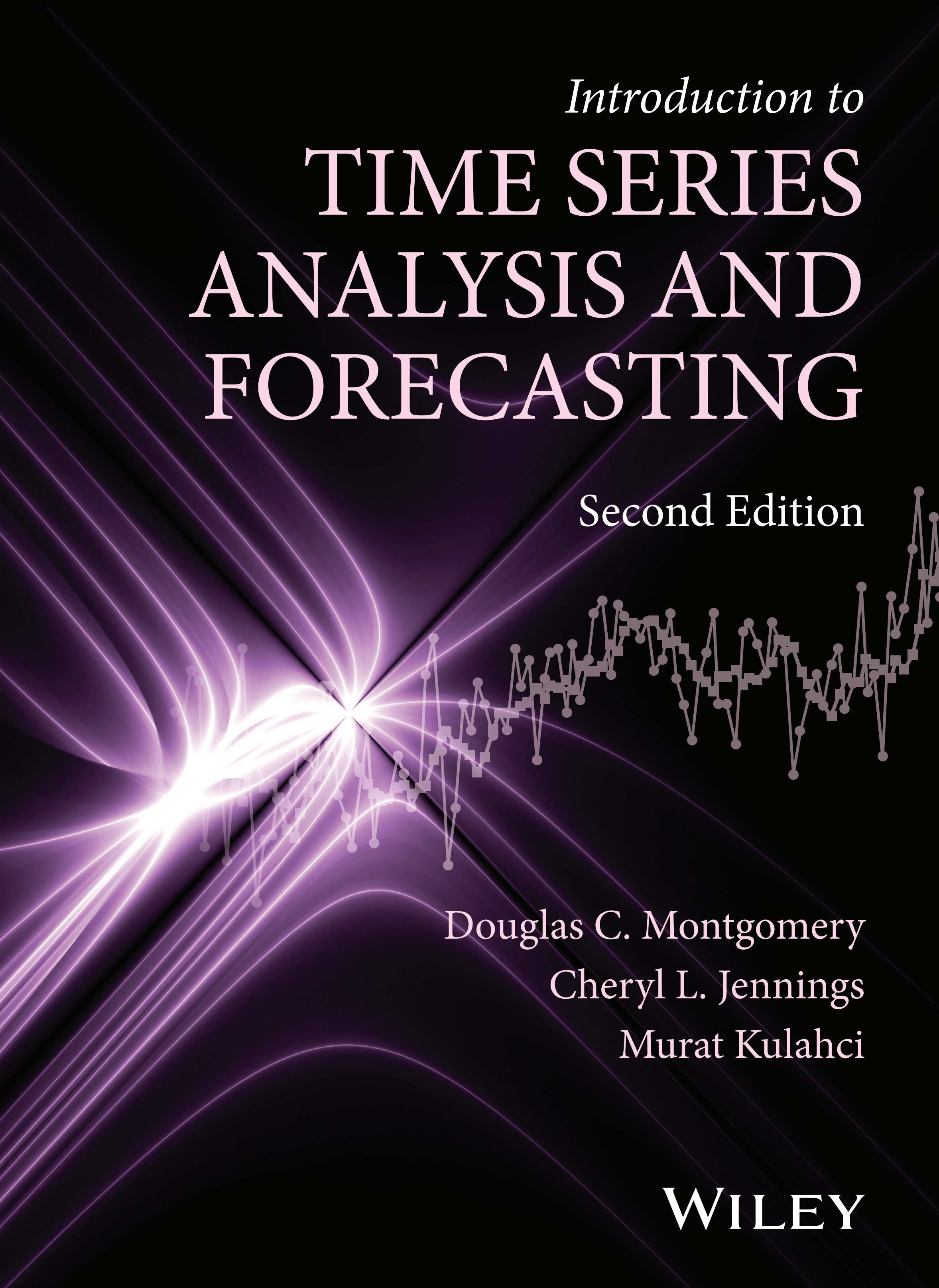
Wiley Series in Probability and Statistics

Introduction to

TIME SERIES ANALYSIS AND FORECASTING

Second Edition



Douglas C. Montgomery
Cheryl L. Jennings
Murat Kulahci

WILEY

INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

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

INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING

Second Edition

DOUGLAS C. MONTGOMERY
Arizona State University
Tempe, Arizona, USA

CHERYL L. JENNINGS
Arizona State University
Tempe, Arizona, USA

MURAT KULAHCI
Technical University of Denmark
Lyngby, Denmark
and
Luleå University of Technology
Luleå, Sweden

WILEY

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

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

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data applied for.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE **xi**

1 INTRODUCTION TO FORECASTING **1**

1.1 The Nature and Uses of Forecasts / 1
1.2 Some Examples of Time Series / 6
1.3 The Forecasting Process / 13
1.4 Data for Forecasting / 16
1.4.1 The Data Warehouse / 16
1.4.2 Data Cleaning / 18
1.4.3 Imputation / 18
1.5 Resources for Forecasting / 19
Exercises / 20

2 STATISTICS BACKGROUND FOR FORECASTING **25**

2.1 Introduction / 25
2.2 Graphical Displays / 26
2.2.1 Time Series Plots / 26
2.2.2 Plotting Smoothed Data / 30
2.3 Numerical Description of Time Series Data / 33
2.3.1 Stationary Time Series / 33

2.3.2 Autocovariance and Autocorrelation Functions / 36
2.3.3 The Variogram / 42
2.4 Use of Data Transformations and Adjustments / 46
2.4.1 Transformations / 46
2.4.2 Trend and Seasonal Adjustments / 48
2.5 General Approach to Time Series Modeling and Forecasting / 61
2.6 Evaluating and Monitoring Forecasting Model Performance / 64
2.6.1 Forecasting Model Evaluation / 64
2.6.2 Choosing Between Competing Models / 74
2.6.3 Monitoring a Forecasting Model / 77
2.7 R Commands for Chapter 2 / 84
Exercises / 96

3 REGRESSION ANALYSIS AND FORECASTING 107

3.1 Introduction / 107
3.2 Least Squares Estimation in Linear Regression Models / 110
3.3 Statistical Inference in Linear Regression / 119
3.3.1 Test for Significance of Regression / 120
3.3.2 Tests on Individual Regression Coefficients and Groups of Coefficients / 123
3.3.3 Confidence Intervals on Individual Regression Coefficients / 130
3.3.4 Confidence Intervals on the Mean Response / 131
3.4 Prediction of New Observations / 134
3.5 Model Adequacy Checking / 136
3.5.1 Residual Plots / 136
3.5.2 Scaled Residuals and PRESS / 139
3.5.3 Measures of Leverage and Influence / 144
3.6 Variable Selection Methods in Regression / 146
3.7 Generalized and Weighted Least Squares / 152
3.7.1 Generalized Least Squares / 153
3.7.2 Weighted Least Squares / 156
3.7.3 Discounted Least Squares / 161
3.8 Regression Models for General Time Series Data / 177

3.8.1 Detecting Autocorrelation: The Durbin–Watson Test / 178
3.8.2 Estimating the Parameters in Time Series Regression Models / 184
3.9 Econometric Models / 205
3.10 R Commands for Chapter 3 / 209
Exercises / 219

4 EXPONENTIAL SMOOTHING METHODS 233

4.1 Introduction / 233
4.2 First-Order Exponential Smoothing / 239
4.2.1 The Initial Value, $\tilde{y}_0$ / 241
4.2.2 The Value of λ / 241
4.3 Modeling Time Series Data / 245
4.4 Second-Order Exponential Smoothing / 247
4.5 Higher-Order Exponential Smoothing / 257
4.6 Forecasting / 259
4.6.1 Constant Process / 259
4.6.2 Linear Trend Process / 264
4.6.3 Estimation of σ_e^2 / 273
4.6.4 Adaptive Updating of the Discount Factor / 274
4.6.5 Model Assessment / 276
4.7 Exponential Smoothing for Seasonal Data / 277
4.7.1 Additive Seasonal Model / 277
4.7.2 Multiplicative Seasonal Model / 280
4.8 Exponential Smoothing of Biosurveillance Data / 286
4.9 Exponential Smoothers and Arima Models / 299
4.10 R Commands for Chapter 4 / 300
Exercises / 311

5 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS 327

5.1 Introduction / 327
5.2 Linear Models for Stationary Time Series / 328
5.2.1 Stationarity / 329
5.2.2 Stationary Time Series / 329

5.3 Finite Order Moving Average Processes / 333

5.3.1 The First-Order Moving Average Process, MA(1) / 334

5.3.2 The Second-Order Moving Average Process, MA(2) / 336

5.4 Finite Order Autoregressive Processes / 337

5.4.1 First-Order Autoregressive Process, AR(1) / 338

5.4.2 Second-Order Autoregressive Process, AR(2) / 341

5.4.3 General Autoregressive Process, AR(*p*) / 346

5.4.4 Partial Autocorrelation Function, PACF / 348

5.5 Mixed Autoregressive–Moving Average Processes / 354

5.5.1 Stationarity of ARMA(*p, q*) Process / 355

5.5.2 Invertibility of ARMA(*p, q*) Process / 355

5.5.3 ACF and PACF of ARMA(*p, q*) Process / 356

5.6 Nonstationary Processes / 363

5.6.1 Some Examples of ARIMA(*p, d, q*) Processes / 363

5.7 Time Series Model Building / 367

5.7.1 Model Identification / 367

5.7.2 Parameter Estimation / 368

5.7.3 Diagnostic Checking / 368

5.7.4 Examples of Building ARIMA Models / 369

5.8 Forecasting Arima Processes / 378

5.9 Seasonal Processes / 383

5.10 Arima Modeling of Biosurveillance Data / 393

5.11 Final Comments / 399

5.12 R Commands for Chapter 5 / 401

Exercises / 412

6 TRANSFER FUNCTIONS AND INTERVENTION MODELS 427

6.1 Introduction / 427

6.2 Transfer Function Models / 428

6.3 Transfer Function–Noise Models / 436

6.4 Cross-Correlation Function / 436

6.5 Model Specification / 438

6.6 Forecasting with Transfer Function–Noise Models / 456

6.7 Intervention Analysis / 462

6.8 R Commands for Chapter 6 / 473

Exercises / 486

7 SURVEY OF OTHER FORECASTING METHODS 493

7.1 Multivariate Time Series Models and Forecasting / 493

7.1.1 Multivariate Stationary Process / 494

7.1.2 Vector ARIMA Models / 494

7.1.3 Vector AR (VAR) Models / 496

7.2 State Space Models / 502

7.3 Arch and Garch Models / 507

7.4 Direct Forecasting of Percentiles / 512

7.5 Combining Forecasts to Improve Prediction Performance / 518

7.6 Aggregation and Disaggregation of Forecasts / 522

7.7 Neural Networks and Forecasting / 526

7.8 Spectral Analysis / 529

7.9 Bayesian Methods in Forecasting / 535

7.10 Some Comments on Practical Implementation and Use of Statistical Forecasting Procedures / 542

7.11 R Commands for Chapter 7 / 545

Exercises / 550

APPENDIX A STATISTICAL TABLES 561

APPENDIX B DATA SETS FOR EXERCISES 581

APPENDIX C INTRODUCTION TO R 627

BIBLIOGRAPHY 631

INDEX 639

PREFACE

Analyzing time-oriented data and forecasting future values of a time series are among the most important problems that analysts face in many fields, ranging from finance and economics to managing production operations, to the analysis of political and social policy sessions, to investigating the impact of humans and the policy decisions that they make on the environment. Consequently, there is a large group of people in a variety of fields, including finance, economics, science, engineering, statistics, and public policy who need to understand some basic concepts of time series analysis and forecasting. Unfortunately, most basic statistics and operations management books give little if any attention to time-oriented data and little guidance on forecasting. There are some very good high level books on time series analysis. These books are mostly written for technical specialists who are taking a doctoral-level course or doing research in the field. They tend to be very theoretical and often focus on a few specific topics or techniques. We have written this book to fill the gap between these two extremes.

We have made a number of changes in this revision of the book. New material has been added on data preparation for forecasting, including dealing with outliers and missing values, use of the variogram and sections on the spectrum, and an introduction to Bayesian methods in forecasting. We have added many new exercises and examples, including new data sets in Appendix B, and edited many sections of the text to improve the clarity of the presentation.

Like the first edition, this book is intended for practitioners who make real-world forecasts. We have attempted to keep the mathematical level modest to encourage a variety of users for the book. Our focus is on short- to medium-term forecasting where statistical methods are useful. Since many organizations can improve their effectiveness and business results by making better short- to medium-term forecasts, this book should be useful to a wide variety of professionals. The book can also be used as a textbook for an applied forecasting and time series analysis course at the advanced undergraduate or first-year graduate level. Students in this course could come from engineering, business, statistics, operations research, mathematics, computer science, and any area of application where making forecasts is important. Readers need a background in basic statistics (previous exposure to linear regression would be helpful, but not essential), and some knowledge of matrix algebra, although matrices appear mostly in the chapter on regression, and if one is interested mainly in the results, the details involving matrix manipulation can be skipped. Integrals and derivatives appear in a few places in the book, but no detailed working knowledge of calculus is required.

Successful time series analysis and forecasting requires that the analyst interact with computer software. The techniques and algorithms are just not suitable to manual calculations. We have chosen to demonstrate the techniques presented using three packages: Minitab®, JMP®, and R, and occasionally SAS®. We have selected these packages because they are widely used in practice and because they have generally good capability for analyzing time series data and generating forecasts. Because R is increasingly popular in statistics courses, we have included a section in each chapter showing the R code necessary for working some of the examples in the chapter. We have also added a brief appendix on the use of R. The basic principles that underlie most of our presentation are not specific to any particular software package. Readers can use any software that they like or have available that has basic statistical forecasting capability. While the text examples do utilize these particular software packages and illustrate some of their features and capability, these features or similar ones are found in many other software packages.

There are three basic approaches to generating forecasts: regression-based methods, heuristic smoothing methods, and general time series models. Because all three of these basic approaches are useful, we give an introduction to all of them. Chapter 1 introduces the basic forecasting problem, defines terminology, and illustrates many of the common features of time series data. Chapter 2 contains many of the basic statistical tools used in analyzing time series data. Topics include plots, numerical summaries of time series data including the autocovariance and autocorrelation functions, transformations, differencing, and decomposing a time series into trend and seasonal components. We also introduce metrics for evaluating forecast errors and methods for evaluating and tracking forecasting performance over time. Chapter 3 discusses regression analysis and its use in forecasting. We discuss both crosssection and time series regression data, least squares and maximum likelihood model fitting, model adequacy checking, prediction intervals, and weighted and generalized least squares. The first part of this chapter covers many of the topics typically seen in an introductory treatment of regression, either in a stand-alone course or as part of another applied statistics course. It should be a reasonable review for many readers. Chapter 4 presents exponential smoothing techniques, both for time series with polynomial components and for seasonal data. We discuss and illustrate methods for selecting the smoothing constant(s), forecasting, and constructing prediction intervals. The explicit time series modeling approach to forecasting that we have chosen to emphasize is the autoregressive integrated moving average (ARIMA) model approach. Chapter 5 introduces ARIMA models and illustrates how to identify and fit these models for both nonseasonal and seasonal time series. Forecasting and prediction interval construction are also discussed and illustrated. Chapter 6 extends this discussion into transfer function models and intervention modeling and analysis. Chapter 7 surveys several other useful topics from time series analysis and forecasting, including multivariate time series problems, ARCH and GARCH models, and combinations of forecasts. We also give some practical advice for using statistical approaches to forecasting and provide some information about realistic expectations. The last two chapters of the book are somewhat higher in level than the first five.

Each chapter has a set of exercises. Some of these exercises involve analyzing the data sets given in Appendix B. These data sets represent an interesting cross section of real time series data, typical of those encountered in practical forecasting problems. Most of these data sets are used in exercises in two or more chapters, an indication that there are usually several approaches to analyzing, modeling, and forecasting a time series. There are other good sources of data for practicing the techniques given in this book. Some of the ones that we have found very interesting and useful include the U.S. Department of Labor—Bureau of Labor Statistics (http://www.bls.gov/data/home.htm), the U.S. Department of Agriculture—National Agricultural Statistics Service, Quick Stats Agricultural Statistics Data (http://www.nass.usda.gov/Data_and_Statistics/Quick_Stats/index.asp), the U.S. Census Bureau (http://www.census.gov), and the U.S.

Department of the Treasury (http://www.treas.gov/offices/domestic-finance/debt-management/interest-rate/). The time series data library created by Rob Hyndman at Monash University (http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/index.htm) and the time series data library at the Mathematics Department of the University of York (http://www.york.ac.uk/depts/maths/data/ts/) also contain many excellent data sets. Some of these sources provide links to other data. Data sets and other materials related to this book can be found at ftp://ftp.wiley.com/public/scitechmed/ timeseries.

We would like to thank the many individuals who provided feedback and suggestions for improvement to the first edition. We found these suggestions most helpful. We are indebted to Clifford Long who generously provided the R codes he used with his students when he taught from the book. We found his codes very helpful in putting the end-of-chapter R code sections together. We also have placed a premium in the book on bridging the gap between theory and practice. We have not emphasized proofs or technical details and have tried to give intuitive explanations of the material whenever possible. The result is a book that can be used with a wide variety of audiences, with different interests and technical backgrounds, whose common interests are understanding how to analyze time-oriented data and constructing good short-term statistically based forecasts.

We express our appreciation to the individuals and organizations who have given their permission to use copyrighted material. These materials are noted in the text. Portions of the output contained in this book are printed with permission of Minitab Inc. All material remains the exclusive property and copyright of Minitab Inc. All rights reserved.

Douglas C. Montgomery
Cheryl L. Jennings
Murat Kulahci

CHAPTER 1

INTRODUCTION TO FORECASTING

> It is difficult to make predictions, especially about the future
>
> NEILS BOHR, *Danish physicist*

1.1 THE NATURE AND USES OF FORECASTS

A **forecast** is a prediction of some future event or events. As suggested by Neils Bohr, making good predictions is not always easy. Famously "bad" forecasts include the following from the book *Bad Predictions*:

- "The population is constant in size and will remain so right up to the end of mankind." *L'Encyclopedie*, 1756.
- "1930 will be a splendid employment year." U.S. Department of Labor, *New Year's Forecast* in 1929, just before the market crash on October 29.
- "Computers are multiplying at a rapid rate. By the turn of the century there will be 220,000 in the U.S." *Wall Street Journal*, 1966.

Introduction to Time Series Analysis and Forecasting, Second Edition.
Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci.

Forecasting is an important problem that spans many fields including business and industry, government, economics, environmental sciences, medicine, social science, politics, and finance. Forecasting problems are often classified as short-term, medium-term, and long-term. Short-term forecasting problems involve predicting events only a few time periods (days, weeks, and months) into the future. Medium-term forecasts extend from 1 to 2 years into the future, and long-term forecasting problems can extend beyond that by many years. Short- and medium-term forecasts are required for activities that range from operations management to budgeting and selecting new research and development projects. Long-term forecasts impact issues such as strategic planning. Short- and medium-term forecasting is typically based on identifying, modeling, and extrapolating the patterns found in historical data. Because these historical data usually exhibit inertia and do not change dramatically very quickly, statistical methods are very useful for short- and medium-term forecasting. This book is about the use of these statistical methods.

Most forecasting problems involve the use of time series data. A **time series** is a time-oriented or chronological sequence of observations on a variable of interest. For example, Figure 1.1 shows the market yield on US Treasury Securities at 10-year constant maturity from April 1953 through December 2006 (data in Appendix B, Table B.1). This graph is called a **time**

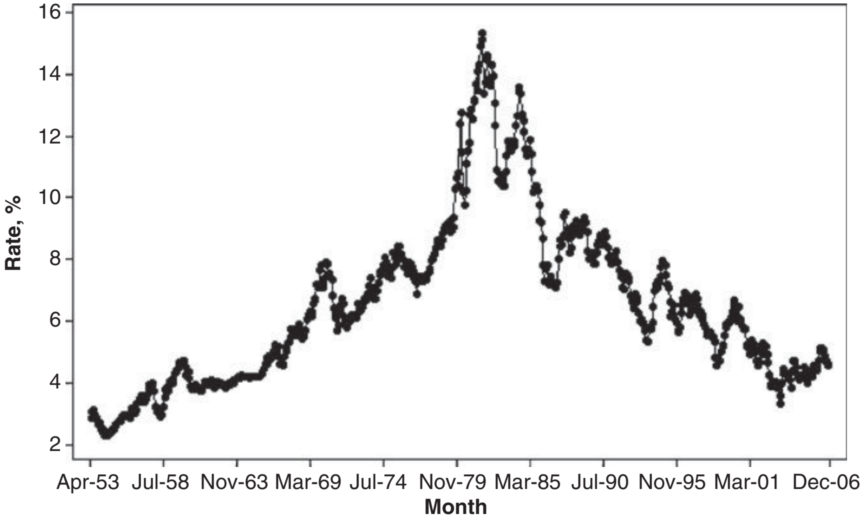


FIGURE 1.1 Time series plot of the market yield on US Treasury Securities at 10-year constant maturity. *Source:* US Treasury.

series plot. The rate variable is collected at equally spaced time periods, as is typical in most time series and forecasting applications. Many business applications of forecasting utilize daily, weekly, monthly, quarterly, or annual data, but any reporting interval may be used. Furthermore, the data may be instantaneous, such as the viscosity of a chemical product at the point in time where it is measured; it may be cumulative, such as the total sales of a product during the month; or it may be a statistic that in some way reflects the activity of the variable during the time period, such as the daily closing price of a specific stock on the New York Stock Exchange.

The reason that forecasting is so important is that prediction of future events is a critical input into many types of planning and decision-making processes, with application to areas such as the following:

1. *Operations Management*. Business organizations routinely use forecasts of product sales or demand for services in order to schedule production, control inventories, manage the supply chain, determine staffing requirements, and plan capacity. Forecasts may also be used to determine the mix of products or services to be offered and the locations at which products are to be produced.
2. *Marketing*. Forecasting is important in many marketing decisions. Forecasts of sales response to advertising expenditures, new promotions, or changes in pricing polices enable businesses to evaluate their effectiveness, determine whether goals are being met, and make adjustments.
3. *Finance and Risk Management*. Investors in financial assets are interested in forecasting the returns from their investments. These assets include but are not limited to stocks, bonds, and commodities; other investment decisions can be made relative to forecasts of interest rates, options, and currency exchange rates. Financial risk management requires forecasts of the volatility of asset returns so that the risks associated with investment portfolios can be evaluated and insured, and so that financial derivatives can be properly priced.
4. *Economics*. Governments, financial institutions, and policy organizations require forecasts of major economic variables, such as gross domestic product, population growth, unemployment, interest rates, inflation, job growth, production, and consumption. These forecasts are an integral part of the guidance behind monetary and fiscal policy, and budgeting plans and decisions made by governments. They are also instrumental in the strategic planning decisions made by business organizations and financial institutions.

5. *Industrial Process Control*. Forecasts of the future values of critical quality characteristics of a production process can help determine when important controllable variables in the process should be changed, or if the process should be shut down and overhauled. Feedback and feedforward control schemes are widely used in monitoring and adjustment of industrial processes, and predictions of the process output are an integral part of these schemes.
6. *Demography*. Forecasts of population by country and regions are made routinely, often stratified by variables such as gender, age, and race. Demographers also forecast births, deaths, and migration patterns of populations. Governments use these forecasts for planning policy and social service actions, such as spending on health care, retirement programs, and antipoverty programs. Many businesses use forecasts of populations by age groups to make strategic plans regarding developing new product lines or the types of services that will be offered.

These are only a few of the many different situations where forecasts are required in order to make good decisions. Despite the wide range of problem situations that require forecasts, there are only two broad types of forecasting techniques—qualitative methods and quantitative methods.

Qualitative forecasting techniques are often subjective in nature and require judgment on the part of experts. Qualitative forecasts are often used in situations where there is little or no historical data on which to base the forecast. An example would be the introduction of a new product, for which there is no relevant history. In this situation, the company might use the expert opinion of sales and marketing personnel to subjectively estimate product sales during the new product introduction phase of its life cycle. Sometimes qualitative forecasting methods make use of marketing tests, surveys of potential customers, and experience with the sales performance of other products (both their own and those of competitors). However, although some data analysis may be performed, the basis of the forecast is subjective judgment.

Perhaps the most formal and widely known qualitative forecasting technique is the **Delphi Method**. This technique was developed by the RAND Corporation (see Dalkey [1967]). It employs a panel of experts who are assumed to be knowledgeable about the problem. The panel members are physically separated to avoid their deliberations being impacted either by social pressures or by a single dominant individual. Each panel member responds to a questionnaire containing a series of questions and returns the information to a coordinator. Following the first questionnaire, subsequent questions are submitted to the panelists along with information about the opinions of the panel as a group. This allows panelists to review their predictions relative to the opinions of the entire group. After several rounds, it is hoped that the opinions of the panelists converge to a consensus, although achieving a consensus is not required and justified differences of opinion can be included in the outcome. Qualitative forecasting methods are not emphasized in this book.

Quantitative forecasting techniques make formal use of historical data and a **forecasting model**. The model formally summarizes patterns in the data and expresses a statistical relationship between previous and current values of the variable. Then the model is used to project the patterns in the data into the future. In other words, the forecasting model is used to extrapolate past and current behavior into the future. There are several types of forecasting models in general use. The three most widely used are regression models, smoothing models, and general time series models. Regression models make use of relationships between the variable of interest and one or more related predictor variables. Sometimes regression models are called **causal forecasting models,** because the predictor variables are assumed to describe the forces that cause or drive the observed values of the variable of interest. An example would be using data on house purchases as a predictor variable to forecast furniture sales. The method of least squares is the formal basis of most regression models. **Smoothing models** typically employ a simple function of previous observations to provide a forecast of the variable of interest. These methods may have a formal statistical basis, but they are often used and justified heuristically on the basis that they are easy to use and produce satisfactory results. General **time series models** employ the statistical properties of the historical data to specify a formal model and then estimate the unknown parameters of this model (usually) by least squares. In subsequent chapters, we will discuss all three types of quantitative forecasting models.

The form of the forecast can be important. We typically think of a forecast as a single number that represents our best estimate of the future value of the variable of interest. Statisticians would call this a **point estimate** or **point forecast.** Now these forecasts are almost always wrong; that is, we experience **forecast error**. Consequently, it is usually a good practice to accompany a forecast with an estimate of how large a forecast error might be experienced. One way to do this is to provide a **prediction interval** (PI) to accompany the point forecast. The PI is a range of values for the future observation, and it is likely to prove far more useful in decision-making than a single number. We will show how to obtain PIs for most of the forecasting methods discussed in the book.

Other important features of the forecasting problem are the **forecast horizon** and the **forecast interval.** The forecast horizon is the number of future periods for which forecasts must be produced. The horizon is often dictated by the nature of the problem. For example, in production planning, forecasts of product demand may be made on a monthly basis. Because of the time required to change or modify a production schedule, ensure that sufficient raw material and component parts are available from the supply chain, and plan the delivery of completed goods to customers or inventory facilities, it would be necessary to forecast up to 3 months ahead. The forecast horizon is also often called the forecast **lead time.** The **forecast interval** is the frequency with which new forecasts are prepared. For example, in production planning, we might forecast demand on a monthly basis, for up to 3 months in the future (the lead time or horizon), and prepare a new forecast each month. Thus the forecast interval is 1 month, the same as the basic period of time for which each forecast is made. If the forecast lead time is always the same length, say, T periods, and the forecast is revised each time period, then we are employing a **rolling** or **moving horizon** forecasting approach. This system updates or revises the forecasts for $T-1$ of the periods in the horizon and computes a forecast for the newest period T. This rolling horizon approach to forecasting is widely used when the lead time is several periods long.

1.2 SOME EXAMPLES OF TIME SERIES

Time series plots can reveal **patterns** such as random, trends, level shifts, periods or cycles, unusual observations, or a combination of patterns. Patterns commonly found in time series data are discussed next with examples of situations that drive the patterns.

The sales of a mature pharmaceutical product may remain relatively flat in the absence of unchanged marketing or manufacturing strategies. Weekly sales of a generic pharmaceutical product shown in Figure 1.2 appear to be constant over time, at about $10{,}400 \times 10^3$ units, in a random sequence with no obvious patterns (data in Appendix B, Table B.2).

To assure conformance with customer requirements and product specifications, the production of chemicals is monitored by many characteristics. These may be input variables such as temperature and flow rate, and output properties such as viscosity and purity.

Due to the continuous nature of chemical manufacturing processes, output properties often are **positively autocorrelated;** that is, a value above the long-run average tends to be followed by other values above the

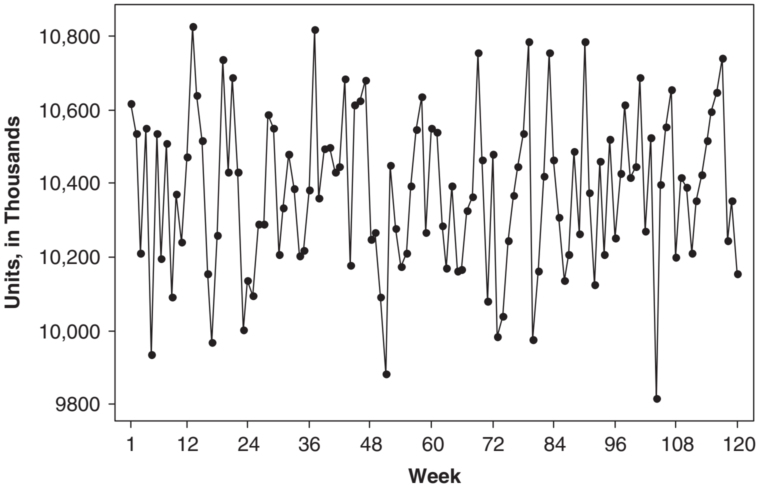


FIGURE 1.2 Pharmaceutical product sales.

average, while a value below the average tends to be followed by other values below the average.

The viscosity readings plotted in Figure 1.3 exhibit autocorrelated behavior, tending to a long-run average of about 85 centipoises (cP), but with a structured, not completely random, appearance (data in Appendix B, Table B.3). Some methods for describing and analyzing autocorrelated data will be described in Chapter 2.

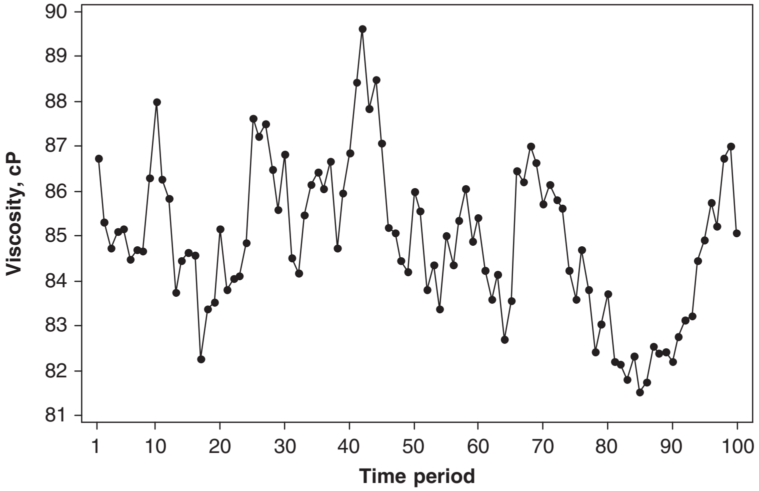


FIGURE 1.3 Chemical process viscosity readings.

The USDA National Agricultural Statistics Service publishes agricultural statistics for many commodities, including the annual production of dairy products such as butter, cheese, ice cream, milk, yogurt, and whey. These statistics are used for market analysis and intelligence, economic indicators, and identification of emerging issues.

Blue and gorgonzola cheese is one of 32 categories of cheese for which data are published. The annual US production of blue and gorgonzola cheeses (in 10^3 lb) is shown in Figure 1.4 (data in Appendix B, Table B.4). Production quadrupled from 1950 to 1997, and the **linear trend** has a constant positive slope with random, year-to-year variation.

The US Census Bureau publishes historic statistics on manufacturers' shipments, inventories, and orders. The statistics are based on North American Industry Classification System (NAICS) code and are utilized for purposes such as measuring productivity and analyzing relationships between employment and manufacturing output.

The manufacture of beverage and tobacco products is reported as part of the nondurable subsector. The plot of monthly beverage product shipments (Figure 1.5) reveals an overall increasing trend, with a distinct **cyclic pattern** that is repeated within each year. January shipments appear to be the lowest, with highs in May and June (data in Appendix B, Table B.5). This monthly, or **seasonal,** variation may be attributable to some cause

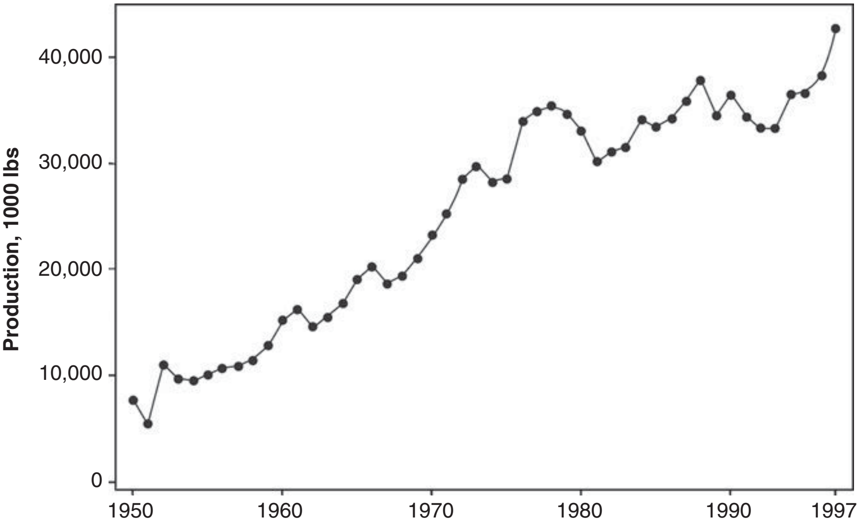


FIGURE 1.4 The US annual production of blue and gorgonzola cheeses. *Source:* USDA–NASS.

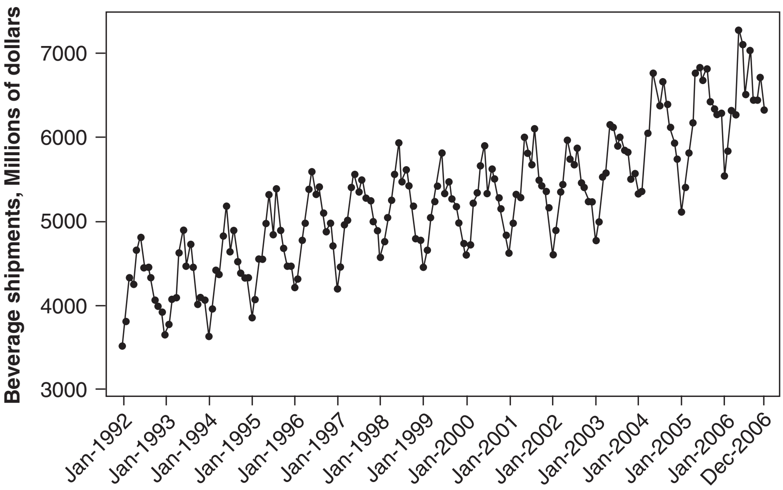


FIGURE 1.5 The US beverage manufacturer monthly product shipments, unadjusted. *Source:* US Census Bureau.

such as the impact of weather on the demand for beverages. Techniques for making seasonal adjustments to data in order to better understand general trends will be discussed in Chapter 2.

To determine whether the Earth is warming or cooling, scientists look at annual mean temperatures. At a single station, the warmest and the coolest temperatures in a day are averaged. Averages are then calculated at stations all over the Earth, over an entire year. The change in global annual mean surface air temperature is calculated from a base established from 1951 to 1980, and the result is reported as an "anomaly."

The plot of the annual mean anomaly in global surface air temperature (Figure 1.6) shows an increasing trend since 1880; however, the slope, or rate of change, varies with time periods (data in Appendix B, Table B.6). While the slope in earlier time periods appears to be constant, slightly increasing, or slightly decreasing, the slope from about 1975 to the present appears much steeper than the rest of the plot.

Business data such as stock prices and interest rates often exhibit **nonstationary** behavior; that is, the time series has no natural mean. The daily closing price adjusted for stock splits of Whole Foods Market (WFMI) stock in 2001 (Figure 1.7) exhibits a combination of patterns for both mean level and slope (data in Appendix B, Table B.7).

While the price is constant in some short time periods, there is no consistent mean level over time. In other time periods, the price changes

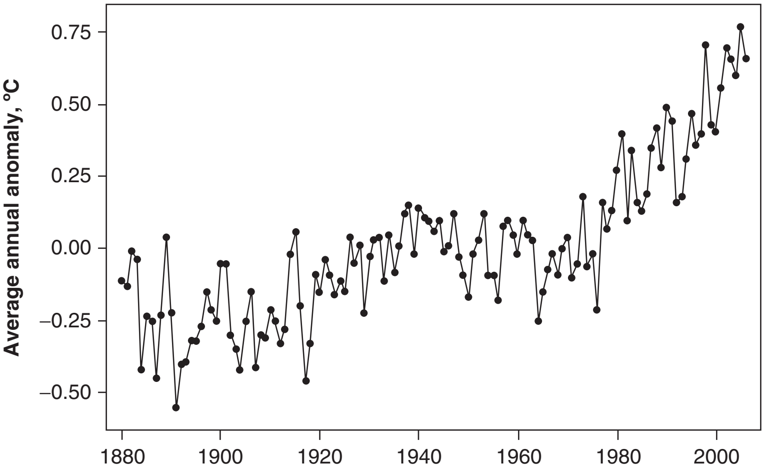


FIGURE 1.6 Global mean surface air temperature annual anomaly. *Source:* NASA-GISS.

at different rates, including occasional abrupt shifts in level. This is an example of nonstationary behavior, which will be discussed in Chapter 2.

The Current Population Survey (CPS) or "household survey" prepared by the US Department of Labor, Bureau of Labor Statistics, contains national data on employment, unemployment, earnings, and other labor market topics by demographic characteristics. The data are used to report

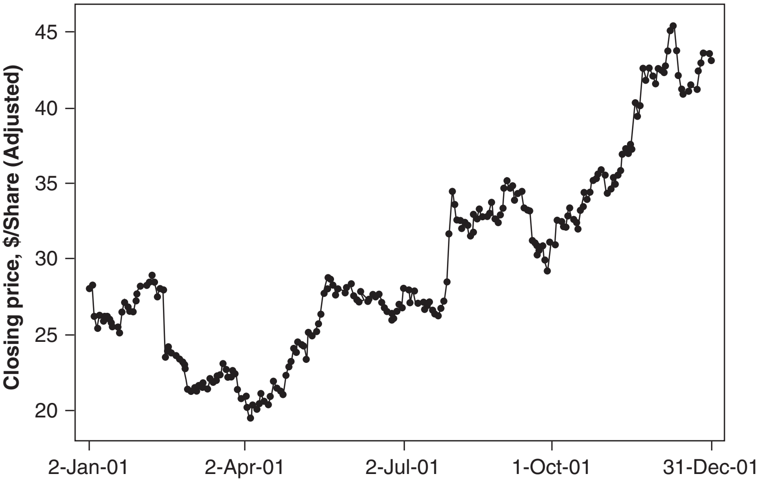


FIGURE 1.7 Whole foods market stock price, daily closing adjusted for splits.

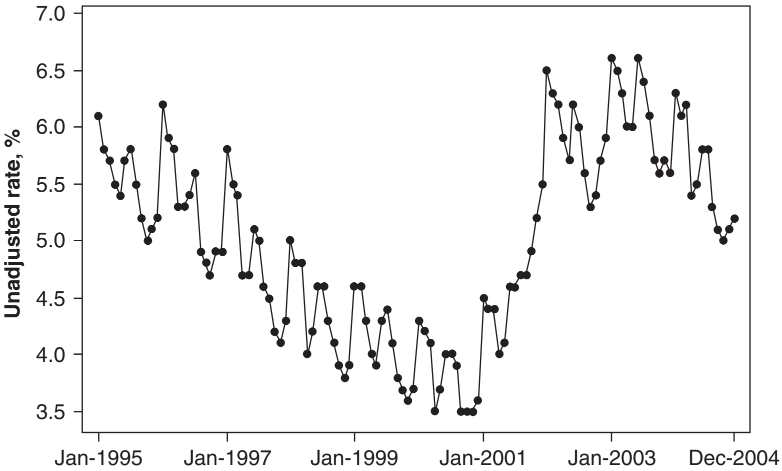


FIGURE 1.8 Monthly unemployment rate—full-time labor force, unadjusted. *Source:* US Department of Labor-BLS.

on the employment situation, for projections with impact on hiring and training, and for a multitude of other business planning activities. The data are reported unadjusted and with seasonal adjustment to remove the effect of regular patterns that occur each year.

The plot of monthly unadjusted unemployment rates (Figure 1.8) exhibits a mixture of patterns, similar to Figure 1.5 (data in Appendix B, Table B.8). There is a distinct cyclic pattern within a year; January, February, and March generally have the highest unemployment rates. The overall level is also changing, from a gradual decrease, to a steep increase, followed by a gradual decrease. The use of seasonal adjustments as described in Chapter 2 makes it easier to observe the nonseasonal movements in time series data.

Solar activity has long been recognized as a significant source of noise impacting consumer and military communications, including satellites, cell phone towers, and electric power grids. The ability to accurately forecast solar activity is critical to a variety of fields. The International Sunspot Number *R* is the oldest solar activity index. The number incorporates both the number of observed sunspots and the number of observed sunspot groups. In Figure 1.9, the plot of annual sunspot numbers reveals cyclic patterns of varying magnitudes (data in Appendix B, Table B.9).

In addition to assisting in the identification of steady-state patterns, time series plots may also draw attention to the occurrence of **atypical events.** Weekly sales of a generic pharmaceutical product dropped due to limited

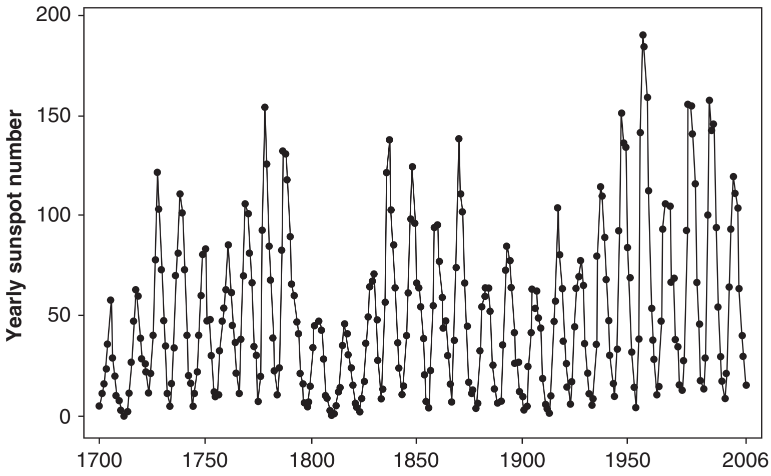


FIGURE 1.9 The international sunspot number. *Source:* SIDC.

availability resulting from a fire at one of the four production facilities. The 5-week reduction is apparent in the time series plot of weekly sales shown in Figure 1.10.

Another type of unusual event may be the failure of the data measurement or collection system. After recording a vastly different viscosity reading at time period 70 (Figure 1.11), the measurement system was

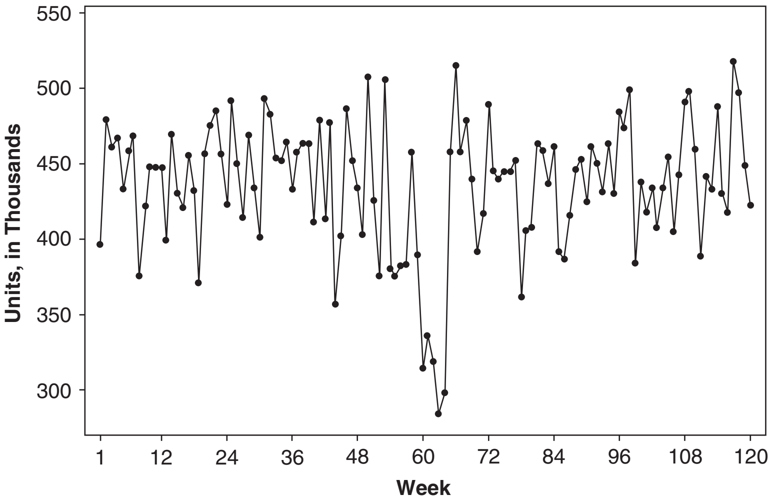


FIGURE 1.10 Pharmaceutical product sales.

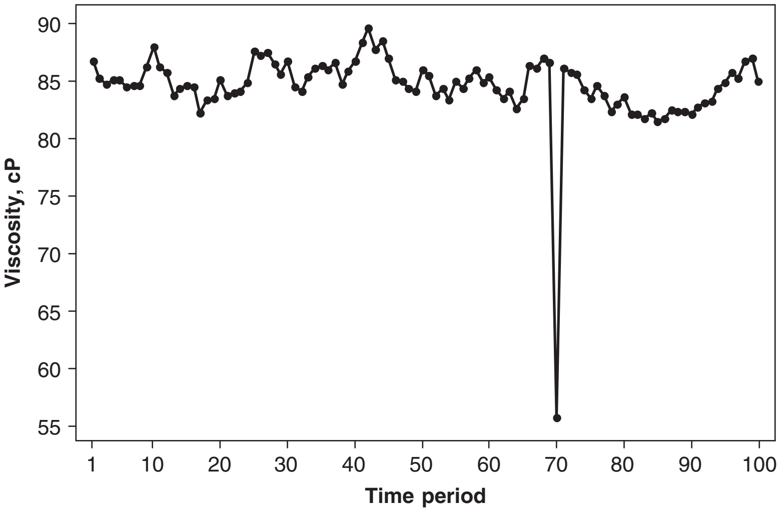


FIGURE 1.11 Chemical process viscosity readings, with sensor malfunction.

checked with a standard and determined to be out of calibration. The cause was determined to be a malfunctioning sensor.

1.3 THE FORECASTING PROCESS

A process is a series of connected activities that transform one or more inputs into one or more outputs. All work activities are performed in processes, and forecasting is no exception. The activities in the forecasting process are:

1. Problem definition
2. Data collection
3. Data analysis
4. Model selection and fitting
5. Model validation
6. Forecasting model deployment
7. Monitoring forecasting model performance

These activities are shown in Figure 1.12.

Problem definition involves developing understanding of how the forecast will be used along with the expectations of the "customer" (the user of

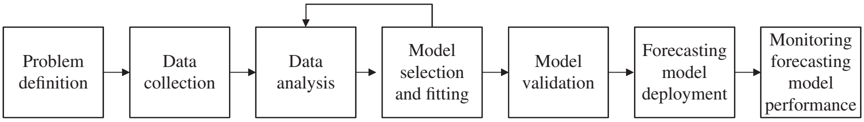

Problem definition
Data collection
Data analysis
Model selection and fitting
Model validation
Forecasting model deployment
Monitoring forecasting model performance

FIGURE 1.12 The forecasting process.

the forecast). Questions that must be addressed during this phase include the desired form of the forecast (e.g., are monthly forecasts required), the forecast horizon or lead time, how often the forecasts need to be revised (the forecast interval), and what level of forecast accuracy is required in order to make good business decisions. This is also an opportunity to introduce the decision makers to the use of prediction intervals as a measure of the risk associated with forecasts, if they are unfamiliar with this approach. Often it is necessary to go deeply into many aspects of the business system that requires the forecast to properly define the forecasting component of the entire problem. For example, in designing a forecasting system for inventory control, information may be required on issues such as product shelf life or other aging considerations, the time required to manufacture or otherwise obtain the products (production lead time), and the economic consequences of having too many or too few units of product available to meet customer demand. When multiple products are involved, the level of aggregation of the forecast (e.g., do we forecast individual products or families consisting of several similar products) can be an important consideration. Much of the ultimate success of the forecasting model in meeting the customer expectations is determined in the problem definition phase.

Data collection consists of obtaining the relevant history for the variable(s) that are to be forecast, including historical information on potential predictor variables.

The key here is "relevant"; often information collection and storage methods and systems change over time and not all historical data are useful for the current problem. Often it is necessary to deal with missing values of some variables, potential outliers, or other data-related problems that have occurred in the past. During this phase, it is also useful to begin planning how the data collection and storage issues in the future will be handled so that the reliability and integrity of the data will be preserved.

Data analysis is an important preliminary step to the selection of the forecasting model to be used. Time series plots of the data should be constructed and visually inspected for recognizable patterns, such as trends and seasonal or other cyclical components. A trend is evolutionary movement, either upward or downward, in the value of the variable. Trends may be long-term or more dynamic and of relatively short duration. Seasonality is the component of time series behavior that repeats on a regular basis, such as each year. Sometimes we will smooth the data to make identification of the patterns more obvious (data smoothing will be discussed in Chapter 2). Numerical summaries of the data, such as the sample mean, standard deviation, percentiles, and autocorrelations, should also be computed and evaluated. Chapter 2 will provide the necessary background to do this. If potential predictor variables are available, scatter plots of each pair of variables should be examined. Unusual data points or potential **outliers** should be identified and flagged for possible further study. The purpose of this preliminary data analysis is to obtain some "feel" for the data, and a sense of how strong the underlying patterns such as trend and seasonality are. This information will usually suggest the initial types of quantitative forecasting methods and models to explore.

Model selection and fitting consists of choosing one or more forecasting models and fitting the model to the data. **By fitting**, we mean estimating the unknown model parameters, usually by the method of least squares. In subsequent chapters, we will present several types of time series models and discuss the procedures of model fitting. We will also discuss methods for evaluating the quality of the model fit, and determining if any of the underlying assumptions have been violated. This will be useful in discriminating between different candidate models.

Model validation consists of an evaluation of the forecasting model to determine how it is likely to perform in the intended application. This must go beyond just evaluating the "fit" of the model to the historical data and must examine what magnitude of forecast errors will be experienced when the model is used to forecast "fresh" or new data. The fitting errors will always be smaller than the forecast errors, and this is an important concept that we will emphasize in this book. A widely used method for validating a forecasting model before it is turned over to the customer is to employ some form of **data splitting,** where the data are divided into two segments—a fitting segment and a forecasting segment. The model is fit to only the fitting data segment, and then forecasts from that model are simulated for the observations in the forecasting segment. This can provide useful guidance on how the forecasting model will perform when exposed to new data and can be a valuable approach for discriminating between competing forecasting models.

Forecasting model deployment involves getting the model and the resulting forecasts in use by the customer. It is important to ensure that the customer understands how to use the model and that generating timely forecasts from the model becomes as routine as possible. Model maintenance, including making sure that data sources and other required information will continue to be available to the customer is also an important issue that impacts the timeliness and ultimate usefulness of forecasts.

Monitoring forecasting model performance should be an ongoing activity after the model has been deployed to ensure that it is still performing satisfactorily. It is the nature of forecasting that conditions change over time, and a model that performed well in the past may deteriorate in performance. Usually performance deterioration will result in larger or more systematic forecast errors. Therefore monitoring of forecast errors is an essential part of good forecasting system design. **Control charts** of forecast errors are a simple but effective way to routinely monitor the performance of a forecasting model. We will illustrate approaches to monitoring forecast errors in subsequent chapters.

1.4 DATA FOR FORECASTING

1.4.1 The Data Warehouse

Developing time series models and using them for forecasting requires data on the variables of interest to decision-makers. The data are the raw materials for the modeling and forecasting process. The terms **data** and **information** are often used interchangeably, but we prefer to use the term data as that seems to reflect a more raw or original form, whereas we think of information as something that is extracted or synthesized from data. The output of a forecasting system could be thought of as information, and that output uses data as an input.

In most modern organizations data regarding sales, transactions, company financial and business performance, supplier performance, and customer activity and relations are stored in a repository known as a **data warehouse**. Sometimes this is a single data storage system; but as the volume of data handled by modern organizations grows rapidly, the data warehouse has become an integrated system comprised of components that are physically and often geographically distributed, such as cloud data storage. The data warehouse must be able to organize, manipulate, and integrate data from multiple sources and different organizational information systems. The basic functionality required includes data extraction, data transformation, and data loading. Data extraction refers to obtaining data from internal sources and from external sources such as third party vendors or government entities and financial service organizations. Once the data are extracted, the transformation stage involves applying rules to prevent duplication of records and dealing with problems such as missing information. Sometimes we refer to the transformation activities as **data**

cleaning. We will discuss some of the important data cleaning operations subsequently. Finally, the data are loaded into the data warehouse where they are available for modeling and analysis.

Data quality has several dimensions. Five important ones that have been described in the literature are accuracy, timeliness, completeness, representativeness, and consistency. Accuracy is probably the oldest dimension of data quality and refers to how close that data conform to its “real” values. Real values are alternative sources that can be used for verification purposes. For example, do sales records match payments to accounts receivable records (although the financial records may occur in later time periods because of payment terms and conditions, discounts, etc.)? Timeliness means that the data are as current as possible. Infrequent updating of data can seriously impact developing a time series model that is going to be used for relatively short-term forecasting. In many time series model applications the time between the occurrence of the real-world event and its entry into the data warehouse must be as short as possible to facilitate model development and use. Completeness means that the data content is complete, with no missing data and no outliers. As an example of representativeness, suppose that the end use of the time series model is to forecast customer demand for a product or service, but the organization only records booked orders and the date of fulfillment. This may not accurately reflect demand, because the orders can be booked before the desired delivery period and the date of fulfillment can take place in a different period than the one required by the customer. Furthermore, orders that are lost because of product unavailability or unsatisfactory delivery performance are not recorded. In these situations demand can differ dramatically from sales. Data cleaning methods can often be used to deal with some problems of completeness. Consistency refers to how closely data records agree over time in format, content, meaning, and structure. In many organizations how data are collected and stored evolves over time; definitions change and even the types of data that are collected change. For example, consider monthly data. Some organizations define “months” that coincide with the traditional calendar definition. But because months have different numbers of days that can induce patterns in monthly data, some organizations prefer to define a year as consisting of 13 “months” each consisting of 4 weeks.

It has been suggested that the output data that reside in the data warehouse are similar to the output of a manufacturing process, where the raw data are the input. Just as in manufacturing and other service processes, the data production process can benefit by the application of quality management and control tools. Jones-Farmer et al. (2014) describe how statistical quality control methods, specifically control charts, can be used to enhance data quality in the data production process.

1.4.2 Data Cleaning

Data cleaning is the process of examining data to detect potential errors, missing data, outliers or unusual values, or other inconsistencies and then correcting the errors or problems that are found. Sometimes errors are the result of recording or transmission problems, and can be corrected by working with the original data source to correct the problem. Effective data cleaning can greatly improve the forecasting process.

Before data are used to develop a time series model, it should be subjected to several different kinds of checks, including but not necessarily limited to the following:

1. Is there missing data?
2. Does the data fall within an expected range?
3. Are there potential outliers or other unusual values?

These types of checks can be automated fairly easily. If this aspect of data cleaning is automated, the rules employed should be periodically evaluated to ensure that they are still appropriate and that changes in the data have not made some of the procedures less effective. However, it is also extremely useful to use graphical displays to assist in identifying unusual data. Techniques such as time series plots, histograms, and scatter diagrams are extremely useful. These and other graphical methods will be described in Chapter 2.

1.4.3 Imputation

Data **imputation** is the process of correcting missing data or replacing outliers with an estimation process. Imputation replaces missing or erroneous values with a “likely” value based on other available information. This enables the analysis to work with statistical techniques which are designed to handle the complete data sets.

Mean value imputation consists of replacing a missing value with the sample average calculated from the nonmissing observations. The big advantage of this method is that it is easy, and if the data does not have any specific trend or seasonal pattern, it leaves the sample mean of the complete data set unchanged. However, one must be careful if there are trends or seasonal patterns, because the sample mean of all of the data may not reflect these patterns. A variation of this is **stochastic mean value imputation**, in which a random variable is added to the mean value to capture some of the noise or variability in the data. The random variable could be assumed to follow a normal distribution with mean zero and standard deviation equal to the standard deviation of the actual observed data. A variation of mean value imputation is to use a subset of the available historical data that reflects any trend or seasonal patterns in the data. For example, consider the time series $y_1, y_2, \ldots, y_T$ and suppose that one observation y_j is missing. We can impute the missing value as

$$y_j^* = \frac{1}{2k}\left(\sum_{t=j-k}^{j-1} y_t + \sum_{t-j+1}^{j+k} y_t\right),$$

where k would be based on the seasonal variability in the data. It is usually chosen as some multiple of the smallest seasonal cycle in the data. So, if the data are monthly and exhibit a monthly cycle, k would be a multiple of 12. **Regression imputation** is a variation of mean value imputation where the imputed value is computed from a model used to predict the missing value. The prediction model does not have to be a linear regression model. For example, it could be a time series model.

Hot deck imputation is an old technique that is also known as the last value carried forward method. The term "hot deck" comes from the use of computer punch cards. The deck of cards was "hot" because it was currently in use. **Cold deck imputation** uses information from a deck of cards not currently in use. In hot deck imputation, the missing values are imputed by using values from similar complete observations. If there are several variables, sort the data by the variables that are most related to the missing observation and then, starting at the top, replace the missing values with the value of the immediately preceding variable. There are many variants of this procedure.

1.5 RESOURCES FOR FORECASTING

There are a variety of good resources that can be helpful to technical professionals involved in developing forecasting models and preparing forecasts. There are three professional journals devoted to forecasting:

- *Journal of Forecasting*
- *International Journal of Forecasting*
- *Journal of Business Forecasting Methods and Systems*

These journals publish a mixture of new methodology, studies devoted to the evaluation of current methods for forecasting, and case studies and applications. In addition to these specialized forecasting journals, there are several other mainstream statistics and operations research/management science journals that publish papers on forecasting, including:

- *Journal of Business and Economic Statistics*
- *Management Science*
- *Naval Research Logistics*
- *Operations Research*
- *International Journal of Production Research*
- *Journal of Applied Statistics*

This is by no means a comprehensive list. Research on forecasting tends to be published in a variety of outlets.

There are several books that are good complements to this one. We recommend Box, Jenkins, and Reinsel (1994); Chatfield (1996); Fuller (1995); Abraham and Ledolter (1983); Montgomery, Johnson, and Gardiner (1990); Wei (2006); and Brockwell and Davis (1991, 2002). Some of these books are more specialized than this one, in that they focus on a specific type of forecasting model such as the autoregressive integrated moving average [ARIMA] model, and some also require more background in statistics and mathematics.

Many statistics software packages have very good capability for fitting a variety of forecasting models. Minitab® Statistical Software, JMP®, the Statistical Analysis System (SAS) and R are the packages that we utilize and illustrate in this book. At the end of most chapters we provide R code for working some of the examples in the chapter. Matlab and S-Plus are also two packages that have excellent capability for solving forecasting problems.

EXERCISES

1.1 Why is forecasting an essential part of the operation of any organization or business?

1.2 What is a time series? Explain the meaning of trend effects, seasonal variations, and random error.

1.3 Explain the difference between a point forecast and an interval forecast.

1.4 What do we mean by a causal forecasting technique?

1.5 Everyone makes forecasts in their daily lives. Identify and discuss a situation where you employ forecasts.

a. What decisions are impacted by your forecasts?

b. How do you evaluate the quality of your forecasts?

c. What is the value to you of a good forecast?

d. What is the harm or penalty associated with a bad forecast?

1.6 What is meant by a rolling horizon forecast?

1.7 Explain the difference between forecast horizon and forecast interval.

1.8 Suppose that you are in charge of capacity planning for a large electric utility. A major part of your job is ensuring that the utility has sufficient generating capacity to meet current and future customer needs. If you do not have enough capacity, you run the risks of brownouts and service interruption. If you have too much capacity, it may cost more to generate electricity.

a. What forecasts do you need to do your job effectively?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

1.9 Your company designs and manufactures apparel for the North American market. Clothing and apparel is a style good, with a relatively limited life. Items not sold at the end of the season are usually sold through off-season outlet and discount retailers. Items not sold through discounting and off-season merchants are often given to charity or sold abroad.

a. What forecasts do you need in this business to be successful?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

d. What are the implications of forecast errors?

1.10 Suppose that you are in charge of production scheduling at a semiconductor manufacturing plant. The plant manufactures about 20 different types of devices, all on 8-inch silicon wafers. Demand for these products varies randomly. When a lot or batch of wafers is started into production, it can take from 4 to 6 weeks before the batch is finished, depending on the type of product. The routing of each batch of wafers through the production tools can be different depending on the type of product.

a. What forecasts do you need in this business to be successful?

b. Are these short-range or long-range forecasts?

c. What data do you need to be able to generate these forecasts?

d. Discuss the impact that forecast errors can potentially have on the efficiency with which your factory operates, including work-in-process inventory, meeting customer delivery schedules, and the cycle time to manufacture product.

1.11 You are the administrator of a large metropolitan hospital that operates the only 24-hour emergency room in the area. You must schedule attending physicians, resident physicians, nurses, laboratory, and support personnel to operate this facility effectively.

a. What measures of effectiveness do you think patients use to evaluate the services that you provide?

b. How are forecasts useful to you in planning services that will maximize these measures of effectiveness?

c. What planning horizon do you need to use? Does this lead to short-range or long-range forecasts?

1.12 Consider an airline that operates a network of flights that serves 200 cities in the continental United States. What long-range forecasts do the operators of the airline need to be successful? What forecasting problems does this business face on a daily basis? What are the consequences of forecast errors for the airline?

1.13 Discuss the potential difficulties of forecasting the daily closing price of a specific stock on the New York Stock Exchange. Would the problem be different (harder, easier) if you were asked to forecast the closing price of a group of stocks, all in the same industry (say, the pharmaceutical industry)?

1.14 Explain how large forecast errors can lead to high inventory levels at a retailer; at a manufacturing plant.

1.15 Your company manufactures and distributes soft drink beverages, sold in bottles and cans at retail outlets such as grocery stores, restaurants and other eating/drinking establishments, and vending machines in offices, schools, stores, and other outlets. Your product line includes about 25 different products, and many of these are produced in different package sizes.

a. What forecasts do you need in this business to be successful?

b. Is the demand for your product likely to be seasonal? Explain why or why not?

c. Does the shelf life of your product impact the forecasting problem?

d. What data do you think that you would need to be able to produce successful forecasts?

CHAPTER 2

STATISTICS BACKGROUND FOR FORECASTING

The future ain't what it used to be.

YOGI BERRA, *New York Yankees catcher*

2.1 INTRODUCTION

This chapter presents some basic statistical methods essential to modeling, analyzing, and forecasting time series data. Both graphical displays and numerical summaries of the properties of time series data are presented. We also discuss the use of data transformations and adjustments in forecasting and some widely used methods for characterizing and monitoring the performance of a forecasting model. Some aspects of how these performance measures can be used to select between competing forecasting techniques are also presented.

Forecasts are based on data or observations on the variable of interest. These data are usually in the form of a **time series**. Suppose that there are T periods of data available, with period T being the most recent. We will let the observation on this variable at time period t be denoted by y_t, $t = 1, 2, \dots, T$. This variable can represent a cumulative quantity, such as the

Introduction to Time Series Analysis and Forecasting, Second Edition.
Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci.

total demand for a product during period t, or an instantaneous quantity, such as the daily closing price of a specific stock on the New York Stock Exchange.

Generally, we will need to distinguish between a **forecast** or **predicted value** of y_t that was made at some previous time period, say, $t - \tau$, and a **fitted value** of y_t that has resulted from estimating the parameters in a time series model to historical data. Note that τ is the forecast lead time. The forecast made at time period $t - \tau$ is denoted by $\hat{y}_t(t - \tau)$. There is a lot of interest in the **lead − 1** forecast, which is the forecast of the observation in period t, y_t, made one period prior, $\hat{y}_t(t - 1)$. We will denote the fitted value of y_t by $\hat{y}_t$.

We will also be interested in analyzing **forecast errors**. The forecast error that results from a forecast of y_t that was made at time period $t - \tau$ is the **lead − τ forecast error**

$$e_t(\tau) = y_t - \hat{y}_t(t - \tau). \tag{2.1}$$

For example, the lead − 1 forecast error is

$$e_t(1) = y_t - \hat{y}_t(t - 1).$$

The difference between the observation y_t and the value obtained by fitting a time series model to the data, or a fitted value $\hat{y}_t$ defined earlier, is called a **residual**, and is denoted by

$$e_t = y_t - \hat{y}_t. \tag{2.2}$$

The reason for this careful distinction between forecast errors and residuals is that models usually fit historical data better than they forecast. That is, the residuals from a model-fitting process will almost always be smaller than the forecast errors that are experienced when that model is used to forecast future observations.

2.2 GRAPHICAL DISPLAYS

2.2.1 Time Series Plots

Developing a forecasting model should always begin with graphical display and analysis of the available data. Many of the broad general features of a time series can be seen visually. This is not to say that analytical tools are not useful, because they are, but the human eye can be a very sophisticated data analysis tool. To paraphrase the great New York Yankees catcher Yogi Berra, “You can observe a lot just by watching.”

The basic graphical display for time series data is the **time series plot**, illustrated in Chapter 1. This is just a graph of y_t versus the time period, t, for $t = 1, 2, \ldots, T$. Features such as trend and seasonality are usually easy to see from the time series plot. It is interesting to observe that some of the classical tools of descriptive statistics, such as the histogram and the stem-and-leaf display, are not particularly useful for time series data because they do not take time order into account.

Example 2.1 Figures 2.1 and 2.2 show time series plots for viscosity readings and beverage production shipments (originally shown in Figures 1.3 and 1.5, respectively). At the right-hand side of each time series plot is a histogram of the data. Note that while the two time series display very different characteristics, the histograms are remarkably similar. Essentially, the histogram summarizes the data across the time dimension, and in so doing, the key time-dependent features of the data are lost. Stem-and-leaf plots and boxplots would have the same issues, losing time-dependent features.

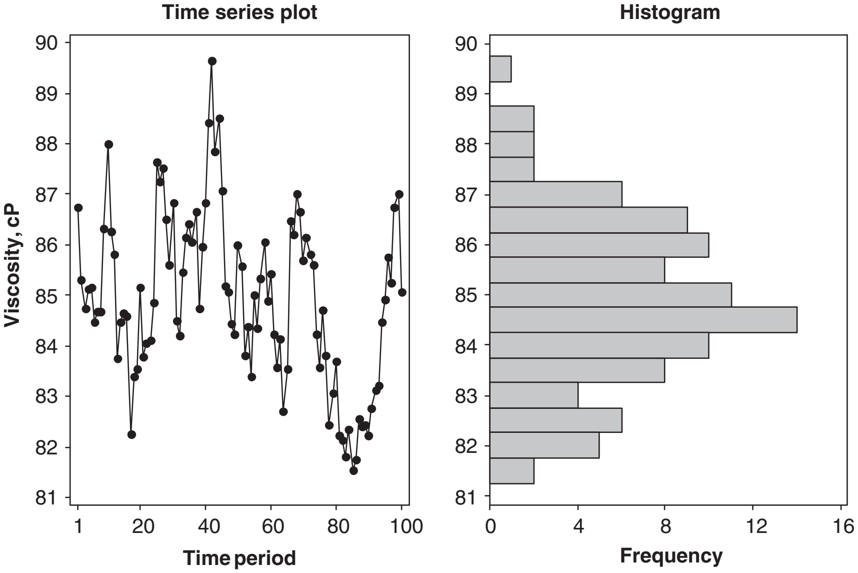


FIGURE 2.1 Time series plot and histogram of chemical process viscosity readings.

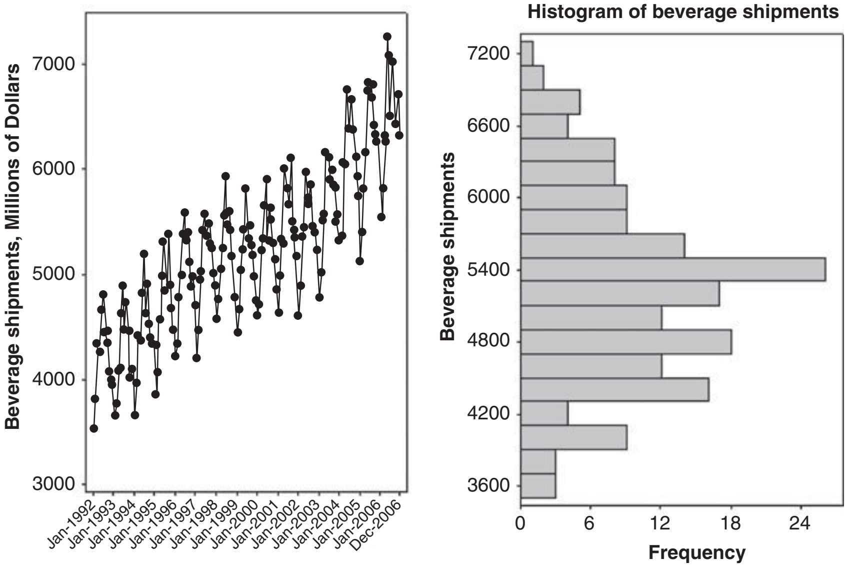


FIGURE 2.2 Time series plot and histogram of beverage production shipments.

When there are two or more variables of interest, **scatter plots** can be useful in displaying the relationship between the variables. For example, Figure 2.3 is a scatter plot of the annual global mean surface air temperature anomaly first shown in Figure 1.6 versus atmospheric CO_2 concentrations. The scatter plot clearly reveals a relationship between the two variables:

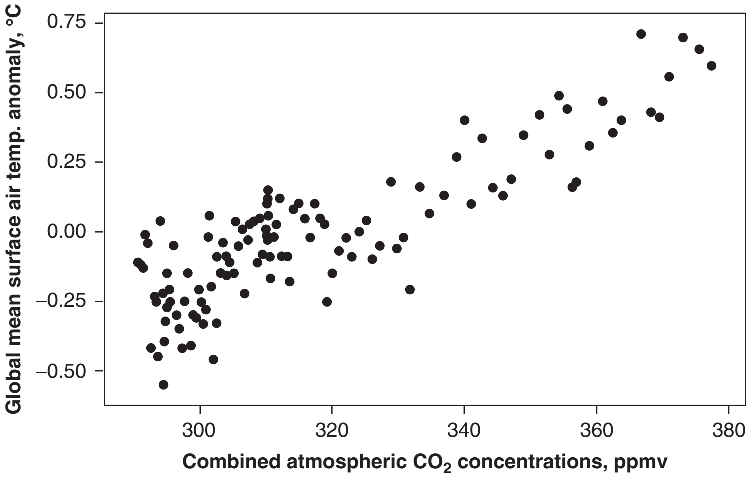


FIGURE 2.3 Scatter plot of temperature anomaly versus CO_2 concentrations. *Sources*: NASA–GISS (anomaly), DOE–DIAC (CO_2).

low concentrations of CO_2 are usually accompanied by negative anomalies, and higher concentrations of CO_2 tend to be accompanied by positive anomalies. Note that this does not imply that higher concentrations of CO_2 actually *cause* higher temperatures. The scatter plot cannot establish a causal relationship between two variables (neither can naive statistical modeling techniques, such as regression), but it is useful in displaying how the variables have varied together in the historical data set.

There are many variations of the time series plot and other graphical displays that can be constructed to show specific features of a time series. For example, Figure 2.4 displays daily price information for Whole Foods Market stock during the first quarter of 2001 (the trading days from January 2, 2001 through March 30, 2001). This chart, created in Excel®, shows the opening, closing, highest, and lowest prices experienced within a trading day for the first quarter. If the opening price was higher than the closing price, the box is filled, whereas if the closing price was higher than the opening price, the box is open. This type of plot is potentially more useful than a time series plot of just the closing (or opening) prices, because it shows the volatility of the stock within a trading day. The volatility of an asset is often of interest to investors because it is a measure of the inherent risk associated with the asset.

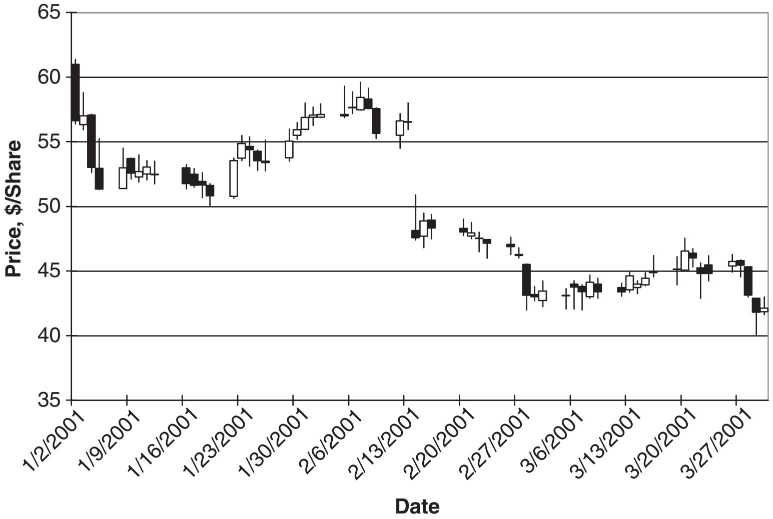


FIGURE 2.4 Open-high/close-low chart of Whole Foods Market stock price. *Source*: `finance.yahoo.com`.

2.2.2 Plotting Smoothed Data

Sometimes it is useful to overlay a **smoothed** version of the original data on the original time series plot to help reveal patterns in the original data. There are several types of data smoothers that can be employed. One of the simplest and most widely used is the ordinary or simple moving average. A simple **moving average** of span N assigns weights $1/N$ to the most recent N observations $y_T, y_{T-1}, \dots, y_{T-N+1}$, and weight zero to all other observations. If we let M_T be the moving average, then the N-span moving average at time period T is

$$M_T = \frac{y_T + y_{T-1} + \cdots + y_{T-N+1}}{N} = \frac{1}{N} \sum_{t=T-N+1}^{T} y_t \tag{2.3}$$

Clearly, as each new observation becomes available it is added into the sum from which the moving average is computed and the oldest observation is discarded. The moving average has less variability than the original observations; in fact, if the variance of an individual observation y_t is σ^2, then assuming that the observations are uncorrelated the variance of the moving average is

$$\text{Var}(M_T) = \text{Var}\left(\frac{1}{N} \sum_{t=T-N+1}^{N} y_t \right) = \frac{1}{N^2} \sum_{t=T-N+1}^{N} \text{Var}(y_t) = \frac{\sigma^2}{N}$$

Sometimes a “centered” version of the moving average is used, such as in

$$M_t = \frac{1}{S+1} \sum_{i=-S}^{S} y_{t-i} \tag{2.4}$$

where the span of the centered moving average is $N = 2S + 1$.

Example 2.2 Figure 2.5 plots the annual global mean surface air temperature anomaly data along with a five-period (a period is 1 year) moving average of the same data. Note that the moving average exhibits less variability than found in the original series. It also makes some features of the data easier to see; for example, it is now more obvious that the global air temperature decreased from about 1940 until about 1975.

Plots of moving averages are also used by analysts to evaluate stock price trends; common MA periods are 5, 10, 20, 50, 100, and 200 days. A time series plot of Whole Foods Market stock price with a 50-day moving

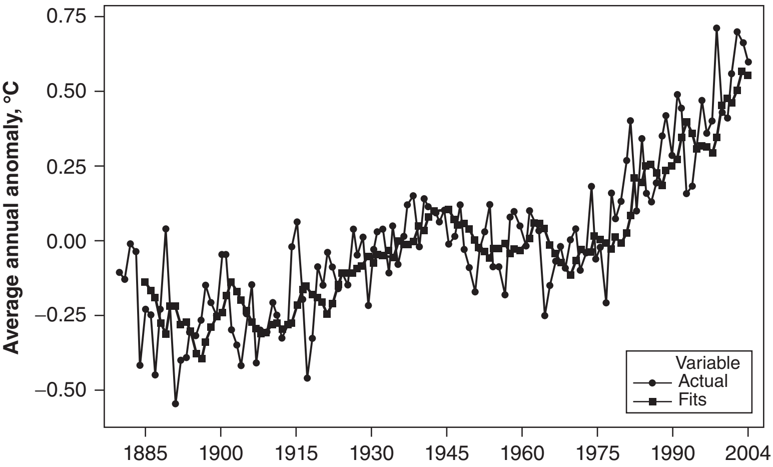


FIGURE 2.5 Time series plot of global mean surface air temperature anomaly, with five-period moving average. *Source*: NASA–GISS.

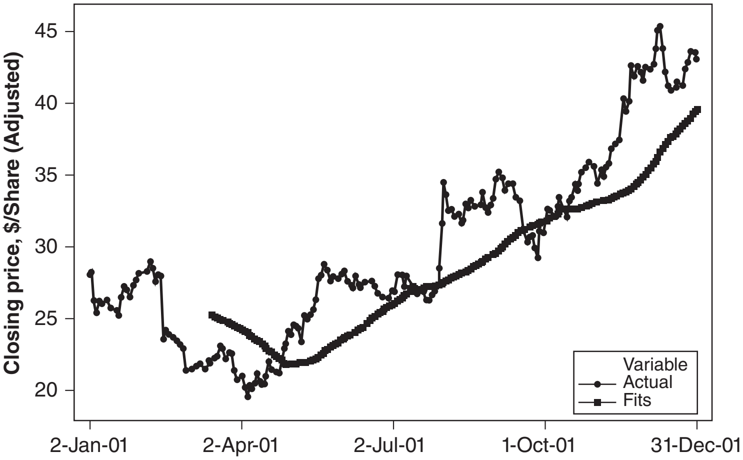


FIGURE 2.6 Time series plot of Whole Foods Market stock price, with 50-day moving average. *Source*: `finance.yahoo.com`.

average is shown in Figure 2.6. The moving average plot smoothes the day-to-day noise and shows a generally increasing trend.

The simple moving average is a **linear data smoother**, or a **linear filter**, because it replaces each observation y_t with a linear combination of the other data points that are near to it in time. The weights in the linear combination are equal, so the linear combination here is an average. Of course, unequal weights could be used. For example, the **Hanning filter** is a weighted, centered moving average

$$M_t^{\mathrm{H}} = 0.25y_{t+1} + 0.5y_t + 0.25y_{t-1}$$

Julius von Hann, a nineteenth century Austrian meteorologist, used this filter to smooth weather data.

An obvious disadvantage of a linear filter such as a moving average is that an unusual or erroneous data point or an outlier will dominate the moving averages that contain that observation, contaminating the moving averages for a length of time equal to the span of the filter. For example, consider the sequence of observations

$$15, 18, 13, 12, 16, 14, 16, 17, 18, 15, 18, 200, 19, 14, 21, 24, 19, 25$$

which increases reasonably steadily from 15 to 25, except for the unusual value 200. Any reasonable smoothed version of the data should also increase steadily from 15 to 25 and not emphasize the value 200. Now even if the value 200 is a legitimate observation, and not the result of a data recording or reporting error (perhaps it should be 20!), it is so unusual that it deserves special attention and should likely not be analyzed along with the rest of the data.

Odd-span **moving medians** (also called **running medians**) are an alternative to moving averages that are effective data smoothers when the time series may be contaminated with unusual values or outliers. The moving median of span N is defined as

$$m_t^{[N]} = med(y_{t-u}, \dots, y_t, \dots, y_{t+u}), \tag{2.5}$$

where $N = 2u + 1$. The median is the middle observation in rank order (or order of value). The moving median of span 3 is a very popular and effective data smoother, where

$$m_t^{[3]} = med(y_{t-1}, y_t, y_{t+1}).$$

This smoother would process the data three values at a time, and replace the three original observations by their median. If we apply this smoother to the data above, we obtain

$$\_\_\_\_, 15, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, 24, \_\_\_\_.$$

This smoothed data are a reasonable representation of the original data, but they conveniently ignore the value 200. The end values are lost when using the moving median, and they are represented by "____".

In general, a moving median will pass monotone sequences of data unchanged. It will follow a step function in the data, but it will eliminate a spike or more persistent upset in the data that has duration of at most u consecutive observations. Moving medians can be applied more than once if desired to obtain an even smoother series of observations. For example, applying the moving median of span 3 to the smoothed data above results in

$$\_\_\_\_, \_\_\_\_, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, \_\_\_\_, \_\_\_\_.$$

These data are now as smooth as it can get; that is, repeated application of the moving median will not change the data, apart from the end values.

If there are a lot of observations, the information loss from the missing end values is not serious. However, if it is necessary or desirable to keep the lengths of the original and smoothed data sets the same, a simple way to do this is to "copy on" or add back the end values from the original data. This would result in the smoothed data:

$$15, 18, 13, 13, 14, 16, 17, 17, 18, 18, 19, 19, 19, 21, 21, 19, 25$$

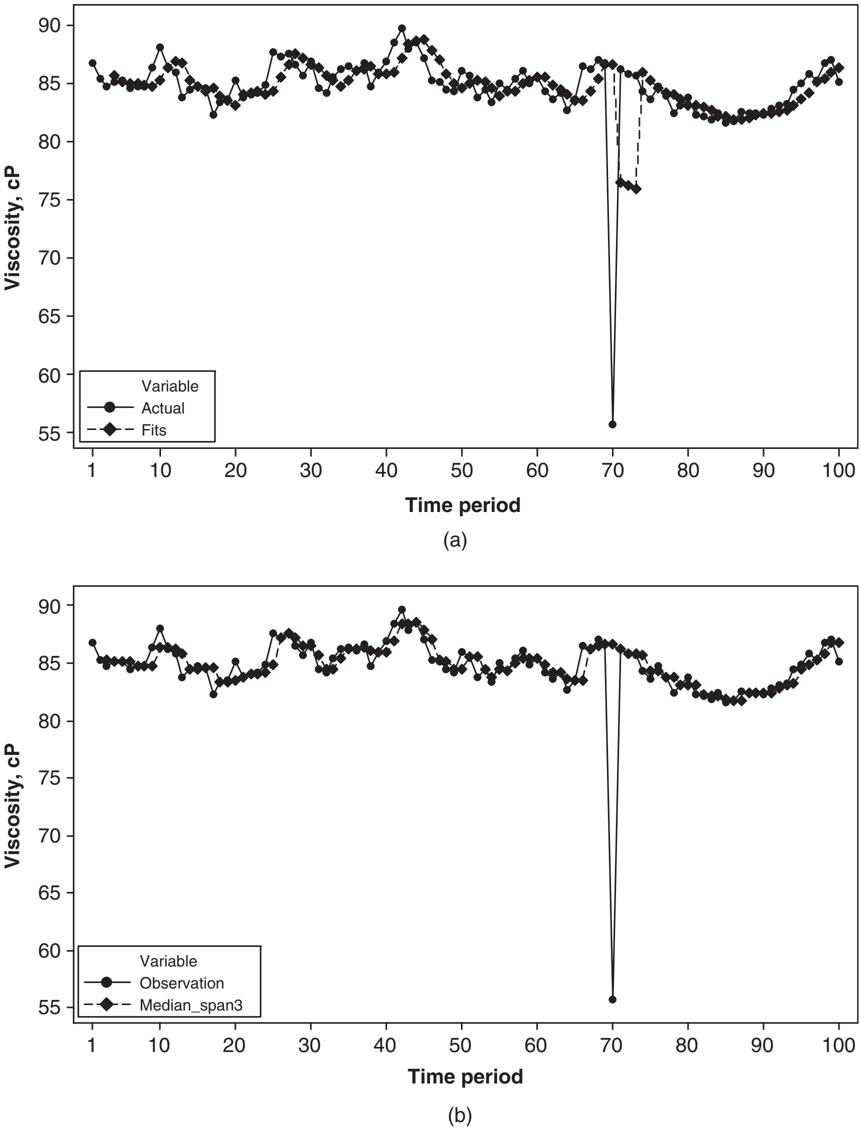
There are also methods for smoothing the end values. Tukey (1979) is a basic reference on this subject and contains many other clever and useful techniques for data analysis.

Example 2.3 The chemical process viscosity readings shown in Figure 1.11 are an example of a time series that benefits from smoothing to evaluate patterns. The selection of a moving median over a moving average, as shown in Figure 2.7, minimizes the impact of the invalid measurements, such as the one at time period 70.

2.3 NUMERICAL DESCRIPTION OF TIME SERIES DATA

2.3.1 Stationary Time Series

A very important type of time series is a **stationary** time series. A time series is said to be **strictly stationary** if its properties are not affected



(a)

(b)

FIGURE 2.7 Viscosity readings with (a) moving average and (b) moving median.

by a change in the time origin. That is, if the joint probability distribution of the observations $y_t, y_{t+1}, \dots, y_{t+n}$ is exactly the same as the joint probability distribution of the observations $y_{t+k}, y_{t+k+1}, \dots, y_{t+k+n}$ then the time series is strictly stationary. When $n = 0$ the stationarity assumption means that the probability distribution of y_t is the same for all time periods

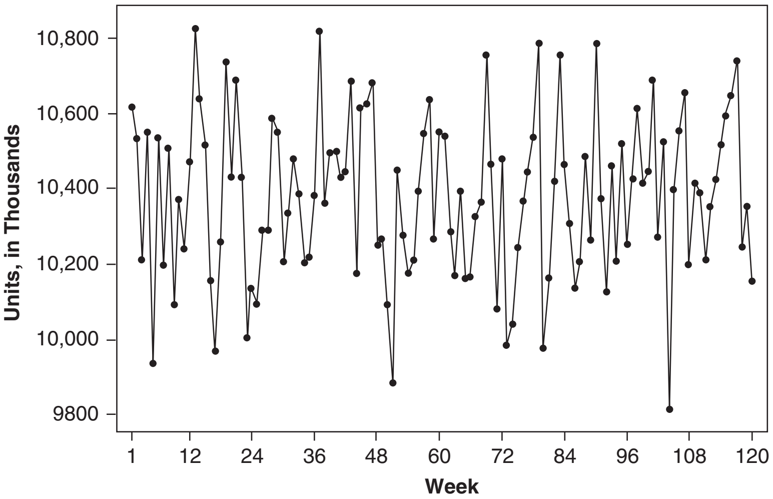


FIGURE 2.8 Pharmaceutical product sales.

and can be written as $f(y)$. The pharmaceutical product sales and chemical viscosity readings time series data originally shown in Figures 1.2 and 1.3, respectively, are examples of stationary time series. The time series plots are repeated in Figures 2.8 and 2.9 for convenience. Note that both time series seem to vary around a fixed level. Based on the earlier definition, this is a characteristic of stationary time series. On the other hand, the Whole

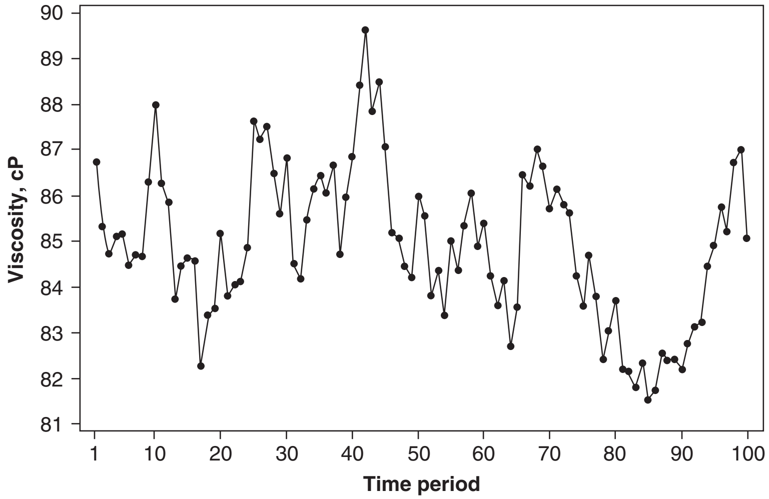


FIGURE 2.9 Chemical process viscosity readings.

Foods Market stock price data in Figure 1.7 tends to wander around or drift, with no obvious fixed level. This is behavior typical of a nonstationary time series.

Stationary implies a type of statistical **equilibrium** or **stability** in the data. Consequently, the time series has a constant mean defined in the usual way as

$$\mu_y = E(y) = \int_{-\infty}^{\infty} y f(y) dy \tag{2.6}$$

and constant variance defined as

$$\sigma_y^2 = \mathrm{Var}(y) = \int_{-\infty}^{\infty} (y - \mu_y)^2 f(y) dy. \tag{2.7}$$

The sample mean and sample variance are used to estimate these parameters. If the observations in the time series are $y_1, y_2, \ldots, y_T$, then the sample mean is

$$\bar{y} = \hat{\mu}_y = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{2.8}$$

and the sample variance is

$$s^2 = \hat{\sigma}_y^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})^2. \tag{2.9}$$

Note that the divisor in Eq. (2.9) is T rather than the more familiar $T - 1$. This is the common convention in many time series applications, and because T is usually not small, there will be little difference between using T instead of $T - 1$.

2.3.2 Autocovariance and Autocorrelation Functions

If a time series is stationary this means that the joint probability distribution of any two observations, say, y_t and y_{t+k}, is the same for any two time periods t and $t + k$ that are separated by the same interval k. Useful information about this joint distribution, and hence about the nature of the time series, can be obtained by plotting a scatter diagram of all of the data pairs y_t, y_{t+k} that are separated by the same interval k. The interval k is called the **lag**.

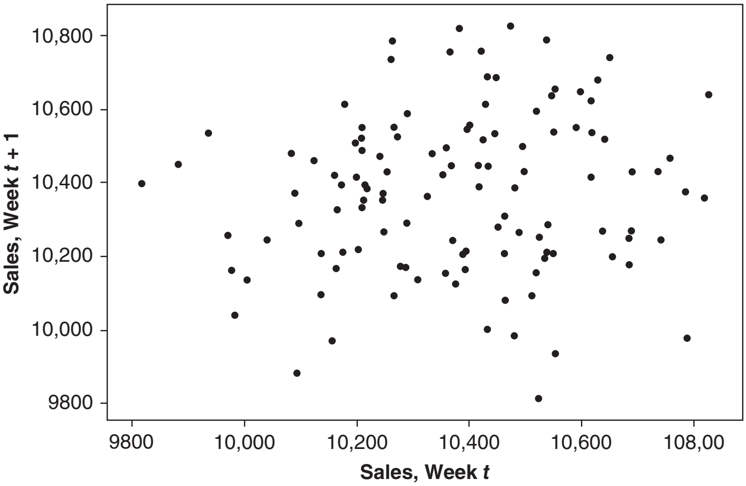


FIGURE 2.10 Scatter diagram of pharmaceutical product sales at lag $k = 1$.

Example 2.4 Figure 2.10 is a scatter diagram for the pharmaceutical product sales for lag $k = 1$ and Figure 2.11 is a scatter diagram for the chemical viscosity readings for lag $k = 1$. Both scatter diagrams were constructed by plotting y_{t+1} versus y_t. Figure 2.10 exhibits little structure; the plotted pairs of adjacent observations y_t, y_{t+1} seem to be **uncorrelated**. That is, the value of y in the current period does not provide any useful information about the value of y that will be observed in the next period. A different story is revealed in Figure 2.11, where we observe that the

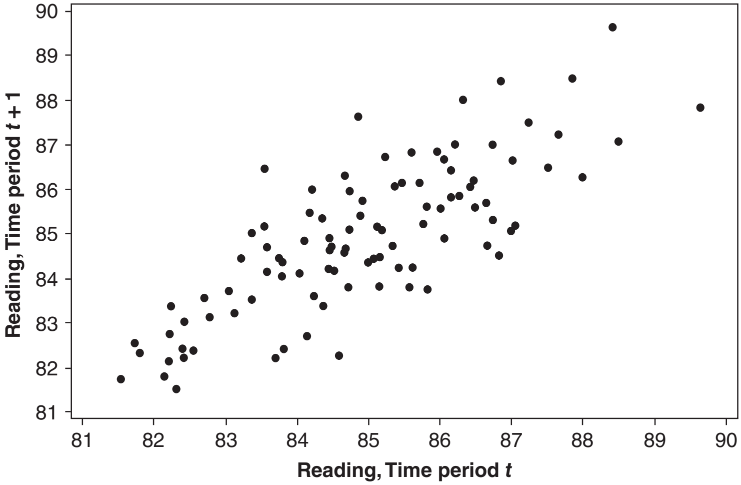


FIGURE 2.11 Scatter diagram of chemical viscosity readings at lag $k = 1$.

pairs of adjacent observations y_{t+1}, y_t are **positively correlated**. That is, a small value of y tends to be followed in the next time period by another small value of y, and a large value of y tends to be followed immediately by another large value of y. Note from inspection of Figures 2.10 and 2.11 that the behavior inferred from inspection of the scatter diagrams is reflected in the observed time series.

The covariance between y_t and its value at another time period, say, y_{t+k} is called the **autocovariance** at lag k, defined by

$$\gamma_k = \text{Cov}(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)]. \tag{2.10}$$

The collection of the values of $\gamma_k, k = 0, 1, 2, \ldots$ is called the **autocovariance function**. Note that the autocovariance at lag $k = 0$ is just the variance of the time series; that is, $\gamma_0 = \sigma_y^2$,which is constant for a stationary time series. The **autocorrelation coefficient** at lag k for a stationary time series is

$$\rho_k = \frac{E[(y_t - \mu)(y_{t+k} - \mu)]}{\sqrt{E[(y_t - \mu)^2]E[(y_{t+k} - \mu)^2]}} = \frac{\text{Cov}(y_t, y_{t+k})}{\text{Var}(y_t)} = \frac{\gamma_k}{\gamma_0}. \tag{2.11}$$

The collection of the values of $\rho_k,\ k = 0, 1, 2, \ldots$ is called the **autocorrelation function (ACF)**. Note that by definition $\rho_0 = 1$. Also, the ACF is independent of the scale of measurement of the time series, so it is a dimensionless quantity. Furthermore, $\rho_k = \rho_{-k}$; that is, the ACF is **symmetric** around zero, so it is only necessary to compute the positive (or negative) half.

If a time series has a finite mean and autocovariance function it is said to be second-order stationary (or weakly stationary of order 2). If, in addition, the joint probability distribution of the observations at all times is multivariate normal, then that would be sufficient to result in a time series that is strictly stationary.

It is necessary to estimate the autocovariance and ACFs from a time series of finite length, say, $y_1, y_2, \ldots, y_T$. The usual estimate of the autocovariance function is

$$c_k = \hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y}), \quad k = 0, 1, 2, \ldots, K \tag{2.12}$$

and the ACF is estimated by the **sample autocorrelation function** (or **sample ACF**)

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0}, \quad k = 0, 1, \ldots, K \tag{2.13}$$

A good general rule of thumb is that at least 50 observations are required to give a reliable estimate of the ACF, and the individual sample autocorrelations should be calculated up to lag K, where K is about $T/4$.

Often we will need to determine if the autocorrelation coefficient at a particular lag is zero. This can be done by comparing the sample autocorrelation coefficient at lag k, r_k, to its standard error. If we make the assumption that the observations are uncorrelated, that is, $\rho_k = 0$ for all k, then the variance of the sample autocorrelation coefficient is

$$\text{Var}(r_k) \cong \frac{1}{T} \tag{2.14}$$

and the standard error is

$$se(r_k) \cong \frac{1}{\sqrt{T}} \tag{2.15}$$

Example 2.5 Consider the chemical process viscosity readings plotted in Figure 2.9; the values are listed in Table 2.1.

The sample ACF at lag $k = 1$ is calculated as

$$\begin{aligned} c_0 &= \frac{1}{100}\sum_{t=1}^{100-0}(y_t - \bar{y})(y_{t+0} - \bar{y}) \\ &= \frac{1}{100}[(86.7418 - 84.9153)(86.7418 - 84.9153) + \cdots \\ &\quad + (85.0572 - 84.9153)(85.0572 - 84.9153)] \\ &= 280.9332 \end{aligned}$$

$$\begin{aligned} c_1 &= \frac{1}{100}\sum_{t=1}^{100-1}(y_t - \bar{y})(y_{t+1} - \bar{y}) \\ &= \frac{1}{100}[(86.7418 - 84.9153)(85.3195 - 84.9153) + \cdots \\ &\quad + (87.0048 - 84.9153)(85.0572 - 84.9153)] \\ &= 220.3137 \end{aligned}$$

$$r_1 = \frac{c_1}{c_0} = \frac{220.3137}{280.9332} = 0.7842$$

A plot and listing of the sample ACFs generated by Minitab for the first 25 lags are displayed in Figures 2.12 and 2.13, respectively.

TABLE 2.1 Chemical Process Viscosity Readings

Time Period	Reading	Time Period	Reading	Time Period	Reading	Time Period	Reading
1	86.7418	26	87.2397	51	85.5722	76	84.7052
2	85.3195	27	87.5219	52	83.7935	77	83.8168
3	84.7355	28	86.4992	53	84.3706	78	82.4171
4	85.1113	29	85.6050	54	83.3762	79	83.0420
5	85.1487	30	86.8293	55	84.9975	80	83.6993
6	84.4775	31	84.5004	56	84.3495	81	82.2033
7	84.6827	32	84.1844	57	85.3395	82	82.1413
8	84.6757	33	85.4563	58	86.0503	83	81.7961
9	86.3169	34	86.1511	59	84.8839	84	82.3241
10	88.0006	35	86.4142	60	85.4176	85	81.5316
11	86.2597	36	86.0498	61	84.2309	86	81.7280
12	85.8286	37	86.6642	62	83.5761	87	82.5375
13	83.7500	38	84.7289	63	84.1343	88	82.3877
14	84.4628	39	85.9523	64	82.6974	89	82.4159
15	84.6476	40	86.8473	65	83.5454	90	82.2102
16	84.5751	41	88.4250	66	86.4714	91	82.7673
17	82.2473	42	89.6481	67	86.2143	92	83.1234
18	83.3774	43	87.8566	68	87.0215	93	83.2203
19	83.5385	44	88.4997	69	86.6504	94	84.4510
20	85.1620	45	87.0622	70	85.7082	95	84.9145
21	83.7881	46	85.1973	71	86.1504	96	85.7609
22	84.0421	47	85.0767	72	85.8032	97	85.2302
23	84.1023	48	84.4362	73	85.6197	98	86.7312
24	84.8495	49	84.2112	74	84.2339	99	87.0048
25	87.6416	50	85.9952	75	83.5737	100	85.0572

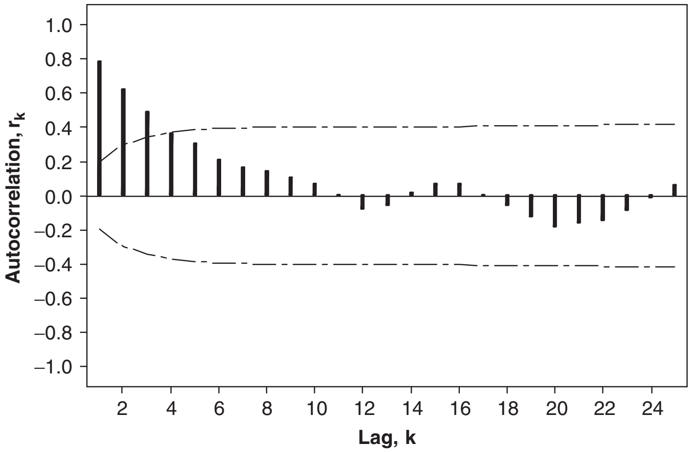


FIGURE 2.12 Sample autocorrelation function for chemical viscosity readings, with 5% significance limits.

```
Autocorrelation function: reading

Lag        ACF      T     LBQ
  1   0.784221   7.84   63.36
  2   0.628050   4.21  104.42
  3   0.491587   2.83  129.83
  4   0.362880   1.94  143.82
  5   0.304554   1.57  153.78
  6   0.208979   1.05  158.52
  7   0.164320   0.82  161.48
  8   0.144789   0.72  163.80
  9   0.103625   0.51  165.01
 10   0.066559   0.33  165.51
 11   0.003949   0.02  165.51
 12  –0.077226  –0.38  166.20
 13  –0.051953  –0.25  166.52
 14   0.020525   0.10  166.57
 15   0.072784   0.36  167.21
 16   0.070753   0.35  167.81
 17   0.001334   0.01  167.81
 18  –0.057435  –0.28  168.22
 19  –0.123122  –0.60  170.13
 20  –0.180546  –0.88  174.29
 21  –0.162466  –0.78  177.70
 22  –0.145979  –0.70  180.48
 23  –0.087420  –0.42  181.50
 24  –0.011579  –0.06  181.51
 25   0.063170   0.30  182.06
```

FIGURE 2.13 Listing of sample autocorrelation functions for first 25 lags of chemical viscosity readings, Minitab session window output (the definition of T and LBQ will be given later).

Note the rate of decrease or decay in ACF values in Figure 2.12 from 0.78 to 0, followed by a sinusoidal pattern about 0. This ACF pattern is typical of stationary time series. The importance of ACF estimates exceeding the 5% significance limits will be discussed in Chapter 5. In contrast, the plot of sample ACFs for a time series of random values with constant mean has a much different appearance. The sample ACFs for pharmaceutical product sales plotted in Figure 2.14 appear randomly positive or negative, with values near zero.

While the ACF is strictly speaking defined only for a stationary time series, the sample ACF can be computed for *any* time series, so a logical question is: What does the sample ACF of a nonstationary time series look like? Consider the daily closing price for Whole Foods Market stock in Figure 1.7. The sample ACF of this time series is shown in Figure 2.15. Note that this sample ACF plot behaves quite differently than the ACF plots in Figures 2.12 and 2.14. Instead of cutting off or tailing off near zero after a few lags, this sample ACF is very **persistent**; that is, it decays very slowly and exhibits sample autocorrelations that are still rather large even at long lags. This behavior is characteristic of a nonstationary time series. Generally, if the sample ACF does not dampen out within about 15 to 20 lags, the time series is nonstationary.

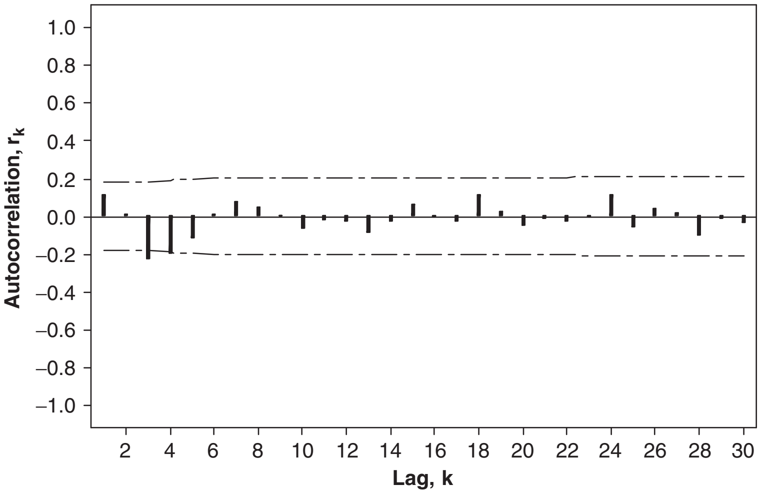


FIGURE 2.14 Autocorrelation function for pharmaceutical product sales, with 5% significance limits.

2.3.3 The Variogram

We have discussed two techniques for determining if a time series is nonstationary, plotting a reasonable long series of the data to see if it drifts or wanders away from its mean for long periods of time, and computing the sample ACF. However, often in practice there is no clear demarcation

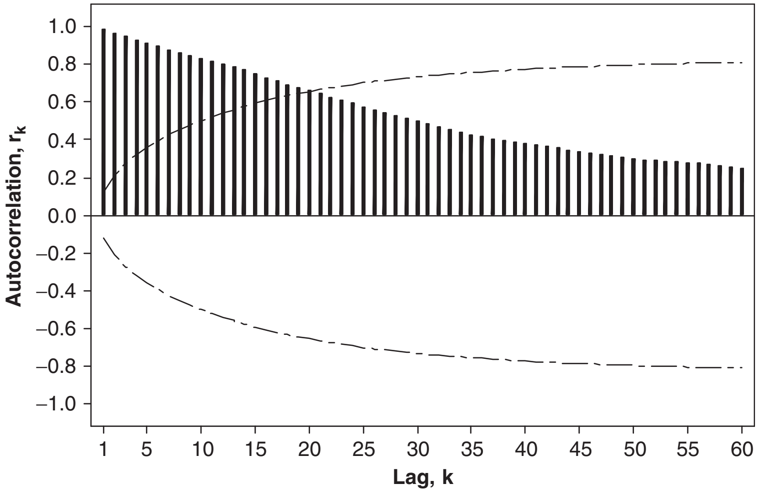


FIGURE 2.15 Autocorrelation function for Whole Foods Market stock price, with 5% significance limits.

between a stationary and a nonstationary process for many real-world time series. An additional diagnostic tool that is very useful is the **variogram**.

Suppose that the time series observations are represented by y_t. The variogram G_k measures variances of the differences between observations that are k lags apart, relative to the variance of the differences that are one time unit apart (or at lag 1). The variogram is defined mathematically as

$$G_k = \frac{\mathrm{Var}\,(y_{t+k} - y_t)}{\mathrm{Var}\,(y_{t+1} - y_t)} \quad k = 1, 2, \ldots \tag{2.16}$$

and the values of G_k are plotted as a function of the lag k. If the time series is stationary, it turns out that

$$G_k = \frac{1 - \rho_k}{1 - \rho_1},$$

but for a stationary time series $\rho_k \to 0$ as k increases, so when the variogram is plotted against lag k, G_k will reach an asymptote $1/(1 - \rho_1)$. However, if the time series is nonstationary, G_k will increase monotonically.

Estimating the variogram is accomplished by simply applying the usual sample variance to the differences, taking care to account for the changing sample sizes when the differences are taken (see Haslett (1997)). Let

$$\begin{aligned} d_t^k &= y_{t+k} - y_t \\ \bar{d}^k &= \frac{1}{T-k} \sum d_t^k. \end{aligned}$$

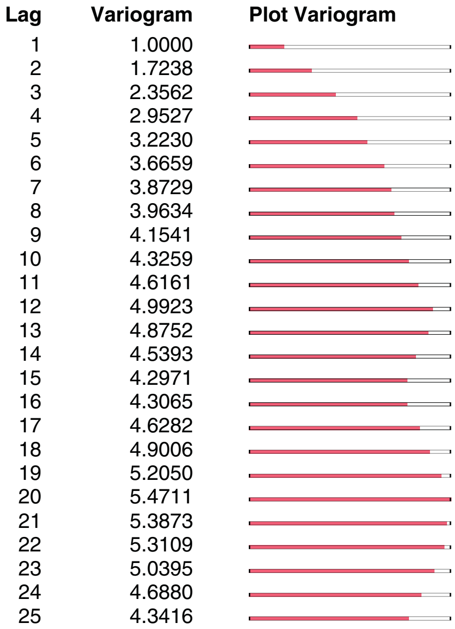
Then an estimate of $\mathrm{Var}\,(y_{t+k} - y_t)$ is

$$s_k^2 = \frac{\sum\limits_{t=1}^{T-k} \left(d_t^k - \bar{d}^k\right)^2}{T - k - 1}.$$

Therefore the sample variogram is given by

$$\hat{G}_k = \frac{s_k^2}{s_1^2} \quad k = 1, 2, \ldots \tag{2.17}$$

To illustrate the use of the variogram, consider the chemical process viscosity data plotted in Figure 2.9. Both the data plot and the sample ACF in



Lag	Variogram	Plot Variogram
1	1.0000	
2	1.7238	
3	2.3562	
4	2.9527	
5	3.2230	
6	3.6659	
7	3.8729	
8	3.9634	
9	4.1541	
10	4.3259	
11	4.6161	
12	4.9923	
13	4.8752	
14	4.5393	
15	4.2971	
16	4.3065	
17	4.6282	
18	4.9006	
19	5.2050	
20	5.4711	
21	5.3873	
22	5.3109	
23	5.0395	
24	4.6880	
25	4.3416	

FIGURE 2.16 JMP output for the sample variogram of the chemical process viscosity data from Figure 2.19.

Figures 2.12 and 2.13 suggest that the time series is stationary. Figure 2.16 is the variogram. Many software packages do not offer the variogram as a standard pull-down menu selection, but the JMP package does. Without software, it is still fairly easy to compute.

Start by computing the successive differences of the time series for a number of lags and then find their sample variances. The ratios of these sample variances to the sample variance of the first differences will produce the sample variogram. The JMP calculations of the sample variogram are shown in Figure 2.16 and a plot is given in Figure 2.17. Notice that the sample variogram generally converges to a stable level and then fluctuates around it. This is consistent with a stationary time series, and it provides additional evidence that the chemical process viscosity data are stationary.

Now let us see what the sample variogram looks like for a nonstationary time series. The Whole Foods Market stock price data from Appendix Table B.7 originally shown in Figure 1.7 are apparently nonstationary, as it wanders about with no obvious fixed level. The sample ACF in Figure 2.15 decays very slowly and as noted previously, gives the impression that the time series is nonstationary. The calculations for the variogram from JMP are shown in Figure 2.18 and the variogram is plotted in Figure 2.19.

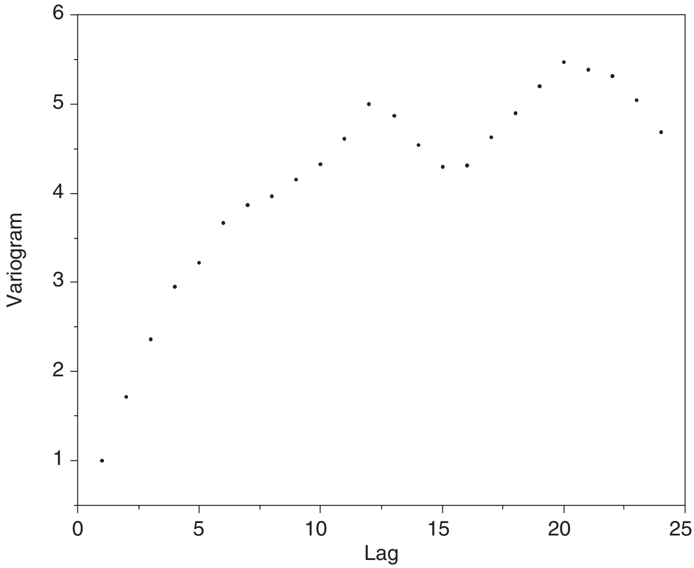
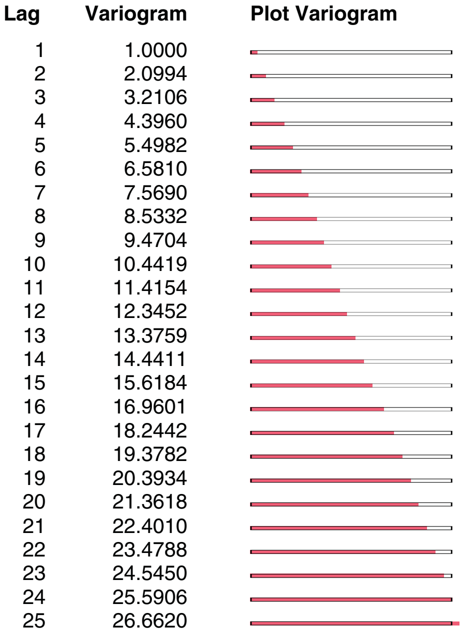


FIGURE 2.17 JMP sample variogram of the chemical process viscosity data from Figure 2.9.



Lag	Variogram	Plot Variogram
1	1.0000	
2	2.0994	
3	3.2106	
4	4.3960	
5	5.4982	
6	6.5810	
7	7.5690	
8	8.5332	
9	9.4704	
10	10.4419	
11	11.4154	
12	12.3452	
13	13.3759	
14	14.4411	
15	15.6184	
16	16.9601	
17	18.2442	
18	19.3782	
19	20.3934	
20	21.3618	
21	22.4010	
22	23.4788	
23	24.5450	
24	25.5906	
25	26.6620	

FIGURE 2.18 JMP output for the sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

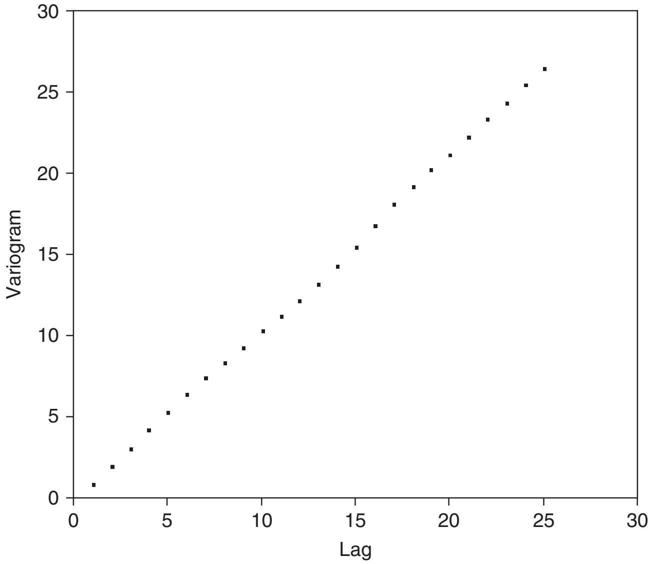


FIGURE 2.19 Sample variogram of the Whole Foods Market stock price data from Figure 1.7 and Appendix Table B.7.

Notice that the sample variogram in Figure 2.19 increases monotonically for all 25 lags. This is a strong indication that the time series is nonstationary.

2.4 USE OF DATA TRANSFORMATIONS AND ADJUSTMENTS

2.4.1 Transformations

Data transformations are useful in many aspects of statistical work, often for stabilizing the variance of the data. Nonconstant variance is quite common in time series data. For example, the International Sunspot Numbers plotted in Figure 2.20a show cyclic patterns of varying magnitudes. The variability from about 1800 to 1830 is smaller than that from about 1830 to 1880; other small periods of constant, but different, variances can also be identified.

A very popular type of data transformation to deal with nonconstant variance is the **power family** of transformations, given by

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases}, \tag{2.18}$$

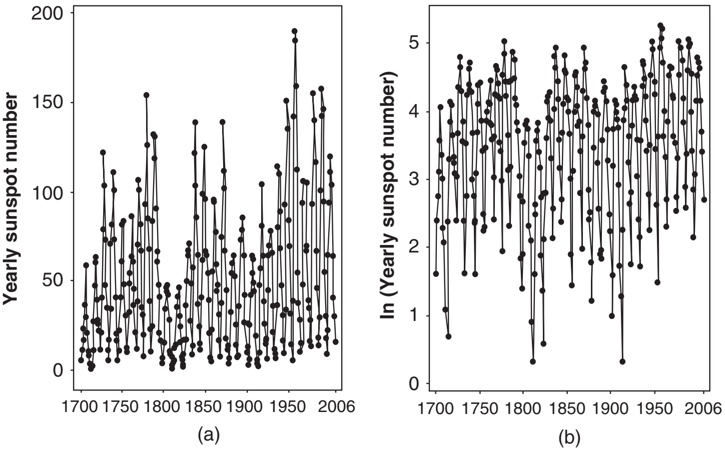


FIGURE 2.20 Yearly International Sunspot Number, (a) untransformed and (b) natural logarithm transformation. *Source*: SIDC.

where $\dot{y} = \exp[(1/T)\sum_{t=1}^{T} \ln y_t]$ is the geometric mean of the observations. If $\lambda = 1$, there is no transformation. Typical values of λ used with time series data are $\lambda = 0.5$ (a square root transformation), $\lambda = 0$ (the log transformation), $\lambda = -0.5$ (reciprocal square root transformation), and $\lambda = -1$ (inverse transformation). The divisor $\dot{y}^{\lambda-1}$ is simply a scale factor that ensures that when different models are fit to investigate the utility of different transformations (values of λ), the residual sum of squares for these models can be meaningfully compared. The reason that $\lambda = 0$ implies a log transformation is that $(y^{\lambda} - 1)/\lambda$ approaches the log of y as λ approaches zero. Often an appropriate value of λ is chosen empirically by fitting a model to $y^{(\lambda)}$ for various values of λ and then selecting the transformation that produces the minimum residual sum of squares.

The log transformation is used frequently in situations where the variability in the original time series increases with the average level of the series. When the standard deviation of the original series increases linearly with the mean, the log transformation is in fact an optimal variance-stabilizing transformation. The log transformation also has a very nice physical interpretation as percentage change. To illustrate this, let the time series be $y_1, y_2, \ldots, y_T$ and suppose that we are interested in the percentage change in y_t, say,

$$x_t = \frac{100(y_t - y_{t-1})}{y_{t-1}},$$

The approximate percentage change in y_t can be calculated from the differences of the log-transformed time series $x_t \cong 100[\ln(y_t) - \ln(y_{t-1})]$ because

$$\begin{aligned} 100[\ln(y_t) - \ln(y_{t-1})] &= 100 \ln\left(\frac{y_t}{y_{t-1}}\right) = 100 \ln\left(\frac{y_{t-1} + (y_t - y_{t-1})}{y_{t-1}}\right) \\ &= 100 \ln\left(1 + \frac{x_t}{100}\right) \cong x_t \end{aligned}$$

since $\ln(1 + z) \cong z$ when z is small.

The application of a natural logarithm transformation to the International Sunspot Number, as shown in Figure 2.20b, tends to stabilize the variance and leaves just a few unusual values.

2.4.2 Trend and Seasonal Adjustments

In addition to transformations, there are also several types of adjustments that are useful in time series modeling and forecasting. Two of the most widely used are **trend adjustments** and **seasonal adjustments**. Sometimes these procedures are called trend and seasonal decomposition.

A time series that exhibits a trend is a **nonstationary** time series. Modeling and forecasting of such a time series is greatly simplified if we can eliminate the trend. One way to do this is to fit a **regression model** describing the trend component to the data and then subtracting it out of the original observations, leaving a set of residuals that are free of trend. The trend models that are usually considered are the linear trend, in which the mean of y_t is expected to change linearly with time as in

$$E(y_t) = \beta_0 + \beta_1 t \tag{2.19}$$

or as a quadratic function of time

$$E(y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 \tag{2.20}$$

or even possibly as an exponential function of time such as

$$E(y_t) = \beta_0 e^{\beta_1 t}. \tag{2.21}$$

The models in Eqs. (2.19)–(2.21) are usually fit to the data by using ordinary least squares.

Example 2.6 We will show how least squares can be used to fit regression models in Chapter 3. However, it would be useful at this point to illustrate how trend adjustment works. Minitab can be used to perform trend adjustment. Consider the annual US production of blue and gorgonzola cheeses

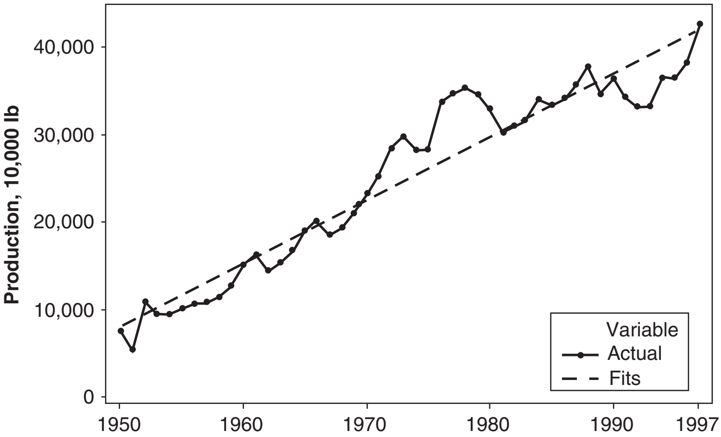


FIGURE 2.21 Blue and gorgonzola cheese production, with fitted regression line. *Source*: USDA–NASS.

shown in Figure 1.4. There is clearly a positive, nearly linear trend. The trend analysis plot in Figure 2.21 shows the original time series with the fitted line.

Plots of the residuals from this model indicate that, in addition to an underlying trend, there is additional structure. The normal probability plot (Figure 2.22a) and histogram (Figure 2.22c) indicate the residuals are

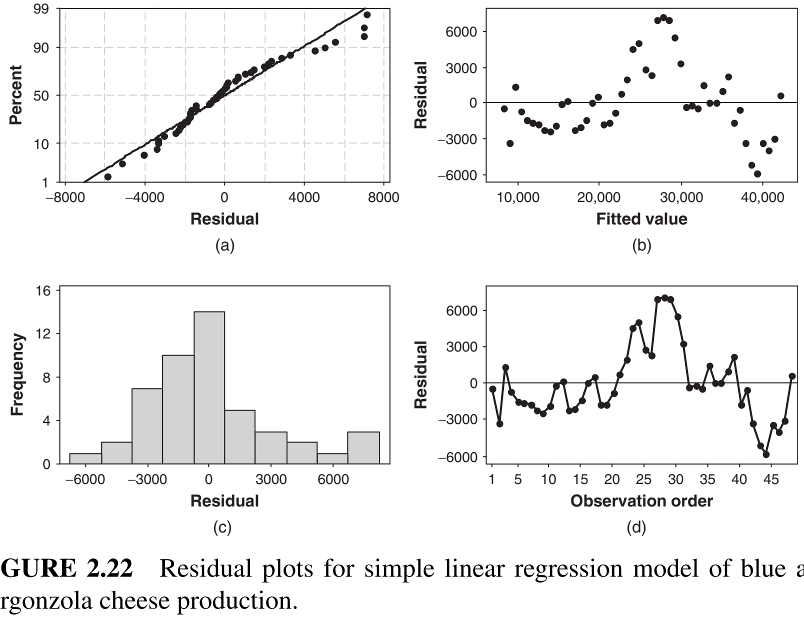


FIGURE 2.22 Residual plots for simple linear regression model of blue and gorgonzola cheese production.

approximately normally distributed. However, the plots of residuals versus fitted values (Figure 2.22b) and versus observation order (Figure 2.22d) indicate nonconstant variance in the last half of the time series. Analysis of model residuals is discussed more fully in Chapter 3.

Another approach to removing trend is by **differencing** the data; that is, applying the difference operator to the original time series to obtain a new time series, say,

$$x_t = y_t - y_{t-1} = \nabla y_t, \tag{2.22}$$

where ∇ is the (backward) difference operator. Another way to write the differencing operation is in terms of a **backshift operator** B, defined as $By_t = y_{t-1}$, so

$$x_t = (1 - B)y_t = \nabla y_t = y_t - y_{t-1} \tag{2.23}$$

with $\nabla = (1 - B)$. Differencing can be performed successively if necessary until the trend is removed; for example, the second difference is

$$x_t = \nabla^2 y_t = \nabla(\nabla y_t) = (1 - B)^2 y_t = (1 - 2B + B^2) = y_t - 2y_{t-1} + y_{t-2} \tag{2.24}$$

In general, powers of the backshift operator and the backward difference operator are defined as

$$\begin{aligned} B^d y_t &= y_{t-d} \\ \nabla^d &= (1 - B)^d \end{aligned} \tag{2.25}$$

Differencing has two advantages relative to fitting a trend model to the data. First, it does not require estimation of any parameters, so it is a more **parsimonious** (i.e., simpler) approach; and second, model fitting assumes that the trend is fixed throughout the time series history and will remain so in the (at least immediate) future. In other words, the trend component, once estimated, is assumed to be **deterministic**. Differencing can allow the trend component to change through time. The first difference accounts for a trend that impacts the change in the mean of the time series, the second difference accounts for changes in the slope of the time series, and so forth. Usually, one or two differences are all that is required in practice to remove an underlying trend in the data.

Example 2.7 Reconsider the blue and gorgonzola cheese production data. A difference of one applied to this time series removes the increasing trend (Figure 2.23) and also improves the appearance of the residuals plotted versus fitted value and observation order when a linear model is fitted to the detrended time series (Figure 2.24). This illustrates that differencing may be a very good alternative to detrending a time series by using a regression model.

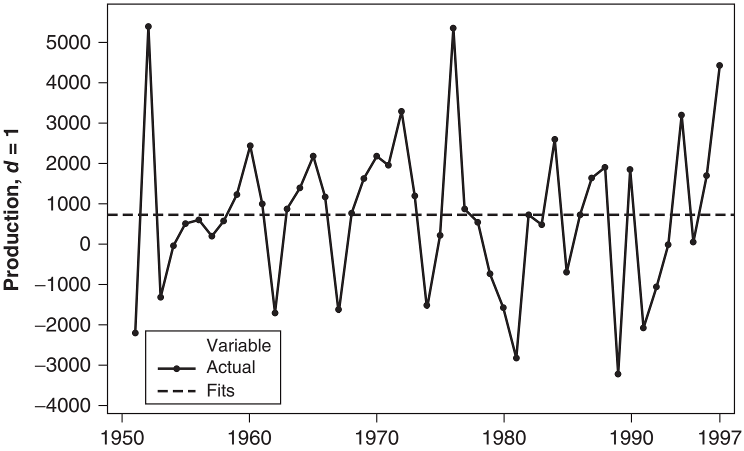


FIGURE 2.23 Blue and gorgonzola cheese production, with one difference. *Source*: USDA–NASS.

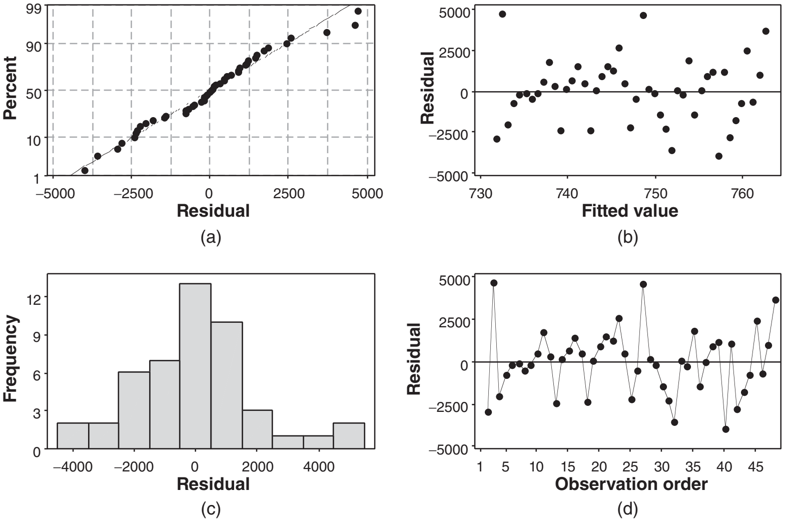


FIGURE 2.24 Residual plots for one difference of blue and gorgonzola cheese production.

Seasonal, or both trend *and* seasonal, components are present in many time series. Differencing can also be used to eliminate seasonality. Define a lag—*d* **seasonal difference** operator as

$$\nabla_d y_t = (1 - B^d) = y_t - y_{t-d}. \tag{2.26}$$

For example, if we had monthly data with an annual season (a very common situation), we would likely use $d = 12$, so the seasonally differenced data would be

$$\nabla_{12} y_t = (1 - B^{12}) y_t = y_t - y_{t-12}.$$

When both trend *and* seasonal components are simultaneously present, we can sequentially difference to eliminate these effects. That is, first seasonally difference to remove the seasonal component and then difference one or more times using the regular difference operator to remove the trend.

Example 2.8 The beverage shipment data shown in Figure 2.2 appear to have a strong monthly pattern—January consistently has the lowest shipments in a year while the peak shipments are in May and June. There is also an overall increasing trend from year to year that appears to be the same regardless of month.

A seasonal difference of twelve followed by a trend difference of one was applied to the beverage shipments, and the results are shown in Figure 2.25. The seasonal differencing removes the monthly pattern (Figure 2.25a), and the second difference of one removes the overall increasing trend (Figure 2.25b). The fitted linear trend line in Figure 2.25b has a slope of virtually zero. Examination of the residual plots in Figure 2.26 does not reveal any problems with the linear trend model fit to the differenced data.

Regression models can also be used to eliminate seasonal (or trend and seasonal components) from time series data. A simple but useful model is

$$E(y_t) = \beta_0 + \beta_1 \sin \frac{2\pi}{d} t + \beta_2 \cos \frac{2\pi}{d} t, \tag{2.27}$$

where d is the period (or length) of the season and $2\pi/d$ is expressed in radians. For example, if we had monthly data and an annual season, then $d = 12$. This model describes a simple, symmetric seasonal pattern that

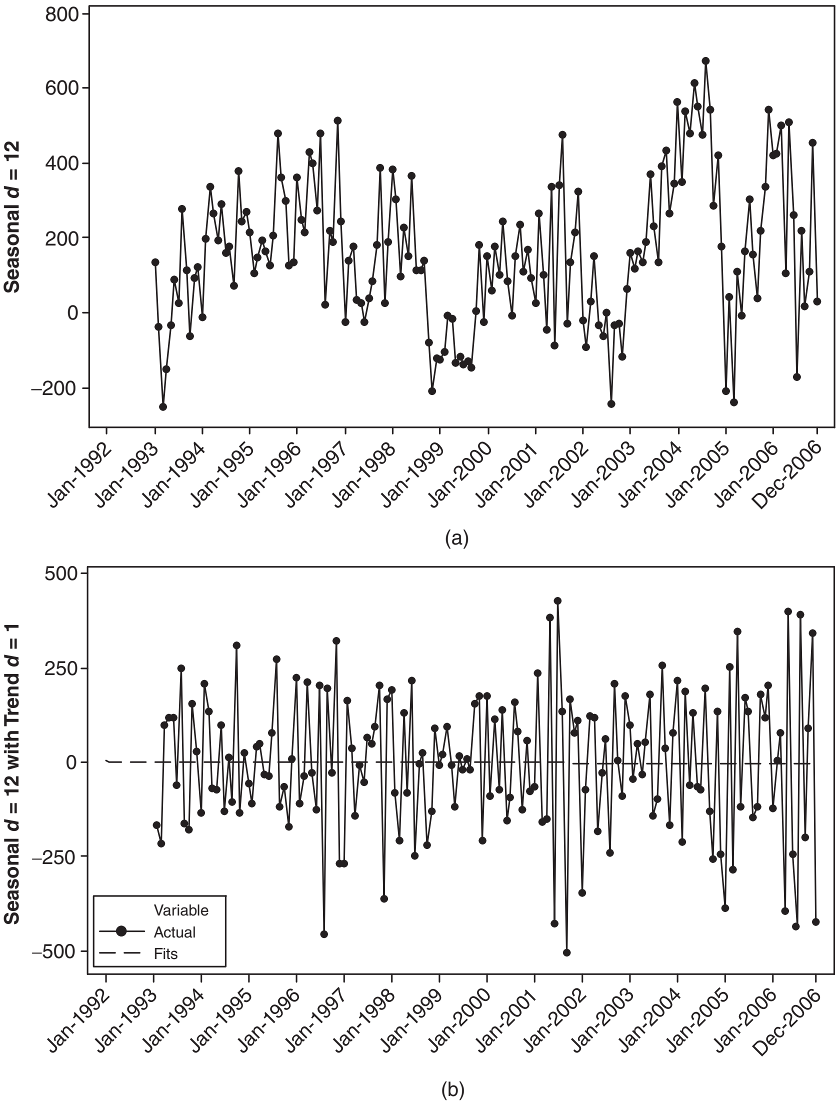


FIGURE 2.25 Time series plots of seasonal- and trend-differenced beverage data.

repeats every 12 periods. The model is actually a sine wave. To see this, recall that a sine wave with amplitude β, phase angle or origin θ, and period or cycle length ω can be written as

$$E(y_t) = \beta \sin \omega(t + \theta). \tag{2.28}$$

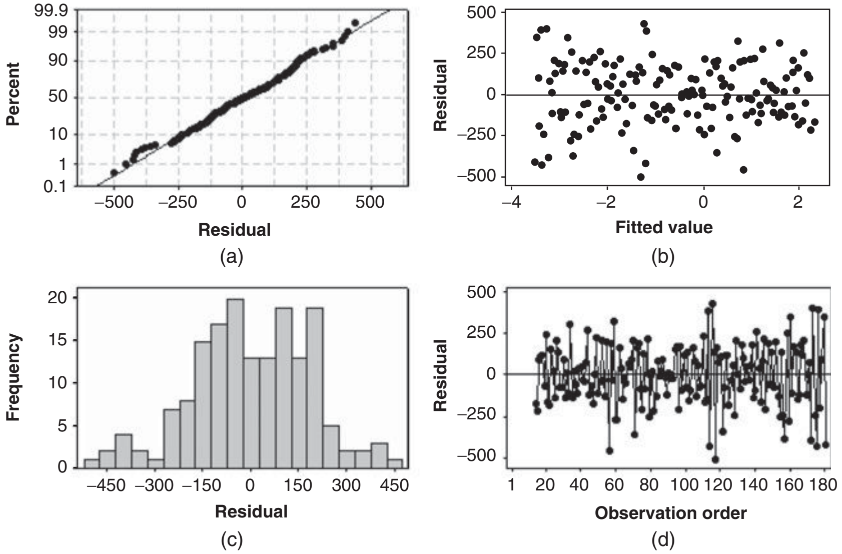


FIGURE 2.26 Residual plots for linear trend model of differenced beverage shipments.

Equation (2.27) was obtained by writing Eq. (2.28) as a sine–cosine pair using the trigonometric identity $\sin(u + v) = \cos u \sin v + \sin u \cos v$ and adding an intercept term β_0:

$$\begin{aligned} E(y_t) &= \beta \sin \omega(t + \theta) \\ &= \beta \cos \omega\theta \sin \omega t + \beta \sin \omega\theta \cos \omega t \\ &= \beta_1 \sin \omega t + \beta_2 \cos \omega t \end{aligned}$$

where $\beta_1 = \beta \cos \omega\theta$ and $\beta_2 = \beta \sin \omega\theta$. Setting $\omega = 2\pi/12$ and adding the intercept term β_0 produces Eq. (2.27). This model is very flexible; for example, if we set $\omega = 2\pi/52$ we can model a yearly seasonal pattern that is observed weekly, if we set $\omega = 2\pi/4$ we can model a yearly seasonal pattern observed quarterly, and if we set $\omega = 2\pi/13$ we can model an annual seasonal pattern observed in 13 four-week periods instead of the usual months.

Equation (2.27) incorporates a single sine wave at the **fundamental frequency** $\omega = 2\pi/12$. In general, we could add **harmonics** of the fundamental frequency to the model in order to model more complex seasonal patterns. For example, a very general model for monthly data and an annual season that uses the fundamental frequency and the first three harmonics is

$$E(y_t) = \beta_0 + \sum_{j=1}^{4} \left(\beta_j \sin \frac{2\pi j}{12} t + \beta_{4+j} \cos \frac{2\pi j}{12} t \right). \tag{2.29}$$

If the data are observed in 13 four-week periods, the model would be

$$E(y_t) = \beta_0 + \sum_{j=1}^{4} \left(\beta_j \sin \frac{2\pi j}{13} t + \beta_{4+j} \cos \frac{2\pi j}{13} t \right). \tag{2.30}$$

There is also a “classical” approach to decomposition of a time series into trend and seasonal components (actually, there are a lot of different decomposition algorithms; here we explain a very simple but useful approach). The general mathematical model for this decomposition is

$$y_t = f(S_t, T_t, \varepsilon_t),$$

where S_t is the seasonal component, T_t is the trend effect (sometimes called the trend-cycle effect), and ε_t is the random error component. There are usually two forms for the function f; an additive model

$$y_t = S_t + T_t + \varepsilon_t$$

and a multiplicative model

$$y_t = S_t T_t \varepsilon_t.$$

The additive model is appropriate if the magnitude (amplitude) of the seasonal variation does not vary with the level of the series, while the multiplicative version is more appropriate if the amplitude of the seasonal fluctuations increases or decreases with the average level of the time series.

Decomposition is useful for breaking a time series down into these component parts. For the additive model, it is relatively easy. First, we would model and remove the trend. A simple linear model could be used to do this, say, $T_t = \beta_0 + \beta_1 t$. Other methods could also be used. Moving averages can be used to isolate a trend and remove it from the original data, as could more sophisticated regression methods. These techniques might be appropriate when the trend is not a straight line over the history of the time series. Differencing could also be used, although it is not typically in the classical decomposition approach.

Once the trend or trend-cycle component is estimated, the series is detrended:

$$y_t - T_t = S_t + \varepsilon_t.$$

Now a seasonal factor can be calculated for each period in the season. For example, if the data are monthly and an annual season is anticipated, we would calculate a season effect for each month in the data set. Then the seasonal indices are computed by taking the average of all of the seasonal factors for each period in the season. In this example, all of the January seasonal factors are averaged to produce a January season index; all of the February seasonal factors are averaged to produce a February season index; and so on. Sometimes medians are used instead of averages. In multiplicative decomposition, ratios are used, so that the data are detrended by

$$\frac{y_t}{T_t} = S_t \varepsilon_t.$$

The seasonal indices are estimated by taking the averages over all of the detrended values for each period in the season.

Example 2.9 The decomposition approach can be applied to the beverage shipment data. Examining the time series plot in Figure 2.2, there is both a strong positive trend as well as month-to-month variation, so the model should include both a trend and a seasonal component. It also appears that the magnitude of the seasonal variation does not vary with the level of the series, so an additive model is appropriate.

Results of a time series decomposition analysis from Minitab of the beverage shipments are in Figure 2.27, showing the original data (labeled “Actual”); along with the fitted trend line (“Trend”) and the predicted values (“Fits”) from the additive model with both the trend and seasonal components.

Details of the seasonal analysis are shown in Figure 2.28. Estimates of the monthly variation from the trend line for each season (seasonal indices) are in Figure 2.28a with boxplots of the actual differences in Figure 2.28b. The percent of variation by seasonal period is in Figure 2.28c, and model residuals by seasonal period are in Figure 2.28d.

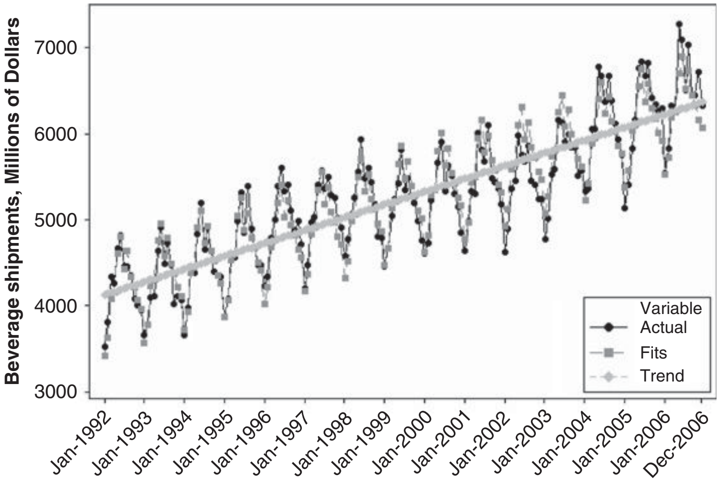


FIGURE 2.27 Time series plot of decomposition model for beverage shipments.

Additional details of the component analysis are shown in Figure 2.29. Figure 2.29a is the original time series, Figure 2.29b is a plot of the time series with the trend removed, Figure 2.29c is a plot of the time series with the seasonality removed, and Figure 2.29d is essentially a residual plot of the detrended and seasonally adjusted data. The wave-like pattern in Figure 2.29d suggests a potential issue with the assumption of constant variance over time.

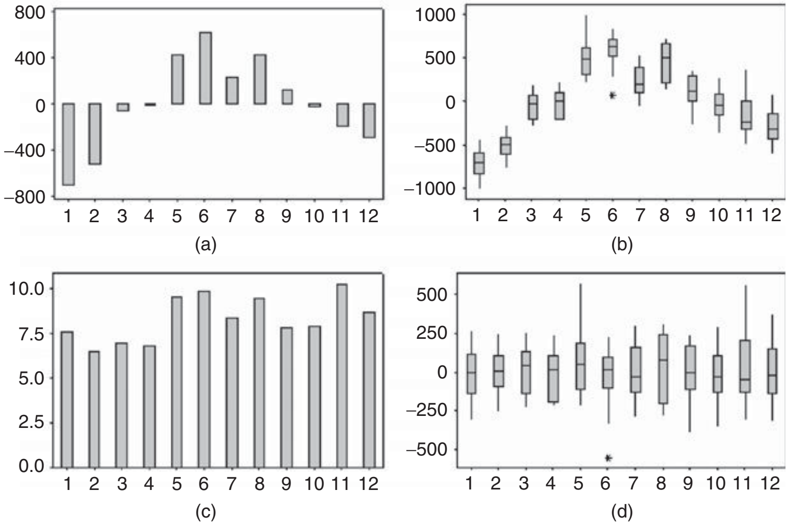
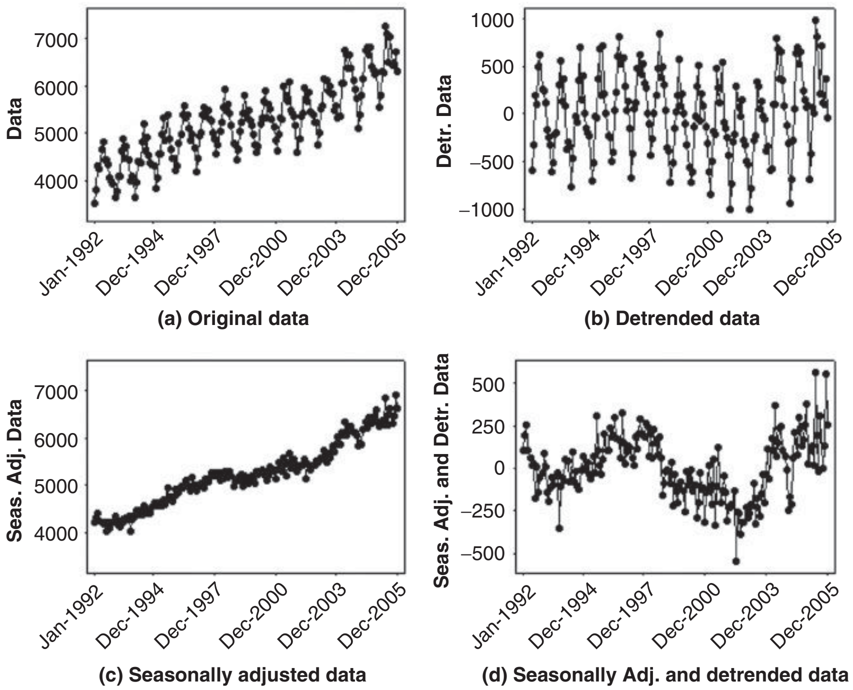


FIGURE 2.28 Seasonal analysis for beverage shipments.



(a) Original data

(b) Detrended data

(c) Seasonally adjusted data

(d) Seasonally Adj. and detrended data

FIGURE 2.29 Component analysis of beverage shipments.

Looking at the normal probability plot and histogram of residuals (Figure 2.30a,c), there does not appear to be an issue with the normality assumption. Figure 2.30d is the same plot as Figure 2.29d. However, variance does seem to increase as the predicted value increases; there is a funnel shape to the residuals plotted in Figure 2.30b. A natural logarithm transformation of the data may stabilize the variance and allow a useful decomposition model to be fit.

Results from the decomposition analysis of the natural log-transformed beverage shipment data are plotted in Figure 2.31, with the transformed data, fitted trend line, and predicted values. Figure 2.32a shows the transformed data, Figure 2.32b the transformed data with the trend removed, Figure 2.32c the transformed data with seasonality removed, and Figure 2.32d the residuals plot of the detrended and seasonally adjusted transformed data. The residual plots in Figure 2.33 indicate that the variance over the range of the predicted values is now stable (Figure 2.33b), and there are no issues with the normality assumption (Figures 2.33a,c). However, there is still a wave-like pattern in the plot of residuals versus time,

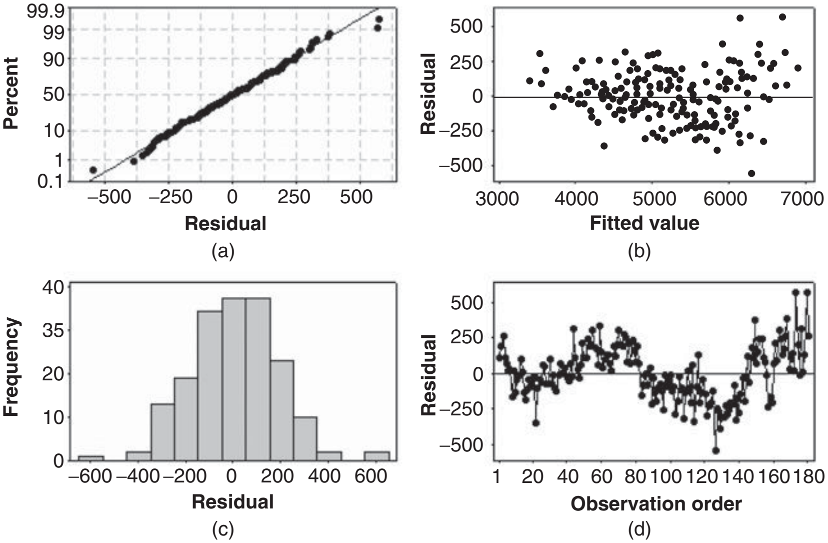


FIGURE 2.30 Residual plots for additive model of beverage shipments.

both Figures 2.32d and 2.33d, indicating that some other structure in the transformed data over time is not captured by the decomposition model. This was not an issue with the model based on seasonal and trend differencing (Figures 2.25 and 2.26), which may be a more appropriate model for monthly beverage shipments.

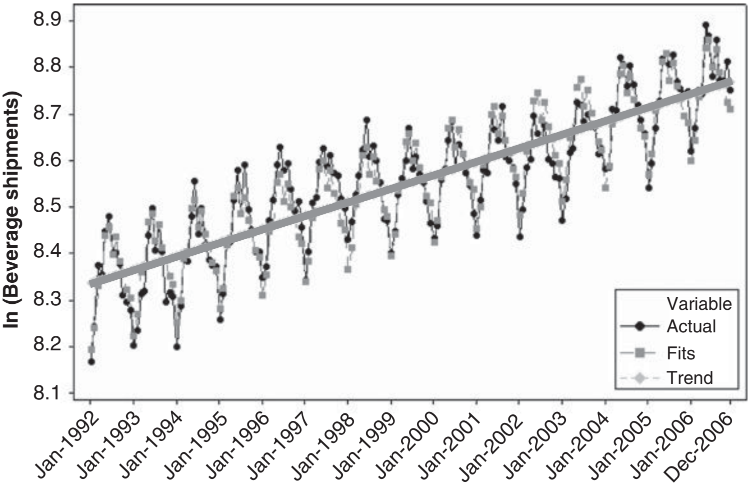


FIGURE 2.31 Time series plot of decomposition model for transformed beverage data.

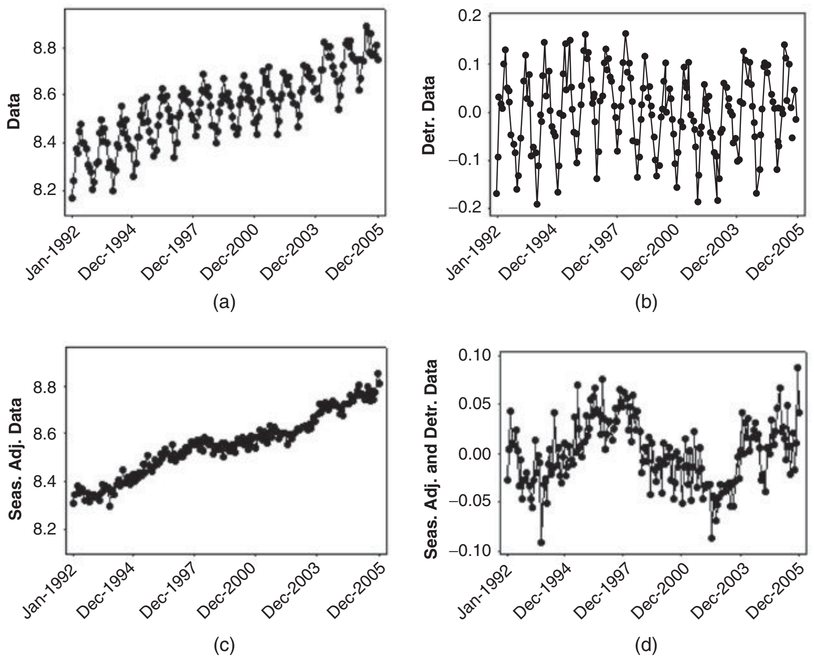


FIGURE 2.32 Component analysis of transformed beverage data.

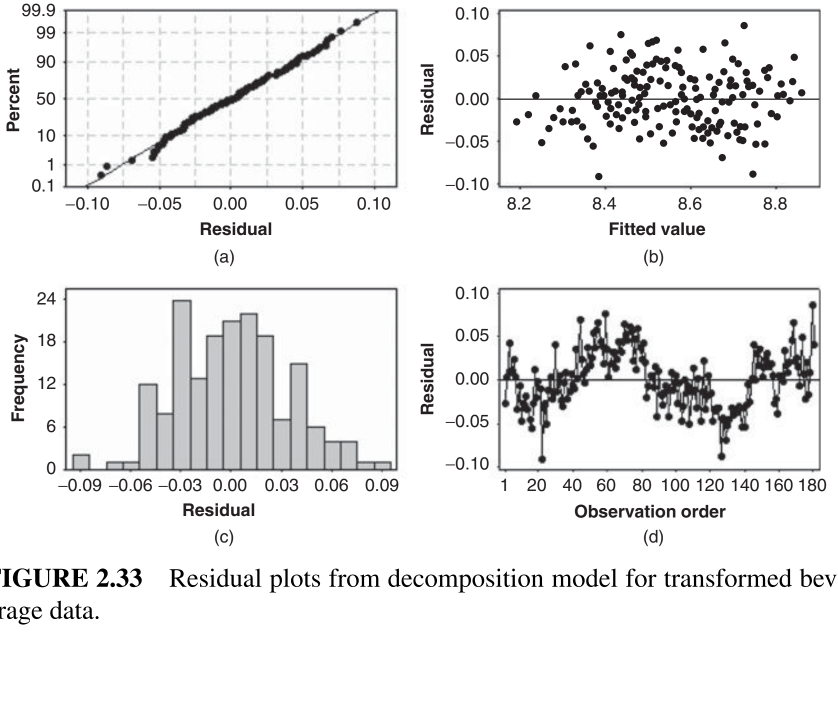


FIGURE 2.33 Residual plots from decomposition model for transformed beverage data.

Another technique for seasonal adjustment that is widely used in modeling and analyzing economic data is the ***X*-11 method**. Much of the development work on this method was done by Julian Shiskin and others at the US Bureau of the Census beginning in the mid-1950s and culminating into the *X*-11 Variant of the Census Method II Seasonal Adjustment Program. References for this work during this period include Shiskin (1958), and Marris (1961). Authoritative documentation for the *X*-11 procedure is in Shiskin, Young, and Musgrave (1967). The *X*-11 method uses symmetric moving averages in an iterative approach to estimating the trend and seasonal components. At the end of the series, however, these symmetric weights cannot be applied. Asymmetric weights have to be used.

JMP (V12 and higher) provides the *X*-11 technique. Figure 2.34 shows the JMP *X*-11 output for the beverage shipment data from Figure 2.2. The upper part of the output contains a plot of the original time series, followed by the sample ACF and PACF. Then Display D10 in the figure shows the final estimates of the seasonal factors by month followed in Display D13 by the irregular or deseasonalized series. The final display is a plot of the original and adjusted time series.

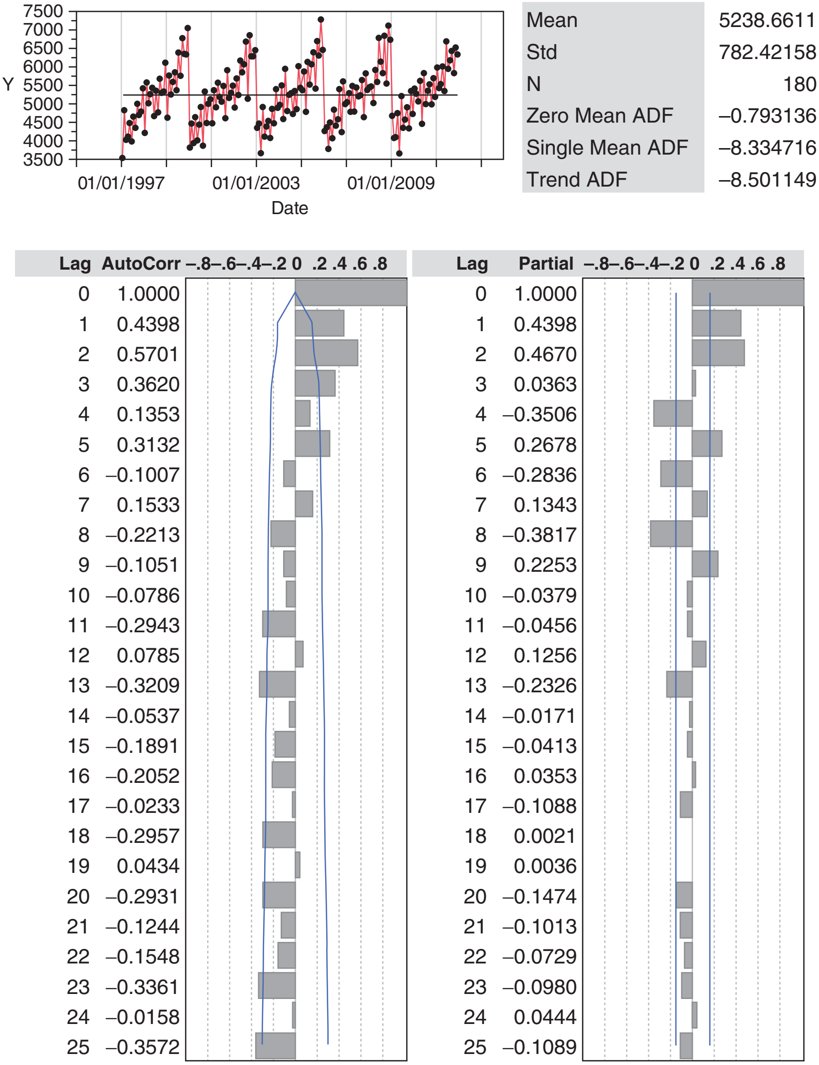
While different variants of the *X*-11 technique have been proposed, the most important method to date has been the *X-11-ARIMA* method developed at Statistics Canada. This method uses Box–Jenkins autoregressive integrated moving average models (which are discussed in Chapter 5) to extend the series. The use of ARIMA models will result in differences in the final component estimates. Details of this method are in Dagum (1980, 1983, 1988).

2.5 GENERAL APPROACH TO TIME SERIES MODELING AND FORECASTING

The techniques that we have been describing form the basis of a general approach to modeling and forecasting time series data. We now give a broad overview of the approach. This should give readers a general understanding of the connections between the ideas we have presented in this chapter and guidance in understanding how the topics in subsequent chapters form a collection of useful techniques for modeling and forecasting time series.

The basic steps in modeling and forecasting a time series are as follows:

1. Plot the time series and determine its basic features, such as whether trends or seasonal behavior or both are present. Look for possible outliers or any indication that the time series has changed with respect

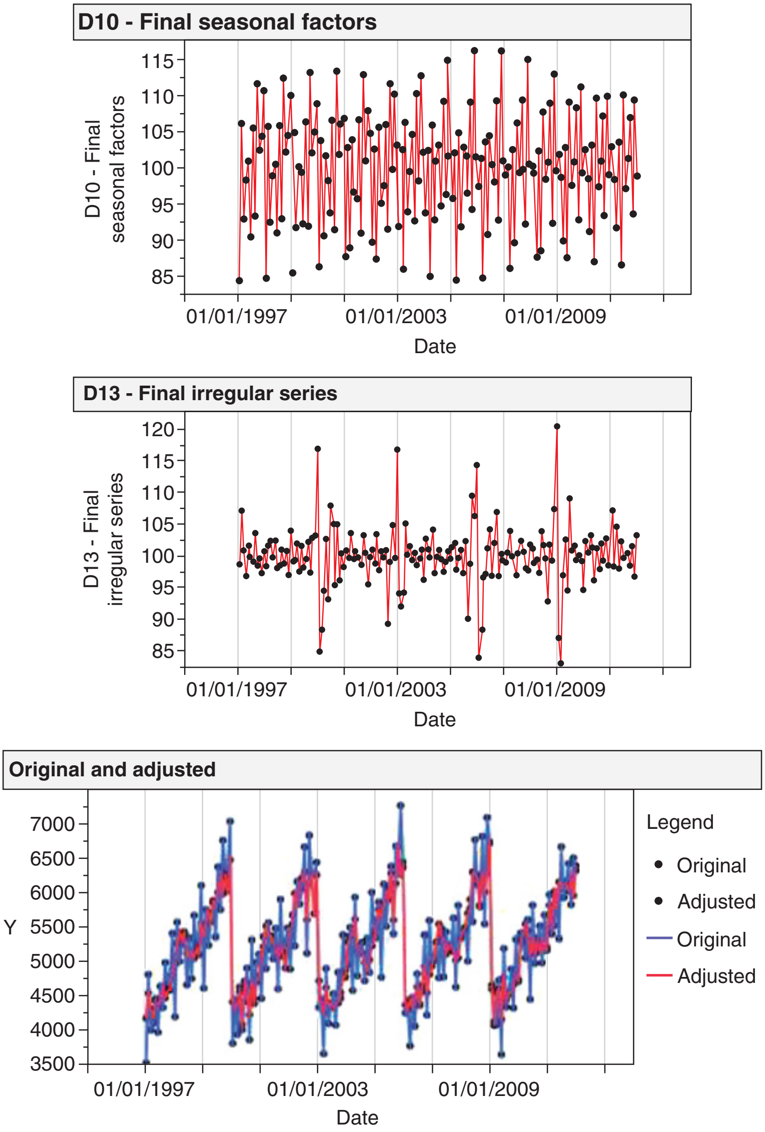


```
Mean               5238.6611
Std                782.42158
N                        180
Zero Mean ADF      –0.793136
Single Mean ADF    –8.334716
Trend ADF          –8.501149

Lag  AutoCorr          Lag  Partial
 0    1.0000            0    1.0000
 1    0.4398            1    0.4398
 2    0.5701            2    0.4670
 3    0.3620            3    0.0363
 4    0.1353            4   –0.3506
 5    0.3132            5    0.2678
 6   –0.1007            6   –0.2836
 7    0.1533            7    0.1343
 8   –0.2213            8   –0.3817
 9   –0.1051            9    0.2253
10   –0.0786           10   –0.0379
11   –0.2943           11   –0.0456
12    0.0785           12    0.1256
13   –0.3209           13   –0.2326
14   –0.0537           14   –0.0171
15   –0.1891           15   –0.0413
16   –0.2052           16    0.0353
17   –0.0233           17   –0.1088
18   –0.2957           18    0.0021
19    0.0434           19    0.0036
20   –0.2931           20   –0.1474
21   –0.1244           21   –0.1013
22   –0.1548           22   –0.0729
23   –0.3361           23   –0.0980
24   –0.0158           24    0.0444
25   –0.3572           25   –0.1089
```

FIGURE 2.34 JMP output for the *X*-11 procedure.

to its basic features (such as trends or seasonality) over the time period history.

2. Eliminate any trend or seasonal components, either by differencing or by fitting an appropriate model to the data. Also consider using data transformations, particularly if the variability in the time series seems to be proportional to the average level of the series. The objective of these operations is to produce a set of stationary residuals.



D10 - Final seasonal factors

D13 - Final irregular series

Original and adjusted

FIGURE 2.34 (*Continued*)

3. Develop a forecasting model for the residuals. It is not unusual to find that there are several plausible models, and additional analysis will have to be performed to determine the best one to deploy. Sometimes potential models can be eliminated on the basis of their fit to the historical data. It is unlikely that a model that fits poorly will produce good forecasts.

4. Validate the performance of the model (or models) from the previous step. This will probably involve some type of split-sample or cross-validation procedure. The objective of this step is to select a model to use in forecasting. We will discuss this more in the next section and illustrate these techniques throughout the book.
5. Also of interest are the differences between the original time series y_t and the values that would be forecast by the model on the original scale. To forecast values on the scale of the original time series y_t, reverse the transformations and any differencing adjustments made to remove trends or seasonal effects.
6. For forecasts of future values in period $T + \tau$ on the original scale, if a transformation was used, say, $x_t = \ln y_t$, then the forecast made at the end of period T for $T + \tau$ would be obtained by reversing the transformation. For the natural log this would be

$$\hat{y}_{T+\tau}(T) = \exp[\hat{x}_{T+\tau}(T)].$$

7. If prediction intervals are desired for the forecast (and we recommend doing this), construct prediction intervals for the residuals and then reverse the transformations made to produce the residuals as described earlier. We will discuss methods for finding prediction intervals for most of the forecasting methods presented in this book.
8. Develop and implement a procedure for monitoring the forecast to ensure that deterioration in performance will be detected reasonably quickly. Forecast monitoring is usually done by evaluating the stream of forecast errors that are experienced. We will present methods for monitoring forecast errors with the objective of detecting changes in performance of the forecasting model.

2.6 EVALUATING AND MONITORING FORECASTING MODEL PERFORMANCE

2.6.1 Forecasting Model Evaluation

We now consider how to evaluate the performance of a forecasting technique for a particular time series or application. It is important to carefully define the meaning of performance. It is tempting to evaluate performance on the basis of the fit of the forecasting or time series model to historical data. There are many statistical measures that describe how well a model fits a given sample of data, and several of these will be described in subsequent chapters. This goodness-of-fit approach often uses the residuals and does not really reflect the capability of the forecasting technique to successfully predict future observations. The user of the forecasts is very concerned about the accuracy of future forecasts, not model goodness of fit, so it is important to evaluate this aspect of any recommended technique. Sometimes forecast accuracy is called "out-of-sample" forecast error, to distinguish it from the residuals that arise from a model-fitting process.

Measure of forecast accuracy should always be evaluated as part of a model validation effort (see step 4 in the general approach to forecasting in the previous section). When more than one forecasting technique seems reasonable for a particular application, these forecast accuracy measures can also be used to discriminate between competing models. We will discuss this more in Section 2.6.2.

It is customary to evaluate forecasting model performance using the one-step-ahead forecast errors

$$e_t(1) = y_t - \hat{y}_t(t-1), \tag{2.31}$$

where $\hat{y}_t(t-1)$ is the forecast of y_t that was made one period prior. Forecast errors at other lags, or at several different lags, could be used if interest focused on those particular forecasts. Suppose that there are n observations for which forecasts have been made and n one-step-ahead forecast errors, $e_t(1), t = 1, 2, \dots, n$. Standard measures of forecast accuracy are the **average error** or **mean error**

$$\text{ME} = \frac{1}{n}\sum_{t=1}^{n} e_t(1), \tag{2.32}$$

the **mean absolute deviation** (or mean absolute error)

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{n} |e_t(1)|, \tag{2.33}$$

and the **mean squared error**

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} [e_t(1)]^2. \tag{2.34}$$

The mean forecast error in Eq. (2.32) is an estimate of the expected value of forecast error, which we would hope to be zero; that is, the forecasting technique produces **unbiased** forecasts. If the mean forecast error differs appreciably from zero, bias in the forecast is indicated. If the mean forecast error drifts away from zero when the forecasting technique is in use, this can be an indication that the underlying time series has changed in some fashion, the forecasting technique has not tracked this change, and now biased forecasts are being generated.

Both the mean absolute deviation (MAD) in Eq. (2.33) and the mean squared error (MSE) in Eq. (2.34) measure the **variability** in forecast errors. Obviously, we want the variability in forecast errors to be small. The MSE is a direct estimator of the variance of the one-step-ahead forecast errors:

$$\hat{\sigma}^2_{e(1)} = \text{MSE} = \frac{1}{n} \sum_{t=1}^{n} [e_t(1)]^2. \tag{2.35}$$

If the forecast errors are normally distributed (this is usually not a bad assumption, and one that is easily checked), the MAD is related to the standard deviation of forecast errors by

$$\hat{\sigma}_{e(1)} = \sqrt{\frac{\pi}{2}}\text{MAD} \cong 1.25\,\text{MAD} \tag{2.36}$$

The one-step-ahead forecast error and its summary measures, the ME, MAD, and MSE, are all scale-dependent measures of forecast accuracy; that is, their values are expressed in terms of the original units of measurement (or in the case of MSE, the square of the original units). So, for example, if we were forecasting demand for electricity in Phoenix during the summer, the units would be megawatts (MW). If the MAD for the forecast error during summer months was 5 MW, we might not know whether this was a large forecast error or a relatively small one. Furthermore, accuracy measures that are scale dependent do not facilitate comparisons of a single forecasting technique across different time series, or comparisons across different time periods. To accomplish this, we need a measure of relative forecast error.

Define the **relative forecast error** (in percent) as

$$re_t(1) = \left(\frac{y_t - \hat{y}_t(t-1)}{y_t} \right) 100 = \left(\frac{e_t(1)}{y_t} \right) 100. \tag{2.37}$$

This is customarily called the **percent forecast error**. The mean percent forecast error (MPE) is

$$\text{MPE} = \frac{1}{n}\sum_{t=1}^{n} re_t(1) \tag{2.38}$$

and the mean absolute percent forecast error (MAPE) is

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} |re_t(1)|. \tag{2.39}$$

Knowing that the relative or percent forecast error or the MAPE is 3% (say) can be much more meaningful than knowing that the MAD is 5 MW. Note that the relative or percent forecast error only makes sense if the time series y_t does not contain zero values.

Example 2.10 Table 2.2 illustrates the calculation of the one-step-ahead forecast error, the absolute errors, the squared errors, the relative (percent) error, and the absolute percent error from a forecasting model for 20 time periods. The last row of columns (3) through (7) display the sums required to calculate the ME, MAD, MSE, MPE, and MAPE.

From Eq. (2.32), the mean (or average) forecast error is

$$\text{ME} = \frac{1}{n}\sum_{t=1}^{n} e_t(1) = \frac{1}{20}(-11.6) = -0.58,$$

the MAD is computed from Eq. (2.33) as

$$\text{MAD} = \frac{1}{n}\sum_{t=1}^{n} |e_t(1)| = \frac{1}{20}(86.6) = 4.33,$$

and the MSE is computed from Eq. (2.34) as

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} [e_t(1)]^2 = \frac{1}{20}(471.8) = 23.59.$$

TABLE 2.2 Calculation of Forecast Accuracy Measures

Time Period	(1) Observed Value y_t	(2) Forecast $\hat{y}_t(t-1)$	(3) Forecast Error $e_t(1)$	(4) Absolute Error $\lvert e_t(1)\rvert$	(5) Squared Error $[e_t(1)]^2$	(6) Relative (%) Error $\left(e_t(1)/y_t\right)100$	(6) Absolute (%) Error $\lvert\left(e_t(1)/y_t\right)100\rvert$
1	47	51.1	−4.1	4.1	16.81	−8.7234	8.723404
2	46	52.9	−6.9	6.9	47.61	−15	15
3	51	48.8	2.2	2.2	4.84	4.313725	4.313725
4	44	48.1	−4.1	4.1	16.81	−9.31818	9.318182
5	54	49.7	4.3	4.3	18.49	7.962963	7.962963
6	47	47.5	−0.5	0.5	0.25	−1.06383	1.06383
7	52	51.2	0.8	0.8	0.64	1.538462	1.538462
8	45	53.1	−8.1	8.1	65.61	−18	18
9	50	54.4	−4.4	4.4	19.36	−8.8	8.8
10	51	51.2	−0.2	0.2	0.04	−0.39216	0.392157
11	49	53.3	−4.3	4.3	18.49	−8.77551	8.77551
12	41	46.5	−5.5	5.5	30.25	−13.4146	13.41463
13	48	53.1	−5.1	5.1	26.01	−10.625	10.625
14	50	52.1	−2.1	2.1	4.41	−4.2	4.2
15	51	46.8	4.2	4.2	17.64	8.235294	8.235294
16	55	47.7	7.3	7.3	53.29	13.27273	13.27273
17	52	45.4	6.6	6.6	43.56	12.69231	12.69231
18	53	47.1	5.9	5.9	34.81	11.13208	11.13208
19	48	51.8	−3.8	3.8	14.44	−7.91667	7.916667
20	52	45.8	6.2	6.2	38.44	11.92308	11.92308
	Totals		*−11.6*	*86.6*	*471.8*	*−35.1588*	*177.3*

Because the MSE estimates the variance of the one-step-ahead forecast errors, we have

$$\hat{\sigma}^2_{e(1)} = \text{MSE} = 23.59$$

and an estimate of the standard deviation of forecast errors is the square root of this quantity, or $\hat{\sigma}_{e(1)} = \sqrt{\text{MSE}} = 4.86$. We can also obtain an estimate of the standard deviation of forecasts errors from the MAD using Eq. (2.36)

$$\hat{\sigma}_{e(1)} \cong 1.25\,\text{MAD} = 1.25(4.33) = 5.41.$$

These two estimates are reasonably similar. The mean percent forecast error, MPE, is computed from Eq. (2.38) as

$$\text{MPE} = \frac{1}{n}\sum_{t=1}^{n} re_t(1) = \frac{1}{20}(-35.1588) = -1.76\%$$

and the mean absolute percent error is computed from Eq. (2.39) as

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} |re_t(1)| = \frac{1}{20}(177.3) = 8.87\%.$$

There is much empirical evidence (and even some theoretical justification) that the distribution of forecast errors can be well approximated by a **normal** distribution. This can easily be checked by constructing a **normal probability plot** of the forecast errors in Table 2.2, as shown in Figure 2.35. The forecast errors deviate somewhat from the straight line, indicating that the normal distribution is not a perfect model for the distribution of forecast errors, but it is not unreasonable. Minitab calculates the Anderson–Darling statistic, a widely used test statistic for normality. The *P*-value is 0.088, so the hypothesis of normality of the forecast errors would not be rejected at the 0.05 level. This test assumes that the observations (in this case the forecast errors) are uncorrelated. Minitab also reports the standard deviation of the forecast errors to be 4.947, a slightly larger value than we computed from the MSE, because Minitab uses the standard method for calculating sample standard deviations.

Note that Eq. (2.31) could have been written as

$$\text{Error} = \text{Observation} - \text{Forecast}.$$

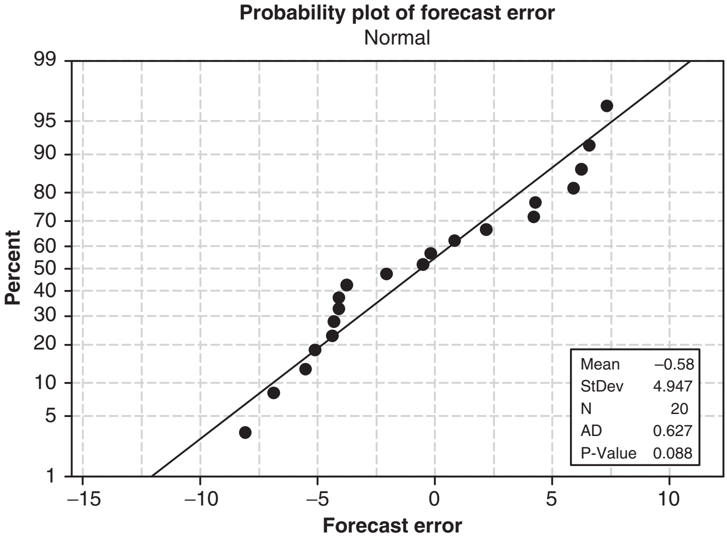


FIGURE 2.35 Normal probability plot of forecast errors from Table 2.2.

Hopefully, the forecasts do a good job of describing the structure in the observations. In an ideal situation, the forecasts would adequately model all of the structure in the data, and the sequence of forecast errors would be structureless. If they are, the sample ACF of the forecast error should look like the ACF of random data; that is, there should not be any large "spikes" on the sample ACF at low lag. Any systematic or nonrandom pattern in the forecast errors will tend to show up as significant spikes on the sample ACF. If the sample ACF suggests that the forecast errors are not random, then this is evidence that the forecasts can be improved by **refining** the forecasting model. Essentially, this would consist of taking the structure out of the forecast errors and putting it into the forecasts, resulting in forecasts that are better prediction of the data.

Example 2.11 Table 2.3 presents a set of 50 one-step-ahead errors from a forecasting model, and Table 2.4 shows the sample ACF of these forecast errors. The sample ACF is plotted in Figure 2.36. This sample ACF was obtained from Minitab. Note that sample autocorrelations for the first 13 lags are computed. This is consistent with our guideline indicating that for T observations only the first $T/4$ autocorrelations should be computed. The sample ACF does not provide any strong evidence to support a claim that there is a pattern in the forecast errors.

TABLE 2.3 One-Step-Ahead Forecast Errors

Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$
1	−0.62	11	−0.49	21	2.90	31	−1.88	41	−3.98
2	−2.99	12	4.13	22	0.86	32	−4.46	42	−4.28
3	0.65	13	−3.39	23	5.80	33	−1.93	43	1.06
4	0.81	14	2.81	24	4.66	34	−2.86	44	0.18
5	−2.25	15	−1.59	25	3.99	35	0.23	45	3.56
6	−2.63	16	−2.69	26	−1.76	36	−1.82	46	−0.24
7	3.57	17	3.41	27	2.31	37	0.64	47	−2.98
8	0.11	18	4.35	28	−2.24	38	−1.55	48	2.47
9	0.59	19	−4.37	29	2.95	39	0.78	49	0.66
10	−0.63	20	2.79	30	6.30	40	2.84	50	0.32

TABLE 2.4 Sample ACF of the One-Step-Ahead Forecast Errors in Table 2.3

Lag	Sample ACF, r_k	Z-Statistic	Ljung–Box Statistic, Q_{LB}
1	0.004656	0.03292	0.0012
2	−0.102647	−0.72581	0.5719
3	0.136810	0.95734	1.6073
4	−0.033988	−0.23359	1.6726
5	0.118876	0.81611	2.4891
6	0.181508	1.22982	4.4358
7	−0.039223	−0.25807	4.5288
8	−0.118989	−0.78185	5.4053
9	0.003400	0.02207	5.4061
10	0.034631	0.22482	5.4840
11	−0.151935	−0.98533	7.0230
12	−0.207710	−1.32163	9.9749
13	0.089387	0.54987	10.5363

If a time series consists of uncorrelated observations and has constant variance, we say that it is **white noise**. If, in addition, the observations in this time series are normally distributed, the time series is **Gaussian white noise**. Ideally, forecast errors are Gaussian white noise. The normal probability plot of the one-step-ahead forecast errors from Table 2.3 are shown in Figure 2.37. This plot does not indicate any serious problem, with the normality assumption, so the forecast errors in Table 2.3 are Gaussian white noise.

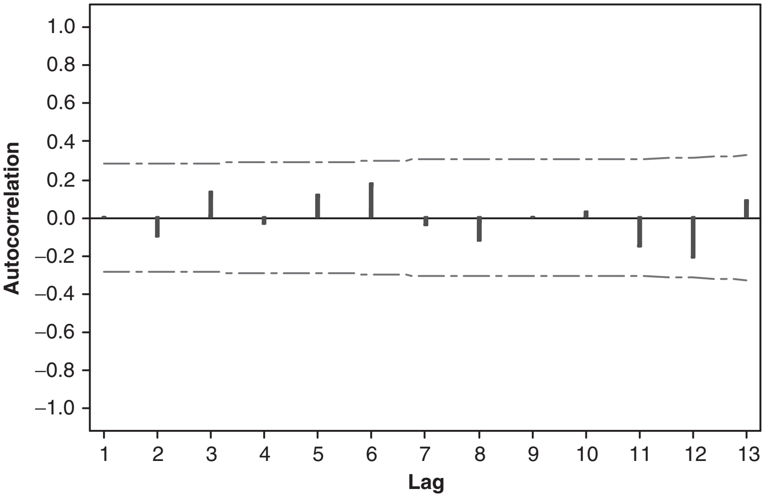


FIGURE 2.36 Sample ACF of forecast errors from Table 2.4.

If a time series is white noise, the distribution of the sample autocorrelation coefficient at lag k in large samples is approximately normal with mean zero and variance $1/T$; that is,

$$r_k \sim N\left(0, \frac{1}{T}\right).$$

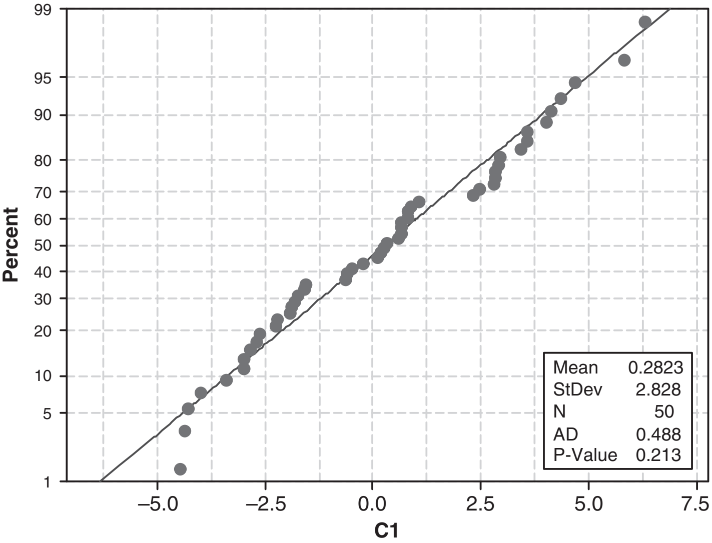


FIGURE 2.37 Normal probability plot of forecast errors from Table 2.3.

Therefore we could test the hypothesis $H_0 : \rho_k = 0$ using the test statistic

$$Z_0 = \frac{r_k}{\sqrt{\frac{1}{T}}} = r_k\sqrt{T}. \tag{2.40}$$

Minitab calculates this Z-statistic (calling it a t-statistic), and it is reported in Table 2.4 for the one-step-ahead forecast errors of Table 2.3 (this is the t-statistic reported in Figure 2.13 for the ACF of the chemical viscosity readings). Large values of this statistic (say, $|Z_0| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution) would indicate that the corresponding autocorrelation coefficient does not equal zero. Alternatively, we could calculate a P-value for this test statistic. Since none of the absolute values of the Z-statistics in Table 2.4 exceeds $Z_{\alpha/2} = Z_{0.025} = 1.96$, we cannot conclude at significance level $\alpha = 0.05$ that any individual autocorrelation coefficient differs from zero.

This procedure is a one-at-a-time test; that is, the significance level applies to the autocorrelations considered individually. We are often interested in evaluating a *set* of autocorrelations jointly to determine if they indicate that the time series is white noise. Box and Pierce (1970) have suggested such a procedure. Consider the square of the test statistic Z_0 in Eq. (2.40). The distribution of $Z_0^2 = r_k^2 T$ is approximately chi-square with one degree of freedom. The Box–Pierce statistic

$$Q_{\text{BP}} = T\sum_{k=1}^{K} r_k^2 \tag{2.41}$$

is distributed approximately as chi-square with Kdegrees of freedom under the null hypothesis that the time series is white noise. Therefore, if $Q_{\text{BP}} > \chi^2_{\alpha,K}$ we would reject the null hypothesis and conclude that the time series is not white noise because some of the autocorrelations are not zero. A P-value approach could also be used. When this test statistic is applied to a set of **residual autocorrelations** the statistic $Q_{\text{BP}} \sim \chi^2_{\alpha,K-p}$, where p is the number of parameters in the model, so the number of degrees of freedom in the chi-square distribution becomes $K - p$. Box and Pierce call this procedure a “Portmanteau” or general **goodness-of-fit statistic** (it is testing the goodness of fit of the ACF to the ACF of white noise). A modification of this test that works better for small samples was devised by Ljung and Box (1978). The Ljung–Box goodness-of-fit statistic is

$$Q_{\text{LB}} = T(T+2)\sum_{k=1}^{K}\left(\frac{1}{T-k}\right) r_k^2. \tag{2.42}$$

Note that the Ljung–Box goodness-of-fit statistic is very similar to the original Box–Pierce statistic, the difference being that the squared sample autocorrelation at lag k is weighted by $(T + 2)/(T - k)$. For large values of T, these weights will be approximately unity, and so the Q_{LB} and Q_{BP} statistics will be very similar.

Minitab calculates the Ljung–Box goodness-of-fit statistic Q_{LB}, and the values for the first 13 sample autocorrelations of the one-step-ahead forecast errors of Table 2.3 are shown in the last column of Table 2.4. At lag 13, the value $Q_{\text{LB}} = 10.5363$, and since $\chi^2_{0.05,13} = 22.36$, there is no strong evidence to indicate that the first 13 autocorrelations of the forecast errors considered jointly differ from zero. If we calculate the P-value for this test statistic, we find that $P = 0.65$. This is a good indication that the forecast errors are white noise. Note that Figure 2.13 also gave values for the Ljung–Box statistic.

2.6.2 Choosing Between Competing Models

There are often several competing models that can be used for forecasting a particular time series. For example, there are several ways to model and forecast trends. Consequently, selecting an appropriate forecasting model is of considerable practical importance. In this section we discuss some general principles of model selection. In subsequent chapters, we will illustrate how these principles are applied in specific situations.

Selecting the model that provides the best fit to historical data generally does not result in a forecasting method that produces the best forecasts of new data. Concentrating too much on the model that produces the best historical fit often results in **overfitting**, or including too many parameters or terms in the model just because these additional terms improve the model fit. In general, the best approach is to select the model that results in the smallest standard deviation (or mean squared error) of the one-step-ahead forecast errors when the model is applied to data that were not used in the fitting process. Some authors refer to this as an **out-of-sample** forecast error standard deviation (or mean squared error). A standard way to measure this out-of-sample performance is by utilizing some form of **data splitting**; that is, divide the time series data into two segments—one for model fitting and the other for performance testing. Sometimes data splitting is called **cross-validation**. It is somewhat arbitrary as to how the data splitting is accomplished. However, a good rule of thumb is to have at least 20 or 25 observations in the performance testing data set.

When evaluating the fit of the model to historical data, there are several criteria that may be of value. The **mean squared error** of the residuals is

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - p} \tag{2.43}$$

where T periods of data have been used to fit a model with p parameters and e_t is the residual from the model-fitting process in period t. The mean squared error s^2 is just the sample variance of the residuals and it is an estimator of the variance of the model errors.

Another criterion is the R-squared statistic

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}. \tag{2.44}$$

The denominator of Eq. (2.44) is just the total sum of squares of the observations, which is constant (not model dependent), and the numerator is just the residual sum of squares. Therefore, selecting the model that maximizes R^2 is equivalent to selecting the model that minimizes the sum of the squared residuals. Large values of R^2 suggest a good fit to the historical data. Because the residual sum of squares always decreases when parameters are added to a model, relying on R^2 to select a forecasting model encourages overfitting or putting in more parameters than are really necessary to obtain good forecasts. A large value of R^2 does not ensure that the out-of-sample one-step-ahead forecast errors will be small.

A better criterion is the "adjusted" R^2 statistic, defined as

$$R^2_{\text{Adj}} = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T - p)}{\sum_{t=1}^{T} (y_t - \bar{y})^2/(T - 1)} = 1 - \frac{s^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2/(T - 1)}. \tag{2.45}$$

The adjustment is a "size" adjustment—that is, adjust for the number of parameters in the model. Note that a model that maximizes the adjusted R^2 statistic is also the model that minimizes the residual mean square.

Two other important criteria are the **Akaike Information Criterion** (**AIC**) (see Akaike (1974)) and the **Schwarz Bayesian Information Criterion** (**abbreviated as BIC or SIC by various authors**) (see Schwarz (1978)):

$$\mathrm{AIC} = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2p}{T} \tag{2.46}$$

and

$$\mathrm{BIC} = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{p\ln(T)}{T}. \tag{2.47}$$

These two criteria penalize the sum of squared residuals for including additional parameters in the model. Models that have small values of the AIC or BIC are considered good models.

One way to evaluate model selection criteria is in terms of **consistency**. A model selection criterion is consistent if it selects the true model when the true model is among those considered with probability approaching unity as the sample size becomes large, and if the true model is not among those considered, it selects the best approximation with probability approaching unity as the sample size becomes large. It turns out that s^2, the adjusted R^2, and the AIC are all inconsistent, because they do not penalize for adding parameters heavily enough. Relying on these criteria tends to result in overfitting. The BIC, which caries a heavier "size adjustment" penalty, is consistent.

Consistency, however, does not tell the complete story. It may turn out that the true model and any reasonable approximation to it are very complex. An **asymptotically efficient** model selection criterion chooses a sequence of models as T(the amount of data available) gets large for which the one-step-ahead forecast error variances approach the one-step-ahead forecast error variance for the true model at least as fast as any other criterion. The AIC is asymptotically efficient but the BIC is not.

There are a number of variations and extensions of these criteria. The AIC is a biased estimator of the discrepancy between all candidate models and the true model. This has led to developing a "corrected" version of AIC:

$$\mathrm{AICc} = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2T(p+1)}{T-p-2}. \tag{2.48}$$

Sometimes we see the first term in the AIC, AICc, or BIC written as $-2\ln L(\boldsymbol{\beta}, \sigma^2)$, where $L(\boldsymbol{\beta}, \sigma^2)$ is the **likelihood function** for the fitted model evaluated at the maximum likelihood estimates of the unknown parameters $\boldsymbol{\beta}$ and σ^2. In this context, AIC, AICc, and SIC are called penalized likelihood criteria.

Many software packages evaluate and print model selection criteria, such as those discussed here. When both AIC and SIC are available, we prefer using SIC. It generally results in smaller, and hence simpler, models, and so its use is consistent with the time-honored model-building principle of **parsimony** (all other things being equal, simple models are preferred to complex ones). We will discuss and illustrate model selection criteria again in subsequent chapters. However, remember that the best way to evaluate a candidate model's potential predictive performance is to use data splitting. This will provide a direct estimate of the one-step-ahead forecast error variance, and this method should always be used, if possible, along with the other criteria that we have discussed here.

2.6.3 Monitoring a Forecasting Model

Developing and implementing procedures to monitor the performance of the forecasting model is an essential component of good forecasting system design. No matter how much effort has been expended in developing the forecasting model, and regardless of how well the model works initially, over time it is likely that its performance will deteriorate. The underlying pattern of the time series may change, either because the internal inertial forces that drive the process may evolve through time, or because of external events such as new customers entering the market. For example, a level change or a slope change could occur in the variable that is being forecasted. It is also possible for the inherent variability in the data to increase. Consequently, performance monitoring is important.

The one-step-ahead forecast errors $e_t(1)$ are typically used for forecast monitoring. The reason for this is that changes in the underlying time series will also typically be reflected in the forecast errors. For example, if a level change occurs in the time series, the sequence of forecast errors will no longer fluctuate around zero; that is, a positive or negative bias will be introduced.

There are several ways to monitor forecasting model performance. The simplest way is to apply **Shewhart control charts** to the forecast errors. A Shewhart control chart is a plot of the forecast errors versus time containing a center line that represents the average (or the target value) of the forecast errors and a set of **control limits** that are designed to provide an indication that the forecasting model performance has changed. The center line is usually taken as either zero (which is the anticipated forecast error for an unbiased forecast) or the average forecast error (ME from Eq. (2.32)), and the control limits are typically placed at three standard deviations of the forecast errors above and below the center line. If the forecast errors plot within the control limits, we assume that the forecasting model performance is satisfactory (or in control), but if one or more forecast errors exceed the control limits, that is a signal that something has happened and the forecast errors are no longer fluctuating around zero. In control chart terminology, we would say that the forecasting process is out of control and some analysis is required to determine what has happened.

The most familiar Shewhart control charts are those applied to data that have been collected in subgroups or samples. The one-step-ahead forecast errors $e_t(1)$ are individual observations. Therefore the Shewhart control chart for individuals would be used for forecast monitoring. On this control chart it is fairly standard practice to estimate the standard deviation of the individual observations using a moving range method. The moving range is defined as the absolute value of the difference between any two successive one-step-ahead forecast errors, say, $|e_t(1) - e_{t-1}(1)|$, and the moving range based on n observations is

$$MR = \sum_{t=2}^{n} |e_t(1) - e_{t-1}(1)|. \tag{2.49}$$

The estimate of the standard deviation of the one-step-ahead forecast errors is based on the average of the moving ranges

$$\hat{\sigma}_{e(1)} = \frac{0.8865MR}{n-1} = \frac{0.8865 \sum_{t=2}^{n} |e_t(1) - e_{t-1}(1)|}{n-1} = 0.8865\overline{MR}, \tag{2.50}$$

where $\overline{MR}$ is the average of the moving ranges. This estimate of the standard deviation would be used to construct the control limits on the control chart for forecast errors. For more details on constructing and interpreting control charts, see Montgomery (2013).

Example 2.12 Minitab can be used to construct Shewhart control charts for individuals. Figure 2.38 shows the Minitab control charts for the one-step-ahead forecast errors in Table 2.3. Note that both an individuals control chart of the one-step-ahead forecast errors and a control chart of the moving ranges of these forecast errors are provided. On the individuals control chart the center line is taken to be the average of the forecast errors ME defined in Eq. (2.30) (denoted $\overline{X}$ in Figure 2.38) and the upper and lower three-sigma control limits are abbreviated as UCL and LCL, respectively. The center line on the moving average control chart is at the average of the moving ranges $\overline{MR} = MR/(n-1)$, the three-sigma upper control limit UCL is at $3.267MR/(n-1)$, and the lower control limit is at zero (for details on how the control limits are derived, see Montgomery (2013)). All of the one-step-ahead forecast errors plot within the control limits (and the moving range also plot within their control limits). Thus there is no reason to suspect that the forecasting model is performing inadequately, at least from the statistical stability viewpoint. Forecast errors that plot outside the control limits would indicate model inadequacy, or possibly the presence of unusual observations such as outliers in the data. An investigation would be required to determine why these forecast errors exceed the control limits.

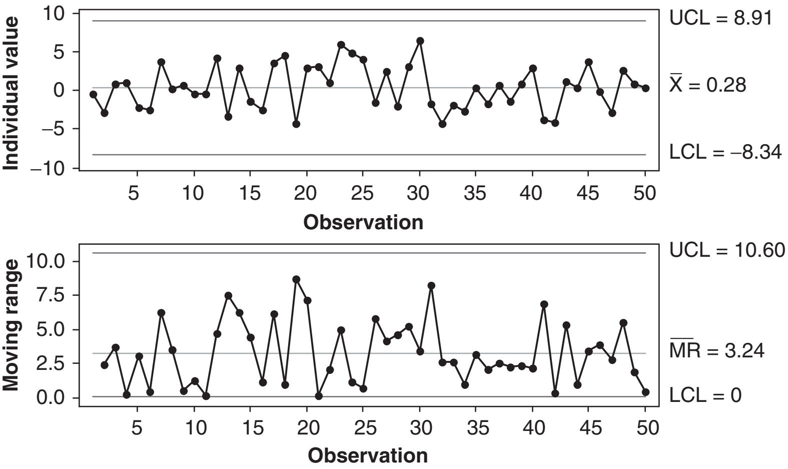


FIGURE 2.38 Individuals and moving range control charts of the one-step-ahead forecast errors in Table 2.3.

Because the control charts in Figure 2.38 exhibit statistical control, we would conclude that there is no strong evidence of statistical inadequacy in the forecasting model. Therefore, these control limits would be retained and used to judge the performance of future forecasts (in other words, we do not recalculate the control limits with each new forecast). However, the stable control chart does not imply that the forecasting performance is satisfactory in the sense that the model results in small forecast errors. In the quality control literature, these two aspects of process performance are referred to as control and capability, respectively. It is possible for the forecasting process to be stable or in statistical control but not capable—that is, produce forecast errors that are unacceptably large.

Two other types of control charts, the cumulative sum (or CUSUM) control chart and the exponentially weighted moving average (or EWMA) control chart, can also be useful for monitoring the performance of a forecasting model. These charts are more effective at detecting smaller changes or disturbances in the forecasting model performance than the individuals control chart. The CUSUM is very effective in detecting level changes in the monitored variable. It works by accumulating deviations of the forecast errors that are above the desired target value T_0 (usually either zero or the average forecast error) with one statistic C^+ and deviations that are below the target with another statistic C^-. The statistics C^+ and C^- are called the upper and lower CUSUMs, respectively. They are computed as follows:

$$\begin{aligned} C_t^+ &= \max[0, e_t(1) - (T_0 + K) + C_{t-1}^+] \\ C_t^- &= \min[0, e_t(1) - (T_0 - K) + C_{t-1}^-] \end{aligned}, \tag{2.51}$$

where the constant K, usually called the reference value, is usually chosen as $K = 0.5\sigma_{e(1)}$ and $\sigma_{e(1)}$ is the standard deviation of the one-step-ahead forecast errors. The logic is that if the forecast errors begin to systematically fall on one side of the target value (or zero), one of the CUSUMs in Eq. (2.51) will increase in magnitude. When this increase becomes large enough, an out-of-control signal is generated. The decision rule is to signal if the statistic C^+ exceeds a decision interval $H = 5\sigma_{e(1)}$ or if C^- exceeds $-H$. The signal indicates that the forecasting model is not performing satisfactorily (Montgomery (2013) discusses the choice of H and K in detail).

Example 2.13 The CUSUM control chart for the forecast errors shown in Table 2.3 is shown in Figure 2.39. This CUSUM chart was constructed

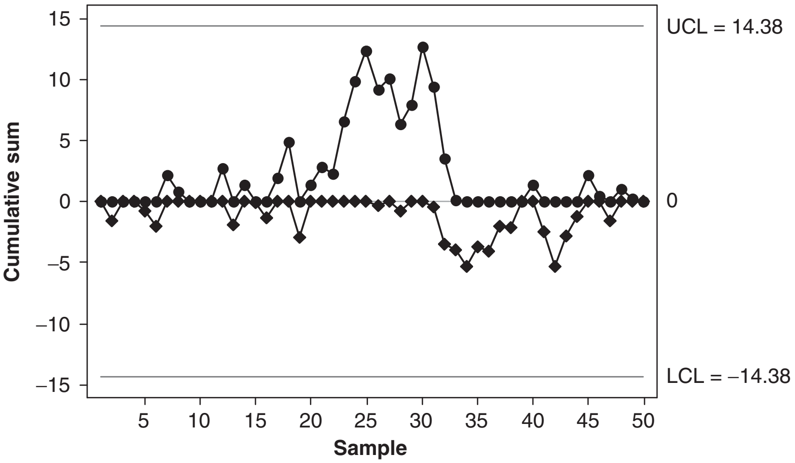


FIGURE 2.39 CUSUM control chart of the one-step-ahead forecast errors in Table 2.3.

using Minitab with a target value of $T = 0$ and $\sigma_{e(1)}$ was estimated using the moving range method described previously, resulting in $H = 5\hat{\sigma}_{e(1)} = 5(0.8865)MR/(T-1) = 5(0.8865)3.24 = 14.36$. Minitab labels H and $-H$ as UCL and LCL, respectively. The CUSUM control chart reveals no obvious forecasting model inadequacies.

A control chart based on the EWMA is also useful for monitoring forecast errors. The EWMA applied to the one-step-ahead forecast errors is

$$\bar{e}_t(1) = \lambda e_t(1) + (1-\lambda)\bar{e}_{t-1}(1), \tag{2.52}$$

where $0 < \lambda < 1$ is a constant (usually called the smoothing constant) and the starting value of the EWMA (required at the first observation) is either $\bar{e}_0(1) = 0$ or the average of the forecast errors. Typical values of the smoothing constant for an EWMA control chart are $0.05 < \lambda < 0.2$.

The EWMA is a weighted average of all current and previous forecast errors, and the weights decrease geometrically with the "age" of the forecast error. To see this, simply substitute recursively for $\bar{e}_{t-1}(1)$, then $\bar{e}_{t-2}(1)$, then $\bar{e}_{t-j}(1)_j$ for $j = 3, 4, \ldots$, until we obtain

$$\bar{e}_n(1) = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j e_{T-j}(1) + (1-\lambda)^n \bar{e}_0(1)$$

and note that the weights sum to unity because

$$\lambda \sum_{j=0}^{n-1} (1 - \lambda)^j = 1 - (1 - \lambda)^n.$$

The standard deviation of the EWMA is

$$\sigma_{\bar{e}_t(1)} = \sigma_{e(1)} \sqrt{\frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]}.$$

So an EWMA control chart for the one-step-ahead forecast errors with a center line of T (the target for the forecast errors) is defined as follows:

$$\begin{aligned} \text{UCL} &= T + 3\sigma_{e(1)} \sqrt{\frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]} \\ \text{Center line} &= T \\ \text{LCL} &= T - 3\sigma_{e(1)} \sqrt{\frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]} \end{aligned} \tag{2.53}$$

Example 2.14 Minitab can be used to construct EWMA control charts. Figure 2.40 is the EWMA control chart of the forecast errors in Table 2.3. This chart uses the mean forecast error as the center line, $\sigma_{e(1)}$ was estimated using the moving range method, and we chose $\lambda = 0.1$. None of the forecast

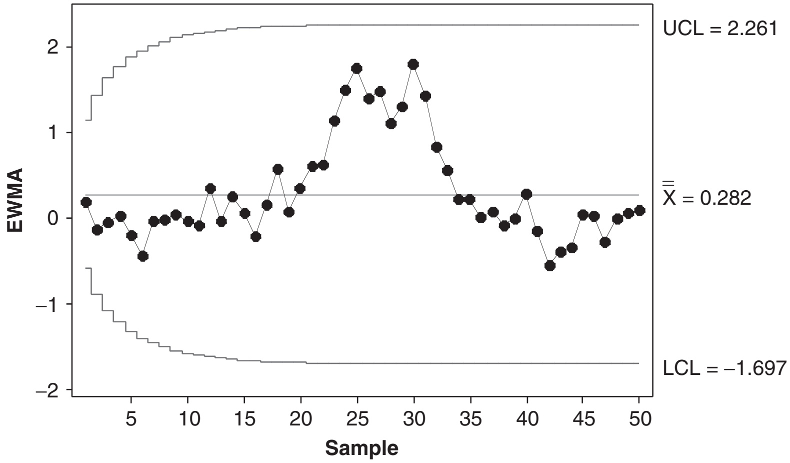


FIGURE 2.40 EWMA control chart of the one-step-ahead forecast errors in Table 2.3.

errors exceeds the control limits so there is no indication of a problem with the forecasting model.

Note from Eq. (2.51) and Figure 2.40 that the control limits on the EWMA control chart increase in width for the first few observations and then stabilize at a constant value because the term $[1 - (1 - \lambda)^{2t}]$ approaches unity as t increases. Therefore steady-state limits for the EWMA control chart are

$$\begin{aligned} \text{UCL} &= T_0 + 3\sigma_{e(1)}\sqrt{\frac{\lambda}{2-\lambda}} \\ \text{Center line} &= T \\ \text{LCL} &= T_0 - 3\sigma_{e(1)}\sqrt{\frac{\lambda}{2-\lambda}}. \end{aligned} \tag{2.54}$$

In addition to control charts, other statistics have been suggested for monitoring the performance of a forecasting model. The most common of these are **tracking signals**. The cumulative error tracking signal (CETS) is based on the cumulative sum of all current and previous forecast errors, say,

$$Y(n) = \sum_{t=1}^{n} e_t(1) = Y(n-1) + e_n(1).$$

If the forecasts are unbiased, we would expect $Y(n)$ to fluctuate around zero. If it differs from zero by very much, it could be an indication that the forecasts are biased. The standard deviation of $Y(n)$, say, $\sigma_{Y(n)}$, will provide a measure of how far $Y(n)$ can deviate from zero due entirely to random variation. Therefore, we would conclude that the forecast is biased if $|Y(n)|$ exceeds some multiple of its standard deviation. To operationalize this, suppose that we have an estimate $\hat{\sigma}_{Y(n)}$ of $\sigma_{Y(n)}$ and form the **cumulative error tracking signal**

$$\text{CETS} = \left| \frac{Y(n)}{\hat{\sigma}_{Y(n)}} \right|. \tag{2.55}$$

If the CETS exceeds a constant, say, K_1, we would conclude that the forecasts are biased and that the forecasting model may be inadequate.

It is also possible to devise a **smoothed error tracking signal** based on the smoothed one-step-ahead forecast errors in Eq. (2.52). This would lead to a ratio

$$\text{SETS} = \left| \frac{\bar{e}_n(1)}{\hat{\sigma}_{\bar{e}_n(1)}} \right|. \tag{2.56}$$

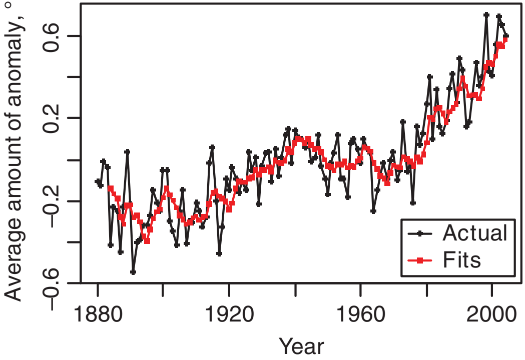
If the SETS exceeds a constant, say, K_2, this is an indication that the forecasts are biased and that there are potentially problems with the forecasting model.

Note that the CETS is very similar to the CUSUM control chart and that the SETS is essentially equivalent to the EWMA control chart. Furthermore, the CUSUM and EWMA are available in standard statistics software (such as Minitab) and the tracking signal procedures are not. So, while tracking signals have been discussed extensively and recommended by some authors, we are not going to encourage their use. Plotting and periodically visually examining a control chart of forecast errors is also very informative, something that is not typically done with tracking signals.

2.7 R COMMANDS FOR CHAPTER 2

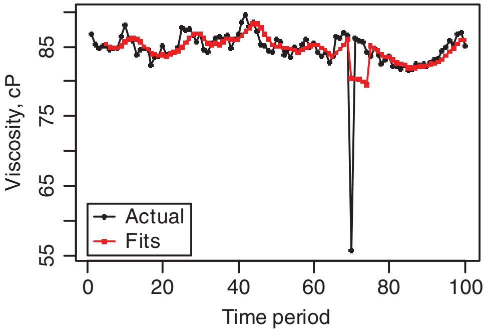
Example 2.15 The data are in the second column of the array called gms.data in which the first column is the year. For moving averages, we use functions from package "zoo."

```
plot(gms.data,type="l",xlab='Year',ylab='Average Amount of
Anomaly, °C')
points(gms.data,pch=16,cex=.5)
lines(gms.data[5:125,1],rollmean(gms.data[,2],5),col="red")
points(gms.data[5:125,1],rollmean(gms.data[,2],5),col="red",pch=15,
cex=.5)
legend(1980,-.3,c("Actual","Fits"), pch=c(16,15),lwd=c(.5,.5),
cex=.55,col=c("black","red"))
```

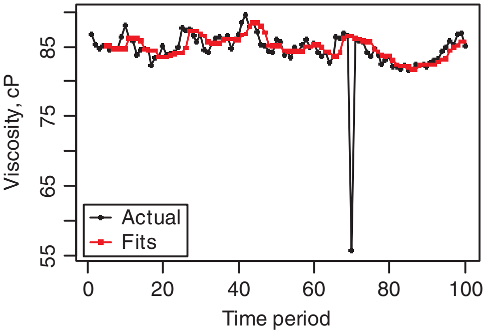


Example 2.16 The data are in the second column of the array called vis.data in which the first column is the time period (or index).

```
# Moving Average
plot(vis.data,type="l",xlab='Time Period',ylab='Viscosity, cP')
points(vis.data,pch=16,cex=.5)
lines(vis.data[5:100,1], rollmean(vis.data[,2],5),col="red")
points(vis.data[5:100,1], rollmean(vis.data[,2],5),col="red",
pch=15,cex=.5)
legend(1,61,c("Actual","Fits"), pch=c(16,15),lwd=c(.5,.5),cex=.55,
col=c("black","red"))
```



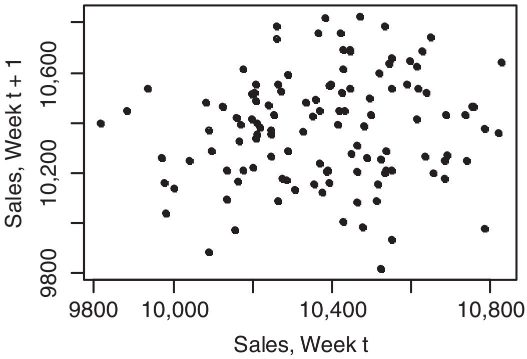
```
# Moving Median
plot(vis.data,type="l",xlab='Time Period',ylab='Viscosity, cP')
points(vis.data,pch=16,cex=.5)
lines(vis.data[5:100,1], rollmedian(vis.data[,2],5),col="red")
points(vis.data[5:100,1], rollmedian(vis.data[,2],5),col="red",
pch=15,cex=.5)
legend(1,61,c("Actual","Fits"), pch=c(16,15),lwd=c(.5,.5),cex=.55,
col=c("black","red"))
```

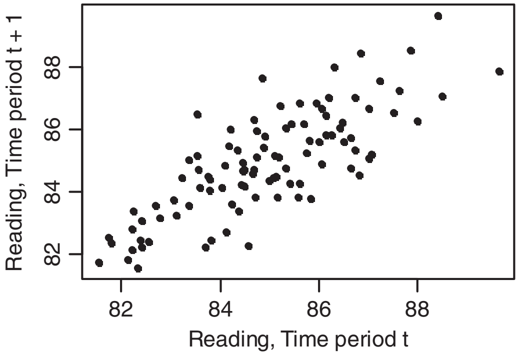


Example 2.17 The pharmaceutical sales data are in the second column of the array called pharma.data in which the first column is the week.

The viscosity data are in the second column of the array called vis.data in which the first column is the year (Note that the 70th observation is corrected).

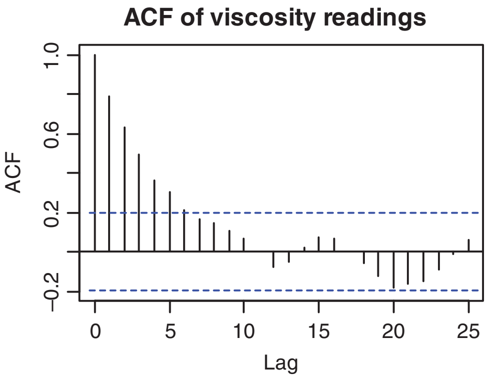
```
nrp<-dim(pharma.data)[1]

nrv<-dim(vis.data)[1]

plot(pharma.data[1:(nrp-1),2], pharma.data[2:nrp,2],type="p",
xlab='Sales, Week t',ylab=' Sales, Week t+1',pch=20,cex=1)

plot(vis.data[1:(nrv-1),2], vis.data[2:nrv,2],type="p", xlab=
'Reading, Time Period t',ylab=' Reading, Time Period t+1',pch=20,
cex=1)
```





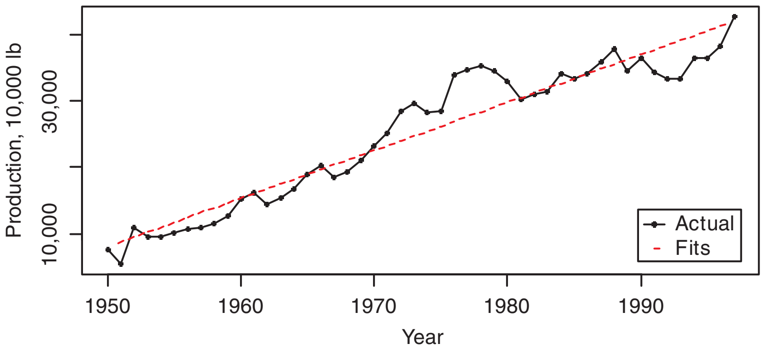
Example 2.18 The viscosity data are in the second column of the array called vis.data in which the first column is the year (Note that the 70th observation is corrected).

```
acf(vis.data[,2], lag.max=25,type="correlation",main="ACF of
viscosity readings")
```

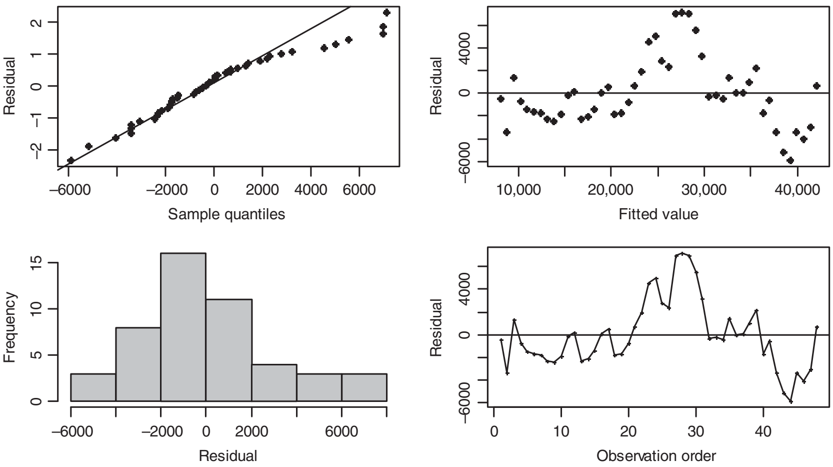


Example 2.19 The cheese production data are in the second column of the array called cheese.data in which the first column is the year.

```
fit.cheese<-lm(cheese.data[,2]~cheese.data[,1])
plot(cheese.data,type="l",xlab='Year',ylab='Production, 10000lb')
points(cheese.data,pch=16,cex=.5)
lines(cheese.data[,1], fit.cheese$fit,col="red",lty=2)
legend(1990,12000,c("Actual","Fits"),
pch=c(16,NA),lwd=c(.5,.5),lty=c(1,2),cex=.55,col=c("black","red"))
```

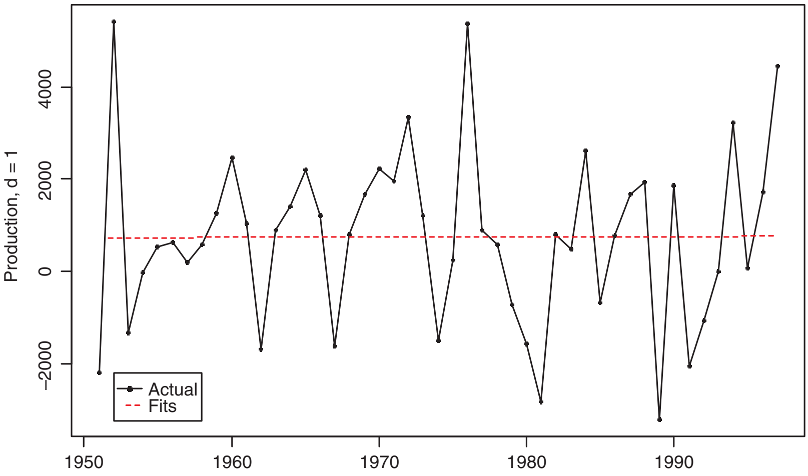


```
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(fit.cheese$res,datax=TRUE,pch=16,xlab='Residual',main='')
qqline(fit.cheese$res,datax=TRUE)
plot(fit.cheese$fit,fit.cheese$res,pch=16, xlab='Fitted Value',
ylab='Residual')
abline(h=0)
hist(fit.cheese$res,col="gray",xlab='Residual',main='')
plot(fit.cheese$res,type="l",xlab='Observation Order',
ylab='Residual')
points(fit.cheese$res,pch=16,cex=.5)
abline(h=0)
```

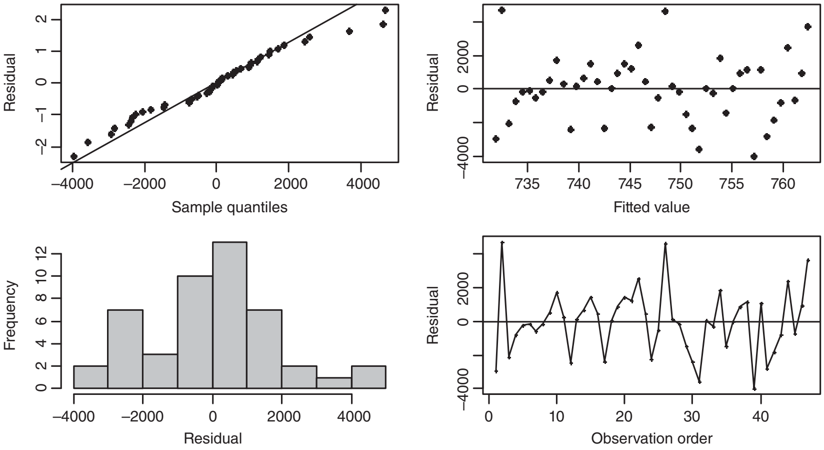


Example 2.20 The cheese production data are in the second column of the array called cheese.data in which the first column is the year.

```
nrc<-dim(cheese.data)[1]
dcheese.data<-cbind(cheese.data[2:nrc,1],diff(cheese.data[,2]))
fit.dcheese<-lm(dcheese.data[,2]~dcheese.data[,1])
plot(dcheese.data,type="l",xlab='',ylab='Production, d=1')
points(dcheese.data,pch=16,cex=.5)
lines(dcheese.data[,1], fit.dcheese$fit,col="red",lty=2)
legend(1952,-2200,c("Actual","Fits"),
pch=c(16,NA),lwd=c(.5,.5),lty=c(1,2),
cex=.75,col=c("black","red"))
```

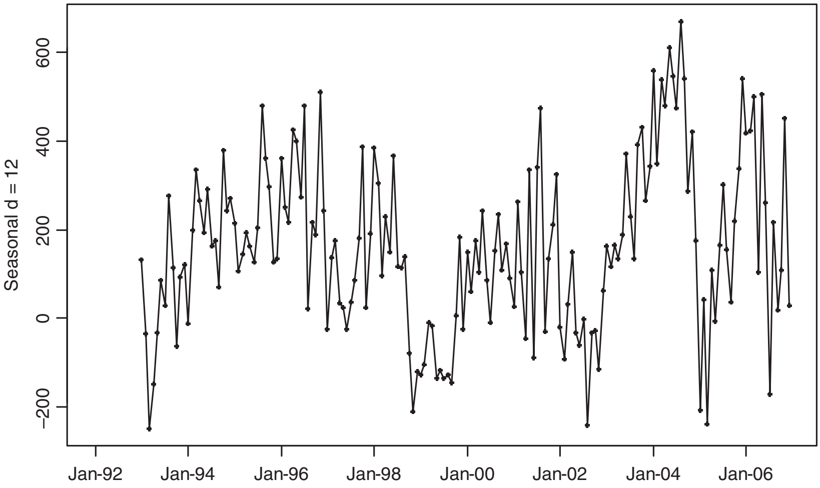


```
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(fit.dcheese$res,datax=TRUE,pch=16,xlab='Residual',main='')
qqline(fit.dcheese$res,datax=TRUE)
plot(fit.dcheese$fit,fit.dcheese$res,pch=16, xlab='Fitted Value',
ylab='Residual')
abline(h=0)
hist(fit.dcheese$res,col="gray",xlab='Residual',main='')
plot(fit.dcheese$res,type="l",xlab='Observation Order',
ylab='Residual')
points(fit.dcheese$res,pch=16,cex=.5)
abline(h=0)
```

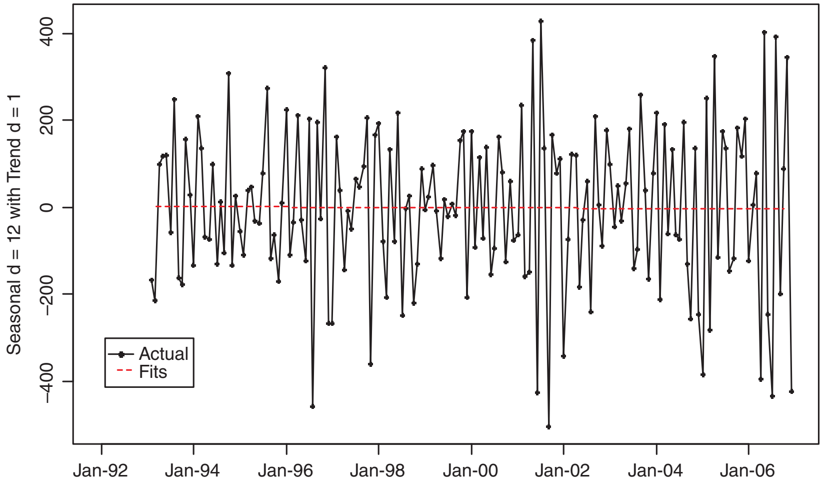


Example 2.21 The beverage sales data are in the second column of the array called bev.data in which the first column is the month of the year.

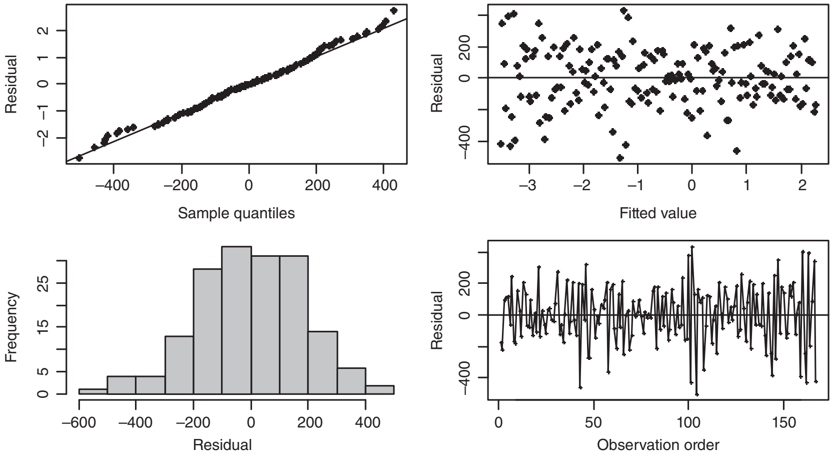
```
nrb<-dim(bev.data)[1]
tt<-1:nrb
dsbev.data<-bev.data
dsbev.data[,2]<- c(array(NA,dim=c(12,1)),diff(bev.data[,2],12))

plot(tt,dsbev.data[,2],type="l",xlab='',ylab='Seasonal d=12',
xaxt='n')axis(1,seq(1,nrb,24),labels=dsbev.data[seq(1,nrb,24),1])
points(tt,dsbev.data[,2],pch=16,cex=.5)
```



```
dstbev.data<-dsbev.data
dstbev.data[,2]<- c(NA,diff(dstbev.data[,2],1))
fit.dstbev<-lm(dstbev.data[,2]~tt)
plot(tt,dstbev.data[,2],type="l",xlab='',ylab='Seasonal d=12 with
Trend d=1',xaxt='n')
axis(1,seq(1,nrb,24),labels=dsbev.data[seq(1,nrb,24),1])
points(tt,dstbev.data[,2],pch=16,cex=.5)
lines(c(array(NA,dim=c(12,1)),fit.dstbev$fit),col="red",lty=2)
legend(2,-300,c("Actual","Fits"),
pch=c(16,NA),lwd=c(.5,.5),lty=c(1,2),cex=.75,col=c("black","red"))
```



```
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(fit.dstbev$res,datax=TRUE,pch=16,xlab='Residual',main='')
qqline(fit.dstbev$res,datax=TRUE)
plot(fit.dstbev$fit,fit.dstbev$res,pch=16, xlab='Fitted Value',
ylab='Residual')
abline(h=0)
hist(fit.dstbev$res,col="gray",xlab='Residual',main='')
plot(fit.dstbev$res,type="l",xlab='Observation Order',
ylab='Residual')
points(fit.dstbev$res,pch=16,cex=.5)
abline(h=0)
```

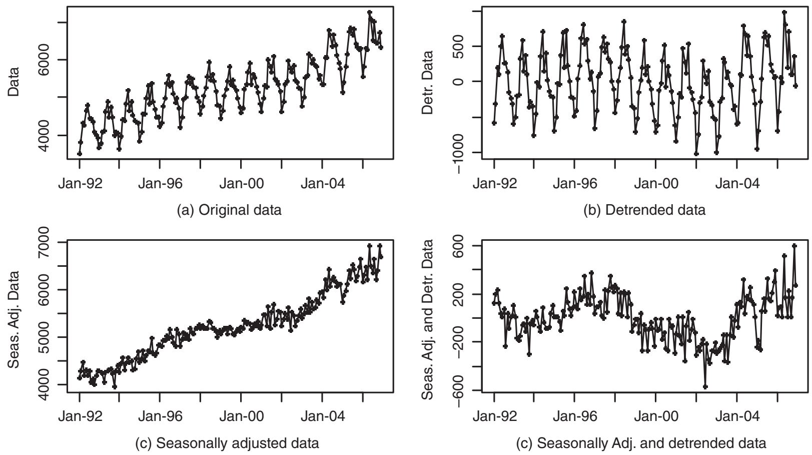


Example 2.22 The beverage sales data are in the second column of the array called bev.data in which the first column is the month of the year.

Software packages use different methods for decomposing a time series. Below we provide the code of doing it in R without using these functions. Note that we use the additive model.

```
nrb<-dim(bev.data)[1]

# De-trend the data
tt<-1:nrb
fit.tbev<-lm(bev.data[,2]~tt)
bev.data.dt<-fit.tbev$res

# Obtain seasonal medians for each month, seasonal period is sp=12
sp<-12
smed<-apply(matrix(bev.data.dt,nrow=sp),1,median)

# Adjust the medians so that their sum is zero
smed<-smed-mean(smed)

# Data without the trend and seasonal components
bev.data.dts<-bev.data.dt-rep(smed,nrb/sp)

# Note that we can also reverse the order, i.e. first take the
 seasonality out
smed2<-apply(matrix(bev.data[,2],nrow=sp),1,median)
smed2<-smed2-mean(smed2)
bev.data.ds<-bev.data[,2]-rep(smed2,nrb/sp)

# To reproduce Figure 2.25

par(mfrow=c(2,2),oma=c(0,0,0,0))
plot(tt,bev.data[,2],type="l",xlab='(a) Original Data',ylab=
'Data',xaxt='n')
axis(1,seq(1,nrb,24),labels=bev.data[seq(1,nrb,24),1])
points(tt,bev.data[,2],pch=16,cex=.75)

plot(tt, bev.data.dt,type="l",xlab='(b) Detrended Data',ylab='Detr.
Data',xaxt='n')
axis(1,seq(1,nrb,24),labels=bev.data[seq(1,nrb,24),1])

points(tt, bev.data.dt,pch=16,cex=.75)
plot(tt, bev.data.ds,type="l",xlab='(c) Seasonally Adjusted Data',
ylab='Seas.
Adj. Data',xaxt='n')
axis(1,seq(1,nrb,24),labels=bev.data[seq(1,nrb,24),1])

points(tt, bev.data.ds,pch=16,cex=.75)
```

```
plot(tt, bev.data.dts,type="l",xlab='(c) Seasonally Adj. and
Detrended Data',ylab='Seas. Adj. and Detr. Data',xaxt='n')
axis(1,seq(1,nrb,24),labels=bev.data[seq(1,nrb,24),1]) points(tt,
bev.data.dts, pch=16,cex=.75)
```



(a) Original data

(b) Detrended data

(c) Seasonally adjusted data

(c) Seasonally Adj. and detrended data

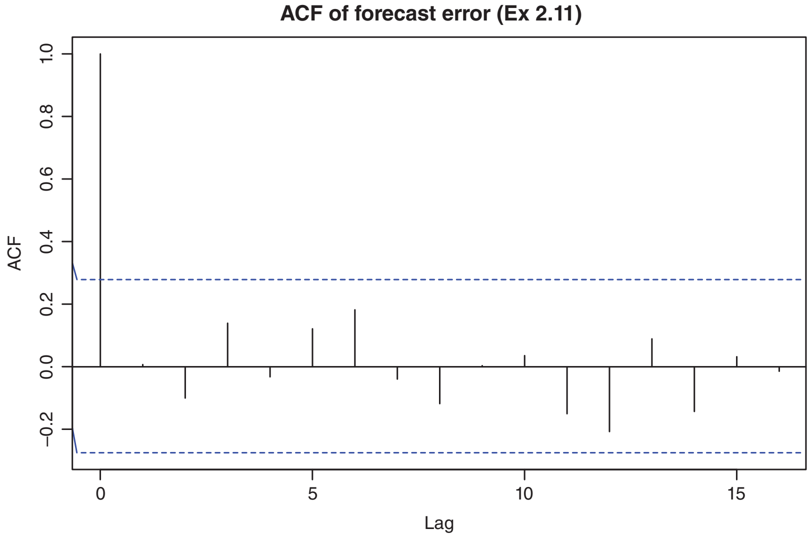
Example 2.23 Functions used to fit a time series model often also provide summary statistics. However, in this example we provide some calculations for a given set of forecast errors as provided in the text.

```
# original data and forecast errors
yt<-c(47,46,51,44,54,47,52,45,50,51,49,41,48,50,51,55,52,53,48,52)
fe<-c(-4.1,-6.9,2.2,-4.1,4.3,-.5,.8,-8.1,-4.4,-.2,-4.3,-5.5,-5.1,
-2.1,4.2,7.3,6.6,5.9,-3.8,6.2)

ME<-mean(fe)
MAD<-mean(abs(fe))
MSE<-mean(fe^2)
ret1<-(fe/yt)*100
MPE<-mean(ret1)
MAPE<-mean(abs(ret1))

> ME
[1] -0.58
> MAD
[1] 4.33
> MSE
[1] 23.59
```

```
> MPE
[1] -1.757938
> MAPE
[1] 8.865001
```

Example 2.24 The forecast error data are in the second column of the array called fe2.data in which the first column is the period.

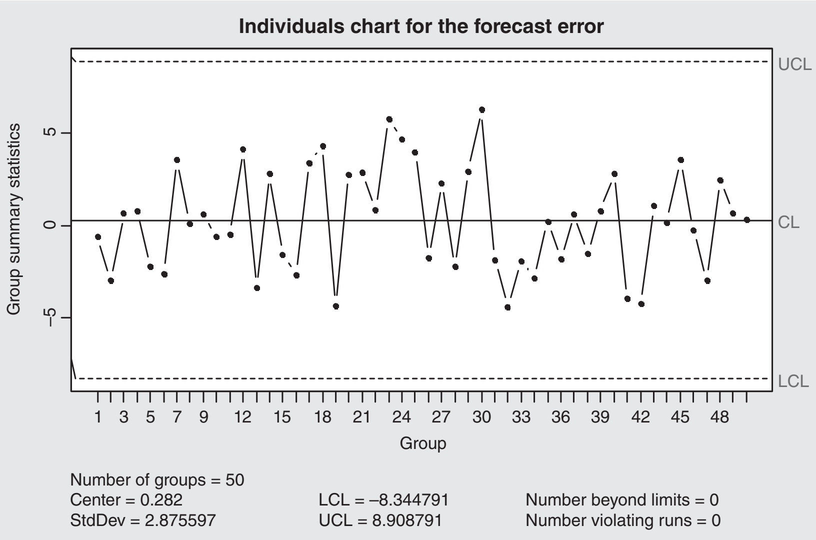
```
acf.fe2<-acf(fe2.data[,2],main='ACF of Forecast Error (Ex 2.11)')
```



```
 # To get the Q_LB statistic, we first define the lag K

K<-13
T<-dim(fe2.data)[1]
QLB<-T*(T+2)*sum((1/(T-1:K))*(acf.fe2$acf[2:(K+1)]^2))

# Upper 5% of χ² distribution with K degrees of freedom
qchisq(.95,K)
```

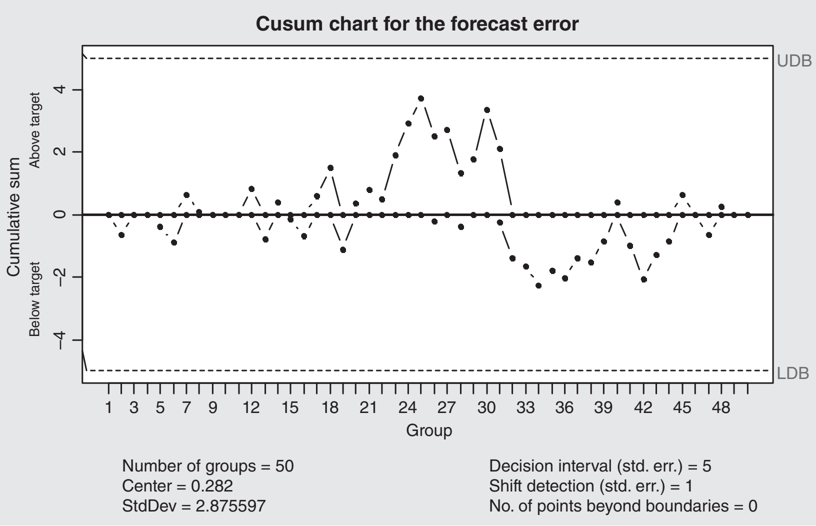
Example 2.25 The forecast error data are in the second column of the array called fe2.data in which the first column is the period.

```
# The following function can be found in qcc package
# Generating the chart for individuals
qcc(fe2.data[,2],type="xbar.one",title="Individuals Chart for the
Forecast Error")
```



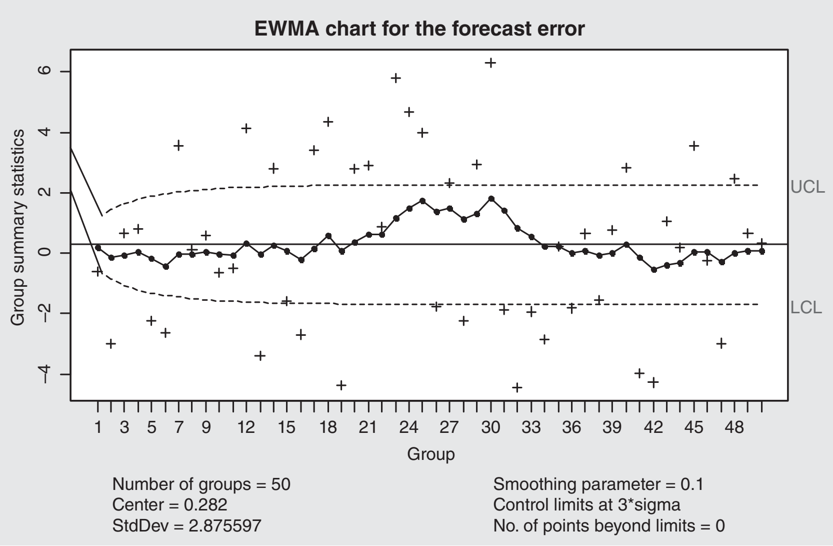
Example 2.26 The forecast error data are in the second column of the array called fe2.data in which the first column is the period.

```
# The following function can be found in qcc package
# Generating the cusum chart

cusum(fe2.data[,2], title='Cusum Chart for the Forecast
Error', sizes=1)
```



Example 2.27 The forecast error data are in the second column of the array called fe2.data in which the first column is the period.

```
# The following function can be found in qcc package
# Generating the EWMA chart
ewma(fe2.data[,2], title='EWMA Chart for the Forecast Error',
lambda=.1,sizes=1)
```



EXERCISES

2.1 Consider the US Treasury Securities rate data in Table B.1 (Appendix B). Find the sample autocorrelation function and the variogram for these data. Is the time series stationary or nonstationary?

2.2 Consider the data on US production of blue and gorgonzola cheeses in Table B.4.

a. Find the sample autocorrelation function and the variogram for these data. Is the time series stationary or nonstationary?

b. Take the first difference of the time series, then find the sample autocorrelation function and the variogram. What conclusions can you draw about the structure and behavior of the time series?

2.3 Table B.5 contains the US beverage product shipments data. Find the sample autocorrelation function and the variogram for these data. Is the time series stationary or nonstationary?

2.4 Table B.6 contains two time series: the global mean surface air temperature anomaly and the global CO_2 concentration. Find the sample autocorrelation function and the variogram for both of these time series. Is either one of the time series stationary?

2.5 Reconsider the global mean surface air temperature anomaly and the global CO_2 concentration time series from Exercise 2.4. Take the first difference of both time series. Find the sample autocorrelation function and variogram of these new time series. Is either one of these differenced time series stationary?

2.6 Find the closing stock price for a stock that interests you for the last 200 trading days. Find the sample autocorrelation function and the variogram for this time series. Is the time series stationary?

2.7 Reconsider the Whole Foods Market stock price data from Exercise 2.6. Take the first difference of the data. Find the sample autocorrelation function and the variogram of this new time series. Is this differenced time series stationary?

2.8 Consider the unemployment rate data in Table B.8. Find the sample autocorrelation function and the variogram for this time series. Is the time series stationary or nonstationary? What conclusions can you draw about the structure and behavior of the time series?

2.9 Table B.9 contains the annual International Sunspot Numbers. Find the sample autocorrelation function and the variogram for this time series. Is the time series stationary or nonstationary?

2.10 Table B.10 contains data on the number of airline miles flown in the United Kingdom. This is strongly seasonal data. Find the sample autocorrelation function for this time series.

a. Is the seasonality apparent in the sample autocorrelation function?

b. Is the time series stationary or nonstationary?

2.11 Reconsider the data on the number of airline miles flown in the United Kingdom from Exercise 2.10. Take the natural logarithm of the data and plot this new time series.

a. What impact has the log transformation had on the time series?

b. Find the autocorrelation function for this time series.

c. Interpret the sample autocorrelation function.

2.12 Reconsider the data on the number of airline miles flown in the United Kingdom from Exercises 2.10 and 2.11. Take the first difference of the natural logarithm of the data and plot this new time series.

a. What impact has the log transformation had on the time series?

b. Find the autocorrelation function for this time series.

c. Interpret the sample autocorrelation function.

2.13 The data on the number of airline miles flown in the United Kingdom in Table B.10 are seasonal. Difference the data at a season lag of 12 months and also apply a first difference to the data. Plot the differenced series. What effect has the differencing had on the time series? Find the sample autocorrelation function and the variogram. What does the sample autocorrelation function tell you about the behavior of the differenced series?

2.14 Table B.11 contains data on the monthly champagne sales in France. This is strongly seasonal data. Find the sample autocorrelation function and variogram for this time series.

a. Is the seasonality apparent in the sample autocorrelation function?

b. Is the time series stationary or nonstationary?

2.15 Reconsider the champagne sales data from Exercise 2.14. Take the natural logarithm of the data and plot this new time series.

a. What impact has the log transformation had on the time series?

b. Find the autocorrelation function and variogram for this time series.

c. Interpret the sample autocorrelation function and variogram.

2.16 Table B.13 contains data on ice cream and frozen yogurt production. Plot the data and calculate both the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the time series and compute the sample autocorrelation function and variogram of the first differences. What impact has differencing had on the time series?

2.17 Table B.14 presents data on CO_2 readings from the Mauna Loa Observatory. Plot the data, then calculate the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the time series and compute the sample autocorrelation function and the variogram of the first differences. What impact has differencing had on the time series?

2.18 Data on violent crime rates are given in Table B.15. Plot the data and calculate the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the time series and compute the sample autocorrelation function and variogram of the first differences. What impact has differencing had on the time series?

2.19 Table B.16 presents data on the US Gross Domestic Product (GDP). Plot the GDP data and calculate the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the GDP time series and compute the sample autocorrelation function and variogram of the first differences. What impact has differencing had on the time series?

2.20 Table B.17 contains information on total annual energy consumption. Plot the energy consumption data and calculate the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the time series and compute the sample autocorrelation function and variogram of the first differences. What impact has differencing had on the time series?

2.21 Data on US coal production are given in Table B.18. Plot the coal production data and calculate the sample autocorrelation function and variogram. Is there an indication of nonstationary behavior in the time series? Now plot the first difference of the time series and compute the sample autocorrelation function and variogram of the first differences. What impact has differencing had on the time series?

2.22 Consider the CO_2 readings from Mauna Loa in Table B.14. Use a six-period moving average to smooth the data. Plot both the smoothed data and the original CO_2 readings on the same axes. What has the moving average done? Repeat the procedure with a three-period moving average. What is the effect of changing the span of the moving average?

2.23 Consider the violent crime rate data in Table B.15. Use a ten-period moving average to smooth the data. Plot both the smoothed data and the original CO_2 readings on the same axes. What has the moving average done? Repeat the procedure with a four-period moving average. What is the effect of changing the span of the moving average?

2.24 Table B.21 contains data from the US Energy Information Administration on monthly average price of electricity for the residential sector in Arizona. Plot the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram. Interpret these graphs.

2.25 Reconsider the residential electricity price data from Exercise 2.24.

a. Plot the first difference of the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram for the differenced data. Interpret these graphs. What impact did differencing have?

b. Now difference the data again at a seasonal lag of 12. Plot the differenced data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram for the differenced data. Interpret these graphs. What impact did regular differencing combined with seasonal differencing have?

2.26 Table B.22 contains data from the Danish Energy Agency on Danish crude oil production. Plot the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram. Interpret these graphs.

2.27 Reconsider the Danish crude oil production data from Exercise 2.26. Plot the first difference of the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram for the differenced data. Interpret these graphs. What impact did differencing have?

2.28 Use a six-period moving average to smooth the first difference of the Danish crude oil production data that you computed in Exercise 2.27. Plot both the smoothed data and the original data on the same axes. What has the moving average done? Does the moving average look like a reasonable forecasting technique for the differenced data?

2.29 Weekly data on positive laboratory test results for influenza are shown in Table B.23. Notice that these data have a number of missing values. Construct a time series plot of the data and comment on any relevant features that you observe.

a. What is the impact of the missing observations on your ability to model and analyze these data?

b. Develop and implement a scheme to estimate the missing values

2.30 Climate data collected from Remote Automated Weather Stations (RAWS) are used to monitor the weather and to assist land management agencies with projects such as monitoring air quality, rating fire danger, and other research purposes. Data from the Western Regional Climate Center for the mean daily solar radiation (in Langleys) at the Zion Canyon, Utah, station are shown in Table B.24.

a. Plot the data and comment on any features that you observe.

b. Calculate and plot the sample ACF and variogram. Comment on the plots.

c. Apply seasonal differencing to the data, plot the data, and construct the sample ACF and variogram. What was the impact of seasonal differencing?

2.31 Table B.2 contains annual US motor vehicle traffic fatalities along with other information. Plot the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram. Interpret these graphs.

2.32 Reconsider the motor vehicle fatality data from Exercise 2.31.

a. Plot the first difference of the data and comment on any features that you observe from the graph. Calculate and plot the sample ACF and variogram for the differenced data. Interpret these graphs. What impact did differencing have?

b. Compute a six-period moving average for the differenced data. Plot the moving average and the original data on the same axes. Does it seem that the six-period moving average would be a good forecasting technique for the differenced data?

2.33 Apply the X-11 seasonal decomposition method (or any other seasonal adjustment technique for which you have software) to the mean daily solar radiation in Table B.24.

2.34 Consider the N-span moving average applied to data that are uncorrelated with mean μ and variance σ^2.

a. Show that the variance of the moving average is $\text{Var}\,(M_t) = \sigma^2/N$.

b. Show that $\text{Cov}(M_t, M_{t+k}) = \sigma^2 \sum_{j=1}^{N-k} (1/N)^2$, for $k < N$.

c. Show that the autocorrelation function is

$$\rho_k = \begin{cases} 1 - \dfrac{|k|}{N}, & k = 1, 2, \ldots, N-1 \\ 0, & k \geq N \end{cases}$$

2.35 Consider an N-span moving average where each observation is weighted by a constant, say, $a_j \geq 0$. Therefore the weighted moving average at the end of period T is

$$M_T^w = \sum_{t=T-N+1}^{T} a_{T+1-t} y_t.$$

a. Why would you consider using a weighted moving average?

b. Show that the variance of the weighted moving average is Var $(M_T^w) = \sigma^2 \sum_{j=i}^{N} a_j^2$.

c. Show that $\text{Cov}(M_T^w, M_{T+k}^w) = \sigma^2 \sum_{j=1}^{N-k} a_j a_{j+k}$, $|k| < N$.

d. Show that the autocorrelation function is

$$\rho_k = \begin{cases} \left(\sum\limits_{j=1}^{N-k} a_j a_{j+k}\right) \Big/ \left(\sum\limits_{j=1}^{N} a_j^2\right), & k = 1, 2, \ldots, N-1 \\ 0, & k \geq N \end{cases}$$

2.36 Consider the Hanning filter. This is a weighted moving average.

a. Find the variance of the weighted moving average for the Hanning filter. Is this variance smaller than the variance of a simple span-3 moving average with equal weights?

b. Find the autocorrelation function for the Hanning filter. Compare this with the autocorrelation function for a simple span-3 moving average with equal weights.

2.37 Suppose that a simple moving average of span N is used to forecast a time series that varies randomly around a constant, that is, $y_t = \mu + \varepsilon_t$, where the variance of the error term is σ^2. The forecast error at lead one is $e_{T+1}(1) = y_{T+1} - M_T$. What is the variance of this lead-one forecast error?

2.38 Suppose that a simple moving average of span N is used to forecast a time series that varies randomly around a constant, that is,

$y_t = \mu + \varepsilon_t$, where the variance of the error term is σ^2. You are interested in forecasting the cumulative value of y over a lead time of L periods, say, $y_{T+1} + y_{T+2} + \cdots + y_{T+L}$.

a. The forecast of this cumulative demand is LM_T. Why?

b. What is the variance of the cumulative forecast error?

2.39 Suppose that a simple moving average of span N is used to forecast a time series that varies randomly around a constant mean, that is, $y_t = \mu + \varepsilon_t$. At the start of period t_1 the process shifts to a new mean level, say, $\mu + \delta$. Show that the expected value of the moving average is

$$E(M_T) = \begin{cases} \mu, & T \leq t_1 - 1 \\ \mu + \dfrac{T - t_1 + 1}{N}\delta, & t_1 \leq T \leq t_1 + N - 2 \cdot \\ \mu + \delta, & T \geq t_1 + N - 1 \end{cases}$$

2.40 Suppose that a simple moving average of span N is used to forecast a time series that varies randomly around a constant mean, that is, $y_t = \mu + \varepsilon_t$. At the start of period t_1 the process experiences a transient; that is, it shifts to a new mean level, say, $\mu + \delta$, but it reverts to its original level μ at the start of period $t_1 + 1$. Show that the expected value of the moving average is

$$E(M_T) = \begin{cases} \mu, & T \leq t_1 - 1 \\ \mu + \dfrac{\delta}{N}, & t_1 \leq T \leq t_1 + N - 1 \cdot \\ \mu, & T \geq t_1 + N \end{cases}$$

2.41 If a simple $N-$span moving average is applied to a time series that has a linear trend, say, $y_t = \beta_0 + \beta_1 t + \varepsilon_t$, the moving average will lag behind the observations. Assume that the observations are uncorrelated and have constant variance. Show that at time T the expected value of the moving average is

$$E(M_T) = \beta_0 + \beta_1 T - \frac{N-1}{2}\beta_1.$$

2.42 Use a three-period moving average to smooth the champagne sales data in Table B.11. Plot the moving average on the same axes as the original data. What impact has this smoothing procedure had on the data?

TABLE E2.1 One-Step-Ahead Forecast Errors for Exercise 2.44

Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$
1	1.83	11	−2.30	21	3.30	31	−0.07
2	−1.80	12	0.65	22	1.036	32	0.57
3	0.09	13	−0.01	23	2.042	33	2.92
4	−1.53	14	−1.11	24	1.04	34	1.99
5	−0.58	15	0.13	25	−0.87	35	1.74
6	0.21	16	−1.07	26	−0.39	36	−0.76
7	1.25	17	0.80	27	−0.29	37	2.35
8	−1.22	18	−1.98	28	2.08	38	−1.91
9	1.32	19	0.02	29	3.36	39	2.22
10	3.63	20	0.25	30	−0.53	40	2.57

2.43 Use a 12-period moving average to smooth the champagne sales data in Table B.11. Plot the moving average on the same axes as the original data. What impact has this smoothing procedure had on the data?

2.44 Table E2.1 contains 40 one-step-ahead forecast errors from a forecasting model.

a. Find the sample ACF of the forecast errors. Interpret the results.

b. Construct a normal probability plot of the forecast errors. Is there evidence to support a claim that the forecast errors are normally distributed?

c. Find the mean error, the mean squared error, and the mean absolute deviation. Is it likely that the forecasting technique produces unbiased forecasts?

2.45 Table E2.2 contains 40 one-step-ahead forecast errors from a forecasting model.

a. Find the sample ACF of the forecast errors. Interpret the results.

b. Construct a normal probability plot of the forecast errors. Is there evidence to support a claim that the forecast errors are normally distributed?

c. Find the mean error, the mean squared error, and the mean absolute deviation. Is it likely that the forecasting method produces unbiased forecasts?

2.46 Exercises 2.44 and 2.45 present information on forecast errors. Suppose that these two sets of forecast errors come from two different

TABLE E2.2 One-Step-Ahead Forecast Errors for Exercise 2.45

Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$	Period, t	$e_t(1)$
1	−4.26	11	3.62	21	−6.24	31	−6.42
2	−3.12	12	−5.08	22	−0.25	32	−8.94
3	−1.87	13	−1.35	23	−3.64	33	−1.76
4	0.98	14	3.46	24	5.49	34	−0.57
5	−5.17	15	−0.19	25	−2.01	35	−10.32
6	0.13	16	−7.48	26	−4.24	36	−5.64
7	1.85	17	−3.61	27	−4.61	37	−1.45
8	−2.83	18	−4.21	28	3.24	38	−5.67
9	0.95	19	−6.49	29	−8.66	39	−4.45
10	7.56	20	4.03	30	−1.32	40	−10.23

forecasting methods applied to the same time series. Which of these two forecasting methods would you recommend for use? Why?

2.47 Consider the forecast errors in Exercise 2.44. Construct individuals and moving range control charts for these forecast errors. Does the forecasting system exhibit stability over this time period?

2.48 Consider the forecast errors in Exercise 2.44. Construct a cumulative sum control chart for these forecast errors. Does the forecasting system exhibit stability over this time period?

2.49 Consider the forecast errors in Exercise 2.45. Construct individuals and moving range control charts for these forecast errors. Does the forecasting system exhibit stability over this time period?

2.50 Consider the forecast errors in Exercise 2.45. Construct a cumulative sum control chart for these forecast errors. Does the forecasting system exhibit stability over this time period?

2.51 Ten additional forecast errors for the forecasting model in Exercise 2.44 are as follows: 5.5358, –2.6183, 0.0130, 1.3543, 12.6980, 2.9007, 0.8985, 2.9240, 2.6663, and –1.6710. Plot these additional 10 forecast errors on the individuals and moving range control charts constructed in Exercise 2.47. Is the forecasting system still working satisfactorily?

2.52 Plot the additional 10 forecast errors from Exercise 2.51 on the cumulative sum control chart constructed in Exercise 2.38. Is the forecasting system still working satisfactorily?

CHAPTER 3

REGRESSION ANALYSIS AND FORECASTING

Weather forecast for tonight: dark

GEORGE CARLIN, *American comedian*

3.1 INTRODUCTION

Regression analysis is a statistical technique for modeling and investigating the relationships between an **outcome** or **response** variable and one or more **predictor** or **regressor** variables. The end result of a regression analysis study is often to generate a model that can be used to forecast or predict future values of the response variable, given specified values of the predictor variables.

The **simple linear regression model** involves a single predictor variable and is written as

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{3.1}$$

where y is the response, x is the predictor variable, β_0 and β_1 are unknown parameters, and ε is an error term. The model parameters or **regression**

Introduction to Time Series Analysis and Forecasting, Second Edition.
Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci.

coefficients β_0 and β_1 have a physical interpretation as the intercept and slope of a straight line, respectively. The slope β_1 measures the change in the mean of the response variable y for a unit change in the predictor variable x. These parameters are typically unknown and must be estimated from a sample of data. The error term ε accounts for deviations of the actual data from the straight line specified by the model equation. We usually think of ε as a statistical error, so we define it as a random variable and will make some assumptions about its distribution. For example, we typically assume that ε is normally distributed with mean zero and variance σ^2, abbreviated $N(0, \sigma^2)$. Note that the variance is assumed constant; that is, it does not depend on the value of the predictor variable (or any other variable).

Regression models often include more than one predictor or regressor variable. If there are k predictors, the **multiple linear regression model** is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{3.2}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_k$ in this model are often called partial regression coefficients because they convey information about the effect on y of the predictor that they multiply, given that all of the other predictors in the model do not change.

The regression models in Eqs. (3.1) and (3.2) are **linear** regression models because they are linear in the unknown parameters (the β's), and not because they necessarily describe linear relationships between the response and the regressors. For example, the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

is a linear regression model because it is linear in the unknown parameters β_0, β_1, and β_2, although it describes a quadratic relationship between y and x. As another example, consider the regression model

$$y_t = \beta_0 + \beta_1 \sin \frac{2\pi}{d} t + \beta_2 \cos \frac{2\pi}{d} t + \varepsilon_t, \tag{3.3}$$

which describes the relationship between a response variable y that varies cyclically with time (hence the subscript t) and the nature of this cyclic variation can be described as a simple sine wave. Regression models such as Eq. (3.3) can be used to remove seasonal effects from time series data (refer to Section 2.4.2 where models like this were introduced). If the period d of the cycle is specified (such as $d = 12$ for monthly data with an annual cycle), then sin $(2\pi/d)t$ and cos $(2\pi/d)t$ are just numbers for each observation on the response variable and Eq. (3.3) is a standard linear regression model.

We will discuss the use of regression models for forecasting or making predictions in two different situations. The first of these is the situation where all of the data are collected on y and the regressors in a single time period (or put another way, the data are not time oriented). For example, suppose that we wanted to develop a regression model to predict the proportion of consumers who will redeem a coupon for purchase of a particular brand of milk (y) as a function of the amount of the discount or face value of the coupon (x). These data are collected over some specified study period (such as a month) and the data do not explicitly vary with time. This type of regression data is called **cross-section data**. The regression model for cross-section data is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \qquad (3.4)$$

where the subscript i is used to denote each individual observation (or case) in the data set and n represents the number of observations. In the other situation the response and the regressors are time series, so the regression model involves **time series data**. For example, the response variable might be hourly CO_2 emissions from a chemical plant and the regressor variables might be the hourly production rate, hourly changes in the concentration of an input raw material, and ambient temperature measured each hour. All of these are time-oriented or time series data.

The regression model for time series data is written as

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \quad t = 1, 2, \ldots, T \qquad (3.5)$$

In comparing Eq. (3.5) to Eq. (3.4), note that we have changed the observation or case subscript from i to t to emphasize that the response and the predictor variables are time series. Also, we have used T instead of n to denote the number of observations in keeping with our convention that, when a time series is used to build a forecasting model, T represents the most recent or last available observation. Equation (3.3) is a specific example of a time series regression model.

The unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ in a linear regression model are typically estimated using the method of **least squares**. We illustrated least squares model fitting in Chapter 2 for removing trend and seasonal effects from time series data. This is an important application of regression models in forecasting, but not the only one. Section 3.1 gives a formal description of the least squares estimation procedure. Subsequent sections deal with statistical inference about the model and its parameters, and with model adequacy checking. We will also describe and illustrate several ways in which regression models are used in forecasting.

3.2 LEAST SQUARES ESTIMATION IN LINEAR REGRESSION MODELS

We begin with the situation where the regression model is used with cross-section data. The model is given in Eq. (3.4). There are $n > k$ observations on the response variable available, say, $y_1, y_2, \ldots, y_n$. Along with each observed response y_i, we will have an observation on each regressor or predictor variable and x_{ij} denotes the ith observation or level of variable x_j. The data will appear as in Table 3.1. We assume that the error term ε in the model has expected value $E(\varepsilon) = 0$ and variance Var $(\varepsilon) = \sigma^2$, and that the errors ε_i, $i = 1, 2, \ldots, n$ are uncorrelated random variables.

The method of least squares chooses the model parameters (the β's) in Eq. (3.4) so that the sum of the squares of the errors, ε_i, is minimized. The least squares function is

$$\begin{aligned} L &= \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2. \end{aligned} \tag{3.6}$$

This function is to be minimized with respect to $\beta_0, \beta_1, \ldots, \beta_k$. Therefore the least squares estimators, say, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\beta_0, \beta_1, \ldots, \beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{3.7}$$

TABLE 3.1 Cross-Section Data for Multiple Linear Regression

Observation	Response, y	x_1	x_2	$\ldots$	x_k
1	y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0, \quad j = 1,2,\ldots,k \tag{3.8}$$

Simplifying Eqs. (3.7) and (3.8), we obtain

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik} = \sum_{i=1}^{n}y_i \tag{3.9}$$

$$\begin{gathered}\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i2}x_{i1} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}x_{i1} = \sum_{i=1}^{n}y_i x_{i1}\\ \vdots \\ \hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}x_{ik} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2}x_{ik} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 = \sum_{i=1}^{n}y_i x_{ik}\end{gathered} \tag{3.10}$$

These equations are called the **least squares normal equations**. Note that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. The solutions to the normal equations will be the least squares estimators of the model regression coefficients.

It is simpler to solve the normal equations if they are expressed in matrix notation. We now give a matrix development of the normal equations that parallels the development of Eq. (3.10). The multiple linear regression model may be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.11}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \cdots & x_{1k} \\ 1 & x_{21} & x_{22} \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In general, **y** is an $(n \times 1)$ vector of the observations, **X** is an $(n \times p)$ matrix of the levels of the regressor variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of random errors. $\mathbf{X}$ is usually called the **model matrix**, because it is the original data table for the problem expanded to the form of the regression model that you desire to fit.

The vector of least squares estimators minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

We can expand the right-hand side of L and obtain

$$L = \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

because $\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$ is a (1×1) matrix, or a scalar, and its transpose $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\boldsymbol{\beta}$ is the same scalar. The least squares estimators must satisfy

$$\left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{0},$$

which simplifies to

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{3.12}$$

In Eq. (3.12) $\mathbf{X}'\mathbf{X}$ is a $(p \times p)$ symmetric matrix and $\mathbf{X}'\mathbf{y}$ is a $(p \times 1)$ column vector. Equation (3.12) is just the matrix form of the least squares normal equations. It is identical to Eq. (3.10). To solve the normal equations, multiply both sides of Eq. (3.12) by the inverse of $\mathbf{X}'\mathbf{X}$ (we assume that this inverse exists). Thus the least squares estimator of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.13}$$

The fitted values of the response variable from the regression model are computed from

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3.14}$$

or in scalar notation,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}, \quad i = 1, 2, \ldots, n \tag{3.15}$$

The difference between the actual observation y_i and the corresponding fitted value is the **residual** $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$. The n residuals can be written as an $(n \times 1)$ vector denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3.16}$$

In addition to estimating the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$, it is also necessary to estimate the variance of the model errors, σ^2. The estimator of this parameter involves the sum of squares of the residuals

$$SS_{\mathrm{E}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

We can show that $E(SS_{\mathrm{E}}) = (n - p)\sigma^2$, so the estimator of σ^2 is the **residual** or **mean square error**

$$\hat{\sigma}^2 = \frac{SS_{\mathrm{E}}}{n - p} \tag{3.17}$$

The method of least squares is not the only way to estimate the parameters in a linear regression model, but it is widely used, and it results in estimates of the model parameters that have nice properties. If the model is correct (it has the right form and includes all of the relevant predictors), the least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of the model parameters $\boldsymbol{\beta}$; that is,

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

The variances and covariances of the estimators $\hat{\boldsymbol{\beta}}$ are contained in a $(p \times p)$ covariance matrix

$$\mathrm{Var}\,(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{3.18}$$

The variances of the regression coefficients are on the main diagonal of this matrix and the covariances are on the off-diagonals.

Example 3.1 A hospital is implementing a program to improve quality and productivity. As part of this program, the hospital is attempting to measure and evaluate patient satisfaction. Table 3.2 contains some of the data that have been collected for a random sample of 25 recently discharged patients. The "severity" variable is an index that measures the severity of

TABLE 3.2 Patient Satisfaction Survey Data

Observation	Age (x_1)	Severity (x_2)	Satisfaction (y)
1	55	50	68
2	46	24	77
3	30	46	96
4	35	48	80
5	59	58	43
6	61	60	44
7	74	65	26
8	38	42	88
9	27	42	75
10	51	50	57
11	53	38	56
12	41	30	88
13	37	31	88
14	24	34	102
15	42	30	88
16	50	48	70
17	58	61	52
18	60	71	43
19	62	62	46
20	68	38	56
21	70	41	59
22	79	66	26
23	63	31	52
24	39	42	83
25	49	40	75

the patient's illness, measured on an increasing scale (i.e., more severe illnesses have higher values of the index), and the response satisfaction is also measured on an increasing scale, with larger values indicating greater satisfaction.

We will fit a multiple linear regression model to the patient satisfaction data. The model is

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

where y = patient satisfaction, x_1 = patient age, and x_2 = illness severity. To solve the least squares normal equations, we will need to set up the $\mathbf{X}'\mathbf{X}$

matrix and the $\mathbf{X}'\mathbf{y}$ vector. The model matrix $\mathbf{X}$ and observation vector $\mathbf{y}$ are

$$\mathbf{X} = \begin{bmatrix} 1 & 55 & 50 \\ 1 & 46 & 24 \\ 1 & 30 & 46 \\ 1 & 35 & 48 \\ 1 & 59 & 58 \\ 1 & 61 & 60 \\ 1 & 74 & 65 \\ 1 & 38 & 42 \\ 1 & 27 & 42 \\ 1 & 51 & 50 \\ 1 & 53 & 38 \\ 1 & 41 & 30 \\ 1 & 37 & 31 \\ 1 & 24 & 34 \\ 1 & 42 & 30 \\ 1 & 50 & 48 \\ 1 & 58 & 61 \\ 1 & 60 & 71 \\ 1 & 62 & 62 \\ 1 & 68 & 38 \\ 1 & 70 & 41 \\ 1 & 79 & 66 \\ 1 & 63 & 31 \\ 1 & 39 & 42 \\ 1 & 49 & 40 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 68 \\ 77 \\ 96 \\ 80 \\ 43 \\ 44 \\ 26 \\ 88 \\ 75 \\ 57 \\ 56 \\ 88 \\ 88 \\ 102 \\ 88 \\ 70 \\ 52 \\ 43 \\ 46 \\ 56 \\ 59 \\ 26 \\ 52 \\ 83 \\ 75 \end{bmatrix}$$

The $\mathbf{X}'\mathbf{X}$ matrix and the $\mathbf{X}'\mathbf{y}$ vector are

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 55 & 46 & \cdots & 49 \\ 50 & 24 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 1 & 55 & 50 \\ 1 & 46 & 24 \\ \vdots & \vdots & \vdots \\ 1 & 49 & 40 \end{bmatrix} = \begin{bmatrix} 25 & 1271 & 1148 \\ 1271 & 69881 & 60814 \\ 1148 & 60814 & 56790 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 55 & 46 & \cdots & 49 \\ 50 & 24 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 68 \\ 77 \\ \vdots \\ 75 \end{bmatrix} = \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix}$$

Using Eq. (3.13), we can find the least squares estimates of the parameters in the regression model as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \begin{bmatrix} 25 & 1271 & 1148 \\ 1271 & 69881 & 60814 \\ 1148 & 60814 & 56790 \end{bmatrix}^{-1} \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix} \\
&= \begin{bmatrix} 0.699946097 & -0.006128086 & -0.007586982 \\ -0.006128086 & 0.00026383 & -0.000158646 \\ -0.007586982 & -0.000158646 & 0.000340866 \end{bmatrix} \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix} \\
&= \begin{bmatrix} 143.4720118 \\ -1.031053414 \\ -0.55603781 \end{bmatrix}
\end{aligned}$$

Therefore the regression model is

$$\hat{y} = 143.472 - 1.031x_1 - 0.556x_2,$$

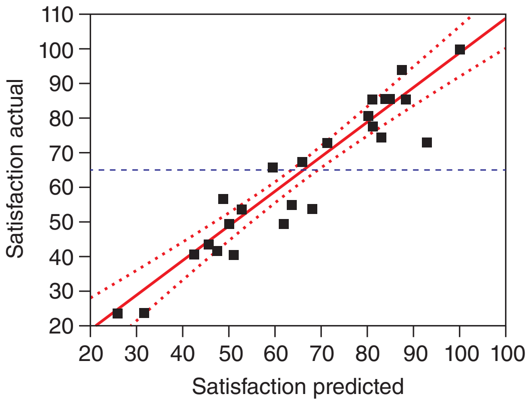
where x_1 = patient age and x_2 = severity of illness, and we have reported the regression coefficients to three decimal places.

Table 3.3 shows the output from the JMP regression routine for the patient satisfaction data. At the top of the table JMP displays a plot of the actual satisfaction data points versus the fitted values from the regression. If the fit is "perfect" then the actual-predicted and the plotted points would lie on a straight 45° line. The points do seem to scatter closely along the 45° line, suggesting that the model is a reasonably good fit to the data. Note that, in addition to the fitted regression model, JMP provides a list of the residuals computed from Eq. (3.16) along with other output that will provide information about the quality of the regression model. This output will be explained in subsequent sections, and we will frequently refer back to Table 3.3.

Example 3.2 Trend Adjustment One way to forecast time series data that contain a linear trend is with a trend adjustment procedure. This involves fitting a model with a linear trend term in time, subtracting the fitted values from the original observations to obtain a set of residuals that are trend-free, then forecast the residuals, and compute the forecast by adding the forecast of the residual value(s) to the estimate of trend. We

TABLE 3.3 JMP Output for the Patient Satisfaction Data in Table 3.2

Actual by Predicted Plot



P < .0001 RSq = 0.90 RMSE = 7.1177

Summary of Fit

RSquare	0.896593
RSquare Adj	0.887192
Root mean square error	7.117667
Mean of response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	9663.694	4831.85	95.3757
Error	22	1114.546	50.66	Prob > F
C. Total	24	10778.240		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |*t*|
Intercept	143.47201	5.954838	24.09	<.0001*
Age	−1.031053	0.115611	−8.92	<.0001*
Severity	−0.556038	0.13141	−4.23	0.0003*

(*continued*)

TABLE 3.3 (*Continued*)

Observation	Age	Severity	Satisfaction	Residual
1	55	50	68	9.03781647
2	46	24	77	−5.6986473
3	30	46	96	9.03732988
4	35	48	80	−0.6953274
5	59	58	43	−7.3896674
6	61	60	44	−3.2154849
7	74	65	26	−5.0316015
8	38	42	88	7.06160595
9	27	42	75	−17.279982
10	51	50	57	−6.0863972
11	53	38	56	−11.696744
12	41	30	88	3.48231247
13	37	31	88	−0.0858634
14	24	34	102	2.17855567
15	42	30	88	4.51336588
16	50	48	70	4.77047378
17	58	61	52	2.24739262
18	60	71	43	0.86987755
19	62	62	46	0.92764409
20	68	38	56	3.76905713
21	70	41	59	10.4992774
22	79	66	26	0.67970337
23	63	31	52	−9.2784746
24	39	42	83	3.09265936
25	49	40	75	4.29111788

described and illustrated trend adjustment in Section 2.4.2, and the basic trend adjustment model introduced there was

$$y_t = \beta_0 + \beta_1 t + \varepsilon, \quad t = 1, 2, \ldots, T.$$

The least squares normal equations for this model are

$$T\hat{\beta}_0 + \hat{\beta}_1 \frac{T(T+1)}{2} = \sum_{t=1}^{T} y_t$$
$$\hat{\beta}_0 \frac{T(T+1)}{2} + \hat{\beta}_1 \frac{T(T+1)(2T+1)}{6} = \sum_{t=1}^{T} t y_t$$

Because there are only two parameters, it is easy to solve the normal equations directly, resulting in the least squares estimators

$$\hat{\beta}_0 = \frac{2(2T+1)}{T(T-1)} \sum_{t=1}^{T} y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} t y_t$$

$$\hat{\beta}_1 = \frac{12}{T(T^2-1)} \sum_{t=1}^{T} t y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} y_t$$

Minitab computes these parameter estimates in its trend adjustment procedure, which we illustrated in Example 2.6. The least squares estimates obtained from this trend adjustment model depend on the point in time at which they were computed, that is, T. Sometimes it may be convenient to keep track of the period of computation and denote the estimates as functions of time, say, $\hat{\beta}_0(T)$ and $\hat{\beta}_1(T)$. The model can be used to predict the next observation by predicting the point on the trend line in period $T + 1$, which is $\hat{\beta}_0(T) + \hat{\beta}_1(T)(T + 1)$, and adding to the trend a forecast of the next residual, say, $\hat{e}_{T+1}(1)$. If the residuals are structureless and have average value zero, the forecast of the next residual would be zero. Then the forecast of the next observation would be

$$\hat{y}_{T+1}(T) = \hat{\beta}_0(T) + \hat{\beta}_1(T)(T+1)$$

When a new observation becomes available, the parameter estimates $\hat{\beta}_0(T)$ and $\hat{\beta}_1(T)$ could be updated to reflect the new information. This could be done by solving the normal equations again. In some situations it is possible to devise simple updating equations so that new estimates $\hat{\beta}_0(T + 1)$ and $\hat{\beta}_1(T + 1)$ can be computed directly from the previous ones $\hat{\beta}_0(T)$ and $\hat{\beta}_1(T)$ without having to directly solve the normal equations. We will show how to do this later.

3.3 STATISTICAL INFERENCE IN LINEAR REGRESSION

In linear regression problems, certain tests of hypotheses about the model parameters and confidence interval estimates of these parameters are helpful in measuring the usefulness of the model. In this section, we describe several important hypothesis-testing procedures and a confidence interval estimation procedure. These procedures require that the errors ε_i in the model are normally and independently distributed with mean zero and variance σ^2, abbreviated NID(0, σ^2). As a result of this assumption, the observations y_i are normally and independently distributed with mean $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ and variance σ^2.

3.3.1 Test for Significance of Regression

The test for significance of regression is a test to determine whether there is a linear relationship between the response variable y and a subset of the predictor or regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{ at least one } \beta_j \neq 0 \end{aligned} \tag{3.19}$$

Rejection of the null hypothesis H_0 in Eq. (3.19) implies that at least one of the predictor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model. The test procedure involves an analysis of variance partitioning of the total sum of squares

$$SS_{\mathrm{T}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.20}$$

into a sum of squares due to the **model** (or to **regression**) and a sum of squares due to **residual** (or **error**), say,

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{E}} \tag{3.21}$$

Now if the null hypothesis in Eq. (3.19) is true and the model errors are normally and independently distributed with constant variance as assumed, then the test statistic for significance of regression is

$$F_0 = \frac{SS_{\mathrm{R}}/k}{SS_{\mathrm{E}}/(n-p)} \tag{3.22}$$

and one rejects H_0 if the test statistic F_0 exceeds the upper tail point of the F distribution with k numerator degrees of freedom and $n - p$ denominator degrees of freedom, $F_{\alpha,k,n-p}$. Table A.4 in Appendix A contains these upper tail percentage points of the F distribution.

Alternatively, we could use the P-value approach to hypothesis testing and thus reject the null hypothesis if the P-value for the statistic F_0 is

TABLE 3.4 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Test Statistic, F_0
Regression	SS_{R}	k	$\dfrac{SS_{\mathrm{R}}}{k}$	$F_0 = \dfrac{SS_{\mathrm{R}}/k}{SS_{\mathrm{E}}/(n-p)}$
Residual (error)	SS_{E}	$n-p$	$\dfrac{SS_{\mathrm{E}}}{n-p}$	
Total	SS_{T}	$n-1$		

less than α. The quantities in the numerator and denominator of the test statistic F_0 are called **mean squares**. Recall that the mean square for error or residual estimates σ^2.

The test for significance of regression is usually summarized in an analysis of variance (ANOVA) table such as Table 3.4. Computational formulas for the sums of squares in the ANOVA are

$$
\begin{aligned}
SS_{\mathrm{T}} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 \\
SS_{\mathrm{R}} &= \hat{\boldsymbol{\beta}}'\mathrm{X}'\mathrm{y} - n\bar{y}^2 \\
SS_{\mathrm{E}} &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}
\end{aligned}
\tag{3.23}
$$

Regression model ANOVA computations are almost always performed using a computer software package. The JMP output in Table 3.3 shows the ANOVA test for significance of regression for the regression model for the patient satisfaction data. The hypotheses in this problem are

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = 0 \\
H_1 &: \text{ at least one } \beta_j \neq 0
\end{aligned}
$$

The reported value of the F-statistic from Eq. (3.22) is

$$F_0 = \frac{9663.694/2}{1114.546/22} = \frac{4831.85}{50.66} = 95.38.$$

and the P-value is reported as <0.0001. The actual P-value is approximately 1.44×10^{-11}, a very small value, so there is strong evidence to reject the null hypothesis and we conclude that either patient age or severity are useful predictors for patient satisfaction.

Table 3.3 also reports the coefficient of multiple determination R^2, first introduced in Section 2.6.2 in the context of choosing between competing forecasting models. Recall that

$$R^2 = \frac{SS_\mathrm{R}}{SS_\mathrm{T}} = 1 - \frac{SS_\mathrm{E}}{SS_\mathrm{T}} \tag{3.24}$$

For the regression model for the patient satisfaction data, we have

$$R^2 = \frac{SS_\mathrm{R}}{SS_\mathrm{T}} = \frac{9663.694}{10778.24} = 0.8966$$

So this model explains about 89.7% of the variability in the data.

The statistic R^2 is a measure of the amount of reduction in the variability of y obtained by using the predictor variables $x_1, x_2, \ldots, x_k$ in the model. It is a measure of how well the regression model fits the data sample. However, as noted in Section 2.6.2, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a variable to the model will never cause a decrease in R^2, even in situations where the additional variable is not statistically significant. In almost all cases, when a variable is added to the regression model R^2 increases. As a result, over reliance on R^2 as a measure of model adequacy often results in **overfitting**; that is, putting too many predictors in the model. In Section 2.6.2 we introduced the adjusted R^2 statistic

$$R^2_\mathrm{Adj} = 1 - \frac{SS_\mathrm{E}/(n-p)}{SS_\mathrm{T}/(n-1)} \tag{3.25}$$

In general, the adjusted R^2 statistic will not always increase as variables are added to the model. In fact, if unnecessary regressors are added, the value of the adjusted R^2 statistic will often decrease. Consequently, models with a large value of the adjusted R^2 statistic are usually considered good regression models. Furthermore, the regression model that maximizes the adjusted R^2 statistic is also the model that minimizes the residual mean square.

JMP reports both R^2 and R^2_{Adj} in Table 3.4. The value of $R^2 = 0.897$ (or 89.7%), and the adjusted R^2 statistic is

$$\begin{aligned} R^2_{\text{Adj}} &= 1 - \frac{SS_{\text{E}}/(n-p)}{SS_{\text{T}}/(n-1)} \\ &= 1 - \frac{1114.546/(25-3)}{10778.24/(25-1)} \\ &= 0.887. \end{aligned}$$

Both R^2 and R^2_{Adj} are very similar, usually a good sign that the regression model does not contain unnecessary predictor variables. It seems reasonable to conclude that the regression model involving patient age and severity accounts for between about 88% and 90% of the variability in the patient satisfaction data.

3.3.2 Tests on Individual Regression Coefficients and Groups of Coefficients

Tests on Individual Regression Coefficients We are frequently interested in testing hypotheses on the individual regression coefficients. These tests would be useful in determining the value or contribution of each predictor variable in the regression model. For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the mean squared error, thereby decreasing the usefulness of the model.

The hypotheses for testing the significance of any individual regression coefficient, say, β_j, are

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{3.26}$$

If the null hypothesis $H_0 : \beta_j = 0$ is not rejected, then this indicates that the predictor variable x_j can be deleted from the model.

The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \tag{3.27}$$

where C_{jj} is the diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$matrix corresponding to the regression coefficient $\hat{\beta}_j$ (in numbering the elements of the matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, it is necessary to number the first row and column as zero so that the first diagonal element C_{00} will correspond to the subscript number on the intercept). The null hypothesis $H_0 : \beta_j = 0$ is rejected if the absolute value of the test statistic $|t_0| > t_{\alpha/2,n-p}$, where $t_{\alpha/2,n-p}$ is the upper $\alpha/2$ percentage point of the t distribution with $n - p$ degrees of freedom. Table A.3 in Appendix A contains these upper tail points of the t distribution. A P-value approach could also be used. This t-test is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all the other regressor variables x_i $(i \neq j)$ that are in the model.

The denominator of Eq. (3.27), $\sqrt{\hat{\sigma}^2 C_{jj}}$, is usually called the **standard error** of the regression coefficient. That is,

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}. \tag{3.28}$$

Therefore an equivalent way to write the t-test statistic in Eq. (3.27) is

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}. \tag{3.29}$$

Most regression computer programs provide the t-test for each model parameter. For example, consider Table 3.3, which contains the JMP output for Example 3.1. The upper portion of this table gives the least squares estimate of each parameter, the standard error, the t statistic, and the corresponding P-value. To illustrate how these quantities are computed, suppose that we wish to test the hypothesis that x_1 = patient age contributes significantly to the model, given that x_2 = severity is included in the regression equation. Stated formally, the hypotheses are

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

The regression coefficient for x_1 = patient age is $\hat{\beta}_1 = -1.0311$. The standard error of this estimated regression coefficient is

$$se(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 C_{11}} = \sqrt{(50.66)(0.00026383)} = 0.1156.$$

which when rounded agrees with the JMP output. (Often manual calculations will differ slightly from those reported by the computer, because the computer carries more decimal places. For instance, in this example if the mean squared error is computed to four decimal places as $MS_E = SS_E/(n-p) = 1114.546/(25-3) = 50.6612$ instead of the two places reported in the JMP output, and this value of the MS_E is used as the estimate $\hat{\sigma}^2$ in calculating the standard error, then the standard error of $\hat{\beta}_1$ will match the JMP output.) The test statistic is computed from Eq. (3.29) as

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-1.031053}{0.115611} = -8.92$$

This is agrees with the results reported by JMP. Because the P-value reported is small, we would conclude that patient age is statistically significant; that is, it is an important predictor variable, given that severity is also in the model. Similarly, because the t-test statistic for x_2 = severity is large, we would conclude that severity is a significant predictor, given that patient age is in the model.

Tests on Groups of Coefficients We may also directly examine the contribution to the regression sum of squares for a particular predictor, say, x_j, or a **group** of predictors, given that other predictors x_i $(i \neq j)$ are included in the model. The procedure for doing this is the general regression significance test or, as it is more often called, the **extra sum of squares method**. This procedure can also be used to investigate the contribution of a *subset* involving several regressor or predictor variables to the model. Consider the regression model with k regressor variables

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.30}$$

where $\mathbf{y}$ is $(n \times 1)$, $\mathbf{X}$ is $(n \times p)$, $\boldsymbol{\beta}$ is $(p \times 1)$, $\boldsymbol{\varepsilon}$ is $(n \times 1)$, and $p = k + 1$. We would like to determine if a subset of the predictor variables $x_1, x_2, \dots, x_r$

($r < k$) contributes significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix},$$

where β_1 is ($r \times 1$) and β_2 is [($p - r$) $\times$ 1]. We wish to test the hypotheses

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_1 = \mathbf{0} \\ H_1 &: \boldsymbol{\beta}_1 \neq \mathbf{0} \end{aligned} \tag{3.31}$$

The model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{3.32}$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ (or the predictor variables) associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ (predictors) associated with $\boldsymbol{\beta}_2$.

For the **full model** (including both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$), we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Also, the regression sum of squares for all predictor variables including the intercept is

$$SS_{\mathrm{R}}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p \text{ degrees of freedom}) \tag{3.33}$$

and the estimate of σ^2 based on this full model is

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n - p} \tag{3.34}$$

$SS_{\mathrm{R}}(\boldsymbol{\beta})$ is called the regression sum of squares due to $\boldsymbol{\beta}$. To find the contribution of the terms in $\boldsymbol{\beta}_1$ to the regression, we fit the model assuming that the null hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ is true. The **reduced model** is found from Eq. (3.32) with $\boldsymbol{\beta}_1 = \mathbf{0}$:

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{3.35}$$

The least squares estimator of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{y}$ and the regression sum of squares for the reduced model is

$$SS_{\mathrm{R}}(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}'_2\mathbf{X}'_2\mathbf{y} \ (p - r \text{ degrees of freedom}) \tag{3.36}$$

The regression sum of squares due to $\boldsymbol{\beta}_1$, given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SS_{\mathrm{R}}(\boldsymbol{\beta}) - SS_{\mathrm{R}}(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'_2\mathbf{X}'_2\mathbf{y} \qquad (3.37)$$

This sum of squares has r degrees of freedom. It is the "**extra sum of squares**" due to $\boldsymbol{\beta}_1$. Note that $SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\ \boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to including the predictor variables $x_1, x_2, \ldots, x_r$ in the model. Now $SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\ \boldsymbol{\beta}_2)$ is independent of the estimate of σ^2 based on the full model from Eq. (3.34), so the null hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{\hat{\sigma}^2}, \qquad (3.38)$$

where $\hat{\sigma}^2$ is computed from Eq. (3.34). If $F_0 > F_{\alpha,r,n-p}$ we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_1$ is not zero, and, consequently, at least one of the predictor variables $x_1, x_2, \ldots, x_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. A P-value approach could also be used in testing this hypothesis. Some authors call the test in Eq. (3.38) a **partial *F* test**.

The partial F test is very useful. We can use it to evaluate the contribution of an individual predictor or regressor x_j as if it were the last variable added to the model by computing

$$SS_{\mathrm{R}}(\beta_j|\boldsymbol{\beta}_i; i \neq j)$$

This is the increase in the regression sum of squares due to adding x_j to a model that already includes $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$. The partial F test on a single variable x_j is equivalent to the t-test in Equation (3.27). The computed value of F_0 will be exactly equal to the square of the t-test statistic t_0. However, the partial F test is a more general procedure in that we can evaluate simultaneously the contribution of more than one predictor variable to the model.

Example 3.3 To illustrate this procedure, consider again the patient satisfaction data from Table 3.2. Suppose that we wish to consider fitting a more elaborate model to this data; specifically, consider the second-order polynomial

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

TABLE 3.5 JMP Output for the Second-Order Model for the Patient Satisfaction Data

Summary of Fit

RSquare	0.900772
RSquare Adj	0.874659
Root mean square error	7.502639
Mean of response	65.52
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	9708.738	1941.75	34.4957
Error	19	1069.502	56.29	Prob > F
C. Total	24	10,778.240		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> |*t*|
Intercept	143.74009	6.774622	21.22	<0.0001*
Age	−0.986524	0.135366	−7.29	<0.0001*
Severity	−0.571637	0.158928	−3.60	0.0019*
(Severity-45.92)*(Age-50.84)	0.0064566	0.016546	0.39	0.7007
(Age-50.84)*(Age-50.84)	−0.00283	0.008588	−0.33	0.7453
(Severity-45.92)*(Severity-45.92)	−0.011368	0.013533	−0.84	0.4113

where x_1 = patient age and x_2 = severity. To fit the model, the model matrix would need to be expanded to include columns for the second-order terms x_1x_2, x_1^2, and x_2^2. The results of fitting this model using JMP are shown in Table 3.5.

Suppose that we want to test the significance of the additional second-order terms. That is, the hypotheses are

$$
\begin{aligned}
&H_0 : \beta_{12} = \beta_{11} = \beta_{22} = 0 \\
&H_1 : \text{ at least one of the parameters } \beta_{12}, \beta_{11}, \text{ or } \beta_{22} \neq 0
\end{aligned}
$$

In the notation used in this section, these second-order terms are the parameters in the vector $\boldsymbol{\beta}_1$. Since the quadratic model is the full model, we can find $SS_R(\boldsymbol{\beta})$ directly from the JMP output in Table 3.5 as

$$SS_R(\boldsymbol{\beta}) = 9708.738$$

with 5 degrees of freedom (because there are five predictors in this model). The reduced model is the model with all of the predictors in the vector $\boldsymbol{\beta}_1$ equal to zero. This reduced model is the original regression model that we fit to the data in Table 3.3. From Table 3.3, we can find the regression sum of squares for the reduced model as

$$SS_{\mathrm{R}}(\boldsymbol{\beta}_2) = 9663.694$$

and this sum of squares has 2 degrees of freedom (the model has two predictors).

Therefore the extra sum of squares for testing the significance of the quadratic terms is just the difference between the regression sums of squares for the full and reduced models, or

$$\begin{aligned} SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) &= SS_{\mathrm{R}}(\boldsymbol{\beta}) - SS_{\mathrm{R}}(\boldsymbol{\beta}_2) \\ &= 9708.738 - 9663.694 \\ &= 45.044 \end{aligned}$$

with $5 - 2 = 3$ degrees of freedom. These three degrees of freedom correspond to the three additional terms in the second-order model. The test statistic from Eq. (3.38) is

$$\begin{aligned} F_0 &= \frac{SS_{\mathrm{R}}(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{\hat{\sigma}^2} \\ &= \frac{45.044/3}{56.29} \\ &= 0.267. \end{aligned}$$

This F-statistic is very small, so there is no evidence against the null hypothesis.

Furthermore, from Table 3.5, we observe that the individual t-statistics for the second-order terms are very small and have large P-values, so there is no reason to believe that the model would be improved by adding any of the second-order terms.

It is also interesting to compare the R^2 and R^2_{Adj} statistics for the two models. From Table 3.3, we find that $R^2 = 0.897$ and $R^2_{\mathrm{Adj}} = 0.887$ for the original two-variable model, and from Table 3.5, we find that $R^2 = 0.901$ and $R^2_{\mathrm{Adj}} = 0.875$ for the quadratic model. Adding the quadratic terms caused the ordinary R^2 to increase slightly (it will never decrease when additional predictors are inserted into the model), but the adjusted R^2 statistic decreased. This decrease in the adjusted R^2 is an indication that the additional variables did not contribute to the explanatory power of the model.

3.3.3 Confidence Intervals on Individual Regression Coefficients

It is often necessary to construct confidence interval (CI) estimates for the parameters in a linear regression and for other quantities of interest from the regression model. The procedure for obtaining these confidence intervals requires that we assume that the model errors are normally and independently distributed with mean zero and variance σ^2, the same assumption made in the two previous sections on hypothesis testing.

Because the least squares estimator $\hat{\boldsymbol{\beta}}$ is a linear combination of the observations, it follows that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector $\boldsymbol{\beta}$ and covariance matrix $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.Then each of the statistics

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad j = 0, 1, \ldots, k \tag{3.39}$$

is distributed as t with $n - p$ degrees of freedom, where C_{jj} is the (jj)th element of the $(\mathbf{X}'\mathbf{X})^{-1}$matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Eq. (3.34). Therefore a $100(1 - \alpha)$ percent confidence interval for an individual regression coefficient $\beta_j, j = 0, 1, \ldots, k$, is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}. \tag{3.40}$$

This CI could also be written as

$$\hat{\beta}_j - t_{\alpha/2,n-p} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p} se(\hat{\beta}_j)$$

because $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$.

Example 3.4 We will find a 95% CI on the regression for patient age in the patient satisfaction data regression model. From the JMP output in

Table 3.3, we find that $\hat{\beta}_1 = -1.0311$ and $se(\hat{\beta}_1) = 0.1156$. Therefore the 95% CI is

$$\begin{gathered}\hat{\beta}_j - t_{\alpha/2,n-p} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p} se(\hat{\beta}_j) \\ -1.0311 - (2.074)(0.1156) \leq \beta_1 \leq -1.0311 + (2.074)(0.1156) \\ -1.2709 \leq \beta_1 \leq -0.7913.\end{gathered}$$

This confidence interval does not include zero; this is equivalent to rejecting (at the 0.05 level of significance) the null hypothesis that the regression coefficient $\beta_1 = 0$.

3.3.4 Confidence Intervals on the Mean Response

We may also obtain a confidence interval on the mean response at a particular combination of the predictor or regressor variables, say, $x_{01}, x_{02}, \ldots, x_{0k}$. We first define a vector that represents this point expanded to model form. Since the standard multiple linear regression model contains the k predictors and an intercept term, this vector is

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The mean response at this point is

$$E[y(\mathbf{x}_0)] = \mu_{y|\mathbf{x}_0} = \mathbf{x}'_0 \boldsymbol{\beta}.$$

The estimator of the mean response at this point is found by substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$

$$\hat{y}(\mathbf{x}_0) = \hat{\mu}_{y|\mathbf{x}_0} = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} \tag{3.41}$$

This estimator is normally distributed because $\hat{\boldsymbol{\beta}}$ is normally distributed and it is also unbiased because $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. The variance of $\hat{y}(\mathbf{x}_0)$ is

$$\text{Var}\ [\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0. \tag{3.42}$$

Therefore, a 100(1 − α) percent CI on the mean response at the point x_{01}, $x_{02}, \ldots, x_{0k}$ is

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{y|\mathbf{x}_0} \le \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}, \tag{3.43}$$

where $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Eq. (3.34). Note that the length of this confidence interval will depend on the location of the point $\mathbf{x}_0$ through the term $\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ in the confidence interval formula. Generally, the length of the CI will increase as the point $\mathbf{x}_0$ moves further from the center of the predictor variable data.

The quantity

$$\sqrt{\text{Var }[\hat{y}(\mathbf{x}_0)]} = \sqrt{\hat{\sigma}^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

used in the confidence interval calculations in Eq. (3.43) is sometimes called the standard error of the fitted response. JMP will calculate and display these standard errors for each individual observation in the sample used to fit the model and for other non-sample points of interest. The next-to-last column of Table 3.6 displays the standard error of the fitted response for the patient satisfaction data. These standard errors can be used to compute the CI in Eq. (3.43).

Example 3.5 Suppose that we want to find a confidence interval on mean patient satisfaction for the point where x_1 = patient age = 55 and x_2 = severity = 50. This is the first observation in the sample, so refer to Table 3.6, the JMP output for the patient satisfaction regression model. For this observation, JMP reports that the "SE Fit" is 1.51 rounded to two decimal places, or in our notation, $\sqrt{\text{Var }[\hat{y}(\mathbf{x}_0)]} = 1.51$. Therefore, if we want to find a 95% CI on the mean patient satisfaction for the case where x_1 = patient age = 55 and x_2 = severity = 50, we would proceed as follows:

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{y|\mathbf{x}_0} \le \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

$$58.96 - 2.074(1.51) \le \mu_{y|\mathbf{x}_0} \le 58.96 + 2.074(1.51)$$

$$55.83 \le \mu_{y|\mathbf{x}_0} \le 62.09.$$

From inspection of Table 3.6, note that the standard errors for each observation are different. This reflects the fact that the length of the CI on

TABLE 3.6 JMP Calculations of the Standard Errors of the Fitted Values and Predicted Responses for the Patient Satisfaction Data

Observation	Age	Severity	Satisfaction	Predicted	Residual	SE (Fit)	SE (Predicted)
1	55	50	68	58.96218	9.037816	1.507444	7.275546
2	46	24	77	82.69865	−5.69865	2.988566	7.719631
3	30	46	96	86.96267	9.03733	2.803259	7.6498
4	35	48	80	80.69533	−0.69533	2.446294	7.526323
5	59	58	43	50.38967	−7.38967	1.962621	7.383296
6	61	60	44	47.21548	−3.21548	2.128407	7.429084
7	74	65	26	31.0316	−5.0316	2.89468	7.683772
8	38	42	88	80.93839	7.061606	1.919979	7.372076
9	27	42	75	92.27998	−17.28	2.895873	7.684221
10	51	50	57	63.0864	−6.0864	1.517813	7.277701
11	53	38	56	67.69674	−11.6967	1.856609	7.355826
12	41	30	88	84.51769	3.482312	2.275784	7.472641
13	37	31	88	88.08586	−0.08586	2.260863	7.468111
14	24	34	102	99.82144	2.178556	2.994326	7.721863
15	42	30	88	83.48663	4.513366	2.277152	7.473058
16	50	48	70	65.22953	4.770474	1.462421	7.266351
17	58	61	52	49.75261	2.247393	2.214287	7.454143
18	60	71	43	42.13012	0.869878	3.21204	7.808866
19	62	62	46	45.07236	0.927644	2.296	7.478823
20	68	38	56	52.23094	3.769057	3.038105	7.738945
21	70	41	59	48.50072	10.49928	2.97766	7.715416
22	79	66	26	25.3203	0.679703	3.24021	7.820495
23	63	31	52	61.27847	−9.27847	3.28074	7.837374
24	39	42	83	79.90734	3.092659	1.849178	7.353954
25	49	40	75	70.70888	4.291118	1.58171	7.291295
—	75	60	—	32.78074	—	2.78991	7.644918
—	60	60	—	48.24654	—	2.120899	7.426937

the mean response depends on the location of the observation. Generally, the standard error increases as the distance of the point from the center of the predictor variable data increases.

In the case where the point of interest $\mathbf{x}_0$ is not one of the observations in the sample, it is necessary to calculate the standard error for that point $\sqrt{\text{Var}\ [\hat{y}(\mathbf{x}_0)]} = \sqrt{\hat{\sigma}^2 \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$, which involves finding $\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ for the observation $\mathbf{x}_0$. This is not too difficult (you can do it in Excel), but it is not necessary, because JMP will provide the CI at any point that you specify. For example, if you want to find a 95% CI on the mean patient satisfaction for the point where x_1 = patient age = 60 and x_2 = severity = 60 (this is not a sample observation), then in the last row of Table 3.6 JMP reports that the estimate of the mean patient satisfaction at the point x_1 = patient age = 60 and x_2 = severity = 60 as $\hat{y}(\mathbf{x}_0) = 48.25$, and the standard error of the fitted response as $\sqrt{\text{Var}\ [\hat{y}(\mathbf{x}_0)]} = \sqrt{\hat{\sigma}^2 \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = 2.12$. Consequently, the 95% CI on the mean patient satisfaction at that point is

$$43.85 \le \mu_{y|\mathbf{x}_0} \le 52.65.$$

3.4 PREDICTION OF NEW OBSERVATIONS

A regression model can be used to predict future observations on the response y corresponding to a particular set of values of the predictor or regressor variables, say, $x_{01}, x_{02}, \dots, x_{0k}$. Let $\mathbf{x}_0$ represent this point, expanded to model form. That is, if the regression model is the standard multiple regression model, then $\mathbf{x}_0$ contains the coordinates of the point of interest and unity to account for the intercept term, so $\mathbf{x}'_0 = [1, x_{01}, x_{02}, \dots, x_{0k}]$. A point estimate of the future observation $y(\mathbf{x}_0)$ at the point $x_{01}, x_{02}, \dots, x_{0k}$ is computed from

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}'_0\hat{\boldsymbol{\beta}} \tag{3.44}$$

The prediction error in using $\hat{y}(\mathbf{x}_0)$ to estimate $y(\mathbf{x}_0)$ is $y(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0)$. Because $\hat{y}(\mathbf{x}_0)$ and $y(\mathbf{x}_0)$ are independent, the variance of this prediction error is

$$\begin{aligned}\text{Var}\ [y(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0)] &= \text{Var}\ [y(\mathbf{x}_0)] + \text{Var}\ [\hat{y}(\mathbf{x}_0)] = \sigma^2 + \sigma^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \\ &= \sigma^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]. \end{aligned} \tag{3.45}$$

If we use $\hat{\sigma}^2$ from Eq. (3.34) to estimate the error variance σ^2, then the ratio

$$\frac{y(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0)}{\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]}}$$

has a t distribution with $n - p$ degrees of freedom. Consequently, we can write the following probability statement:

$$P\left(-t_{\alpha/2,n-p} \leq \frac{y(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0)}{\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]}} \leq t_{\alpha/2,n-p}\right) = 1 - \alpha$$

This probability statement can be rearranged as follows:

$$P\left(\begin{array}{l}\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]} \leq y(\mathbf{x}_0) \\ \leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]}\end{array}\right) = 1 - \alpha.$$

Therefore, the probability is $1 - \alpha$ that the future observation falls in the interval

$$\begin{aligned}\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]} &\leq y(\mathbf{x}_0) \\ &\leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\left[1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]}\end{aligned} \tag{3.46}$$

This statement is called a $100(1 - \alpha)$ percent **prediction interval** (**PI**) for the future observation $y(\mathbf{x}_0)$ at the point $x_{01}, x_{02}, \ldots, x_{0k}$. The expression in the square tool in Eq. (3.46) is often called the standard error of the predicted response.

The PI formula in Eq. (3.46) looks very similar to the formula for the CI on the mean, Eq. (3.43). The difference is the “1” in the variance of the prediction error under the square root. This will make PI longer than the corresponding CI at the same point. It is reasonable that the PI should be longer, as the CI is an interval estimate on the mean of the response distribution at a specific point, while the PI is an interval estimate on a single future observation from the response distribution at that point. There should be more variability associated with an individual observation than with an estimate of the mean, and this is reflected in the additional length of the PI.

Example 3.6 JMP will compute the standard errors of the predicted response so it is easy to construct the prediction interval in Eq. (3.46). To illustrate, suppose that we want a 95% PI on a future observation of patient satisfaction for a patient whose age is 75 and with severity of illness 60. In the next to last row of Table 3.6 JMP predicted value of satisfaction at this new observation as $\hat{y}(\mathbf{x}_0) = 32.78$, and the standard error of the predicted response is 7.65. Then from Eq. (3.46) the prediction interval is

$$16.93 \leq y(\mathbf{x}_0) \leq 48.64.$$

This example provides us with an opportunity to compare prediction and confidence intervals. First, note that from Table 3.6 the standard error of the fit at this point is smaller than the standard error of the prediction. Therefore, the PI is longer than the corresponding CI. Now compare the length of the CI and the PI for this point with the length of the CI and the PI for the point x_1 = patient age = 60 and x_2 = severity = 60 from Example 3.4. The intervals are longer for the point in this example because this point with x_1 = patient age = 75 and x_2 = severity = 60 is further from the center of the predictor variable data than the point in Example 3.4, where x_1 = patient age = 60 and x_2 = severity = 60.

3.5 MODEL ADEQUACY CHECKING

3.5.1 Residual Plots

An important part of any data analysis and model-building procedure is checking the adequacy of the model. We know that all models are wrong, but a model that is a reasonable fit to the data used to build it and that does not seriously ignore or violate any of the underlying model-building assumptions can be quite useful. Model adequacy checking is particularly important in building regression models for purposes of forecasting, because forecasting will almost always involve some extrapolation or projection of the model into the future, and unless the model is reasonable the forecasting process is almost certainly doomed to failure.

Regression model residuals, originally defined in Eq. (2.2), are very useful in model adequacy checking and to get some sense of how well the regression model assumptions of normally and independently distributed model errors with constant variance are satisfied. Recall that if y_i is the observed value of the response variable and if the corresponding fitted value from the model is $\hat{y}_i$, then the residuals are

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Residual plots are the primary approach to model adequacy checking. The simplest way to check the adequacy of the normality assumption on the model errors is to construct a normal probability plot of the residuals. In Section 2.6.1 we introduced and used the normal probability plot of forecast errors to check for the normality of forecast errors. The use of the normal probability plot for regression residuals follows the same approach. To check the assumption of constant variance, plot the residuals versus the fitted values from the model. If the constant variance assumption is satisfied, this plot should exhibit a random scatter of residuals around zero. Problems with the equal variance assumption usually show up as a **pattern** on this plot. The most common pattern is an outward-opening funnel or megaphone pattern, indicating that the variance of the observations is increasing as the mean increases. Data **transformations** (see Section 2.4.1) are useful in stabilizing the variance. The log transformation is frequently useful in forecasting applications. It can also be helpful to plot the residuals against each of the predictor or regressor variables in the model. Any deviation from random scatter on these plots can indicate how well the model fits a particular predictor.

When the data are a time series, it is also important to plot the residuals versus **time order**. As usual, the anticipated pattern on this plot is random scatter. Trends, cycles, or other patterns in the plot of residuals versus time indicate model inadequacies, possibly due to missing terms or some other model specification issue. A funnel-shaped pattern that increases in width with time is an indication that the variance of the time series is increasing with time. This happens frequently in economic time series data, and in data that span a long period of time. Log transformations are often useful in stabilizing the variance of these types of time series.

Example 3.7 Table 3.3 presents the residuals for the regression model for the patient satisfaction data from Example 3.1. Figure 3.1 presents plots of these residuals. The plot in the upper left-hand portion of the display is a normal probability plot of the residuals. The residuals lie generally along a straight line, so there is no obvious reason to be concerned with

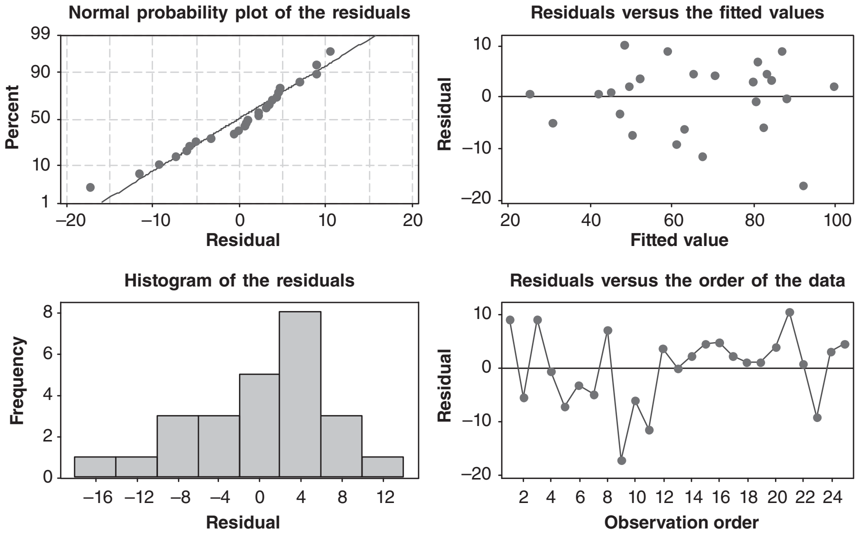


FIGURE 3.1 Plots of residuals for the patient satisfaction model.

the normality assumption. There is a very mild indication that one of the residuals (in the lower tail) may be slightly larger than expected, so this could be an indication of an outlier (a very mild one). The lower left plot is a histogram of the residuals. Histograms are more useful for large samples of data than small ones, so since there are only 25 residuals, this display is probably not as reliable as the normal probability plot. However, the histogram does not give any serious indication of nonnormality. The upper right is a plot of residuals versus the fitted values. This plot indicates essentially random scatter in the residuals, the ideal pattern. If this plot had exhibited a funnel shape, it could indicate problems with the equality of variance assumption. The lower right is a plot of the observations in the order of the data. If this was the order in which the data were collected, or if the data were a time series, this plot could reveal information about how the data may be changing over time. For example, a funnel shape on this plot might indicate that the variability of the observations was changing with time.

In addition to residual plots, other model diagnostics are frequently useful in regression. The following sections introduce and briefly illustrate some of these procedures. For more complete presentations, see Montgomery, Peck, and Vining (2012) and Myers (1990).

3.5.2 Scaled Residuals and PRESS

Standardized Residuals Many regression model builders prefer to work with **scaled residuals** in contrast to the ordinary least squares (OLS) residuals. These scaled residuals frequently convey more information than do the ordinary residuals. One type of scaled residual is the **standardized residual**,

$$d_i = \frac{e_i}{\hat{\sigma}}, \tag{3.47}$$

where we generally use $\hat{\sigma} = \sqrt{MS_{\mathrm{E}}}$ in the computation. The standardized residuals have mean zero and approximately unit variance; consequently, they are useful in looking for **outliers**. Most of the standardized residuals should lie in the interval $-3 \le d_i \le +3$, and any observation with a standardized residual outside this interval is potentially unusual with respect to its observed response. These outliers should be carefully examined because they may represent something as simple as a data-recording error or something of more serious concern, such as a region of the predictor or regressor variable space where the fitted model is a poor approximation to the true response.

Studentized Residuals The standardizing process in Eq. (3.47) scales the residuals by dividing them by their approximate average standard deviation. In some data sets, residuals may have standard deviations that differ greatly. We now present a scaling that takes this into account. The vector of fitted values $\hat{y}_i$ that corresponds to the observed values y_i is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{3.48}$$

The $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is usually called the "hat" matrix because it maps the vector of observed values into a vector of fitted values. The hat matrix and its properties play a central role in regression analysis.

The residuals from the fitted model may be conveniently written in matrix notation as

$$\mathbf{e} = \mathbf{y} - \hat{y} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{3.49}$$

and the covariance matrix of the residuals is

$$\mathrm{Cov}(\mathbf{e}) = V\,[(\mathbf{I} - \mathbf{H})\mathbf{y}] = \sigma^2(\mathbf{I} - \mathbf{H}).$$

The matrix **I** – **H** is in general not diagonal, so the residuals from a linear regression model have different variances and are correlated. The variance of the *i*th residual is

$$V(e_i) = \sigma^2(1 - h_{ii}), \tag{3.50}$$

where h_{ii} is the *i*th diagonal element of the hat matrix **H**. Because $0 \leq h_{ii} \leq 1$ using the mean squared error MS_{E} to estimate the variance of the residuals actually overestimates the true variance. Furthermore, it turns out that h_{ii} is a measure of the location of the *i*th point in the predictor variable or *x*-space; the variance of the residual e_i depends on where the point $\mathbf{x}_i$ lies. As h_{ii} increases, the observation $\mathbf{x}_i$ lies further from the center of the region containing the data. Therefore residuals near the center of the **x**-space have larger variance than do residuals at more remote locations. Violations of model assumptions are more likely at remote points, so these violations may be hard to detect from inspection of the ordinary residuals e_i (or the standardized residuals d_i) because their residuals will usually be smaller.

We recommend taking this inequality of variance into account when scaling the residuals. We suggest plotting the **studentized residuals as**:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \tag{3.51}$$

with $\hat{\sigma}^2 = MS_{\mathrm{E}}$ instead of the ordinary residuals or the standardized residuals. The studentized residuals have unit variance (i.e., $V(r_i) = 1$) regardless of the location of the observation $\mathbf{x}_i$ when the form of the regression model is correct. In many situations the variance of the residuals stabilizes, particularly for large data sets. In these cases, there may be little difference between the standardized and studentized residuals. Thus standardized and studentized residuals often convey equivalent information. However, because any point with a large residual and a large hat diagonal h_{ii} is potentially highly influential on the least squares fit, examination of the studentized residuals is generally recommended.

Table 3.7 displays the residuals, the studentized residuals, hat diagonals h_{ii}, and several other diagnostics for the regression model for the patient satisfaction data in Example 3.1. These quantities were computer generated using JMP. To illustrate the calculations, consider the first observation.

TABLE 3.7 Residuals and Other Diagnostics for the Regression Model for the Patient Satisfaction Data in Example 3.1

Observation	Residuals	Studentized Residuals	R-Student	h_{ii}	Cook's Distance
1	9.0378	1.29925	1.32107	0.044855	0.026424
2	−5.6986	−0.88216	−0.87754	0.176299	0.055521
3	9.0373	1.38135	1.41222	0.155114	0.116772
4	−0.6953	−0.10403	−0.10166	0.118125	0.000483
5	−7.3897	−1.08009	−1.08440	0.076032	0.031999
6	−3.2155	−0.47342	−0.46491	0.089420	0.007337
7	−5.0316	−0.77380	−0.76651	0.165396	0.039553
8	7.0616	1.03032	1.03183	0.072764	0.027768
9	−17.2800	−2.65767	−3.15124	0.165533	0.467041
10	−6.0864	−0.87524	−0.87041	0.045474	0.012165
11	−11.6967	−1.70227	−1.78483	0.068040	0.070519
12	3.4823	0.51635	0.50757	0.102232	0.010120
13	−0.0859	−0.01272	−0.01243	0.100896	0.000006
14	2.1786	0.33738	0.33048	0.176979	0.008159
15	4.5134	0.66928	0.66066	0.102355	0.017026
16	4.7705	0.68484	0.67634	0.042215	0.006891
17	2.2474	0.33223	0.32541	0.096782	0.003942
18	0.8699	0.13695	0.13386	0.203651	0.001599
19	0.9276	0.13769	0.13458	0.104056	0.000734
20	3.7691	0.58556	0.57661	0.182192	0.025462
21	10.4993	1.62405	1.69133	0.175015	0.186511
22	0.6797	0.10725	0.10481	0.207239	0.001002
23	−9.2785	−1.46893	−1.51118	0.212456	0.194033
24	3.0927	0.44996	0.44165	0.067497	0.004885
25	4.2911	0.61834	0.60945	0.049383	0.006621

The studentized residual is calculated as follows:

$$\begin{aligned} r_1 &= \frac{e_1}{\sqrt{\hat{\sigma}^2(1-h_{11})}} \\ &= \frac{e_1}{\hat{\sigma}\sqrt{(1-h_{11})}} \\ &= \frac{9.0378}{7.11767\sqrt{1-0.044855}} \\ &= 1.2992 \end{aligned}$$

which agrees approximately with the value reported by JMP in Table 3.7. Large values of the studentized residuals are usually an indication of potential unusual values or outliers in the data. Absolute values of the studentized residuals that are larger than three or four indicate potentially problematic observations. Note that none of the studentized residuals in Table 3.7 is this large. The largest studentized residual, −2.65767, is associated with observation 9. This observation does show up on the normal probability plot of residuals in Figure 3.1 as a very mild outlier, but there is no indication of a significant problem with this observation.

PRESS Another very useful residual scaling can be based on the prediction error sum of squares or PRESS. To calculate PRESS, we select an observation—for example, i. We fit the regression model to the remaining $n - 1$ observations and use this equation to predict the withheld observation y_i. Denoting this predicted value by $\hat{y}_{(i)}$, we may now find the prediction error for the ith observation as

$$e_{(i)} = y_i - \hat{y}_{(i)} \tag{3.52}$$

The prediction error is often called the ith PRESS residual. Note that the prediction error for the ith observation differs from the ith residual because observation i was not used in calculating the ith prediction value $\hat{y}_{(i)}$. This procedure is repeated for each observation $i = 1, 2, \dots, n$, producing a set of n PRESS residuals $e_{(1)}, e_{(2)}, \dots, e_{(n)}$. Then the PRESS statistic is defined as the sum of squares of the n PRESS residuals or

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 \tag{3.53}$$

Thus PRESS is a form of **data splitting** (discussed in Chapter 2), since it uses each possible subset of $n - 1$ observations as an estimation data set, and every observation in turn is used to form a prediction data set. Generally, small values of PRESS imply that the regression model will be useful in predicting new observations. To get an idea about how well the model will predict new data, we can calculate an R^2-like statistic called the R^2 for prediction

$$R^2_{\text{Prediction}} = 1 - \frac{\text{PRESS}}{SS_{\text{T}}} \tag{3.54}$$

Now PRESS will always be larger than the residual sum of squares and, because the ordinary $R^2 = 1 - (SS_E/SS_T)$, if the value of the $R^2_{\text{Prediction}}$ is not much smaller than the ordinary R^2, this is a good indication about potential model predictive performance.

It would initially seem that calculating PRESS requires fitting n different regressions. However, it is possible to calculate PRESS from the results of a single least squares fit to all n observations. It turns out that the ith PRESS residual is

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}, \tag{3.55}$$

where e_i is the OLS residual. The hat matrix diagonals are directly calculated as a routine part of solving the least squares normal equations. Therefore PRESS is easily calculated as

$$\text{PRESS} = \sum_{i=1}^{n} \frac{e_i^2}{1 - h_{ii}} \tag{3.56}$$

JMP will calculate the PRESS statistic for a regression model and the R^2 for prediction based on PRESS from Eq. (3.54). The value of PRESS is PRESS = 1484.93 and the R^2 for prediction is

$$\begin{aligned} R^2_{\text{Prediction}} &= 1 - \frac{\text{PRESS}}{SS_T} \\ &= 1 - \frac{1484.93}{10778.2} \\ &= 0.8622. \end{aligned}$$

That is, this model would be expected to account for about 86.22% of the variability in new data.

R-Student The studentized residual r_i discussed earlier is often considered an outlier diagnostic. It is customary to use the mean squared error MS_E as an estimate of σ^2 in computing r_i. This is referred to as **internal scaling** of the residual because MS_E is an internally generated estimate of σ^2 obtained from fitting the model to all n observations. Another approach would be to use an estimate of σ^2 based on a data set with the ith observation removed. We denote the estimate of σ^2 so obtained by $S^2_{(i)}$. We can show that

$$S^2_{(i)} = \frac{(n-p)MS_{\mathrm{E}} - e_i^2/(1-h_{ii})}{n-p-1} \tag{3.57}$$

The estimate of σ^2 in Eq. (3.57) is used instead of MS_{E} to produce an **externally studentized residual**, usually called **R-student**, given by

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} \tag{3.58}$$

In many situations, t_i will differ little from the studentized residual r_i. However, if the ith observation is influential, then $S^2_{(i)}$ can differ significantly from MS_{E}, and consequently the R-student residual will be more sensitive to this observation. Furthermore, under the standard assumptions, the R-student residual t_i has a t-distribution with $n - p - 1$ degrees of freedom. Thus R-student offers a more formal procedure for investigating potential outliers by comparing the absolute magnitude of the residual t_i to an appropriate percentage point of t_{n-p-1}.

JMP will compute the R-student residuals. They are shown in Table 3.7 for the regression model for the patient satisfaction data. The largest value of R-student is for observation 9, $t_9 = -3.15124$. This is another indication that observation 9 is a very mild outlier.

3.5.3 Measures of Leverage and Influence

In building regression models, we occasionally find that a small subset of the data exerts a disproportionate influence on the fitted model. That is, estimates of the model parameters or predictions may depend more on the influential subset than on the majority of the data. We would like to locate these influential points and assess their impact on the model. If these influential points really are "bad" values, they should be eliminated. On the other hand, there may be nothing wrong with these points, but if they control key model properties, we would like to know it because it could affect the use of the model. In this section we describe and illustrate some useful measures of influence.

The disposition of points in the predictor variable space is important in determining many properties of the regression model. In particular, remote observations potentially have disproportionate leverage on the parameter estimates, predicted values, and the usual summary statistics.

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is very useful in identifying influential observations. As noted earlier, $\mathbf{H}$ determines the variances and covariances of the predicted response and the residuals because

$$\text{Var}\,(\hat{\mathbf{y}}) = \sigma^2\mathbf{H} \quad \text{and} \quad \text{Var}\,(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

The elements h_{ij} of the hat matrix $\mathbf{H}$ may be interpreted as the amount of leverage exerted by the observation y_j on the predicted value $\hat{y}_i$. Thus inspection of the elements of $\mathbf{H}$ can reveal points that are potentially influential by virtue of their location in x-space.

Attention is usually focused on the diagonal elements of the hat matrix h_{ii}. It can be shown that $\sum_{i=1}^{n} h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p$, so the average size of the diagonal elements of the $\mathbf{H}$ matrix is p/n. A widely used rough guideline is to compare the diagonal elements $h_{ii}$to twice their average value $2p/n$, and if any hat diagonal exceeds this value to consider that observation as a high-leverage point.

JMP will calculate and save the values of the hat diagonals. Table 3.7 displays the hat diagonals for the regression model for the patient satisfaction data in Example 3.1. Since there are $p = 3$ parameters in the model and $n = 25$ observations, twice the average size of a hat diagonal for this problem is

$$2p/n = 2(3)/25 = 0.24.$$

The largest hat diagonal, 0.212456, is associated with observation 23. This does not exceed twice the average size of a hat diagonal, so there are no high-leverage observations in these data.

The hat diagonals will identify points that are potentially influential due to their location in x-space. It is desirable to consider both the location of the point and the response variable in measuring influence. Cook (1977, 1979) has suggested using a measure of the squared distance between the least squares estimate based on all n points $\hat{\boldsymbol{\beta}}$ and the estimate obtained by deleting the ith point, say, $\hat{\beta}_{(i)}$. This distance measure can be expressed as

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{pMS_{\text{E}}}, \qquad i = 1, 2, \ldots, n \tag{3.59}$$

A reasonable cutoff for D_i is unity. That is, we usually consider observations for which $D_i > 1$ to be influential. Cook's distance statistic D_i is actually calculated from

$$D_i = \frac{r_i^2}{p} \frac{V[\hat{y}(\mathbf{x}_i)]}{V(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}} \tag{3.60}$$

Note that, apart from the constant p, D_i is the product of the square of the ith studentized residual and the ratio $h_{ii}/(1 - h_{ii})$. This ratio can be shown to be the distance from the vector $\mathbf{x}_i$ to the centroid of the remaining data. Thus D_i is made up of a component that reflects how well the regression model fits the ith observation y_i and a component that measures how far that point is from the rest of the data. Either component (or both) may contribute to a large value of D_i.

JMP will calculate and save the values of Cook's distance statistic D_i. Table 3.7 displays the values of Cook's distance statistic for the regression model for the patient satisfaction data in Example 3.1. The largest value, 0.467041, is associated with observation 9. This value was calculated from Eq. (3.60) as follows:

$$\begin{aligned} D_i &= \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}} \\ &= \frac{(-2.65767)^2}{3} \frac{0.165533}{1 - 0.165533} \\ &= 0.467041. \end{aligned}$$

This does not exceed twice the cutoff of unity, so there are no influential observations in these data.

3.6 VARIABLE SELECTION METHODS IN REGRESSION

In our treatment of regression we have concentrated on fitting the full regression model. Actually, in most applications of regression the analyst will have a very good idea about the general form of the model he/she wishes to fit, but there may be uncertainty about the exact structure of the model. For example, we may not know if all of the predictor variables are really necessary. These applications of regression frequently involve a moderately large or large set of **candidate predictors**, and the objective of the analyst here is to fit a regression model to the “best subset” of these candidates. This can be a complex problem, as these data sets frequently have outliers, strong correlations between subsets of the variables, and other complicating features.

There are several techniques that have been developed for selecting the best subset regression model. Generally, these methods are either **stepwise-type** variable selection methods or **all possible regressions**. Stepwise-type methods build a regression model by either adding or removing a predictor variable to the basic model at each step. The forward selection version of the procedure begins with a model containing none of the candidate predictor variables and sequentially inserts variables into the model one-at-a-time until a final equation is produced. The criterion for entering a variable into the equation is that the t-statistic for that variable must be significant. The process is continued until there are no remaining candidate predictors that qualify for entry into the equation. In backward elimination, the procedure begins with all of the candidate predictor variables in the equation, and then variables are removed one-at-a-time to produce a final equation. The criterion for removing a variable is usually based on the t-statistic, with the variable having the smallest t-statistic considered for removal first. Variables are removed until all of the predictors remaining in the model have significant t-statistics. Stepwise regression usually consists of a combination of forward and backward stepping. There are many variations of the basic procedures.

In all possible regressions with K candidate predictor variables, the analyst examines all 2^K possible regression equations to identify the ones with potential to be a useful model. Obviously, as K becomes even moderately large, the number of possible regression models quickly becomes formidably large. Efficient algorithms have been developed that implicitly rather than explicitly examine all of these equations. Typically, only the equations that are found to be “best” according to some criterion (such as minimum MS_{E} or AICc) at each subset size are displayed. For more discussion of variable selection methods, see textbooks on regression such as Montgomery, Peck, and Vining (2012) or Myers (1990).

Example 3.8 Table 3.8 contains an expanded set of data for the hospital patient satisfaction data introduced in Example 3.1. In addition to the patient age and illness severity data, there are two additional regressors, an indicator of whether the patent is a surgical patient (1) or a medical patient (0), and an index indicating the patient’s anxiety level. We will use these data to illustrate how variable selection methods in regression can be used to help the analyst build a regression model.

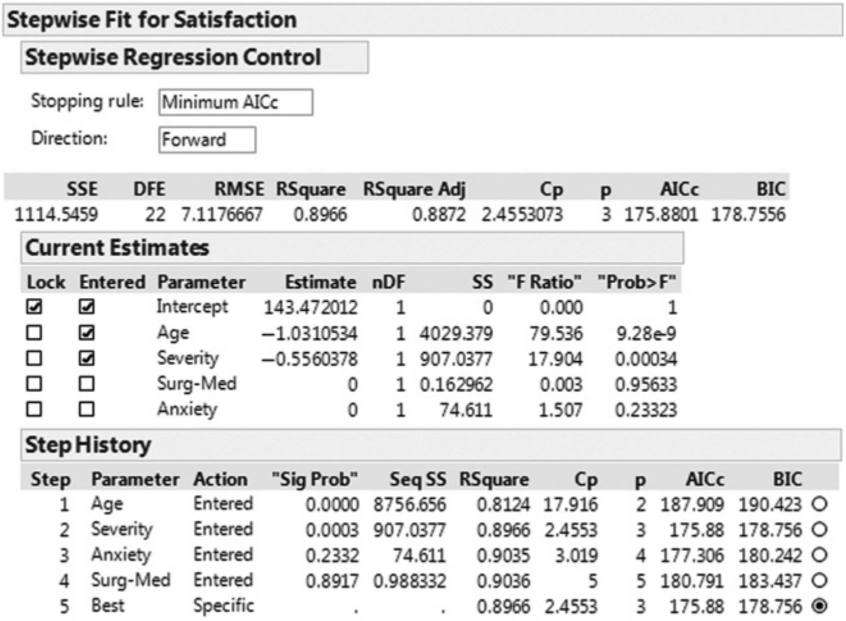
TABLE 3.8 Expanded Patient Satisfaction Data

Observation	Age	Severity	Surgical–Medical	Anxiety	Satisfaction
1	55	50	0	2.1	68
2	46	24	1	2.8	77
3	30	46	1	3.3	96
4	35	48	1	4.5	80
5	59	58	0	2.0	43
6	61	60	0	5.1	44
7	74	65	1	5.5	26
8	38	42	1	3.2	88
9	27	42	0	3.1	75
10	51	50	1	2.4	57
11	53	38	1	2.2	56
12	41	30	0	2.1	88
13	37	31	0	1.9	88
14	24	34	0	3.1	102
15	42	30	0	3.0	88
16	50	48	1	4.2	70
17	58	61	1	4.6	52
18	60	71	1	5.3	43
19	62	62	0	7.2	46
20	68	38	0	7.8	56
21	70	41	1	7.0	59
22	79	66	1	6.2	26
23	63	31	1	4.1	52
24	39	42	0	3.5	83
25	49	40	1	2.1	75

We will illustrate the forward selection procedure first. The JMP output that results from applying forward selection to these data is shown in Table 3.9. We used the AICc criterion for selecting the best model. The forward selection algorithm inserted the predictor patient age first, then severity, then anxiety, and finally surg-med was inserted into the equation. The best model based on the minimum value of AICc contained age and severity.

Table 3.10 presents the results of applying the JMP backward elimination procedure to the patient satisfaction data. Once again the AICc criterion was chosen to select the final model. The procedure begins with all four predictors in the model, then the surgical–medical indicator variable was removed, followed by the anxiety predictor, followed by severity.

TABLE 3.9 JMP Forward Selection for the Patient Satisfaction Data in Table 3.8



Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping rule: Minimum AICc

Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1114.5459	22	7.1176667	0.8966	0.8872	2.4553073	3	175.8801	178.7556

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	143.472012	1	0	0.000	1
□	☑	Age	−1.0310534	1	4029.379	79.536	9.28e-9
□	☑	Severity	−0.5560378	1	907.0377	17.904	0.00034
□	□	Surg-Med	0	1	0.162962	0.003	0.95633
□	□	Anxiety	0	1	74.611	1.507	0.23323

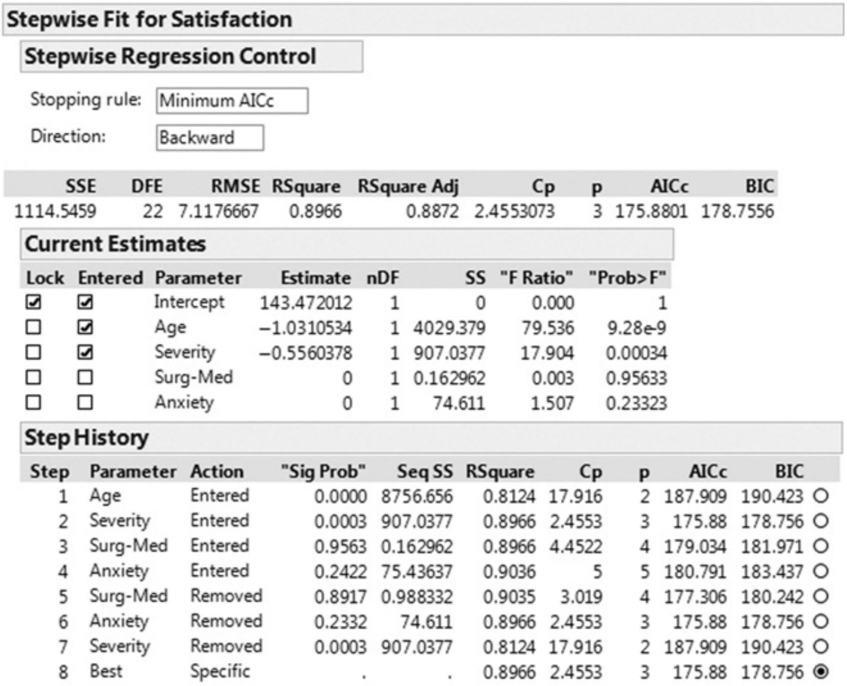
Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Age	Entered	0.0000	8756.656	0.8124	17.916	2	187.909	190.423	○
2	Severity	Entered	0.0003	907.0377	0.8966	2.4553	3	175.88	178.756	○
3	Anxiety	Entered	0.2332	74.611	0.9035	3.019	4	177.306	180.242	○
4	Surg-Med	Entered	0.8917	0.988332	0.9036	5	5	180.791	183.437	○
5	Best	Specific	.	.	0.8966	2.4553	3	175.88	178.756	●

However, removing severity causes an increase in AICc so it is added back to the model. The algorithm concluded with both patient age and severity in the model. Note that in this example, the forward selection procedure produced the same model as the backward elimination procedure. This does not always happen, so it is usually a good idea to investigate different model-building techniques for a problem.

Table 3.11 is the JMP stepwise regression algorithm applied to the patient satisfaction data, JMP calls the stepwise option “mixed” variable selection. The default significance levels of 0.25 to enter or remove variables from the model were used. At the first step, patient age is entered in the model. Then severity is entered as the second variable. This is followed by anxiety as the third variable. At that point, none of the remaining predictors met the 0.25 significance level criterion to enter the model, so stepwise regression terminated with age, severity and anxiety as the model predictors. This is not the same model found by backwards elimination and forward selection.

TABLE 3.10 JMP Backward Elimination for the Patient Satisfaction Data in Table 3.8



Stepwise Fit for Satisfaction

Stepwise Regression Control

Stopping rule: Minimum AICc

Direction: Backward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1114.5459	22	7.1176667	0.8966	0.8872	2.4553073	3	175.8801	178.7556

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	143.472012	1	0	0.000	1
☐	☑	Age	−1.0310534	1	4029.379	79.536	9.28e-9
☐	☑	Severity	−0.5560378	1	907.0377	17.904	0.00034
☐	☐	Surg-Med	0	1	0.162962	0.003	0.95633
☐	☐	Anxiety	0	1	74.611	1.507	0.23323

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Age	Entered	0.0000	8756.656	0.8124	17.916	2	187.909	190.423	○
2	Severity	Entered	0.0003	907.0377	0.8966	2.4553	3	175.88	178.756	○
3	Surg-Med	Entered	0.9563	0.162962	0.8966	4.4522	4	179.034	181.971	○
4	Anxiety	Entered	0.2422	75.43637	0.9036	5	5	180.791	183.437	○
5	Surg-Med	Removed	0.8917	0.988332	0.9035	3.019	4	177.306	180.242	○
6	Anxiety	Removed	0.2332	74.611	0.8966	2.4553	3	175.88	178.756	○
7	Severity	Removed	0.0003	907.0377	0.8124	17.916	2	187.909	190.423	○
8	Best	Specific	.	.	0.8966	2.4553	3	175.88	178.756	●

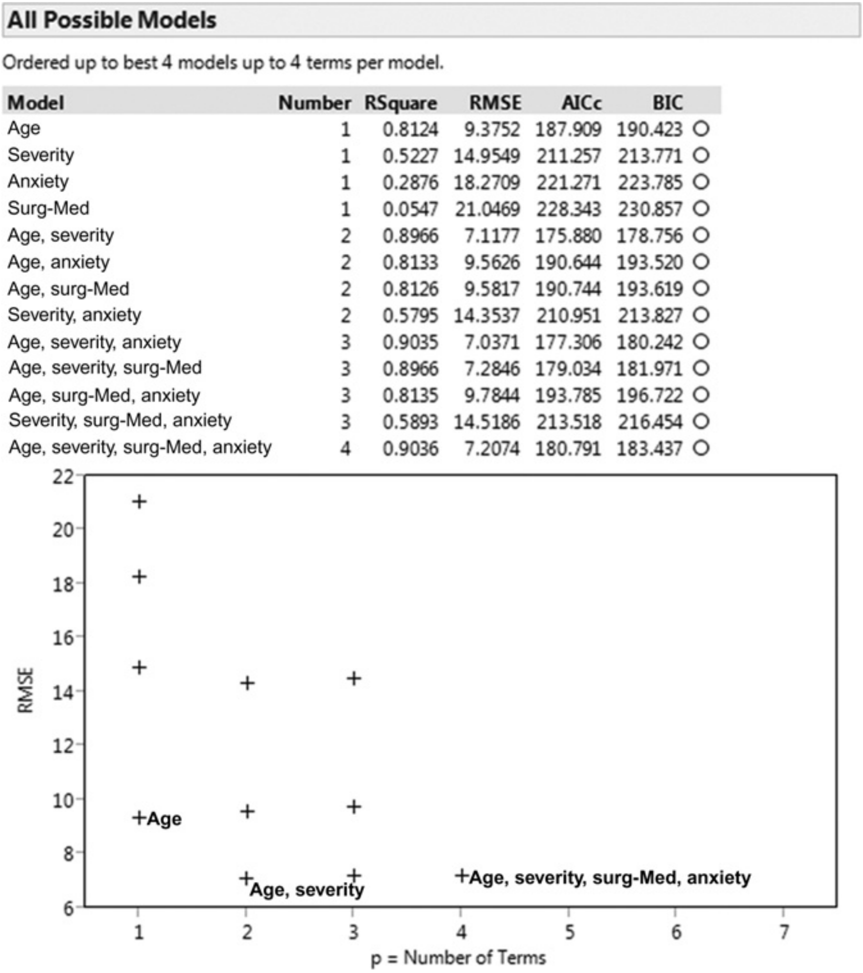
Table 3.12 shows the results of applying the JMP all possible regressions algorithm to the patient satisfaction data. Since there are $k = 4$ predictors, there are 16 possible regression equations. JMP shows the best four of each subset size, along with the full (four-variable) model. For each model, JMP presents the value of R^2, the square root of the mean squared error (RMSE), and the AICc and BIC statistics.

The model with the smallest value of AICc and BIC is the two-variable model with age and severity. The model with the smallest value of the mean squared error (or its square root, RMSE) is the three-variable model with age, severity, and anxiety. Both of these models were found using the stepwise-type algorithms. Either one of these models is likely to be a good regression model describing the effects of the predictor variables on patient satisfaction.

TABLE 3.11 JMP Stepwise (Mixed) Variable Selection for the Patient Satisfaction Data in Table 3.8

```
Stepwise Fit for Satisfaction
Stepwise Regression Control

Stopping rule:    P-Value Threshold

                  Prob to Enter 0.25

                  Prob to Leave 0.25
Direction: Mixed

SSE       DFE     RMSE       RSquare   RSquare Adj    Cp          p       AICc      BIC
1039.935   21     7.0370954   0.9035     0.8897     3.0190257     4     177.3058 180.2422

Current Estimates
Lock      Entered   Parameter  Estimate       nDF       SS        "F Ratio"  "Prob>F"
[X]         [X]     Intercept 143.895206       1         0          0.000        1
[ ]         [X]     Age       -1.1135376       1      3492.683     70.530    3.75e-8
[ ]         [X]     Severity  -0.5849193       1      971.8361     19.625    0.00023
[ ]         [ ]     Surg-Med       0           1      0.988332     0.019     0.89167
[ ]         [X]     Anxiety    1.2961695       1       74.611      1.507     0.23323

Step History
Step     Parameter Action     "Sig Prob"    Seq SS      RSquare      Cp          p       AICc      BIC
1        Age       Entered      0.0000     8756.656     0.8124     17.916        2      187.909  190.423  ()
2        Severity  Entered      0.0003     907.0377     0.8966     2.4553        3       175.88  178.756  ()
3        Anxiety   Entered      0.2332      74.611      0.9035     3.019         4      177.306  180.242 (X)
```

TABLE 3.12 JMP All Possible Models Regression for the Patient Satisfaction Data in Table 3.8



All Possible Models

Ordered up to best 4 models up to 4 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC	
Age	1	0.8124	9.3752	187.909	190.423	○
Severity	1	0.5227	14.9549	211.257	213.771	○
Anxiety	1	0.2876	18.2709	221.271	223.785	○
Surg-Med	1	0.0547	21.0469	228.343	230.857	○
Age, severity	2	0.8966	7.1177	175.880	178.756	○
Age, anxiety	2	0.8133	9.5626	190.644	193.520	○
Age, surg-Med	2	0.8126	9.5817	190.744	193.619	○
Severity, anxiety	2	0.5795	14.3537	210.951	213.827	○
Age, severity, anxiety	3	0.9035	7.0371	177.306	180.242	○
Age, severity, surg-Med	3	0.8966	7.2846	179.034	181.971	○
Age, surg-Med, anxiety	3	0.8135	9.7844	193.785	196.722	○
Severity, surg-Med, anxiety	3	0.5893	14.5186	213.518	216.454	○
Age, severity, surg-Med, anxiety	4	0.9036	7.2074	180.791	183.437	○

3.7 GENERALIZED AND WEIGHTED LEAST SQUARES

In Section 3.4 we discussed methods for checking the adequacy of a linear regression model. Analysis of the model residuals is the basic methodology. A common defect that shows up in fitting regression models is nonconstant variance. That is, the variance of the observations is not constant but changes in some systematic way with each observation. This problem is often identified from a plot of residuals versus the fitted values. Transformation of the response variable is a widely used method for handling the inequality of variance problem.

Another technique for dealing with nonconstant error variance is to fit the model using the method of **weighted least squares** (WLS). In this method of estimation the deviation between the observed and expected values of y_i is multiplied by a **weight** w_i that is inversely proportional to the variance of y_i. For the case of simple linear regression, the WLS function is

$$L = \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_i)^2, \tag{3.61}$$

where $w_i = 1/\sigma_i^2$ and σ_i^2 is the variance of the ith observation y_i. The resulting least squares normal equations are

$$\begin{aligned}\hat{\beta}_0 \sum_{i=1}^{n} w_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i &= \sum_{i=1}^{n} w_i y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} w_i x_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i^2 &= \sum_{i=1}^{n} w_i x_i y_i\end{aligned} \tag{3.62}$$

Solving Eq. (3.62) will produce WLS estimates of the model parameters β_0 and β_1.

In this section we give a development of WLS for the multiple regression model. We begin by considering a slightly more general situation concerning the structure of the model errors.

3.7.1 Generalized Least Squares

The assumptions that we have made concerning the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and Var $(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$; that is, the errors have expected value zero and constant variance, and they are uncorrelated. For testing hypotheses and constructing confidence and prediction intervals we also assume that the errors are normally distributed, in which case they are also independent. As we have observed, there are situations where these assumptions are unreasonable. We will now consider the modifications that are necessary to the OLS procedure when $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and Var $(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a known $n \times n$ matrix. This situation has a simple interpretation; if $\mathbf{V}$ is diagonal but with unequal diagonal elements, then the observations

y are **uncorrelated** but have **unequal variances**, while if some of the off-diagonal elements of **V** are nonzero, then the observations are **correlated**. When the model is

$$\begin{gathered}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}\end{gathered} \tag{3.63}$$

the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is no longer appropriate. The OLS estimator is unbiased because

$$E(\hat{\boldsymbol{\beta}}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\beta} = \boldsymbol{\beta}$$

but the covariance matrix of $\hat{\boldsymbol{\beta}}$ is not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Instead, the covariance matrix is

$$\begin{aligned}\mathrm{Var}\,(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\,[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Practically, this implies that the variances of the regression coefficients are larger than we expect them to be.

This problem can be avoided if we estimate the model parameters with a technique that takes the correct variance structure in the errors into account. We will develop this technique by transforming the model to a new set of observations that satisfy the standard least squares assumptions. Then we will use OLS on the transformed observations.

Because $\sigma^2\mathbf{V}$ is the covariance matrix of the errors, **V** must be nonsingular and positive definite, so there exists an $n \times n$ nonsingular symmetric matrix **K** defined such that

$$\mathbf{K}'\mathbf{K} = \mathbf{K}\mathbf{K} = \mathbf{V}$$

The matrix **K** is often called the **square root** of **V**. Typically, the error variance σ^2 is unknown, in which case **V** represents the known (or assumed) structure of the variances and covariances among the random errors apart from the constant σ^2.

Define the new variables

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \quad \mathbf{B} = \mathbf{K}^{-1}\mathbf{X}, \quad \text{and} \quad \boldsymbol{\delta} = \mathbf{K}^{-1}\boldsymbol{\varepsilon} \tag{3.64}$$

so that the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ becomes, upon multiplication by $\mathbf{K}^{-1}$,

$$\mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}^{-1}\boldsymbol{\varepsilon}$$

or

$$\mathbf{z} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\delta} \tag{3.65}$$

The errors in the transformed model Eq. (3.65) have zero expectation because $E(\boldsymbol{\delta}) = E(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Furthermore, the covariance matrix of $\boldsymbol{\delta}$ is

$$\begin{aligned}
\mathrm{Var}\ (\boldsymbol{\delta}) &= V(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) \\
&= \mathbf{K}^{-1}V(\boldsymbol{\varepsilon})\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}\mathbf{K}^{-1} \\
&= \sigma^2\mathbf{I}
\end{aligned}$$

Thus the elements of the vector of errors $\boldsymbol{\delta}$ have mean zero and constant variance and are uncorrelated. Since the errors $\boldsymbol{\delta}$ in the model in Eq. (3.65) satisfy the usual assumptions, we may use OLS to estimate the parameters. The least squares function is

$$\begin{aligned}
L &= \boldsymbol{\delta}'\boldsymbol{\delta} \\
&= (\mathbf{K}^{-1}\boldsymbol{\varepsilon})'\mathbf{K}^{-1}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}'\mathrm{K}^{-1}\mathbf{K}^{-1}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}'\mathrm{V}^{-1}\boldsymbol{\varepsilon} \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

The corresponding normal equations are

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{3.66}$$

In Equation (3.66) $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is the **generalized least squares (GLS) estimator** of the model parameters $\boldsymbol{\beta}$. The solution to the GLS normal equations is

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{3.67}$$

The GLS estimator is an unbiased estimator for the model parameters $\boldsymbol{\beta}$, and the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is

$$\text{Var}\,(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{3.68}$$

The GLS estimator is a best linear unbiased estimator of the model parameters $\boldsymbol{\beta}$, where "best" means minimum variance.

3.7.2 Weighted Least Squares

Weighted least squares or WLS is a special case of GLS where the n response observations y_i do not have the same variances but are uncorrelated. Therefore the matrix $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix},$$

where σ_i^2 is the variance of the ith observation y_i, $i = 1, 2, \ldots, n$. Because the weight for each observation should be the reciprocal of the variance of that observation, it is convenient to define a diagonal matrix of weights $\mathbf{W} = \mathbf{V}^{-1}$. Clearly, the weights are the main diagonals of the matrix $\mathbf{W}$. Therefore the WLS criterion is

$$L = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.69}$$

and the WLS normal equations are

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\hat{\boldsymbol{\beta}}_{\text{WLS}} = \mathbf{X}'\mathbf{W}\mathbf{y}. \tag{3.70}$$

The WLS estimator is

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{3.71}$$

The WLS estimator is an unbiased estimator for the model parameters $\boldsymbol{\beta}$, and the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is

$$\text{Var}\,(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}. \tag{3.72}$$

To use WLS, the weights w_i must be known. Sometimes prior knowledge or experience or information from an underlying theoretical model can be used to determine the weights. For example, suppose that a significant source of error is measurement error and different observations are measured by different instruments of unequal but known or well-estimated accuracy. Then the weights could be chosen inversely proportional to the variances of measurement error.

In most practical situations, however, the analyst learns about the inequality of variance problem from the residual analysis for the original model that was fit using OLS. For example, the plot of the OLS residuals e_i versus the fitted values $\hat{y}_i$ may exhibit an outward-opening funnel shape, suggesting that the variance of the observations is increasing with the mean of the response variable y. Plots of the OLS residuals versus the predictor variables may indicate that the variance of the observations is a function of one of the predictors. In these situations we can often use estimates of the weights. There are several approaches that could be used to estimate the weights. We describe two of the most widely used methods.

Estimation of a Variance Equation In the first method, suppose that analysis of the OLS residuals indicates that the variance of the ith observation is a function of one or more predictors or the mean of y. The squared OLS residual e_i^2 is an estimator of the variance of the ith observation σ_i^2 if the form of the regression model is correct. Furthermore, the absolute value of the residual $|e_i|$ is an estimator of the standard deviation σ_i (because $\sigma_i = |\sqrt{\sigma_i^2}|$). Consequently, we can find a **variance equation** or a regression model relating σ_i^2 to appropriate predictor variables by the following process:

1. Fit the model relating y to the predictor variables using OLS and find the OLS residuals.
2. Use residual analysis to determine potential relationships between σ_i^2 and either the mean of y or some of the predictor variables.
3. Regress the squared OLS residuals on the appropriate predictors to obtain an equation for predicting the variance of each observation, say, $\hat{s}_i^2 = f(x)$ or $\hat{s}_i^2 = f(y)$.
4. Use the fitted values from the estimated variance function to obtain estimates of the weights, $w_i = 1/\hat{s}_i^2, i = 1, 2, \ldots, n$.
5. Use the estimated weights as the diagonal elements of the matrix $\mathbf{W}$ in the WLS procedure.

As an alternative to estimating a variance equation in step 3 above, we could use the absolute value of the OLS residual and fit an equation that relates the standard deviation of each observation to the appropriate regressors. This is the preferred approach if there are potential outliers in the data, because the absolute value of the residuals is less affected by outliers than the squared residuals.

When using the five-step procedure outlined above, it is a good idea to compare the estimates of the model parameters obtained from the WLS fit to those obtained from the original OLS fit. Because both methods produce unbiased estimators, we would expect to find that the point estimates of the parameters from both analyses are very similar. If the WLS estimates differ significantly from their OLS counterparts, it is usually a good idea to use the new WLS residuals and reestimate the variance equation to produce a new set of weights and a revised set of WLS estimates using these new weights. This procedure is called **iteratively reweighted least squares** (IRLS). Usually one or two iterations are all that is required to produce stable estimates of the model parameters.

Using Replicates or Nearest Neighbors The second approach to estimating the weights makes use of **replicate observations** or **nearest neighbors**. Exact replicates are sample observations that have exactly the same values of the predictor variables. Suppose that there are replicate observations at each of the combination of levels of the predictor variables. The weights w_i can be estimated directly as the reciprocal of the sample variances at each combination of these levels. Each observation in a replicate group would receive the same weight. This method works best when there are a moderately large number of observations in each replicate group, because small samples do not produce reliable estimates of the variance.

Unfortunately, it is fairly unusual to find groups of replicate observations in most regression-modeling situations. It is especially unusual to find them in time series data. An alternative is to look for observations with **similar** x-levels, which can be thought of as a nearest-neighbor group of observations. The observations in a nearest-neighbor group can be considered as pseudoreplicates and the sample variance for all of the observations in each nearest-neighbor group can be computed. The reciprocal of a sample variance would be used as the weight for all observations in the nearest-neighbor group.

Sometimes these nearest-neighbor groups can be identified visually by inspecting the scatter plots of y versus the predictor variables or from plots of the predictor variables versus each other. Analytical methods can also be used to find these nearest-neighbor groups. One nearest-neighbor algorithm is described in Montgomery, Peck, and Vining (2012). These authors also present a complete example showing how the nearest-neighbor approach can be used to estimate the weights for a WLS analysis.

Statistical Inference in WLS In WLS the variances σ_i^2 are almost always unknown and must be estimated. Since statistical inference on the model parameters as well as confidence intervals and prediction intervals on the response are usually necessary, we should consider the effect of using estimated weights on these procedures. Recall that the covariance matrix of the model parameters in WLS was given in Eq. (3.72). This covariance matrix plays a central role in statistical inference. Obviously, when estimates of the weights are substituted into Eq. (3.72) an estimated covariance matrix is obtained. Generally, the impact of using estimated weights is modest, provided that the sample size is not very small. In these situations, statistical tests, confidence intervals, and prediction intervals should be considered as approximate rather than exact.

Example 3.9 Table 3.13 contains 28 observations on the strength of a connector and the age in weeks of the glue used to bond the components of the connector together. A scatter plot of the strength versus age, shown in Figure 3.2, suggests that there may be a linear relationship between strength and age, but there may also be a problem with nonconstant variance in the data. The regression model that was fit to these data is

$$\hat{y} = 25.936 + 0.3759x,$$

where $x =$ weeks.

The residuals from this model are shown in Table 3.13. Figure 3.3 is a plot of the residuals versus weeks. The pronounced outward-opening funnel shape on this plot confirms the inequality of variance problem. Figure 3.4 is a plot of the absolute value of the residuals from this model versus week. There is an indication that a linear relationship may exist between the absolute value of the residuals and weeks, although there is evidence of one outlier in the data. Therefore it seems reasonable to fit a model relating the absolute value of the residuals to weeks. Since the absolute value of a residual is the residual standard deviation, the predicted values from this equation could be used to determine weights for the regression model relating strength to weeks. This regression model is

$$\hat{s}_i = -5.854 + 0.29852x.$$

TABLE 3.13 Connector Strength Data

Observation	Weeks	Strength	Residual	Absolute Residual	Weights
1	20	34	0.5454	0.5454	73.9274
2	21	35	1.1695	1.1695	5.8114
3	23	33	−1.5824	1.5824	0.9767
4	24	36	1.0417	1.0417	0.5824
5	25	35	−0.3342	0.3342	0.3863
6	28	34	−2.4620	2.4620	0.1594
7	29	37	0.1621	0.1621	0.1273
8	30	34	−3.2139	3.2139	0.1040
9	32	42	4.0343	4.0343	0.0731
10	33	35	−3.3416	3.3416	0.0626
11	35	33	−6.0935	6.0935	0.0474
12	37	46	6.1546	6.1546	0.0371
13	38	43	2.7787	2.7787	0.0332
14	40	32	−8.9731	8.9731	0.0270
15	41	37	−4.3491	4.3491	0.0245
16	43	50	7.8991	7.8991	0.0205
17	44	34	−8.4769	8.4769	0.0189
18	45	54	11.1472	11.1472	0.0174
19	46	49	5.7713	5.7713	0.0161
20	48	55	11.0194	11.0194	0.0139
21	50	40	−4.7324	4.7324	0.0122
22	51	33	−12.1084	12.1084	0.0114
23	52	56	10.5157	10.5157	0.0107
24	55	58	11.3879	11.3879	0.0090
25	56	45	−1.9880	1.9880	0.0085
26	57	33	−14.3639	14.3639	0.0080
27	59	60	11.8842	11.8842	0.0072
28	60	35	−13.4917	13.4917	0.0069

The weights would be equal to the inverse of the square of the fitted value for each s_i. These weights are shown in Table 3.13. Using these weights to fit a new regression model to strength using WLS results in

$$\hat{y} = 27.545 + 0.32383x$$

Note that the weighted least squares model does not differ very much from the OLS model. Because the parameter estimates did not change very much, this is an indication that it is not necessary to iteratively reestimate the standard deviation model and obtain new weights.

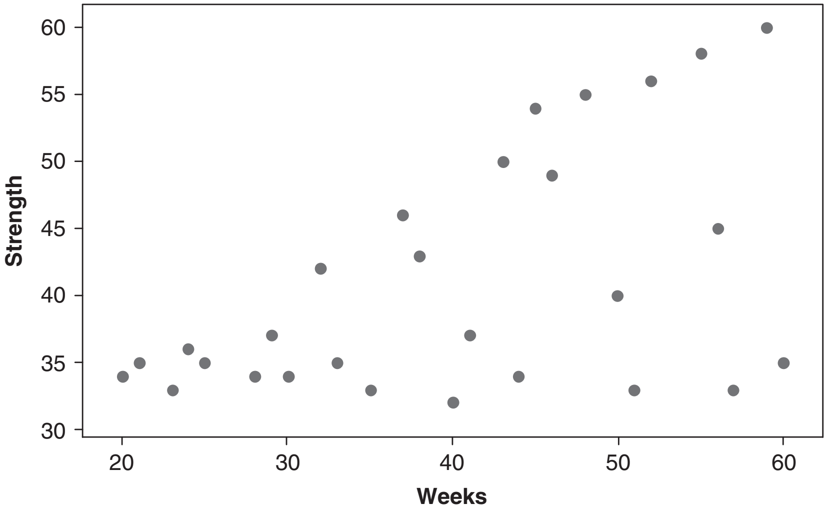


FIGURE 3.2 Scatter diagram of connector strength versus age from Table 3.12.

3.7.3 Discounted Least Squares

Weighted least squares is typically used in situations where the variance of the observations is not constant. We now consider a different situation where a WLS-type procedure is also appropriate. Suppose that the predictor variables in the regression model are only functions of time. As

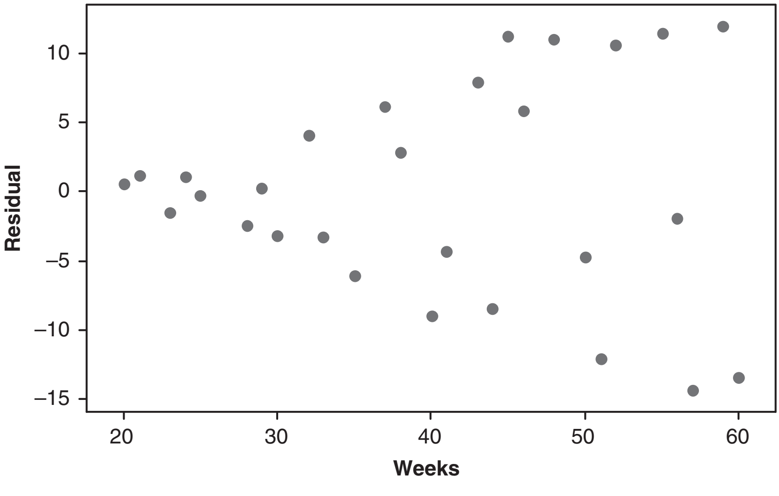


FIGURE 3.3 Plot of residuals versus weeks.

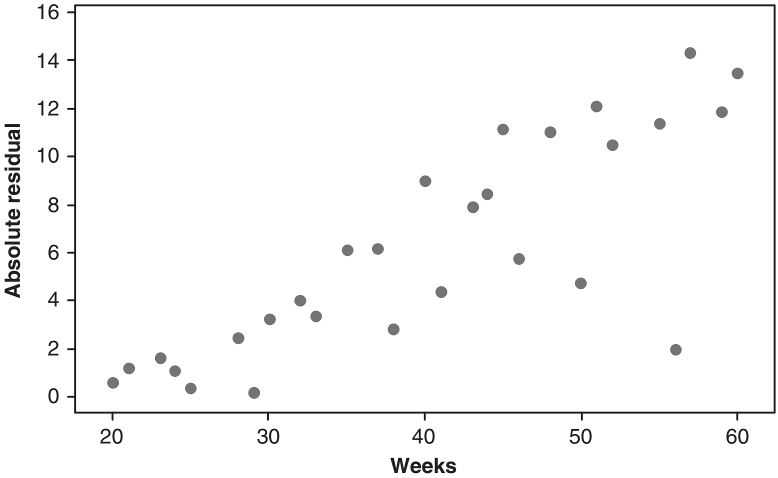


FIGURE 3.4 Scatter plot of absolute residuals versus weeks.

an illustration, consider the linear regression model with a linear trend in time:

$$y_t = \beta_0 + \beta_1 t + \varepsilon, \quad t = 1, 2, \ldots, T \tag{3.73}$$

This model was introduced to illustrate trend adjustment in a time series in Section 2.4.2 and Example 3.2. As another example, the regression model

$$y_t = \beta_0 + \beta_1 \sin \frac{2\pi}{d} t + \beta_2 \cos \frac{2\pi}{d} t + \varepsilon \tag{3.74}$$

describes the relationship between a response variable y that varies cyclically or periodically with time where the cyclic variation is modeled as a simple sine wave. A very general model for these types of situations could be written as

$$y_t = \beta_0 + \beta_1 x_1(t) + \cdots + \beta_k x_k(t) + \varepsilon_t, \quad t = 1, 2, \ldots, T, \tag{3.75}$$

where the predictors $x_1(t), x_2(t), \ldots, x_k(t)$ are mathematical functions of time, t. In these types of models it is often logical to believe that older observations are of less value in predicting the future observations at periods $T + 1, T + 2, \ldots,$ than are the observations that are close to the current time period, T. In other words, if you want to predict the value of y at time

$T + 1$ given that you are at the end of time period T (or $\hat{y}_{T+1}(T)$), it is logical to assume that the more recent observations such as y_T, y_{T-1}, and y_{T-2} carry much more useful information than do older observations such as y_{T-20}. Therefore it seems reasonable to weight the observations in the regression model so that recent observations are weighted more heavily than older observations. A very useful variation of WLS, called **discounted least squares**, can be used to do this. Discounted least squares also lead to a relatively simple way to update the estimates of the model parameters after each new observation in the time series.

Suppose that the model for observation y_t is given by Eq. (3.75):

$$\begin{aligned} y_t &= \beta_1 x_1(t) + \cdots + \beta_p x_p(t) + \varepsilon_t \\ &= \mathbf{x}(t)'\boldsymbol{\beta}, \qquad t = 1, 2, \ldots, T, \end{aligned}$$

where $\mathbf{x}(t)' = [x_1(t), x_2(t), \ldots, x_p(t)]$ and $\boldsymbol{\beta}' = [\beta_1, \beta_2, \ldots, \beta_p]$. This model could have an intercept term, in which case $x_1(t) = 1$ and the final model term could be written as $\beta_k x_k(t)$ as in Eq. (3.75). In matrix form, Eq. (3.75) is

$$\mathbf{y} = \mathbf{X}(T)\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.76}$$

where $\mathbf{y}$ is a $T \times 1$ vector of the observations, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the model parameters, $\boldsymbol{\varepsilon}$ is a $T \times 1$ vector of the errors, and $\mathbf{X}(T)$ is the $T \times p$ matrix

$$\mathbf{X}(T) = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_p(1) \\ x_1(2) & x_2(2) & \cdots & x_p(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(T) & x_2(T) & \cdots & x_p(T) \end{bmatrix}$$

Note that the tth row of $\mathbf{X}(T)$ contains the values of the predictor variables that correspond to the tth observation of the response, y_t.

We will estimate the parameters in Eq. (3.76) using WLS. However, we are going to choose the weights so that they decrease in magnitude with time. Specifically, let the weight for observation y_{T-j} be θ^j, where $0 < \theta < 1$. We are also going to shift the origin of time with each new observation so that T is the current time period. Therefore the WLS criterion is

$$\begin{aligned} L &= \sum_{j=0}^{T-1} w_j \left[y_{T-j} - (\beta_1(T)x_1(-j) + \cdots + \beta_p(T)x_k(-j))\right]^2 \\ &= \sum_{j=0}^{T-1} w_j \left[y_{T-j} - \mathbf{x}(-j)\boldsymbol{\beta}(T)\right]^2, \end{aligned} \tag{3.77}$$

where $\boldsymbol{\beta}(T)$ indicates that the vector of regression coefficients is estimated at the end of time period T, and $\mathbf{x}(-j)$ indicates that the predictor variables, which are just mathematical functions of time, are evaluated at $-j$. This is just WLS with a $T \times T$ diagonal weight matrix

$$\mathbf{W} = \begin{bmatrix} \theta^{T-1} & 0 & 0 & \cdots & 0 \\ 0 & \theta^{T-2} & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \cdots & & \theta & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

By analogy with Eq. (3.70), the WLS normal equations are

$$\mathbf{X}(T)'\mathbf{W}\mathbf{X}(T)\hat{\boldsymbol{\beta}}(T) = \mathbf{X}(T)'\mathbf{W}\mathbf{y}$$

or

$$\mathbf{G}(T)\hat{\boldsymbol{\beta}}(T) = \mathbf{g}(T), \tag{3.78}$$

where

$$\begin{aligned} \mathbf{G}(\mathrm{T}) &= \mathbf{X}(T)'\mathbf{W}\mathbf{X}(T) \\ \mathbf{g}(T) &= \mathbf{X}(T)'\mathbf{W}\mathbf{y} \end{aligned} \tag{3.79}$$

The solution to the WLS normal equations is

$$\hat{\boldsymbol{\beta}}(T) = \mathbf{G}(T)^{-1}\mathbf{g}(T), \tag{3.80}$$

$\hat{\boldsymbol{\beta}}(T)$ is called the **discounted least squares estimator** of $\boldsymbol{\beta}$.

In many important applications, the discounted least squares estimator can be simplified considerably. Assume that the predictor variables $x_i(t)$ in the model are functions of time that have been chosen so that their values at time period $t + 1$ are linear combinations of their values at the previous time period. That is,

$$x_i(t+1) = L_{i1}x_1(t) + L_{i2}x_2(t) + \cdots + L_{ip}x_p(t), \qquad i = 1, 2, \ldots, p \quad (3.81)$$

In matrix form,

$$\mathbf{x}(t+1) = \mathbf{L}\mathbf{x}(t), \quad (3.82)$$

where $\mathbf{L}$ is the $p \times p$ matrix of the constants L_{ij} in Eq. (3.81). The transition property in Eq. (3.81) holds for polynomial, trigonometric, and certain exponential functions of time. This transition relationship implies that

$$\mathbf{x}(t) = \mathbf{L}^t\mathbf{x}(0) \quad (3.83)$$

Consider the matrix $\mathbf{G}(T)$ in the normal equations (3.78). We can write

$$\begin{aligned}
\mathbf{G}(\mathrm{T}) &= \sum_{j=0}^{T-1} \theta^j \mathbf{x}(-j)\mathbf{x}(-j)' \\
&= \mathbf{G}(T-1) + \theta^{T-1}\mathbf{x}(-(T-1))\mathbf{x}(-(T-1))'
\end{aligned}$$

If the predictor variables $x_i(t)$ in the model are polynomial, trigonometric, or certain exponential functions of time, the matrix $\mathbf{G}(T)$ approaches a steady-state limiting value $\mathbf{G}$,where

$$\mathbf{G} = \sum_{j=0}^{\infty} \theta^j \mathbf{x}(-j)\mathbf{x}(-j)' \quad (3.84)$$

Consequently, the inverse of $\mathbf{G}$ would only need to be computed once. The right-hand side of the normal equations can also be simplified. We can write

$$\begin{aligned}
\mathbf{g}(\mathrm{T}) &= \sum_{j=0}^{T-1} \theta^j y_{T-j}\mathbf{x}(-j) \\
&= y_T\mathbf{x}(0) + \sum_{j=1}^{T-1} \theta^j y_{T-j}\mathbf{x}(-j) \\
&= y_T\mathbf{x}(0) + \theta \sum_{j=1}^{T-1} \theta^{j-1} y_{T-j}\mathbf{L}^{-1}\mathbf{x}(-j+1) \\
&= y_T\mathbf{x}(0) + \theta\mathbf{L}^{-1} \sum_{k=0}^{T-2} \theta^k y_{T-1-k}\mathbf{x}(-k) \\
&= y_T\mathbf{x}(0) + \theta\mathbf{L}^{-1}\mathbf{g}(T-1)
\end{aligned}$$

So the discounted least squares estimator can be written as

$$\hat{\boldsymbol{\beta}}(T) = \mathbf{G}^{-1}\mathbf{g}(T)$$

This can also be simplified. Note that

$$\begin{aligned}\hat{\boldsymbol{\beta}}(T) &= \mathbf{G}^{-1}\mathbf{g}(T)\\ &= \mathbf{G}^{-1}[y_T\mathbf{x}(0) + \theta\mathbf{L}^{-1}\mathbf{g}(T-1)]\\ &= \mathbf{G}^{-1}[y_T\mathbf{x}(0) + \theta\mathbf{L}^{-1}\mathbf{G}\hat{\boldsymbol{\beta}}(T-1)]\\ &= y_T\mathbf{G}^{-1}\mathbf{x}(0) + \theta\mathbf{G}^{-1}\mathbf{L}^{-1}\mathbf{G}\hat{\boldsymbol{\beta}}(T-1)\end{aligned}$$

or

$$\hat{\boldsymbol{\beta}}(T) = \mathbf{h}y_T + \mathbf{Z}\hat{\boldsymbol{\beta}}(T-1) \tag{3.85}$$

where

$$\mathbf{h} = \mathbf{G}^{-1}\mathbf{x}(0) \tag{3.86}$$

and

$$\mathbf{Z} = \theta\mathbf{G}^{-1}\mathbf{L}^{-1}\mathbf{G} \tag{3.87}$$

The right-hand side of Eq. (3.85) can still be simplified because

$$\begin{aligned}\mathbf{L}^{-1}\mathbf{G} &= \mathbf{L}^{-1}\mathbf{G}(\mathbf{L}')^{-1}\mathbf{L}'\\ &= \sum_{j=0}^{\infty}\theta^{j}\mathbf{L}^{-1}\mathbf{x}(-j)\mathbf{x}(-j)'(\mathbf{L}')^{-1}\mathbf{L}'\\ &= \sum_{j=0}^{\infty}\theta^{j}[\mathbf{L}^{-1}\mathbf{x}(-j)][\mathbf{L}^{-1}\mathbf{x}(-j)]'\mathbf{L}'\\ &= \sum_{j=0}^{\infty}\theta^{j}\mathbf{x}(-j-1)\mathbf{x}(-j-1)'\mathbf{L}'\end{aligned}$$

and letting $k = j + 1$,

$$\begin{aligned}\mathbf{L}^{-1}\mathbf{G} &= \theta^{-1}\sum_{k=1}^{\infty}\theta^{k}\mathbf{x}(-k)\mathbf{x}(-k)'\mathbf{L}'\\ &= \theta^{-1}[\mathbf{G} - \mathbf{x}(0)\mathbf{x}(0)']\mathbf{L}'\end{aligned}$$

Substituting for $\mathbf{L}^{-1}\mathbf{G}$ on the right-hand side of Eq. (3.87) results in

$$\begin{aligned}\mathbf{Z} &= \theta\mathbf{G}^{-1}\theta^{-1}[\mathbf{G} - \mathbf{x}(0)\mathbf{x}(0)']\mathbf{L}' \\ &= [\mathbf{I} - \mathbf{G}^{-1}\mathbf{x}(0)\mathbf{x}(0)']\mathbf{L}' \\ &= \mathbf{L}' - \mathbf{h}\mathbf{x}(0)\mathbf{L}' \\ &= \mathbf{L}' - \mathbf{h}[\mathbf{L}\mathbf{x}(0)]' \\ &= \mathbf{L}' - \mathbf{h}\mathbf{x}(1)'\end{aligned}$$

Now the vector of discounted least squares parameter estimates at the end of time period T in Eq. (3.85) is

$$\begin{aligned}\hat{\boldsymbol{\beta}}(T) &= \mathbf{h}y_T + \mathbf{Z}\hat{\boldsymbol{\beta}}(T-1) \\ &= \mathbf{h}y_T + [\mathbf{L}' - \mathbf{h}\mathbf{x}(1)']\hat{\boldsymbol{\beta}}(T-1) \\ &= \mathbf{L}'\hat{\boldsymbol{\beta}}(T-1) + \mathbf{h}[y_T - \mathbf{x}(1)'\hat{\boldsymbol{\beta}}(T-1)].\end{aligned}$$

But $\mathbf{x}(1)'\boldsymbol{\beta}(T-1) = \hat{y}_T(T-1)$ is the forecast of y_T computed at the end of the previous time period, $T-1$, so the discounted least squares vector of parameter estimates computed at the end of time period t is

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\mathrm{T}) &= \mathbf{L}'\hat{\boldsymbol{\beta}}(T-1) + \mathbf{h}[y_T - \hat{y}_T(T-1)] \\ &= \mathbf{L}'\hat{\boldsymbol{\beta}}(T-1) + \mathbf{h}e_t(1).\end{aligned} \tag{3.88}$$

The last line in Eq. (3.88) is an extremely important result; it states that in discounted least squares the vector of parameter estimates computed at the end of time period T can be computed as a simple linear combination of the estimates made at the end of the previous time period $T-1$ and the one-step-ahead forecast error for the observation in period T. Note that there are really two things going on in estimating $\boldsymbol{\beta}$ by discounted least squares: the origin of time is being shifted to the end of the current period, and the estimates of the model parameters are being modified to reflect the forecast error in the current time period. The first and second terms on the right-hand side of Eq. (3.88) accomplish these objectives, respectively.

When discounted least squares estimation is started up, an initial estimate of the parameters is required at time period zero, say, $\hat{\boldsymbol{\beta}}(0)$. This could be found by a standard least squares (or WLS) analysis of historical data.

Because the origin of time is shifted to the end of the current time period, forecasting is easy with discounted least squares. The forecast of the observation at a future time period $T + \tau$, made at the end of time period T, is

$$\begin{aligned}\hat{y}_{T+\tau}(T) &= \hat{\boldsymbol{\beta}}(T)'\mathbf{x}(\tau)\\ &= \sum_{j=1}^{p} \hat{\beta}_j(T)x_j(\tau).\end{aligned} \tag{3.89}$$

Example 3.10 Discounted Least Squares and the Linear Trend Model To illustrate the discounted least squares procedure, let us consider the linear trend model:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad t = 1, 2, \dots, T$$

To write the parameter estimation equations in Eq. (3.88), we need the transition matrix **L**. For the linear trend model, this matrix is

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

Therefore the parameter estimation equations are

$$\hat{\boldsymbol{\beta}}(\mathrm{T}) = \mathbf{L}'\hat{\boldsymbol{\beta}}(\mathrm{T} - 1) + \mathbf{h}\mathrm{e}_{\mathrm{T}}(1)$$

$$\begin{bmatrix} \hat{\beta}_0(T) \\ \hat{\beta}_1(T) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0(T-1) \\ \hat{\beta}_1(T-1) \end{bmatrix} + \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} e_T(1)$$

or

$$\begin{aligned}\hat{\beta}_0(T) &= \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + h_1 e_1(T)\\ \hat{\beta}_1(T) &= \hat{\beta}_1(T-1) + h_2 e_T(1)\end{aligned} \tag{3.90}$$

The elements of the vector **h** are found from Eq. (3.86):

$$\begin{aligned}\mathbf{h} &= \mathbf{G}^{-1}\mathbf{x}(0)\\ &= \mathbf{G}^{-1}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\end{aligned}$$

The steady-state matrix **G** is found as follows:

$$\mathbf{G}(T) = \sum_{j=0}^{T-1} \theta^j \mathbf{x}(-j)\mathbf{x}(-j)'$$

$$= \sum_{j=0}^{T-1} \theta^j \begin{bmatrix} 1 \\ -j \end{bmatrix} \begin{bmatrix} 1 & -j \end{bmatrix}$$

$$= \sum_{j=0}^{T-1} \theta^j \begin{bmatrix} 1 & -j \\ -j & +j^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=0}^{T-1} \theta^j & -\sum_{j=0}^{T-1} j\theta^j \\ -\sum_{j=0}^{T-1} j\theta^j & \sum_{j=0}^{T-1} j^2\theta^j \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1-\theta^T}{1-\theta} & -\dfrac{\theta(1-\theta^T)}{1-\theta} \\ -\dfrac{\theta(1-\theta^T)}{1-\theta} & \dfrac{\theta(1+\theta)(1-\theta^T)}{(1-\theta)^3} \end{bmatrix}$$

The steady-state value of **G**(T) is found by taking the limit as $T \to \infty$, which results in

$$\mathbf{G} = \lim_{T\to\infty} \mathbf{G}(T)$$

$$= \begin{bmatrix} \dfrac{1}{1-\theta} & -\dfrac{\theta}{1-\theta} \\ -\dfrac{\theta}{1-\theta} & \dfrac{\theta(1+\theta)}{(1-\theta)^3} \end{bmatrix}$$

The inverse of **G** is

$$\mathbf{G}^{-1} = \begin{bmatrix} 1-\theta^2 & (1-\theta)^2 \\ (1-\theta)^2 & \dfrac{(1-\theta)^2}{\theta} \end{bmatrix}.$$

Therefore, the vector **h** is

$$\begin{aligned}\mathbf{h} &= \mathbf{G}^{-1}\mathbf{x}(0)\\ &= \mathbf{G}^{-1}\begin{bmatrix}1\\0\end{bmatrix}\\ &= \begin{bmatrix}1-\theta^2 & (1-\theta)^2\\ (1-\theta)^2 & \dfrac{(1-\theta)^2}{\theta}\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}\\ &= \begin{bmatrix}1-\theta^2\\(1-\theta)^2\end{bmatrix}.\end{aligned}$$

Substituting the elements of the vector **h** into Eq. (3.90) we obtain the parameter estimating equations for the linear trend model as

$$\begin{aligned}\hat{\beta}_0(T) &= \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + (1-\theta^2)e_T(1)\\ \hat{\beta}_1(T) &= \hat{\beta}_1(T-1) + (1-\theta)^2 e_T(1)\end{aligned}$$

Inspection of these equations illustrates the twin aspects of discounted least squares; shifting the origin of time, and updating the parameter estimates. In the first equation, the updated intercept at time T consists of the old intercept plus the old slope (this shifts the origin of time to the end of the current period T), plus a fraction of the current forecast error (this revises or updates the estimate of the intercept). The second equation revises the slope estimate by adding a fraction of the current period forecast error to the previous estimate of the slope.

To illustrate the computations, suppose that we are forecasting a time series with a linear trend and we have initial estimates of the slope and intercept at time $t = 0$ as

$$\hat{\beta}_0(0) = 50 \quad \text{and} \quad \hat{\beta}_1(0) = 1.5$$

These estimates could have been obtained by regression analysis of historical data.

Assume that $\theta = 0.9$, so that $1 - \theta^2 = 1 - (0.9)^2 = 0.19$ and $(1 - \theta)^2 = (1 - 0.9)^2 = 0.01$. The forecast for time period $t = 1$, made at the end of time period $t = 0$, is computed from Eq. (3.89):

$$\begin{aligned}\hat{y}_1(0) &= \hat{\boldsymbol{\beta}}(0)'\mathbf{x}(1)\\ &= \hat{\beta}_0(0) + \hat{\beta}_1(0)\\ &= 50 + 1.5\\ &= 51.5\end{aligned}$$

Suppose that the actual observation in time period 1 is $y_1 = 52$. The forecast error in time period 1 is

$$\begin{aligned}e_1(1) &= y_1 - \hat{y}_1(0)\\ &= 52 - 51.5\\ &= 0.5.\end{aligned}$$

The updated estimates of the model parameter computed at the end of time period 1 are now

$$\begin{aligned}\hat{\beta}_0(1) &= \hat{\beta}_0(0) + \hat{\beta}_1(0) + 0.19e_1(0)\\ &= 50 + 1.5 + 0.19(0.5)\\ &= 51.60\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_1(1) &= \hat{\beta}_1(0) + 0.01e_1(0)\\ &= 1.5 + 0.01(0.5)\\ &= 1.55\end{aligned}$$

The origin of time is now $T = 1$. Therefore the forecast for time period 2 made at the end of period 1 is

$$\begin{aligned}\hat{y}_2(1) &= \hat{\beta}_0(1) + \hat{\beta}_1(1)\\ &= 51.6 + 1.55\\ &= 53.15.\end{aligned}$$

If the observation in period 2 is $y_2 = 55$, we would update the parameter estimates exactly as we did at the end of time period 1. First, calculate the forecast error:

$$\begin{aligned}e_2(1) &= y_2 - \hat{y}_2(1)\\ &= 55 - 53.15\\ &= 1.85\end{aligned}$$

Second, revise the estimates of the model parameters:

$$\begin{aligned}\hat{\beta}_0(2) &= \hat{\beta}_0(1) + \hat{\beta}_1(1) + 0.19e_2(1)\\ &= 51.6 + 1.55 + 0.19(1.85)\\ &= 53.50\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_1(2) &= \hat{\beta}_1(1) + 0.01e_2(1)\\ &= 1.55 + 0.01(1.85)\\ &= 1.57\end{aligned}$$

The forecast for period 3, made at the end of period 2, is

$$\begin{aligned}\hat{y}_3(2) &= \hat{\beta}_0(2) + \hat{\beta}_1(2)\\ &= 53.50 + 1.57\\ &= 55.07.\end{aligned}$$

Suppose that a forecast at a longer lead time than one period is required. If a forecast for time period 5 is required at the end of time period 2, then because the forecast lead time is $\tau = 5 - 2 = 3$, the desired forecast is

$$\begin{aligned}\hat{y}_5(2) &= \hat{\beta}_0(2) + \hat{\beta}_1(2)3\\ &= 53.50 + 1.57(3)\\ &= 58.21.\end{aligned}$$

In general, the forecast for any lead time τ, computed at the current origin of time (the end of time period 2), is

$$\begin{aligned}\hat{y}_5(2) &= \hat{\beta}_0(2) + \hat{\beta}_1(2)\tau\\ &= 53.50 + 1.57\tau.\end{aligned}$$

When the discounted least squares procedure is applied to a linear trend model as in Example 3.9, the resulting forecasts are equivalent to the forecasts produced by a method called **double exponential smoothing**. Exponential smoothing is a popular and very useful forecasting technique and will be discussed in detail in Chapter 4.

Discounted least squares can be applied to more complex models. For example, suppose that the model is a polynomial of degree k. The transition matrix for this model is a square $(k+1)\times(k+1)$ matrix in which the diagonal elements are unity, the elements immediately to the left of the diagonal are also unity, and all other elements are zero. In this polynomial, the term of degree r is written as

$$\beta_r\binom{t}{r} = \beta_r\frac{t!}{(t-r)!r!}$$

In the next example we illustrate discounted least squares for a simple seasonal model.

Example 3.11 A Simple Seasonal Model Suppose that a time series can be modeled as a linear trend with a superimposed sine wave to represent a seasonal pattern that is observed monthly. The model is a variation of the one shown in Eq. (3.3):

$$y_t = \beta_0 + \beta_1 t + \beta_2 \sin\frac{2\pi}{d}t + \beta_3 \cos\frac{2\pi}{d}t + \varepsilon \tag{3.91}$$

Since this model represents monthly data, $d = 12$, Eq. (3.91) becomes

$$y_t = \beta_0 + \beta_1 t + \beta_2 \sin\frac{2\pi}{12}t + \beta_3 \cos\frac{2\pi}{12}t + \varepsilon \tag{3.92}$$

The transition matrix $\mathbf{L}$ for this model, which contains a mixture of polynomial and trigonometric terms, is

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & \cos\frac{2\pi}{12} & \sin\frac{2\pi}{12} \\ 0 & 0 & -\sin\frac{2\pi}{12} & \cos\frac{2\pi}{12} \end{bmatrix}.$$

Note that $\mathbf{L}$ has a **block diagonal structure**, with the first block containing the elements for the polynomial portion of the model and the second block containing the elements for the trigonometric terms, and the remaining elements of the matrix are zero. The parameter estimation equations for this model are

$$\hat{\boldsymbol{\beta}}(\mathrm{T}) = \mathbf{L}'\hat{\boldsymbol{\beta}}(\mathrm{T}-1) + \mathbf{h}\mathrm{e}_{\mathrm{T}}(1)$$

$$\begin{bmatrix} \hat{\beta}_0(T) \\ \hat{\beta}_1(T) \\ \hat{\beta}_2(T) \\ \hat{\beta}_3(T) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & \cos\frac{2\pi}{12} & \sin\frac{2\pi}{12} \\ 0 & 0 & -\sin\frac{2\pi}{12} & \cos\frac{2\pi}{12} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0(T-1) \\ \hat{\beta}_1(T-1) \\ \hat{\beta}_2(T-1) \\ \hat{\beta}_3(T-1) \end{bmatrix} + \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} e_T(1)$$

or

$$\begin{aligned}
\hat{\beta}_0(T) &= \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + h_1 e_T(1) \\
\hat{\beta}_1(T) &= \hat{\beta}_1(T-1) + h_2 e_T(1) \\
\hat{\beta}_2(T) &= \cos\frac{2\pi}{12}\hat{\beta}_2(T-1) - \sin\frac{2\pi}{12}\hat{\beta}_3(T-1) + h_3 e_T(1) \\
\hat{\beta}_3(T) &= \sin\frac{2\pi}{12}\hat{\beta}_2(T-1) + \cos\frac{2\pi}{12}\hat{\beta}_3(T-1) + h_4 e_T(1)
\end{aligned}$$

and since $2\pi/12 = 30°$, these equations become

$$\begin{aligned}
\hat{\beta}_0(T) &= \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + h_1 e_T(1) \\
\hat{\beta}_1(T) &= \hat{\beta}_1(T-1) + h_2 e_T(1) \\
\hat{\beta}_2(T) &= 0.866\hat{\beta}_2(T-1) - 0.5\hat{\beta}_3(T-1) + h_3 e_T(1) \\
\hat{\beta}_3(T) &= 0.5\hat{\beta}_2(T-1) + 0.866\hat{\beta}_3(T-1) + h_4 e_T(1)
\end{aligned}$$

The steady-state **G** matrix for this model is

$$\mathbf{G} = \begin{bmatrix}
\sum_{k=0}^{\infty}\theta^k & -\sum_{k=0}^{\infty}k\theta^k & -\sum_{k=0}^{\infty}\theta^k\sin\omega k & \sum_{k=0}^{\infty}\theta^k\cos\omega k \\
 & \sum_{k=0}^{\infty}k^2\theta^k & \sum_{k=0}^{\infty}k\theta^k\sin\omega k & -\sum_{k=0}^{\infty}k\theta^k\cos\omega k \\
 & & \sum_{k=0}^{\infty}\theta^k\sin\omega k\sin\omega k & -\sum_{k=0}^{\infty}\theta^k\sin\omega k\cos\omega k \\
 & & & \sum_{k=0}^{\infty}\theta^k\cos\omega k\cos\omega k
\end{bmatrix}$$

where we have let $\omega = 2\pi/12$. Because $\mathbf{G}$ is symmetric, we only need to show the upper half of the matrix. It turns out that there are closed-form expressions for all of the entries in $\mathbf{G}$. We will evaluate these expressions for $\theta = 0.9$. This gives the following:

$$\sum_{k=0}^{\infty} \theta^k = \frac{1}{1-\theta} = \frac{1}{1-0.9} = 10$$

$$\sum_{k=0}^{\infty} k\theta^k = \frac{\theta}{(1-\theta)^2} = \frac{0.9}{(1-0.9)^2} = 90$$

$$\sum_{k=0}^{\infty} k^2\theta^k = \frac{\theta(1+\theta)}{(1-\theta)^3} = \frac{0.9(1+0.9)}{(1-0.9)^3} = 1710$$

for the polynomial terms and

$$\sum_{k=0}^{\infty} \theta^k \sin \omega k = \frac{\theta \sin \omega}{1-2\theta\cos\omega+\theta^2} = \frac{(0.9)0.5}{1-2(0.9)0.866+(0.9)^2} = 1.79$$

$$\sum_{k=0}^{\infty} \theta^k \cos \omega k = \frac{1-\theta \cos \omega}{1-2\theta\cos\omega+\theta^2} = \frac{1-(0.9)0.866}{1-2(0.9)0.866+(0.9)^2} = 0.8824$$

$$\begin{aligned}\sum_{k=0}^{\infty} k\theta^k \sin \omega k &= \frac{\theta(1-\theta^2)\sin\omega}{(1-2\theta\cos\omega+\theta^2)^2} = \frac{0.9[1-(0.9)^2]0.5}{[1-2(0.9)0.866+(0.9)^2]^2}\\ &= 1.368\end{aligned}$$

$$\begin{aligned}\sum_{k=0}^{\infty} k\theta^k \cos \omega k &= \frac{2\theta^2-\theta(1+\theta^2)\cos\omega}{(1-2\theta\cos\omega+\theta^2)^2} = \frac{2(0.9)^2-0.9[1+(0.9)^2]0.866}{[1-2(0.9)0.866+(0.9)^2]^2}\\ &= 3.3486\end{aligned}$$

$$\begin{aligned}\sum_{k=0}^{\infty} \theta^k \sin \omega k \sin \omega k &= -\frac{1}{2}\left[\frac{1-\theta\cos(2\omega)}{1-2\theta\cos(2\omega)+\theta^2} - \frac{1-\theta\cos(0)}{1-2\theta\cos(0)+\theta^2}\right]\\ &\quad -\frac{1}{2}\left[\frac{1-0.9(0.5)}{1-2(0.9)0.5+(0.9)^2} - \frac{1-0.9(1)}{1-2(0.9)(1)+(0.9)^2}\right]\\ &= 4.7528\end{aligned}$$

$$\begin{aligned}
\sum_{k=0}^{\infty} \theta^k \sin \omega k \cos \omega k &= \frac{1}{2}\left[\frac{\theta \sin(2\omega)}{1-2\theta\cos(2\omega)+\theta^2} - \frac{\theta \sin(0)}{1-2\theta\cos(0)+\theta^2}\right] \\
&= \frac{1}{2}\left[\frac{0.9(0.866)}{1-2(0.9)0.5+(0.9)^2} + \frac{0.9(0)}{1-2(0.9)1+(0.9)^2}\right] \\
&= 0.4284
\end{aligned}$$

$$\begin{aligned}
\sum_{k=0}^{\infty} \theta^k \cos \omega k \cos \omega k &= \frac{1}{2}\left[\frac{1-\theta \cos(2\omega)}{1-2\theta\cos(2\omega)+\theta^2} + \frac{1-\theta \cos(0)}{1-2\theta\cos(0)+\theta^2}\right] \\
&= \frac{1}{2}\left[\frac{1-0.9(0.5)}{1-2(0.9)0.5+(0.9)^2} + \frac{1-0.9(1)}{1-2(0.9)(1)+(0.9)^2}\right] \\
&= 5.3022
\end{aligned}$$

for the trignometric terms. Therefore the **G** matrix is

$$\mathbf{G} = \begin{bmatrix} 10 & -90 & -1.79 & 0.8824 \\ & 1740 & 1.368 & -3.3486 \\ & & 4.7528 & -0.4284 \\ & & & 5.3022 \end{bmatrix}$$

and $\mathbf{G}^{-1}$ is

$$\mathbf{G}^{-1} = \begin{bmatrix} 0.214401 & 0.01987 & 0.075545 & -0.02264 \\ 0.01987 & 0.001138 & 0.003737 & -0.00081 \\ 0.075545 & 0.003737 & 0.238595 & 0.009066 \\ -0.02264 & -0.00081 & 0.009066 & 0.192591 \end{bmatrix}$$

where we have shown the entire matrix. The **h** vector is

$$\begin{aligned}
\mathbf{h} &= \mathbf{G}^{-1}\mathbf{x}(0) \\
&= \begin{bmatrix} 0.214401 & 0.01987 & 0.075545 & -0.02264 \\ 0.01987 & 0.001138 & 0.003737 & -0.00081 \\ 0.075545 & 0.003737 & 0.238595 & 0.009066 \\ -0.02264 & -0.00081 & 0.009066 & 0.192591 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.191762 \\ 0.010179 \\ 0.084611 \\ 0.169953 \end{bmatrix}
\end{aligned}$$

Therefore the discounted least squares parameter estimation equations are

$$\begin{aligned}
\hat{\beta}_0(T) &= \hat{\beta}_0(T-1) + \hat{\beta}_1(T-1) + 0.191762e_T(1)\\
\hat{\beta}_1(T) &= \hat{\beta}_1(T-1) + 0.010179e_T(1)\\
\hat{\beta}_2(T) &= \cos\frac{2\pi}{12}\hat{\beta}_2(T-1) - \sin\frac{2\pi}{12}\hat{\beta}_3(T-1) + 0.084611e_T(1)\\
\hat{\beta}_3(T) &= \sin\frac{2\pi}{12}\hat{\beta}_2(T-1) + \cos\frac{2\pi}{12}\hat{\beta}_3(T-1) + 0.169953e_T(1)
\end{aligned}$$

3.8 REGRESSION MODELS FOR GENERAL TIME SERIES DATA

Many applications of regression in forecasting involve both predictor and response variables that are time series. Regression models using time series data occur relatively often in economics, business, and many fields of engineering. The assumption of uncorrelated or independent errors that is typically made for cross-section regression data is often not appropriate for time series data. Usually the errors in time series data exhibit some type of autocorrelated structure. You might find it useful at this point to review the discussion of autocorrelation in time series data from Chapter 2.

There are several **sources** of autocorrelation in time series regression data. In many cases, the cause of autocorrelation is the failure of the analyst to include one or more important predictor variables in the model. For example, suppose that we wish to regress the annual sales of a product in a particular region of the country against the annual advertising expenditures for that product. Now the growth in the population in that region over the period of time used in the study will also influence the product sales. If population size is not included in the model, this may cause the errors in the model to be positively autocorrelated, because if the per capita demand for the product is either constant or increasing with time, population size is positively correlated with product sales.

The presence of autocorrelation in the errors has several effects on the OLS regression procedure. These are summarized as follows:

1. The OLS regression coefficients are still unbiased, but they are no longer minimum-variance estimates. We know this from our study of GLS in Section 3.7.
2. When the errors are positively autocorrelated, the residual mean square may seriously underestimate the error variance σ^2.

Consequently, the standard errors of the regression coefficients may be too small. As a result, confidence and prediction intervals are shorter than they really should be, and tests of hypotheses on individual regression coefficients may be misleading, in that they may indicate that one or more predictor variables contribute significantly to the model when they really do not. Generally, underestimating the error variance σ^2 gives the analyst a false impression of precision of estimation and potential forecast accuracy.

3. The confidence intervals, prediction intervals, and tests of hypotheses based on the t and Fdistributions are, strictly speaking, no longer exact procedures.

There are three approaches to dealing with the problem of autocorrelation. If autocorrelation is present because of one or more omitted predictors and if those predictor variable(s) can be identified and included in the model, the observed autocorrelation should disappear. Alternatively, the WLS or GLS methods discussed in Section 3.7 could be used if there were sufficient knowledge of the autocorrelation structure. Finally, if these approaches cannot be used, the analyst must turn to a model that specifically incorporates the autocorrelation structure. These models usually require special parameter estimation techniques. We will provide an introduction to these procedures in Section 3.8.2.

3.8.1 Detecting Autocorrelation: The Durbin–Watson Test

Residual plots can be useful for the detection of autocorrelation. The most useful display is the plot of residuals versus time. If there is positive autocorrelation, residuals of identical sign occur in clusters: that is, there are not enough changes of sign in the pattern of residuals. On the other hand, if there is negative autocorrelation, the residuals will alternate signs too rapidly.

Various **statistical tests** can be used to detect the presence of autocorrelation. The test developed by Durbin and Watson (1950, 1951, 1971) is a very widely used procedure. This test is based on the assumption that the errors in the regression model are generated by a **first-order autoregressive process** observed at equally spaced time periods; that is,

$$\varepsilon_t = \phi\varepsilon_{t-1} + a_t, \tag{3.93}$$

where ε_t is the error term in the model at time period t, a_t is an NID$(0, \sigma_a^2)$ random variable, and ϕ is a parameter that defines the relationship between successive values of the model errors ε_t and ε_{t-1}. We will require that $|\phi| < 1$, so that the model error term in time period t is equal to a fraction of the error experienced in the immediately preceding period plus a normally and independently distributed random shock or disturbance that is unique to the current period. In time series regression models ϕ is sometimes called the **autocorrelation parameter**. Thus a simple linear regression model with **first-order autoregressive errors** would be

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t = \phi\varepsilon_{t-1} + a_t, \tag{3.94}$$

where y_t and x_t are the observations on the response and predictor variables at time period t.

When the regression model errors are generated by the first-order autoregressive process in Eq. (3.93), there are several interesting properties of these errors. By successively substituting for $\varepsilon_t, \varepsilon_{t-1}, \ldots$ on the right-hand side of Eq. (3.93) we obtain

$$\varepsilon_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$$

In other words, the error term in the regression model for period t is just a linear combination of all of the current and previous realizations of the NID$(0, \sigma^2)$ random variables a_t. Furthermore, we can show that

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ \text{Var}\ (\varepsilon_t) &= \sigma^2 = \sigma_a^2 \left(\frac{1}{1-\phi^2} \right) \\ \text{Cov}\ (\varepsilon_t, \varepsilon_{t\pm j}) &= \phi^j \sigma_a^2 \left(\frac{1}{1-\phi^2} \right) \end{aligned} \tag{3.95}$$

That is, the errors have zero mean and constant variance but have a nonzero covariance structure unless $\phi = 0$.

The **autocorrelation** between two errors that are one period apart, or the **lag one autocorrelation**, is

$$\rho_1 = \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t+1})}{\sqrt{\text{Var}\ (\varepsilon_t)}\sqrt{\text{Var}\ (\varepsilon_t)}}$$

$$= \frac{\phi\sigma_a^2\left(\frac{1}{1-\phi^2}\right)}{\sqrt{\sigma_a^2\left(\frac{1}{1-\phi^2}\right)}\sqrt{\sigma_a^2\left(\frac{1}{1-\phi^2}\right)}}$$
$$= \phi$$

The autocorrelation between two errors that are k periods apart is

$$\rho_k = \phi^k, \quad i = 1, 2, \ldots$$

This is called the **autocorrelation function** (refer to Section 2.3.2). Recall that we have required that $|\phi| < 1$. When ϕ is positive, all error terms are positively correlated, but the magnitude of the correlation decreases as the errors grow further apart. Only if $\phi = 0$ are the model errors uncorrelated.

Most time series regression problems involve data with positive autocorrelation. The Durbin–Watson test is a statistical test for the presence of positive autocorrelation in regression model errors. Specifically, the hypotheses considered in the Durbin–Watson test are

$$\begin{aligned} H_0 &: \phi = 0 \\ H_1 &: \phi > 0 \end{aligned} \tag{3.96}$$

The Durbin–Watson **test statistic** is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = \frac{\sum_{t=2}^{T} e_t^2 + \sum_{t=2}^{T} e_{t-1}^2 - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \approx 2(1 - r_1), \tag{3.97}$$

where the e_t, $t = 1, 2, \ldots, T$ are the residuals from an OLS regression of y_t on x_t. In Eq. (3.97) r_1 is the lag one autocorrelation between the residuals, so for uncorrelated errors the value of the Durbin–Watson statistic should be approximately 2. Statistical testing is necessary to determine just how far away from 2 the statistic must fall in order for us to conclude that the assumption of uncorrelated errors is violated. Unfortunately, the distribution of the Durbin–Watson test statistic d depends on the **X** matrix, and this makes critical values for a statistical test difficult to obtain. However, Durbin and Watson (1951) show that d lies between lower and upper bounds, say, d_{L} and d_{U} , such that if d is outside these limits, a conclusion regarding the hypotheses in Eq. (3.96) can be reached. The decision procedure is as follows:

$$\begin{gathered}\text{If } d < d_{\mathrm{L}} \text{ reject } H_0 : \phi = 0\\ \text{If } d > d_{\mathrm{U}} \text{ do not reject } H_0 : \phi = 0\\ \text{If } d_{\mathrm{L}} \leq d \leq d_{\mathrm{U}} \text{ the test is inconclusive}\end{gathered}$$

Table A.5 in Appendix A gives the bounds d_{L} and d_{U} for a range of sample sizes, various numbers of predictors, and three type I error rates ($\alpha = 0.05$, $\alpha = 0.025$, and $\alpha = 0.01$). It is clear that small values of the test statistic d imply that $H_0 : \phi = 0$ should be rejected because positive autocorrelation indicates that successive error terms are of similar magnitude, and the differences in the residuals $e_t - e_{t-1}$ will be small. Durbin and Watson suggest several procedures for resolving inconclusive results. A reasonable approach in many of these inconclusive situations is to analyze the data as if there were positive autocorrelation present to see if any major changes in the results occur.

Situations where negative autocorrelation occurs are not often encountered. However, if a test for negative autocorrelation is desired, one can use the statistic $4 - d$, where d is defined in Eq. (3.97). Then the decision rules for testing the hypotheses $H_0 : \phi = 0$ versus $H_1 : \phi < 0$ are the same as those used in testing for positive autocorrelation. It is also possible to test a two-sided alternative hypothesis ($H_0 : \phi = 0$ versus $H_1 : \phi \neq 0$) by using both of the one-sided tests simultaneously. If this is done, the two-sided procedure has type I error 2α, where α is the type I error used for each individual one-sided test.

Example 3.12 Montgomery, Peck, and Vining (2012) present an example of a regression model used to relate annual regional advertising expenses to annual regional concentrate sales for a soft drink company. Table 3.14 presents the 20 years of these data used by Montgomery, Peck, and Vining (2012). The authors assumed that a straight-line relationship was appropriate and fit a simple linear regression model by OLS. The Minitab output for this model is shown in Table 3.15 and the residuals are shown in the last column of Table 3.14. Because these are time series data, there is a possibility that autocorrelation may be present. The plot of residuals versus time, shown in Figure 3.5, has a pattern indicative of potential autocorrelation; there is a definite upward trend in the plot, followed by a downward trend.

TABLE 3.14 Soft Drink Concentrate Sales Data

Year	Sales (Units)	Expenditures (10^3 Dollars)	Residuals
1	3083	75	−32.3298
2	3149	78	−26.6027
3	3218	80	2.2154
4	3239	82	−16.9665
5	3295	84	−1.1484
6	3374	88	−2.5123
7	3475	93	−1.9671
8	3569	97	11.6691
9	3597	99	−0.5128
10	3725	104	27.0324
11	3794	109	−4.4224
12	3959	115	40.0318
13	4043	120	23.5770
14	4194	127	33.9403
15	4318	135	−2.7874
16	4493	144	−8.6060
17	4683	153	0.5753
18	4850	161	6.8476
19	5005	170	−18.9710
20	5236	182	−29.0625

We will use the Durbin–Watson test for

$$H_0 : \phi = 0$$
$$H_1 : \phi > 0$$

The test statistic is calculated as follows:

$$\begin{aligned} d &= \frac{\sum_{t=2}^{20} (e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} \\ &= \frac{[-26.6027 - (-32.3298)]^2 + [2.2154 - (-26.6027)]^2 + \cdots + [-29.0625 - (-18.9710)]^2}{(-32.3298)^2 + (-26.6027)^2 + \cdots + (-29.0625)^2} \\ &= 1.08 \end{aligned}$$

Minitab will also calculate and display the Durbin–Watson statistic. Refer to the Minitab output in Table 3.15. If we use a significance level

TABLE 3.15 Minitab Output for the Soft Drink Concentrate Sales Data

Regression Analysis: Sales Versus Expenditures

```
The regression equation is
Sales = 1609 + 20.1 Expenditures

Predictor        Coef  SE Coef       T      P
Constant      1608.51    17.02   94.49  0.000
Expenditures  20.0910   0.1428  140.71  0.000

S = 20.5316   R-Sq = 99.9%   R-Sq(adj) = 99.9%

Analysis of Variance

Source          DF       SS       MS         F      P
Regression       1  8346283  8346283  19799.11  0.000
Residual Error  18     7588      422
Total           19  8353871

Unusual Observations

Obs  Expenditures    Sales      Fit  SE Fit  Residual  St Resid
 12           115  3959.00  3918.97    4.59     40.03      2.00R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.08005
```

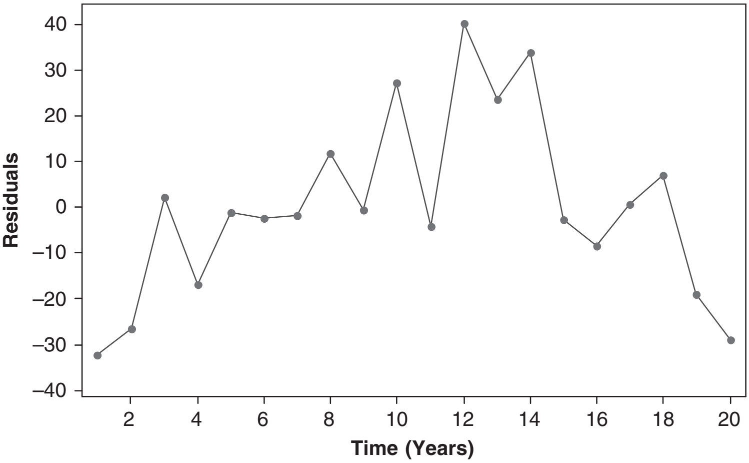


FIGURE 3.5 Plot of residuals versus time for the soft drink concentrate sales model.

of 0.05, Table A.5 in Appendix A gives the critical values corresponding to one predictor variable and 20 observations as $d_{\mathrm{L}} = 1.20$ and $d_{\mathrm{U}} = 1.41$. Since the calculated value of the Durbin–Watson statistic $d = 1.08$ is less than $d_{\mathrm{L}} = 1.20$, we reject the null hypothesis and conclude that the errors in the regression model are positively autocorrelated.

3.8.2 Estimating the Parameters in Time Series Regression Models

A significant value of the Durbin–Watson statistic or a suspicious residual plot indicates a potential problem with auto correlated model errors. This could be the result of an actual time dependence in the errors or an "artificial" time dependence caused by the omission of one or more important predictor variables. If the apparent autocorrelation results from missing predictors and if these missing predictors can be identified and incorporated into the model, the apparent autocorrelation problem may be eliminated. This is illustrated in the following example.

Example 3.13 Table 3.16 presents an expanded set of data for the soft drink concentrate sales problem introduced in Example 3.12. Because it is reasonably likely that regional population affects soft drink sales, Montgomery, Peck, and Vining (2012) provided data on regional population for each of the study years. Table 3.17 is the Minitab output for a regression model that includes as the predictor variables advertising expenditures and population. Both of these predictor variables are highly significant. The last column of Table 3.16 shows the residuals from this model. Minitab calculates the Durbin–Watson statistic for this model as $d = 3.05932$, and the 5% critical values are $d_{\mathrm{L}} = 1.10$ and $d_{\mathrm{U}} = 1.54$, and since d is greater than d_{U}, we conclude that there is no evidence to reject the null hypothesis. That is, there is no indication of autocorrelation in the errors.

Figure 3.6 is a plot of the residuals from this regression model in time order. This plot shows considerable improvement when compared to the plot of residuals from the model using only advertising expenditures as the predictor. Therefore, we conclude that adding the new predictor population size to the original model has eliminated an apparent problem with autocorrelation in the errors.

The Cochrane–Orcutt Method When the observed autocorrelation in the model errors cannot be removed by adding one or more new predictor variables to the model, it is necessary to take explicit account of the autocorrelative structure in the model and use an appropriate parameter

TABLE 3.16 Expanded Soft Drink Concentrate Sales Data for Example 3.13

Year	Sales (Units)	Expenditures (10^3 Dollars)	Population	Residuals
1	3083	75	825,000	−4.8290
2	3149	78	830,445	−3.2721
3	3218	80	838,750	14.9179
4	3239	82	842,940	−7.9842
5	3295	84	846,315	5.4817
6	3374	88	852,240	0.7986
7	3475	93	860,760	−4.6749
8	3569	97	865,925	6.9178
9	3597	99	871,640	−11.5443
10	3725	104	877,745	14.0362
11	3794	109	886,520	−23.8654
12	3959	115	894,500	17.1334
13	4043	120	900,400	−0.9420
14	4194	127	904,005	14.9669
15	4318	135	908,525	−16.0945
16	4493	144	912,160	−13.1044
17	4683	153	917,630	1.8053
18	4850	161	922,220	13.6264
19	5005	170	925,910	−3.4759
20	5236	182	929,610	0.1025

estimation method. A very good and widely used approach is the procedure devised by Cochrane and Orcutt (1949).

We will describe the Cochrane–Orcutt method for the simple linear regression model with first-order autocorrelated errors given in Eq. (3.94). The procedure is based on transforming the response variable so that $y_t' = y_t - \phi y_{t-1}$. Substituting for y_t and y_{t-1}, the model becomes

$$\begin{aligned} y_t' &= y_t - \phi y_{t-1} \\ &= \beta_0 + \beta_1 x_t + \varepsilon_t - \phi(\beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}) \\ &= \beta_0(1-\phi) + \beta_1(x_t - \phi x_{t-1}) + \varepsilon_t - \phi\varepsilon_{t-1} \\ &= \beta_0' + \beta_1 x_t' + a_t, \end{aligned} \tag{3.98}$$

where $\beta_0' = \beta_0(1-\phi)$ and $x_t' = x_t - \phi x_{t-1}$. Note that the error terms a_t in the transformed or reparameterized model are independent random variables. Unfortunately, this new reparameterized model contains an unknown

TABLE 3.17 Minitab Output for the Soft Drink Concentrate Data in Example 3.13

Regression Analysis: Sales Versus Expenditures, Population

```
The regression equation is
Sales = 320 + 18.4 Expenditures + 0.00168 Population

Predictor          Coef     SE Coef      T      P
Constant          320.3       217.3   1.47  0.159
Expenditures    18.4342      0.2915  63.23  0.000
Population    0.0016787  0.0002829   5.93  0.000

S = 12.0557   R-Sq = 100.0%   R-Sq(adj) = 100.0%

Analysis of Variance

Source          DF       SS       MS         F      P
Regression       2  8351400  4175700  28730.40  0.000
Residual Error  17     2471      145
Total           19  8353871

Source        DF   Seq SS
Expenditures   1  8346283
Population     1     5117

Unusual Observations

Obs  Expenditures    Sales      Fit  SE Fit  Residual  St Resid
 11           109  3794.00  3817.87    4.27    -23.87    -2.12R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 3.05932
```

parameter ϕ and it is also no longer linear in the unknown parameters because it involves products of ϕ, β_0, and β_1. However, the first-order autoregressive process $\varepsilon_t = \phi\varepsilon_{t-1} + a_t$ can be viewed as a simple linear regression through the origin and the parameter ϕ can be estimated by obtaining the residuals of an OLS regression of y_t on x_t and then regressing e_t on e_{t-1}. The OLS regression of e_t on e_{t-1} results in

$$\hat{\phi} = \frac{\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{y=1}^{T} e_t^2} \tag{3.99}$$

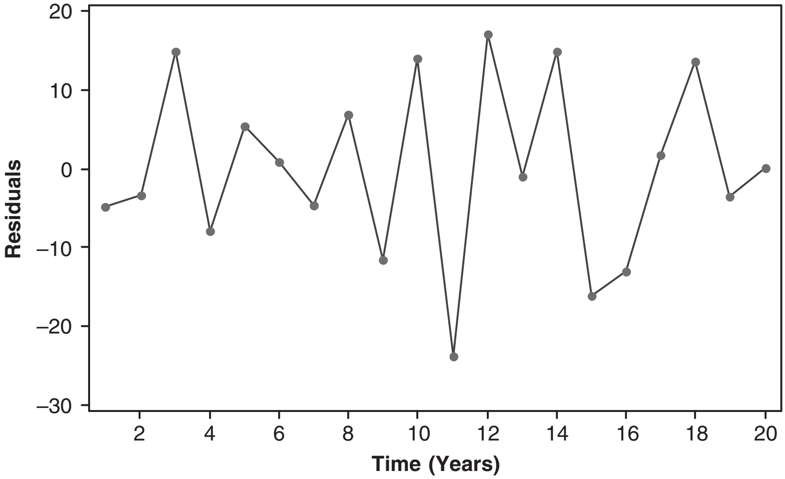


FIGURE 3.6 Plot of residuals versus time for the soft drink concentrate sales model in Example 3.13.

Using $\hat{\phi}$ as an estimate of ϕ, we can calculate the transformed response and predictor variables as

$$y'_t = y_t - \hat{\phi} y_{t-1}$$
$$x'_t = x_t - \hat{\phi} x_{t-1}$$

Now apply OLS to the transformed data. This will result in estimates of the transformed slope $\hat{\beta}'_0$, the intercept $\hat{\beta}_1$, and a new set of residuals. The Durbin–Watson test can be applied to these new residuals from the reparameterized model. If this test indicates that the new residuals are uncorrelated, then no additional analysis is required. However, if positive autocorrelation is still indicated, then another iteration is necessary. In the second iteration ϕ is estimated with new residuals that are obtained by using the regression coefficients from the reparameterized model with the original regressor and response variables. This iterative procedure may be continued as necessary until the residuals indicate that the error terms in the reparameterized model are uncorrelated. Usually only one or two iterations are sufficient to produce uncorrelated errors.

Example 3.14 Montgomery, Peck, and Vining (2012) give data on the market share of a particular brand of toothpaste for 30 time periods and the corresponding selling price per pound. These data are shown in

TABLE 3.18 Toothpaste Market Share Data

Time	Market Share	Price	Residuals	y_t'	x_t'	Residuals
1	3.63	0.97	0.281193			
2	4.20	0.95	0.365398	2.715	0.553	−0.189435
3	3.33	0.99	0.466989	1.612	0.601	0.392201
4	4.54	0.91	−0.266193	3.178	0.505	−0.420108
5	2.89	0.98	−0.215909	1.033	0.608	−0.013381
6	4.87	0.90	−0.179091	3.688	0.499	−0.058753
7	4.90	0.89	−0.391989	2.908	0.522	−0.268949
8	5.29	0.86	−0.730682	3.286	0.496	−0.535075
9	6.18	0.85	−0.083580	4.016	0.498	0.244473
10	7.20	0.82	0.207727	4.672	0.472	0.256348
11	7.25	0.79	−0.470966	4.305	0.455	−0.531811
12	6.09	0.83	−0.659375	3.125	0.507	−0.423560
13	6.80	0.81	−0.435170	4.309	0.471	−0.131426
14	8.65	0.77	0.443239	5.869	0.439	0.635804
15	8.43	0.76	−0.019659	4.892	0.445	−0.192552
16	8.29	0.80	0.811932	4.842	0.489	0.847507
17	7.18	0.83	0.430625	3.789	0.503	0.141344
18	7.90	0.79	0.179034	4.963	0.451	0.027093
19	8.45	0.76	0.000341	5.219	0.437	−0.063744
20	8.23	0.78	0.266136	4.774	0.469	0.284026

Table 3.18. A simple linear regression model is fit to these data, and the resulting Minitab output is in Table 3.19. The residuals are shown in Table 3.18. The Durbin–Watson statistic for the residuals from this model is $d = 1.13582$ (see the Minitab output), and the 5% critical values are $d_{\mathrm{L}} = 1.20$ and $d_{\mathrm{U}} = 1.41$, so there is evidence to support the conclusion that the residuals are positively autocorrelated.

We will use the Cochrane–Orcutt method to estimate the model parameters. The autocorrelation coefficient can be estimated using the residuals in Table 3.18 and Eq. (3.99) as follows:

$$\begin{aligned}\hat{\phi} &= \frac{\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{y=1}^{T} e_t^2}\\ &= \frac{1.3547}{3.3083}\\ &= 0.409\end{aligned}$$

TABLE 3.19 Minitab Regression Results for the Toothpaste Market Share Data

Regression Analysis: Market Share Versus Price

```
The regression equation is
Market Share = 26.9 - 24.3 Price

Predictor     Coef  SE Coef       T      P
Constant    26.910    1.110   24.25  0.000
Price      -24.290    1.298  -18.72  0.000

S = 0.428710   R-Sq = 95.1%   R-Sq(adj) = 94.8%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  64.380  64.380  350.29  0.000
Residual Error  18   3.308   0.184
Total           19  67.688

Durbin-Watson statistic = 1.13582
```

The transformed variables are computed according to

$$y'_t = y_t - 0.409y_{t-1}$$
$$x'_t = x_t - 0.409x_{t-1}$$

for $t = 2, 3, \ldots, 20$. These transformed variables are also shown in Table 3.18. The Minitab results for fitting a regression model to the transformed data are summarized in Table 3.20. The residuals from the transformed model are shown in the last column of Table 3.18. The Durbin–Watson statistic for the transformed model is $d = 2.15671$, and the 5% critical values from Table A.5 in Appendix A are $d_{\mathrm{L}} = 1.18$ and $d_{\mathrm{U}} = 1.40$, so we conclude that there is no problem with autocorrelated errors in the transformed model. The Cochrane–Orcutt method has been effective in removing the autocorrelation.

The slope in the transformed model β'_1 is equal to the slope in the original model, β_1. A comparison of the slopes in the two models in Tables 3.19 and 3.20 shows that the two estimates are very similar. However, if the standard errors are compared, the Cochrane–Orcutt method produces an estimate of the slope that has a larger standard error than the standard error of the OLS estimate. This reflects the fact that if the errors are autocorrelated and

TABLE 3.20 Minitab Regression Results for Fitting the Transformed Model to the Toothpaste Sales Data

Regression Analysis: y′ Versus x′

```
The regression equation is
y-prime = 16.1 - 24.8 x-prime

Predictor      Coef  SE Coef       T      P
Constant    16.1090   0.9610   16.76  0.000
x-prime     -24.774    1.934  -12.81  0.000

S = 0.390963   R-Sq = 90.6%   R-Sq(adj) = 90.1%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  25.080  25.080  164.08  0.000
Residual Error  17   2.598   0.153
Total           18  27.679

Unusual Observations

Obs  x-prime  y-prime     Fit  SE Fit  Residual  St Resid
  2    0.601   1.6120  1.2198  0.2242    0.3922      1.22 X
  4    0.608   1.0330  1.0464  0.2367   -0.0134     -0.04 X
 15    0.489   4.8420  3.9945  0.0904    0.8475      2.23R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 2.15671
```

OLS is used, the standard errors of the model coefficients are likely to be underestimated.

The Maximum Likelihood Approach There are other alternatives to the Cochrane–Orcutt method. A popular approach is to use the method of **maximum likelihood** to estimate the parameters in a time series regression model. We will concentrate on the simple linear regression model with first-order autoregressive errors

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t = \phi\varepsilon_{t-1} + a_t \tag{3.100}$$

One reason that the method of maximum likelihood is so attractive is that, unlike the Cochrane–Orcutt method, it can be used in situations where the autocorrelative structure of the errors is more complicated than first-order autoregressive.

For readers unfamiliar with maximum likelihood estimation, we will present a simple example. Consider the time series model

$$y_t = \mu + a_t, \tag{3.101}$$

where a_t is $N(0, \sigma^2)$ and μ is unknown. This is a time series model for a process that varies randomly around a fixed level (μ) and for which there is no autocorrelation. We will estimate the unknown parameter μ using the method of maximum likelihood.

Suppose that there are T observations available, $y_1, y_2, \dots, y_T$. The probability distribution of any observation is normal, that is,

$$\begin{aligned} f(y_t) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-[(y_t-\mu)/\sigma]^2/2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-(a_t/\sigma)^2/2} \end{aligned}$$

The **likelihood function** is just the joint probability density function of the sample. Because the observations $y_1, y_2, \dots, y_T$ are independent, the likelihood function is just the product of the individual density functions, or

$$\begin{aligned} l(y_t, \mu) &= \prod_{t=1}^{T} f(y_t) \\ &= \prod_{t=1}^{T} \frac{1}{\sigma\sqrt{2\pi}} e^{-(a_t/\sigma)^2/2} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^T \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t^2\right) \end{aligned} \tag{3.102}$$

The **maximum likelihood estimator** of μ is the value of the parameter that maximizes the likelihood function. It is often easier to work with the log-likelihood, and this causes no problems because the value of μ that maximizes the likelihood function also maximizes the log-likelihood.

The log-likelihood is

$$\ln l(y_t \mu) = -\frac{T}{2}\ln(2\pi) - T\ln\sigma - \frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t^2$$

Suppose that σ^2 is known. Then to maximize the log-likelihood we would choose the estimate of μ that minimizes

$$\sum_{t=1}^{T} a_t^2 = \sum_{t=1}^{T} (y_t - \mu)^2$$

Note that this is just the error sum of squares from the model in Eq. (3.101). So, in the case of normally distributed errors, the maximum likelihood estimator of μ is identical to the least squares estimator of μ. It is easy to show that this estimator is just the sample average; that is,

$$\hat{\mu} = \bar{y}$$

Suppose that the mean of the model in Eq. (3.101) is a linear regression function of time, say,

$$\mu = \beta_0 + \beta_1 t$$

so that the model is

$$y_t = \mu + a_t = \beta_0 + \beta_1 t + a_t$$

with independent and normally distributed errors. The likelihood function for this model is identical to Eq. (3.102), and, once again, the maximum likelihood estimators of the model parameters β_0 and β_1 are found by minimizing the error sum of squares from the model. Thus when the errors are normally and independently distributed, the maximum likelihood estimators of the model parameters β_0 and β_1 in the linear regression model are identical to the least squares estimators.

Now let us consider the simple linear regression model with first-order autoregressive errors, first introduced in Eq. (3.94), and repeated for convenience below:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t = \phi\varepsilon_{t-1} + a_t$$

Recall that the a's are normally and independently distributed with mean zero and variance σ_a^2 and ϕ is the autocorrelation parameter. Write this equation for y_{t-1} and subtract ϕy_{t-1} from y_t. This results in

$$y_t - \phi y_{t-1} = (1 - \phi)\beta_0 + \beta_1(x_t - \phi x_{t-1}) + a_t$$

or

$$\begin{aligned} y_t &= \phi y_{t-1} + (1 - \phi)\beta_0 + \beta_1(x_t - \phi x_{t-1}) + a_t \\ &= \mu(\mathbf{z}_t, \boldsymbol{\theta}) + a_t, \end{aligned} \tag{3.103}$$

where $\mathbf{z}'_t = [y_{t-1}, x_t]$ and $\boldsymbol{\theta}' = [\phi, \beta_0, \beta_1]$. We can think of $\mathbf{z}_t$ as a vector of predictor variables and $\boldsymbol{\theta}$ as the vector of regression model parameters. Since y_{t-1} appears on the right-hand side of the model in Eq. (3.103), the index of time must run from 2, 3, … , T. At time period $t = 2$, we treat y_1 as an observed predictor.

Because the a's are normally and independently distributed, the joint probability density of the a's is

$$\begin{aligned} f(a_2, a_3, \dots, a_T) &= \prod_{t=2}^{T} \frac{1}{\sigma_a\sqrt{2\pi}} e^{-(a_t/\sigma_a)^2/2} \\ &= \left(\frac{1}{\sigma_a\sqrt{2\pi}}\right)^{T-1} \exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=1}^{T} a_t^2\right) \end{aligned}$$

and the likelihood function is obtained from this joint distribution by substituting for the a's:

$$\begin{aligned} l(y_t, \phi, \beta_0, \beta_1) = &\left(\frac{1}{\sigma_a\sqrt{2\pi}}\right)^{T-1} \\ &\exp\left(-\frac{1}{2\sigma_a^2}\sum_{t=2}^{T}\{y_t - [\phi y_{t-1} + (1 - \phi)\beta_0 + \beta_1(x_t - \phi x_{t-1})]\}^2\right) \end{aligned}$$

The log-likelihood is

$$\begin{aligned} \ln l(y_t, \phi, \beta_0, \beta_1) = &-\frac{T-1}{2}\ln(2\pi) - (T-1)\ln\sigma_a \\ &- \frac{1}{2\sigma_a^2}\sum_{t=2}^{T}\{y_t - [\phi y_{t-1} + (1 - \phi)\beta_0 + \beta_1(x_t - \phi x_{t-1})]\}^2 \end{aligned}$$

This log-likelihood is maximized with respect to the parameters ϕ, β_0, and β_1 by minimizing the quantity

$$SS_{\mathrm{E}} = \sum_{t=2}^{T} \{y_t - [\phi y_{t-1} + (1-\phi)\beta_0 + \beta_1(x_t - \phi x_{t-1})]\}^2, \quad (3.104)$$

which is the error sum of squares for the model. Therefore the maximum likelihood estimators of ϕ, β_0, and β_1 are also least squares estimators.

There are two important points about the maximum likelihood (or least squares) estimators. First, the sum of squares in Eq. (3.104) is conditional on the initial value of the time series, y_1. Therefore the maximum likelihood (or least squares) estimators found by minimizing this conditional sum of squares are conditional maximum likelihood (or conditional least squares) estimators. Second, because the model involves products of the parameters ϕ and β_0, the model is no longer linear in the unknown parameters. That is, it is not a linear regression model and consequently we cannot give an explicit closed-form solution for the parameter estimators. Iterative methods for fitting nonlinear regression models must be used. These procedures work by linearizing the model about a set of initial guesses for the parameters, solving the linearized model to obtain improved parameter estimates, then using the improved estimates to define a new linearized model, which leads to new parameter estimates and so on. The details of fitting nonlinear models by least squares are discussed in Montgomery, Peck, and Vining (2012).

Suppose that we have obtained a set of parameter estimates, say, $\hat{\boldsymbol{\theta}}' = [\hat{\phi}, \hat{\beta}_0, \hat{\beta}_1]$. The maximum likelihood estimate of σ_a^2 is computed as

$$\hat{\sigma}_a^2 = \frac{SS_{\mathrm{E}}(\hat{\boldsymbol{\theta}})}{n-1}, \quad (3.105)$$

where $SS_{\mathrm{E}}(\hat{\boldsymbol{\theta}})$ is the error sum of squares in Eq. (3.104) evaluated at the conditional maximum likelihood (or conditional least squares) parameter estimates $\hat{\boldsymbol{\theta}}' = [\hat{\phi}, \hat{\beta}_0, \hat{\beta}_1]$. Some authors (and computer programs) use an adjusted number of degrees of freedom in the denominator to account for the number of parameters that have been estimated. If there are k predictors, then the number of estimated parameters will be $p = k + 3$, and the formula for estimating σ_a^2 is

$$\hat{\sigma}_a^2 = \frac{SS_{\mathrm{E}}(\hat{\boldsymbol{\theta}})}{n-p-1} = \frac{SS_{\mathrm{E}}(\hat{\boldsymbol{\theta}})}{n-k-4} \quad (3.106)$$

In order to test hypotheses about the model parameters and to find confidence intervals, standard errors of the model parameters are needed. The standard errors are usually found by expanding the nonlinear model in a first-order Taylor series around the final estimates of the parameters $\hat{\boldsymbol{\theta}}' = [\hat{\phi}, \hat{\beta}_0, \hat{\beta}_1]$. This results in

$$y_t \approx \mu(\mathbf{z}_t, \hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \left. \frac{\partial \mu(\mathbf{z}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\theta = \hat{\theta}} + a_t$$

The column vector of derivatives, $\partial \mu(\mathbf{z}_t, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, is found by differentiating the model with respect to each parameter in the vector $\boldsymbol{\theta}' = [\phi, \beta_0, \beta_1]$. This vector of derivatives is

$$\frac{\partial \mu(\mathbf{z}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} 1 - \phi \\ x_t - \phi x_{t-1} \\ y_{t-1} - \beta_0 - \beta_1 x_{t-1} \end{bmatrix}$$

This vector is evaluated for each observation at the set of conditional maximum likelihood parameter estimates $\hat{\boldsymbol{\theta}}' = [\hat{\phi}, \hat{\beta}_0, \hat{\beta}_1]$ and assembled into an $\mathbf{X}$ matrix. Then the covariance matrix of the parameter estimates is found from

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma_a^2 (\mathbf{X}'\mathbf{X})^{-1}$$

When σ_a^2 is replaced by the estimate $\hat{\sigma}_a^2$ from Eq. (3.106) an estimate of the covariance matrix results, and the standard errors of the model parameters are the main diagonals of the covariance matrix.

Example 3.15 We will fit the regression model with time series errors in Eq. (3.104) to the toothpaste market share data originally analyzed in Example 3.14. We will use a widely available software package, SAS (the Statistical Analysis System). The SAS procedure for fitting regression models with time series errors is SAS PROC AUTOREG. Table 3.21 contains the output from this software program for the toothpaste market share data. Note that the autocorrelation parameter (or the lag one autocorrelation) is estimated to be 0.4094, which is very similar to the value obtained by the Cochrane–Orcutt method. The overall R^2 for this model is 0.9601, and we can show that the residuals exhibit no autocorrelative structure, so this is likely a reasonable model for the data.

There is, of course, some possibility that a more complex autocorrelation structure than first-order may exist. SAS PROC AUTOREG can fit

TABLE 3.21 SAS PROC AUTOREG Output for the Toothpaste Market Share Data, Assuming First-Order Autoregressive Errors

```
                                        The SAS System

                                     The AUTOREG Procedure

                                    Dependent Variable    y


                                Ordinary Least Squares Estimates

SSE                  3.30825739    DFE                     18
MSE                     0.18379    Root MSE           0.42871
SBC                   26.762792    AIC             24.7713275
Regress R-Square         0.9511    Total R-Square      0.9511
Durbin-Watson            1.1358    Pr < DW             0.0098
Pr > DW                  0.9902
NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is
the p-value for testing negative autocorrelation.

Standard                                                          Approx    Variable
Variable        DF      Estimate         Error     t Value     Pr > |t|     Label

Intercept        1       26.9099        1.1099       24.25       <.0001
x                1      -24.2898        1.2978      -18.72       <.0001    x

                                 Estimates of Autocorrelations

Lag    Covariance     Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0         0.1654        1.000000    |                    |********************|
1         0.0677        0.409437    |                    |********            |

                                  Preliminary MSE      0.1377

                             Estimates of Autoregressive Parameters

Standard
Lag     Coefficient           Error    t Value

1       -0.409437        0.221275      -1.85

Algorithm converged.
```

more complex patterns. Since there is obviously first-order autocorrelation present, an obvious possibility is that the autocorrelation might be second-order autoregressive, as in

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + a_t,$$

where the parameters ϕ_1 and ϕ_2 are autocorrelations at lags one and two, respectively. The output from SAS PROC AUTOREG for this model is in Table 3.22. The t-statistic for the lag two autocorrelation is not significant so there is no reason to believe that this more complex autocorrelative

TABLE 3.21 (*Continued*)

```
                                 The SAS System

                              The AUTOREG Procedure

                          Maximum Likelihood Estimates


SSE                  2.69864377     DFE                      17
MSE                     0.15874     Root MSE            0.39843
SBC                  25.8919447     AIC              22.9047479
Regress R-Square         0.9170     Total R-Square       0.9601
Durbin-Watson            1.8924     Pr < DW              0.3472
Pr > DW                  0.6528
NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is
the p-value for testing negative autocorrelation.

Standard                                                    Approx    Variable
Variable        DF     Estimate        Error    t Value    Pr > |t|    Label

Intercept        1      26.3322       1.4777      17.82      <.0001
x                1     -23.5903       1.7222     -13.70      <.0001    x
AR1              1      -0.4323       0.2203      -1.96      0.0663

                    Autoregressive parameters assumed given.

Standard                Approx    Variable
Variable        DF     Estimate        Error    t Value    Pr > |t|    Label


Intercept        1      26.3322       1.4776      17.82      <.0001
x                1     -23.5903       1.7218     -13.70      <.0001    x
```

structure is necessary to adequately model the data. The model with first-order autoregressive errors is satisfactory.

Forecasting and Prediction Intervals We now consider how to obtain forecasts at any lead time using a time series model. It is very tempting to ignore the autocorrelation in the data when forecasting, and simply substitute the conditional maximum likelihood estimates into the regression equation:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

Now suppose that we are at the end of the current time period, T, and we wish to obtain a forecast for period $T + 1$. Using the above equation, this results in

$$\hat{y}_{T+1}(T) = \hat{\beta}_0 + \hat{\beta}_1 x_{T+1},$$

assuming that the value of the predictor variable in the next time period x_{T+1} is known.

TABLE 3.22 SAS PROC AUTOREG Output for the Toothpaste Market Share Data, Assuming Second-Order Autoregressive Errors

```
                                   The SAS System

                                The AUTOREG Procedure

                               Dependent Variable    y
                                                     y


                          Ordinary Least Squares Estimates

SSE                   3.30825739    DFE                       18
MSE                      0.18379    Root MSE             0.42871
SBC                    26.762792    AIC               24.7713275
Regress R-Square          0.9511    Total R-Square        0.9511
Durbin-Watson             1.1358    Pr < DW               0.0098
Pr > DW                   0.9902
NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is
the p-value for testing negative autocorrelation.


Standard                  Approx    Variable
Variable        DF      Estimate       Error    t Value    Pr > |t|    Label

Intercept        1       26.9099      1.1099      24.25      <.0001
x                1      -24.2898      1.2978     -18.72      <.0001    x


                           Estimates of Autocorrelations

Lag    Covariance     Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

0        0.1654       1.000000     |                    |********************|
1        0.0677       0.409437     |                    |********            |
2        0.0223       0.134686     |                    |***                 |


                            Preliminary MSE      0.1375


                      Estimates of Autoregressive Parameters


Standard
Lag     Coefficient           Error    t Value

1       -0.425646        0.249804      -1.70
2        0.039590        0.249804       0.16

Algorithm converged.
```

TABLE 3.22 (*Continued*)

```
                                 The SAS System

                             The AUTOREG Procedure

                          Maximum Likelihood Estimates

SSE                  2.69583958    DFE                       16
MSE                     0.16849    Root MSE             0.41048
SBC                  28.8691217    AIC               24.8861926
Regress R-Square         0.9191    Total R-Square        0.9602
Durbin-Watson            1.9168    Pr < DW               0.3732
Pr > DW                  0.6268
NOTE: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is
the p-value for testing negative autocorrelation.

Standard                 Approx    Variable
Variable         DF    Estimate       Error    t Value    Pr > |t|    Label

Intercept         1     26.3406      1.5493      17.00      <.0001
x                 1    -23.6025      1.8047     -13.08      <.0001    x
AR1               1     -0.4456      0.2562      -1.74      0.1012
AR2               1      0.0297      0.2617       0.11      0.9110

                         Autoregressive parameters assumed given.

Standard                 Approx    Variable
Variable         DF    Estimate       Error    t Value    Pr > |t|    Label

Intercept         1     26.3406      1.5016      17.54      <.0001
x                 1    -23.6025      1.7502     -13.49      <.0001    x
```

Unfortunately, this naive approach is not correct. From Eq. (3.103), we know that the observation at time period t is

$$y_t = \phi y_{t-1} + (1-\phi)\beta_0 + \beta_1(x_t - \phi x_{t-1}) + a_t \tag{3.107}$$

So at the end of the current time period T the next observation is

$$y_{T+1} = \phi y_T + (1-\phi)\beta_0 + \beta_1(x_{T+1} - \phi x_T) + a_{T+1}$$

Assume that the future value of the regressor variable x_{T+1} is known. Obviously, at the end of the current time period, both y_T and x_T are known. The random error at time T+1, a_{T+1}, has not been observed yet, and because we have assumed that the expected value of the errors is zero, the best estimate we can make of a_{T+1} is $a_{T+1} = 0$. This suggests that a reasonable forecast of the observation in time period $T+1$ that we can make at the end of the current time period T is

$$\hat{y}_{T+1}(T) = \hat{\phi} y_T + (1 - \hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+1} - \hat{\phi} x_T) \qquad (3.108)$$

Note that this forecast is likely to be very different from the naive forecast obtained by ignoring the autocorrelation.

To find a **prediction interval** on the forecast, we need to find the variance of the prediction error. The one-step-ahead forecast error is

$$y_{T+1} - \hat{y}_{T+1}(T) = a_{T+1},$$

assuming that all of the parameters in the forecasting model are known. The variance of the one-step-ahead forecast error is

$$\text{Var}\ (a_{T+1}) = \sigma_a^2$$

Using the variance of the one-step-ahead forecast error, we can construct a $100(1-\alpha)$ percent prediction interval for the lead-one forecast from Eq. (3.107). The PI is

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\sigma_a,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. To actually compute an interval, we must replace σ_a by an estimate, resulting in

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\hat{\sigma}_a \qquad (3.109)$$

as the PI. Because σ_a and the model parameters in the forecasting equation have been replaced by estimates, the probability level on the PI in Eq. (3.109) is only approximate.

Now suppose that we want to forecast two periods ahead assuming that we are at the end of the current time period, T. Using Eq. (3.107), we can write the observation at time period $T+2$ as

$$\begin{aligned} y_{T+2} &= \phi y_{T+1} + (1-\phi)\beta_0 + \beta_1(x_{T+2} - \phi x_{T+1}) + a_{T+2} \\ &= \phi[\phi y_T + (1-\phi)\beta_0 + \beta_1(x_{T+1} - \phi x_T) + a_{T+1}] + (1-\phi)\beta_0 \\ &\quad + \beta_1(x_{T+2} - \phi x_{T+1}) + a_{T+2} \end{aligned}$$

Assume that the future value of the regressor variables x_{T+1} and x_{T+2} are known. At the end of the current time period, both y_T and x_T are known. The random errors at time $T + 1$ and $T + 2$ have not been observed yet, and because we have assumed that the expected value of the errors is zero, the best estimate we can make of both a_{T+1} and a_{T+2} is zero. This suggests that the forecast of the observation in time period $T + 2$ made at the end of the current time period T is

$$\begin{aligned}\hat{y}_{T+2}(T) &= \hat{\phi}\hat{y}_{T+1}(T) + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+2} - \hat{\phi}x_{T+1}) \\ &= \hat{\phi}[\hat{\phi}y_T + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+1} - \hat{\phi}x_T)] \\ &\quad +(1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+2} - \hat{\phi}x_{T+1})\end{aligned} \tag{3.110}$$

The two-step-ahead forecast error is

$$y_{T+2} - \hat{y}_{T+2}(T) = a_{T+2} + \phi a_{T+1},$$

assuming that all estimated parameters are actually known. The variance of the two-step-ahead forecast error is

$$\begin{aligned}\text{Var}\ (a_{T+2} + \phi a_{T+1}) &= \sigma_a^2 + \phi^2\sigma_a^2 \\ &= (1+\phi^2)\sigma_a^2\end{aligned}$$

Using the variance of the two-step-ahead forecast error, we can construct a $100(1-\alpha)$ percent PI for the lead-one forecast from Eq. (3.107):

$$\hat{y}_{T+2}(T) \pm z_{\alpha/2}(1+\phi^2)^{1/2}\sigma_a$$

To actually compute the PI, both σ_a and ϕ must be replaced by estimates, resulting in

$$\hat{y}_{T+2}(T) \pm z_{\alpha/2}(1+\hat{\phi}^2)^{1/2}\hat{\sigma}_a \tag{3.111}$$

as the PI. Because σ_a and ϕ have been replaced by estimates, the probability level on the PI in Eq. (3.111) is only approximate.

In general, if we want to forecast τ periods ahead, the forecasting equation is

$$\hat{y}_{T+\tau}(T) = \hat{\phi}\hat{y}_{T+\tau-1}(T) + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+\tau} - \hat{\phi}x_{T+\tau-1}) \tag{3.112}$$

The τ-step-ahead forecast error is (assuming that the estimated model parameters are known)

$$y_{T+\tau} - \hat{y}_{T+\tau}(T) = a_{T+\tau} + \phi a_{T+\tau-1} + \cdots + \phi^{\tau-1} a_{T+1}$$

and the variance of the τ-step-ahead forecast error is

$$\begin{aligned} V(a_{T+\tau} + \phi a_{T+\tau-1} + \cdots + \phi^{\tau-1} a_{T+1}) &= (1 + \phi^2 + \cdots + \phi^{2(\tau-1)})\sigma_a^2 \\ &= \frac{1-\phi^{2\tau}}{1+\phi^2}\sigma_a^2 \end{aligned}$$

A $100(1-\alpha)$ percent PI for the lead-τ forecast from Eq. (3.112) is

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2}\left(\frac{1-\phi^{2\tau}}{1+\phi^2}\right)^{1/2}\sigma_a$$

Replacing σ_a and ϕ by estimates, the approximate $100(1-\alpha)$ percent PI is actually computed from

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2}\left(\frac{1-\hat{\phi}^{2\tau}}{1+\hat{\phi}^2}\right)^{1/2}\hat{\sigma}_a \tag{3.113}$$

The Case Where the Predictor Variable Must Also Be Forecast

In the preceding discussion, we assumed that in order to make forecasts, any necessary values of the predictor variable in future time periods $T + \tau$ are known. This is often (probably usually) an unrealistic assumption. For example, if you are trying to forecast how many new vehicles will be registered in the state of Arizona in some future year $T + \tau$ as a function of the state population in year $T + \tau$, it is pretty unlikely that you will know the state population in that future year.

A straightforward solution to this problem is to replace the required future values of the predictor variable in future time periods $T + \tau$ by forecasts of these values. For example, suppose that we are forecasting one period ahead. From Eq. (3.108) we know that the forecast for y_{T+1} is

$$\hat{y}_{T+1}(T) = \hat{\phi} y_T + (1-\hat{\phi})\hat{\beta}_0 + \hat{\beta}_1(x_{T+1} - \hat{\phi} x_T)$$

But the future value of x_{T+1} is not known. Let $\hat{x}_{T+1}(T)$ be an unbiased forecast of x_{T+1}, made at the end of the current time period T. Now the forecast for y_{T+1} is

$$\hat{y}_{T+1}(T) = \hat{\phi} y_T + (1 - \hat{\phi})\hat{\beta}_0 + \hat{\beta}_1[\hat{x}_{T+1}(T) - \hat{\phi} x_T] \qquad (3.114)$$

If we assume that the model parameters are known, the one-step-ahead forecast error is

$$y_{T+1} - \hat{y}_{T+1}(T) = a_{T+1} + \beta_1[x_{T+1} - \hat{x}_{T+1}(T)]$$

and the variance of this forecast error is

$$\text{Var}\,(a_{T+1}) = \sigma_a^2 + \beta_1^2 \sigma_x^2(1), \qquad (3.115)$$

where $\sigma_x^2(1)$ is the variance of the one-step-ahead forecast error for the predictor variable x and we have assumed that the random error a_{T+1} in period $T+1$ is independent of the error in forecasting the predictor variable. Using the variance of the one-step-ahead forecast error from Eq. (3.115), we can construct a $100(1 - \alpha)$ percent prediction interval for the lead-one forecast from Eq. (3.114). The PI is

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\left[\sigma_a^2 + \beta_1^2 \sigma_x^2(1)\right]^{1/2},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. To actually compute an interval, we must replace the parameters β_1, σ_a^2, and $\sigma_x^2(1)$ by estimates, resulting in

$$\hat{y}_{T+1}(T) \pm z_{\alpha/2}\left[\hat{\sigma}_a^2 + \hat{\beta}_1^2 \hat{\sigma}_x^2(1)\right]^{1/2} \qquad (3.116)$$

as the PI. Because the parameters have been replaced by estimates, the probability level on the PI in Eq. (3.116) is only approximate.

In general, if we want to forecast τ periods ahead, the forecasting equation is

$$\hat{y}_{T+\tau}(T) = \hat{\phi}\hat{y}_{T+\tau-1}(T) + (1 - \hat{\phi})\hat{\beta}_0 + \hat{\beta}_1[\hat{x}_{T+\tau}(T) - \hat{\phi}\hat{x}_{T+\tau-1}(T)] \quad (3.117)$$

The τ-step-ahead forecast error is, assuming that the model parameters are known,

$$y_{T+\tau} - \hat{y}_{T+\tau}(T) = a_{T+\tau} + \phi a_{T+\tau-1} + \cdots + \phi^{\tau-1} a_{T+1} + \beta_1[x_{T+\tau} - \hat{x}_{T+\tau}(T)]$$

and the variance of the τ-step-ahead forecast error is

$$\begin{aligned}
&\text{Var}\,(a_{T+\tau} + \phi a_{T+\tau-1} + \cdots + \phi^{\tau-1} a_{T+1} + \beta_1[x_{T+\tau} - \hat{x}_{T+\tau}(t)]) \\
&= (1 + \phi^2 + \cdots + \phi^{2(\tau-1)})\sigma_a^2 + \beta_1^2\sigma_x^2(\tau) \\
&= \frac{1-\phi^{2\tau}}{1+\phi^2}\sigma_a^2 + \beta_1^2\sigma_x^2(\tau),
\end{aligned}$$

where $\sigma_x^2(\tau)$ is the variance of the τ-step-ahead forecast error for the predictor variable x. A 100(1 − α) percent PI for the lead-τ forecast from Eq. (3.117) is

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2}\left(\frac{1-\phi^{2\tau}}{1+\phi^2}\sigma_a^2 + \beta_1^2\sigma_x^2(\tau)\right)^{1/2}$$

Replacing all of the unknown parameters by estimates, the approximate 100(1 − α) percent PI is actually computed from

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2}\left(\frac{1-\hat{\phi}^{2\tau}}{1+\hat{\phi}^2}\hat{\sigma}_a^2 + \hat{\beta}_1^2\hat{\sigma}_x^2(\tau)\right)^{1/2} \qquad (3.118)$$

Alternate Forms of the Model The regression model with autocorrelated errors

$$y_t = \phi y_{t-1} + (1-\phi)\beta_0 + \beta_1(x_t - \phi x_{t-1}) + a_t$$

is a very useful model for forecasting time series regression data. However, when using this model there are two alternatives that should be considered. The first of these is

$$y_t = \phi y_{t-1} + \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + a_t \qquad (3.119)$$

This model removes the requirement that the regression coefficient for the lagged predictor variable x_{t-1} be equal to $-\beta_1\phi$. An advantage of this model is that it can be fit by OLS. Another alternative model to consider is to simply drop the lagged value of the predictor variable from Eq. (3.119), resulting in

$$y_t = \phi y_{t-1} + \beta_0 + \beta_1 x_t + a_t \qquad (3.120)$$

Often just including the lagged value of the response variable is sufficient and Eq. (3.120) will be satisfactory.

The choice between models should always be a data-driven decision. The different models can be fit to the available data, and model selection can be based on the criteria that we have discussed previously, such as model adequacy checking and residual analysis, and (if enough data are available to do some data splitting) forecasting performance over a test or trial period of data.

Example 3.16 Reconsider the toothpaste market share data originally presented in Example 3.14 and modeled with a time series regression model with first-order autoregressive errors in Example 3.15. First we will try fitting the model in Eq. (3.119). This model simply relaxes the restriction that the regression coefficient for the lagged predictor variable x_{t-1} (price in this example) be equal to $-\beta_1\phi$. Since this is just a linear regression model, we can fit it using Minitab. Table 3.23 contains the Minitab results.

This model is a good fit to the data. The Durbin–Watson statistic is $d = 2.04203$, which indicates no problems with autocorrelation in the residuals. However, note that the t-statistic for the lagged predictor variable (price) is not significant ($P = 0.217$), indicating that this variable could be removed from the model. If x_{t-1} is removed, the model becomes the one in Eq. (3.120). The Minitab output for this model is in Table 3.24.

This model is also a good fit to the data. Both predictors, the lagged variable y_{t-1} and x_t, are significant. The Durbin–Watson statistic does not indicate any significant problems with autocorrelation. It seems that either of these models would be reasonable for the toothpaste market share data. The advantage of these models relative to the time series regression model with autocorrelated errors is that they can be fit by OLS. In this example, including a lagged response variable and a lagged predictor variable has essentially eliminated any problems with autocorrelated errors.

3.9 ECONOMETRIC MODELS

The field of **econometrics** involves the unified study of economics, economic data, mathematics, and statistical models. The term econometrics is generally credited to the Norwegian economist Ragnar Frisch (1895–1973) who was one of the founders of the Econometric Society and the founding editor of the important journal *Econometrica* in 1933. Frisch was a co-winner of the first Nobel Prize in Economic Sciences in 1969. For

TABLE 3.23 Minitab Results for Fitting Model (3.119) to the Toothpaste Market Share Data

Regression Analysis: y Versus y_{t-1}, x, x_{t-1}

```
The regression equation is
y = 16.1 + 0.425 y(t-1) - 22.2 x + 7.56 x(t-1)


Predictor     Coef  SE Coef      T      P
Constant    16.100    6.095   2.64  0.019
y(t-1)      0.4253   0.2239   1.90  0.077
x          -22.250    2.488  -8.94  0.000
x(t-1)       7.562    5.872   1.29  0.217

S = 0.402205   R-Sq = 96.0%   R-Sq(adj) = 95.2%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       3  58.225  19.408  119.97  0.000
Residual Error  15   2.427   0.162
Total           18  60.651

Source  DF  Seq SS
y(t-1)   1  44.768
x        1  13.188
x(t-1)   1   0.268

Durbin-Watson statistic = 2.04203
```

introductory books on econometrics, see Greene (2011) and Woodridge (2011).

Econometric models assume that the quantities being studied are random variables and regression modeling techniques are widely used in the field to describe the relationships between these quantities. Typically, an analyst may want to quantify the impact of one set of variables on another variable. For example, one may want to investigate the effect of education on income; that is, what is the change in earnings that result from increasing a worker's education, while holding other variables such as age and gender constant. Large-scale, comprehensive econometric models of macroeconomic relationships are used by government agencies and central banks to evaluate economic activity and to provide guidance on economic

TABLE 3.24 Minitab Results for Fitting Model (3.120) to the Toothpaste Market Share Data

Regression Analysis: y Versus y_{t-1}, x

```
The regression equation is
y = 23.3 + 0.162 y(t-1) - 21.2 x

Predictor      Coef  SE Coef      T      P
Constant     23.279    2.515   9.26  0.000
y(t-1)      0.16172  0.09238   1.75  0.099
x           -21.181    2.394  -8.85  0.000

S = 0.410394   R-Sq = 95.6%   R-Sq(adj) = 95.0%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  57.956  28.978  172.06  0.000
Residual Error  16   2.695   0.168
Total           18  60.651

Source  DF  Seq SS
y(t-1)   1  44.768
x        1  13.188

Durbin-Watson statistic = 1.61416
```

policies. For example, the United States Federal Reserve Bank has maintained macroeconometric models for forecasting and quantitative policy and macroeconomic analysis for over 40 years. The Fed focuses on both the US economy and the global economy.

There are several types of data used in econometric modeling. **Time-series data** are used in many applications. Typical examples include aggregates of economic quantities, such as GDP, asset or commodity prices, and interest rates. As we have discussed earlier in this chapter, time series such as these are characterized by serial correlation. A lot of aggregate economic data are only available at a relatively low sampling frequency, such as monthly, quarterly, or in some cases annually. One exception is financial data, which may be available at very high frequency, such as hourly, daily, or even by individual transaction. **Cross-sectional data** consist of observations taken at the same point in time. In econometric work, surveys are a typical source of cross-sectional data. In typical applications, the surveys