Bentler, P. and Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88, 588–606.
- Bentler, P. and Weeks, D. (1980). Linear structural equations with latent variables. *Psychometrika* 45, 289–307.
- Berkovec, J. (1985). New car sales and used car stocks: a model of the automobile market. *Rand Journal of Economics* 16, 195–214.
- Bhat, C. (1996a). A hazard-based duration model of shopping activity with non-parametric baseline specification and non-parametric control for unobserved heterogeneity. *Transportation Research Part B* 30, 189–207.
- Bhat, C. (1996b). A generalized multiple durations proportional hazard model with an application to activity behavior during the evening work-to-home commute. *Transportation Research Part B* 30, 465–480.
- Bhat, C. (2000). Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling, *Transportation Research Part A* 32, 495–507.
- Bhat, C. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B* 37, 837–855.
- Bhat, C. (2005). A multiple discrete-continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B* 39, 679–707.
- Bhat, C. (2008). The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B* 42, 274–303.
- Birnbaum, Z. and Hall, R. (1960). Small sample distributions for multi-sample statistics of the Smirnov type. *Annals of Mathematical Statistics* 22, 592–596.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, NY.
- Boersch-Supan, A. and Hajivassiliou, V. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* 58, 347–368.
- Bollen, K. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika* 51, 375–377.
- Bollen, K. and Stine, R. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research* 21, 205–229.

- Bollen, K. and Long, J. Eds. (1993). *Testing Structural Equation Models*. Sage, Newbury Park, CA.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31, 307–327.
- Box, G. and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA.
- Brannas, K. (1994). Estimation and testing in Integer-valued AR(1) models. *Umeå Economic Studies*, Paper Number 335 (revised).
- Brannas, K. (1995). Explanatory variables in the AR(1) count data model. *Umeå Economic Studies*, Paper Number 381.
- Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
- Breusch, T. (1987). Maximum likelihood estimation of random effects models. *Journal of Econometrics* 36, 383–389.
- Breusch, T. and Pagan, A. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–1294.
- Brijs, T., Karlis, D., Wets, G. (2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention* 40, 1180–1190.
- Brooks, S. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7, 434–455.
- Brown, G. and Mood, A. (1951). On median tests for linear hypotheses. In *Proc. 2nd Berkeley Symp. on Mathematical Statistics and Probability*, J. Neyman, Ed. University of California Press, Berkeley. 159–166.
- Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematics and Statistical Psychology* 37, 62–83.
- Browne, M. and Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research* 24, 445–455.
- Browne, M. and Cudeck, R. (1993). Alternative ways of assessing model fit. In *Testing Structural Equation Models*, K. Bollen and J. Long, Eds. Sage, Newbury Park, CA, 136–162.
- Brownstone, D. and Train, K. (1999). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics* 89, 109–129.
- Byrne, B. (1989). *A Primer of LISREL: Basic Applications and Programming for Confirmatory Factor Analytic Methods*. Springer-Verlag, NY.
- Cameron, C. and Windmeijer, F. (1993). R-squared measures for count data regression models with applications to health care utilization. Working Paper 93–24, Department of Economics, University of California, Davis.
- Capon, J. (1961). Asymptotic efficiency of certain locally most powerful rank tests. *Annals of Mathematical Statistics* 32, 88–100.
- Carlin, B. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, NY.
- Carmines, E. and McIver, J. (1981). Analyzing models with unobserved variables. In *Social Measurement: Current Issues*, G. Bohrnstedt and E. Borgatta, Eds. Sage, Beverly Hills, CA.
- Carson, J. and Mannering, F. (2001). The effects of ice warning signs on accident frequencies and severities. *Accident Analysis and Prevention* 33, 99–109.
- Cattell, R. and Burdsal, C. (1975). The radial parcel double factoring design: a solution to the item-vs.-parcel controversy. *Multivariate Behavioral Research* 10, 165–179.

- Chamberlain, G. (1978). Omitted variable bias in panel data: estimating the returns to schooling. *Annales de l'INSEE* 30(1), 49–82.
- Chatfield, C. (1992). A commentary on errors measures. *International Journal of Forecasting* 8, 100–102.
- Chen, R. and Tsay, R. (1993a). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* 88, 298–308.
- Chen, R. and Tsay, R. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association* 88, 955–967.
- Cheng, B. and Titterington, D. (1994). Neural networks: a review from a statistical perspective. *Statistical Science* 9, 2–54.
- Chou, C., Bentler, P., Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology* 44, 347–357.
- Chu, C. and Durango-Cohen, P. (2008). Estimation of dynamic performance models for transportation infrastructure using panel data. *Transportation Research Part B: Methodological* 42(1), 57–81.
- Clemens, M. and Hendry, D. (1993). On the limitations of the comparison mean square forecast errors. *Journal of Forecasting* 12, 615–637.
- Cochran, W. (1950). The comparison of percentages in matched samples. *Biometrika* 37, 256–266.
- Cochrane, D. and Orcutt, G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* 44, 32–61.
- Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley & Sons, NY.
- Congdon, P. (2003). *Applied Bayesian Statistical Models*. John Wiley & Sons, NY.
- Conover, W. (1980). *Practical Nonparametric Statistics*, 2nd ed. John Wiley & Sons, NY.
- Cosslett, S. (1981). Efficient estimation of discrete-choice models. In *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski and D. McFadden, Eds. MIT Press, Cambridge, MA.
- Cox, D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B* 34, 187–200.
- Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC, Boca Raton, FL.
- Cox, D. and Stuart, A. (1955). Some quick tests for trend in location and dispersion. *Biometrika* 42, 80–95.
- Crotte, A., Noland, R., Graham, D. (2009). Is the Mexico City metro an inferior good? *Transport Policy* 16, 40–45.
- Cryer, J. (1986). *Time Series Analysis*. Duxbury Press, Boston, MA.
- Daganzo, C. (1979). *Multinomial Probit: The Theory and its Application to Demand Forecasting*. Academic Press, NY.
- Damm, D. and Lerman, S. (1981). A theory of activity scheduling behavior. *Environment and Planning* 13A, 703–718.
- Daniel, C. and Wood, F. (1980). *Fitting Equations to Data*. John Wiley & Sons, NY.
- Daniel, W. (1978). *Applied Nonparametric Statistics*. Houghton Mifflin, Boston, MA.
- Daniels, H. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society B* 12, 171–181.
- Dargay, J. and Hanly, M. (2002). The demand for local bus services in England. *Journal of Transport Economics and Policy* 36, 79–91.

- David, F. and Barton, D. (1958). A test for birth-order effects. *Annals of Human Eugenics* 22, 250–257.
- Dee, G. (1999). State alcohol policies, teen drinking, and traffic fatalities. *Journal of Public Economics* 72, 289–315.
- De Groot, C. and Wurtz, D. (1991). Analysis of univariate time series with connectionist nets: a case study of two classical examples. *Neurocomputing* 3, 177–192.
- de Veaux, R. (1990). Finding transformations for regression using the ACE algorithm. In *Modern Methods of Data Analysis*, J. Fox and J.S. Long, Eds. Sage, Newbury Park, CA, 177–208.
- Derrig, R., Segui-Gomez, M., Abtahi, A., Liu, L. (2002). The effect of population safety belt usage rates on motor vehicle-related fatalities. *Accident Analysis and Prevention* 34, 101–110.
- Devore, J. and Farnum, N. (1999). *Applied Statistics for Engineers and Scientists*. Duxbury Press, NY.
- Diamond, P. and Hausman, J. (1984). The retirement and unemployment behavior of older men. In *Retirement and Economic Behavior*, H. Aaron and G. Burtless, Eds. Brookings Institute, Washington, DC.
- Dougherty, M. (1995). A review of neural networks applied to transport. *Transportation Research Part C* 3, 247–260.
- Dubin, J. and McFadden, D. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52, 345–362.
- Durbin, J. (1951). Incomplete blocks in ranking experiments. *British Journal of Psychology*. (Stat. Sec.) 4, 85–90.
- Durbin, J. (1960). Estimation of parameters in time-series regression model. *Journal of the Royal Statistical Society Series B* 22, 139–153.
- Durbin, J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrics* 38, 410–421.
- Durbin, J. and Watson, G. (1951). Testing for serial correlation in least squares regression-II. *Biometrika* 37, 159–178.
- Edgington, E. (1961). Probability table for number of runs of signs or first differences in ordered series. *Journal of the American Statistical Association* 56, 156–159.
- Efron, B. (1977). Efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 72, 557–565.
- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science* 1, 54–74.
- Emerson, J. and Stoto, M. (1983). Transforming data. In *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller, and J. Tukey, Eds. John Wiley & Sons, NY, 97–127.
- Eluru, N., Bhat, C., and Hensher, D. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention* 40, 1033–1054.
- Enders, W. (2004). *Applied Econometric Time Series*, 2nd ed. John Wiley & Sons, NY.
- Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of UK inflation. *Econometrica* 50, 987–1008.
- Engle, R. and Russell, J. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* 66, 1127–1162.
- Erdil, E. and Yetkiner, I. (2009). The Granger-causality between health care expenditure and output: a panel data approach. *Applied Economics* 41, 511–518.

- Eskelanda, G. and Feyziolub, T. (1997). Is demand for polluting goods manageable? An econometric study of car ownership and use in Mexico. *Journal of Development Economics* 53, 423–445.
- Ettema, D. and Timmermans, H. (1997). *Activity-Based Approaches to Travel Analysis*. Pergamon/Elsevier, Amsterdam.
- Falk, M., Marohn, F., Michel, R., Hofmann, D. and Macke, M. (2006). A First Course on Time Series Analysis—Examples with SAS, University of Würzburg, Germany. Version 2006.Feb. 01.
- FHWA. (1997). *Status of the Nation's Surface Transportation System: Condition and Performance*. Federal Highway Administration, Washington, DC.
- Finch, J., Curran, P., West, S. (1994). The effects of model and data characteristics on the accuracy of parameter estimates and standard errors in confirmatory factor analysis. Unpublished manuscript. University of North Carolina, Chapel Hill, NC.
- Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, NY.
- Fomunung, I., Washington, S., Guensler, R., Bachman, W. (2000). Validation of the MEASURE automobile emissions model: a statistical analysis. *Journal of Transportation Statistics* 3, 65–84.
- Freund, J.E. and Ansari, A. (1957). *Two-Way Rank Sum Tests for Variance*. Technical Report to Office of Ordnance Research and National Science Foundation 34. Virginia Polytechnic Institute, Blacksburg, VA.
- Freund, R. and Wilson, W. (1997). *Statistical Methods*. Academic Press, NY.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32, 675–701.
- Garber, N. (1991). *Impacts of Differential Speed Limits on Highway Speeds and Accidents*. AAA Foundation for Traffic Safety, Washington, DC.
- Gastwirth, J. (1965). Percentile modifications of two-sample rank tests. *Journal of the American Statistical Association* 60, 1127–1141.
- Geary, R. (1966). A note on residual covariance and estimation efficiency in regression. *American Statistician* 20, 30–31.
- Gelman, A., Carlin, J., Stern, H., Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, NY.
- Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *The Journal of Time Series Analysis* 4, 221–238.
- Geweke, J., Keane, M., Runkle, D. (1994). Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics* 76, 609–632.
- Gibbons, J. (1985a). *Nonparametric Methods for Quantitative Analysis*, 2nd ed. American Sciences Press, Syracuse, NY.
- Gibbons, J. (1985b). *Nonparametric Statistical Inference*, 2nd ed. Marcel Dekker, NY.
- Gibbons, J. and Gastwirth, J. (1970). Properties of the percentile modified rank tests. *Annals of the Institute of Statistical Mathematics*, Suppl. 6, 95–114.
- Gilbert, C. (1992). A duration model of automobile ownership. *Transportation Research Part B* 26, 97–114.
- Gillen, M. and Martin, B. (2009). Price volatility in airline markets. *Transportation Research Part E* 45, 693–709.
- Glasser, G. and Winter, R. (1961). Critical values of the rank correlation coefficient for testing the hypothesis of independence. *Biometrika* 48, 444–448.

- Glen, D. and Martin, B. (2004). A Survey of the modeling of dry bulk and tanker markets. *Research in Transportation Economics* 12, 19–64.
- Glenberg, A. (1996). *Learning from Data: An Introduction to Statistical Reasoning*, 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.
- Godfrey, L. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge University Press, NY.
- Goldfeld, S. and Quandt, R. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association* 60, 539–547.
- Goldfeld, S. and Quandt, R. (1972). *Nonlinear Methods in Econometrics*. North-Holland, Publishing Company, Amsterdam and London.
- Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* 52, 681–700.
- Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* 37, 424–438.
- Granger, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–130.
- Granger, C. and Andersen, A. (1978). *An Introduction to Bilinear Time Series Models*. Vandenhoeck and Ruprecht, Göttingen.
- Granger, C. and Engle, P. (1984). Dynamic model specification with equilibrium constraints: co-integration and error correction. Draft manuscript.
- Granger, C. and Engle, R. (1987). Co-integration and error-correction: representation, estimation and testing. *Econometrica* 55, 251–276.
- Granger, C. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series and Analysis* 1(1), 15–29.
- Granger, C. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–120.
- Granger, C. and Weiss, A. (1983). Time series analysis of error-correction models. In *Studies in Econometrics, Time Series and Multivariate Statistics in Honor of T.W. Anderson*, S. Karlin, T. Amemiya, and I.A. Goodman, Eds. Academic Press, San Diego, 255–278.
- Green, M. and Symons, M. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases* 36, 715–724.
- Greene, W. (1990a). *Econometric Analysis*. Macmillan, NY.
- Greene, W. (1990b). A gamma distributed stochastic frontier model. *Journal of Econometrics* 46, 141–163.
- Greene, W. (1995b). Count data. Manuscript, Department of Economics, Stern School of Business, New York University, NY.
- Greene, W. (2007). *LIMDEP Version 9.0*. Econometric Software, Inc., Plainview, NY.
- Greene, W. (2008). *Econometric Analysis*, 6th ed. Prentice-Hall, Upper Saddle River, NJ.
- Greene, D. and Hu, P. (1984). The influence of the price of gasoline on vehicle use in multivehicle households. *Transportation Research Record* 988, 19–23.
- Griffiths, W., Hill, R., Judge, G. (1993). *Learning and Practicing Econometrics*. John Wiley & Sons, NY.
- Gujarati, D. (1992). *Essentials of Econometrics*. McGraw-Hill, NY.
- Gullikson, H. and Tukey, J. (1958). Reliability for the law of comparative judgment. *Psychometrika* 23, 95–110.
- Gumbel, E. (1958). *Statistics of Extremes*. Columbia University Press, NY.

- Gurmu, S. and Trivedi, P. (1994). Recent developments in models of event counts: a survey. Department of Economics, University of Virginia, Charlottesville.
- Haider, S. and Chatti, K. (2009). Effect of design and site factors on fatigue cracking of new flexible pavements in the LTPP SPS-1 experiment. *International Journal of Pavement Engineering* 10, 133–147.
- Haldar, A. and Mahadevan, S. (2000). *Probability, Reliability, and Statistical Methods in Engineering Design*. John Wiley & Sons, NY.
- Halton, J. (1960). On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals,' *Numerische Mathematik* 2, 84–90.
- Hamed, M. and Mannering, F. (1993). Modeling travelers' postwork activity involvement: toward a new methodology. *Transportation Science* 17, 381–394.
- Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–384.
- Han, A. and Hausman, J. (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* 5, 1–28.
- Harley, H. and Rao, J. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93–108.
- Harrod, S. (2009). Capacity factors of mixed speed railway network. *Transportation Research Part E* 45, 830–841.
- Harvey, A. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44, 461–465.
- Harwood, D., Council, F., Hauer, E., Hughes, W., Vogt, A. (2000). *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Federal Highway Administration Report, FHWA-RD-99-207, Washington, DC.
- Hausman, J., Hall, B., Griliches, Z. (1984). Economic models for count data with an application to the patents-R&D relationship. *Econometrica* 52, 909–938.
- Hausman, J. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–1271.
- Hausman, J. and Taylor, W. (1981). Panel data and unobservable individual effects. *Econometrica* 49, 1377–1398.
- Hay, J. (1980). Occupational choice and occupational earnings. Ph.D. dissertation. Yale University, New Haven, CT.
- Haykin, S. (1999). *NN: A Comprehensive Foundation*. Macmillan, NY.
- Heckman, J. (1976). The common structure of statistical models for truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–492.
- Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931–960.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–162.
- Heckman, J. (1981). Statistical models for discrete panel data. In *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski and D. McFadden, Eds. MIT Press, Cambridge, MA.
- Heckman, J. and Borjas, G. (1980). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Econometrica* 47, 247–283.
- Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271–320.

- Heckman, J. and McCurdy, T. (1980). A life-cycle model of female labor supply. *Review of Economic Studies* 47, 47–74.
- Hensher, D. (1998). The timing of change for automobile transactions: competing risk multispell specification. In *Travel Behaviour Research: Updating the State of Play*, J.D. Ortuzar, D. Hensher, and S. Jara-Diaz, Eds. Elsevier, Amsterdam, 487–506.
- Hensher, D. (2001). The sensitivity of the valuation of travel time savings to the specification of unobserved effects. *Transportation Research Part E* 37, 129–142.
- Hensher, D. and Mannering, F. (1994). Hazard-based duration models and their application to transport analysis. *Transport Reviews* 14, 63–82.
- Hensher, D. and Milthorpe, F. (1987). Selectivity correction in discrete-continuous choice analysis: with empirical evidence for vehicle choice and use. *Regional Science and Urban Economics* 17, 123–150.
- Hilbe, J. (2009). *Logistic regression models*. Chapman & Hall/CRC.
- Hildreth, C. and Lu, J. (1960). Demand relations with autocorrelated disturbances. Technical Bulletin 276. Michigan State University, Agriculture Experiment Station.
- Hoeffding, W. (1951). Optimum nonparametric tests. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, J. Neuman, Ed. University of California Press, Berkeley, CA, 83–92.
- Hogg, R. and Craig, A. (1994). *Introduction to Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, NJ.
- Hollander, M. (1970). A distribution-free test for parallelism. *Journal of the American Statistical Association* 65, 387–394.
- Honore, B. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60, 533–565.
- Hornik, K., Stinchcombe, M., White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366.
- Hosking, J. (1981). Fractional differencing. *Biometrika* 68(1), 165–176.
- Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression. Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, NY.
- Hoyle, R. Ed. (1995). *Structural Equation Modeling: Concepts, Issues, and Applications*. Sage, Thousand Oaks, CA.
- Hsiao, C. (1975). Some estimation methods for a random coefficients model. *Econometrica* 43, 305–325.
- Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge University Press, Cambridge, UK.
- Hu, L., Bentler, P., Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin* 112, 351–362.
- Hui, W. (1990). Proportional hazard Weibull mixtures. Working paper, Department of Economics, Australian National University, Canberra.
- Iman, R., Quade, D., Alexander, D. (1975). Exact probability levels for the Kruskal-Wallis test. In *Selected Tables in Mathematical Statistics*, Vol. 3. Institute of Mathematical Statistics, American Mathematical Society, Providence, RI, 329–384.
- Ingram, D. and Kleinman, J. (1989). Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine* 8, 525–538.
- Ishak, S. and Al-Deek, H. (2002). Performance evaluation of short-term time series traffic prediction model. *Journal of Transportation Engineering* 128, 490–498.
- Issarayangyun, T. and Greaves, S. (2007). Analysis of minute-by-minute exposure to fine particulates inside a car: a time-series modelling approach. *Transportation Research Part D* 12, 347–357.

- Jiang, X. and Adeli, H. (2004). Wavelet-packet autocorrelation function method for traffic flow pattern analysis. *Computer-Aided Civil and Infrastructure Engineering* 19, 324–337.
- Jiang, X. and Adeli, H. (2005). Dynamic wavelet neural network model for traffic flow forecasting. *Journal of Transportation Engineering* 131, 771–779.
- Johansson, P. (1996). Speed limitation and motorway casualties: a time series count data regression approach. *Accident Analysis and Prevention* 28, 73–87.
- Johnson, N. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 1. John Wiley & Sons, NY.
- Johnson, R. (1994). *Miller & Freund's Probability and Statistics for Engineers*, 5th ed. Prentice-Hall, Englewood Cliffs, NJ.
- Johnson, R. and Wichern, D. (1992). *Multivariate Statistical Analysis*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.
- Joreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183–202.
- Kaarsemaker, L. and van Wijngaarden, A. (1953). Tables for use in rank correlation. *Statistica Neerlandica* 7, 41–54.
- Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, NY.
- Kamarianakis, Y. and Prastacos, P. (2003). Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. *Transportation Research Record: Journal of the Transportation Research Board* 1857, 74–84.
- Kamat, A. (1956). A two-sample distribution-free test. *Biometrika* 43, 377–387.
- Kang, S. (1985). A note on the equivalence of specification tests in the two-factor multivariate variance components model. *Journal of Econometrics* 28, 193–203.
- Kanussanos, M., Visvikis, I., Batchelor, R. (2004). Over-the-counter forward contracts and spot price volatility in shipping. *Transportation Research Part E* 40, 273–296.
- Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481.
- Karlaftis, M. (2010). Ownership and competition in European transit: assessing efficiency. *Transportmetrica* (Forthcoming).
- Karlaftis, M. and McCarthy, P. (1998). Operating subsidies and performance: an empirical study. *Transportation Research Part A* 32, 359–375.
- Karlaftis, M. and Sinha, K. (1997). Modeling approach for rolling stock deterioration prediction. *Journal of Transportation Engineering* 12, 227–228.
- Karlaftis, M. and Tarko, A. (1998). Heterogeneity considerations in accident modeling. *Accident Analysis and Prevention* 30, 425–433.
- Karlaftis, M. and Vlahogianni, E. (2009). Memory properties and fractional integration in transportation time-series. *Transportation Research Part C* 17, 444–453.
- Karlaftis, M. and Vlahogianni, E. (2010). Statistical methods versus neural networks in transportation research: similarities and some insights. *Transportation Research Part C* (Forthcoming).
- Karlaftis, M., Golias, J., Papadimitriou, S. (2002). Transit quality as an integrated traffic management strategy: measuring perceived service quality. *Journal of Public Transportation* 4, 13–26.
- Karlaftis, M., McCarthy, P., Sinha, K. (1999). System size and cost structure of transit industry, *Journal of Transportation Engineering* 125, 208–215.
- Katz, L. (1986). Layoffs, recall and the duration of unemployment. Working paper 1825, National Bureau of Economic Research, Cambridge, MA.

- Keane, M. (1994). A computationally practical simulation estimator for panel data. *Econometrica* 62, 95–116.
- Keller, G. and Warrack, B. (1997). *Statistics for Management and Economics*, 4th ed. Duxbury Press, Belmont, CA.
- Kendall, M. (1962). *Rank Correlation Methods*, 3rd ed. Hafner, NY.
- Kennan, D. (1985). A Tukey non-additivity-type test for time series nonlinearity. *Biometrika* 72, 39–44.
- Kennedy, P. (1998). *A Guide to Econometrics*. MIT Press, Cambridge, MA.
- Kharoufeh, J. and Goulias, K. (2002). Non-parametric identification of daily activity durations using kernel density estimators. *Transportation Research Part B* 36, 59–82.
- Kiefer, N. (1988). Economic duration data and hazard functions. *Journal of Economic Literature* 26, 646–679.
- Kim, P. and Jennrich, R. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion D_{mn} ($m = n$). In *Selected Tables in Mathematical Statistics*, Vol. 1. Institute of Mathematical Statistics, American Mathematics Society, Providence, RI, 79–170.
- Kitagawa, G. and Gerch, W. (1984). A smoothness priors modeling of time series with trend and seasonality. *Journal of the American Statistical Association* 79, 378–389.
- Kline, R. (1998). *Principles and Practice of Structural Equation Modeling*. Guilford Press, NY.
- Klotz, J. (1962). Nonparametric tests for scale. *Annals of Mathematical Statistics* 33, 495–512.
- Kmenta, J. (1997). *Elements of Econometrics*. 2nd ed. Macmillan Press, New York, NY.
- Koppelman, F. (1975). Travel prediction with models of individualistic choice behavior. Ph.D. dissertation, Department of Civil Engineering, MIT, Cambridge, MA.
- Kotegoda, N. and Rosso, R. (1997). *Probability, Statistics, and Reliability for Civil and Environmental Engineers*. McGraw-Hill, NY.
- Kruskal, W. and Wallis, W. (1952). Use of ranks in the one-criterion analysis of variance. *Journal of the American Statistical Association* 47, 583–621, 1952; errata, 48, 905–911.
- Kuan, C. and White, H. (1994). Artificial neural networks: an econometric perspective. *Econometric Reviews* 13(1), 1–9.
- Lachapelle, U. and Frank, L. (2009). Transit and health: mode of transport, employer-sponsored public transit pass programs, and physical activity. *Journal of Public Health Policy* 30, S73–S94.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1–14.
- Lancaster, T. (1968). Grouping estimators on heteroscedastic data. *Journal of the American Statistical Association* 63, 182–191.
- Larson, R. and Marx, M. (1986). *An Introduction to Mathematical Statistics and Its Applications*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.
- Lau, J., Hung, W., Yuen, D., Cheung, C. (2009). Long-memory characteristics of urban roadside air quality. *Transportation Research Part D* 14, 353–359.
- Laubscher, N. and Odeh, R. (1976). A confidence interval for the scale parameter based on Sukhatme's two-sample statistic. *Communications in Statistics—Theory and Methods* 14, 1393–1407.
- Laubscher, N., Odeh, R., Steffens, F., deLange, E. (1968). Exact critical values for Mood's distribution-free test statistic for dispersion and its normal approximation. *Technometrics* 10, 497–508.

- Lee, E. (1992). *Statistical Methods for Survival Data Analysis*, 2nd ed. John Wiley & Sons, NY.
- Lee, H-Y., Lin, K., Wu, J. (2002). Pitfalls in using Granger causality tests to find an engine of growth. *Applied Economics Letters* 9, 411–414.
- Lee, J. and Mannerling, F. (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis and Prevention* 34, 149–161.
- Lillard, L. and Willis, R. (1978). Dynamic aspects of earning mobility. *Econometrica* 46, 985–1012.
- Link, H., Gotze, W., Himanen, V. (2009). Estimating the marginal costs of airport operation using multivariate time series models with correlated error terms. *Journal of Air Transport Management* 15, 41–46.
- Liu, J. and Brockwell, P.J. (1988). On the general bilinear time-series model. *Journal of Applied Probability* 25, 553–564.
- Ljung, G. and Box, G. (1978). On a measure of lack of fit in time series models. *Biometrika* 65, 297–303.
- Loizides, J. and Tsionas, E. (2002). Productivity growth in European railways: a new approach. *Transportation Research Part A* 36, 633–644.
- Lord, D., Washington, S., Ivan, J. (2004). Poisson, Poisson-gamma, and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37, 35–46.
- MacCallum, R. (1990). The need for alternative measures of fit in covariance structure modeling. *Multivariate Behavioral Research* 25, 157–162.
- Madanat, S., Karlaftis, M., McCarthy, P. (1997). Probabilistic infrastructure deterioration models with panel data. *ASCE Journal of Infrastructure Systems* 3, 4–9.
- Maddala, G. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, UK.
- Maddala, G. (1988). *Introduction to Econometrics*. Macmillan, NY.
- Maghsoodloo, S. (1975). Estimates of the quantiles of Kendall's partial rank correlation coefficient and additional quantile estimates. *Journal of Statistical Computing and Simulation* 4, 155–164.
- Maghsoodloo, S. and Pallos, L. (1981). Asymptotic behavior of Kendall's partial rank correlation coefficient and additional quantile estimates. *Journal of Statistical Computing and Simulation* 13, 41–48.
- Magnus, J. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics* 19, 239–285.
- Makridakis, S., Wheelwright, S., McGee, V. (1989). *Forecasting: Methods and Applications*, 3rd ed. John Wiley & Sons, NY.
- Malyshkina, N. and Mannerling, F. (2009). Markov switching multinomial logit model: an application to accident-injury severities. *Accident Analysis and Prevention* 41, 829–838.
- Malyshkina, N. and Mannerling, F. (2010). Zero-state Markov switching count-data models: an empirical assessment. *Accident Analysis and Prevention* 42, 122–130.
- Malyshkina, N., Mannerling, F., Tarko, A. (2009). Markov switching negative binomial models: an application to vehicle accident frequencies. *Accident Analysis and Prevention* 41, 217–226.
- Manly, B. (1986). *Multivariate Statistical Methods, A Primer*. Chapman & Hall, NY.
- Mann, H. (1945). Nonparametric tests against trend. *Econometrica* 13, 245–259.

- Mann, H. and Whitney, D. (1947). On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 50–60.
- Mannerling, F. (1983). An econometric analysis of vehicle use in multivehicle households. *Transportation Research Part A* 17, 183–189.
- Mannerling, F. (1986a). Selectivity bias in models of discrete/continuous choice: an empirical analysis. *Transportation Research Record* 1085, 58–62.
- Mannerling, F. (1986b). A note on endogenous variables in household vehicle utilization equations. *Transportation Research Part B* 20, 1–6.
- Mannerling, F. (1993). Male/female driver characteristics and accident risk: some new evidence. *Accident Analysis and Prevention* 25, 77–84.
- Mannerling, F. (2007). Effects of interstate speed limits on driving speeds: Some new evidence. Compendium of papers CD-ROM, Transportation Research Board 86th Annual Meeting, Paper No. 07-0120, Washington DC.
- Mannerling, F. and Hensher, D. (1987). Discrete/continuous econometric models and their application to transport analysis. *Transport Reviews* 7, 227–244.
- Mannerling, F. and Hamed, M. (1990a). Occurrence, frequency, and duration of commuters' work-to-home departure delay. *Transportation Research Part B* 24, 99–109.
- Mannerling, F. and Hamed, M. (1990b). Commuter welfare approach to high occupancy vehicle lane evaluation: an exploratory analysis. *Transportation Research Part A* 24, 371–379.
- Mannerling, F. and Harrington, I. (1981). Use of density function and Monte Carlo simulation techniques in the evaluation of policy impacts on travel demand. *Transportation Research Record* 801, 8–15.
- Mannerling, F. and Winston, C. (1985). Dynamic empirical analysis of household vehicle ownership and utilization. *Rand Journal of Economics* 16, 215–236.
- Mannerling, F. and Winston, C. (1987a). Economic effects of voluntary export restrictions. In *Blind Intersection: Policy and the Automobile Industry*. Brookings Institution, Washington, DC, 61–67.
- Mannerling, F. and Winston, C. (1987b). U.S. automobile market demand. In *Blind Intersection: Policy and the Automobile Industry*. Brookings Institution, Washington, DC, 36–60.
- Mannerling, F., Abu-Eisheh, S., Arnadottir, A. (1990). Dynamic traffic equilibrium with discrete/continuous econometric models. *Transportation Science* 24, 105–116.
- Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* 45, 1977–1988.
- Manski, C. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski and D. McFadden, Eds. MIT Press, Cambridge, MA.
- Marsh, H. and Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: first- and higher-order factor models and their invariance across groups. *Psychological Bulletin* 97, 562–582.
- Masten, S. (2007). Do states upgrading to primary enforcement of safety belt laws experience increased daytime and nighttime belt use? *Accident Analysis and Prevention* 39, 1131–1139.
- Mathsoft. (1999). *S-Plus 2000 Statistical Software*. Mathsoft, Inc., Bellevue, WA.
- McCarthy, P. (2001). *Transportation Economics Theory and Practice: A Case Study Approach*. Blackwell, Boston, MA.

- McCulloch, R. and Tsay, R. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* 88, 968–978.
- McFadden, D. (1978). Modeling the choice of residential location. In *Spatial Interaction Theory and Residential Location*, A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, Eds. North-Holland, Amsterdam.
- McFadden, D. (1981). Econometric models of probabilistic choice. In *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski and D. McFadden, Eds. MIT Press, Cambridge, MA.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57, 995–1026.
- McFadden, D. and Ruud, P. (1994). Estimation by simulation. *Review of Economics and Statistics* 76, 591–608.
- McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15, 447–470.
- McKelvey, R. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4, 103–120.
- McKenzie, E. (1985). Some simple models for discrete variable time series. *Water Resources Bulletin* 21, 645–650.
- McLeod, A. and Li, W. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* 4, 269–273.
- McNemar, Q. (1962). *Psychological Statistics*. John Wiley & Sons, NY.
- Mehta, C. and Patel, N. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* 78, 427–434.
- Meyer, B. (1990). Unemployment insurance and unemployment spells. *Econometrica* 58, 757–782.
- Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.
- Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention* 26, 471–482.
- Miaou, S. and Lum, H. (1993). Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis and Prevention* 25, 689–709.
- Milton, J. and Mannering, F. (1998). The relationship among highway geometrics, traffic-related elements and motor vehicle accident frequencies, *Transportation* 25, 395–413.
- Milton, J., Shankar, V., Mannering, F. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis and Prevention* 40, 260–266.
- Molenaar, K., Park, J-I., Washington, S. (2009). Framework for measuring corporate safety culture and its impact on construction safety performance. *ASCE Journal of Construction Engineering and Management* 135, 488–496.
- Molenaar, K., Washington, S., Diekmann, J. (2000). A structural equation model of contract dispute potential. *ASCE Journal of Construction Engineering and Management* 124, 268–277.
- Mood, A. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics* 25, 514–522.
- Moran, P. (1951). Partial and multiple rank correlation. *Biometrika* 38, 26–32.

- Moses, L. (1963). Rank tests for dispersion. *Annals of Mathematical Statistics* 34, 973–983.
- Moulton, B. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics* 32, 385–397.
- Mulaik, S., James, L., Van Alstine, J., Bennett, N., Lind, S., Stilwell, C. (1989). Evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin* 105, 430–445.
- Mullahey, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* 33, 341–365.
- Mundlak, Y. (1978). On the pooling of time series and cross-section data. *Econometrica* 46, 69–85.
- Murray, M. (1994). A drunk and her dog: an illustration of cointegration and error correction. *The American Statistician* 48, 37–39.
- Muthen, B. (1984). A general structural equation model with dichotomous, ordered Categorical, and continuous latent variable indicators. *Psychometrika* 49, 115–132.
- Myers, R. (1990). *Classical and Modern Regression with Applications*, 2nd ed. Duxbury Press, Belmont, CA.
- Nam, D. and Mannerling, F. (2000). Hazard-based analysis of highway incident duration. *Transportation Research Part A* 34, 85–102.
- National Technical University of Athens, Department of Transportation Planning and Engineering (1996). <http://www/transport.ntua.gr/map>. Accessed May 6, 2010.
- Nerlove, M. (1971). A note on error components models. *Econometrica* 39, 383–396.
- Neter, J., Kutner, M., Nachtsheim, C., Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th ed. Irwin, Boston, MA.
- Nicholson, W. (1978). *Microeconomic Theory*. Dryden Press, Hinsdale, IL.
- Niemeier, D. (1997). Accessibility: an evaluation using consumer welfare. *Transportation* 24, 377–396.
- Noland, R. and Karlaftis, M. (2005). Sensitivity of crash models to alternative specifications. *Transportation Research part E*, Vol. 41, 439–458.
- Norris, J. (1997). *Markov Chains*. Cambridge University Press, Cambridge.
- Oaks, D. (1977). The asymptotic information in censored survival data. *Biometrika* 64, 441–448.
- Odaki, M. (1993). On the invertibility of fractionally differenced ARIMA processes. *Biometrika* 80(13), 703–709.
- Olmstead, P. (1946). Distribution of sample arrangements for runs up and down. *Annals of Mathematical Statistics* 17, 24–33.
- Park, R. (1966). Estimation with heteroscedastic error terms. *Econometrica* 34, 888.
- Pedhazur, E. and Pedhazur S. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Peeta, S. and Anastassopoulos, I. (2002). Automatic real-time detection and correction of erroneous detector data with Fourier transforms for on-line traffic control architectures. *Transportation Research Record* 1811, 1–16.
- Peterson, A. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: applications to competing risks. *Proceedings of the National Academy of Sciences U.S.A.* 73, 11–13.
- Phlips, L. (1974). *Applied Consumption Analysis*. North-Holland, Amsterdam.
- Pindyck, R. and Rubinfeld, D. (1990). *Econometric Models and Economic Forecasts*. McGraw-Hill, NY.

- Pindyck, R. and Rubinfeld, D. (1997). *Econometric Models & Economic Forecasts*, 4th ed. McGraw-Hill, NY.
- Pinjari, A. and Bhat, C. (2010). A multiple discrete-continuous nested extreme value (MDCNEV) model: formulation and application to non-worker activity time-use and timing behavior on weekdays *Transportation Research Part B* 44, 562–583.
- Poch, M. and Mannering, F. (1996). Negative binomial analysis of intersection accident frequencies. *Journal of Transportation Engineering* 122, 391–401.
- Principe, J., Euliano, N., Lefebvre, C. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley & Sons, Inc.
- Qiao, F. (2005). Application to wavelet: a tutorial, transportation research board. Workshop188.http://cseweb.ucsd.edu/~baden/Doc/wavelets/qiao_wavelet_intro.pdf. Accessed December 29, 2009.
- Quddus, M. (2007). Time series count data models: an empirical application to traffic accidents. In *Proceedings of the 86th Transportation Research Board Annual Meeting*, Washington, DC.
- Quddus, M. (2008). Time series count data models: an empirical application to traffic accidents. *Accident Analysis and Prevention* 40, 1732–1741.
- Ramsey, J. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society Series B*, 31, 350–371.
- Rao, C. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. John Wiley & Sons, NY.
- Reznick, S. (1994). Adventures in Stochastic Process. Birkhauser, Basel.
- Rhyne, A. and Steel, R. (1965). Tables for a treatment versus control multiple comparisons sign test. *Technometrics* 7, 293–306.
- Rigdon, E., Schumacker, R., Wothke, W. (1998). A comparative review of interaction and nonlinear modeling. In *Interaction and Nonlinear Effects in Structural Equation Modeling*, R. Schumacker and G. Marcoulides, Eds. Lawrence Erlbaum Associates, Mahwah, NJ.
- Ripley, B. (1993). Statistical aspects of neural networks. In: O.E. Barndoff-Nielsen Jensen, J.L. Jensen and W.S. Kendall, Editors, *Networks and Chaos—Statistical and Probabilistic Aspects*, Chapman & Hall, London (1993), pp. 40–123.
- Ripley, B. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society, Series B* 56, 409–456.
- Ripley, B. (1996). *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.
- Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer. NY.
- Romilly, P. (2001). Subsidy and local bus service deregulation in Britain, a re-evaluation. *Journal of Transport Economics and Policy* 35, 161–194.
- Rosenbaum, S. (1953). Tables for a nonparametric test of dispersion. *Annals of Mathematical Statistics* 24, 663–668.
- Rosenkrantz, W. (1997). *Introduction to Probability and Statistics for Scientists and Engineers*. McGraw-Hill, NY.
- Sarle, W. (1994). Neural networks and statistical models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, April, 1–13.
- Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. *Quality and Quantity* 24, 367–386.
- Savolainen, P. and Mannering F. (2007). Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. *Accident Analysis and Prevention* 39, 955–963.

- Schumacker, R. and Marcoulides, G. Eds. (1998). *Interaction and Non-Linear Effects in Structural Equation Modeling*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Shafizadeh, K. and Mannering, F. (2006). Statistical modeling of user perceptions of infrastructure condition: An application to the case of highway roughness. *Journal of Transportation Engineering* 132, 133–140.
- Shankar, V. and Mannering, F. (1998). Modeling the endogeneity of lane-mean speeds and lane-speed deviations: a structural equations approach. *Transportation Research Part A* 32, 311–322.
- Shankar, V., Albin, R., Milton, J., Mannering, F. (1998). Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. *Transportation Research Record* 1635, 44–48.
- Shankar, V., Mannering, F., Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural accident frequencies. *Accident Analysis and Prevention* 27, 371–389.
- Shankar, V., Milton, J., Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention* 29, 829–837.
- Shumway, R. and Stoffer, S. (2000). *Time Series Analysis and Its Applications*. Springer-Verlag, NY.
- Siegel, S. and Tukey, J. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association* 55, 429–445; errata, 56, 1005, 1961.
- Sims, C. (1972). Money, income, and causality. *American Economic Review, American Economic Association* 62(4), 540–552.
- Small, K. and Hsiao, C. (1985). Multinomial logit specification tests. *International Economic Review* 26, 619–627.
- Small, K. and Rosen, H. (1981). Applied welfare economics with discrete choice models. *Econometrica* 49, 105–130.
- Smith, B., Williams, B., Oswald, K. (2002). Comparison of parametric and non-parametric models for traffic flow forecasting. *Transportation Research Part C* 10(4), 303–321.
- Smith, L. (1998). *Linear Algebra*, 3rd ed. Springer, NY.
- Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology* 15, 73–101.
- Spiegel, M. and Stephens, L. (1998). *Schaum's Outlines: Statistics*, 3rd ed. McGraw-Hill International, NY.
- Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B* 64, 583–640.
- Stathopoulos, A. and Karlaftis, M. (2001a). Spectral and cross-spectral analysis of urban traffic flows. In *Proceedings of the 4th IEEE Conference on Intelligent Transportation Systems*, August 25–29, Oakland, CA.
- Stathopoulos, A. and Karlaftis, M. (2001b). Temporal and spatial variations of real-time traffic data in urban areas. *Transportation Research Record* 1768, 135–140.
- Stathopoulos, A. and Karlaftis, M. (2002). Modeling the duration of urban traffic congestion. *Journal of Transportation Engineering* 128, 587–590.

- Stathopoulos, A. and Karlaftis, M. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C* 11, 121–135.
- Steel, R. (1959a). A multiple comparison sign test: treatments versus control. *Journal of the American Statistical Association* 54, 767–775.
- Steel, R. (1959b). A multiple comparison sign test: treatments versus control. *Biometrics* 15, 560–572.
- Steel, R. (1960). A rank sum test for comparing all sets of treatments. *Technometrics* 2, 197–207.
- Steel, R. (1961). Some rank sum multiple comparisons tests. *Biometrics* 17, 539–552.
- Steel, R., Torrie, J., Dickey, D. (1997). *Principles and Procedures of Statistics: A Biometrical Approach*, 3rd ed. McGraw-Hill, NY.
- Steiger, J. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research* 25, 173–180.
- Steiger, J., Shapiro, A., Browne, M. (1985). On the multivariate asymptotic distribution and sequential chi-square statistics. *Psychometrika* 50, 253–263.
- Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature* 35, 2006–2039.
- Sukhatme, B. (1957). On certain two-sample nonparametric tests for variances. *Annals of Mathematical Statistics* 28, 188–194.
- Swamy, P. (1971). *Statistical Inference in Random Coefficient Regression Models*. Springer-Verlag, NY.
- Swamy, P. and Arora, S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica* 40, 261–275.
- Swed, F. and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics* 14, 66–87.
- Taylor, W. (1980). Small sample considerations in estimation from panel data. *Journal of Econometrics* 13, 203–223.
- Teng, H. and Qi, Y. (2003). Application of wavelet technique to freeway incident detection. *Transportation Research Part C* 11, 289–308.
- Terasvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* 89, 208–218.
- Terry, M. (1952). Some rank order tests which are most powerful against specific parametric alternatives. *Annals of Mathematical Statistics* 14, 66–87.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae* 12, 85–91.
- Theil, H. (1978). *Introduction to Econometrics*. Prentice-Hall, Englewood Cliffs, NJ.
- Tjøstheim, D. (1986). Some doubly stochastic time series models. *Journal of Time Series Analysis* 7, 51–72.
- Tjøstheim, D. (1994). Non-linear time series: a selective review. *Scandinavian Journal of Statistics* 21, 97–130.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36.
- Tong, H. (1983). Threshold models in nonlinear time series analysis. Lecture Notes in Statistics, Springer-Verlag, NY.
- Tong, H. (1990). *A Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford, UK.
- Train, K. (1986). *Qualitative Choice Analysis: Theory, Econometrics and an Application to Automobile Demand*. MIT Press, Cambridge, MA.

- Train, K. (1999). Halton sequences for mixed logit. Working Paper, Department of Economics, University of California, Berkley.
- Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press. Cambridge, UK.
- Tsay, R. (1986). Nonlinearity tests for time series. *Biometrika* 73, 461–466.
- Tsay, R. (2002). *Analysis of Financial Time series*. John Wiley & Sons, NY.
- Tukey, J. (1959). A quick, compact, two-sample test to Duchworth's specifications. *Technometrics* 1, 31–48.
- Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.
- Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review* 79, 281–299.
- Ulfarsson, G., and Mannering, F. (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis and Prevention* 36, 135–147.
- U.S. DOT. (1997). *National Transportation Statistics*. Bureau of Transportation Statistics, Washington, DC.
- Van der Waerden, B. (1952). Order tests for the two-sample problem and their power, I, II, III. *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen, Series A 55(Indagationes Mathematicae) 14*. 453–459.
- Van der Waerden, B. (1953). Order Tests for the Two-Sample Problem and Their Power, I, II, III. *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen, Series A 55 Indagationes Mathematicae* 15, 303–310, 311–316; errata, 15, 80, 1953.
- Van der Waerden, B. and Nievergelt, E. (1956). *Tafeln zum Vergleich zweier Stichproben mittels X-test und Zeichentest*. Springer, Berlin.
- Van Harn, K. and Steutel, F. (1977). Generalized renewal sequences and infinitely divisible lattice distributions. *Stochastic Processes and their Applications* 5, 47–55.
- Van Harn, K. and Steutel, F. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability* 7, 893–899.
- Vardeman, S. (1994). *Statistics for Engineering Problem Solving*. PWS, Boston, MA.
- Vardeman, S. and Jobe, J. (1994). *Basic Engineering Data Collection and Analysis*. Duxbury/Thomas Learning, NY.
- Vogt, A. (1999). Accident models for rural intersections: four-lane by two-lane stop-controlled and two-lane by two-lane signalized. Federal Highway Administration Report FHWA-RD-99-128. Washington, DC.
- Vogt, A. and Bared, J. (1998). Accident prediction models for two-lane rural roads: segments and intersections. Federal Highway Administration Report FHWA-RD-98-133, Washington, DC.
- Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–334.
- Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics* 11, 147–162.
- Wallace, T. and Hussain, A. (1969). The use of error components models in combining cross-section and time-series data. *Econometrica* 37, 55–72.
- Wansbeek, T. and Kapteyn, A. (1982). A simple way to obtain the spectral decomposition of variance components models for balanced data. *Communications in Statistics A11*, 2105–2112.
- Wansbeek, T. and Kapteyn, A. (1983). A note on spectral decomposition and maximum likelihood estimation of ANOVA models with balanced data. *Statistics and Probability Letters* 1, 213–215.

- Wansbeek, T. and Kapteyn, A. (1989). Estimation of the error components model with incomplete panels. *Journal of Econometrics* 41, 341–361.
- Washington, S. (2000a). Conducting statistical tests of hypotheses: five common misconceptions found in transportation research. *Transportation Research Record* 1665, 1–6.
- Washington, S. (2000b). Iteratively specified tree-based regression models: theoretical development and example applied to trip generation. *Journal of Transportation Engineering* 126, 482–491.
- Washington, S. and Wolf, J. (1997). Hierarchical tree-based versus linear regression: theory and example applied to trip generation. *Transportation Research Record* 1581, 82–88.
- Washington, S., Metarko, J., Fomunung, I., Ross, R., Julian, F., Moran, E. (1999). An inter-regional comparison: fatal crashes in the southeastern and non-southeastern United States: preliminary findings. *Accident Analysis and Prevention* 31, 135–146.
- Washington, S., Wolf, J., Guensler, R. (1997). A binary recursive partitioning method for modeling hot-stabilized emissions from motor vehicles. *Transportation Research Record* 1587, 96–105.
- Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous populations of any form. *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen* 51, 252–261.
- White, D. and Washington, S. (2001). Safety restraint use rate as a function of law enforcement and other factors: preliminary analysis. *Transportation Research Record* 1779, 109–115.
- White, H. (1980). A heteroscedastic-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48, 817–838.
- Whittaker, J., Garside, S., Lindevelde, K. (1997). Tracking and predicting network traffic process. *International Journal of Forecasting* 13, 51–61.
- WingBUGS, Version 1.4.3. (2007). Imperial College and Medical Research Council, UK.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* 1, 80–83.
- Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics* 3, 119–122.
- Wilcoxon, F. (1949). *Some Rapid Approximate Statistical Procedures*. American Cyanamid Co., Stanford Research Laboratory, Stanford, CA.
- Wilcoxon, F., Katti, S., Wilcox, R. (1972). Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. In *Selected Tables in Mathematical Statistics*. Vol. 1, Institute of Mathematical Statistics, American Mathematical Society, Providence, RI, 171–259.
- Williams, B. and Hoel, L. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: a theoretical basis and empirical results. *Journal of Transportation Engineering* 129, 583–716.
- Williams, B., Durvasula, P., Brown, D. (1998). Urban traffic flow prediction: application of seasonal ARIMA and exponential smoothing models. *Transportation Research Record* 1644, 132–144.
- Winston, C. and Manner, F. (1984). Consumer demand for automobile safety. *American Economic Review* 74, 316–319.
- Wolf, J., Guensler, R., Washington, S. (1998). High emitting vehicle characterization using regression tree analysis. *Transportation Research Record* 1641, 58–65.

- Wong, D., Pitfield, D., Caves, R., Appleyard, A. (2009). The development of more risk-sensitive and flexible airport safety area strategy: Part I. The development of an improved accident frequency model. *Safety Science* 47, 903–912.
- Yamamoto, T., Hashiji, J. and Shankar, V. (2008). Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis and Prevention* 40, 1320–1329.
- Yamamoto, T., Kitamura, R., Kimura, S. (1999). Competing-risks-duration model of household vehicle transactions with indicators of changes in explanatory variables. *Transportation Research Record* 1676, 116–123.
- Yamamoto, T., Madre, J.-L., Kitamura, R. (2004). An analysis of the effects of French vehicle inspection program and grant for scrappage on household vehicle transaction. *Transportation Research Part B* 38(10), 905–926.
- Yannis, G. and Karlaftis, M. (2010). An investigation of weather effects, traffic volume and speed on daily traffic accidents and fatalities. *Accident Analysis and Prevention* (Forthcoming).
- Young, P. and Ord, K. (2004). Monitoring transportation indicators, and an analysis of the effects of September 11, 2001. *Journal of Transportation Statistics* 7, 69–85.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and test for aggregation bias. *Journal of the American Statistical Association* 57, 348–368.

Praise for the First Edition

“... an excellent addition to a practicing transportation analyst’s library as well as a perfect companion to a first-year graduate modeling or methods course ... this text adroitly fills a very important niche between practice and theory. ...”

—*Journal of Transportation Statistics*

“It is well done and well organized, and provides good coverage of all the essential elements of statistical and econometric methods and models applied to transportation ... I would highly recommend it to anyone engaged in transportation research. I suspect it will be the definitive text on statistics in transportation for some years to come ...”

—*Technometrics*

“... an outstanding and unique contribution to the existing transportation literature. I have no doubt that the book will serve as an important resource for transportation practitioners and researchers ...”

—*Journal of Transportation Engineering*

Describing tools commonly used in the field, **Statistical and Econometric Methods for Transportation Data Analysis, Second Edition** provides an understanding of a broad range of analytical tools required to solve transportation problems. It includes a wide breadth of examples and case studies covering applications in various aspects of transportation planning, engineering, safety, and economics. Each chapter clearly presents fundamental concepts and principles and includes numerous references for those seeking additional technical details and applications.

This second edition contains new chapters on logistic regression, ordered probability models, random-parameter models, and Bayesian statistical modeling. Along with new examples and data sets, it also offers an explicit treatment of frequency domain time series analysis, including Fourier and wavelets analysis methods.



CRC Press

Taylor & Francis Group
an informa business

www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487

270 Madison Avenue
New York, NY 10016

2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

C285X

ISBN: 978-1-4200-8285-2

9 781420 082852