Jonathan D. Cryer • Kung-Sik Chan

Time Series Analysis

With Applications in R

Second Edition

Springer

CONTENTS

CHAPTER 1

INTRODUCTION

Data obtained from observations collected sequentially over time are extremely common. In business, we observe weekly interest rates, daily closing stock prices, monthly price indices, yearly sales figures, and so forth. In meteorology, we observe daily high and low temperatures, annual precipitation and drought indices, and hourly wind speeds. In agriculture, we record annual figures for crop and livestock production, soil erosion, and export sales. In the biological sciences, we observe the electrical activity of the heart at millisecond intervals. In ecology, we record the abundance of an animal species. The list of areas in which time series are studied is virtually endless. The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanism that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series and, possibly, other related series or factors.

This chapter will introduce a variety of examples of time series from diverse areas of application. A somewhat unique feature of time series and their models is that we usually cannot assume that the observations arise independently from a common population (or from populations with different means, for example). Studying models that incorporate dependence is the key concept in time series analysis.

1.1 Examples of Time Series

In this section, we introduce a number of examples that will be pursued in later chapters.

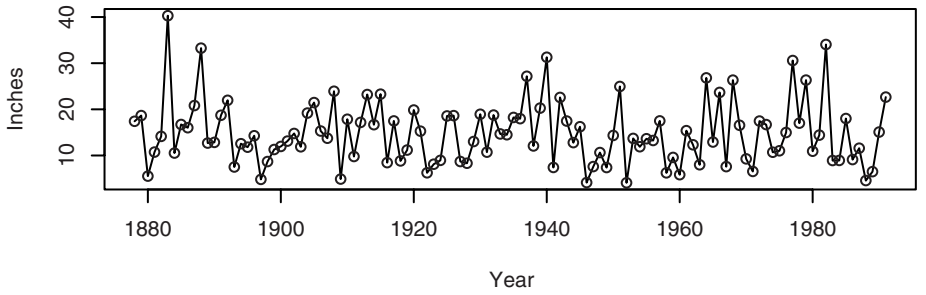
Annual Rainfall in Los Angeles

Exhibit 1.1 displays a time series plot of the annual rainfall amounts recorded in Los Angeles, California, over more than 100 years. The plot shows considerable variation in rainfall amount over the years—some years are low, some high, and many are in-between in value. The year 1883 was an exceptionally wet year for Los Angeles, while 1983 was quite dry. For analysis and modeling purposes we are interested in whether or not consecutive years are related in some way. If so, we might be able to use one year's rainfall value to help forecast next year's rainfall amount. One graphical way to investigate that question is to pair up consecutive rainfall values and plot the resulting scatterplot of pairs.

Exhibit 1.2 shows such a scatterplot for rainfall. For example, the point plotted near the lower right-hand corner shows that the year of extremely high rainfall, 40 inches in 1883, was followed by a middle of the road amount (about 12 inches) in 1884. The point

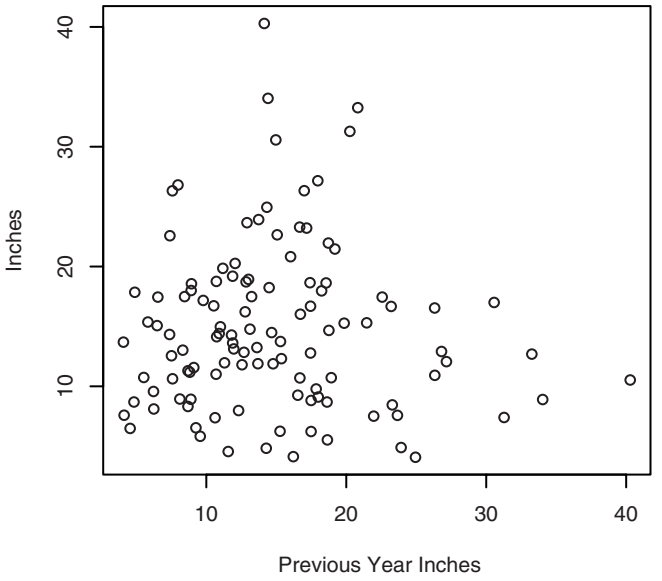
near the top of the display shows that the 40 inch year was preceded by a much more typical year of about 15 inches.

Exhibit 1.1 Time Series Plot of Los Angeles Annual Rainfall




```
> library(TSA)
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(larain); plot(larain,ylab='Inches',xlab='Year',type='o')
```

Exhibit 1.2 Scatterplot of LA Rainfall versus Last Year's LA Rainfall



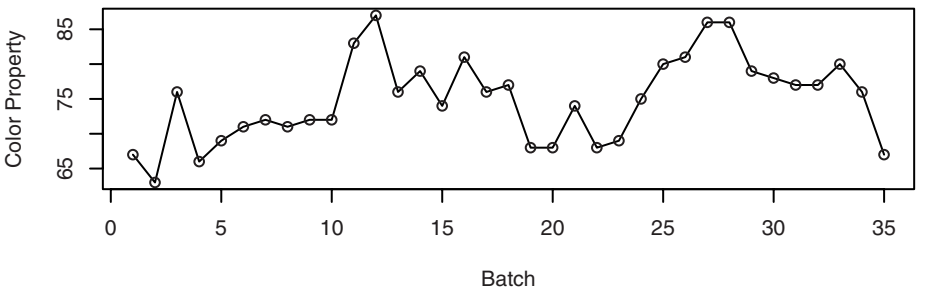

```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=larain,x=zlag(larain),ylab='Inches',
    xlab='Previous Year Inches')
```

The main impression that we obtain from this plot is that there is little if any information about this year's rainfall amount from last year's amount. The plot shows no "trends" and no general tendencies. There is little correlation between last year's rainfall amount and this year's amount. From a modeling or forecasting point of view, this is not a very interesting time series!

An Industrial Chemical Process

As a second example, we consider a time series from an industrial chemical process. The variable measured here is a color property from consecutive batches in the process. Exhibit 1.3 shows a time series plot of these color values. Here values that are neighbors in time tend to be similar in size. It seems that neighbors are related to one another.

Exhibit 1.3 Time Series Plot of Color Property from a Chemical Process

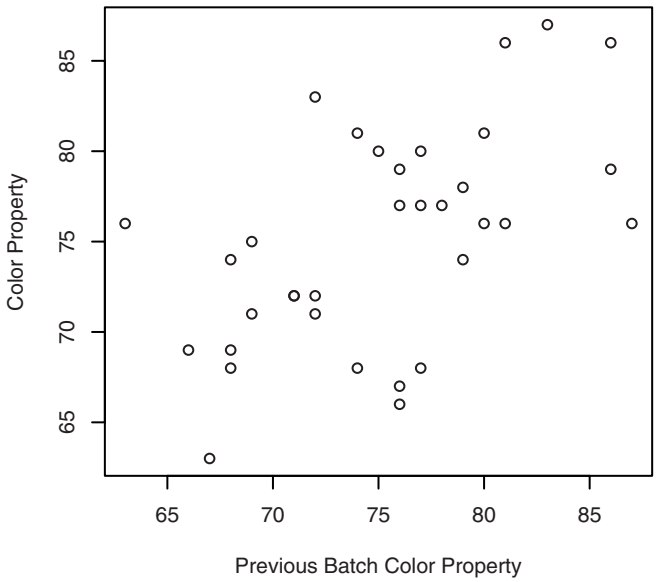



```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(color)
> plot(color,ylab='Color Property',xlab='Batch',type='o')
```

This can be seen better by constructing the scatterplot of neighboring pairs as we did with the first example.

Exhibit 1.4 displays the scatterplot of the neighboring pairs of color values. We see a slight upward trend in this plot—low values tend to be followed in the next batch by low values, middle-sized values tend to be followed by middle-sized values, and high values tend to be followed by high values. The trend is apparent but is not terribly strong. For example, the correlation in this scatterplot is about 0.6.

Exhibit 1.4 Scatterplot of Color Value versus Previous Color Value

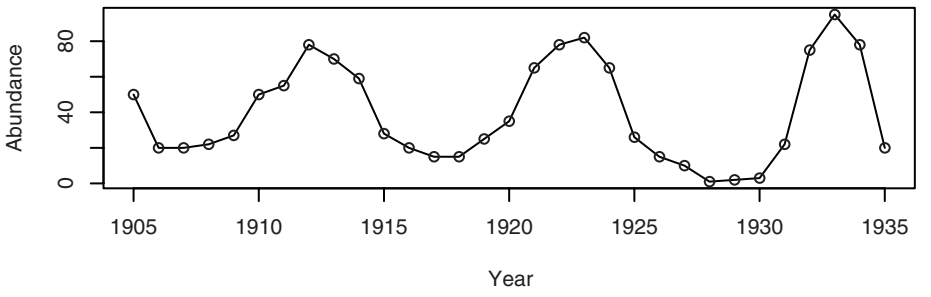



```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=color,x=zlag(color),ylab='Color Property',
    xlab='Previous Batch Color Property')
```

Annual Abundance of Canadian Hare

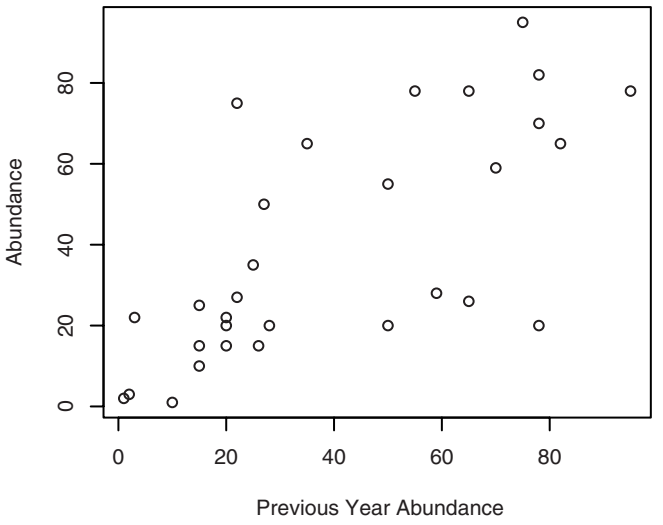
Our third example concerns the annual abundance of Canadian hare. Exhibit 1.5 gives the time series plot of this abundance over about 30 years. Neighboring values here are very closely related. Large changes in abundance do not occur from one year to the next. This neighboring correlation is seen clearly in Exhibit 1.6 where we have plotted abundance versus the previous year's abundance. As in the previous example, we see an upward trend in the plot—low values tend to be followed by low values in the next year, middle-sized values by middle-sized values, and high values by high values.

Exhibit 1.5 Abundance of Canadian Hare




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(hare); plot(hare,ylab='Abundance',xlab='Year',type='o')
```

Exhibit 1.6 Hare Abundance versus Previous Year's Hare Abundance

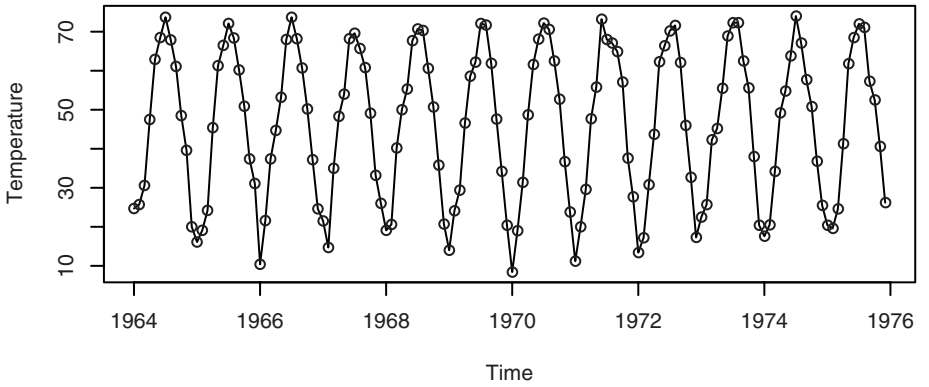



```
> win.graph(width=3, height=3,pointsize=8)
> plot(y=hare,x=zlag(hare),ylab='Abundance',
   xlab='Previous Year Abundance')
```

Monthly Average Temperatures in Dubuque, Iowa

The average monthly temperatures (in degrees Fahrenheit) over a number of years recorded in Dubuque, Iowa, are shown in Exhibit 1.7.

Exhibit 1.7 Average Monthly Temperatures, Dubuque, Iowa




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(tempdub); plot(tempdub,ylab='Temperature',type='o')
```

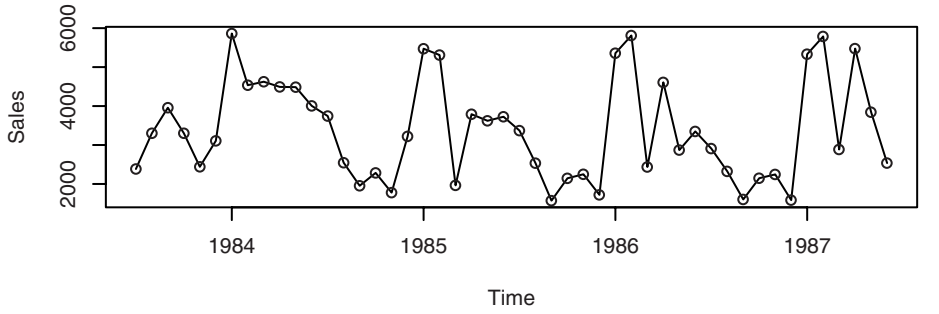
This time series displays a very regular pattern called **seasonality**. Seasonality for monthly values occurs when observations twelve months apart are related in some manner or another. All Januarys and Februarys are quite cold but they are similar in value and different from the temperatures of the warmer months of June, July, and August, for example. There is still variation among the January values and variation among the June values. Models for such series must accommodate this variation while preserving the similarities. Here the reason for the seasonality is well understood—the Northern Hemisphere's changing inclination toward the sun.

Monthly Oil Filter Sales

Our last example for this chapter concerns the monthly sales to dealers of a specialty oil filter for construction equipment manufactured by John Deere. When these data were first presented to one of the authors, the manager said, "There is no reason to believe that these sales are seasonal." Seasonality would be present if January values tended to be related to other January values, February values tended to be related to other February values, and so forth. The time series plot shown in Exhibit 1.8 is not designed to display seasonality especially well. Exhibit 1.9 gives the same plot but amended to use meaningful plotting symbols. In this plot, all January values are plotted with the character J, all Februarys with F, all Marches with M, and so forth.[†] With these plotting symbols, it is much easier to see that sales for the winter months of January and February all tend to be high, while sales in September, October, November, and December are gener-

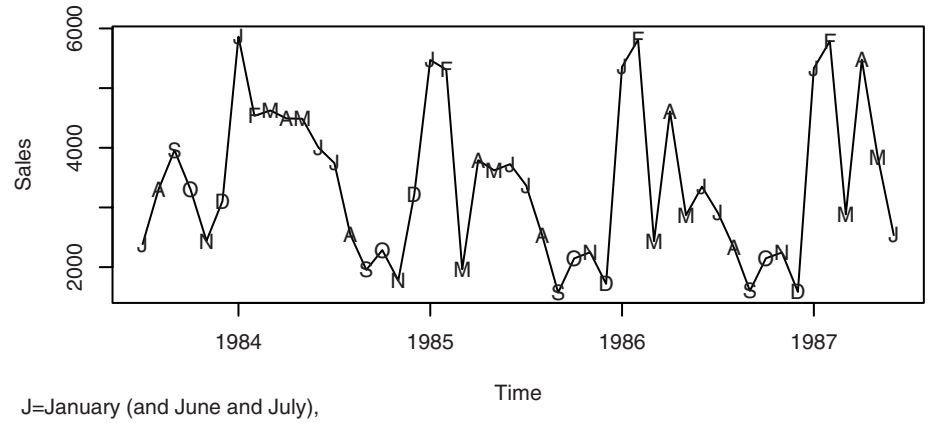
ally quite low. The seasonality in the data is much easier to see from this modified time series plot.

Exhibit 1.8 Monthly Oil Filter Sales




```
> data(oilfilters); plot(oilfilters,type='o',ylab='Sales')
```

Exhibit 1.9 Monthly Oil Filter Sales with Special Plotting Symbols



J=January (and June and July),
F=February, M=March (and May), and so forth

```
> plot(oilfilters,type='l',ylab='Sales')
> points(y=oilfilters,x=time(oilfilters),
    pch=as.vector(season(oilfilters)))
```

[†] In reading the plot, you will still have to distinguish between Januarys, Junes, and Julys, between Marches and Mays, and Aprils and Augusts, but this is easily done by looking at neighboring plotting characters.

In general, our goal is to emphasize plotting methods that are appropriate and useful for finding patterns that will lead to suitable models for our time series data. In later chapters, we will consider several different ways to incorporate seasonality into time series models.

1.2 A Model-Building Strategy

Finding appropriate models for time series is a nontrivial task. We will develop a multistep model-building strategy espoused so well by Box and Jenkins (1976). There are three main steps in the process, each of which may be used several times:

1. **model specification** (or identification)
2. **model fitting**, and
3. **model diagnostics**

In model specification (or identification), the classes of time series models are selected that may be appropriate for a given observed series. In this step we look at the time plot of the series, compute many different statistics from the data, and also apply any knowledge of the subject matter in which the data arise, such as biology, business, or ecology. It should be emphasized that the model chosen at this point is *tentative* and subject to revision later on in the analysis.

In choosing a model, we shall attempt to adhere to the **principle of parsimony**; that is, the model used should require the smallest number of parameters that will adequately represent the time series. Albert Einstein is quoted in Parzen (1982, p. 68) as remarking that "everything should be made as simple as possible but not simpler."

The model will inevitably involve one or more parameters whose values must be estimated from the observed series. Model fitting consists of finding the best possible estimates of those unknown parameters within a given model. We shall consider criteria such as least squares and maximum likelihood for estimation.

Model diagnostics is concerned with assessing the quality of the model that we have specified and estimated. How well does the model fit the data? Are the assumptions of the model reasonably well satisfied? If no inadequacies are found, the modeling may be assumed to be complete, and the model may be used, for example, to forecast future values. Otherwise, we choose another model in the light of the inadequacies found; that is, we return to the model specification step. In this way, we cycle through the three steps until, ideally, an acceptable model is found.

Because the computations required for each step in model building are intensive, we shall rely on readily available statistical software to carry out the calculations and do the plotting.

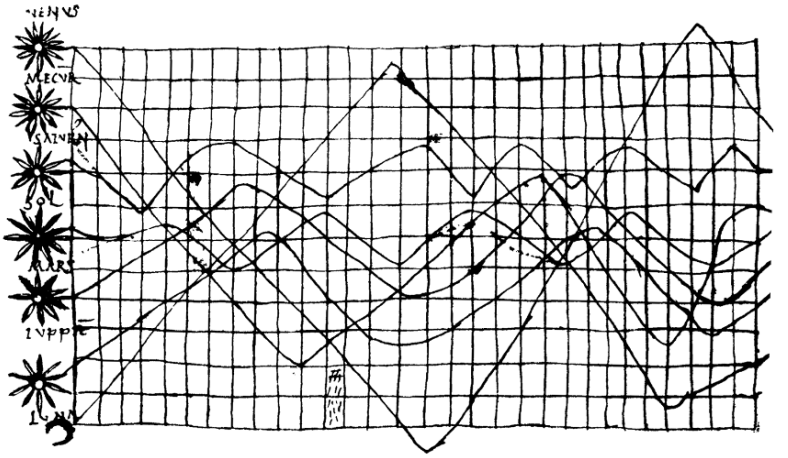
1.3 Time Series Plots in History

According toTufte (1983, p. 28), "The time-series plot is the most frequently used form of graphic design. With one dimension marching along to the regular rhythm of sec-

onds, minutes, hours, days, weeks, months, years, or millennia, the natural ordering of the time scale gives this design a strength and efficiency of interpretation found in no other graphic arrangement."

Exhibit 1.10 reproduces what appears to be the oldest known example of a time series plot, dating from the tenth (or possibly eleventh) century and showing the inclinations of the planetary orbits.† Commenting on this artifact, Tufte says "It appears as a mysterious and isolated wonder in the history of data graphics, since the next extant graphic of a plotted time-series shows up some 800 years later."

Exhibit 1.10 A Tenth-Century Time Series Plot



1.4 An Overview of the Book

Chapter 2 develops the basic ideas of mean, covariance, and correlation functions and ends with the important concept of stationarity. Chapter 3 discusses trend analysis and investigates how to estimate and check common deterministic trend models, such as those for linear time trends and seasonal means.

Chapter 4 begins the development of parametric models for stationary time series, namely the so-called autoregressive moving average (ARMA) models (also known as Box-Jenkins models). These models are then generalized in Chapter 5 to encompass certain types of stochastic nonstationary cases—the ARIMA models.

Chapters 6, 7, and 8 form the heart of the model-building strategy for ARIMA modeling. Techniques are presented for tentatively specifying models (Chapter 6), efficiently estimating the model parameters using least squares and maximum likelihood (Chapter 7), and determining how well the models fit the data (Chapter 8).

Chapter 9 thoroughly develops the theory and methods of minimum mean square error forecasting for ARIMA models. Chapter 10 extends the ideas of Chapters 4

† From Tufte (1983, p. 28).

through 9 to stochastic seasonal models. The remaining chapters cover selected topics and are of a somewhat more advanced nature.

EXERCISES

1.1 Use software to produce the time series plot shown in Exhibit 1.2, on page 2. The data are in the file named larain.[†]

1.2 Produce the time series plot displayed in Exhibit 1.3, on page 3. The data file is named color.

1.3 Simulate a completely random process of length 48 with independent, normal values. Plot the time series plot. Does it look "random"? Repeat this exercise several times with a new simulation each time.

1.4 Simulate a completely random process of length 48 with independent, chi-square distributed values, each with 2 degrees of freedom. Display the time series plot. Does it look "random" and nonnormal? Repeat this exercise several times with a new simulation each time.

1.5 Simulate a completely random process of length 48 with independent, *t*-distributed values each with 5 degrees of freedom. Construct the time series plot. Does it look "random" and nonnormal? Repeat this exercise several times with a new simulation each time.

1.6 Construct a time series plot with monthly plotting symbols for the Dubuque temperature series as in Exhibit 1.9, on page 7. The data are in the file named tempdub.

[†] If you have installed the R package TSA, available for download at www.r-project.org, the larain data are accessed by the R command: data(larain). An ASCII file of the data is also available on the book Website at www.stat.uiowa.edu/~kchan/TSA.htm.

CHAPTER 2

FUNDAMENTAL CONCEPTS

This chapter describes the fundamental concepts in the theory of time series models. In particular, we introduce the concepts of stochastic processes, mean and covariance functions, stationary processes, and autocorrelation functions.

2.1 Time Series and Stochastic Processes

The sequence of random variables $\{Y_t: t = 0, \pm1, \pm2, \pm3, \ldots\}$ is called a **stochastic process** and serves as a model for an observed time series. It is known that the complete probabilistic structure of such a process is determined by the set of distributions of all finite collections of the Y's. Fortunately, we will not have to deal explicitly with these multivariate distributions. Much of the information in these joint distributions can be described in terms of means, variances, and covariances. Consequently, we concentrate our efforts on these first and second moments. (If the joint distributions of the Y's are multivariate *normal* distributions, then the first and second moments completely determine all the joint distributions.)

2.2 Means, Variances, and Covariances

For a stochastic process $\{Y_t: t = 0, \pm1, \pm2, \pm3, \ldots\}$, the **mean function** is defined by

$$\mu_t = E(Y_t) \qquad \text{for } t = 0, \pm1, \pm2, \ldots \tag{2.2.1}$$

That is, μ_t is just the expected value of the process at time t. In general, μ_t can be different at each time point t.

The **autocovariance function**, $\gamma_{t,s}$, is defined as

$$\gamma_{t,s} = Cov(Y_t, Y_s) \qquad \text{for } t, s = 0, \pm1, \pm2, \ldots \tag{2.2.2}$$

where $Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$.

The **autocorrelation function**, $\rho_{t,s}$, is given by

$$\rho_{t,s} = Corr(Y_t, Y_s) \qquad \text{for } t, s = 0, \pm1, \pm2, \ldots \tag{2.2.3}$$

where

$$Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \tag{2.2.4}$$

We review the basic properties of expectation, variance, covariance, and correlation in Appendix A on page 24.

Recall that both covariance and correlation are measures of the (linear) dependence between random variables but that the unitless correlation is somewhat easier to interpret. The following important properties follow from known results and our definitions:

$$\left.\begin{aligned} \gamma_{t,t} &= Var(Y_t) & \rho_{t,t} &= 1 \\ \gamma_{t,s} &= \gamma_{s,t} & \rho_{t,s} &= \rho_{s,t} \\ |\gamma_{t,s}| &\le \sqrt{\gamma_{t,t}\gamma_{s,s}} & |\rho_{t,s}| &\le 1 \end{aligned}\right\} \tag{2.2.5}$$

Values of $\rho_{t,s}$ near ± 1 indicate strong (linear) dependence, whereas values near zero indicate weak (linear) dependence. If $\rho_{t,s} = 0$, we say that Y_t and Y_s are *uncorrelated*.

To investigate the covariance properties of various time series models, the following result will be used repeatedly: If $c_1, c_2, \ldots, c_m$ and $d_1, d_2, \ldots, d_n$ are constants and $t_1, t_2, \ldots, t_m$ and $s_1, s_2, \ldots, s_n$ are time points, then

$$Cov\left[\sum_{i=1}^{m} c_i Y_{t_i}, \sum_{j=1}^{n} d_j Y_{s_j}\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} c_i d_j Cov(Y_{t_i}, Y_{s_j}) \tag{2.2.6}$$

The proof of Equation (2.2.6), though tedious, is a straightforward application of the linear properties of expectation. As a special case, we obtain the well-known result

$$Var\left[\sum_{i=1}^{n} c_i Y_{t_i}\right] = \sum_{i=1}^{n} c_i^2 Var(Y_{t_i}) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} c_i c_j Cov(Y_{t_i}, Y_{t_j}) \tag{2.2.7}$$

The Random Walk

Let $e_1, e_2, \ldots$ be a sequence of independent, identically distributed random variables each with zero mean and variance σ_e^2. The observed time series, $\{Y_t: t = 1, 2, \ldots\}$, is constructed as follows:

$$\left.\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_t &= e_1 + e_2 + \cdots + e_t \end{aligned}\right\} \tag{2.2.8}$$

Alternatively, we can write

$$Y_t = Y_{t-1} + e_t \tag{2.2.9}$$

with "initial condition" $Y_1 = e_1$. If the e's are interpreted as the sizes of the "steps" taken (forward or backward) along a number line, then Y_t is the position of the "random walker" at time t. From Equation (2.2.8), we obtain the mean function

$$\mu_t = E(Y_t) = E(e_1 + e_2 + \cdots + e_t) = E(e_1) + E(e_2) + \cdots + E(e_t)$$
$$= 0 + 0 + \cdots + 0$$

so that

$$\mu_t = 0 \quad \text{for all } t \tag{2.2.10}$$

We also have

$$Var(Y_t) = Var(e_1 + e_2 + \cdots + e_t) = Var(e_1) + Var(e_2) + \cdots + Var(e_t)$$
$$= \sigma_e^2 + \sigma_e^2 + \cdots + \sigma_e^2$$

so that

$$Var(Y_t) = t\sigma_e^2 \tag{2.2.11}$$

Notice that the process variance increases linearly with time.

To investigate the covariance function, suppose that $1 \le t \le s$. Then we have

$$\gamma_{t,s} = Cov(Y_t, Y_s) = Cov(e_1 + e_2 + \cdots + e_t, e_1 + e_2 + \cdots + e_t + e_{t+1} + \cdots + e_s)$$

From Equation (2.2.6), we have

$$\gamma_{t,s} = \sum_{i=1}^{s} \sum_{j=1}^{t} Cov(e_i, e_j)$$

However, these covariances are zero unless $i = j$, in which case they equal $Var(e_i) = \sigma_e^2$. There are exactly t of these so that $\gamma_{t,s} = t\sigma_e^2$.

Since $\gamma_{t,s} = \gamma_{s,t}$, this specifies the autocovariance function for all time points t and s and we can write

$$\gamma_{t,s} = t\sigma_e^2 \qquad \text{for } 1 \le t \le s \tag{2.2.12}$$

The autocorrelation function for the random walk is now easily obtained as

$$\rho_{t,s} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \sqrt{\frac{t}{s}} \qquad \text{for } 1 \le t \le s \tag{2.2.13}$$

The following numerical values help us understand the behavior of the random walk.

$$\rho_{1,2} = \sqrt{\frac{1}{2}} = 0.707 \qquad \rho_{8,9} = \sqrt{\frac{8}{9}} = 0.943$$

$$\rho_{24,25} = \sqrt{\frac{24}{25}} = 0.980 \qquad \rho_{1,25} = \sqrt{\frac{1}{25}} = 0.200$$

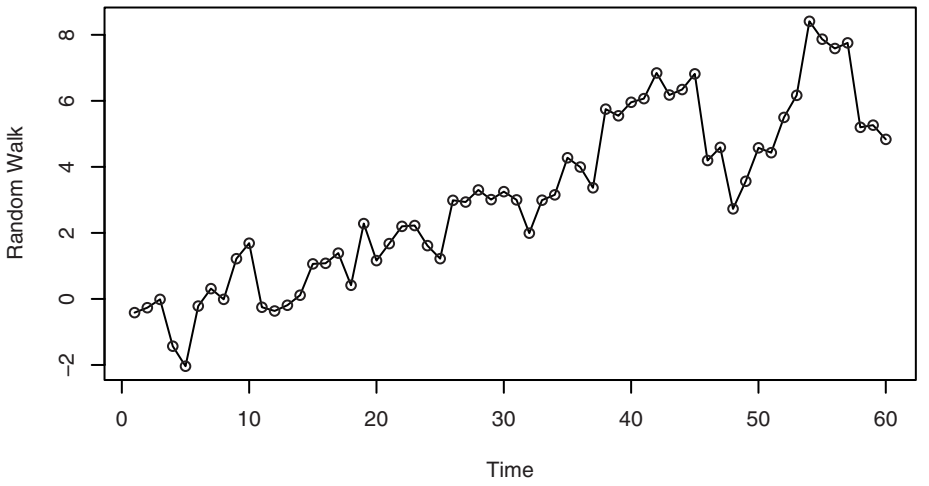
The values of Y at neighboring time points are more and more strongly and positively correlated as time goes by. On the other hand, the values of Y at distant time points are less and less correlated.

A simulated random walk is shown in Exhibit 2.1 where the e's were selected from a standard normal distribution. Note that even though the theoretical mean function is

zero for all time points, the fact that the variance increases over time and that the correlation between process values nearby in time is nearly 1 indicate that we should expect long excursions of the process away from the mean level of zero.

The simple random walk process provides a good model (at least to a first approximation) for phenomena as diverse as the movement of common stock price, and the position of small particles suspended in a fluid—so-called Brownian motion.

Exhibit 2.1 Time Series Plot of a Random Walk




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> data(rwalk) # rwalk contains a simulated random walk
> plot(rwalk,type='o',ylab='Random Walk')
```

A Moving Average

As a second example, suppose that $\{Y_t\}$ is constructed as

$$Y_t = \frac{e_t + e_{t-1}}{2} \tag{2.2.14}$$

where (as always throughout this book) the e's are assumed to be independent and identically distributed with zero mean and variance σ_e^2. Here

$$\mu_t = E(Y_t) = E\left\{\frac{e_t + e_{t-1}}{2}\right\} = \frac{E(e_t) + E(e_{t-1})}{2}$$

$$= 0$$

and

$$Var(Y_t) = Var\left\{\frac{e_t + e_{t-1}}{2}\right\} = \frac{Var(e_t) + Var(e_{t-1})}{4}$$
$$= 0.5\sigma_e^2$$

Also

$$Cov(Y_t, Y_{t-1}) = Cov\left\{\frac{e_t + e_{t-1}}{2}, \frac{e_{t-1} + e_{t-2}}{2}\right\}$$
$$= \frac{Cov(e_t, e_{t-1}) + Cov(e_t, e_{t-2}) + Cov(e_{t-1}, e_{t-1})}{4}$$
$$+ \frac{Cov(e_{t-1}, e_{t-2})}{4}$$
$$= \frac{Cov(e_{t-1}, e_{t-1})}{4} \qquad \text{(as all the other covariances are zero)}$$
$$= 0.25\sigma_e^2$$

or

$$\gamma_{t,\,t-1} = 0.25\sigma_e^2 \qquad \text{for all } t \tag{2.2.15}$$

Furthermore,

$$Cov(Y_t, Y_{t-2}) = Cov\left\{\frac{e_t + e_{t-1}}{2}, \frac{e_{t-2} + e_{t-3}}{2}\right\}$$
$$= 0 \qquad \text{since the } e'\text{s are independent.}$$

Similarly, $Cov(Y_t, Y_{t-k}) = 0$ for $k > 1$, so we may write

$$\gamma_{t,\,s} = \begin{cases} 0.5\sigma_e^2 & \text{for } |t-s| = 0 \\ 0.25\sigma_e^2 & \text{for } |t-s| = 1 \\ 0 & \text{for } |t-s| > 1 \end{cases}$$

For the autocorrelation function, we have

$$\rho_{t,\,s} = \begin{cases} 1 & \text{for } |t-s| = 0 \\ 0.5 & \text{for } |t-s| = 1 \\ 0 & \text{for } |t-s| > 1 \end{cases} \tag{2.2.16}$$

since $0.25\sigma_e^2 / 0.5\sigma_e^2 = 0.5$.

Notice that $\rho_{2,1} = \rho_{3,2} = \rho_{4,3} = \rho_{9,8} = 0.5$. Values of Y precisely one time unit apart have exactly the same correlation no matter where they occur in time. Furthermore, $\rho_{3,1} = \rho_{4,2} = \rho_{t,\,t-2}$ and, more generally, $\rho_{t,\,t-k}$ is the same for all values of t. This leads us to the important concept of stationarity.

2.3 Stationarity

To make statistical inferences about the structure of a stochastic process on the basis of an observed record of that process, we must usually make some simplifying (and presumably reasonable) assumptions about that structure. The most important such assumption is that of **stationarity**. The basic idea of stationarity is that the probability laws that govern the behavior of the process do not change over time. In a sense, the process is in statistical equilibrium. Specifically, a process $\{Y_t\}$ is said to be **strictly stationary** if the joint distribution of $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ is the same as the joint distribution of $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ for all choices of time points $t_1, t_2, \ldots, t_n$ and all choices of time lag k.

Thus, when $n = 1$ the (univariate) distribution of Y_t is the same as that of Y_{t-k} for all t and k; in other words, the Y's are (marginally) identically distributed. It then follows that $E(Y_t) = E(Y_{t-k})$ for all t and k so that the mean function is constant for all time. Additionally, $Var(Y_t) = Var(Y_{t-k})$ for all t and k so that the variance is also constant over time.

Setting $n = 2$ in the stationarity definition we see that the bivariate distribution of Y_t and Y_s must be the same as that of Y_{t-k} and Y_{s-k} from which it follows that $Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k})$ for all t, s, and k. Putting $k = s$ and then $k = t$, we obtain

$$\begin{aligned}\gamma_{t,s} &= Cov(Y_{t-s}, Y_0)\\ &= Cov(Y_0, Y_{s-t})\\ &= Cov(Y_0, Y_{|t-s|})\\ &= \gamma_{0,|t-s|}\end{aligned}$$

That is, the covariance between Y_t and Y_s depends on time only through the time difference $|t-s|$ and not otherwise on the actual times t and s. Thus, for a stationary process, we can simplify our notation and write

$$\gamma_k = Cov(Y_t, Y_{t-k}) \qquad \text{and} \qquad \rho_k = Corr(Y_t, Y_{t-k}) \tag{2.3.1}$$

Note also that

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

The general properties given in Equation (2.2.5) now become

$$\left.\begin{array}{ll}\gamma_0 = Var(Y_t) & \rho_0 = 1\\ \gamma_k = \gamma_{-k} & \rho_k = \rho_{-k}\\ |\gamma_k| \le \gamma_0 & |\rho_k| \le 1\end{array}\right\} \tag{2.3.2}$$

If a process is strictly stationary and has finite variance, then the covariance function must depend only on the time lag.

A definition that is similar to that of strict stationarity but is mathematically weaker

is the following: A stochastic process $\{Y_t\}$ is said to be **weakly** (or **second-order**) **stationary** if

1. The mean function is constant over time, and
2. $\gamma_{t,\,t-k} = \gamma_{0,\,k}$ for all time t and lag k

In this book the term stationary when used alone will always refer to this weaker form of stationarity. However, if the joint distributions for the process are all multivariate normal distributions, it can be shown that the two definitions coincide. For stationary processes, we usually only consider $k \geq 0$.

White Noise

A very important example of a stationary process is the so-called **white noise** process, which is defined as a sequence of independent, identically distributed random variables $\{e_t\}$. Its importance stems not from the fact that it is an interesting model itself but from the fact that many useful processes can be constructed from white noise. The fact that $\{e_t\}$ is strictly stationary is easy to see since

$$\begin{aligned}
&Pr(e_{t_1} \leq x_1, e_{t_2} \leq x_2, \ldots, e_{t_n} \leq x_n) \\
&= Pr(e_{t_1} \leq x_1)Pr(e_{t_2} \leq x_2)\cdots Pr(e_{t_n} \leq x_n) \qquad \text{(by independence)} \\
&= Pr(e_{t_1-k} \leq x_1)Pr(e_{t_2-k} \leq x_2)\cdots Pr(e_{t_n-k} \leq x_n) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(identical distributions)} \\
&= Pr(e_{t_1-k} \leq x_1, e_{t_2-k} \leq x_2, \ldots, e_{t_n-k} \leq x_n) \qquad \text{(by independence)}
\end{aligned}$$

as required. Also, $\mu_t = E(e_t)$ is constant and

$$\gamma_k = \begin{cases} Var(e_t) & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

Alternatively, we can write

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases} \tag{2.3.3}$$

The term white noise arises from the fact that a frequency analysis of the model shows that, in analogy with white light, all frequencies enter equally. We usually assume that the white noise process has mean zero and denote $Var(e_t)$ by σ_e^2.

The moving average example, on page 14, where $Y_t = (e_t + e_{t-1})/2$, is another example of a stationary process constructed from white noise. In our new notation, we have for the moving average process that

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ 0.5 & \text{for } |k| = 1 \\ 0 & \text{for } |k| \geq 2 \end{cases}$$

Random Cosine Wave

As a somewhat different example,† consider the process defined as follows:

$$Y_t = \cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right] \qquad \text{for } t = 0, \pm 1, \pm 2, \ldots$$

where Φ is selected (once) from a uniform distribution on the interval from 0 to 1. A sample from such a process will appear highly deterministic since Y_t will repeat itself identically every 12 time units and look like a perfect (discrete time) cosine curve. However, its maximum will not occur at $t = 0$ but will be determined by the random phase Φ. The phase Φ can be interpreted as the fraction of a complete cycle completed by time $t = 0$. Still, the statistical properties of this process can be computed as follows:

$$\begin{aligned} E(Y_t) &= E\left\{\cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right]\right\} \\ &= \int_0^1 \cos\left[2\pi\left(\frac{t}{12} + \phi\right)\right] d\phi \\ &= \frac{1}{2\pi}\sin\left[2\pi\left(\frac{t}{12} + \phi\right)\right]\Bigg|_{\phi = 0}^{1} \\ &= \frac{1}{2\pi}\left[\sin\left(2\pi\frac{t}{12} + 2\pi\right) - \sin\left(2\pi\frac{t}{12}\right)\right] \end{aligned}$$

But this is zero since the sines must agree. So $\mu_t = 0$ for all t.

Also

$$\begin{aligned} \gamma_{t,s} &= E\left\{\cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right]\cos\left[2\pi\left(\frac{s}{12} + \Phi\right)\right]\right\} \\ &= \int_0^1 \cos\left[2\pi\left(\frac{t}{12} + \phi\right)\right]\cos\left[2\pi\left(\frac{s}{12} + \phi\right)\right] d\phi \\ &= \frac{1}{2}\int_0^1 \left\{\cos\left[2\pi\left(\frac{t-s}{12}\right)\right] + \cos\left[2\pi\left(\frac{t+s}{12} + 2\phi\right)\right]\right\} d\phi \\ &= \frac{1}{2}\left\{\cos\left[2\pi\left(\frac{t-s}{12}\right)\right] + \frac{1}{4\pi}\sin\left[2\pi\left(\frac{t+s}{12} + 2\phi\right)\right]\Bigg|_{\phi = 0}^{1}\right\} \\ &= \frac{1}{2}\cos\left[2\pi\left(\frac{|t-s|}{12}\right)\right] \end{aligned}$$

† This example contains optional material that is not needed in order to understand most of the remainder of this book. It will be used in Chapter 13, Introduction to Spectral Analysis.

So the process is stationary with autocorrelation function

$$\rho_k = \cos\left(2\pi\frac{k}{12}\right) \qquad \text{for } k = 0, \pm1, \pm2, \ldots \tag{2.3.4}$$

This example suggests that it will be difficult to assess whether or not stationarity is a reasonable assumption for a given time series on the basis of the time sequence plot of the observed data.

The random walk of page 12, where $Y_t = e_1 + e_2 + \cdots + e_t$, is also constructed from white noise but is *not* stationary. For example, the variance function, $Var(Y_t) = t\sigma_e^2$, is *not* constant; furthermore, the covariance function $\gamma_{t,s} = t\sigma_e^2$ for $0 \le t \le s$ does not depend only on time lag. However, suppose that instead of analyzing $\{Y_t\}$ directly, we consider the differences of successive Y-values, denoted ∇Y_t. Then $\nabla Y_t = Y_t - Y_{t-1} = e_t$, so the *differenced series*, $\{\nabla Y_t\}$, is stationary. This represents a simple example of a technique found to be extremely useful in many applications. Clearly, many real time series cannot be reasonably modeled by stationary processes since they are not in statistical equilibrium but are evolving over time. However, we can frequently transform nonstationary series into stationary series by simple techniques such as differencing. Such techniques will be vigorously pursued in the remaining chapters.

2.4 Summary

In this chapter we have introduced the basic concepts of stochastic processes that serve as models for time series. In particular, you should now be familiar with the important concepts of mean functions, autocovariance functions, and autocorrelation functions. We illustrated these concepts with the basic processes: the random walk, white noise, a simple moving average, and a random cosine wave. Finally, the fundamental concept of stationarity introduced here will be used throughout the book.

EXERCISES

2.1 Suppose $E(X) = 2$, $Var(X) = 9$, $E(Y) = 0$, $Var(Y) = 4$, and $Corr(X,Y) = 0.25$. Find:
(a) $Var(X + Y)$.
(b) $Cov(X, X + Y)$.
(c) $Corr(X + Y, X - Y)$.

2.2 If X and Y are dependent but $Var(X) = Var(Y)$, find $Cov(X + Y, X - Y)$.

2.3 Let X have a distribution with mean μ and variance σ^2, and let $Y_t = X$ for all t.
(a) Show that $\{Y_t\}$ is strictly and weakly stationary.
(b) Find the autocovariance function for $\{Y_t\}$.
(c) Sketch a "typical" time plot of Y_t.

2.4 Let $\{e_t\}$ be a zero mean white noise process. Suppose that the observed process is $Y_t = e_t + \theta e_{t-1}$, where θ is either 3 or 1/3.
(a) Find the autocorrelation function for $\{Y_t\}$ both when $\theta = 3$ and when $\theta = 1/3$.
(b) You should have discovered that the time series is stationary regardless of the value of θ and that the autocorrelation functions are the same for $\theta = 3$ and $\theta = 1/3$. For simplicity, suppose that the process mean is known to be zero and the variance of Y_t is known to be 1. You observe the series $\{Y_t\}$ for $t = 1, 2, \ldots, n$ and suppose that you can produce good estimates of the autocorrelations ρ_k. Do you think that you could determine which value of θ is correct (3 or 1/3) based on the estimate of ρ_k? Why or why not?

2.5 Suppose $Y_t = 5 + 2t + X_t$, where $\{X_t\}$ is a zero-mean stationary series with autocovariance function γ_k.
(a) Find the mean function for $\{Y_t\}$.
(b) Find the autocovariance function for $\{Y_t\}$.
(c) Is $\{Y_t\}$ stationary? Why or why not?

2.6 Let $\{X_t\}$ be a stationary time series, and define $Y_t = \begin{cases} X_t & \text{for } t \text{ odd} \\ X_t + 3 & \text{for } t \text{ even.} \end{cases}$
(a) Show that $Cov(Y_t, Y_{t-k})$ is free of t for all lags k.
(b) Is $\{Y_t\}$ stationary?

2.7 Suppose that $\{Y_t\}$ is stationary with autocovariance function γ_k.
(a) Show that $W_t = \nabla Y_t = Y_t - Y_{t-1}$ is stationary by finding the mean and autocovariance function for $\{W_t\}$.
(b) Show that $U_t = \nabla^2 Y_t = \nabla[Y_t - Y_{t-1}] = Y_t - 2Y_{t-1} + Y_{t-2}$ is stationary. (You need not find the mean and autocovariance function for $\{U_t\}$.)

2.8 Suppose that $\{Y_t\}$ is stationary with autocovariance function γ_k. Show that for any fixed positive integer n and any constants $c_1, c_2, \ldots, c_n$, the process $\{W_t\}$ defined by $W_t = c_1 Y_t + c_2 Y_{t-1} + \cdots + c_n Y_{t-n+1}$ is stationary. (Note that Exercise 2.7 is a special case of this result.)

2.9 Suppose $Y_t = \beta_0 + \beta_1 t + X_t$, where $\{X_t\}$ is a zero-mean stationary series with autocovariance function γ_k and β_0 and β_1 are constants.
(a) Show that $\{Y_t\}$ is not stationary but that $W_t = \nabla Y_t = Y_t - Y_{t-1}$ is stationary.
(b) In general, show that if $Y_t = \mu_t + X_t$, where $\{X_t\}$ is a zero-mean stationary series and μ_t is a polynomial in t of degree d, then $\nabla^m Y_t = \nabla(\nabla^{m-1} Y_t)$ is stationary for $m \geq d$ and nonstationary for $0 \leq m < d$.

2.10 Let $\{X_t\}$ be a zero-mean, unit-variance stationary process with autocorrelation function ρ_k. Suppose that μ_t is a nonconstant function and that σ_t is a positive-valued nonconstant function. The observed series is formed as $Y_t = \mu_t + \sigma_t X_t$.
(a) Find the mean and covariance function for the $\{Y_t\}$ process.
(b) Show that the autocorrelation function for the $\{Y_t\}$ process depends only on the time lag. Is the $\{Y_t\}$ process stationary?
(c) Is it possible to have a time series with a constant mean and with $Corr(Y_t, Y_{t-k})$ free of t but with $\{Y_t\}$ *not* stationary?

2.11 Suppose $Cov(X_t, X_{t-k}) = \gamma_k$ is free of t but that $E(X_t) = 3t$.
(a) Is $\{X_t\}$ stationary?
(b) Let $Y_t = 7 - 3t + X_t$. Is $\{Y_t\}$ stationary?

2.12 Suppose that $Y_t = e_t - e_{t-12}$. Show that $\{Y_t\}$ is stationary and that, for $k > 0$, its autocorrelation function is nonzero only for lag $k = 12$.

2.13 Let $Y_t = e_t - \theta(e_{t-1})^2$. For this exercise, assume that the white noise series is normally distributed.
(a) Find the autocorrelation function for $\{Y_t\}$.
(b) Is $\{Y_t\}$ stationary?

2.14 Evaluate the mean and covariance function for each of the following processes. In each case, determine whether or not the process is stationary.
(a) $Y_t = \theta_0 + te_t$.
(b) $W_t = \nabla Y_t$, where Y_t is as given in part (a).
(c) $Y_t = e_t e_{t-1}$. (You may assume that $\{e_t\}$ is normal white noise.)

2.15 Suppose that X is a random variable with zero mean. Define a time series by $Y_t = (-1)^t X$.
(a) Find the mean function for $\{Y_t\}$.
(b) Find the covariance function for $\{Y_t\}$.
(c) Is $\{Y_t\}$ stationary?

2.16 Suppose $Y_t = A + X_t$, where $\{X_t\}$ is stationary and A is random but independent of $\{X_t\}$. Find the mean and covariance function for $\{Y_t\}$ in terms of the mean and autocovariance function for $\{X_t\}$ and the mean and variance of A.

2.17 Let $\{Y_t\}$ be stationary with autocovariance function γ_k. Let $\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$. Show that

$$Var(\bar{Y}) = \frac{\gamma_0}{n} + \frac{2}{n}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\gamma_k$$

$$= \frac{1}{n}\sum_{k=-n+1}^{n-1}\left(1 - \frac{|k|}{n}\right)\gamma_k$$

2.18 Let $\{Y_t\}$ be stationary with autocovariance function γ_k. Define the sample variance as $S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \bar{Y})^2$.

(a) First show that $\sum_{t=1}^{n}(Y_t - \mu)^2 = \sum_{t=1}^{n}(Y_t - \bar{Y})^2 + n(\bar{Y} - \mu)^2$.

(b) Use part (a) to show that

(c) $E(S^2) = \frac{n}{n-1}\gamma_0 - \frac{n}{n-1}Var(\bar{Y}) = \gamma_0 - \frac{2}{n-1}\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\gamma_k$.

(Use the results of Exercise 2.17 for the last expression.)

(d) If $\{Y_t\}$ is a white noise process with variance γ_0, show that $E(S^2) = \gamma_0$.

2.19 Let $Y_1 = \theta_0 + e_1$, and then for $t > 1$ define Y_t recursively by $Y_t = \theta_0 + Y_{t-1} + e_t$. Here θ_0 is a constant. The process $\{Y_t\}$ is called a **random walk with drift**.

(a) Show that Y_t may be rewritten as $Y_t = t\theta_0 + e_t + e_{t-1} + \cdots + e_1$.

(b) Find the mean function for Y_t.

(c) Find the autocovariance function for Y_t.

2.20 Consider the standard random walk model where $Y_t = Y_{t-1} + e_t$ with $Y_1 = e_1$.

(a) Use the representation of Y_t above to show that $\mu_t = \mu_{t-1}$ for $t > 1$ with initial condition $\mu_1 = E(e_1) = 0$. Hence show that $\mu_t = 0$ for all t.

(b) Similarly, show that $Var(Y_t) = Var(Y_{t-1}) + \sigma_e^2$ for $t > 1$ with $Var(Y_1) = \sigma_e^2$ and hence $Var(Y_t) = t\sigma_e^2$.

(c) For $0 \le t \le s$, use $Y_s = Y_t + e_{t+1} + e_{t+2} + \cdots + e_s$ to show that $Cov(Y_t, Y_s) = Var(Y_t)$ and, hence, that $Cov(Y_t, Y_s) = \min(t, s)\sigma_e^2$.

2.21 For a random walk with random starting value, let $Y_t = Y_0 + e_t + e_{t-1} + \cdots + e_1$ for $t > 0$, where Y_0 has a distribution with mean μ_0 and variance σ_0^2. Suppose further that $Y_0, e_1, \ldots, e_t$ are independent.

(a) Show that $E(Y_t) = \mu_0$ for all t.

(b) Show that $Var(Y_t) = t\sigma_e^2 + \sigma_0^2$.

(c) Show that $Cov(Y_t, Y_s) = \min(t, s)\sigma_e^2 + \sigma_0^2$.

(d) Show that $Corr(Y_t, Y_s) = \sqrt{\dfrac{t\sigma_a^2 + \sigma_0^2}{s\sigma_a^2 + \sigma_0^2}}$ for $0 \le t \le s$.

2.22 Let $\{e_t\}$ be a zero-mean white noise process, and let c be a constant with $|c| < 1$. Define Y_t recursively by $Y_t = cY_{t-1} + e_t$ with $Y_1 = e_1$.

(a) Show that $E(Y_t) = 0$.

(b) Show that $Var(Y_t) = \sigma_e^2(1 + c^2 + c^4 + \cdots + c^{2t-2})$. Is $\{Y_t\}$ stationary?

(c) Show that

$$Corr(Y_t, Y_{t-1}) = c\sqrt{\frac{Var(Y_{t-1})}{Var(Y_t)}} \quad \text{and, in general,}$$

$$Corr(Y_t, Y_{t-k}) = c^k\sqrt{\frac{Var(Y_{t-k})}{Var(Y_t)}} \qquad \text{for } k > 0$$

Hint: Argue that Y_{t-1} is independent of e_t. Then use

$$Cov(Y_t, Y_{t-1}) = Cov(cY_{t-1} + e_t, Y_{t-1})$$

(d) For large t, argue that

$$Var(Y_t) \approx \frac{\sigma_e^2}{1 - c^2} \qquad \text{and} \qquad Corr(Y_t, Y_{t-k}) \approx c^k \qquad \text{for } k > 0$$

so that $\{Y_t\}$ could be called **asymptotically stationary**.

(e) Suppose now that we alter the initial condition and put $Y_1 = \dfrac{e_1}{\sqrt{1 - c^2}}$. Show that now $\{Y_t\}$ is stationary.

2.23 Two processes $\{Z_t\}$ and $\{Y_t\}$ are said to be **independent** if for any time points $t_1, t_2, \ldots, t_m$ and $s_1, s_2, \ldots, s_n$ the random variables $\{Z_{t_1}, Z_{t_2}, \ldots, Z_{t_m}\}$ are independent of the random variables $\{Y_{s_1}, Y_{s_2}, \ldots, Y_{s_n}\}$. Show that if $\{Z_t\}$ and $\{Y_t\}$ are independent stationary processes, then $W_t = Z_t + Y_t$ is stationary.

2.24 Let $\{X_t\}$ be a time series in which we are interested. However, because the measurement process itself is not perfect, we actually observe $Y_t = X_t + e_t$. We assume that $\{X_t\}$ and $\{e_t\}$ are independent processes. We call X_t the **signal** and e_t the **measurement noise** or **error process**.

If $\{X_t\}$ is stationary with autocorrelation function ρ_k, show that $\{Y_t\}$ is also stationary with

$$Corr(Y_t, Y_{t-k}) = \frac{\rho_k}{1 + \sigma_e^2/\sigma_X^2} \qquad \text{for } k \geq 1$$

We call σ_X^2/σ_e^2 the **signal-to-noise ratio**, or SNR. Note that the larger the SNR, the closer the autocorrelation function of the observed process $\{Y_t\}$ is to the autocorrelation function of the desired signal $\{X_t\}$.

2.25 Suppose $Y_t = \beta_0 + \sum_{i=1}^{k} [A_i \cos(2\pi f_i t) + B_i \sin(2\pi f_i t)]$, where $\beta_0, f_1, f_2, \ldots, f_k$ are constants and $A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_k$ are independent random variables with zero means and variances $Var(A_i) = Var(B_i) = \sigma_i^2$. Show that $\{Y_t\}$ is stationary and find its covariance function.

2.26 Define the function $\Gamma_{t,s} = \frac{1}{2}E[(Y_t - Y_s)^2]$. In geostatistics, $\Gamma_{t,s}$ is called the *semivariogram*.

(a) Show that for a stationary process $\Gamma_{t,s} = \gamma_0 - \gamma_{|t-s|}$.

(b) A process is said to be *intrinsically stationary* if $\Gamma_{t,s}$ depends only on the time difference $|t - s|$. Show that the random walk process is intrinsically stationary.

2.27 For a fixed, positive integer r and constant ϕ, consider the time series defined by $Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots + \phi^r e_{t-r}$.

(a) Show that this process is stationary for any value of ϕ.

(b) Find the autocorrelation function.

2.28 (Random cosine wave extended) Suppose that

$$Y_t = R\cos(2\pi(ft + \Phi)) \qquad \text{for } t = 0, \pm 1, \pm 2, \ldots$$

where $0 < f < ½$ is a fixed frequency and R and Φ are uncorrelated random variables and with Φ uniformly distributed on the interval (0,1).

(a) Show that $E(Y_t) = 0$ for all t.

(b) Show that the process is stationary with $\gamma_k = \frac{1}{2}E(R^2)\cos(2\pi f k)$.

Hint: Use the calculations leading up to Equation (2.3.4), on page 19.

2.29 (Random cosine wave extended further) Suppose that

$$Y_t = \sum_{j=1}^{m} R_j \cos[2\pi(f_j t + \Phi_j)] \qquad \text{for } t = 0, \pm 1, \pm 2, \ldots$$

where $0 < f_1 < f_2 < \cdots < f_m < ½$ are m fixed frequencies, and R_1, Φ_1, R_2, Φ_2,…, R_m, Φ_m are uncorrelated random variables with each Φ_j uniformly distributed on the interval (0,1).

(a) Show that $E(Y_t) = 0$ for all t.

(b) Show that the process is stationary with $\gamma_k = \frac{1}{2}\sum_{j=1}^{m} E(R_j^2)\cos(2\pi f_j k)$.
Hint: Do Exercise 2.28 first.

2.30 (Mathematical statistics required) Suppose that

$$Y_t = R\cos[2\pi(ft + \Phi)] \qquad \text{for } t = 0, \pm 1, \pm 2, \ldots$$

where R and Φ are independent random variables and f is a fixed frequency. The phase Φ is assumed to be uniformly distributed on $(0,1)$, and the amplitude R has a Rayleigh distribution with pdf $f(r) = re^{-r^2/2}$ for $r > 0$. Show that for each time point t, Y_t has a normal distribution. (Hint: Let $Y = R\cos[2\pi(ft + \Phi)]$ and $X = R\sin[2\pi(ft + \Phi)]$. Now find the joint distribution of X and Y. It can also be shown that all of the finite dimensional distributions are multivariate normal and hence the process is strictly stationary.)

Appendix A: Expectation, Variance, Covariance, and Correlation

In this appendix, we define expectation for continuous random variables. However, all of the properties described hold for all types of random variables, discrete, continuous, or otherwise. Let X have probability density function $f(x)$ and let the pair (X,Y) have joint probability density function $f(x,y)$.

The **expected value** of X is defined as $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

(If $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$; otherwise $E(X)$ is undefined.) $E(X)$ is also called the **expectation** of X or the **mean** of X and is often denoted μ or μ_X.

Properties of Expectation

If $h(x)$ is a function such that $\int_{-\infty}^{\infty} |h(x)|f(x)dx < \infty$, it may be shown that

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

Similarly, if $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y)|f(x,y)dxdy < \infty$, it may be shown that

$$E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy \tag{2.A.1}$$

As a corollary to Equation (2.A.1), we easily obtain the important result

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{2.A.2}$$

We also have

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy \tag{2.A.3}$$

The **variance** of a random variable X is defined as

$$Var(X) = E\{[X - E(X)]^2\} \tag{2.A.4}$$

(provided $E(X^2)$ exists). The variance of X is often denoted by σ^2 or σ_X^2.

Properties of Variance

$$Var(X) \geq 0 \tag{2.A.5}$$

$$Var(a + bX) = b^2 Var(X) \tag{2.A.6}$$

If X and Y are independent, then

$$Var(X + Y) = Var(X) + Var(Y) \tag{2.A.7}$$

In general, it may be shown that

$$Var(X) = E(X^2) - [E(X)]^2 \tag{2.A.8}$$

The positive square root of the variance of X is called the **standard deviation** of X and is often denoted by σ or σ_X. The random variable $(X - \mu_X)/\sigma_X$ is called the **standardized version** of X. The mean and standard deviation of a standardized variable are always zero and one, respectively.

The **covariance** of X and Y is defined as $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Properties of Covariance

$$Cov(a + bX, c + dY) = bdCov(X, Y) \tag{2.A.9}$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) \tag{2.A.10}$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \tag{2.A.11}$$

$$Cov(X, X) = Var(X) \tag{2.A.12}$$

$$Cov(X, Y) = Cov(Y, X) \tag{2.A.13}$$

If X and Y are independent,

$$Cov(X, Y) = 0 \tag{2.A.14}$$

The **correlation coefficient** of X and Y, denoted by $Corr(X, Y)$ or ρ, is defined as

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Alternatively, if X^* is a standardized X and Y^* is a standardized Y, then $\rho = E(X^*Y^*)$.

Properties of Correlation

$$-1 \leq Corr(X, Y) \leq 1 \tag{2.A.15}$$

$$Corr(a + bX, c + dY) = sign(bd)Corr(X, Y)$$

$$\text{where } sign(bd) = \begin{cases} 1 \text{ if } bd > 0 \\ 0 \text{ if } bd = 0 \\ -1 \text{ if } bd < 0 \end{cases} \tag{2.A.16}$$

$Corr(X, Y) = \pm 1$ if and only if there are constants a and b such that $Pr(Y = a + bX) = 1$.

CHAPTER 3

TRENDS

In a general time series, the mean function is a totally arbitrary function of time. In a stationary time series, the mean function must be constant in time. Frequently we need to take the middle ground and consider mean functions that are relatively simple (but not constant) functions of time. These trends are considered in this chapter.

3.1 Deterministic Versus Stochastic Trends

"Trends" can be quite elusive. The same time series may be viewed quite differently by different analysts. The simulated random walk shown in Exhibit 2.1 might be considered to display a general upward trend. However, we know that the random walk process has zero mean for all time. The perceived trend is just an artifact of the strong positive correlation between the series values at nearby time points and the increasing variance in the process as time goes by. A second and third simulation of exactly the same process might well show completely different "trends." We ask you to produce some additional simulations in the exercises. Some authors have described such trends as **stochastic trends** (see Box, Jenkins, and Reinsel, 1994), although there is no generally accepted definition of a stochastic trend.

The average monthly temperature series plotted in Exhibit 1.7 on page 6, shows a cyclical or seasonal trend, but here the reason for the trend is clear—the Northern Hemisphere's changing inclination toward the sun. In this case, a possible model might be $Y_t = \mu_t + X_t$, where μ_t is a deterministic function that is periodic with period 12; that is μ_t, should satisfy

$$\mu_t = \mu_{t-12} \qquad \text{for all } t$$

We might assume that X_t, the unobserved variation around μ_t, has zero mean for all t so that indeed μ_t is the mean function for the observed series Y_t. We could describe this model as having a **deterministic trend** as opposed to the stochastic trend considered earlier. In other situations we might hypothesize a deterministic trend that is linear in time (that is, $\mu_t = \beta_0 + \beta_1 t$) or perhaps a quadratic time trend, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$. Note that an implication of the model $Y_t = \mu_t + X_t$ with $E(X_t) = 0$ for all t is that the deterministic trend μ_t applies for all time. Thus, if $\mu_t = \beta_0 + \beta_1 t$, we are assuming that the *same* linear time trend applies forever. We should therefore have good reasons for assuming such a model—not just because the series looks somewhat linear over the time period observed.

In this chapter, we consider methods for modeling deterministic trends. Stochastic trends will be discussed in Chapter 5, and stochastic seasonal models will be discussed in Chapter 10. Many authors use the word trend only for a slowly changing mean function, such as a linear time trend, and use the term seasonal component for a mean function that varies cyclically. We do not find it useful to make such distinctions here.

3.2 Estimation of a Constant Mean

We first consider the simple situation where a constant mean function is assumed. Our model may then be written as

$$Y_t = \mu + X_t \tag{3.2.1}$$

where $E(X_t) = 0$ for all t. We wish to estimate μ with our observed time series $Y_1, Y_2, \ldots, Y_n$. The most common estimate of μ is the sample mean or average defined as

$$\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t \tag{3.2.2}$$

Under the minimal assumptions of Equation (3.2.1), we see that $E(\bar{Y}) = \mu$; therefore $\bar{Y}$ is an unbiased estimate of μ. To investigate the precision of $\bar{Y}$ as an estimate of μ, we need to make further assumptions concerning X_t.

Suppose that $\{Y_t\}$, (or, equivalently, $\{X_t\}$ of Equation (3.2.1)) is a stationary time series with autocorrelation function ρ_k. Then, by Exercise 2.17, we have

$$\begin{aligned} Var(\bar{Y}) &= \frac{\gamma_0}{n}\left[\sum_{k=-n+1}^{n-1}\left(1-\frac{|k|}{n}\right)\rho_k\right] \\ &= \frac{\gamma_0}{n}\left[1+2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\rho_k\right] \end{aligned} \tag{3.2.3}$$

Notice that the first factor, γ_0/n, is the process (population) variance divided by the sample size—a concept with which we are familiar in simpler random sampling contexts. If the series $\{X_t\}$ of Equation (3.2.1) is just white noise, then $\rho_k = 0$ for $k > 0$ and $Var(\bar{Y})$ reduces to simply γ_0/n.

In the (stationary) moving average model $Y_t = e_t - \frac{1}{2}e_{t-1}$, we find that $\rho_1 = -0.4$ and $\rho_k = 0$ for $k > 1$. In this case, we have

$$\begin{aligned} Var(\bar{Y}) &= \frac{\gamma_0}{n}\left[1+2\left(1-\frac{1}{n}\right)(-0.4)\right] \\ &= \frac{\gamma_0}{n}\left[1-0.8\left(\frac{n-1}{n}\right)\right] \end{aligned}$$

For values of n usually occurring in time series ($n > 50$, say), the factor $(n-1)/n$ will be close to 1, so that we have

$$Var(\bar{Y}) \approx 0.2\frac{\gamma_0}{n}$$

We see that the negative correlation at lag 1 has improved the estimation of the mean compared with the estimation obtained in the white noise (random sample) situation. Because the series tends to oscillate back and forth across the mean, the sample mean obtained is more precise.

On the other hand, if $\rho_k \geq 0$ for all $k \geq 1$, we see from Equation (3.2.3) that $Var(\bar{Y})$ will be larger than γ_0/n. Here the positive correlations make estimation of the mean *more* difficult than in the white noise case. In general, some correlations will be positive and some negative, and Equation (3.2.3) must be used to assess the total effect.

For many stationary processes, the autocorrelation function decays quickly enough with increasing lags that

$$\sum_{k=0}^{\infty} |\rho_k| < \infty \tag{3.2.4}$$

(The random cosine wave of Chapter 2 is an exception.)

Under assumption (3.2.4) and given a large sample size n, the following useful approximation follows from Equation (3.2.3) (See Anderson, 1971, p. 459, for example)

$$Var(\bar{Y}) \approx \frac{\gamma_0}{n}\left[\sum_{k=-\infty}^{\infty} \rho_k\right] \qquad \text{for large } n \tag{3.2.5}$$

Notice that to this approximation the variance is inversely proportional to the sample size n.

As an example, suppose that $\rho_k = \phi^{|k|}$ for all k, where ϕ is a number strictly between -1 and $+1$. Summing a geometric series yields

$$Var(\bar{Y}) \approx \frac{(1+\phi)}{(1-\phi)}\frac{\gamma_0}{n} \tag{3.2.6}$$

For a nonstationary process (but with a constant mean), the precision of the sample mean as an estimate of μ can be strikingly different. As a useful example, suppose that in Equation (3.2.1) $\{X_t\}$ is a random walk process as described in Chapter 2. Then directly from Equation (2.2.8) we have

$$Var(\bar{Y}) = \frac{1}{n^2}Var\left[\sum_{i=1}^{n} Y_i\right]$$

$$= \frac{1}{n^2}Var\left[\sum_{i=1}^{n}\sum_{j=1}^{i} e_j\right]$$

$$= \frac{1}{n^2} Var(e_1 + 2e_2 + 3e_3 + \cdots + ne_n)$$

$$= \frac{\sigma_e^2}{n^2} \sum_{k=1}^{n} k^2$$

so that

$$Var(\bar{Y}) = \sigma_e^2 (2n+1)\frac{(n+1)}{6n} \tag{3.2.7}$$

Notice that in this special case the variance of our estimate of the mean actually *increases* as the sample size n increases. Clearly this is unacceptable, and we need to consider other estimation techniques for nonstationary series.

3.3 Regression Methods

The classical statistical method of regression analysis may be readily used to estimate the parameters of common nonconstant mean trend models. We shall consider the most useful ones: linear, quadratic, seasonal means, and cosine trends.

Linear and Quadratic Trends in Time

Consider the deterministic time trend expressed as

$$\mu_t = \beta_0 + \beta_1 t \tag{3.3.1}$$

where the *slope* and *intercept*, β_1 and β_0 respectively, are unknown parameters. The classical least squares (or regression) method is to choose as estimates of β_1 and β_0 values that minimize

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t)]^2$$

The solution may be obtained in several ways, for example, by computing the partial derivatives with respect to both β's, setting the results equal to zero, and solving the resulting linear equations for the β's. Denoting the solutions by $\hat{\beta}_0$ and $\hat{\beta}_1$, we find that

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (Y_t - \bar{Y})(t - \bar{t})}{\sum_{t=1}^{n} (t - \bar{t})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{t} \tag{3.3.2}$$

where $\bar{t} = (n+1)/2$ is the average of $1, 2, \ldots, n$. These formulas can be simplified somewhat, and various versions of the formulas are well-known. However, we assume that

the computations will be done by statistical software and we will not pursue other expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ here.

Example

Consider the random walk process that was shown in Exhibit 2.1. Suppose we (mistakenly) treat this as a linear time trend and estimate the slope and intercept by least-squares regression. Using statistical software we obtain Exhibit 3.1.

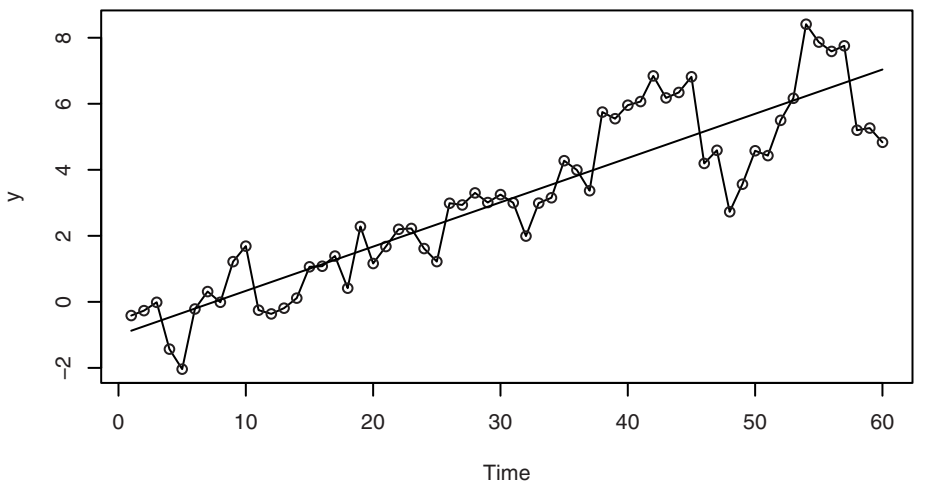
Exhibit 3.1 Least Squares Regression Estimates for Linear Time Trend

	Estimate	Std. Error	*t* value	*Pr*(>|*t*|)
Intercept	−1.008	0.2972	−3.39	0.00126
Time	0.1341	0.00848	15.82	< 0.0001

```
> data(rwalk)
> model1=lm(rwalk~time(rwalk))
> summary(model1)
```

So here the estimated slope and intercept are $\hat{\beta}_1 = 0.1341$ and $\hat{\beta}_0 = -1.008$, respectively. Exhibit 3.2 displays the random walk with the least squares regression trend line superimposed. We will interpret more of the regression output later in Section 3.5 on page 40 and see that fitting a line to these data is not appropriate.

Exhibit 3.2 Random Walk with Linear Time Trend




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> plot(rwalk,type='o',ylab='y')
> abline(model1) # add the fitted least squares line from model1
```

Cyclical or Seasonal Trends

Consider now modeling and estimating seasonal trends, such as for the average monthly temperature data in Exhibit 1.7. Here we assume that the observed series can be represented as

$$Y_t = \mu_t + X_t$$

where $E(X_t) = 0$ for all t.

The most general assumption for μ_t with monthly seasonal data is that there are 12 constants (parameters), β_1, β_2,…, and β_{12}, giving the expected average temperature for each of the 12 months. We may write

$$\mu_t = \begin{cases} \beta_1 & \text{for } t = 1, 13, 25, \ldots \\ \beta_2 & \text{for } t = 2, 14, 26, \ldots \\ \vdots & \\ \beta_{12} & \text{for } t = 12, 24, 36, \ldots \end{cases} \tag{3.3.3}$$

This is sometimes called a **seasonal means** model.

As an example of this model consider the average monthly temperature data shown in Exhibit 1.7 on page 6. To fit such a model, we need to set up indicator variables (sometimes called dummy variables) that indicate the month to which each of the data points pertains. The procedure for doing this will depend on the particular statistical software that you use. We also need to note that the model as stated does not contain an intercept term, and the software will need to know this also. Alternatively, we could use an intercept and leave out any one of the β's in Equation (3.3.3).

Exhibit 3.3 displays the results of fitting the seasonal means model to the temperature data. Here the t-values and $Pr(>|t|)$-values reported are of little interest since they relate to testing the null hypotheses that the β's are zero—not an interesting hypothesis in this case.

Exhibit 3.3 Regression Results for the Seasonal Means Model

	Estimate	Std. Error	*t*-value	*Pr*(>|*t*|)
January	16.608	0.987	16.8	< 0.0001
February	20.650	0.987	20.9	< 0.0001
March	32.475	0.987	32.9	< 0.0001
April	46.525	0.987	47.1	< 0.0001
May	58.092	0.987	58.9	< 0.0001
June	67.500	0.987	68.4	< 0.0001
July	71.717	0.987	72.7	< 0.0001

	Estimate	Std. Error	*t*-value	*Pr*(>|*t*|)
August	69.333	0.987	70.2	< 0.0001
September	61.025	0.987	61.8	< 0.0001
October	50.975	0.987	51.6	< 0.0001
November	36.650	0.987	37.1	< 0.0001
December	23.642	0.987	24.0	< 0.0001

```
> data(tempdub)
> month.=season(tempdub) # period added to improve table display
> model2=lm(tempdub~month.-1) # -1 removes the intercept term
> summary(model2)
```

Exhibit 3.4 shows how the results change when we fit a model *with* an intercept term. The software omits the January coefficient in this case. Now the February coefficient is interpreted as the difference between February and January average temperatures, the March coefficient is the difference between March and January average temperatures, and so forth. Once more, the *t*-values and *Pr*(>|*t*|) (*p*-values) are testing hypotheses of little interest in this case. Notice that the Intercept coefficient plus the February coefficient here equals the February coefficient displayed in Exhibit 3.3.

Exhibit 3.4 Results for Seasonal Means Model with an Intercept

	Estimate	Std. Error	*t*-value	*Pr*(>|*t*|)
Intercept	16.608	0.987	16.83	< 0.0001
February	4.042	1.396	2.90	0.00443
March	15.867	1.396	11.37	< 0.0001
April	29.917	1.396	21.43	< 0.0001
May	41.483	1.396	29.72	< 0.0001
June	50.892	1.396	36.46	< 0.0001
July	55.108	1.396	39.48	< 0.0001
August	52.725	1.396	37.78	< 0.0001
September	44.417	1.396	31.82	< 0.0001
October	34.367	1.396	24.62	< 0.0001
November	20.042	1.396	14.36	< 0.0001
December	7.033	1.396	5.04	< 0.0001

```
> model3=lm(tempdub~month.) # January is dropped automatically
> summary(model3)
```

Cosine Trends

The seasonal means model for monthly data consists of 12 independent parameters and does not take the shape of the seasonal trend into account at all. For example, the fact that the March and April means are quite similar (and different from the June and July means) is not reflected in the model. In some cases, seasonal trends can be modeled economically with cosine curves that incorporate the smooth change expected from one time period to the next while still preserving the seasonality.

Consider the cosine curve with equation

$$\mu_t = \beta\cos(2\pi ft + \Phi) \tag{3.3.4}$$

We call β (> 0) the *amplitude*, f the *frequency*, and Φ the *phase* of the curve. As t varies, the curve oscillates between a maximum of β and a minimum of $-\beta$. Since the curve repeats itself exactly every $1/f$ time units, $1/f$ is called the period of the cosine wave. As noted in Chapter 2, Φ serves to set the arbitrary origin on the time axis. For monthly data with time indexed as 1, 2,…, the most important frequency is $f = 1/12$, because such a cosine wave will repeat itself every 12 months. We say that the *period* is 12.

Equation (3.3.4) is inconvenient for estimation because the parameters β and Φ do not enter the expression linearly. Fortunately, a trigonometric identity is available that reparameterizes (3.3.4) more conveniently, namely

$$\beta\cos(2\pi ft + \Phi) = \beta_1\cos(2\pi ft) + \beta_2\sin(2\pi ft) \tag{3.3.5}$$

where

$$\beta = \sqrt{\beta_1^2 + \beta_2^2}\,, \qquad \Phi = \operatorname{atan}(-\beta_2/\beta_1) \tag{3.3.6}$$

and, conversely,

$$\beta_1 = \beta\cos(\Phi), \qquad \beta_2 = \beta\sin(\Phi) \tag{3.3.7}$$

To estimate the parameters β_1 and β_2 with regression techniques, we simply use $\cos(2\pi ft)$ and $\sin(2\pi ft)$ as regressors or predictor variables.

The simplest such model for the trend would be expressed as

$$\mu_t = \beta_0 + \beta_1\cos(2\pi ft) + \beta_2\sin(2\pi ft) \tag{3.3.8}$$

Here the constant term, β_0, can be meaningfully thought of as a cosine with frequency zero.

In any practical example, we must be careful how we measure time, as our choice of time measurement will affect the values of the frequencies of interest. For example, if we have monthly data but use 1, 2, 3,... as our time scale, then 1/12 would be the most interesting frequency, with a corresponding period of 12 months. However, if we measure time by year and fractional year, say 1980 for January, 1980.08333 for February of 1980, and so forth, then a frequency of 1 corresponds to an annual or 12 month periodicity.

Exhibit 3.5 is an example of fitting a cosine curve at the fundamental frequency to the average monthly temperature series.

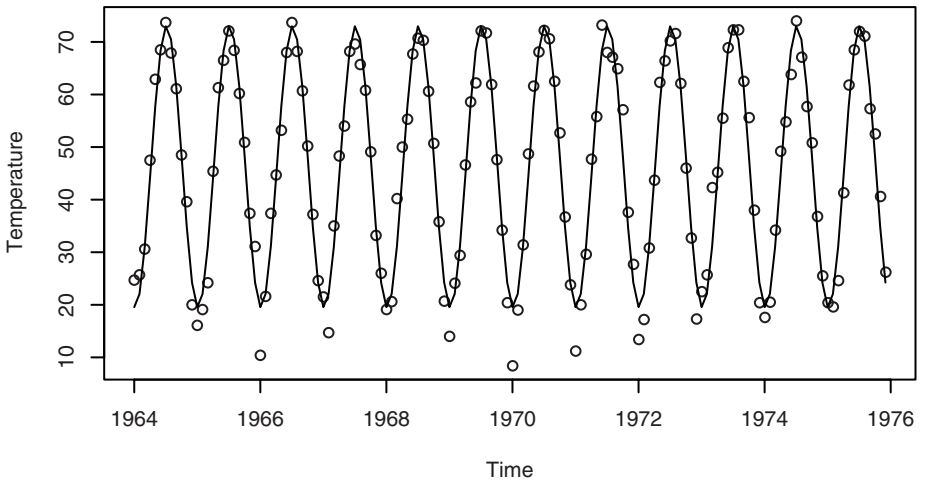
Exhibit 3.5 Cosine Trend Model for Temperature Series

Coefficient	Estimate	Std. Error	*t*-value	*Pr*(>|*t*|)
Intercept	46.2660	0.3088	149.82	< 0.0001
$\cos(2\pi t)$	−26.7079	0.4367	−61.15	< 0.0001
$\sin(2\pi t)$	−2.1697	0.4367	−4.97	<0.0001

```
> har.=harmonic(tempdub,1)
> model4=lm(tempdub~har.)
> summary(model4)
```

In this output, time is measured in years, with 1964 as the starting value and a frequency of 1 per year. A graph of the time series values together with the fitted cosine curve is shown in Exhibit 3.6. The trend fits the data quite well with the exception of most of the January values, where the observations are lower than the model would predict.

Exhibit 3.6 Cosine Trend for the Temperature Series




```
> win.graph(width=4.875, height=2.5,pointsize=8)
> plot(ts(fitted(model4),freq=12,start=c(1964,1)),
    ylab='Temperature',type='l',
> ylim=range(c(fitted(model4),tempdub))); points(tempdub)
> # ylim ensures that the y axis range fits the raw data and the
    fitted values
```

Additional cosine functions at other frequencies will frequently be used to model cyclical trends. For monthly series, the higher harmonic frequencies, such as 2/12 and 3/12, are especially pertinent and will sometimes improve the fit at the expense of add-

ing more parameters to the model. In fact, it may be shown that any periodic trend with period 12 may be expressed exactly by the sum of six pairs of cosine-sine functions. These ideas are discussed in detail in Fourier analysis or spectral analysis. We pursue these ideas further in Chapters 13 and 14.

3.4 Reliability and Efficiency of Regression Estimates

We assume that the series is represented as $Y_t = \mu_t + X_t$, where μ_t is a deterministic trend of the kind considered above and $\{X_t\}$ is a zero-mean stationary process with autocovariance and autocorrelation functions γ_k and ρ_k, respectively. Ordinary regression estimates parameters in a linear model according to the criterion of least squares regardless of whether we are fitting linear time trends, seasonal means, cosine curves, or whatever.

We first consider the easiest case—the seasonal means. As mentioned earlier, the least squares estimates of the seasonal means are just seasonal averages; thus, if we have N (complete) years of monthly data, we can write the estimate for the mean for the jth season as

$$\hat{\beta}_j = \frac{1}{N}\sum_{i=0}^{N-1} Y_{j+12i}$$

Since $\hat{\beta}_j$ is an average like $\overline{Y}$ but uses only every 12th observation, Equation (3.2.3) can be easily modified to give $Var(\hat{\beta}_j)$. We replace n by N (years) and ρ_k by ρ_{12k} to get

$$Var(\hat{\beta}_j) = \frac{\gamma_0}{N}\left[1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho_{12k}\right] \qquad \text{for } j = 1, 2, ..., 12 \tag{3.4.1}$$

We notice that if $\{X_t\}$ is white noise, then $Var(\hat{\beta}_j)$ reduces to γ_0/N, as expected. Furthermore, if several ρ_k are nonzero but $\rho_{12k} = 0$, then we still have $Var(\hat{\beta}_j) = \gamma_0/N$. In any case, only the seasonal autocorrelations, ρ_{12}, ρ_{24}, ρ_{36},..., enter into Equation (3.4.1). Since N will rarely be very large (except perhaps for quarterly data), approximations like those shown in Equation (3.2.5) will usually not be useful.

We turn now to the cosine trends expressed as in Equation (3.3.8). For any frequency of the form $f = m/n$, where m is an integer satisfying $1 \le m < n/2$, explicit expressions are available for the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, the amplitudes of the cosine and sine:

$$\hat{\beta}_1 = \frac{2}{n}\sum_{t=1}^{n}\left[\cos\left(\frac{2\pi mt}{n}\right)Y_t\right], \qquad \hat{\beta}_2 = \frac{2}{n}\sum_{t=1}^{n}\left[\sin\left(\frac{2\pi mt}{n}\right)Y_t\right] \tag{3.4.2}$$

(These are effectively the correlations between the time series $\{Y_t\}$ and the cosine and sine waves with frequency m/n.)

Because these are linear functions of $\{Y_t\}$, we may evaluate their variances using Equation (2.2.6). We find

$$Var(\hat{\beta}_1) = \frac{2\gamma_0}{n}\left[1 + \frac{4}{n}\sum_{s=2}^{n}\sum_{t=1}^{s-1}\cos\left(\frac{2\pi mt}{n}\right)\cos\left(\frac{2\pi ms}{n}\right)\rho_{s-t}\right] \tag{3.4.3}$$

where we have used the fact that $\sum_{t=1}^{n}[\cos(2\pi mt/n)]^2 = n/2$. However, the double sum in Equation (3.4.3) does not, in general, reduce further. A similar expression holds for $Var(\hat{\beta}_2)$ if we replace the cosines by sines.

If $\{X_t\}$ is white noise, we get just $2\gamma_0/n$. If $\rho_1 \neq 0$, $\rho_k = 0$ for $k > 1$, and $m/n = 1/12$, then the variance reduces to

$$Var(\hat{\beta}_1) = \frac{2\gamma_0}{n}\left[1 + \frac{4\rho_1}{n}\sum_{t=1}^{n-1}\cos\left(\frac{\pi t}{6}\right)\cos\left(\frac{\pi t+1}{6}\right)\right] \tag{3.4.4}$$

To illustrate the effect of the cosine terms, we have calculated some representative values:

n	$Var(\hat{\beta}_1)$
25	$\left(\frac{2\gamma_0}{n}\right)(1 + 1.71\rho_1)$
50	$\left(\frac{2\gamma_0}{n}\right)(1 + 1.75\rho_1)$
500	$\left(\frac{2\gamma_0}{n}\right)(1 + 1.73\rho_1)$
∞	$\left(\frac{2\gamma_0}{n}\right)\left(1 + 2\rho_1\cos\left(\frac{\pi}{6}\right)\right) = \left(\frac{2\gamma_0}{n}\right)(1 + 1.732\rho_1)$ (3.4.5)

If $\rho_1 = -0.4$, then the large sample multiplier in Equation (3.4.5) is $1+1.732(-0.4) = 0.307$ and the variance is reduced by about 70% when compared with the white noise case.

In some circumstances, seasonal means and cosine trends could be considered as competing models for a cyclical trend. If the simple cosine model is an adequate model, how much do we lose if we use the less parsimonious seasonal means model? To approach this problem, we must first consider how to compare the models. The parameters themselves are not directly comparable, but we can compare the estimates of the trend at comparable time points.

Consider the two estimates for the trend in January; that is, μ_1. With seasonal means, this estimate is just the January average, which has variance given by Equation (3.4.1). With the cosine trend model, the corresponding estimate is

$$\hat{\mu}_1 = \hat{\beta}_0 + \hat{\beta}_1 \cos\left(\frac{2\pi}{12}\right) + \hat{\beta}_2 \sin\left(\frac{2\pi}{12}\right)$$

To compute the variance of this estimate, we need one more fact: With this model, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are uncorrelated.† This follows from the orthogonality relationships of the cosines and sines involved. See Bloomfield (1976) or Fuller (1996) for more details. For the cosine model, then, we have

$$Var(\hat{\mu}_1) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1)\left[\cos\left(\frac{2\pi}{12}\right)\right]^2 + Var(\hat{\beta}_2)\left[\sin\left(\frac{2\pi}{12}\right)\right]^2 \qquad (3.4.6)$$

For our first comparison, assume that the stochastic component is white noise. Then the variance of our estimate in the seasonal means model is just γ_0/N. For the cosine model, we use Equation (3.4.6), and Equation (3.4.4) and its sine equivalent, to obtain

$$Var(\hat{\mu}_1) = \frac{\gamma_0}{n}\left\{1 + 2\left[\cos\left(\frac{\pi}{6}\right)\right]^2 + 2\left[\sin\left(\frac{\pi}{6}\right)\right]^2\right\}$$

$$= 3\frac{\gamma_0}{n}$$

since $(\cos\theta)^2 + (\sin\theta)^2 = 1$. Thus the ratio of the standard deviation in the cosine model to that in the seasonal means model is

$$\sqrt{\frac{3\gamma_0/n}{\gamma_0/N}} = \sqrt{\frac{3N}{n}}$$

In particular, for the monthly temperature series, we have $n = 144$ and $N = 12$; thus, the ratio is

$$\sqrt{\frac{3(12)}{144}} = 0.5$$

Thus, in the cosine model, we estimate the January effect with a standard deviation that is only half as large as it would be if we estimated with a seasonal means model—a substantial gain. (Of course, this assumes that the cosine trend plus white noise model is the correct model.)

Suppose now that the stochastic component is such that $\rho_1 \neq 0$ but $\rho_k = 0$ for $k > 1$. With a seasonal means model, the variance of the estimated January effect will be unchanged (see Equation (3.4.1) on page 36). For the cosine trend model, if we have a reasonably large sample size, we may use Equation (3.4.5), an identical expression for $Var(\hat{\beta}_2)$, and Equation (3.2.3) on page 28 for $Var(\hat{\beta}_0)$ to obtain

† This assumes that 1/12 is a "Fourier frequency"; that is, it is of the form m/n. Otherwise, these estimates are only approximately uncorrelated.

$$Var(\hat{\mu}_1) = \frac{\gamma_0}{n}\left\{1 + 2\rho_1 + 2\left[1 + 2\rho_1 \cos\left(\frac{2\pi}{12}\right)\right]\right\}$$

$$= \frac{\gamma_0}{n}\left\{3 + 2\rho_1\left[1 + 2\cos\left(\frac{\pi}{6}\right)\right]\right\}$$

If $\rho_1 = -0.4$, then we have $0.814\gamma_0/n$, and the ratio of the standard deviation in the cosine case to the standard deviation in the seasonal means case is

$$\sqrt{\left[\frac{(0.814\gamma_0)/n}{\gamma_0/N}\right]} = \sqrt{\frac{0.814N}{n}}$$

If we take $n = 144$ and $N = 12$, the ratio is

$$\sqrt{\frac{0.814(12)}{144}} = 0.26$$

a very substantial reduction indeed!

We now turn to linear time trends. For these trends, an alternative formula to Equation (3.3.2) on page 30 for $\hat{\beta}_1$ is more convenient. It can be shown that the least squares estimate of the slope may be written

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})Y_t}{\sum_{t=1}^{n}(t-\bar{t})^2} \tag{3.4.7}$$

Since the estimate is a linear combination of Y-values, some progress can be made in evaluating its variance. We have

$$Var(\hat{\beta}_1) = \frac{12\gamma_0}{n(n^2-1)}\left[1 + \frac{24}{n(n^2-1)}\sum_{s=2}^{n}\sum_{t=1}^{s-1}(t-\bar{t})(s-\bar{t})\rho_{s-t}\right] \tag{3.4.8}$$

where we have used $\sum_{t=1}^{n}(t-\bar{t})^2 = n(n^2-1)/12$. Again the double sum does not in general reduce.

To illustrate the effect of Equation (3.4.8), consider again the case where $\rho_1 \neq 0$ but $\rho_k = 0$ for $k > 1$. Then, after some algebraic manipulation, again involving the sum of consecutive integers and their squares, Equation (3.4.8) can be reduced to

$$Var(\hat{\beta}_1) = \frac{12\gamma_0}{n(n^2-1)}\left[1 + 2\rho_1\left(1 - \frac{3}{n}\right)\right]$$

For large n, we can neglect the $3/n$ term and use

$$Var(\hat{\beta}_1) = \frac{12\gamma_0(1+2\rho_1)}{n(n^2-1)} \tag{3.4.9}$$

If $\rho_1 = -0.4$, then $1 + 2\rho_1 = 0.2$, and then the variance of $\hat{\beta}_1$ is only 20% of what it would be if $\{X_t\}$ were white noise. Of course, if $\rho_1 > 0$, then the variance would be larger than for the white noise case.

We turn now to comparing the least squares estimates with the so-called **best linear unbiased estimates** (BLUE) or the **generalized least squares** (GLS) estimates. If the stochastic component $\{X_t\}$ is not white noise, estimates of the unknown parameters in the trend function may be made; they are linear functions of the data, are unbiased, and have the smallest variances among all such estimates—the so-called BLUE or GLS estimates. These estimates and their variances can be expressed fairly explicitly by using certain matrices and their inverses. (Details may be found in Draper and Smith (1981).) However, constructing these estimates requires complete knowledge of the covariance function of the stochastic component, a function that is unknown in virtually all real applications. It is possible to iteratively estimate the covariance function for $\{X_t\}$ based on a preliminary estimate of the trend. The trend is then estimated again using the estimated covariance function for $\{X_t\}$ and thus iterated to an approximate BLUE for the trend. This method will not be pursued here, however.

Fortunately, there are some results based on large sample sizes that support the use of the simpler least squares estimates for the types of trends that we have considered. In particular, we have the following result (see Fuller (1996), pp. 476–480, for more details): We assume that the trend is either a polynomial in time, a trigonometric polynomial, seasonal means, or a linear combination of these. Then, for a very general stationary stochastic component $\{X_t\}$, the least squares estimates for the trend have the same variance as the best linear unbiased estimates for large sample sizes.

Although the simple least squares estimates may be asymptotically efficient, it does not follow that the estimated standard deviations of the coefficients as printed out by all regression routines are correct. We shall elaborate on this point in the next section. We also caution the reader that the result above is restricted to certain kinds of trends and cannot, in general, be extended to regression on arbitrary predictor variables, such as other time series. For example, Fuller (1996, pp. 518–522) shows that if $Y_t = \beta Z_t + X_t$, where $\{X_t\}$ has a simple stochastic structure but $\{Z_t\}$ is also a stationary series, then the least squares estimate of β can be very inefficient and biased even for large samples.

3.5 Interpreting Regression Output

We have already noted that the standard regression routines calculate least squares estimates of the unknown regression coefficients—the betas. As such, the estimates are reasonable under minimal assumptions on the stochastic component $\{X_t\}$. However, some of the properties of the regression output depend heavily on the usual regression assumption that $\{X_t\}$ is white noise, and some depend on the further assumption that $\{X_t\}$ is approximately normally distributed. We begin with the items that depend least on the assumptions.

Consider the regression output shown in Exhibit 3.7. We shall write $\hat{\mu}_t$ for the estimated trend regardless of the assumed parametric form for μ_t. For example, for the linear time trend, we have $\mu_t = \beta_0 + \beta_1 t$. For each t, the unobserved stochastic component

X_t can be estimated (predicted) by $Y_t - \hat{\mu}_t$. If the $\{X_t\}$ process has constant variance, then we can estimate the standard deviation of X_t, namely $\sqrt{\gamma_0}$, by the **residual standard deviation**

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2} \tag{3.5.1}$$

where p is the number of parameters estimated in μ_t and $n - p$ is the so-called *degrees of freedom* for s. The value of s gives an absolute measure of the goodness of fit of the estimated trend—the smaller the value of s, the better the fit. However, a value of s of, say, 60.74 is somewhat difficult to interpret.

A unitless measure of the goodness of fit of the trend is the value of R^2, also called the **coefficient of determination** or multiple R-squared. One interpretation of R^2 is that it is the square of the sample correlation coefficient between the observed series and the estimated trend. It is also the fraction of the variation in the series that is explained by the estimated trend. Exhibit 3.7 is a more complete regression output when fitting the straight line to the random walk data. This extends what we saw in Exhibit 3.1 on page 31.

Exhibit 3.7 Regression Output for Linear Trend Fit of Random Walk

	Estimate	Std. Error	*t*-value	*Pr*(>|*t*|)
Intercept	−1.007888	0.297245	−3.39	0.00126
Time	0.134087	0.008475	15.82	< 0.0001

Residual standard error	1.137	with 58 degrees of freedom
Multiple *R*-Squared	0.812	
Adjusted *R*-squared	0.809	
***F*-statistic**	250.3	with 1 and 58 df; p-value < 0.0001

```
> model1=lm(rwalk~time(rwalk))
> summary(model1)
```

According to Exhibit 3.7, about 81% of the variation in the random walk series is explained by the linear time trend. The adjusted R-squared value is a small adjustment to R^2 that yields an approximately unbiased estimate based on the number of parameters estimated in the trend. It is useful for comparing models with different numbers of parameters. Various formulas for computing R^2 may be found in any book on regression, such as Draper and Smith (1981). The standard deviations of the coefficients labeled Std. Error on the output need to be interpreted carefully. They are appropriate only when the stochastic component is white noise—the usual regression assumption.

For example, in Exhibit 3.7 the value 1.137 is obtained from the square root of the value given by Equation (3.4.8) when $\rho_k = 0$ for $k > 0$ and with γ_0 estimated by s^2, that is, to within rounding,

$$0.008475 = \sqrt{\frac{12(1.137)^2}{60(60^2 - 1)}}$$

The important point is that these standard deviations assume a white noise stochastic component that will rarely be true for time series.

The *t*-values or *t*-ratios shown in Exhibit 3.7 are just the estimated regression coefficients, each divided by their respective standard errors. If the stochastic component is normally distributed white noise, then these ratios provide appropriate test statistics for checking the significance of the regression coefficients. In each case, the null hypothesis is that the corresponding unknown regression coefficient is zero. The significance levels and *p*-values are determined from the *t*-distribution with $n - p$ degrees of freedom.

3.6 Residual Analysis

As we have already noted, the unobserved stochastic component $\{X_t\}$ can be estimated, or predicted, by the **residual**

$$\hat{X}_t = Y_t - \hat{\mu}_t \tag{3.6.1}$$

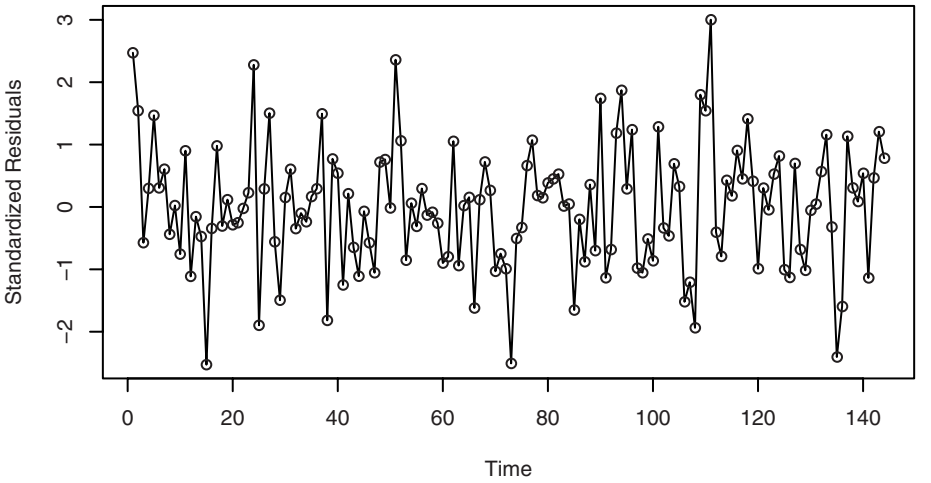
Predicted is really a better term. We reserve the term estimate for the guess of an unknown parameter and the term predictor for an estimate of an unobserved random variable. We call $\hat{X}_t$ the residual corresponding to the *t*th observation. If the trend model is reasonably correct, then the residuals should behave roughly like the true stochastic component, and various assumptions about the stochastic component can be assessed by looking at the residuals. If the stochastic component is white noise, then the residuals should behave roughly like independent (normal) random variables with zero mean and standard deviation *s*. Since a least squares fit of any trend containing a constant term automatically produces residuals with a zero mean, we might consider standardizing the residuals as $\hat{X}_t / s$. However, most statistics software will produce standardized residuals using a more complicated standard error in the denominator that takes into account the specific regression model being fit.

With the residuals or standardized residuals in hand, the next step is to examine various residual plots. We first look at the plot of the residuals over time. If the data are possibly seasonal, we should use plotting symbols as we did in Exhibit 1.9 on page 7, so that residuals associated with the same season can be identified easily.

We will use the monthly average temperature series which we fitted with seasonal means as our first example to illustrate some of the ideas of residual analysis. Exhibit 1.7 on page 6 shows the time series plot of that series. Exhibit 3.8 shows a time series plot for the standardized residuals of the monthly temperature data fitted by seasonal means. If the stochastic component is white noise and the trend is adequately modeled, we would expect such a plot to suggest a rectangular scatter with no discernible trends whatsoever. There are no striking departures from randomness apparent in this display.

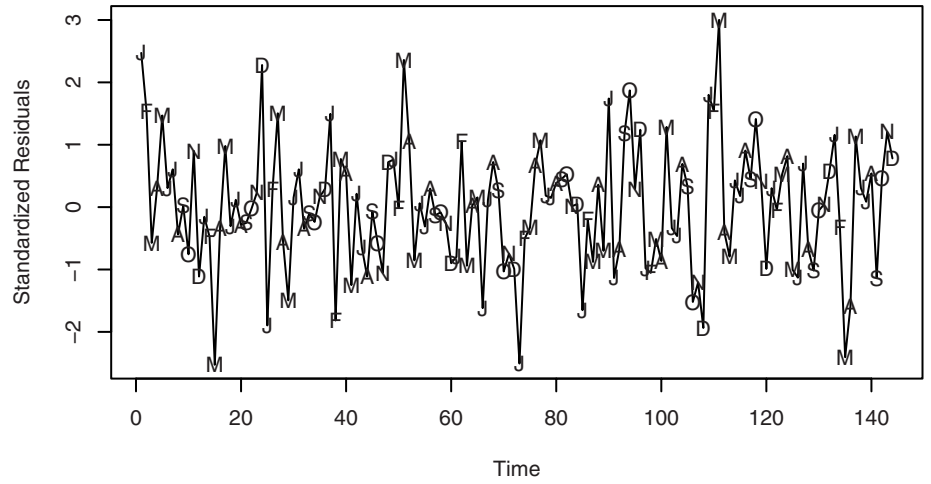
Exhibit 3.9 repeats the time series plot but now with seasonal plotting symbols. Again there are no apparent patterns relating to different months of the year.

Exhibit 3.8 Residuals versus Time for Temperature Seasonal Means




```
> plot(y=rstudent(model3),x=as.vector(time(tempdub)),
    xlab='Time',ylab='Standardized Residuals',type='o')
```

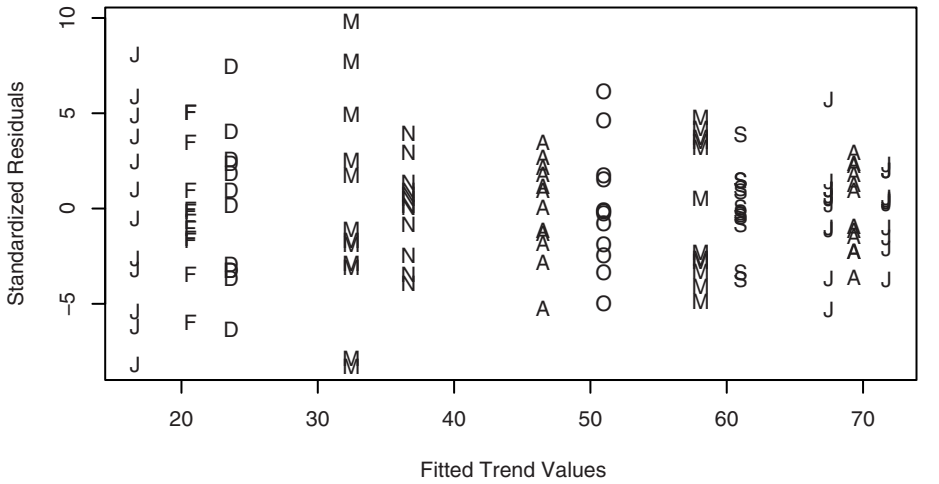
Exhibit 3.9 Residuals versus Time with Seasonal Plotting Symbols




```
> plot(y=rstudent(model3),x=as.vector(time(tempdub)),xlab='Time',
> ylab='Standardized Residuals',type='l')
> points(y=rstudent(model3),x=as.vector(time(tempdub)),
    pch=as.vector(season(tempdub)))
```

Next we look at the standardized residuals versus the corresponding trend estimate, or fitted value, as in Exhibit 3.10. Once more we are looking for patterns. Are small residuals associated with small fitted trend values and large residuals with large fitted trend values? Is there less variation for residuals associated with certain sized fitted trend values or more variation with other fitted trend values? There is somewhat more variation for the March residuals and less for November, but Exhibit 3.10 certainly does not indicate any dramatic patterns that would cause us to doubt the seasonal means model.

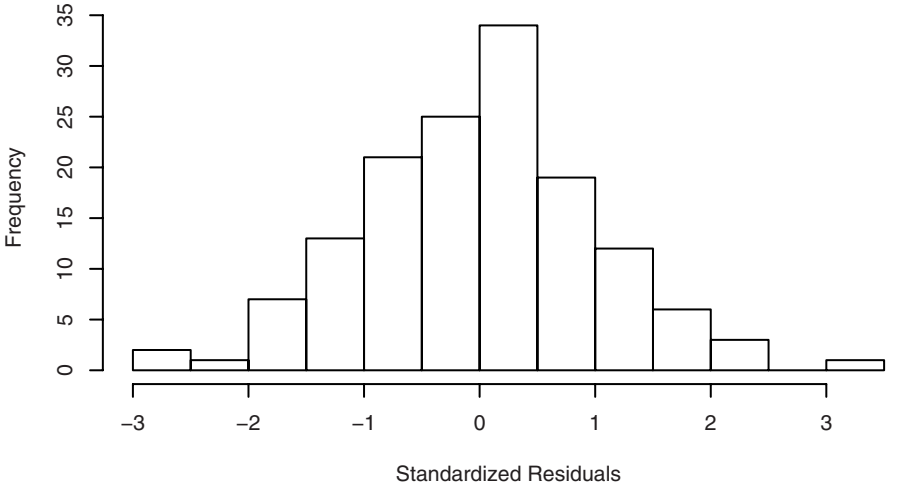
Exhibit 3.10 Standardized Residuals versus Fitted Values for the Temperature Seasonal Means Model




```
> plot(y=rstudent(model3),x=as.vector(fitted(model3)),
    xlab='Fitted Trend Values',
> ylab='Standardized Residuals',type='n')
> points(y=rstudent(model3),x=as.vector(fitted(model3)),
    pch=as.vector(season(tempdub)))
```

Gross nonnormality can be assessed by plotting a histogram of the residuals or standardized residuals. Exhibit 3.11 displays a frequency histogram of the standardized residuals from the seasonal means model for the temperature series. The plot is somewhat symmetric and tails off at both the high and low ends as a normal distribution does.

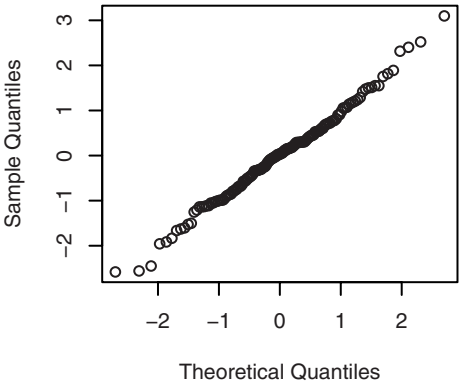
Exhibit 3.11 Histogram of Standardized Residuals from Seasonal Means Model




```
> hist(rstudent(model3),xlab='Standardized Residuals')
```

Normality can be checked more carefully by plotting the so-called normal scores or quantile-quantile (QQ) plot. Such a plot displays the quantiles of the data versus the theoretical quantiles of a normal distribution. With normally distributed data, the QQ plot looks approximately like a straight line. Exhibit 3.12 shows the QQ normal scores plot for the standardized residuals from the seasonal means model for the temperature series. The straight-line pattern here supports the assumption of a normally distributed stochastic component in this model.

Exhibit 3.12 Q-Q Plot: Standardized Residuals of Seasonal Means Model




```
> win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(rstudent(model3))
```

An excellent test of normality is known as the Shapiro-Wilk test.[†] It essentially calculates the correlation between the residuals and the corresponding normal quantiles. The lower this correlation, the more evidence we have against normality. Applying that test to these residuals gives a test statistic of $W = 0.9929$ with a p-value of 0.6954. We cannot reject the null hypothesis that the stochastic component of this model is normally distributed.

Independence in the stochastic component can be tested in several ways. The **runs test** examines the residuals in sequence to look for patterns—patterns that would give evidence against independence. Runs above or below their median are counted. A small number of runs would indicate that neighboring residuals are positively dependent and tend to "hang together" over time. On the other hand, too many runs would indicate that the residuals oscillate back and forth across their median. Then neighboring residuals are negatively dependent. So either too few or too many runs lead us to reject independence. Performing a runs test[‡] on these residuals produces the following values: observed runs = 65, expected runs = 72.875, which leads to a p-value of 0.216 and we cannot reject independence of the stochastic component in this seasonal means model.

The Sample Autocorrelation Function

Another very important diagnostic tool for examining dependence is the sample autocorrelation function. Consider any sequence of data $Y_1, Y_2, \ldots, Y_n$—whether residuals, standardized residuals, original data, or some transformation of data. Tentatively assuming stationarity, we would like to estimate the autocorrelation function ρ_k for a variety of lags $k = 1, 2, \ldots$. The obvious way to do this is to compute the sample correlation between the pairs k units apart in time. That is, among (Y_1, Y_{1+k}), (Y_2, Y_{2+k}), (Y_3, Y_{3+k}),..., and (Y_{n-k}, Y_n). However, we modify this slightly, taking into account that we are assuming stationarity, which implies a common mean and variance for the series. With this in mind, we define the **sample autocorrelation function**, r_k, at lag k as

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \qquad \text{for } k = 1, 2, \ldots \tag{3.6.2}$$

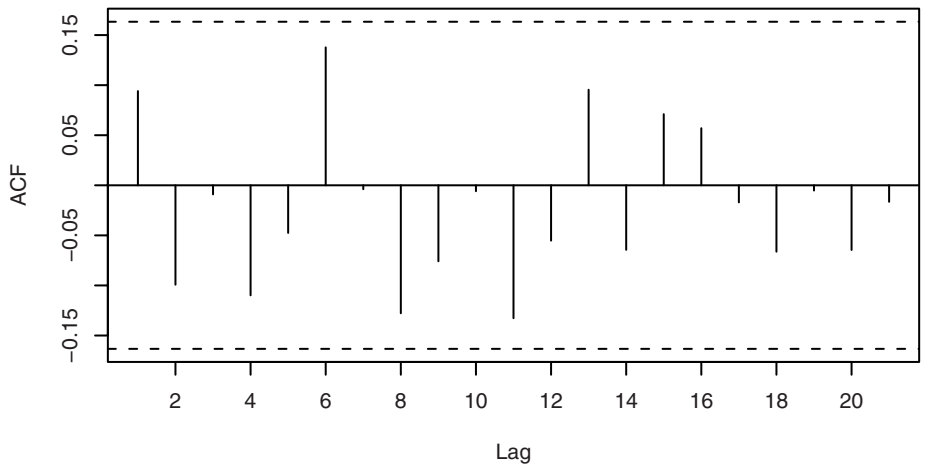
Notice that we used the "grand mean," $\bar{Y}$, in all places and have also divided by the "grand sum of squares" rather than the product of the two separate standard deviations used in the ordinary correlation coefficient. We also note that the denominator is a sum of n squared terms while the numerator contains only $n - k$ cross products. For a variety of reasons, this has become the standard definition for the sample autocorrelation function. A plot of r_k versus lag k is often called a **correlogram**.

[†] Royston, P. (1982) "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples." *Applied Statistics*, **31**, 115–124.

[‡] R code: `runs(rstudent(model3))`

In our present context, we are interested in discovering possible dependence in the stochastic component; therefore the sample autocorrelation function for the standardized residuals is of interest. Exhibit 3.13 displays the sample autocorrelation for the standardized residuals from the seasonal means model of the temperature series. All values are within the horizontal dashed lines, which are placed at zero plus and minus two approximate standard errors of the sample autocorrelations, namely $\pm 2/\sqrt{n}$. The values of r_k are, of course, estimates of ρ_k. As such, they have their own sampling distributions, standard errors, and other properties. For now we shall use r_k as a descriptive tool and defer discussion of those topics until Chapters 6 and 8. According to Exhibit 3.13, for k = 1, 2,..., 21, none of the hypotheses $\rho_k = 0$ can be rejected at the usual significance levels, and it is reasonable to infer that the stochastic component of the series is white noise.

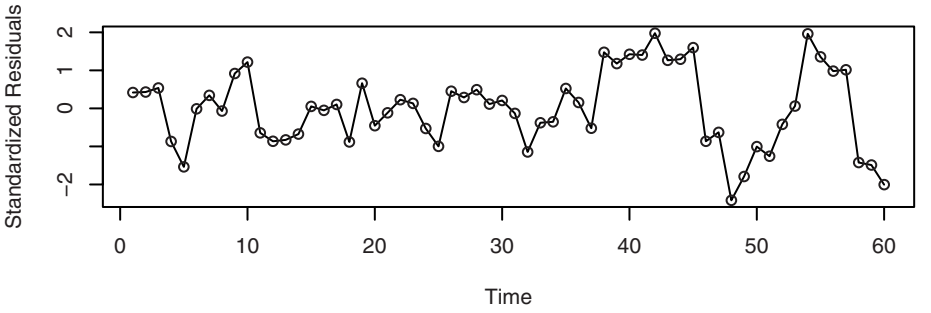
Exhibit 3.13 Sample Autocorrelation of Residuals of Seasonal Means Model




```
> win.graph(width=4.875,height=3,pointsize=8)
> acf(rstudent(model3))
```

As a second example consider the standardized residuals from fitting a straight line to the random walk time series. Recall Exhibit 3.2 on page 31, which shows the data and fitted line. A time series plot of the standardized residuals is shown in Exhibit 3.14.

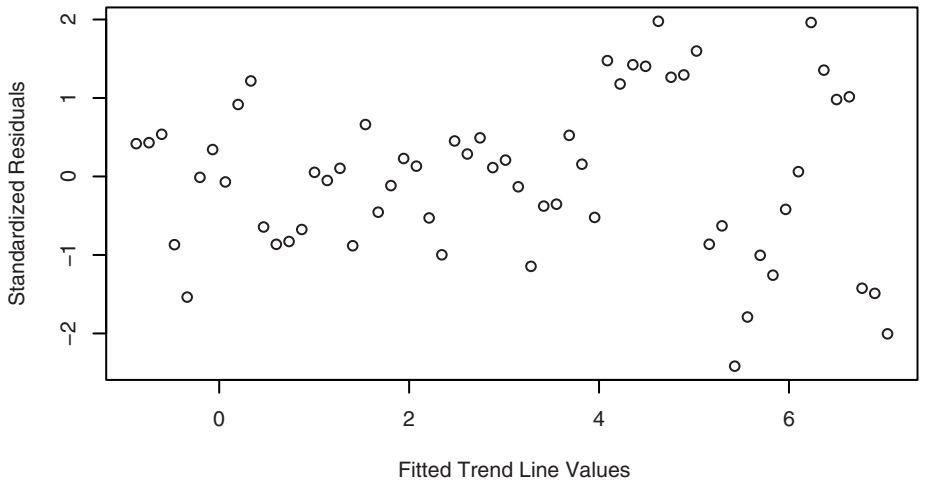
Exhibit 3.14 Residuals from Straight Line Fit of the Random Walk




```
> plot(y=rstudent(model1),x=as.vector(time(rwalk)),
    ylab='Standardized Residuals',xlab='Time',type='o')
```

In this plot, the residuals "hang together" too much for white noise—the plot is too smooth. Furthermore, there seems to be more variation in the last third of the series than in the first two-thirds. Exhibit 3.15 shows a similar effect with larger residuals associated with larger fitted values.

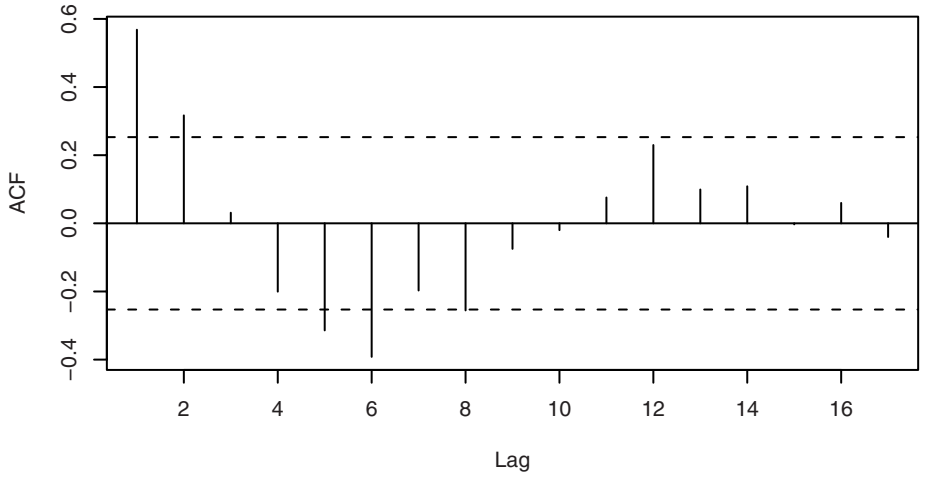
Exhibit 3.15 Residuals versus Fitted Values from Straight Line Fit




```
> win.graph(width=4.875, height=3,pointsize=8)
> plot(y=rstudent(model1),x=fitted(model1),
    ylab='Standardized Residuals',xlab='Fitted Trend Line Values',
    type='p')
```

The sample autocorrelation function of the standardized residuals, shown in Exhibit 3.16, confirms the smoothness of the time series plot that we observed in Exhibit 3.14. The lag 1 and lag 2 autocorrelations exceed two standard errors above zero and the lag 5 and lag 6 autocorrelations more than two standard errors below zero. This is not what we expect from a white noise process.

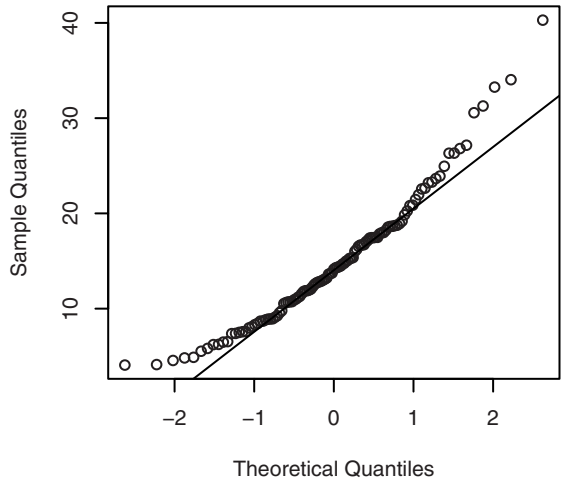
Exhibit 3.16 Sample Autocorrelation of Residuals from Straight Line Model




```
> acf(rstudent(model1))
```

Finally, we return to the annual rainfall in Los Angeles shown in Exhibit 1.1 on page 2. We found no evidence of dependence in that series, but we now look for evidence against normality. Exhibit 3.17 displays the normal quantile-quantile plot for that series. We see considerable curvature in the plot. A line passing through the first and third normal quartiles helps point out the departure from a straight line in the plot.

Exhibit 3.17 Quantile-Quantile Plot of Los Angeles Annual Rainfall Series




```
> win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(larain); qqline(larain)
```

3.7 Summary

This chapter is concerned with describing, modeling, and estimating deterministic trends in time series. The simplest deterministic "trend" is a constant-mean function. Methods of estimating a constant mean were given but, more importantly, assessment of the accuracy of the estimates under various conditions was considered. Regression methods were then pursued to estimate trends that are linear or quadratic in time. Methods for modeling cyclical or seasonal trends came next, and the reliability and efficiency of all of these regression methods were investigated. The final section began our study of residual analysis to investigate the quality of the fitted model. This section also introduced the important sample autocorrelation function, which we will revisit throughout the remainder of the book.

EXERCISES

3.1 Verify Equation (3.3.2) on page 30, for the least squares estimates of β_0 and of β_1 when the model $Y_t = \beta_0 + \beta_1 t + X_t$ is considered.

3.2 Suppose $Y_t = \mu + e_t - e_{t-1}$. Find $Var(\overline{Y})$. Note any unusual results. In particular, compare your answer to what would have been obtained if $Y_t = \mu + e_t$. (Hint: You may avoid Equation (3.2.3) on page 28 by first doing some algebraic simplification on $\sum_{t=1}^{n} (e_t - e_{t-1})$.)

3.3 Suppose $Y_t = \mu + e_t + e_{t-1}$. Find $Var(\overline{Y})$. Compare your answer to what would have been obtained if $Y_t = \mu + e_t$. Describe the effect that the autocorrelation in $\{Y_t\}$ has on $Var(\overline{Y})$.

3.4 The data file hours contains monthly values of the average hours worked per week in the U.S. manufacturing sector for July 1982 through June 1987.

(a) Display and interpret the time series plot for these data.

(b) Now construct a time series plot that uses separate plotting symbols for the various months. Does your interpretation change from that in part (a)?

3.5 The data file wages contains monthly values of the average hourly wages (in dollars) for workers in the U.S. apparel and textile products industry for July 1981 through June 1987.

(a) Display and interpret the time series plot for these data.

(b) Use least squares to fit a linear time trend to this time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

(c) Construct and interpret the time series plot of the standardized residuals from part (b).

(d) Use least squares to fit a quadratic time trend to the wages time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

(e) Construct and interpret the time series plot of the standardized residuals from part (d).

3.6 The data file beersales contains monthly U.S. beer sales (in millions of barrels) for the period January 1975 through December 1990.

(a) Display and interpret the plot the time series plot for these data.

(b) Now construct a time series plot that uses separate plotting symbols for the various months. Does your interpretation change from that in part (a)?

(c) Use least squares to fit a seasonal-means trend to this time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

(d) Construct and interpret the time series plot of the standardized residuals from part (c). Be sure to use proper plotting symbols to check on seasonality in the standardized residuals.

(e) Use least squares to fit a seasonal-means plus quadratic time trend to the beer sales time series. Interpret the regression output. Save the standardized residuals from the fit for further analysis.

(f) Construct and interpret the time series plot of the standardized residuals from part (e). Again use proper plotting symbols to check for any remaining seasonality in the residuals.

3.7 The data file winnebago contains monthly unit sales of recreational vehicles from Winnebago, Inc., from November 1966 through February 1972.

(a) Display and interpret the time series plot for these data.

(b) Use least squares to fit a line to these data. Interpret the regression output. Plot the standardized residuals from the fit as a time series. Interpret the plot.

(c) Now take natural logarithms of the monthly sales figures and display and

interpret the time series plot of the transformed values.

(d) Use least squares to fit a line to the logged data. Display and interpret the time series plot of the standardized residuals from this fit.

(e) Now use least squares to fit a seasonal-means plus linear time trend to the logged sales time series and save the standardized residuals for further analysis. Check the statistical significance of each of the regression coefficients in the model.

(f) Display the time series plot of the standardized residuals obtained in part (e). Interpret the plot.

3.8 The data file retail lists total U.K. (United Kingdom) retail sales (in billions of pounds) from January 1986 through March 2007. The data are not "seasonally adjusted," and year 2000 = 100 is the base year.

(a) Display and interpret the time series plot for these data. Be sure to use plotting symbols that permit you to look for seasonality.

(b) Use least squares to fit a seasonal-means plus linear time trend to this time series. Interpret the regression output and save the standardized residuals from the fit for further analysis.

(c) Construct and interpret the time series plot of the standardized residuals from part (b). Be sure to use proper plotting symbols to check on seasonality.

3.9 The data file prescrip gives monthly U.S. prescription costs for the months August 1986 to March 1992. These data are from the State of New Jersey's Prescription Drug Program and are the cost per prescription claim.

(a) Display and interpret the time series plot for these data. Use plotting symbols that permit you to look for seasonality.

(b) Calculate and plot the sequence of month-to-month percentage changes in the prescription costs. Again, use plotting symbols that permit you to look for seasonality.

(c) Use least squares to fit a cosine trend with fundamental frequency 1/12 to the percentage change series. Interpret the regression output. Save the standardized residuals.

(d) Plot the sequence of standardized residuals to investigate the adequacy of the cosine trend model. Interpret the plot.

3.10 (Continuation of Exercise 3.4) Consider the hours time series again.

(a) Use least squares to fit a quadratic trend to these data. Interpret the regression output and save the standardized residuals for further analysis.

(b) Display a sequence plot of the standardized residuals and interpret. Use monthly plotting symbols so that possible seasonality may be readily identified.

(c) Perform the Runs test of the standardized residuals and interpret the results.

(d) Calculate and interpret the sample autocorrelations for the standardized residuals.

(e) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.11 (Continuation of Exercise 3.5) Return to the wages series.
(a) Consider the residuals from a least squares fit of a quadratic time trend.
(b) Perform a runs test on the standardized residuals and interpret the results.
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.
(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.12 (Continuation of Exercise 3.6) Consider the time series in the data file beersales.
(a) Obtain the residuals from the least squares fit of the seasonal-means plus quadratic time trend model.
(b) Perform a runs test on the standardized residuals and interpret the results.
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.
(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.13 (Continuation of Exercise 3.7) Return to the winnebago time series.
(a) Calculate the least squares residuals from a seasonal-means plus linear time trend model on the logarithms of the sales time series.
(b) Perform a runs test on the standardized residuals and interpret the results.
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.
(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.14 (Continuation of Exercise 3.8) The data file retail contains U.K. monthly retail sales figures.
(a) Obtain the least squares residuals from a seasonal-means plus linear time trend model.
(b) Perform a runs test on the standardized residuals and interpret the results.
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.
(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.15 (Continuation of Exercise 3.9) Consider again the prescrip time series.
(a) Save the standardized residuals from a least squares fit of a cosine trend with fundamental frequency 1/12 to the percentage change time series.
(b) Perform a runs test on the standardized residuals and interpret the results.
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.
(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.

3.16 Suppose that a stationary time series, $\{Y_t\}$, has an autocorrelation function of the form $\rho_k = \phi^k$ for $k > 0$, where ϕ is a constant in the range (−1,+1).

(a) Show that $Var(\bar{Y}) = \frac{\gamma_0}{n}\left[\frac{1+\phi}{1-\phi} - \frac{2\phi(1-\phi^n)}{n(1-\phi)^2}\right]$.

(Hint: Use Equation (3.2.3) on page 28, the finite geometric sum $\sum_{k=0}^{n} \phi^k = \frac{1-\phi^{n+1}}{1-\phi}$, and the related sum $\sum_{k=0}^{n} k\phi^{k-1} = \frac{d}{d\phi}\left[\sum_{k=0}^{n} \phi^k\right]$.)

(b) If n is large, argue that $Var(\bar{Y}) \approx \frac{\gamma_0}{n}\left[\frac{1+\phi}{1-\phi}\right]$.

(c) Plot $(1+\phi)/(1-\phi)$ for ϕ over the range −1 to +1. Interpret the plot in terms of the precision in estimating the process mean.

3.17 Verify Equation (3.2.6) on page 29. (Hint: You will need the fact that $\sum_{k=0}^{\infty} \phi^k = \frac{1}{1-\phi}$ for $-1 < \phi < +1$.)

3.18 Verify Equation (3.2.7) on page 30. (Hint: You will need the two sums $\sum_{t=1}^{n} t = \frac{n(n+1)}{2}$ and $\sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}$.)

CHAPTER 4

MODELS FOR STATIONARY TIME SERIES

This chapter discusses the basic concepts of a broad class of parametric time series models—the autoregressive moving average (ARMA) models. These models have assumed great importance in modeling real-world processes.

4.1 General Linear Processes

We will always let $\{Y_t\}$ denote the observed time series. From here on we will also let $\{e_t\}$ represent an unobserved white noise series, that is, a sequence of identically distributed, zero-mean, independent random variables. For much of our work, the assumption of independence could be replaced by the weaker assumption that the $\{e_t\}$ are uncorrelated random variables, but we will not pursue that slight generality.

A *general linear process*, $\{Y_t\}$, is one that can be represented as a weighted linear combination of present and past white noise terms as

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots \tag{4.1.1}$$

If the right-hand side of this expression is truly an infinite series, then certain conditions must be placed on the ψ-weights for the right-hand side to be meaningful mathematically. For our purposes, it suffices to assume that

$$\sum_{i=1}^{\infty} \psi_i^2 < \infty \tag{4.1.2}$$

We should also note that since $\{e_t\}$ is unobservable, there is no loss in the generality of Equation (4.1.2) if we assume that the coefficient on e_t is 1; effectively, $\psi_0 = 1$.

An important nontrivial example to which we will return often is the case where the ψ's form an exponentially decaying sequence

$$\psi_j = \phi^j$$

where ϕ is a number strictly between -1 and $+1$. Then

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots$$

For this example,

$$E(Y_t) = E(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots) = 0$$

so that $\{Y_t\}$ has a constant mean of zero. Also,

$$\begin{aligned} Var(Y_t) &= Var(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots) \\ &= Var(e_t) + \phi^2 Var(e_{t-1}) + \phi^4 Var(e_{t-2}) + \cdots \\ &= \sigma_e^2(1 + \phi^2 + \phi^4 + \cdots) \\ &= \frac{\sigma_e^2}{1-\phi^2} \text{ (by summing a geometric series)} \end{aligned}$$

Furthermore,

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots, e_{t-1} + \phi e_{t-2} + \phi^2 e_{t-3} + \cdots) \\ &= Cov(\phi e_{t-1}, e_{t-1}) + Cov(\phi^2 e_{t-2}, \phi e_{t-2}) + \cdots \\ &= \phi\sigma_e^2 + \phi^3\sigma_e^2 + \phi^5\sigma_e^2 + \cdots \\ &= \phi\sigma_e^2(1 + \phi^2 + \phi^4 + \cdots) \\ &= \frac{\phi\sigma_e^2}{1-\phi^2} \text{ (again summing a geometric series)} \end{aligned}$$

Thus

$$Corr(Y_t, Y_{t-1}) = \left[\frac{\phi\sigma_e^2}{1-\phi^2}\right] \Big/ \left[\frac{\sigma_e^2}{1-\phi^2}\right] = \phi$$

In a similar manner, we can find $Cov(Y_t, Y_{t-k}) = \dfrac{\phi^k\sigma_e^2}{1-\phi^2}$

and thus

$$Corr(Y_t, Y_{t-k}) = \phi^k \tag{4.1.3}$$

It is important to note that the process defined in this way is stationary—the autocovariance structure depends only on time lag and not on absolute time. For a general linear process, $Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots$, calculations similar to those done above yield the following results:

$$E(Y_t) = 0 \qquad \gamma_k = Cov(Y_t, Y_{t-k}) = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k} \qquad k \geq 0 \tag{4.1.4}$$

with $\psi_0 = 1$. A process with a nonzero mean μ may be obtained by adding μ to the right-hand side of Equation (4.1.1). Since the mean does not affect the covariance properties of a process, we assume a zero mean until we begin fitting models to data.

4.2 Moving Average Processes

In the case where only a finite number of the ψ-weights are nonzero, we have what is called a moving average process. In this case, we change notation[†] somewhat and write

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{4.2.1}$$

We call such a series a **moving average of order *q*** and abbreviate the name to MA(q). The terminology moving average arises from the fact that Y_t is obtained by applying the weights $1, -\theta_1, -\theta_2, \ldots, -\theta_q$ to the variables $e_t, e_{t-1}, e_{t-2}, \ldots, e_{t-q}$ and then moving the weights and applying them to $e_{t+1}, e_t, e_{t-1}, \ldots, e_{t-q+1}$ to obtain Y_{t+1} and so on. Moving average models were first considered by Slutsky (1927) and Wold (1938).

The First-Order Moving Average Process

We consider in detail the simple but nevertheless important moving average process of order 1, that is, the MA(1) series. Rather than specialize the formulas in Equation (4.1.4), it is instructive to rederive the results. The model is $Y_t = e_t - \theta e_{t-1}$. Since only one θ is involved, we drop the redundant subscript 1. Clearly $E(Y_t) = 0$ and $Var(Y_t) = \sigma_e^2(1 + \theta^2)$. Now

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(e_t - \theta e_{t-1}, e_{t-1} - \theta e_{t-2}) \\ &= Cov(-\theta e_{t-1}, e_{t-1}) = -\theta\sigma_e^2 \end{aligned}$$

and

$$\begin{aligned} Cov(Y_t, Y_{t-2}) &= Cov(e_t - \theta e_{t-1}, e_{t-2} - \theta e_{t-3}) \\ &= 0 \end{aligned}$$

since there are no e's with subscripts in common between Y_t and Y_{t-2}. Similarly, $Cov(Y_t, Y_{t-k}) = 0$ whenever $k \geq 2$; that is, *the process has no correlation beyond lag* 1. This fact will be important later when we need to choose suitable models for real data.

In summary, for an MA(1) model $Y_t = e_t - \theta e_{t-1}$,

$$\left.\begin{aligned} E(Y_t) &= 0 \\ \gamma_0 &= Var(Y_t) = \sigma_e^2(1 + \theta^2) \\ \gamma_1 &= -\theta\sigma_e^2 \\ \rho_1 &= (-\theta)/(1 + \theta^2) \\ \gamma_k &= \rho_k = 0 \quad \text{for } k \geq 2 \end{aligned}\right\} \tag{4.2.2}$$

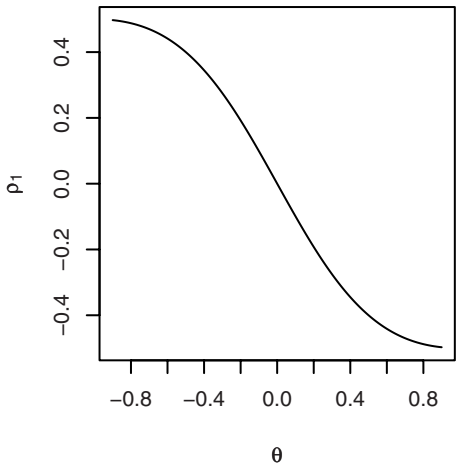
[†] The reason for this change will be evident later on. Some statistical software, for example R, uses plus signs before the thetas. Check with yours to see which convention it uses.

Some numerical values for ρ_1 versus θ in Equation (4.2.2) help illustrate the possibilities. Note that the ρ_1 values for negative θ can be obtained by simply negating the value given for the corresponding positive θ-value.

θ	$\rho_1 = -\theta/(1+\theta^2)$	θ	$\rho_1 = -\theta/(1+\theta^2)$
0.1	–0.099	0.6	–0.441
0.2	–0.192	0.7	–0.470
0.3	–0.275	0.8	–0.488
0.4	–0.345	0.9	–0.497
0.5	–0.400	1.0	–0.500

A calculus argument shows that the largest value that ρ_1 can attain is $\rho_1 = ½$ when $\theta = -1$ and the smallest value is $\rho_1 = -½$, which occurs when $\theta = +1$ (see Exercise 4.3). Exhibit 4.1 displays a graph of the lag 1 autocorrelation values for θ ranging from -1 to $+1$.

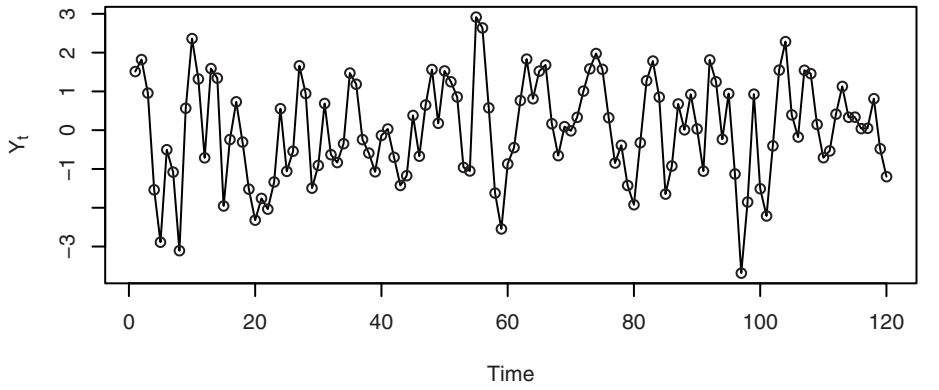
Exhibit 4.1 Lag 1 Autocorrelation of an MA(1) Process for Different θ



Exercise 4.4 asks you to show that when any nonzero value of θ is replaced by $1/\theta$, the *same* value for ρ_1 is obtained. For example, ρ_1 is the same for $\theta = ½$ as for $\theta = 1/(½) = 2$. If we knew that an MA(1) process had $\rho_1 = 0.4$, we still could not tell the precise value of θ. We will return to this troublesome point when we discuss *invertibility* in Section 4.5 on page 79.

Exhibit 4.2 shows a time plot of a simulated MA(1) series with $\theta = -0.9$ and normally distributed white noise. Recall from Exhibit 4.1 that $\rho_1 = 0.4972$ for this model; thus there is moderately strong positive correlation at lag 1. This correlation is evident in the plot of the series since consecutive observations tend to be closely related. If an observation is above the mean level of the series, then the next observation also tends to be above the mean. The plot is relatively smooth over time, with only occasional large fluctuations.

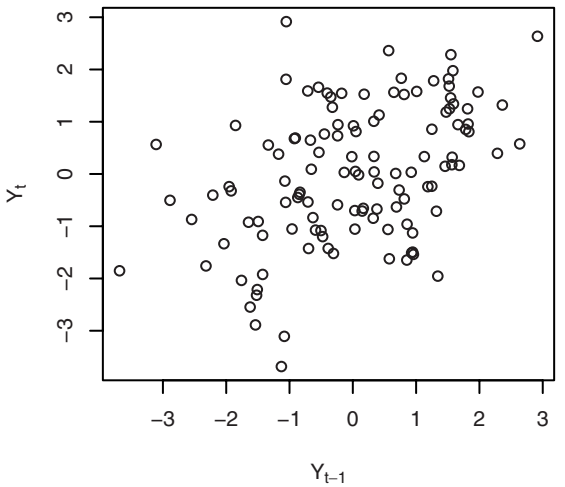
Exhibit 4.2 Time Plot of an MA(1) Process with $\theta = -0.9$




```
> win.graph(width=4.875,height=3,pointsize=8)
> data(ma1.2.s); plot(ma1.2.s,ylab=expression(Y[t]),type='o')
```

The lag 1 autocorrelation is even more apparent in Exhibit 4.3, which plots Y_t versus Y_{t-1}. Note the moderately strong upward trend in this plot.

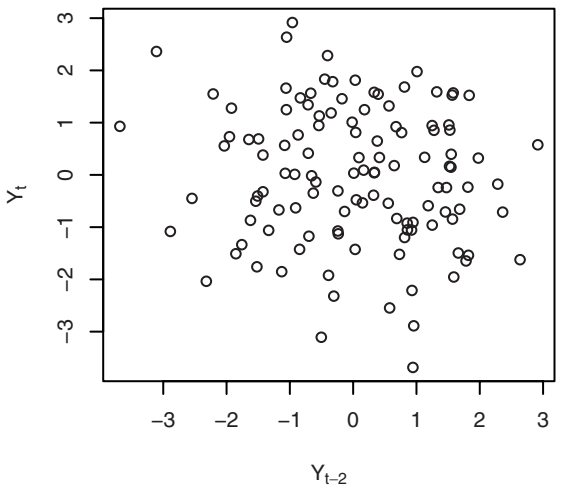
Exhibit 4.3 Plot of Y_t versus Y_{t-1} for MA(1) Series in Exhibit 4.2




```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=ma1.2.s,x=zlag(ma1.2.s),ylab=expression(Y[t]),
    xlab=expression(Y[t-1]),type='p')
```

The plot of Y_t versus Y_{t-2} in Exhibit 4.4 gives a strong visualization of the zero autocorrelation at lag 2 for this model.

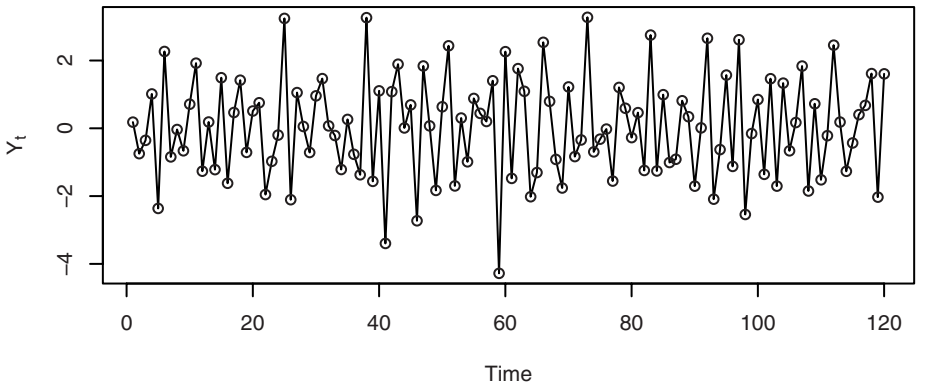
Exhibit 4.4 Plot of Y_t versus Y_{t-2} for MA(1) Series in Exhibit 4.2




```
> plot(y=ma1.2.s,x=zlag(ma1.2.s,2),ylab=expression(Y[t]),
    xlab=expression(Y[t-2]),type='p')
```

A somewhat different series is shown in Exhibit 4.5. This is a simulated MA(1) series with $\theta = +0.9$. Recall from Exhibit 4.1 that $\rho_1 = -0.497$ for this model; thus there is moderately strong negative correlation at lag 1. This correlation can be seen in the plot of the series since consecutive observations tend to be on opposite sides of the zero mean. If an observation is above the mean level of the series, then the next observation tends to be below the mean. The plot is quite jagged over time—especially when compared with the plot in Exhibit 4.2.

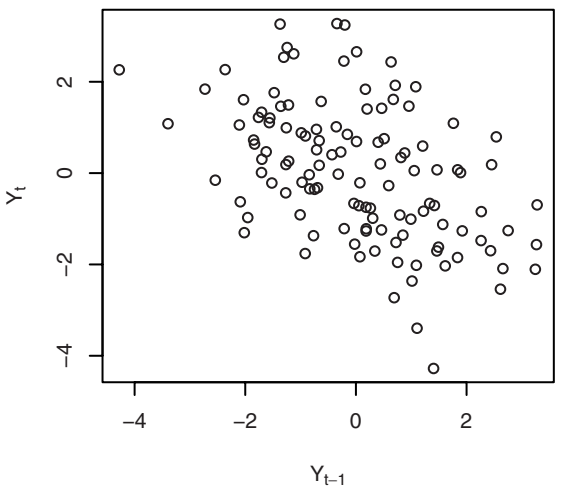
Exhibit 4.5 Time Plot of an MA(1) Process with $\theta = +0.9$




```
> win.graph(width=4.875,height=3,pointsize=8)
> data(ma1.1.s)
> plot(ma1.1.s,ylab=expression(Y[t]),type='o')
```

The negative lag 1 autocorrelation is even more apparent in the lag plot of Exhibit 4.6.

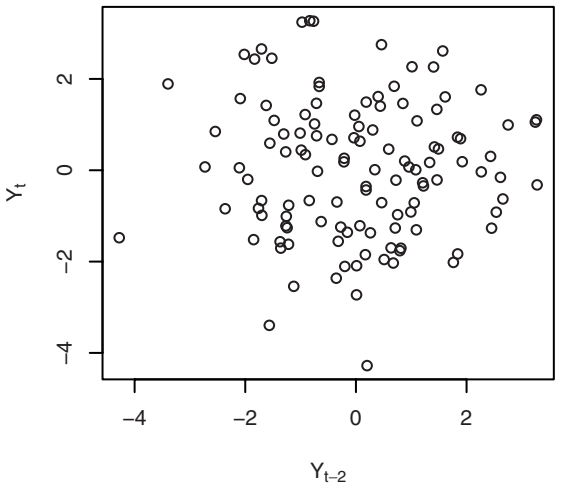
Exhibit 4.6 Plot of Y_t versus Y_{t-1} for MA(1) Series in Exhibit 4.5




```
> win.graph(width=3, height=3,pointsize=8)
> plot(y=ma1.1.s,x=zlag(ma1.1.s),ylab=expression(Y[t]),
    xlab=expression(Y[t-1]),type='p')
```

The plot of Y_t versus Y_{t-2} in Exhibit 4.7 displays the zero autocorrelation at lag 2 for this model.

Exhibit 4.7 Plot of Y_t versus Y_{t-2} for MA(1) Series in Exhibit 4.5




```
> plot(y=ma1.1.s,x=zlag(ma1.1.s,2),ylab=expression(Y[t]),
    xlab=expression(Y[t-2]),type='p')
```

MA(1) processes have no autocorrelation beyond lag 1, but by increasing the order of the process, we can obtain higher-order correlations.

The Second-Order Moving Average Process

Consider the moving average process of order 2:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

Here

$$\gamma_0 = Var(Y_t) = Var(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}) = (1 + \theta_1^2 + \theta_2^2)\sigma_e^2$$

$$\begin{aligned}\gamma_1 &= Cov(Y_t, Y_{t-1}) = Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3}) \\ &= Cov(-\theta_1 e_{t-1}, e_{t-1}) + Cov(-\theta_1 e_{t-2}, -\theta_2 e_{t-2}) \\ &= [-\theta_1 + (-\theta_1)(-\theta_2)]\sigma_e^2 \\ &= (-\theta_1 + \theta_1\theta_2)\sigma_e^2\end{aligned}$$

and

$$\begin{aligned}\gamma_2 &= Cov(Y_t, Y_{t-2}) = Cov(e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}, e_{t-2} - \theta_1 e_{t-3} - \theta_2 e_{t-4}) \\ &= Cov(-\theta_2 e_{t-2}, e_{t-2}) \\ &= -\theta_2 \sigma_e^2\end{aligned}$$

Thus, for an MA(2) process,

$$\begin{aligned}\rho_1 &= \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \\ \rho_2 &= \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2} \\ \rho_k &= 0 \quad \text{for } k = 3, 4,...\end{aligned} \tag{4.2.3}$$

For the specific case $Y_t = e_t - e_{t-1} + 0.6e_{t-2}$, we have

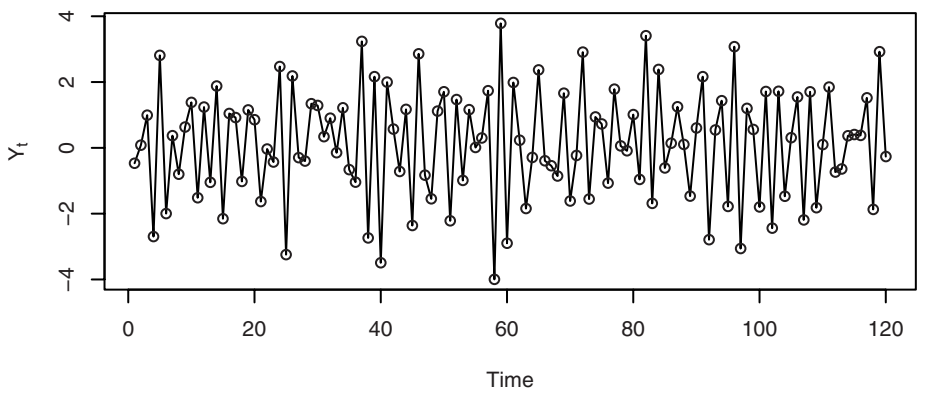
$$\rho_1 = \frac{-1 + (1)(-0.6)}{1 + (1)^2 + (-0.6)^2} = \frac{-1.6}{2.36} = -0.678$$

and

$$\rho_2 = \frac{0.6}{2.36} = 0.254$$

A time plot of a simulation of this MA(2) process is shown in Exhibit 4.8. The series tends to move back and forth across the mean in one time unit. This reflects the fairly strong negative autocorrelation at lag 1.

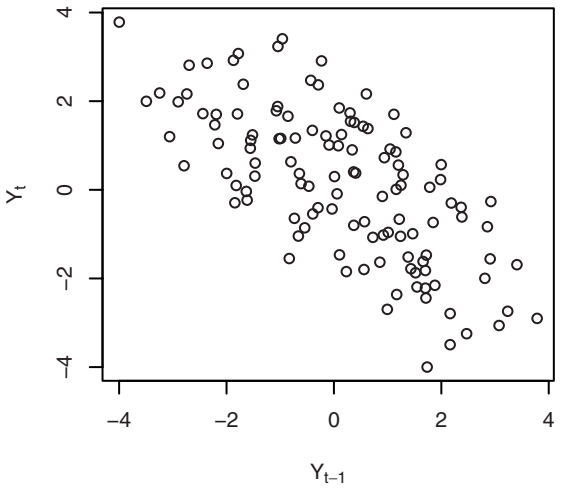
Exhibit 4.8 Time Plot of an MA(2) Process with $\theta_1 = 1$ and $\theta_2 = -0.6$




```
> win.graph(width=4.875, height=3,pointsize=8)
> data(ma2.s); plot(ma2.s,ylab=expression(Y[t]),type='o')
```

The plot in Exhibit 4.9 reflects that negative autocorrelation quite dramatically.

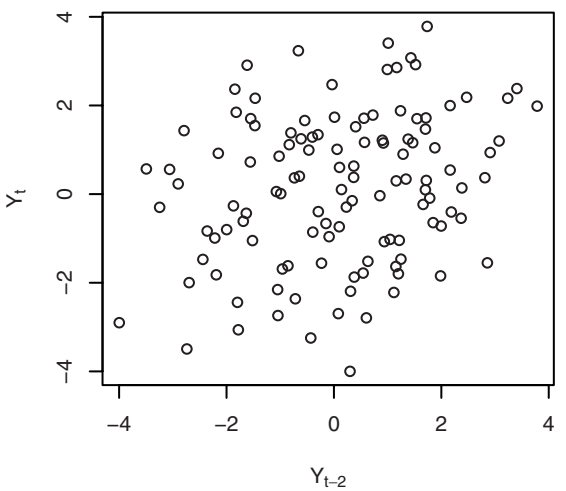
Exhibit 4.9 Plot of Y_t versus Y_{t-1} for MA(2) Series in Exhibit 4.8




```
> win.graph(width=3,height=3,pointsize=8)
> plot(y=ma2.s,x=zlag(ma2.s),ylab=expression(Y[t]),
    xlab=expression(Y[t-1]),type='p')
```

The weak positive autocorrelation at lag 2 is displayed in Exhibit 4.10.

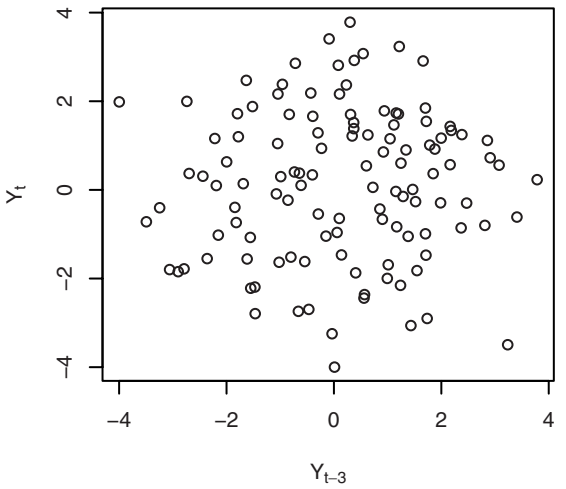
Exhibit 4.10 Plot of Y_t versus Y_{t-2} for MA(2) Series in Exhibit 4.8




```
> plot(y=ma2.s,x=zlag(ma2.s,2),ylab=expression(Y[t]),
    xlab=expression(Y[t-2]),type='p')
```

Finally, the lack of autocorrelation at lag 3 is apparent from the scatterplot in Exhibit 4.11.

Exhibit 4.11 Plot of Y_t versus Y_{t-3} for MA(2) Series in Exhibit 4.8




```
> plot(y=ma2.s,x=zlag(ma2.s,3),ylab=expression(Y[t]),
    xlab=expression(Y[t-3]),type='p')
```

The General MA(*q*) Process

For the general MA(q) process $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$, similar calculations show that

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_e^2 \tag{4.2.4}$$

and

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} & \text{for } k = 1, 2, \ldots, q \\ 0 \qquad \text{for } k > q \end{cases} \tag{4.2.5}$$

where the numerator of ρ_q is just $-\theta_q$. The autocorrelation function "cuts off" after lag q; that is, it is zero. Its shape can be almost anything for the earlier lags. Another type of process, the autoregressive process, provides models for alternative autocorrelation patterns.

4.3 Autoregressive Processes

Autoregressive processes are as their name suggests—regressions on themselves. Specifically, a pth-order **autoregressive process** $\{Y_t\}$ satisfies the equation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \tag{4.3.1}$$

The current value of the series Y_t is a linear combination of the p most recent past values of itself plus an "innovation" term e_t that incorporates everything new in the series at time t that is not explained by the past values. Thus, for every t, we assume that e_t is independent of $Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots$. Yule (1926) carried out the original work on autoregressive processes.[†]

The First-Order Autoregressive Process

Again, it is instructive to consider the first-order model, abbreviated AR(1), in detail. Assume the series is stationary and satisfies

$$Y_t = \phi Y_{t-1} + e_t \tag{4.3.2}$$

where we have dropped the subscript 1 from the coefficient ϕ for simplicity. As usual, in these initial chapters, we assume that the process mean has been subtracted out so that the series mean is zero. The conditions for stationarity will be considered later.

We first take variances of both sides of Equation (4.3.2) and obtain

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_e^2$$

Solving for γ_0 yields

$$\gamma_0 = \frac{\sigma_e^2}{1 - \phi^2} \tag{4.3.3}$$

Notice the immediate implication that $\phi^2 < 1$ or that $|\phi| < 1$. Now take Equation (4.3.2), multiply both sides by Y_{t-k} ($k = 1, 2, \ldots$), and take expected values

$$E(Y_{t-k} Y_t) = \phi E(Y_{t-k} Y_{t-1}) + E(e_t Y_{t-k})$$

or

$$\gamma_k = \phi \gamma_{k-1} + E(e_t Y_{t-k})$$

Since the series is assumed to be stationary with zero mean, and since e_t is independent of Y_{t-k}, we obtain

$$E(e_t Y_{t-k}) = E(e_t) E(Y_{t-k}) = 0$$

and so

[†] Recall that we are assuming that Y_t has zero mean. We can always introduce a nonzero mean by replacing Y_t by $Y_t - \mu$ throughout our equations.

$$\gamma_k = \phi\gamma_{k-1} \qquad \text{for } k = 1, 2, 3,... \tag{4.3.4}$$

Setting $k = 1$, we get $\gamma_1 = \phi\gamma_0 = \phi\sigma_e^2/(1-\phi^2)$. With $k = 2$, we obtain $\gamma_2 = \phi^2\sigma_e^2/(1-\phi^2)$. Now it is easy to see that in general

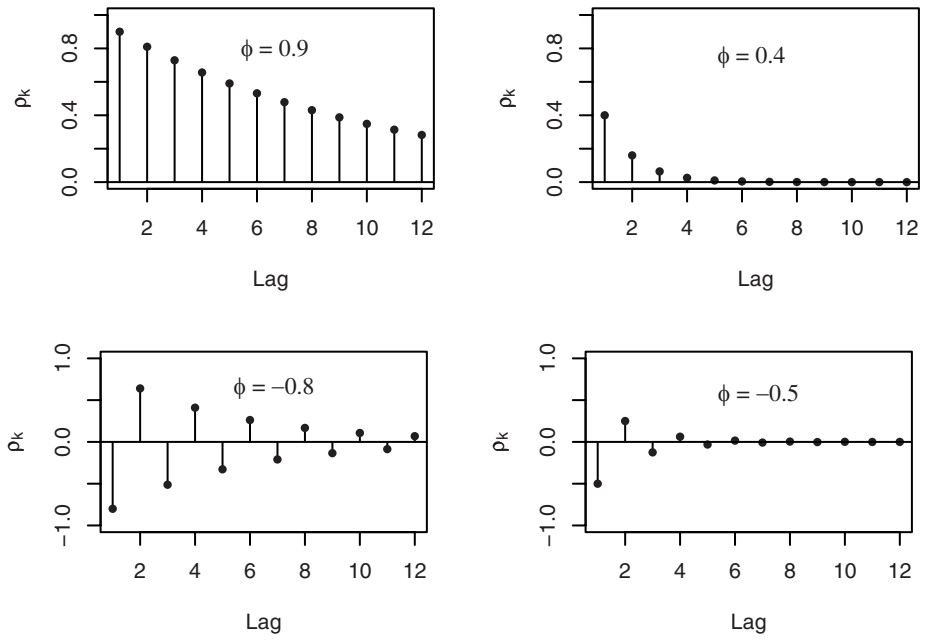
$$\gamma_k = \phi^k \frac{\sigma_e^2}{1-\phi^2} \tag{4.3.5}$$

and thus

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k \qquad \text{for } k = 1, 2, 3,... \tag{4.3.6}$$

Since $|\phi| < 1$, the magnitude of the autocorrelation function decreases exponentially as the number of lags, k, increases. If $0 < \phi < 1$, all correlations are positive; if $-1 < \phi < 0$, the lag 1 autocorrelation is negative ($\rho_1 = \phi$) and the signs of successive autocorrelations alternate from positive to negative, with their magnitudes decreasing exponentially. Portions of the graphs of several autocorrelation functions are displayed in Exhibit 4.12.

Exhibit 4.12 Autocorrelation Functions for Several AR(1) Models

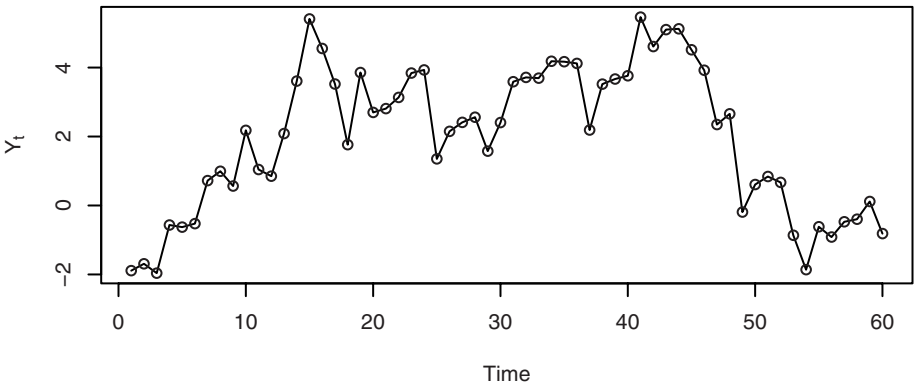


Notice that for ϕ near ± 1, the exponential decay is quite slow (for example, $(0.9)^6 = 0.53$), but for smaller ϕ, the decay is quite rapid (for example, $(0.4)^6 = 0.00410$). With ϕ near ± 1, the strong correlation will extend over many lags and produce a relatively

smooth series if ϕ is positive and a very jagged series if ϕ is negative.

Exhibit 4.13 displays the time plot of a simulated AR(1) process with $\phi = 0.9$. Notice how infrequently the series crosses its theoretical mean of zero. There is a lot of inertia in the series—it hangs together, remaining on the same side of the mean for extended periods. An observer might claim that the series has several trends. We know that in fact the theoretical mean is zero for all time points. The illusion of trends is due to the strong autocorrelation of neighboring values of the series.

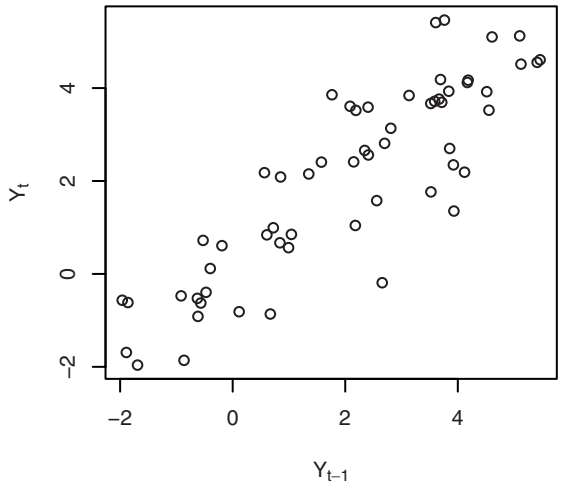
Exhibit 4.13 Time Plot of an AR(1) Series with $\phi = 0.9$




```
> win.graph(width=4.875, height=3,pointsize=8)
> data(ar1.s); plot(ar1.s,ylab=expression(Y[t]),type='o')
```

The smoothness of the series and the strong autocorrelation at lag 1 are depicted in the lag plot shown in Exhibit 4.14.

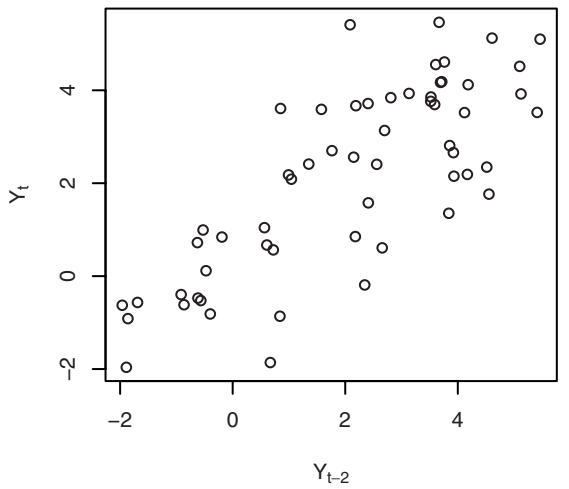
Exhibit 4.14 Plot of Y_t versus Y_{t-1} for AR(1) Series of Exhibit 4.13




```
> win.graph(width=3, height=3,pointsize=8)
> plot(y=ar1.s,x=zlag(ar1.s),ylab=expression(Y[t]),
    xlab=expression(Y[t-1]),type='p')
```

This AR(1) model also has strong positive autocorrelation at lag 2, namely $\rho_2 = (0.9)^2 = 0.81$. Exhibit 4.15 shows this quite well.

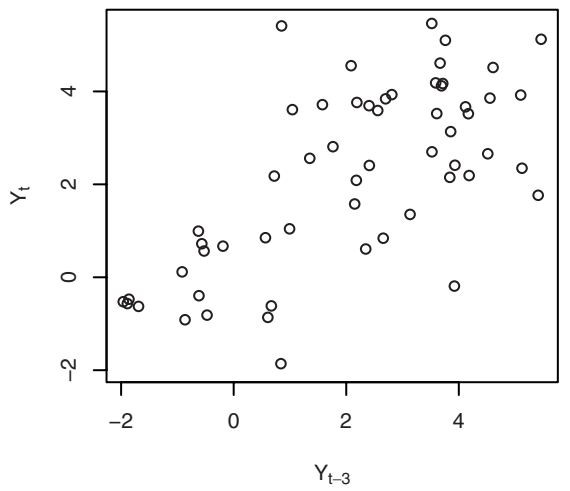
Exhibit 4.15 Plot of Y_t versus Y_{t-2} for AR(1) Series of Exhibit 4.13




```
> plot(y=ar1.s,x=zlag(ar1.s,2),ylab=expression(Y[t]),
    xlab=expression(Y[t-2]),type='p')
```

Finally, at lag 3, the autocorrelation is still quite high: $\rho_3 = (0.9)^3 = 0.729$. Exhibit 4.16 confirms this for this particular series.

Exhibit 4.16 Plot of Y_t versus Y_{t-3} for AR(1) Series of Exhibit 4.13




```
> plot(y=ar1.s,x=zlag(ar1.s,3),ylab=expression(Y[t]),
    xlab=expression(Y[t-3]),type='p')
```

The General Linear Process Version of the AR(1) Model

The recursive definition of the AR(1) process given in Equation (4.3.2) is extremely useful for interpretating the model. For other purposes, it is convenient to express the AR(1) model as a general linear process as in Equation (4.1.1). The recursive definition is valid for all t. If we use this equation with t replaced by $t-1$, we get $Y_{t-1} = \phi Y_{t-2} + e_{t-1}$. Substituting this into the original expression gives

$$\begin{aligned} Y_t &= \phi(\phi Y_{t-2} + e_{t-1}) + e_t \\ &= e_t + \phi e_{t-1} + \phi^2 Y_{t-2} \end{aligned}$$

If we repeat this substitution into the past, say $k-1$ times, we get

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \cdots + \phi^{k-1} e_{t-k+1} + \phi^k Y_{t-k} \tag{4.3.7}$$

Assuming $|\phi| < 1$ and letting k increase without bound, it seems reasonable (this is almost a rigorous proof) that we should obtain the infinite series representation

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots \tag{4.3.8}$$

This is in the form of the general linear process of Equation (4.1.1) with $\psi_j = \phi^j$, which we already investigated in Section 4.1 on page 55. Note that this representation reemphasizes the need for the restriction $|\phi| < 1$.

Stationarity of an AR(1) Process

It can be shown that, subject to the restriction that e_t be independent of Y_{t-1}, Y_{t-2}, Y_{t-3},… and that $\sigma_e^2 > 0$, the solution of the AR(1) defining recursion $Y_t = \phi Y_{t-1} + e_t$ will be stationary if and only if $|\phi| < 1$. The requirement $|\phi| < 1$ is usually called the **stationarity condition** for the AR(1) process (See Box, Jenkins, and Reinsel, 1994, p. 54; Nelson, 1973, p. 39; and Wei, 2005, p. 32) even though more than stationarity is involved. See especially Exercises 4.16, 4.18, and 4.25.

At this point, we should note that the autocorrelation function for the AR(1) process has been derived in two different ways. The first method used the general linear process representation leading up to Equation (4.1.3). The second method used the defining recursion $Y_t = \phi Y_{t-1} + e_t$ and the development of Equations (4.3.4), (4.3.5), and (4.3.6). A third derivation is obtained by multiplying both sides of Equation (4.3.7) by Y_{t-k}, taking expected values of both sides, and using the fact that e_t, e_{t-1}, e_{t-2},…, $e_{t-(k-1)}$ are independent of Y_{t-k}. The second method should be especially noted since it will generalize nicely to higher-order processes.

The Second-Order Autoregressive Process

Now consider the series satisfying

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \tag{4.3.9}$$

where, as usual, we assume that e_t is independent of Y_{t-1}, Y_{t-2}, Y_{t-3},… . To discuss stationarity, we introduce the **AR characteristic polynomial**

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2$$

and the corresponding **AR characteristic equation**

$$1 - \phi_1 x - \phi_2 x^2 = 0$$

We recall that a quadratic equation always has two roots (possibly complex).

Stationarity of the AR(2) Process

It may be shown that, subject to the condition that e_t is independent of Y_{t-1}, Y_{t-2}, Y_{t-3},..., a stationary solution to Equation (4.3.9) exists if and only if the roots of the AR characteristic equation exceed 1 in absolute value (modulus). We sometimes say that the roots should lie outside the unit circle in the complex plane. This statement will generalize to the *p*th-order case without change.†

† It also applies in the first-order case, where the AR characteristic equation is just $1 - \phi x = 0$ with root $1/\phi$, which exceeds 1 in absolute value if and only if $|\phi| < 1$.

In the second-order case, the roots of the quadratic characteristic equation are easily found to be

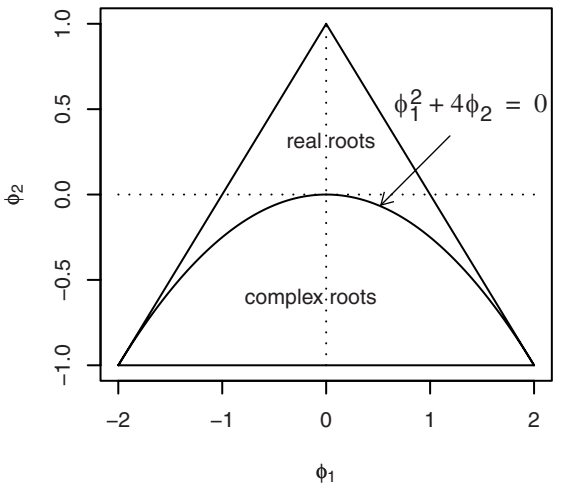
$$\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \tag{4.3.10}$$

For stationarity, we require that these roots exceed 1 in absolute value. In Appendix B, page 84, we show that this will be true if and only if three conditions are satisfied:

$$\phi_1 + \phi_2 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad \text{and} \qquad |\phi_2| < 1 \tag{4.3.11}$$

As with the AR(1) model, we call these the **stationarity conditions** for the AR(2) model. This stationarity region is displayed in Exhibit 4.17.

Exhibit 4.17 Stationarity Parameter Region for AR(2) Process



The Autocorrelation Function for the AR(2) Process

To derive the autocorrelation function for the AR(2) case, we take the defining recursive relationship of Equation (4.3.9), multiply both sides by Y_{t-k}, and take expectations. Assuming stationarity, zero means, and that e_t is independent of Y_{t-k}, we get

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} \qquad \text{for } k = 1, 2, 3, \ldots \tag{4.3.12}$$

or, dividing through by γ_0,

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} \qquad \text{for } k = 1, 2, 3, \ldots \tag{4.3.13}$$

Equations (4.3.12) and/or (4.3.13) are usually called the **Yule-Walker equations**, especially the set of two equations obtained for $k = 1$ and 2. Setting $k = 1$ and using $\rho_0 = 1$ and $\rho_{-1} = \rho_1$, we get $\rho_1 = \phi_1 + \phi_2\rho_1$ and so

$$\rho_1 = \frac{\phi_1}{1-\phi_2} \tag{4.3.14}$$

Using the now known values for ρ_1 (and ρ_0), Equation (4.3.13) can be used with $k = 2$ to obtain

$$\begin{aligned}\rho_2 &= \phi_1\rho_1 + \phi_2\rho_0 \\ &= \frac{\phi_2(1-\phi_2)+\phi_1^2}{1-\phi_2}\end{aligned} \tag{4.3.15}$$

Successive values of ρ_k may be easily calculated numerically from the recursive relationship of Equation (4.3.13).

Although Equation (4.3.13) is very efficient for calculating autocorrelation values numerically from given values of ϕ_1 and ϕ_2, for other purposes it is desirable to have a more explicit formula for ρ_k. The form of the explicit solution depends critically on the roots of the characteristic equation $1 - \phi_1 x - \phi_2 x^2 = 0$. Denoting the reciprocals of these roots by G_1 and G_2, it is shown in Appendix B, page 84, that

$$G_1 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} \qquad \text{and} \qquad G_2 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

For the case $G_1 \neq G_2$, it can be shown that we have

$$\rho_k = \frac{(1-G_2^2)G_1^{k+1} - (1-G_1^2)G_2^{k+1}}{(G_1 - G_2)(1 + G_1G_2)} \qquad \text{for } k \geq 0 \tag{4.3.16}$$

If the roots are complex (that is, if $\phi_1^2 + 4\phi_2 < 0$), then ρ_k may be rewritten as

$$\rho_k = R^k \frac{\sin(\Theta k + \Phi)}{\sin(\Phi)} \qquad \text{for } k \geq 0 \tag{4.3.17}$$

where $R = \sqrt{-\phi_2}$ and Θ and Φ are defined by $\cos(\Theta) = \phi_1/(2\sqrt{-\phi_2})$ and $\tan(\Phi) = [(1-\phi_2)/(1+\phi_2)]$.

For completeness, we note that if the roots are equal ($\phi_1^2 + 4\phi_2 = 0$), then we have

$$\rho_k = \left(1 + \frac{1+\phi_2}{1-\phi_2}k\right)\left(\frac{\phi_1}{2}\right)^k \qquad \text{for } k = 0, 1, 2, \ldots \tag{4.3.18}$$

A good discussion of the derivations of these formulas can be found in Fuller (1996, Section 2.5).

The specific details of these formulas are of little importance to us. We need only note that the autocorrelation function can assume a wide variety of shapes. In all cases, the magnitude of ρ_k dies out exponentially fast as the lag k increases. In the case of complex roots, ρ_k displays a damped sine wave behavior with **damping factor** R, $0 \leq R < 1$, **frequency** Θ, and **phase** Φ. Illustrations of the possible shapes are given in Exhibit 4.18. (The R function `ARMAacf` discussed on page 450 is useful for plotting.)

Exhibit 4.18 Autocorrelation Functions for Several AR(2) Models

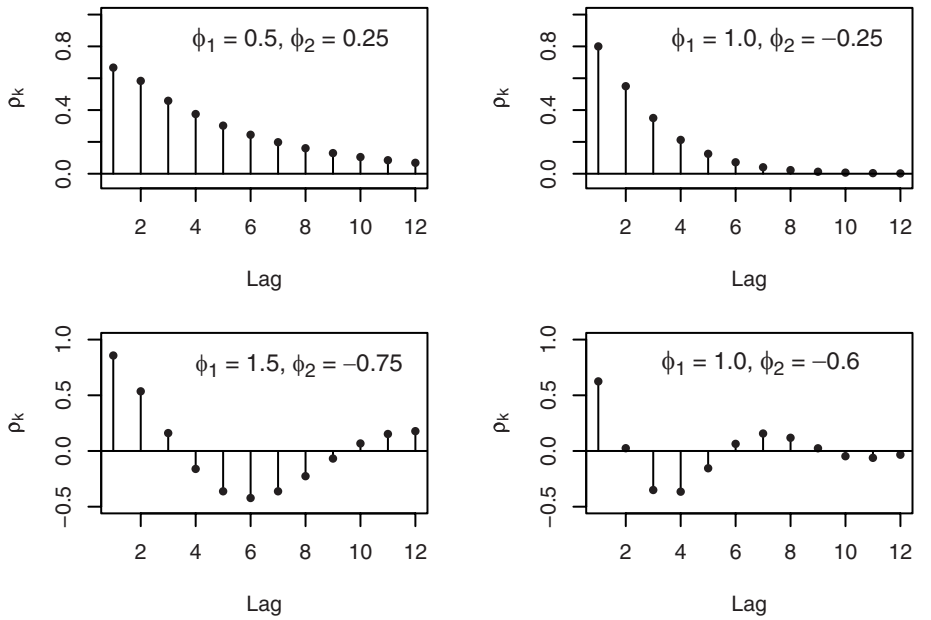
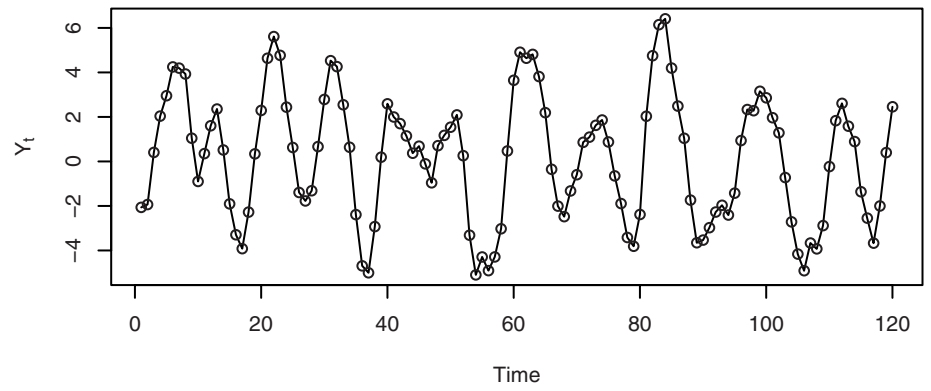


Exhibit 4.19 displays the time plot of a simulated AR(2) series with $\phi_1 = 1.5$ and $\phi_2 = -0.75$. The periodic behavior of ρ_k shown in Exhibit 4.18 is clearly reflected in the nearly periodic behavior of the series with the same period of 360/30 = 12 time units. If Θ is measured in radians, $2\pi/\Theta$ is sometimes called the **quasi-period** of the AR(2) process.

Exhibit 4.19 Time Plot of an AR(2) Series with $\phi_1 = 1.5$ and $\phi_2 = -0.75$




```
> win.graph(width=4.875,height=3,pointsize=8)
> data(ar2.s); plot(ar2.s,ylab=expression(Y[t]),type='o')
```

The Variance for the AR(2) Model

The process variance γ_0 can be expressed in terms of the model parameters ϕ_1, ϕ_2, and σ_e^2 as follows: Taking the variance of both sides of Equation (4.3.9) yields

$$\gamma_0 = (\phi_1^2 + \phi_2^2)\gamma_0 + 2\phi_1\phi_2\gamma_1 + \sigma_e^2 \tag{4.3.19}$$

Setting $k = 1$ in Equation (4.3.12) gives a second linear equation for γ_0 and γ_1, $\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1$, which can be solved simultaneously with Equation (4.3.19) to obtain

$$\begin{aligned}\gamma_0 &= \frac{(1-\phi_2)\sigma_e^2}{(1-\phi_2)(1-\phi_1^2-\phi_2^2)-2\phi_2\phi_1^2} \\ &= \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_e^2}{(1-\phi_2)^2-\phi_1^2}\end{aligned} \tag{4.3.20}$$

The ψ-Coefficients for the AR(2) Model

The ψ-coefficients in the general linear process representation for an AR(2) series are more complex than for the AR(1) case. However, we can substitute the general linear process representation using Equation (4.1.1) for Y_t, for Y_{t-1}, and for Y_{t-2} into $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$. If we then equate coefficients of e_j, we get the recursive relationships

$$\left.\begin{aligned}\psi_0 &= 1 \\ \psi_1 - \phi_1\psi_0 &= 0 \\ \psi_j - \phi_1\psi_{j-1} - \phi_2\psi_{j-2} &= 0 \quad \text{for } j = 2, 3, \ldots\end{aligned}\right\} \tag{4.3.21}$$

These may be solved recursively to obtain $\psi_0 = 1$, $\psi_1 = \phi_1$, $\psi_2 = \phi_1^2 + \phi_2$, and so on. These relationships provide excellent numerical solutions for the ψ-coefficients for given numerical values of ϕ_1 and ϕ_2.

One can also show that, for $G_1 \neq G_2$, an explicit solution is

$$\psi_j = \frac{G_1^{j+1} - G_2^{j+1}}{G_1 - G_2} \tag{4.3.22}$$

where, as before, G_1 and G_2 are the reciprocals of the roots of the AR characteristic equation. If the roots are complex, Equation (4.3.22) may be rewritten as

$$\psi_j = R^j\left\{\frac{\sin[(j+1)\Theta]}{\sin(\Theta)}\right\} \tag{4.3.23}$$

a damped sine wave with the same damping factor R and frequency Θ as in Equation (4.3.17) for the autocorrelation function.

For completeness, we note that if the roots are equal, then

$$\psi_j = (1+j)\phi_1^j \tag{4.3.24}$$

The General Autoregressive Process

Consider now the pth-order autoregressive model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \tag{4.3.25}$$

with AR characteristic polynomial

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \tag{4.3.26}$$

and corresponding AR characteristic equation

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0 \tag{4.3.27}$$

As noted earlier, assuming that e_t is independent of $Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots$ a stationary solution to Equation (4.3.27) exists if and only if the p roots of the AR characteristic equation each exceed 1 in absolute value (modulus). Other relationships between polynomial roots and coefficients may be used to show that the following two inequalities are necessary for stationarity. That is, for the roots to be greater than 1 in modulus, it is necessary, but not sufficient, that both

$$\left.\begin{array}{l} \phi_1 + \phi_2 + \cdots + \phi_p < 1 \\ \text{and} \qquad |\phi_p| < 1 \end{array}\right\} \tag{4.3.28}$$

Assuming stationarity and zero means, we may multiply Equation (4.3.25) by Y_{t-k}, take expectations, divide by γ_0, and obtain the important recursive relationship

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \phi_3 \rho_{k-3} + \cdots + \phi_p \rho_{k-p} \qquad \text{for } k \geq 1 \tag{4.3.29}$$

Putting $k = 1, 2, \ldots$, and p into Equation (4.3.29) and using $\rho_0 = 1$ and $\rho_{-k} = \rho_k$, we get the general **Yule-Walker equations**

$$\left.\begin{array}{l} \rho_1 = \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 + \cdots + \phi_p \rho_{p-1} \\ \rho_2 = \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 + \cdots + \phi_p \rho_{p-2} \\ \quad \vdots \\ \rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \phi_3 \rho_{p-3} + \cdots + \phi_p \end{array}\right\} \tag{4.3.30}$$

Given numerical values for $\phi_1, \phi_2, \ldots, \phi_p$, these linear equations can be solved to obtain numerical values for $\rho_1, \rho_2, \ldots, \rho_p$. Then Equation (4.3.29) can be used to obtain numerical values for ρ_k at any number of higher lags.

Noting that

$$E(e_t Y_t) = E[e_t(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t)] = E(e_t^2) = \sigma_e^2$$

we may multiply Equation (4.3.25) by Y_t, take expectations, and find

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \cdots + \phi_p \gamma_p + \sigma_e^2$$

which, using $\rho_k = \gamma_k/\gamma_0$, can be written as

$$\gamma_0 = \frac{\sigma_e^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \cdots - \phi_p\rho_p} \tag{4.3.31}$$

and express the process variance γ_0 in terms of the parameters σ_e^2, $\phi_1, \phi_2, \ldots, \phi_p$, and the now known values of $\rho_1, \rho_2, \ldots, \rho_p$. Of course, explicit solutions for ρ_k are essentially impossible in this generality, but we can say that ρ_k will be a linear combination of exponentially decaying terms (corresponding to the real roots of the characteristic equation) and damped sine wave terms (corresponding to the complex roots of the characteristic equation).

Assuming stationarity, the process can also be expressed in the general linear process form of Equation (4.1.1), but the ψ-coefficients are complicated functions of the parameters $\phi_1, \phi_2, \ldots, \phi_p$. The coefficients can be found numerically; see Appendix C on page 85.

4.4 The Mixed Autoregressive Moving Average Model

If we assume that the series is partly autoregressive and partly moving average, we obtain a quite general time series model. In general, if

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{4.4.1}$$

we say that $\{Y_t\}$ is a mixed **autoregressive moving average** process of orders p and q, respectively; we abbreviate the name to ARMA(p,q). As usual, we discuss an important special case first.[†]

The ARMA(1,1) Model

The defining equation can be written

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \tag{4.4.2}$$

To derive Yule-Walker type equations, we first note that

$$\begin{aligned} E(e_t Y_t) &= E[e_t(\phi Y_{t-1} + e_t - \theta e_{t-1})] \\ &= \sigma_e^2 \end{aligned}$$

and

[†] In mixed models, we assume that there are no common factors in the autoregressive and moving average polynomials. If there were, we could cancel them and the model would reduce to an ARMA model of lower order. For ARMA(1,1), this means $\theta \neq \phi$.

$$\begin{aligned} E(e_{t-1}Y_t) &= E[e_{t-1}(\phi Y_{t-1} + e_t - \theta e_{t-1})] \\ &= \phi\sigma_e^2 - \theta\sigma_e^2 \\ &= (\phi - \theta)\sigma_e^2 \end{aligned}$$

If we multiply Equation (4.4.2) by Y_{t-k} and take expectations, we have

$$\left.\begin{aligned} \gamma_0 &= \phi\gamma_1 + [1 - \theta(\phi - \theta)]\sigma_e^2 \\ \gamma_1 &= \phi\gamma_0 - \theta\sigma_e^2 \\ \gamma_k &= \phi\gamma_{k-1} \quad \text{for } k \geq 2 \end{aligned}\right\} \tag{4.4.3}$$

Solving the first two equations yields

$$\gamma_0 = \frac{(1 - 2\phi\theta + \theta^2)}{1 - \phi^2}\sigma_e^2 \tag{4.4.4}$$

and solving the simple recursion gives

$$\rho_k = \frac{(1 - \theta\phi)(\phi - \theta)}{1 - 2\theta\phi + \theta^2}\phi^{k-1} \quad \text{for } k \geq 1 \tag{4.4.5}$$

Note that this autocorrelation function decays exponentially as the lag k increases. The damping factor is ϕ, but the decay starts from initial value ρ_1, which also depends on θ. This is in contrast to the AR(1) autocorrelation, which also decays with damping factor ϕ but always from initial value $\rho_0 = 1$. For example, if $\phi = 0.8$ and $\theta = 0.4$, then $\rho_1 = 0.523$, $\rho_2 = 0.418$, $\rho_3 = 0.335$, and so on. Several shapes for ρ_k are possible, depending on the sign of ρ_1 and the sign of ϕ.

The general linear process form of the model can be obtained in the same manner that led to Equation (4.3.8). We find

$$Y_t = e_t + (\phi - \theta)\sum_{j=1}^{\infty}\phi^{j-1}e_{t-j} \ , \tag{4.4.6}$$

that is,

$$\psi_j = (\phi - \theta)\phi^{j-1} \qquad \text{for } j \geq 1$$

We should now mention the obvious stationarity condition $|\phi| < 1$, or equivalently the root of the AR characteristic equation $1 - \phi x = 0$ must exceed unity in absolute value.

For the general ARMA(p,q) model, we state the following facts without proof: Subject to the condition that e_t is independent of $Y_{t-1}, Y_{t-2}, Y_{t-3},\ldots$, a stationary solution to Equation (4.4.1) exists if and only if all the roots of the AR characteristic equation $\phi(x) = 0$ exceed unity in modulus.

If the stationarity conditions are satisfied, then the model can also be written as a general linear process with ψ-coefficients determined from

$$\left.\begin{aligned}
\psi_0 &= 1 \\
\psi_1 &= -\theta_1 + \phi_1 \\
\psi_2 &= -\theta_2 + \phi_2 + \phi_1\psi_1 \\
&\vdots \\
\psi_j &= -\theta_j + \phi_p\psi_{j-p} + \phi_{p-1}\psi_{j-p+1} + \cdots + \phi_1\psi_{j-1}
\end{aligned}\right\} \tag{4.4.7}$$

where we take $\psi_j = 0$ for $j < 0$ and $\theta_j = 0$ for $j > q$.

Again assuming stationarity, the autocorrelation function can easily be shown to satisfy

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p} \qquad \text{for } k > q \tag{4.4.8}$$

Similar equations can be developed for $k = 1, 2, 3, \ldots, q$ that involve $\theta_1, \theta_2, \ldots, \theta_q$. An algorithm suitable for numerical computation of the complete autocorrelation function is given in Appendix C on page 85. (This algorithm is implemented in the R function named `ARMAacf`.)

4.5 Invertibility

We have seen that for the MA(1) process we get exactly the same autocorrelation function if θ is replaced by $1/\theta$. In the exercises, we find a similar problem with nonuniqueness for the MA(2) model. This lack of uniqueness of MA models, given their autocorrelation functions, must be addressed before we try to infer the values of parameters from observed time series. It turns out that this nonuniqueness is related to the seemingly unrelated question stated next.

An autoregressive process can always be reexpressed as a general linear process through the ψ-coefficients so that an AR process may also be thought of as an infinite-order moving average process. However, for some purposes, the autoregressive representations are also convenient. Can a moving average model be reexpressed as an autoregression?

To fix ideas, consider an MA(1) model:

$$Y_t = e_t - \theta e_{t-1} \tag{4.5.1}$$

First rewriting this as $e_t = Y_t + \theta e_{t-1}$ and then replacing t by $t - 1$ and substituting for e_{t-1} above, we get

$$\begin{aligned}
e_t &= Y_t + \theta(Y_{t-1} + \theta e_{t-2}) \\
&= Y_t + \theta Y_{t-1} + \theta^2 e_{t-2}
\end{aligned}$$

If $|\theta| < 1$, we may continue this substitution "infinitely" into the past and obtain the expression [compare with Equations (4.3.7) and (4.3.8)]

$$e_t = Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \cdots$$

or

$$Y_t = (-\theta Y_{t-1} - \theta^2 Y_{t-2} - \theta^3 Y_{t-3} - \cdots) + e_t \tag{4.5.2}$$

If $|\theta| < 1$, we see that the MA(1) model can be inverted into an infinite-order autoregressive model. We say that the MA(1) model is invertible if and only if $|\theta| < 1$.

For a general MA(q) or ARMA(p,q) model, we define the **MA characteristic polynomial** as

$$\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \theta_3 x^3 - \cdots - \theta_q x^q \tag{4.5.3}$$

and the corresponding **MA characteristic equation**

$$1 - \theta_1 x - \theta_2 x^2 - \theta_3 x^3 - \cdots - \theta_q x^q = 0 \tag{4.5.4}$$

It can be shown that the MA(q) model is **invertible**; that is, there are coefficients π_j such that

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t \tag{4.5.5}$$

if and only if the roots of the MA characteristic equation exceed 1 in modulus. (Compare this with stationarity of an AR model.)

It may also be shown that there is only one set of parameter values that yield an invertible MA process with a given autocorrelation function. For example, $Y_t = e_t + 2e_{t-1}$ and $Y_t = e_t + \frac{1}{2}e_{t-1}$ both have the same autocorrelation function, but only the second one with root -2 is invertible. From here on, we will restrict our attention to the physically sensible class of invertible models.

For a general ARMA(p,q) model, we require both stationarity and invertibility.

4.6 Summary

This chapter introduces the simple but very useful autoregressive, moving average (ARMA) time series models. The basic statistical properties of these models were derived in particular for the important special cases of moving averages of orders 1 and 2 and autoregressive processes of orders 1 and 2. Stationarity and invertibility issues have been pursued for these cases. Properties of mixed ARMA models have also been investigated. You should be well-versed in the autocorrelation properties of these models and the various representations of the models.

EXERCISES

4.1 Use first principles to find the autocorrelation function for the stationary process defined by

$$Y_t = 5 + e_t - \tfrac{1}{2}e_{t-1} + \tfrac{1}{4}e_{t-2}$$

4.2 Sketch the autocorrelation functions for the following MA(2) models with parameters as specified:
(a) $\theta_1 = 0.5$ and $\theta_2 = 0.4$.
(b) $\theta_1 = 1.2$ and $\theta_2 = -0.7$.
(c) $\theta_1 = -1$ and $\theta_2 = -0.6$.

4.3 Verify that for an MA(1) process

$$\max_{-\infty < \theta < \infty} \rho_1 = 0.5 \qquad \text{and} \qquad \min_{-\infty < \theta < \infty} \rho_1 = -0.5$$

4.4 Show that when θ is replaced by $1/\theta$, the autocorrelation function for an MA(1) process does not change.

4.5 Calculate and sketch the autocorrelation functions for each of the following AR(1) models. Plot for sufficient lags that the autocorrelation function has nearly died out.
(a) $\phi_1 = 0.6$.
(b) $\phi_1 = -0.6$.
(c) $\phi_1 = 0.95$. (Do out to 20 lags.)
(d) $\phi_1 = 0.3$.

4.6 Suppose that $\{Y_t\}$ is an AR(1) process with $-1 < \phi < +1$.
(a) Find the autocovariance function for $W_t = \nabla Y_t = Y_t - Y_{t-1}$ in terms of ϕ and σ_e^2.
(b) In particular, show that $Var(W_t) = 2\sigma_e^2/(1+\phi)$.

4.7 Describe the important characteristics of the autocorrelation function for the following models: (a) MA(1), (b) MA(2), (c) AR(1), (d) AR(2), and (e) ARMA(1,1).

4.8 Let $\{Y_t\}$ be an AR(2) process of the special form $Y_t = \phi_2 Y_{t-2} + e_t$. Use first principles to find the range of values of ϕ_2 for which the process is stationary.

4.9 Use the recursive formula of Equation (4.3.13) to calculate and then sketch the autocorrelation functions for the following AR(2) models with parameters as specified. In each case, specify whether the roots of the characteristic equation are real or complex. If the roots are complex, find the damping factor, R, and frequency, Θ, for the corresponding autocorrelation function when expressed as in Equation (4.3.17), on page 73.
(a) $\phi_1 = 0.6$ and $\phi_2 = 0.3$.
(b) $\phi_1 = -0.4$ and $\phi_2 = 0.5$.
(c) $\phi_1 = 1.2$ and $\phi_2 = -0.7$.
(d) $\phi_1 = -1$ and $\phi_2 = -0.6$.
(e) $\phi_1 = 0.5$ and $\phi_2 = -0.9$.
(f) $\phi_1 = -0.5$ and $\phi_2 = -0.6$.

4.10 Sketch the autocorrelation functions for each of the following ARMA models:
(a) ARMA(1,1) with $\phi = 0.7$ and $\theta = 0.4$.
(b) ARMA(1,1) with $\phi = 0.7$ and $\theta = -0.4$.

4.11 For the ARMA(1,2) model $Y_t = 0.8Y_{t-1} + e_t + 0.7e_{t-1} + 0.6e_{t-2}$, show that
(a) $\rho_k = 0.8\rho_{k-1}$ for $k > 2$.
(b) $\rho_2 = 0.8\rho_1 + 0.6\sigma_e^2/\gamma_0$.

4.12 Consider two MA(2) processes, one with $\theta_1 = \theta_2 = 1/6$ and another with $\theta_1 = -1$ and $\theta_2 = 6$.
(a) Show that these processes have the same autocorrelation function.
(b) How do the roots of the corresponding characteristic polynomials compare?

4.13 Let $\{Y_t\}$ be a stationary process with $\rho_k = 0$ for $k > 1$. Show that we must have $|\rho_1| \le ½$. (Hint: Consider $Var(Y_{n+1} + Y_n + \cdots + Y_1)$ and then $Var(Y_{n+1} - Y_n + Y_{n-1} - \cdots \pm Y_1)$. Use the fact that both of these must be nonnegative for *all* n.)

4.14 Suppose that $\{Y_t\}$ is a zero mean, stationary process with $|\rho_1| < 0.5$ and $\rho_k = 0$ for $k > 1$. Show that $\{Y_t\}$ must be representable as an MA(1) process. That is, show that there is a white noise sequence $\{e_t\}$ such that $Y_t = e_t - \theta e_{t-1}$, where ρ_1 is correct and e_t is uncorrelated with Y_{t-k} for $k > 0$. (Hint: Choose θ such that $|\theta| < 1$ and $\rho_1 = -\theta/(1 + \theta^2)$; then let $e_t = \sum_{j=0}^{\infty} \theta^j Y_{t-j}$. If we assume that $\{Y_t\}$ is a normal process, e_t will also be normal, and zero correlation is equivalent to independence.)

4.15 Consider the AR(1) model $Y_t = \phi Y_{t-1} + e_t$. Show that if $|\phi| = 1$ the process cannot be stationary. (Hint: Take variances of both sides.)

4.16 Consider the "nonstationary" AR(1) model $Y_t = 3Y_{t-1} + e_t$.
(a) Show that $Y_t = -\sum_{j=1}^{\infty} (\frac{1}{3})^j e_{t+j}$ satisfies the AR(1) equation.
(b) Show that the process defined in part (a) is stationary.
(c) In what way is this solution unsatisfactory?

4.17 Consider a process that satisfies the AR(1) equation $Y_t = ½Y_{t-1} + e_t$.
(a) Show that $Y_t = 10(½)^t + e_t + ½e_{t-1} + (½)^2 e_{t-2} + \cdots$ is a solution of the AR(1) equation.
(b) Is the solution given in part (a) stationary?

4.18 Consider a process that satisfies the zero-mean, "stationary" AR(1) equation $Y_t = \phi Y_{t-1} + e_t$ with $-1 < \phi < +1$. Let c be any nonzero constant, and define $W_t = Y_t + c\phi^t$.
(a) Show that $E(W_t) = c\phi^t$.
(b) Show that $\{W_t\}$ satisfies the "stationary" AR(1) equation $W_t = \phi W_{t-1} + e_t$.
(c) Is $\{W_t\}$ stationary?

4.19 Consider an MA(6) model with $\theta_1 = 0.5$, $\theta_2 = -0.25$, $\theta_3 = 0.125$, $\theta_4 = -0.0625$, $\theta_5 = 0.03125$, and $\theta_6 = -0.015625$. Find a much simpler model that has nearly the same ψ-weights.

4.20 Consider an MA(7) model with $\theta_1 = 1$, $\theta_2 = -0.5$, $\theta_3 = 0.25$, $\theta_4 = -0.125$, $\theta_5 = 0.0625$, $\theta_6 = -0.03125$, and $\theta_7 = 0.015625$. Find a much simpler model that has nearly the same ψ-weights.

4.21 Consider the model $Y_t = e_{t-1} - e_{t-2} + 0.5e_{t-3}$.
(a) Find the autocovariance function for this process.
(b) Show that this is a certain ARMA(p,q) process in disguise. That is, identify values for p and q and for the θ's and ϕ's such that the ARMA(p,q) process has the same statistical properties as $\{Y_t\}$.

4.22 Show that the statement "The roots of $1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0$ are greater than 1 in absolute value" is equivalent to the statement "The roots of $x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - \cdots - \phi_p = 0$ are less than 1 in absolute value." (Hint: If G is a root of one equation, is $1/G$ a root of the other?)

4.23 Suppose that $\{Y_t\}$ is an AR(1) process with $\rho_1 = \phi$. Define the sequence $\{b_t\}$ as $b_t = Y_t - \phi Y_{t+1}$.
(a) Show that $Cov(b_t, b_{t-k}) = 0$ for all t and k.
(b) Show that $Cov(b_t, Y_{t+k}) = 0$ for all t and $k > 0$.

4.24 Let $\{e_t\}$ be a zero-mean, unit-variance white noise process. Consider a process that begins at time $t = 0$ and is defined recursively as follows. Let $Y_0 = c_1 e_0$ and $Y_1 = c_2 Y_0 + e_1$. Then let $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$ for $t > 1$ as in an AR(2) process.
(a) Show that the process mean is zero.
(b) For particular values of ϕ_1 and ϕ_2 within the stationarity region for an AR(2) model, show how to choose c_1 and c_2 so that both $Var(Y_0) = Var(Y_1)$ and the lag 1 autocorrelation between Y_1 and Y_0 match that of a stationary AR(2) process with parameters ϕ_1 and ϕ_2.
(c) Once the process $\{Y_t\}$ is generated, show how to transform it to a new process that has any desired mean and variance. (This exercise suggests a convenient method for simulating stationary AR(2) processes.)

4.25 Consider an "AR(1)" process satisfying $Y_t = \phi Y_{t-1} + e_t$, where ϕ can be ***any*** number and $\{e_t\}$ is a white noise process such that e_t is independent of the past $\{Y_{t-1}, Y_{t-2}, \ldots\}$. Let Y_0 be a random variable with mean μ_0 and variance σ_0^2.
(a) Show that for $t > 0$ we can write

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots + \phi^{t-1} e_1 + \phi^t Y_0.$$

(b) Show that for $t > 0$ we have $E(Y_t) = \phi^t \mu_0$.
(c) Show that for $t > 0$

$$Var(Y_t) = \begin{cases} \dfrac{1-\phi^{2t}}{1-\phi^2}\sigma_e^2 + \phi^{2t}\sigma_0^2 & \text{for } \phi \neq 1 \\ t\sigma_e^2 + \sigma_0^2 & \text{for } \phi = 1 \end{cases}$$

(d) Suppose now that $\mu_0 = 0$. Argue that, if $\{Y_t\}$ is stationary, we must have $\phi \neq 1$.
(e) Continuing to suppose that $\mu_0 = 0$, show that, if $\{Y_t\}$ is stationary, then $Var(Y_t) = \sigma_e^2/(1-\phi^2)$ and so we must have $|\phi| < 1$.

Appendix B: The Stationarity Region for an AR(2) Process

In the second-order case, the roots of the quadratic characteristic polynomial are easily found to be

$$\frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \tag{4.B.1}$$

For stationarity we require that these roots exceed 1 in absolute value. We now show that this will be true if and only if three conditions are satisfied:

$$\phi_1 + \phi_2 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad \text{and} \qquad |\phi_2| < 1 \tag{4.B.2}$$

Proof: Let the reciprocals of the roots be denoted G_1 and G_2. Then

$$G_1 = \frac{2\phi_2}{-\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}} = \frac{2\phi_2}{-\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}\left[\frac{-\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}\right]$$

$$= \frac{2\phi_2(-\phi_1 + \sqrt{\phi_1^2 + 4\phi_2})}{\phi_1^2 - (\phi_1^2 + 4\phi_2)} = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

Similarly,

$$G_2 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

We now divide the proof into two cases corresponding to real and complex roots. The roots will be real if and only if $\phi_1^2 + 4\phi_2 \geq 0$.

I. **Real Roots**: $|G_i| < 1$ for $i = 1$ and 2 if and only if

$$-1 < \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} < \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} < 1$$

or

$$-2 < \phi_1 - \sqrt{\phi_1^2 + 4\phi_2} < \phi_1 + \sqrt{\phi_1^2 + 4\phi_2} < 2.$$

Consider just the first inequality. Now $-2 < \phi_1 - \sqrt{\phi_1^2 + 4\phi_2}$ if and only if $\sqrt{\phi_1^2 + 4\phi_2} < \phi_1 + 2$ if and only if $\phi_1^2 + 4\phi_2 < \phi_1^2 + 4\phi_1 + 4$ if and only if $\phi_2 < \phi_1 + 1$, or $\phi_2 - \phi_1 < 1$.

The inequality $\phi_1 + \sqrt{\phi_1^2 + 4\phi_2} < 2$ is treated similarly and leads to $\phi_2 + \phi_1 < 1$.

These equations together with $\phi_1^2 + 4\phi_2 \geq 0$ define the stationarity region for the real root case shown in Exhibit 4.17.

II. **Complex Roots**: Now $\phi_1^2 + 4\phi_2 < 0$. Here G_1 and G_2 will be complex conjugates and $|G_1| = |G_2| < 1$ if and only if $|G_1|^2 < 1$. But $|G_1|^2 = [\phi_1^2 + (-\phi_1^2 - 4\phi_2)]/4 = -\phi_2$ so that $\phi_2 > -1$. This together with the inequality $\phi_1^2 + 4\phi_2 < 0$ defines the part of the stationarity region for complex roots shown in Exhibit 4.17 and establishes Equation (4.3.11). This completes the proof.

Appendix C: The Autocorrelation Function for ARMA(*p*,*q*)

Let $\{Y_t\}$ be a stationary, invertible ARMA(p,q) process. Recall that we can always write such a process in general linear process form as

$$Y_t = \sum_{j=0}^{\infty} \psi_j e_{t-j} \tag{4.C.1}$$

where the ψ-weights can be obtained recursively from Equations (4.4.7), on page 79. We then have

$$E(Y_{t+k}e_t) = E\left(\sum_{j=0}^{\infty} \psi_j e_{t+k-j}e_t\right) = \psi_k\sigma_e^2 \text{ for } k \geq 0 \tag{4.C.2}$$

Thus the autocovariance must satisfy

$$\begin{aligned}\gamma_k &= E(Y_{t+k}Y_t) = E\left[\left(\sum_{j=1}^{p} \phi_j Y_{t+k-j} - \sum_{j=0}^{q} \theta_j e_{t+k-j}\right) Y_t\right] \\ &= \sum_{j=1}^{p} \phi_j \gamma_{k-j} - \sigma_e^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}\end{aligned} \tag{4.C.3}$$

where $\theta_0 = -1$ and the last sum is absent if $k > q$. Setting $k = 0, 1, \ldots, p$ and using $\gamma_{-k} = \gamma_k$ leads to $p + 1$ linear equations in $\gamma_0, \gamma_1, \ldots, \gamma_p$.

$$\left.\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p - \sigma_e^2(\theta_0 + \theta_1\psi_1 + \cdots + \theta_q\psi_q) \\ \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 + \cdots + \phi_p\gamma_{p-1} - \sigma_e^2(\theta_1 + \theta_2\psi_1 + \cdots + \theta_q\psi_{q-1}) \\ &\vdots \\ \gamma_p &= \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + \cdots + \phi_p\gamma_0 - \sigma_e^2(\theta_p + \theta_{p+1}\psi_1 + \cdots + \theta_q\psi_{q-p})\end{aligned}\right\} \tag{4.C.4}$$

where $\theta_j = 0$ if $j > q$.

For a given set of parameter values σ_e^2, ϕ's, and θ's (and hence ψ's), we can solve the linear equations to obtain $\gamma_0, \gamma_1, \ldots, \gamma_p$. The values of γ_k for $k > p$ can then be evaluated from the recursion in Equations (4.4.8), on page 79. Finally, ρ_k is obtained from $\rho_k = \gamma_k/\gamma_0$.

CHAPTER 5

MODELS FOR NONSTATIONARY TIME SERIES

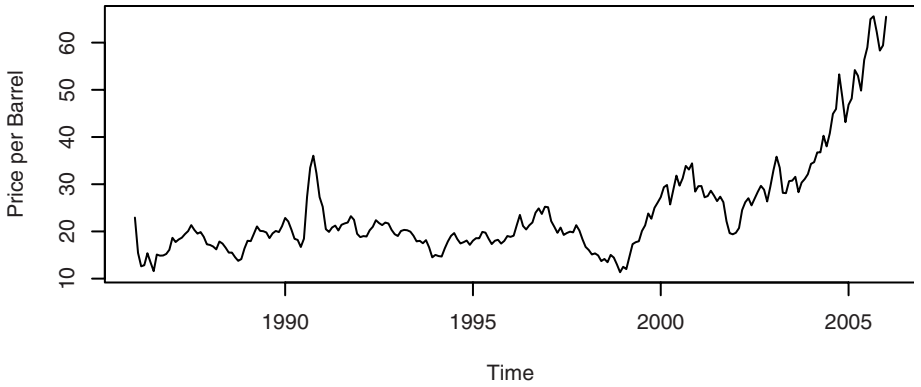
Any time series without a constant mean over time is nonstationary. Models of the form

$$Y_t = \mu_t + X_t$$

where μ_t is a nonconstant mean function and X_t is a zero-mean, stationary series, were considered in Chapter 3. As stated there, such models are reasonable only if there are good reasons for believing that the deterministic trend is appropriate "forever." That is, just because a segment of the series looks like it is increasing (or decreasing) approximately linearly, do we believe that the linearity is intrinsic to the process and will persist in the future? Frequently in applications, particularly in business and economics, we cannot legitimately assume a deterministic trend. Recall the random walk displayed in Exhibit 2.1, on page 14. The time series appears to have a strong upward trend that might be linear in time. However, also recall that the random walk process has a constant, zero mean and contains no deterministic trend at all.

As an example consider the monthly price of a barrel of crude oil from January 1986 through January 2006. Exhibit 5.1 displays the time series plot. The series displays considerable variation, especially since 2001, and a stationary model does not seem to be reasonable. We will discover in Chapters 6, 7, and 8 that no deterministic trend model works well for this series but one of the nonstationary models that have been described as containing stochastic trends does seem reasonable. This chapter discusses such models. Fortunately, as we shall see, many stochastic trends can be modeled with relatively few parameters.

Exhibit 5.1 Monthly Price of Oil: January 1986–January 2006




```
> win.graph(width=4.875,height=3,pointsize=8)
> data(oil.price)
> plot(oil.price, ylab='Price per Barrel',type='l')
```

5.1 Stationarity Through Differencing

Consider again the AR(1) model

$$Y_t = \phi Y_{t-1} + e_t \tag{5.1.1}$$

We have seen that assuming e_t is a true "innovation" (that is, e_t is uncorrelated with $Y_{t-1}, Y_{t-2},\ldots$), we must have $|\phi| < 1$. What can we say about solutions to Equation (5.1.1) if $|\phi| \geq 1$? Consider in particular the equation

$$Y_t = 3Y_{t-1} + e_t \tag{5.1.2}$$

Iterating into the past as we have done before yields

$$Y_t = e_t + 3e_{t-1} + 3^2 e_{t-2} + \cdots + 3^{t-1} e_1 + 3^t Y_0 \tag{5.1.3}$$

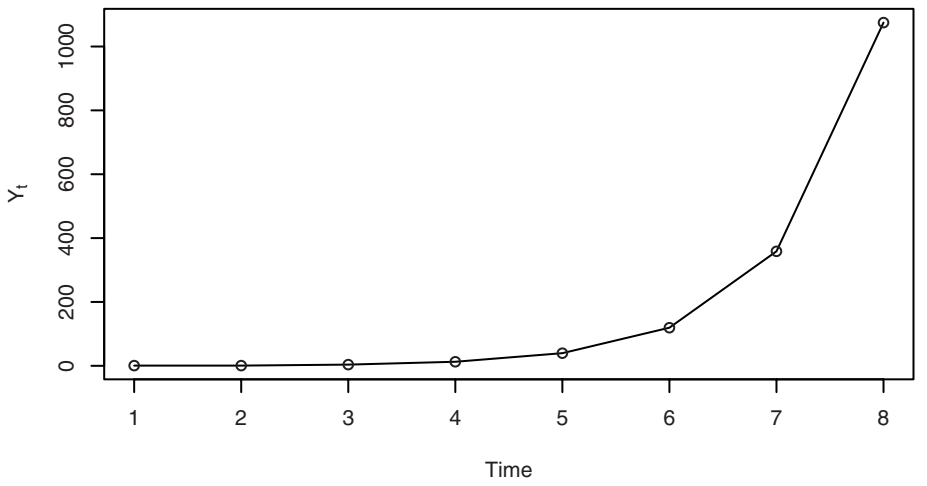
We see that the influence of distant past values of Y_t and e_t does not die out—indeed, the weights applied to Y_0 and e_1 grow exponentially large. In Exhibit 5.2, we show the values for a very short simulation of such a series. Here the white noise sequence was generated as standard normal variables and we used $Y_0 = 0$ as an initial condition.

Exhibit 5.2 Simulation of the Explosive "AR(1) Model" $Y_t = 3Y_{t-1} + e_t$

t	1	2	3	4	5	6	7	8
e_t	0.63	−1.25	1.80	1.51	1.56	0.62	0.64	−0.98
Y_t	0.63	0.64	3.72	12.67	39.57	119.33	358.63	1074.91

Exhibit 5.3 shows the time series plot of this explosive AR(1) simulation.

Exhibit 5.3 An Explosive "AR(1)" Series




```
> data(explode.s)
> plot(explode.s,ylab=expression(Y[t]),type='o')
```

The explosive behavior of such a model is also reflected in the model's variance and covariance functions. These are easily found to be

$$Var(Y_t) = \frac{1}{8}(9^t - 1)\sigma_e^2 \tag{5.1.4}$$

and

$$Cov(Y_t, Y_{t-k}) = \frac{3^k}{8}(9^{t-k} - 1)\sigma_e^2 \tag{5.1.5}$$

respectively. Notice that we have

$$Corr(Y_t, Y_{t-k}) = 3^k\left(\frac{9^{t-k} - 1}{9^t - 1}\right) \approx 1 \qquad \text{for large } t \text{ and moderate } k$$

The same general exponential growth or explosive behavior will occur for any ϕ such that $|\phi| > 1$. A more reasonable type of nonstationarity obtains when $\phi = 1$. If $\phi = 1$, the AR(1) model equation is

$$Y_t = Y_t + e_t \tag{5.1.6}$$

This is the relationship satisfied by the random walk process of Chapter 2 (Equation (2.2.9) on page 12). Alternatively, we can rewrite this as

$$\nabla Y_t = e_t \tag{5.1.7}$$

where $\nabla Y_t = Y_t - Y_{t-1}$ is the **first difference** of Y_t. The random walk then is easily extended to a more general model whose first difference is some stationary process—not just white noise.

Several somewhat different sets of assumptions can lead to models whose first difference is a stationary process. Suppose

$$Y_t = M_t + X_t \tag{5.1.8}$$

where M_t is a series that is changing only slowly over time. Here M_t could be either deterministic or stochastic. If we assume that M_t is approximately constant over every two consecutive time points, we might estimate (predict) M_t at t by choosing β_0 so that

$$\sum_{j=0}^{1} (Y_{t-j} - \beta_{0,t})^2$$

is minimized. This clearly leads to

$$\hat{M}_t = \frac{1}{2}(Y_t + Y_{t-1})$$

and the "detrended" series at time t is then

$$Y_t - \hat{M}_t = Y_t - \frac{1}{2}(Y_t + Y_{t-1}) = \frac{1}{2}(Y_t - Y_{t-1}) = \frac{1}{2}\nabla Y_t$$

This is a constant multiple of the first difference, ∇Y_t.[†]

A second set of assumptions might be that M_t in Equation (5.1.8) is stochastic and changes slowly over time governed by a random walk model. Suppose, for example, that

$$Y_t = M_t + e_t \qquad \text{with} \qquad M_t = M_{t-1} + \varepsilon_t \tag{5.1.9}$$

where $\{e_t\}$ and $\{\varepsilon_t\}$ are independent white noise series. Then

$$\begin{aligned} \nabla Y_t &= \nabla M_t + \nabla e_t \\ &= \varepsilon_t + e_t - e_{t-1} \end{aligned}$$

which would have the autocorrelation function of an MA(1) series with

$$\rho_1 = -\{1/[2 + (\sigma_\varepsilon^2/\sigma_e^2)]\} \tag{5.1.10}$$

In either of these situations, we are led to the study of ∇Y_t as a stationary process.

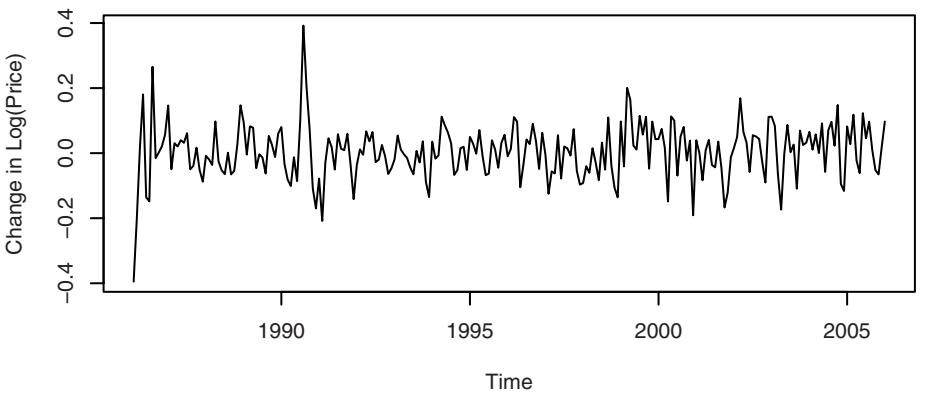
Returning to the oil price time series, Exhibit 5.4 displays the time series plot of the differences of logarithms of that series.[‡] The differenced series looks much more stationary when compared with the original time series shown in Exhibit 5.1, on page 88.

[†] A more complete labeling of this difference would be that it is a **first difference at lag 1**.

[‡] In Section 5.4 on page 98 we will see why logarithms are often a convenient transformation.

(We will also see later that there are outliers in this series that need to be considered to produce an adequate model.)

Exhibit 5.4 The Difference Series of the Logs of the Oil Price Time




```
> plot(diff(log(oil.price)),ylab='Change in Log(Price)',type='l')
```

We can also make assumptions that lead to stationary second-difference models. Again we assume that Equation (5.1.8) on page 90, holds, but now assume that M_t is linear in time over three consecutive time points. We can now estimate (predict) M_t at the middle time point t by choosing $\beta_{0,t}$ and $\beta_{1,t}$ to minimize

$$\sum_{j=-1}^{1} (Y_{t-j} - (\beta_{0,t} + j\beta_{1,t}))^2$$

The solution yields

$$\hat{M}_t = \frac{1}{3}(Y_{t+1} + Y_t + Y_{t-1})$$

and thus the detrended series is

$$\begin{aligned} Y_t - \hat{M}_t &= Y_t - \left(\frac{Y_{t+1} + Y_t + Y_{t-1}}{3}\right) \\ &= \left(-\frac{1}{3}\right)(Y_{t+1} - 2Y_t + Y_{t-1}) \\ &= \left(-\frac{1}{3}\right)\nabla(\nabla Y_{t+1}) \\ &= \left(-\frac{1}{3}\right)\nabla^2(Y_{t+1}) \end{aligned}$$

a constant multiple of the centered **second difference** of Y_t. Notice that we have differenced twice, but both differences are at lag 1.

Alternatively, we might assume that

$$Y_t = M_t + e_t, \quad \text{where} \quad M_t = M_{t-1} + W_t \quad \text{and} \quad W_t = W_{t-1} + \varepsilon_t \tag{5.1.11}$$

with $\{e_t\}$ and $\{\varepsilon_t\}$ independent white noise time series. Here the stochastic trend M_t is such that its "rate of change," ∇M_t, is changing slowly over time. Then

$$\nabla Y_t = \nabla M_t + \nabla e_t = W_t + \nabla e_t$$

and

$$\begin{aligned}\nabla^2 Y_t &= \nabla W_t + \nabla^2 e_t \\ &= \varepsilon_t + (e_t - e_{t-1}) - (e_{t-1} - e_{t-2}) \\ &= \varepsilon_t + e_t - 2e_{t-1} + e_{t-2}\end{aligned}$$

which has the autocorrelation function of an MA(2) process. The important point is that the second difference of the nonstationary process $\{Y_t\}$ is stationary. This leads us to the general definition of the important integrated autoregressive moving average time series models.

5.2 ARIMA Models

A time series $\{Y_t\}$ is said to follow an **integrated autoregressive moving average** model if the dth difference $W_t = \nabla^d Y_t$ is a stationary ARMA process. If $\{W_t\}$ follows an ARMA(p,q) model, we say that $\{Y_t\}$ is an ARIMA(p,d,q) process. Fortunately, for practical purposes, we can usually take $d = 1$ or at most 2.

Consider then an ARIMA$(p,1,q)$ process. With $W_t = Y_t - Y_{t-1}$, we have

$$\begin{aligned}W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \\ - \cdots - \theta_q e_{t-q}\end{aligned} \tag{5.2.1}$$

or, in terms of the observed series,

$$\begin{aligned}Y_t - Y_{t-1} = \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \cdots + \phi_p(Y_{t-p} - Y_{t-p-1}) \\ + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}\end{aligned}$$

which we may rewrite as

$$\begin{aligned}Y_t = (1+\phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + (\phi_3 - \phi_2)Y_{t-3} + \cdots \\ + (\phi_p - \phi_{p-1})Y_{t-p} - \phi_p Y_{t-p-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}\end{aligned} \tag{5.2.2}$$

We call this the **difference equation form** of the model. Notice that it appears to be an ARMA$(p+1,q)$ process. However, the characteristic polynomial satisfies

$$\begin{aligned}1 - (1+\phi_1)x - (\phi_2 - \phi_1)x^2 - (\phi_3 - \phi_2)x^3 - \cdots - (\phi_p - \phi_{p-1})x^p + \phi_p x^{p+1} \\ = (1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p)(1 - x)\end{aligned}$$

which can be easily checked. This factorization clearly shows the root at $x = 1$, which implies nonstationarity. The remaining roots, however, are the roots of the characteristic polynomial of the *stationary* process ∇Y_t.

Explicit representations of the observed series in terms of either W_t or the white noise series underlying W_t are more difficult than in the stationary case. Since nonstationary processes are not in statistical equilibrium, we cannot assume that they go infinitely into the past or that they start at $t = -\infty$. However, we can and shall assume that they start at some time point $t = -m$, say, where $-m$ is earlier than time $t = 1$, at which point we first observed the series. For convenience, we take $Y_t = 0$ for $t < -m$. The difference equation $Y_t - Y_{t-1} = W_t$ can be solved by summing both sides from $t = -m$ to $t = t$ to get the representation

$$Y_t = \sum_{j=-m}^{t} W_j \tag{5.2.3}$$

for the ARIMA(p,1,q) process.

The ARIMA(p,2,q) process can be dealt with similarly by summing twice to get the representations

$$\begin{aligned} Y_t &= \sum_{j=-m}^{t} \sum_{i=-m}^{j} W_i \\ &= \sum_{j=0}^{t+m} (j+1) W_{t-j} \end{aligned} \tag{5.2.4}$$

These representations have limited use but can be used to investigate the covariance properties of ARIMA models and also to express Y_t in terms of the white noise series $\{e_t\}$. We defer the calculations until we evaluate specific cases.

If the process contains no autoregressive terms, we call it an integrated moving average and abbreviate the name to IMA(d,q). If no moving average terms are present, we denote the model as ARI(p,d). We first consider in detail the important IMA(1,1) model.

The IMA(1,1) Model

The simple IMA(1,1) model satisfactorily represents numerous time series, especially those arising in economics and business. In difference equation form, the model is

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1} \tag{5.2.5}$$

To write Y_t explicitly as a function of present and past noise values, we use Equation (5.2.3) and the fact that $W_t = e_t - \theta e_{t-1}$ in this case. After a little rearrangement, we can write

$$Y_t = e_t + (1-\theta)e_{t-1} + (1-\theta)e_{t-2} + \cdots + (1-\theta)e_{-m} - \theta e_{-m-1} \tag{5.2.6}$$

Notice that in contrast to our stationary ARMA models, the weights on the white noise terms *do not die out* as we go into the past. Since we are assuming that $-m < 1$ and $0 < t$, we may usefully think of Y_t as mostly an equally weighted accumulation of a large number of white noise values.

From Equation (5.2.6), we can easily derive variances and correlations. We have

$$Var(Y_t) = [1 + \theta^2 + (1-\theta)^2(t+m)]\sigma_e^2 \tag{5.2.7}$$

and

$$\begin{aligned} Corr(Y_t, Y_{t-k}) &= \frac{1-\theta+\theta^2+(1-\theta)^2(t+m-k)}{[Var(Y_t)Var(Y_{t-k})]^{1/2}} \\ &\approx \sqrt{\frac{t+m-k}{t+m}} \\ &\approx 1 \qquad \text{for large } m \text{ and moderate } k \end{aligned} \tag{5.2.8}$$

We see that as t increases, $Var(Y_t)$ increases and could be quite large. Also, the correlation between Y_t and Y_{t-k} will be strongly positive for many lags $k = 1, 2, \ldots$.

The IMA(2,2) Model

The assumptions of Equation (5.1.11) led to an IMA(2,2) model. In difference equation form, we have

$$\nabla^2 Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

or

$$Y_t = 2Y_{t-1} - Y_{t-2} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \tag{5.2.9}$$

The representation of Equation (5.2.4) may be used to express Y_t in terms of $e_t, e_{t-1}, \ldots$. After some tedious algebra, we find that

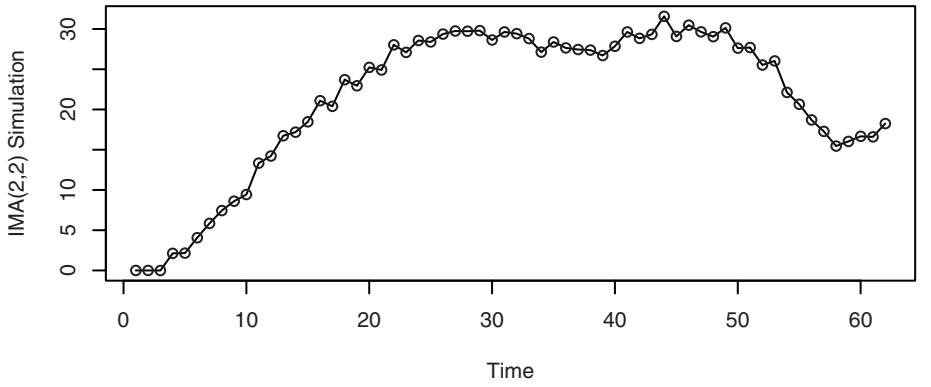
$$\begin{aligned} Y_t = e_t + \sum_{j=1}^{t+m} \psi_j e_{t-j} - [(t+m+1)\theta_1 + (t+m)\theta_2]e_{-m-1} \\ -(t+m+1)\theta_2 e_{-m-2} \end{aligned} \tag{5.2.10}$$

where $\psi_j = 1 + \theta_2 + (1 - \theta_1 - \theta_2)j$ for $j = 1, 2, 3, \ldots, t + m$. Once more we see that the ψ-weights do not die out but form a linear function of j.

Again, variances and correlations for Y_t can be obtained from the representation given in Equation (5.2.10), but the calculations are tedious. We shall simply note that the variance of Y_t increases rapidly with t and again $Corr(Y_t, Y_{t-k})$ is nearly 1 for all moderate k.

The results of a simulation of an IMA(2,2) process are displayed in Exhibit 5.5. Notice the smooth change in the process values (and the unimportance of the zero-mean function). The increasing variance and the strong, positive neighboring correlations dominate the appearance of the time series plot.

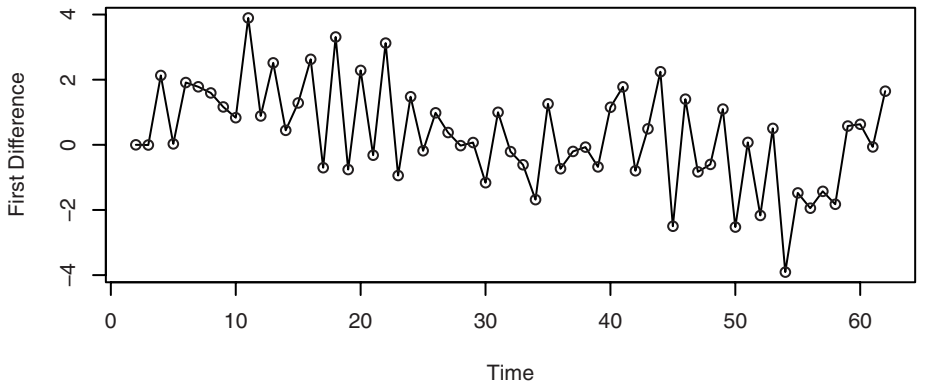
Exhibit 5.5 Simulation of an IMA(2,2) Series with $\theta_1 = 1$ and $\theta_2 = -0.6$




```
> data(ima22.s)
> plot(ima22.s,ylab='IMA(2,2) Simulation',type='o')
```

Exhibit 5.6 shows the time series plot of the first difference of the simulated series. This series is also nonstationary, as it is governed by an IMA(1,2) model.

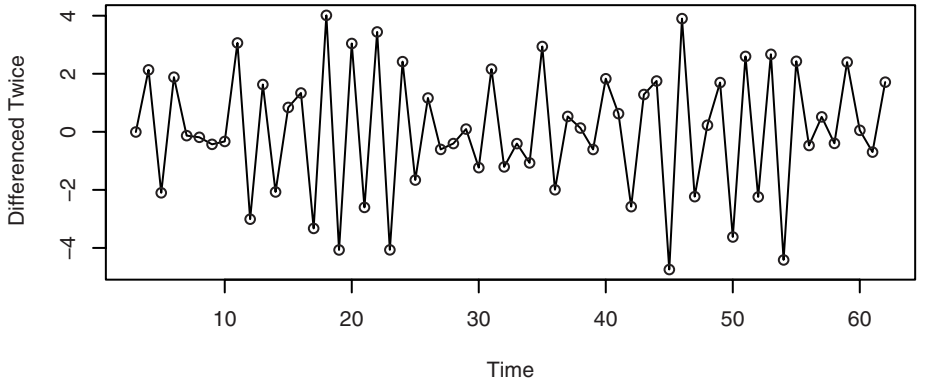
Exhibit 5.6 First Difference of the Simulated IMA(2,2) Series




```
> plot(diff(ima22.s),ylab='First Difference',type='o')
```

Finally, the second differences of the simulated IMA(2,2) series values are plotted in Exhibit 5.7. These values arise from a stationary MA(2) model with $\theta_1 = 1$ and $\theta_2 = -0.6$. From Equation (4.2.3) on page 63, the theoretical autocorrelations for this model are $\rho_1 = -0.678$ and $\rho_2 = 0.254$. These correlation values seem to be reflected in the appearance of the time series plot.

Exhibit 5.7 Second Difference of the Simulated IMA(2,2) Series




```
> plot(diff(ima22.s,difference=2),ylab='Differenced
    Twice',type='o')
```

The ARI(1,1) Model

The ARI(1,1) process will satisfy

$$Y_t - Y_{t-1} = \phi(Y_{t-1} - Y_{t-2}) + e_t \tag{5.2.11}$$

or

$$Y_t = (1+\phi)Y_{t-1} - \phi Y_{t-2} + e_t \tag{5.2.12}$$

where $|\phi| < 1$.[†]

To find the ψ-weights in this case, we shall use a technique that will generalize to arbitrary ARIMA models. It can be shown that the ψ-weights can be obtained by equating like powers of x in the identity:

$$\begin{aligned}(1-\phi_1 x - \phi_2 x^2 - \cdots - \phi^p x^p)(1-x)^d(1+\psi_1 x + \psi_2 x^2 + \psi_3 x^3 + \cdots) \\ = (1-\theta_1 x - \theta_2 x^2 - \theta_3 x^3 - \cdots - \theta_q x^q)\end{aligned} \tag{5.2.13}$$

In our case, this relationship reduces to

$$(1-\phi x)(1-x)(1+\psi_1 x + \psi_2 x^2 + \psi_3 x^3 + \cdots) = 1$$

or

$$[1-(1+\phi)x + \phi x^2](1+\psi_1 x + \psi_2 x^2 + \psi_3 x^3 + \cdots) = 1$$

Equating like powers of x on both sides, we obtain

[†] Notice that this looks like a special AR(2) model. However, one of the roots of the corresponding AR(2) characteristic polynomial is 1, and this is not allowed in stationary AR(2) models.

$$-(1+\phi)+\psi_1 = 0$$
$$\phi-(1+\phi)\psi_1+\psi_2 = 0$$

and, in general,

$$\psi_k = (1+\phi)\psi_{k-1}-\phi\psi_{k-2} \qquad \text{for } k \geq 2 \tag{5.2.14}$$

with $\psi_o = 1$ and $\psi_1 = 1+\phi$. This recursion with starting values allows us to compute as many ψ-weights as necessary. It can also be shown that in this case an explicit solution to the recursion is given as

$$\psi_k = \frac{1-\phi^{k+1}}{1-\phi} \quad \text{for } k \geq 1 \tag{5.2.15}$$

(It is easy, for example, to show that this expression satisfies Equation (5.2.14).

5.3 Constant Terms in ARIMA Models

For an ARIMA(p,d,q) model, $\nabla^d Y_t = W_t$ is a stationary ARMA(p,q) process. Our standard assumption is that stationary models have a zero mean; that is, we are actually working with deviations from the constant mean. A nonzero constant mean, μ, in a stationary ARMA model $\{W_t\}$ can be accommodated in either of two ways. We can assume that

$$W_t-\mu = \phi_1(W_{t-1}-\mu)+\phi_2(W_{t-2}-\mu)+\cdots+\phi_p(W_{t-p}-\mu) + e_t-\theta_1 e_{t-1}-\theta_2 e_{t-2}-\cdots-\theta_q e_{t-q}$$

Alternatively, we can introduce a constant term θ_0 into the model as follows:

$$W_t = \theta_0+\phi_1 W_{t-1}+\phi_2 W_{t-2}+\cdots+\phi_p W_{t-p} + e_t-\theta_1 e_{t-1}-\theta_2 e_{t-2}-\cdots-\theta_q e_{t-q}$$

Taking expected values on both sides of the latter expression, we find that

$$\mu = \theta_0+(\phi_1+\phi_2+\cdots+\phi_p)\mu$$

so that

$$\mu = \frac{\theta_0}{1-\phi_1-\phi_2-\cdots-\phi_p} \tag{5.3.16}$$

or, conversely, that

$$\theta_0 = \mu(1-\phi_1-\phi_2-\cdots-\phi_p) \tag{5.3.17}$$

Since the alternative representations are equivalent, we shall use whichever parameterization is convenient.

What will be the effect of a nonzero mean for W_t on the undifferenced series Y_t? Consider the IMA(1,1) case with a constant term. We have

$$Y_t = Y_{t-1} + \theta_0 + e_t - \theta e_{t-1}$$

or

$$W_t = \theta_0 + e_t - \theta e_{t-1}$$

Either by substituting into Equation (5.2.3) on page 93 or by iterating into the past, we find that

$$Y_t = e_t + (1-\theta)e_{t-1} + (1-\theta)e_{t-2} + \cdots + (1-\theta)e_{-m} - \theta e_{-m-1} + (t+m+1)\theta_0 \tag{5.3.18}$$

Comparing this with Equation (5.2.6), we see that we have an added *linear deterministic time trend* $(t+m+1)\theta_0$ with slope θ_0.

An equivalent representation of the process would then be

$$Y_t = Y_t' + \beta_0 + \beta_1 t$$

where Y_t' is an IMA(1,1) series with $E(\nabla Y_t') = 0$ and $E(\nabla Y_t) = \beta_1$.

For a general ARIMA(p,d,q) model where $E(\nabla^d Y_t) \neq 0$, it can be argued that $Y_t = Y_t' + \mu_t$, where μ_t is a deterministic polynomial of degree d and Y_t' is ARIMA(p,d,q) with $EY_t' = 0$. With $d = 2$ and $\theta_0 \neq 0$, a quadratic trend would be implied.

5.4 Other Transformations

We have seen how differencing can be a useful transformation for achieving stationarity. However, the logarithm transformation is also a useful method in certain circumstances. We frequently encounter series where increased dispersion seems to be associated with higher levels of the series—the higher the level of the series, the more variation there is around that level and conversely.

Specifically, suppose that $Y_t > 0$ for all t and that

$$E(Y_t) = \mu_t \qquad \text{and} \qquad \sqrt{Var(Y_t)} = \mu_t \sigma \tag{5.4.1}$$

Then

$$E[\log(Y_t)] \approx \log(\mu_t) \qquad \text{and} \qquad Var(\log(Y_t)) \approx \sigma^2 \tag{5.4.2}$$

These results follow from taking expected values and variances of both sides of the (Taylor) expansion

$$\log(Y_t) \approx \log(\mu_t) + \frac{Y_t - \mu_t}{\mu_t}$$

In words, if the standard deviation of the series is proportional to the level of the series, then transforming to logarithms will produce a series with approximately constant variance over time. Also, if the level of the series is changing roughly exponentially, the

log-transformed series will exhibit a linear time trend. Thus, we might then want to take first differences. An alternative set of assumptions leading to differences of logged data follows.

Percentage Changes and Logarithms

Suppose Y_t tends to have relatively stable percentage changes from one time period to the next. Specifically, assume that

$$Y_t = (1 + X_t)Y_{t-1}$$

where $100X_t$ is the percentage change (possibly negative) from Y_{t-1} to Y_t. Then

$$\begin{aligned} \log(Y_t) - \log(Y_{t-1}) &= \log\left(\frac{Y_t}{Y_{t-1}}\right) \\ &= \log(1 + X_t) \end{aligned}$$

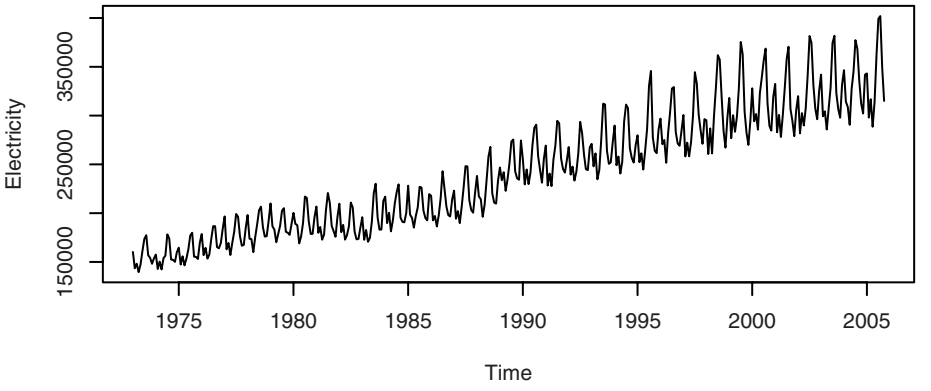
If X_t is restricted to, say, $|X_t| < 0.2$ (that is, the percentage changes are at most ±20%), then, to a good approximation, $\log(1+X_t) \approx X_t$. Consequently,

$$\nabla[\log(Y_t)] \approx X_t \tag{5.4.3}$$

will be relatively stable and perhaps well-modeled by a stationary process. Notice that we take logs first and then compute first differences—the order does matter. In financial literature, the differences of the (natural) logarithms are usually called **returns**.

As an example, consider the time series shown in Exhibit 5.8. This series gives the total monthly electricity generated in the United States in millions of kilowatt-hours. The higher values display considerably more variation than the lower values.

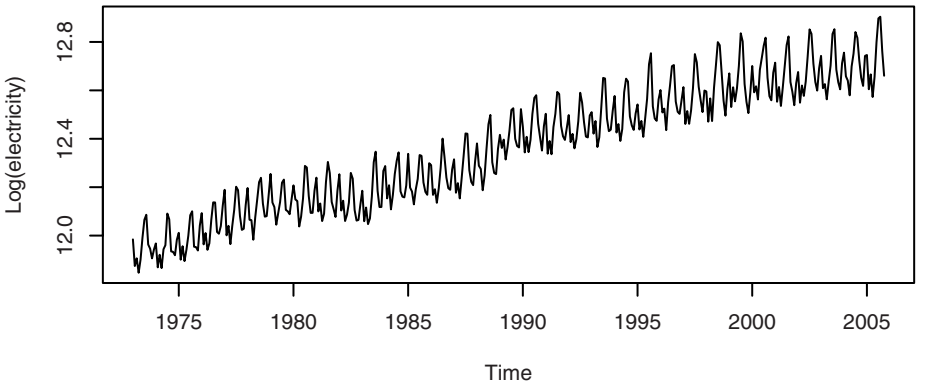
Exhibit 5.8 U.S. Electricity Generated by Month




```
> data(electricity); plot(electricity)
```

Exhibit 5.9 displays the time series plot of the logarithms of the electricity values. Notice how the amount of variation around the upward trend is now much more uniform across high and low values of the series.

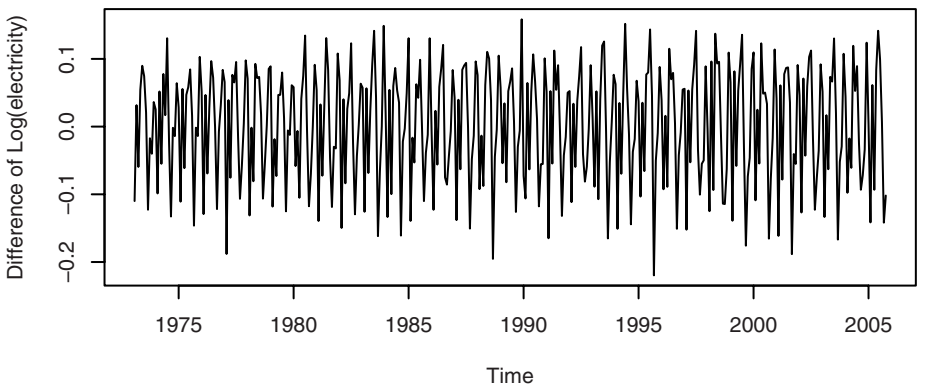
Exhibit 5.9 Time Series Plot of Logarithms of Electricity Values




```
> plot(log(electricity),ylab='Log(electricity)')
```

The differences of the logarithms of the electricity values are displayed in Exhibit 5.10. On the basis of this plot, we might well consider a stationary model as appropriate.

Exhibit 5.10 Difference of Logarithms for Electricity Time Series




```
> plot(diff(log(electricity)),
   ylab='Difference of Log(electricity)')
```

Power Transformations

A flexible family of transformations, the **power transformations**, was introduced by Box and Cox (1964). For a given value of the parameter λ, the transformation is defined by

$$g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log x & \text{for } \lambda = 0 \end{cases} \tag{5.4.4}$$

The term x^{λ} is the important part of the first expression, but subtracting 1 and dividing by λ makes $g(x)$ change smoothly as λ approaches zero. In fact, a calculus argument[†] shows that as $\lambda \to 0$, $(x^{\lambda} - 1)/\lambda \to \log(x)$. Notice that $\lambda = \frac{1}{2}$ produces a square root transformation useful with Poisson-like data, and $\lambda = -1$ corresponds to a reciprocal transformation.

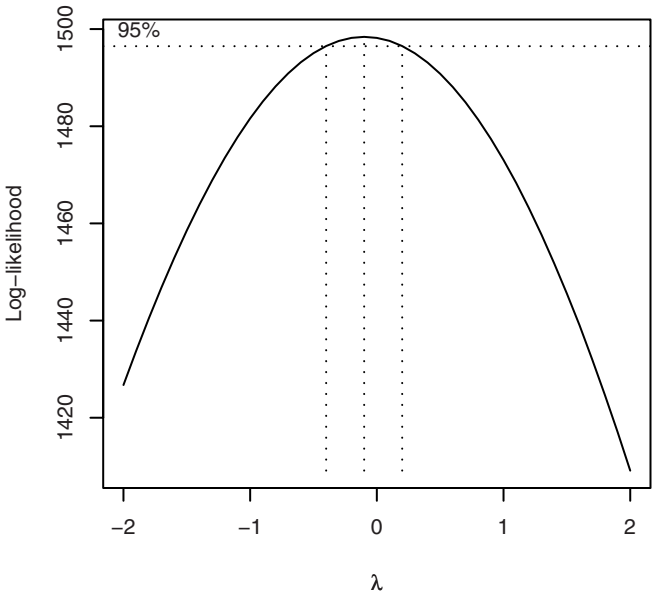
The power transformation applies only to positive data values. If some of the values are negative or zero, a positive constant may be added to all of the values to make them all positive before doing the power transformation. The shift is often determined subjectively. For example, for nonnegative catch data in biology, the occurrence of zeros is often dealt with by adding a constant equal to the smallest positive data value to all of the data values. An alternative approach consists of using transformations applicable to any data—positive or not. A drawback of this alternative approach is that interpretations of such transformations are often less straightforward than the interpretations of the power transformations. See Yeo and Johnson (2000) and the references contained therein.

We can consider λ as an additional parameter in the model to be estimated from the observed data. However, precise estimation of λ is usually not warranted. Evaluation of a range of transformations based on a grid of λ values, say ± 1, $\pm 1/2$, $\pm 1/3$, $\pm 1/4$, and 0, will usually suffice and may have some intuitive meaning.

Software allows us to consider a range of lambda values and calculate a log-likelihood value for each lambda value based on a normal likelihood function. A plot of these values is shown in Exhibit 5.11 for the electricity data. The 95% confidence interval for λ contains the value of $\lambda = 0$ quite near its center and strongly suggests a logarithmic transformation ($\lambda = 0$) for these data.

[†] Exercise (5.17) asks you to verify this.

Exhibit 5.11 Log-likelihood versus Lambda




```
> BoxCox.ar(electricity)
```

5.5 Summary

This chapter introduced the concept of differencing to induce stationarity on certain nonstationary processes. This led to the important integrated autoregressive moving average models (ARIMA). The properties of these models were then thoroughly explored. Other transformations, namely percentage changes and logarithms, were then considered. More generally, power transformations or Box-Cox transformations were introduced as useful transformations to stationarity and often normality.

EXERCISES

5.1 Identify the following as specific ARIMA models. That is, what are p, d, and q and what are the values of the parameters (the ϕ's and θ's)?

(a) $Y_t = Y_{t-1} - 0.25Y_{t-2} + e_t - 0.1e_{t-1}$.

(b) $Y_t = 2Y_{t-1} - Y_{t-2} + e_t$.

(c) $Y_t = 0.5Y_{t-1} - 0.5Y_{t-2} + e_t - 0.5e_{t-1} + 0.25e_{t-2}$.

5.2 For each of the ARIMA models below, give the values for $E(\nabla Y_t)$ and $Var(\nabla Y_t)$.

(a) $Y_t = 3 + Y_{t-1} + e_t - 0.75e_{t-1}$.

(b) $Y_t = 10 + 1.25Y_{t-1} - 0.25Y_{t-2} + e_t - 0.1e_{t-1}$.

(c) $Y_t = 5 + 2Y_{t-1} - 1.7Y_{t-2} + 0.7Y_{t-3} + e_t - 0.5e_{t-1} + 0.25e_{t-2}$.

5.3 Suppose that $\{Y_t\}$ is generated according to $Y_t = e_t + ce_{t-1} + ce_{t-2} + ce_{t-3} + \cdots + ce_0$ for $t > 0$.

(a) Find the mean and covariance functions for $\{Y_t\}$. Is $\{Y_t\}$ stationary?

(b) Find the mean and covariance functions for $\{\nabla Y_t\}$. Is $\{\nabla Y_t\}$ stationary?

(c) Identify $\{Y_t\}$ as a specific ARIMA process.

5.4 Suppose that $Y_t = A + Bt + X_t$, where $\{X_t\}$ is a random walk. First suppose that A and B are constants.

(a) Is $\{Y_t\}$ stationary?

(b) Is $\{\nabla Y_t\}$ stationary?

Now suppose that A and B are random variables that are independent of the random walk $\{X_t\}$.

(c) Is $\{Y_t\}$ stationary?

(d) Is $\{\nabla Y_t\}$ stationary?

5.5 Using the simulated white noise values in Exhibit 5.2, on page 88, verify the values shown for the explosive process Y_t.

5.6 Consider a stationary process $\{Y_t\}$. Show that if $\rho_1 < ½$, ∇Y_t has a larger variance than does Y_t.

5.7 Consider two models:

A: $Y_t = 0.9Y_{t-1} + 0.09Y_{t-2} + e_t$.

B: $Y_t = Y_{t-1} + e_t - 0.1e_{t-1}$.

(a) Identify each as a specific ARIMA model. That is, what are p, d, and q and what are the values of the parameters, ϕ's and θ's?

(b) In what ways are the two models different?

(c) In what ways are the two models similar? (Compare ψ-weights and π-weights.)

5.8 Consider a nonstationary "AR(1)" process defined as a solution to Equation (5.1.2) on page 88, with $|\phi| > 1$.
(a) Derive an equation similar to Equation (5.1.3) on page 88, for this more general case. Use $Y_0 = 0$ as an initial condition.
(b) Derive an equation similar to Equation (5.1.4) on page 89, for this more general case.
(c) Derive an equation similar to Equation (5.1.5) on page 89, for this more general case.
(d) Is it true that for any $|\phi| > 1$, $Corr(Y_t, Y_{t-k}) \approx 1$ for large t and moderate k?

5.9 Verify Equation (5.1.10) on page 90.

5.10 Nonstationary ARIMA series can be simulated by first simulating the corresponding stationary ARMA series and then "integrating" it (really partially summing it). Use statistical software to simulate a variety of IMA(1,1) and IMA(2,2) series with a variety of parameter values. Note any stochastic "trends" in the simulated series.

5.11 The data file winnebago contains monthly unit sales of recreational vehicles (RVs) from Winnebago, Inc., from November 1966 through February 1972.
(a) Display and interpret the time series plot for these data.
(b) Now take natural logarithms of the monthly sales figures and display the time series plot of the transformed values. Describe the effect of the logarithms on the behavior of the series.
(c) Calculate the fractional relative changes, $(Y_t - Y_{t-1})/Y_{t-1}$, and compare them with the differences of (natural) logarithms, $\nabla\log(Y_t) = \log(Y_t) - \log(Y_{t-1})$. How do they compare for smaller values and for larger values?

5.12 The data file SP contains quarterly Standard & Poor's Composite Index stock price values from the first quarter of 1936 through the fourth quarter of 1977.
(a) Display and interpret the time series plot for these data.
(b) Now take natural logarithms of the quarterly values and display and the time series plot of the transformed values. Describe the effect of the logarithms on the behavior of the series.
(c) Calculate the (fractional) relative changes, $(Y_t - Y_{t-1})/Y_{t-1}$, and compare them to the differences of (natural) logarithms, $\nabla\log(Y_t)$. How do they compare for smaller values and for larger values?

5.13 The data file airpass contains monthly U.S. air passenger miles flown from January 1949 through December 1960. This is a classic time series analyzed in Box and Jenkins (1976).
(a) Display and interpret the time series plot for these data.
(b) Now take natural logarithms of the monthly values and display and the time series plot of the transformed values. Describe the effect of the logarithms on the behavior of the series.
(c) Calculate the (fractional) relative changes, $(Y_t - Y_{t-1})/Y_{t-1}$, and compare them to the differences of (natural) logarithms, $\nabla\log(Y_t)$. How do they compare for smaller values and for larger values?

5.14 Consider the annual rainfall data for Los Angeles shown in Exhibit 1.1, on page 2. The quantile-quantile normal plot of these data, shown in Exhibit 3.17, on page 50, convinced us that the data were not normal. The data are in the file larain.

(a) Use software to produce a plot similar to Exhibit 5.11, on page 102, and determine the "best" value of λ for a power transformation of the data.

(b) Display a quantile-quantile plot of the transformed data. Are they more normal?

(c) Produce a time series plot of the transformed values.

(d) Use the transformed values to display a plot of Y_t versus Y_{t-1} as in Exhibit 1.2, on page 2. Should we expect the transformation to change the dependence or lack of dependence in the series?

5.15 Quarterly earnings per share for the Johnson & Johnson Company are given in the data file named JJ. The data cover the years from 1960 through 1980.

(a) Display a time series plot of the data. Interpret the interesting features in the plot.

(b) Use software to produce a plot similar to Exhibit 5.11, on page 102, and determine the "best" value of λ for a power transformation of these data.

(c) Display a time series plot of the transformed values. Does this plot suggest that a stationary model might be appropriate?

(d) Display a time series plot of the differences of the transformed values. Does this plot suggest that a stationary model might be appropriate for the differences?

5.16 The file named gold contains the daily price of gold (in dollars per troy ounce) for the 252 trading days of year 2005.

(a) Display the time series plot of these data. Interpret the plot.

(b) Display the time series plot of the differences of the logarithms of these data. Interpret this plot.

(c) Calculate and display the sample ACF for the differences of the logarithms of these data and argue that the logarithms appear to follow a random walk model.

(d) Display the differences of logs in a histogram and interpret.

(e) Display the differences of logs in a quantile-quantile normal plot and interpret.

5.17 Use calculus to show that, for any fixed $x > 0$, as $\lambda \to 0$, $(x^{\lambda} - 1)/\lambda \to \log x$.

Appendix D: The Backshift Operator

Many other books and much of the time series literature use what is called the **backshift operator** to express and manipulate ARIMA models. The backshift operator, denoted B, operates on the time index of a series and shifts time back one time unit to form a new series.[†] In particular,

$$BY_t = Y_{t-1}$$

The backshift operator is linear since for any constants a, b, and c and series Y_t and X_t, it is easy to see that

$$B(aY_t + bX_t + c) = aBY_t + bBX_t + c$$

Consider now the MA(1) model. In terms of B, we can write

$$\begin{aligned} Y_t &= e_t - \theta e_{t-1} = e_t - \theta B e_t = (1 - \theta B)e_t \\ &= \theta(B)e_t \end{aligned}$$

where $\theta(B)$ is the MA characteristic polynomial "evaluated" at B.

Since BY_t is itself a time series, it is meaningful to consider BBY_t. But clearly $BBY_t = BY_{t-1} = Y_{t-2}$, and we can write

$$B^2Y_t = Y_{t-2}$$

More generally, we have

$$B^mY_t = Y_{t-m}$$

for any positive integer m. For a general MA(q) model, we can then write

$$\begin{aligned} Y_t &= e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \\ &= e_t - \theta_1 B e_t - \theta_2 B^2 e_t - \cdots - \theta_q B^q e_t \\ &= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)e_t \end{aligned}$$

or

$$Y_t = \theta(B)e_t$$

where, again, $\theta(B)$ is the MA characteristic polynomial evaluated at B.

For autoregressive models AR(p), we first move all of the terms involving Y to the left-hand side

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} = e_t$$

and then write

$$Y_t - \phi_1 B Y_t - \phi_2 B^2 Y_t - \cdots - \phi_p B^p Y_t = e_t$$

or

[†] Sometimes B is called a **Lag operator**.

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)Y_t = e_t$$

which can be expressed as

$$\phi(B)Y_t = e_t$$

where $\phi(B)$ is the AR characteristic polynomial evaluated at B.

Combining the two, the general ARMA(p,q) model may be written compactly as

$$\phi(B)Y_t = \theta(B)e_t$$

Differencing can also be conveniently expressed in terms of B. We have

$$\begin{aligned}\nabla Y_t &= Y_t - Y_{t-1} = Y_t - BY_t \\ &= (1-B)Y_t\end{aligned}$$

with second differences given by

$$\nabla^2 Y_t = (1-B)^2 Y_t$$

Effectively, $\nabla = 1 - B$ and $\nabla^2 = (1 - B)^2$.

The general ARIMA(p,d,q) model is expressed concisely as

$$\phi(B)(1-B)^d Y_t = \theta(B)e_t$$

In the literature, one must carefully distinguish from the context the use of B as a backshift operator and its use as an ordinary real (or complex) variable. For example, the stationarity condition is frequently given by stating that the roots of $\phi(B) = 0$ must be greater than 1 in absolute value or, equivalently, must lie outside the unit circle in the complex plane. Here B is to be treated as a dummy variable in an equation rather than as the backshift operator.

CHAPTER 6

MODEL SPECIFICATION

We have developed a large class of parametric models for both stationary and nonstationary time series—the ARIMA models. We now begin our study and implementation of *statistical inference* for such models. The subjects of the next three chapters, respectively, are:

1. how to choose appropriate values for p, d, and q for a given series;
2. how to estimate the parameters of a specific ARIMA(p,d,q) model;
3. how to check on the appropriateness of the fitted model and improve it if needed.

Our overall strategy will first be to decide on reasonable—but tentative—values for p, d, and q. Having done so, we shall estimate the ϕ's, θ's, and σ_e for that model in the most efficient way. Finally, we shall look critically at the fitted model thus obtained to check its adequacy, in much the same way that we did in Section 3.6 on page 42. If the model appears inadequate in some way, we consider the nature of the inadequacy to help us select another model. We proceed to estimate that new model and check it for adequacy.

With a few iterations of this model-building strategy, we hope to arrive at the best possible model for a given series. The book by George E. P. Box and G. M. Jenkins (1976) so popularized this technique that many authors call the procedure the "Box-Jenkins method." We begin by continuing our investigation of the properties of the sample autocorrelation function.

6.1 Properties of the Sample Autocorrelation Function

Recall from page 46 the definition of the sample or estimated autocorrelation function. For the observed series $Y_1, Y_2, \ldots, Y_n$, we have

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \qquad \text{for } k = 1, 2, \ldots \tag{6.1.1}$$

Our goal is to recognize, to the extent possible, patterns in r_k that are characteristic of the known patterns in ρ_k for common ARMA models. For example, we know that $\rho_k = 0$ for $k > q$ in an MA(q) model. However, as the r_k are only estimates of the ρ_k, we

need to investigate their sampling properties to facilitate the comparison of estimated correlations with theoretical correlations.

From the definition of r_k, a ratio of quadratic functions of possibly dependent variables, it should be apparent that the sampling properties of r_k will *not* be obtained easily. Even the expected value of r_k is difficult to determine—recall that the expected value of a ratio is *not* the ratio of the respective expected values. We shall be content to accept a general large-sample result and consider its implications in special cases. Bartlett (1946) carried out the original work. We shall take a more general result from Anderson (1971). A recent discussion of these results may be found in Shumway and Stoffer (2006, p. 519).

We suppose that

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j e_{t-j}$$

where the e_t are independent and identically distributed with zero means and finite, nonzero, common variances. We assume further that

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \qquad \text{and} \qquad \sum_{j=0}^{\infty} j\psi_j^2 < \infty$$

(These will be satisfied by any stationary ARMA model.)

Then, for any fixed m, the joint distribution of

$$\sqrt{n}(r_1 - \rho_1),\ \sqrt{n}(r_2 - \rho_2),\ \ldots,\ \sqrt{n}(r_m - \rho_m)$$

approaches, as $n \to \infty$, a joint normal distribution with zero means, variances c_{jj}, and covariances c_{ij},where

$$c_{ij} = \sum_{k=-\infty}^{\infty} (\rho_{k+i}\rho_{k+j} + \rho_{k-i}\rho_{k+j} - 2\rho_i\rho_k\rho_{k+j} - 2\rho_j\rho_k\rho_{k+i} + 2\rho_i\rho_j\rho_k^2) \qquad (6.1.2)$$

For large n, we would say that r_k is approximately normally distributed with mean ρ_k and variance c_{kk}/n. Furthermore, $Corr(r_k, r_j) \approx c_{kj}/\sqrt{c_{kk}c_{jj}}$. Notice that the approximate variance of r_k is inversely proportional to the sample size, but $Corr(r_k, r_j)$ is approximately *constant* for large n.

Since Equation (6.1.2) is clearly difficult to interpret in its present generality, we shall consider some important special cases and simplifications. Suppose first that $\{Y_t\}$ is white noise. Then Equation (6.1.2) reduces considerably, and we obtain

$$Var(r_k) \approx \frac{1}{n} \text{ and } Corr(r_k, r_j) \approx 0 \text{ for } k \neq j \qquad (6.1.3)$$

Next suppose that $\{Y_t\}$ is generated by an AR(1) process with $\rho_k = \phi^k$ for $k > 0$. Then, after considerable algebra and summing several geometric series, Equation (6.1.2) with $i = j$ yields

$$Var(r_k) \approx \frac{1}{n}\left[\frac{(1+\phi^2)(1-\phi^{2k})}{1-\phi^2} - 2k\phi^{2k}\right] \qquad (6.1.4)$$

In particular,

$$Var(r_1) \approx \frac{1-\phi^2}{n} \tag{6.1.5}$$

Notice that the closer ϕ is to ± 1, the more precise our estimate of ρ_1 $(= \phi)$ becomes.

For large lags, the terms in Equation (6.1.4) involving ϕ^k may be ignored, and we have

$$Var(r_k) \approx \frac{1}{n}\left[\frac{1+\phi^2}{1-\phi^2}\right] \text{ for large } k \tag{6.1.6}$$

Notice that here, in contrast to Equation (6.1.5), values of ϕ close to ± 1 imply large variances for r_k. Thus we should not expect nearly as precise estimates of $\rho_k = \phi^k \approx 0$ for large k as we do of $\rho_k = \phi^k$ for small k.

For the AR(1) model, Equation (6.1.2) can also be simplified (after much algebra) for general $0 < i < j$ as

$$c_{ij} = \frac{(\phi^{j-i} - \phi^{j+i})(1+\phi^2)}{1-\phi^2} + (j-i)\phi^{j-i} - (j+i)\phi^{j+i} \tag{6.1.7}$$

In particular, we find

$$Corr(r_1, r_2) \approx 2\phi\sqrt{\frac{1-\phi^2}{1+2\phi^2-3\phi^4}} \tag{6.1.8}$$

Based on Equations (6.1.4) through (6.1.8), Exhibit 6.1 gives approximate standard deviations and correlations for several lags and a few values of ϕ in AR(1) models.

Exhibit 6.1 Large Sample Results for Selected r_k from an AR(1) Model

ϕ	$\sqrt{Var(r_1)}$	$\sqrt{Var(r_2)}$	$Corr(r_1, r_2)$	$\sqrt{Var(r_{10})}$
±0.9	$0.44/\sqrt{n}$	$0.807/\sqrt{n}$	±0.97	$2.44/\sqrt{n}$
±0.7	$0.71/\sqrt{n}$	$1.12/\sqrt{n}$	±0.89	$1.70/\sqrt{n}$
±0.4	$0.92/\sqrt{n}$	$1.11/\sqrt{n}$	±0.66	$1.18/\sqrt{n}$
±0.2	$0.98/\sqrt{n}$	$1.04/\sqrt{n}$	±0.38	$1.04/\sqrt{n}$

For the MA(1) case, Equation (6.1.2) simplifies as follows:

$$c_{11} = 1 - 3\rho_1^2 + 4\rho_1^4 \quad \text{and} \quad c_{kk} = 1 + 2\rho_1^2 \text{ for } k > 1 \tag{6.1.9}$$

Furthermore,

$$c_{12} = 2\rho_1(1-\rho_1^2) \tag{6.1.10}$$

Based on these expressions, Exhibit 6.2 lists large-sample standard deviations and correlations for the sample autocorrelations for several lags and several θ-values. Notice again that the sample autocorrelations can be highly correlated and that the standard deviation of r_k is larger for $k > 1$ than for $k = 1$.

Exhibit 6.2 Large-Sample Results for Selected r_k from an MA(1) Model

θ	$\sqrt{Var(r_1)}$	$\sqrt{Var(r_k)}$ for $k > 1$	$Corr(r_1, r_2)$
±0.9	$0.71/\sqrt{n}$	$1.22/\sqrt{n}$	∓0.86
±0.7	$0.73/\sqrt{n}$	$1.20/\sqrt{n}$	∓0.84
±0.5	$0.79/\sqrt{n}$	$1.15/\sqrt{n}$	∓0.74
±0.4	$0.89/\sqrt{n}$	$1.11/\sqrt{n}$	∓0.53

For a general MA(q) process and $i = j = k$, Equation (6.1.2) reduces to

$$c_{kk} = 1 + 2\sum_{j=1}^{q} \rho_j^2 \quad \text{for } k > q$$

so that

$$Var(r_k) = \frac{1}{n}\left[1 + 2\sum_{j=1}^{q} \rho_j^2\right] \quad \text{for } k > q \tag{6.1.11}$$

For an observed time series, we can replace ρ's by r's, take the square root, and obtain an estimated standard deviation of r_k, that is, the **standard error** of r_k for large lags. A test of the hypothesis that the series is MA(q) could be carried out by comparing r_k to plus and minus two standard errors. We would reject the null hypothesis if and only if r_k lies outside these bounds. In general, we should not expect the sample autocorrelation to mimic the true autocorrelation in great detail. Thus, we should not be surprised to see ripples or "trends" in r_k that have no counterparts in the ρ_k.

6.2 The Partial and Extended Autocorrelation Functions

Since for MA(q) models the autocorrelation function is zero for lags beyond q, the sample autocorrelation is a good indicator of the order of the process. However, the autocorrelations of an AR(p) model do not become zero after a certain number of lags—they die off rather than cut off. So a different function is needed to help determine the order of autoregressive models. Such a function may be defined as the correlation between Y_t and Y_{t-k} *after removing the effect of the intervening variables* $Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-k+1}$. This coefficient is called the partial autocorrelation at lag k and will be denoted by ϕ_{kk}. (The reason for the seemingly redundant double subscript on ϕ_{kk} will become apparent later on in this section.)

There are several ways to make this definition precise. If $\{Y_t\}$ is a normally distributed time series, we can let

$$\phi_{kk} = Corr(Y_t, Y_{t-k} | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}) \tag{6.2.1}$$

That is, ϕ_{kk} is the correlation in the bivariate distribution of Y_t and Y_{t-k} conditional on $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}$.

An alternative approach, not based on normality, can be developed in the following way. Consider predicting Y_t based on a linear function of the intervening variables Y_{t-1}, $Y_{t-2},\ldots, Y_{t-k+1}$, say, $\beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_{k-1} Y_{t-k+1}$, with the β's chosen to minimize the mean square error of prediction. If we assume that the β's have been so chosen and then think backward in time, it follows from stationarity that the best "predictor" of Y_{t-k} based on the same $Y_{t-1}, Y_{t-2},\ldots, Y_{t-k+1}$ will be $\beta_1 Y_{t-k+1} + \beta_2 Y_{t-k+2} + \cdots + \beta_{k-1} Y_{t-1}$. The **partial autocorrelation function** at lag k is then defined to be the correlation between the prediction errors; that is,

$$\begin{aligned}\phi_{kk} = Corr(&Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - \cdots - \beta_{k-1} Y_{t-2},\\ &Y_{t-k} - \beta_1 Y_{t-k+1} - \beta_2 Y_{t-k+2} - \cdots - \beta_{k-1} Y_{t-1})\end{aligned} \tag{6.2.2}$$

(For normally distributed series, it can be shown that the two definitions coincide.) By convention, we take $\phi_{11} = 1$.

As an example, consider ϕ_{22}. It is shown in Appendix F on page 218 that the best linear prediction of Y_t based on Y_{t-1} alone is just $\rho_1 Y_{t-1}$. Thus, according to Equation (6.2.2), we will obtain ϕ_{22} by computing

$$Cov(Y_t - \rho_1 Y_{t-1}, Y_{t-2} - \rho_1 Y_{t-1}) = \gamma_0(\rho_2 - \rho_1^2 - \rho_1^2 + \rho_1^2) = \gamma_0(\rho_2 - \rho_1^2)$$

Since

$$\begin{aligned}Var(Y_t - \rho_1 Y_{t-1}) &= Var(Y_{t-2} - \rho_1 Y_{t-1})\\ &= \gamma_0(1 + \rho_1^2 - 2\rho_1^2)\\ &= \gamma_0(1 - \rho_1^2)\end{aligned}$$

we have that, for any stationary process, the lag 2 partial autocorrelation can be expressed as

$$\phi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{6.2.3}$$

Consider now an AR(1) model. Recall that $\rho_k = \phi^k$ so that

$$\phi_{22} = \frac{\phi^2 - \phi^2}{1 - \phi^2} = 0$$

We shall soon see that for the AR(1) case, $\phi_{kk} = 0$ for all $k > 1$. Thus the partial autocorrelation is nonzero for lag 1, the order of the AR(1) process, but is zero for all lags greater than 1. We shall show this to be generally the case for AR(p) models. Sometimes we say that the partial autocorrelation function for an AR(p) process *cuts off* after the lag exceeds the order of the process.

Consider a general AR(p) case. It will be shown in Chapter 9 that the best linear predictor of Y_t based on a linear function of the variables $Y_{t-1}, Y_{t-2},\ldots, Y_p,\ldots, Y_{t-k+1}$ for $k > p$ is $\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}$. Also, the best linear predictor of Y_{t-k} is some function of $Y_{t-1}, Y_{t-2},\ldots, Y_p,\ldots, Y_{t-k+1}$, call it $h(Y_{t-1}, Y_{t-2},\ldots, Y_p,\ldots, Y_{t-k+1})$. So the covariance between the two prediction errors is

$$
\begin{aligned}
&Cov(Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p}, \\
&\qquad\qquad Y_{t-k} - h(Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_{t-1})) \\
&= Cov(e_t, Y_{t-k} - h(Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_{t-1})) \\
&= 0 \quad \text{since } e_t \text{ is independent of } Y_{t-k}, Y_{t-k+1}, Y_{t-k+2}, \ldots, Y_{t-1}
\end{aligned}
$$

Thus we have established the key fact that, for an AR(p) model,

$$\phi_{kk} = 0 \quad \text{for} \quad k > p \tag{6.2.4}$$

For an MA(1) model, Equation (6.2.3) quickly yields

$$\phi_{22} = \frac{-\theta^2}{1 + \theta^2 + \theta^4} \tag{6.2.5}$$

Furthermore, for the MA(1) case, it may be shown that

$$\phi_{kk} = -\frac{\theta^k(1 - \theta^2)}{1 - \theta^{2(k+1)}} \quad \text{for} \quad k \geq 1 \tag{6.2.6}$$

Notice that the partial autocorrelation of an MA(1) model never equals zero but essentially decays to zero exponentially fast as the lag increases—rather like the autocorrelation function of the AR(1) process. More generally, it can be shown that the partial autocorrelation of an MA(q) model behaves very much like the *autocorrelation* of an AR(q) model.

A general method for finding the partial autocorrelation function for any stationary process with autocorrelation function ρ_k is as follows (see Anderson 1971, pp. 187–188, for example). For a given lag k, it can be shown that the ϕ_{kk} satisfy the Yule-Walker equations (which first appeared in Chapter 4 on page 79):

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \phi_{k3}\rho_{j-3} + \cdots + \phi_{kk}\rho_{j-k} \quad \text{for } j = 1, 2, \ldots, k \tag{6.2.7}$$

More explicitly, we can write these k linear equations as

$$
\left.
\begin{aligned}
\phi_{k1} + \quad \rho_1\phi_{k2} + \quad \rho_2\phi_{k3} + \cdots + \rho_{k-1}\phi_{kk} &= \rho_1 \\
\rho_1\phi_{k1} + \quad \phi_{k2} + \quad \rho_1\phi_{k3} + \cdots + \rho_{k-2}\phi_{kk} &= \rho_2 \\
\vdots \qquad\qquad & \\
\rho_{k-1}\phi_{k1} + \rho_{k-2}\phi_{k2} + \rho_{k-3}\phi_{k3} + \cdots + \phi_{kk} &= \rho_k
\end{aligned}
\right\} \tag{6.2.8}
$$

Here we are treating $\rho_1, \rho_2, \ldots, \rho_k$ as given and wish to solve for $\phi_{k1}, \phi_{k2}, \ldots, \phi_{kk}$ (discarding all but ϕ_{kk}).

These equations yield ϕ_{kk} for any stationary process. However, if the process is in fact AR(p), then since for $k = p$ Equations (6.2.8) are just the Yule-Walker equations (page 79), which the AR(p) model is known to satisfy, we must have $\phi_{pp} = \phi_p$. In addition, as we have already seen by an alternative derivation, $\phi_{kk} = 0$ for $k > p$. Thus the partial autocorrelation effectively displays the correct order p of an autoregressive process as the highest lag k before ϕ_{kk} becomes zero.

The Sample Partial Autocorrelation Function

For an observed time series, we need to be able to estimate the partial autocorrelation function at a variety of lags. Given the relationships in Equations (6.2.8), an obvious method is to estimate the ρ's with sample autocorrelations, the corresponding r's, and then solve the resulting linear equations for $k = 1, 2, 3, \ldots$ to get estimates of ϕ_{kk}. We call the estimated function the **sample partial autocorrelation function** (sample PACF) and denote it by $\hat{\phi}_{kk}$.

Levinson (1947) and Durbin (1960) gave an efficient method for obtaining the solutions to Equations (6.2.8) for either theoretical or sample partial autocorrelations. They showed independently that Equations (6.2.8) can be solved recursively as follows:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\,\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\,\rho_j} \tag{6.2.9}$$

where

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} \quad \text{for } j = 1, 2, \ldots, k-1$$

For example, using $\phi_{11} = \rho_1$ to get started, we have

$$\phi_{22} = \frac{\rho_2 - \phi_{11}\rho_1}{1 - \phi_{11}\rho_1} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

(as before) with $\phi_{21} = \phi_{11} - \phi_{22}\phi_{11}$, which is needed for the next step. Then

$$\phi_{33} = \frac{\rho_3 - \phi_{21}\rho_2 - \phi_{22}\rho_1}{1 - \phi_{21}\rho_1 - \phi_{22}\rho_2}$$

We may thus calculate numerically as many values for ϕ_{kk} as desired. As stated, these recursive equations give us the theoretical partial autocorrelations, but by replacing ρ's with r's, we obtain the estimated or sample partial autocorrelations.

To assess the possible magnitude of the sample partial autocorrelations, Quenoulle (1949) has shown that, under the hypothesis that an AR(p) model is correct, the sample partial autocorrelations at lags greater than p are approximately normally distributed with zero means and variances $1/n$. Thus, for $k > p$, $\pm 2/\sqrt{n}$ can be used as critical limits on $\hat{\phi}_{kk}$ to test the null hypothesis that an AR(p) model is correct.

Mixed Models and the Extended Autocorrelation Function

Exhibit 6.3 summarizes the behavior of the autocorrelation and partial autocorrelation functions that is useful in specifying models.

Exhibit 6.3 General Behavior of the ACF and PACF for ARMA Models

	AR(p)	MA(q)	ARMA(p,q), $p>0$, and $q>0$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

The Extended Autocorrelation Function

The sample ACF and PACF provide effective tools for identifying pure AR(p) or MA(q) models. However, for a mixed ARMA model, its theoretical ACF and PACF have infinitely many nonzero values, making it difficult to identify mixed models from the sample ACF and PACF. Many graphical tools have been proposed to make it easier to identify the ARMA orders, for example, the corner method (Becuin et al., 1980), the extended autocorrelation (EACF) method (Tsay and Tiao, 1984), and the smallest canonical correlation (SCAN) method (Tsay and Tiao, 1985), among others. We shall outline the EACF method, which seems to have good sampling properties for moderately large sample sizes according to a comparative simulation study done by W. S. Chan (1999).

The EACF method uses the fact that if the AR part of a mixed ARMA model is known, "filtering out" the autoregression from the observed time series results in a pure MA process that enjoys the cutoff property in its ACF. The AR coefficients may be estimated by a finite sequence of regressions. We illustrate the procedure for the case where the true model is an ARMA(1,1) model:

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$$

In this case, a simple linear regression of Y_t on Y_{t-1} results in an inconsistent estimator of ϕ, even with infinitely many data. Indeed, the theoretical regression coefficient equals $\rho_1 = (\phi - \theta)(1 - \phi\theta)/(1 - 2\phi\theta + \theta^2)$, not ϕ. But the residuals from this regression do contain information about the error process $\{e_t\}$. A second multiple regression is performed that consists of regressing Y_t on Y_{t-1} and on the lag 1 of the residuals from the first regression. The coefficient of Y_{t-1} in the second regression, denoted by $\tilde{\phi}$, turns out to be a consistent estimator of ϕ. Define $W_t = Y_t - \tilde{\phi} Y_{t-1}$, which is then approximately an MA(1) process. For an ARMA(1,2) model, a third regression that regresses Y_t on its lag 1, the lag 1 of the residuals from the second regression, and the lag 2 of the residuals from the first regression leads to the coefficient of Y_{t-1} being a consistent estimator of ϕ. Similarly, the AR coefficients of an ARMA(p,q) model can be consistently estimated via a sequence of q regressions.

As the AR and MA orders are unknown, an iterative procedure is required. Let

$$W_{t,k,j} = Y_t - \tilde{\phi}_1 Y_{t-1} - \cdots - \tilde{\phi}_k Y_{t-k} \tag{6.2.10}$$

be the autoregressive residuals defined with the AR coefficients estimated iteratively assuming the AR order is k and the MA order is j. The sample autocorrelations of $W_{t,k,j}$ are referred to as the extended sample autocorrelations. For $k = p$ and $j \geq q$, $\{W_{t,k,j}\}$ is approximately an MA(q) model, so that its theoretical autocorrelations of lag $q + 1$ or

higher are equal to zero. For $k > p$, an overfitting problem occurs, and this increases the MA order for the W process by the minimum of $k - p$ and $j - q$. Tsay and Tiao (1984) suggested summarizing the information in the sample EACF by a table with the element in the kth row and jth column equal to the symbol X if the lag $j + 1$ sample correlation of $W_{t,k,j}$ is significantly different from 0 (that is, if its magnitude is greater than $1.96/\sqrt{n-j-k}$ since the sample autocorrelation is asymptotically $N(0,1/(n-k-j))$ if the W's are approximately an MA(j) process) and 0 otherwise. In such a table, an MA(p,q) process will have a theoretical pattern of a triangle of zeroes, with the upper left-hand vertex corresponding to the ARMA orders. Exhibit 6.4 displays the schematic pattern for an ARMA(1,1) model. The upper left-hand vertex of the triangle of zeros is marked with the symbol 0^* and is located in the $p = 1$ row and $q = 1$ column—an indication of an ARMA(1,1) model.

Exhibit 6.4 Theoretical Extended ACF (EACF) for an ARMA(1,1) Model

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	0*	0	0	0	0	0	0	0	0	0	0	0	0
2	x	x	0	0	0	0	0	0	0	0	0	0	0	0
3	x	x	x	0	0	0	0	0	0	0	0	0	0	0
4	x	x	x	x	0	0	0	0	0	0	0	0	0	0
5	x	x	x	x	x	0	0	0	0	0	0	0	0	0
6	x	x	x	x	x	x	0	0	0	0	0	0	0	0
7	x	x	x	x	x	x	x	0	0	0	0	0	0	0

Of course, the *sample* EACF will never be this clear-cut. Displays like Exhibit 6.4 will contain 8×14 = 112 different estimated correlations, and some will be statistically significantly different from zero by chance (see Exhibit 6.17 on page 124, for an example). We will illustrate the use of the EACF in the next two sections and throughout the remainder of the book.

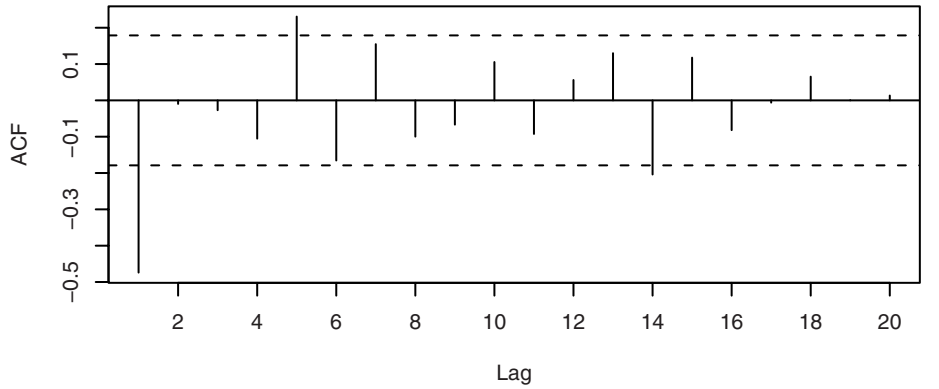
6.3 Specification of Some Simulated Time Series

To illustrate the theory of Sections 6.1 and 6.2, we shall consider the sample autocorrelation and sample partial correlation of some simulated time series.

Exhibit 6.5 displays a graph of the sample autocorrelation out to lag 20 for the simulated time series that we first saw in Exhibit 4.5 on page 61. This series, of length 120, was generated from an MA(1) model with $\theta = 0.9$. From Exhibit 4.1 on page 58, the theoretical autocorrelation at lag 1 is –0.4972. The estimated or sample value shown at lag 1 on the graph is –0.474. Using Exhibit 6.2 on page 112, the approximate standard error

of this estimate is $0.71/\sqrt{n} = 0.71/\sqrt{120} = 0.065$, so the estimate is well within two standard errors of the true value.

Exhibit 6.5 Sample Autocorrelation of an MA(1) Process with $\theta = 0.9$

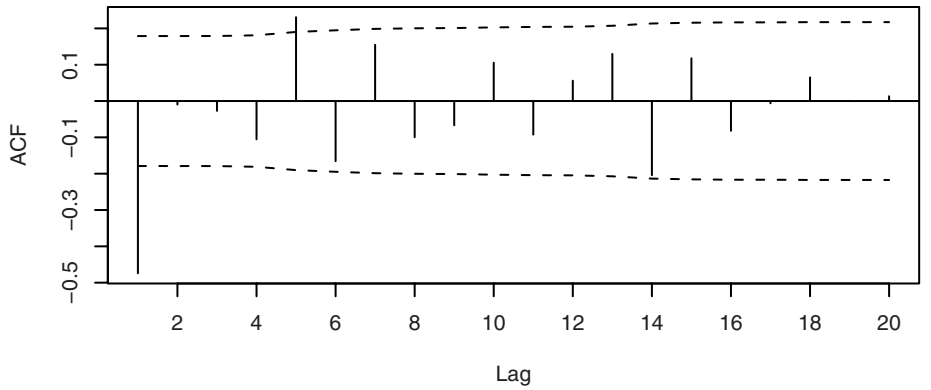



```
> data(ma1.1.s)
> win.graph(width=4.875,height=3,pointsize=8)
> acf(ma1.1.s,xaxp=c(0,20,10))
```

The dashed horizontal lines in Exhibit 6.5, plotted at $\pm 2/\sqrt{n} = \pm 0.1826$, are intended to give critical values for testing whether or not the autocorrelation coefficients are significantly different from zero. These limits are based on the approximate large sample standard error that applies to a white noise process, namely $1/\sqrt{n}$. Notice that the sample ACF values exceed these rough critical values at lags 1, 5, and 14. Of course, the true autocorrelations at lags 5 and 14 are both zero.

Exhibit 6.6 displays the same sample ACF but with critical bounds based on plus and minus two of the more complex standard errors implied by Equation (6.1.11) on page 112. In using Equation (6.1.11), we replace ρ's by r's, let q equal 1, 2, 3,… successively, and take the square root to obtain these standard errors.

Exhibit 6.6 Alternative Bounds for the Sample ACF for the MA(1) Process

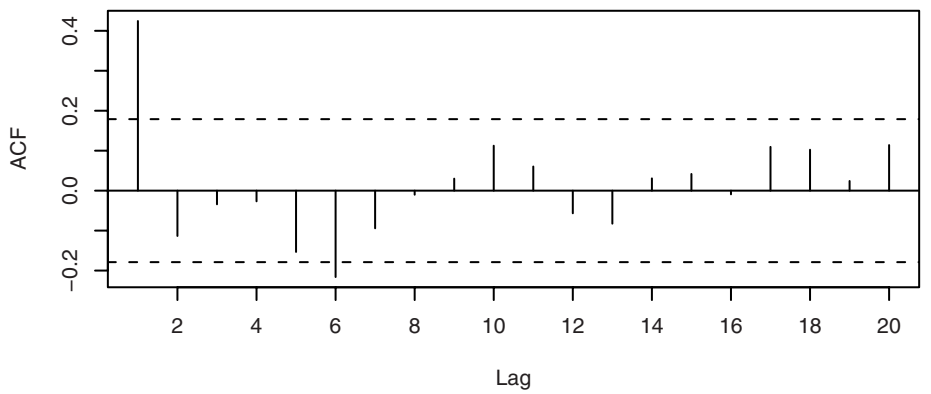



```
> acf(ma1.1.s,ci.type='ma',xaxp=c(0,20,10))
```

Now the sample ACF value at lag 14 is insignificant and the one at lag 5 is just barely significant. The lag 1 autocorrelation is still highly significant, and the information given in these two plots taken together leads us to consider an MA(1) model for this series. Remember that the model is tentative at this point and we would certainly want to consider other "nearby" alternative models when we carry out model diagnostics.

As a second example, Exhibit 6.7 shows the sample ACF for the series shown in Exhibit 4.2 on page 59, generated by an MA(1) model with $\theta = -0.9$. The critical values based on the very approximate standard errors point to an MA(1) model for this series also.

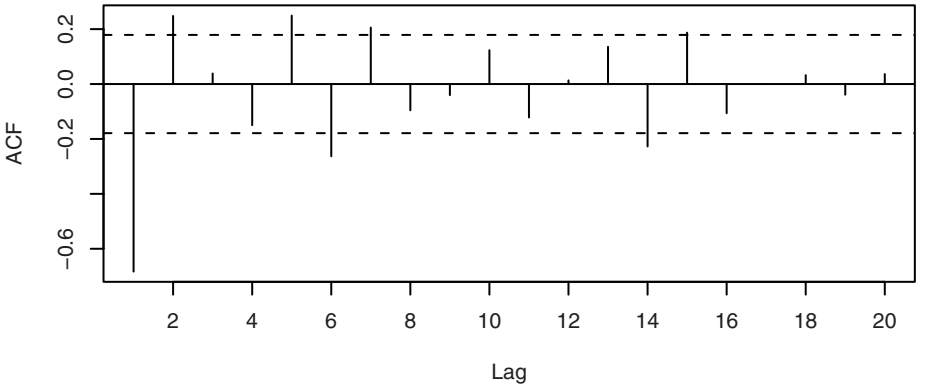
Exhibit 6.7 Sample Autocorrelation for an MA(1) Process with $\theta = -0.9$




```
> data(ma1.2.s); acf(ma1.2.s,xaxp=c(0,20,10))
```

For our third example, we use the data shown in Exhibit 4.8 on page 63, which were simulated from an MA(2) model with $\theta_1 = 1$ and $\theta_2 = -0.6$. The sample ACF displays significance at lags 1, 2, 5, 6, 7, and 14 when we use the simple standard error bounds.

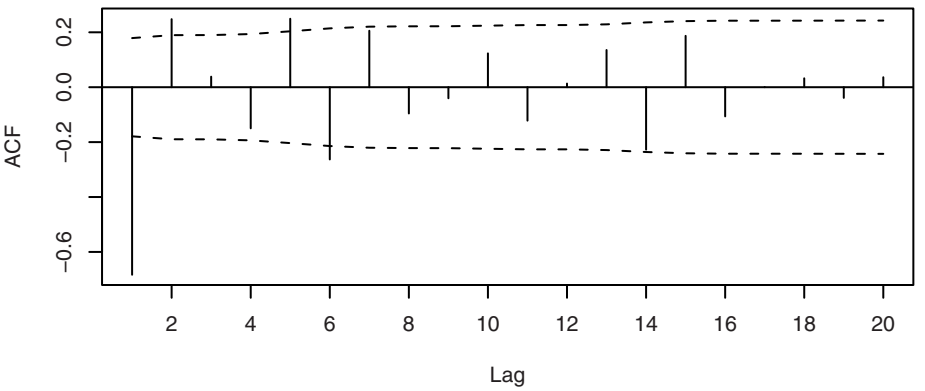
Exhibit 6.8 Sample ACF for an MA(2) Process with $\theta_1 = 1$ and $\theta_2 = -0.6$




```
> data(ma2.s); acf(ma2.s,xaxp=c(0,20,10))
```

Exhibit 6.9 displays the sample ACF with the more sophisticated standard error bounds. Now the lag 2 ACF is no longer significant, and it appears that an MA(1) may be applicable. We will have to wait until we get further along in the model-building process to see that the MA(2) model—the correct one—is the most appropriate model for these data.

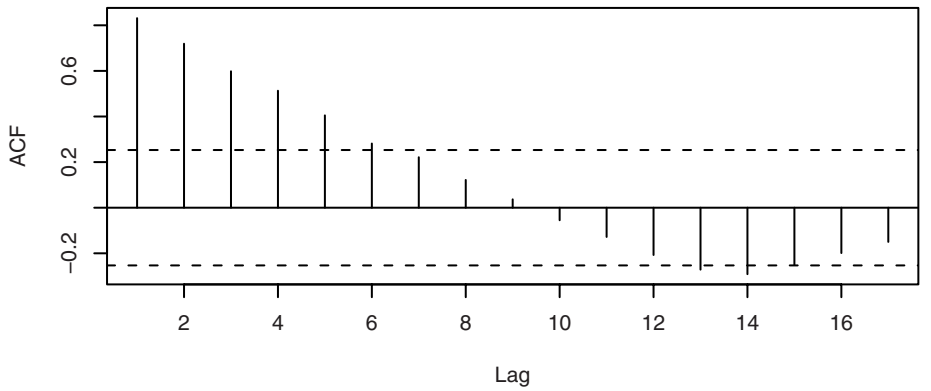
Exhibit 6.9 Alternative Bounds for the Sample ACF for the MA(2) Process




```
> acf(ma2.s,ci.type='ma',xaxp=c(0,20,10))
```

How do these techniques work for autoregressive models? Exhibit 6.10 gives the sample ACF for the simulated AR(1) process we saw in Exhibit 4.13 on page 68. The positive sample ACF values at lags 1, 2, and 3 reflect the strength of the lagged relationships that we saw earlier in Exhibits 4.14, 4.15, and 4.16. However, notice that the sample ACF decreases more linearly than exponentially as theory suggests. Also contrary to theory, the sample ACF goes negative at lag 10 and remains so for many lags.

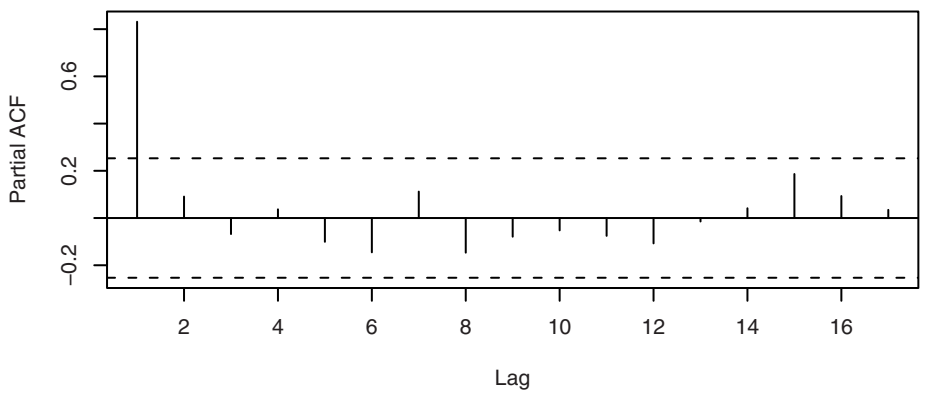
Exhibit 6.10 Sample ACF for an AR(1) Process with $\phi = 0.9$




```
> data(ar1.s); acf(ar1.s,xaxp=c(0,20,10))
```

The sample partial autocorrelation (PACF) shown in Exhibit 6.11, gives a much clearer picture about the nature of the generating model. Based on this graph, we would certainly entertain an AR(1) model for this time series.

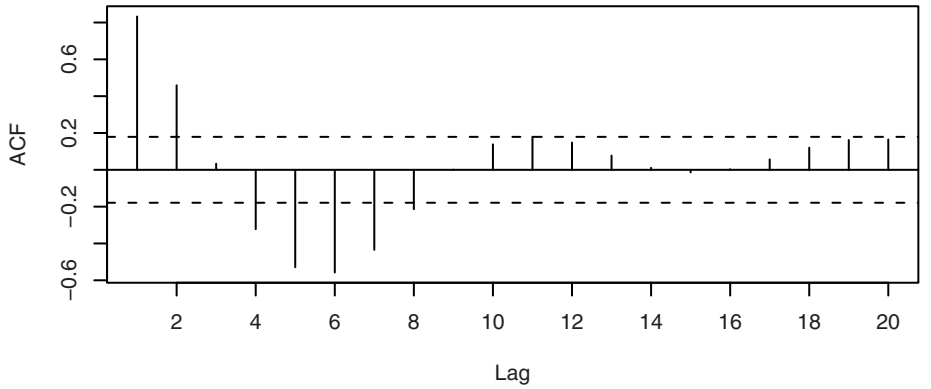
Exhibit 6.11 Sample Partial ACF for an AR(1) Process with $\phi = 0.9$



```
> pacf(ar1.s,xaxp=c(0,20,10))
```

Exhibit 6.12 displays the sample ACF for our AR(2) time series. The time series plot for this series was shown in Exhibit 4.19 on page 74. The sample ACF does look somewhat like the damped wave that Equation (4.3.17) on page 73, and Exhibit 4.18 suggest. However, the sample ACF does not damp down nearly as quickly as theory predicts.

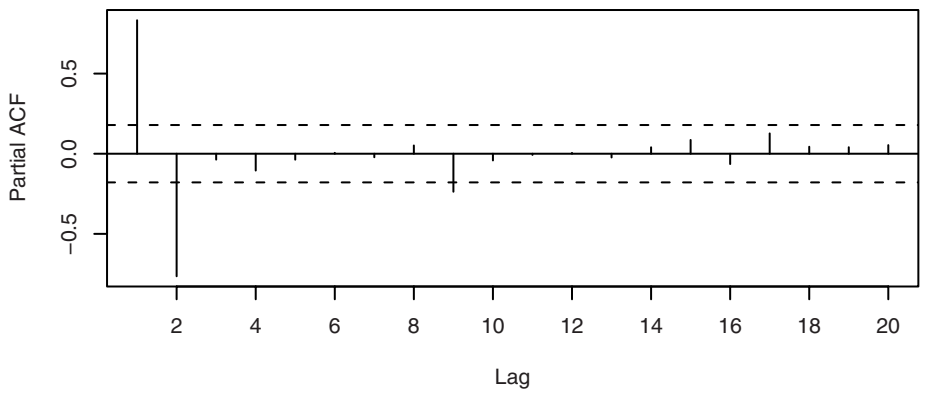
Exhibit 6.12 Sample ACF for an AR(2) Process with $\phi_1 = 1.5$ and $\phi_2 = -0.75$




```
> acf(ar2.s,xaxp=c(0,20,10))
```

The sample PACF in Exhibit 6.13 gives a strong indication that we should consider an AR(2) model for these data. The seemingly significant sample PACF at lag 9 would need to be investigated further during model diagnostics.

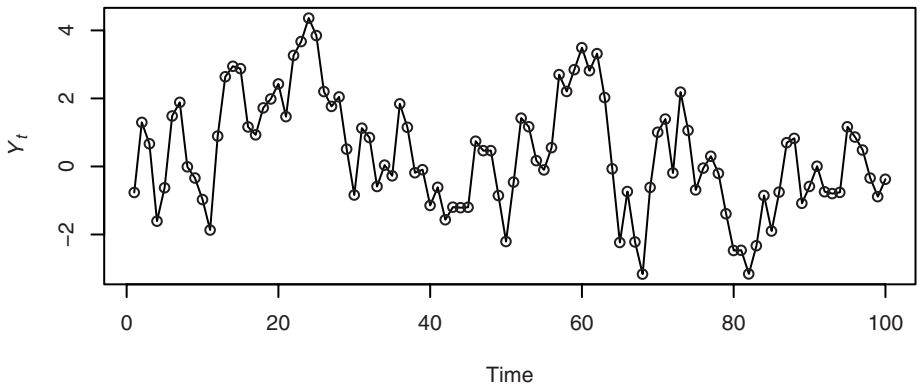
Exhibit 6.13 Sample PACF for an AR(2) Process with $\phi_1 = 1.5$ and $\phi_2 = -0.75$



```
> pacf(ar2.s,xaxp=c(0,20,10))
```

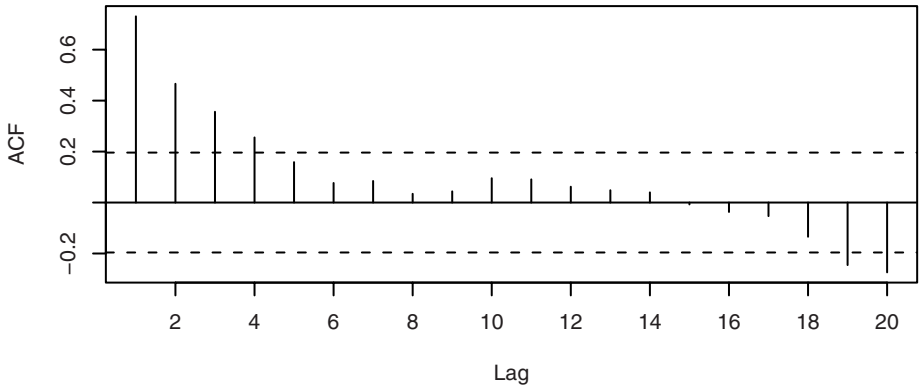
As a final example, we simulated 100 values of a mixed ARMA(1,1) model with $\phi = 0.6$ and $\theta = -0.3$. The time series plot is shown in Exhibit 6.14 and the sample ACF and PACFs are shown in Exhibit 6.15 and Exhibit 6.16, respectively. These seem to indicate that an AR(1) model should be specified.

Exhibit 6.14 Simulated ARMA(1,1) Series with $\phi = 0.6$ and $\theta = -0.3$.



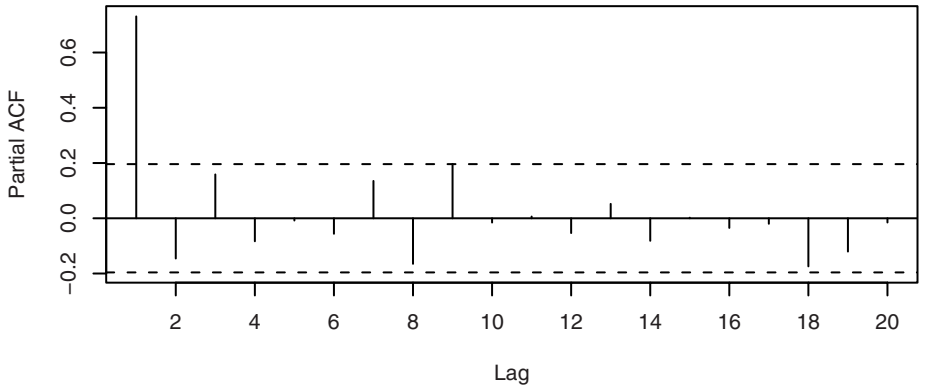

```
> data(arma11.s)
> plot(arma11.s, type='o',ylab=expression(Y[t]))
```

Exhibit 6.15 Sample ACF for Simulated ARMA(1,1) Series




```
> acf(arma11.s,xaxp=c(0,20,10))
```

Exhibit 6.16 Sample PACF for Simulated ARMA(1,1) Series




```
> pacf(arma11.s,xaxp=c(0,20,10))
```

However, the triangular region of zeros shown in the sample EACF in Exhibit 6.17 indicates quite clearly that a mixed model with $q = 1$ and with $p = 1$ or 2 would be more appropriate. We will illustrate further uses of the EACF when we specify some real series in Section 6.6.

Exhibit 6.17 Sample EACF for Simulated ARMA(1,1) Series

AR / MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	o	o	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	o	x	o	o	o	o	o	o	o	o	o	o	o
5	x	o	o	o	o	o	o	o	o	o	o	o	o	o
6	x	o	o	o	x	o	o	o	o	o	o	o	o	o
7	x	o	o	o	x	o	o	o	o	o	o	o	o	o

```
> eacf(arma11.s)
```

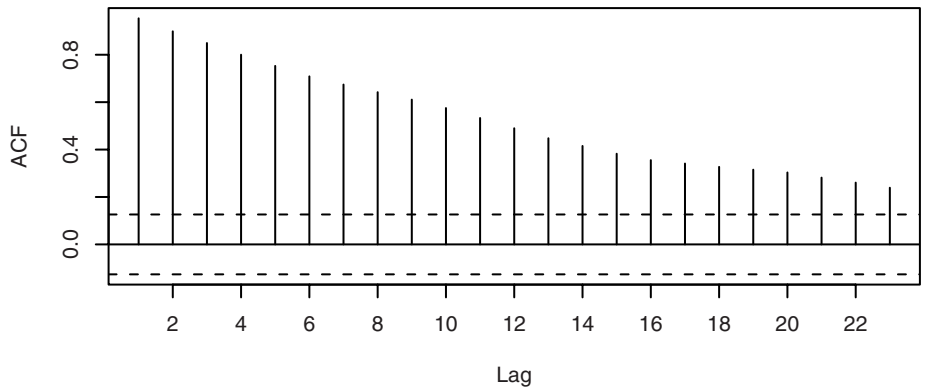
6.4 Nonstationarity

As indicated in Chapter 5, many series exhibit nonstationarity that can be explained by integrated ARMA models. The nonstationarity will frequently be apparent in the time series plot of the series. A review of Exhibits 5.1, 5.5, and 5.8 is recommended here.

The sample ACF computed for nonstationary series will also usually indicate the nonstationarity. The definition of the sample autocorrelation function implicitly *assumes* stationarity; for example, we use lagged products of deviations from the overall mean, and the denominator assumes a constant variance over time. Thus it is not at all clear what the sample ACF is estimating for a nonstationary process. Nevertheless, for nonstationary series, the sample ACF typically fails to die out rapidly as the lags increase. This is due to the tendency for nonstationary series to drift slowly, either up or down, with apparent "trends." The values of r_k need not be large even for low lags, but often they are.

Consider the oil price time series shown in Exhibit 5.1 on page 88. The sample ACF for the logarithms of these data is displayed in Exhibit 6.18. All values shown are "significantly far from zero," and the only pattern is perhaps a linear decrease with increasing lag. The sample PACF (not shown) is also indeterminate.

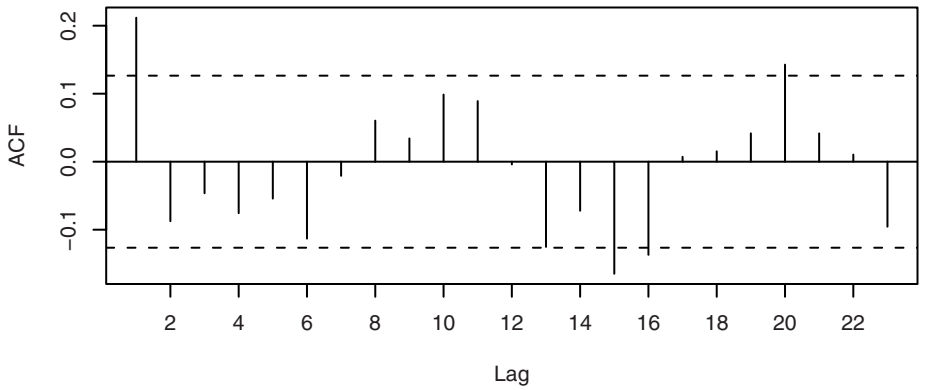
Exhibit 6.18 Sample ACF for the Oil Price Time Series




```
> data(oil.price)
> acf(as.vector(oil.price),xaxp=c(0,24,12))
```

The sample ACF computed on the first differences of the logs of the oil price series is shown in Exhibit 6.19. Now the pattern emerges much more clearly—after differencing, a moving average model of order 1 seems appropriate. The model for the original oil price series would then be a nonstationary IMA(1,1) model. (The "significant" ACF at lags 15, 16, and 20 are ignored for now.)

Exhibit 6.19 Sample ACF for the Difference of the Log Oil Price Series




```
> acf(diff(as.vector(log(oil.price))),xaxp=c(0,24,12))
```

If the first difference of a series and its sample ACF do not appear to support a stationary ARMA model, then we take another difference and again compute the sample ACF and PACF to look for characteristics of a stationary ARMA process. Usually one or at most two differences, perhaps combined with a logarithm or other transformation, will accomplish this reduction to stationarity. Additional properties of the sample ACF computed on nonstationary data are given in Wichern (1973), Roy (1977), and Hasza (1980). See also Box, Jenkins, and Reinsel (1994, p. 218).

Overdifferencing

From Exercise 2.6 on page 20, we know that the difference of any stationary time series is also stationary. However, overdifferencing introduces unnecessary correlations into a series and will complicate the modeling process.

For example, suppose our observed series, $\{Y_t\}$, is in fact a random walk so that one difference would lead to a very simple white noise model

$$\nabla Y_t = Y_t - Y_{t-1} = e_t$$

However, if we difference once more (that is, overdifference) we have

$$\nabla^2 Y_t = e_t - e_{t-1}$$

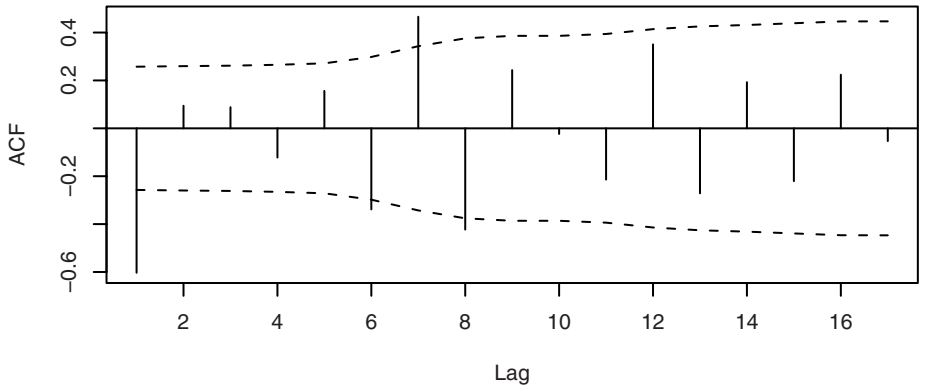
which is an MA(1) model but with $\theta = 1$. If we take two differences in this situation we unnecessarily have to estimate the unknown value of θ. Specifying an IMA(2,1) model would not be appropriate here. The random walk model, which can be thought of as IMA(1,1) with $\theta = 0$, is the correct model.[†] Overdifferencing also creates a noninvert-

[†] The random walk model can also be thought of as an ARI(1,1) with $\phi = 0$ or as a nonstationary AR(1) with $\phi = 1$.

ible model—see Section 4.5 on page 79.[†] Noninvertible models also create serious problems when we attempt to estimate their parameters—see Chapter 7.

To illustrate overdifferencing, consider the random walk shown in Exhibit 2.1 on page 14. Taking one difference should lead to white noise—a very simple model. If we mistakenly take two differences (that is, overdifference) and compute the sample ACF, we obtain the graph shown in Exhibit 6.20. Based on this plot, we would likely specify at least an IMA(2,1) model for the original series and then estimate the unnecessary MA parameter. We also have a significant sample ACF value at lag 7 to think about and deal with.

Exhibit 6.20 Sample ACF of Overdifferenced Random Walk

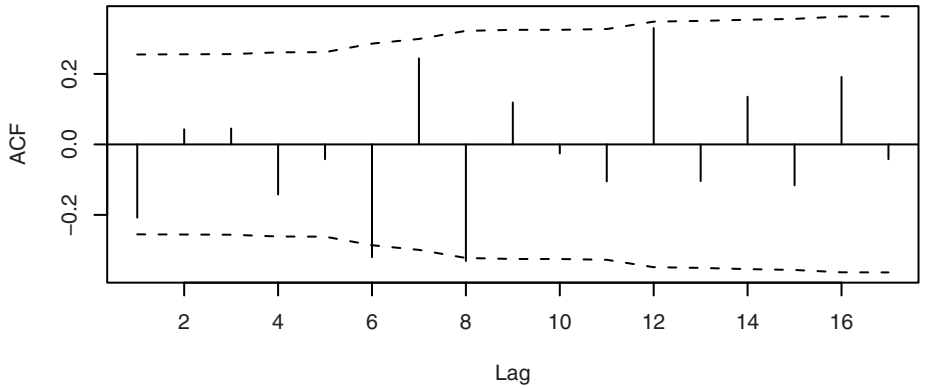



```
> data(rwalk)
> acf(diff(rwalk,difference=2),ci.type='ma', xaxp=c(0,18,9))
```

In contrast, Exhibit 6.21 displays the sample ACF of the *first* difference of the random walk series. Viewing this graph, we would likely want to consider the correct model—the first difference looks very much like white noise.

[†] In backshift notation, if the correct model is $\phi(B)(1-B)Y_t = \theta(B)e_t$, overdifferencing leads to $\phi(B)(1-B)^2Y_t = \theta(B)(1-B)e_t = \theta'(B)e_t$, say, where $\theta'(B) = (1-B)\theta(B)$ and the "forbidden" root in $\theta'(B)$ at $B = 1$ is obvious.

Exhibit 6.21 Sample ACF of Correctly Differenced Random Walk




```
> acf(diff(rwalk),ci.type='ma',xaxp=c(0,18,9))
```

To avoid overdifferencing, we recommend looking carefully at each difference in succession and keeping the principle of parsimony always in mind—*models should be simple, but not too simple*.

The Dickey-Fuller Unit-Root Test

While the approximate linear decay of the sample ACF is often taken as a symptom that the underlying time series is nonstationary and requires differencing, it is also useful to quantify the evidence of nonstationarity in the data-generating mechanism. This can be done via hypothesis testing. Consider the model

$$Y_t = \alpha Y_{t-1} + X_t \text{ for } t = 1, 2, \ldots$$

where $\{X_t\}$ is a stationary process. The process $\{Y_t\}$ is nonstationary if the coefficient $\alpha = 1$, but it is stationary if $|\alpha| < 1$. Suppose that $\{X_t\}$ is an AR(k) process: $X_t = \phi_1 X_{t-1} + \cdots + \phi_k X_{t-k} + e_t$. Under the null hypothesis that $\alpha = 1$, $X_t = Y_t - Y_{t-1}$. Letting $a = \alpha - 1$, we have

$$\begin{aligned} Y_t - Y_{t-1} &= (\alpha - 1)Y_{t-1} + X_t \\ &= aY_{t-1} + \phi_1 X_{t-1} + \cdots + \phi_k X_{t-k} + e_t \\ &= aY_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \cdots + \phi_k(Y_{t-k} - Y_{t-k-1}) + e_t \end{aligned} \tag{6.4.1}$$

where $a = 0$ under the hypothesis that Y_t is difference nonstationary. On the other hand, if $\{Y_t\}$ is stationary so that $-1 < \alpha < 1$, then it can be verified that Y_t still satisfies an equation similar to the equation above but with different coefficients; for example, $a = (1 - \phi_1 - \cdots - \phi_k)(1 - \alpha) < 0$. Indeed, $\{Y_t\}$ is then an AR($k + 1$) process whose AR characteristic equation is given by $\Phi(x)(1 - \alpha x) = 0$, where $\Phi(x) = 1 - \phi_1 x - \cdots - \phi_k x^k$. So, the null hypothesis corresponds to the case where the AR characteristic polynomial has a unit root and the alternative hypothesis states that it has no unit roots. *Consequently, the*

test for differencing amounts to testing for a unit root in the AR characteristic polynomial of $\{Y_t\}$.

By the analysis above, the null hypothesis that $\alpha = 1$ (equivalently $a = 0$) can be tested by regressing the first difference of the observed time series on lag 1 of the observed series and on the past k lags of the first difference of the observed series. We then test whether the coefficient $a = 0$—the null hypothesis being that the process is difference nonstationary. That is, the process is nonstationary but becomes stationary after first differencing. The alternative hypothesis is that $a < 0$ and hence $\{Y_t\}$ is stationary. The augmented Dickey-Fuller (ADF) test statistic is the t-statistic of the estimated coefficient of a from the method of least squares regression. However, the ADF test statistic is not approximately t-distributed under the null hypothesis; instead, it has a certain nonstandard large-sample distribution under the null hypothesis of a unit root. Fortunately, percentage points of this limit (null) distribution have been tabulated; see Fuller (1996).

In practice, even after first differencing, the process may not be a finite-order AR process, but it may be closely approximated by some AR process with the AR order increasing with the sample size. Said and Dickey (1984) (see also Chang and Park, 2002) showed that with the AR order increasing with the sample size, the ADF test has the same large-sample null distribution as the case where the first difference of the time series is a finite-order AR process. Often, the approximating AR order can be first estimated based on some information criteria (for example, AIC or BIC) before carrying out the ADF test. See Section 6.5 on page 130 for more information on the AIC and BIC criteria.

In some cases, the process may be trend nonstationary in the sense that it has a deterministic trend (for example, some linear trend) but otherwise is stationary. A unit-root test may be conducted with the aim of discerning difference stationarity from trend stationarity. This can be done by carrying out the ADF test with the detrended data. Equivalently, this can be implemented by regressing the first difference on the covariates defining the trend, the lag 1 of the original data, and the past lags of the first difference of the original data. The t-statistic based on the coefficient estimate of the lag 1 of the original data furnishes the ADF test statistic, which has another nonstandard large-sample null distribution. See Phillips and Xiao (1998) for a survey of unit root testing.

We now illustrate the ADF test with the simulated random walk shown in Exhibit 2.1 on page 14. First, we consider testing the null hypothesis of unit root versus the alternative hypothesis that the time series is stationary with unknown mean. Hence, the regression defined by Equation (6.4.1) is augmented with an intercept to allow for the possibly nonzero mean under the alternative hypothesis. (For the alternative hypothesis that the process is a stationary process of zero mean, the ADF test statistic can be obtained by running the unaugmented regression defined by Equation (6.4.1).) To carry out the test, it remains to determine k.[†] According to the AIC with the first difference of the data, k is found to be 8, in which case the ADF test statistic becomes –0.601, with the p-value being greater than 0.1.[‡] On the other hand, setting $k = 0$ (the true order) leads

[†] R code: `ar(diff(rwalk))`

to the ADF statistic –1.738, with p-value still greater than 0.1.[†] Thus, there is strong evidence supporting the unit-root hypothesis. Second, recall that the simulated random walk appears to have a linear trend. Hence, linear trend plus stationary error forms another reasonable alternative to the null hypothesis of unit root (difference nonstationarity). For this test, we include both an intercept term and the covariate time in the regression defined by Equation (6.4.1). With $k = 8$, the ADF test statistic equals –2.289 with p-value greater than 0.1; that is, we do not reject the null hypothesis of unit root. On the other hand, setting $k = 0$, the true order that is unknown in practice, the ADF test statistic becomes –3.49 with p-value equal to 0.0501.[‡] Hence, there is weak evidence that the process is linear-trend nonstationary; that is, the process equals linear time trend plus stationary error, contrary to the truth that the process is a random walk, being difference nonstationary! This example shows that with a small sample size, it may be hard to differentiate between trend nonstationarity and difference nonstationarity.

6.5 Other Specification Methods

A number of other approaches to model specification have been proposed since Box and Jenkins' seminal work. One of the most studied is **Akaike's** (1973) **Information Criterion** (**AIC**). This criterion says to select the model that minimizes

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2k \tag{6.5.1}$$

where $k = p + q + 1$ if the model contains an intercept or constant term and $k = p + q$ otherwise. Maximum likelihood estimation is discussed in Chapter 7. The addition of the term $2(p + q + 1)$ or $2(p + q)$ serves as a "penalty function" to help ensure selection of parsimonious models and to avoid choosing models with too many parameters.

The AIC is an estimator of the average Kullback-Leibler divergence of the estimated model from the true model. Let $p(y_1,y_2,\ldots,y_n)$ be the true pdf of $Y_1, Y_2, \ldots, Y_n$, and $q_\theta(y_1,y_2,\ldots,y_n)$ be the corresponding pdf under the model with parameter θ. The Kullback-Leibler divergence of q_θ from p is defined by the formula

$$D(p, q_\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p(y_1, y_2, \ldots, y_n)\log\left[\frac{p(y_1, y_2, \ldots, y_n)}{q_\theta(y_1, y_2, \ldots, y_n)}\right]dy_1 dy_2 \ldots dy_n$$

The AIC estimates $E[D(p, q_{\hat{\theta}})]$, where $\hat{\theta}$ is the maximum likelihood estimator of the vector parameter θ. However, the AIC is a biased estimator, and the bias can be appreciable for large parameter per data ratios. Hurvich and Tsai (1989) showed that the bias can be approximately eliminated by adding another nonstochastic penalty term to the AIC, resulting in the corrected AIC, denoted by AIC_C and defined by the formula

---

[‡] R code: `library(uroot); ADF.test(rwalk,selectlags=list (mode=c(1,2,3,4,5,6,7,8),Pmax=8),itsd=c(1,0,0))`

[†] `ADF.test(rwalk,selectlags=list(mode=c(1,2,3,4,5,6,7,8),Pmax=8), itsd=c(1,1,0))`

[‡] `ADF.test(rwalk,selectlags=list(Pmax=0),itsd=c(1,1,0))`

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2(k+1)(k+2)}{n-k-2} \tag{6.5.2}$$

Here n is the (effective) sample size and again k is the total number of parameters as above excluding the noise variance. Simulation results by Hurvich and Tsai (1989) suggest that for cases with k/n greater than 10%, the $\mathrm{AIC_C}$ outperforms many other model selection criteria, including both the AIC and BIC.

Another approach to determining the ARMA orders is to select a model that minimizes the Schwarz **Bayesian Information Criterion** (**BIC**) defined as

$$\mathrm{BIC} = -2\log(\text{maximum likelihood}) + k\log(n) \tag{6.5.3}$$

If the true process follows an ARMA(p,q) model, then it is known that the orders specified by minimizing the BIC are consistent; that is, they approach the true orders as the sample size increases. However, if the true process is not a finite-order ARMA process, then minimizing AIC among an increasingly large class of ARMA models enjoys the appealing property that it will lead to an optimal ARMA model that is closest to the true process among the class of models under study.†

Regardless of whether we use the AIC or BIC, the methods require carrying out maximum likelihood estimation. However, maximum likelihood estimation for an ARMA model is prone to numerical problems due to multimodality of the likelihood function and the problem of overfitting when the AR and MA orders exceed the true orders. Hannan and Rissanen (1982) proposed an interesting and practical solution to this problem. Their procedure consists of first fitting a high-order AR process with the order determined by minimizing the AIC. The second step uses the residuals from the first step as proxies for the unobservable error terms. Thus, an ARMA(k,j) model can be approximately estimated by regressing the time series on its own lags 1 to k together with the lags 1 to j of the residuals from the high order autoregression; the BIC of this autoregressive model is an estimate of the BIC obtained with maximum likelihood estimation. Hannan and Rissanen (1982) demonstrated that minimizing the approximate BIC still leads to consistent estimation of the ARMA orders.

Order determination is related to the problem of finding the subset of nonzero coefficients of an ARMA model with sufficiently high ARMA orders. A subset ARMA(p,q) model is an ARMA(p,q) model with a subset of its coefficients known to be zero. For example, the model

$$Y_t = 0.8Y_{t-12} + e_t + 0.7e_{t-12} \tag{6.5.4}$$

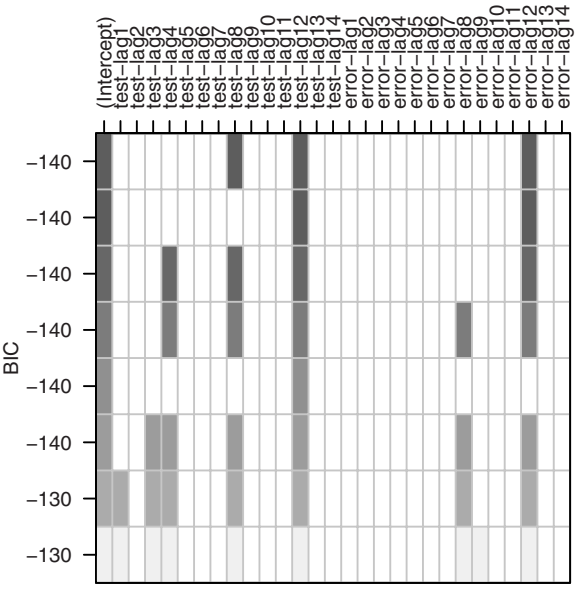
is a subset ARMA(12,12) model useful for modeling some monthly seasonal time series. For ARMA models of very high orders, such as the preceding ARMA(12,12) model, finding a subset ARMA model that adequately approximates the underlying process is more important from a practical standpoint than simply determining the ARMA orders. The method of Hannan and Rissanen (1982) for estimating the ARMA orders can be extended to solving the problem of finding an optimal subset ARMA model.

† Closeness is measured in terms of the Kullback-Leibler divergence—a measure of disparity between models. See Shibata (1976) and the discussion in Stenseth et al. (2004).

Indeed, several model selection criteria (including AIC and BIC) of the subset ARMA(p,q) models (2^{p+q} of them!) can be approximately, exhaustively, and quickly computed by the method of regression by leaps and bounds (Furnival and Wilson, 1974) applied to the subset regression of Y_t on its own lags and on lags of the residuals from a high-order autoregression of $\{Y_t\}$.

It is prudent to examine a few best subset ARMA models (in terms of, for example, BIC) in order to arrive at some helpful tentative models for further study. The pattern of which lags of the observed time series and which of the error process enter into the various best subset models can be summarized succinctly in a display like that shown in Exhibit 6.22. This table is based on a simulation of the ARMA(12,12) model shown in Equation (6.5.4). Each row in the exhibit corresponds to a subset ARMA model where the cells of the variables selected for the model are shaded. The models are sorted according to their BIC, with better models (lower BIC) placed in higher rows and with darker shades. The top row tells us that the subset ARMA(14,14) model with the smallest BIC contains only lags 8 and 12 of the observed time series and lag 12 of the error process. The next best model contains lag 12 of the time series and lag 8 of the errors, while the third best model contains lags 4, 8, and 12 of the time series and lag 12 of the errors. In our simulated time series, the second best model is the true subset model. However, the BIC values for these three models are all very similar, and all three (plus the fourth best model) are worthy of further study. However, lag 12 of the time series and that of the errors are the two variables most frequently found in the various subset models summarized in the exhibit, suggesting that perhaps they are the more important variables, as we know they are!

Exhibit 6.22 Best Subset ARMA Selection Based on BIC



```
> set.seed(92397)
> test=arima.sim(model=list(ar=c(rep(0,11),.8),
    ma=c(rep(0,11),0.7)),n=120)
> res=armasubsets(y=test,nar=14,nma=14,y.name='test',
    ar.method='ols')
> plot(res)
```

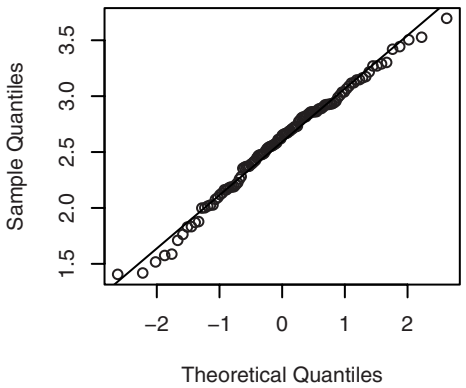
6.6 Specification of Some Actual Time Series

Consider now specification of models for some of the actual time series that we saw in earlier chapters.

The Los Angeles Annual Rainfall Series

Annual total rainfall amounts for Los Angeles were shown in Exhibit 1.1 on page 2. In Chapter 3, we noted in Exhibit 3.17 on page 50, that rainfall amounts were not normally distributed. As is shown in Exhibit 6.23, taking logarithms improves the normality dramatically.

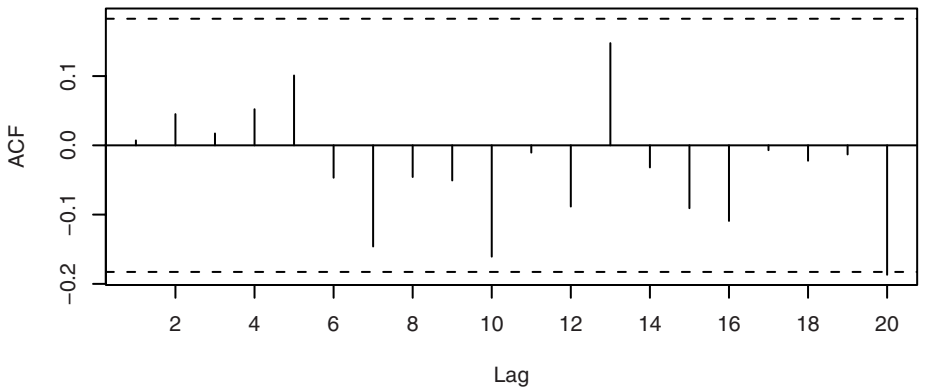
Exhibit 6.23 QQ Normal Plot of the Logarithms of LA Annual Rainfall




```
> data(larain); win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(log(larain)); qqline(log(larain))
```

Exhibit 6.24 displays the sample autocorrelations for the logarithms of the annual rainfall series.

Exhibit 6.24 Sample ACF of the Logarithms of LA Annual Rainfall



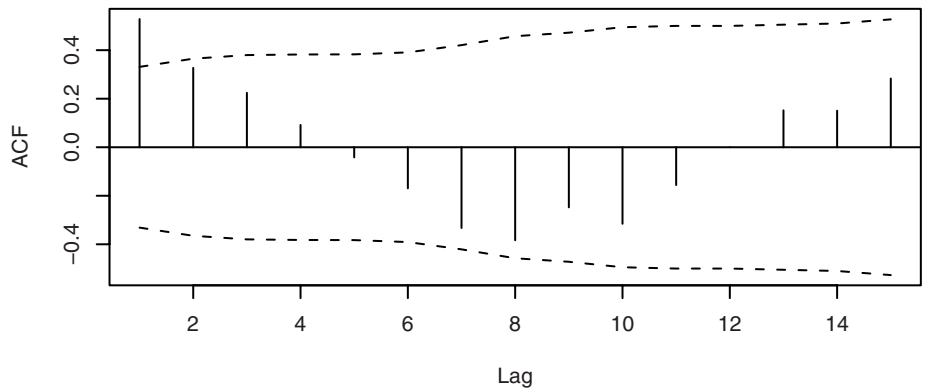

```
> win.graph(width=4.875,height=3,pointsize=8)
> acf(log(larain),xaxp=c(0,20,10))
```

The log transformation has improved the normality, but there is no discernable dependence in this time series. We could model the logarithm of annual rainfall amount as independent, normal random variables with mean 2.58 and standard deviation 0.478. Both these values are in units of log(inches).

The Chemical Process Color Property Series

The industrial chemical process color property displayed in Exhibit 1.3 on page 3, shows more promise of interesting time series modeling—especially in light of the dependence of successive batches shown in Exhibit 1.4 on page 4. The sample ACF plotted in Exhibit 6.25 might at first glance suggest an MA(1) model, as only the lag 1 autocorrelation is significantly different from zero.

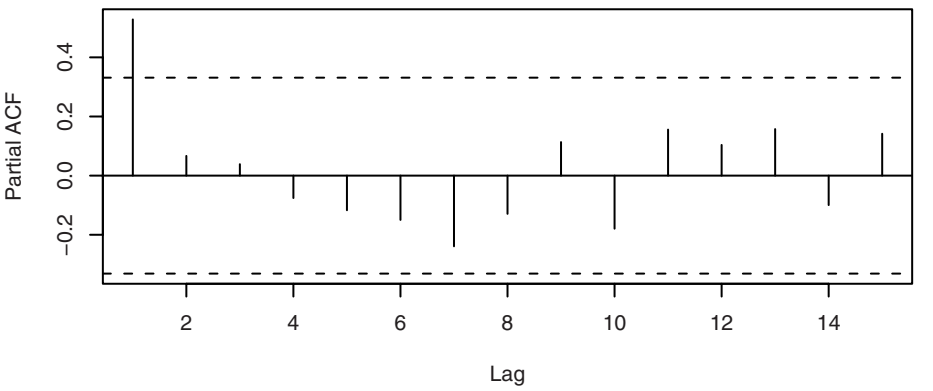
Exhibit 6.25 Sample ACF for the Color Property Series




```
> data(color); acf(color,ci.type='ma')
```

However, the damped sine wave appearance of the plot encourages us to look further at the sample partial autocorrelation. Exhibit 6.26 displays that plot, and now we see clearly that an AR(1) model is worthy of first consideration. As always, our specified models are tentative and subject to modification during the model diagnostics stage of model building.

Exhibit 6.26 Sample Partial ACF for the Color Property Series

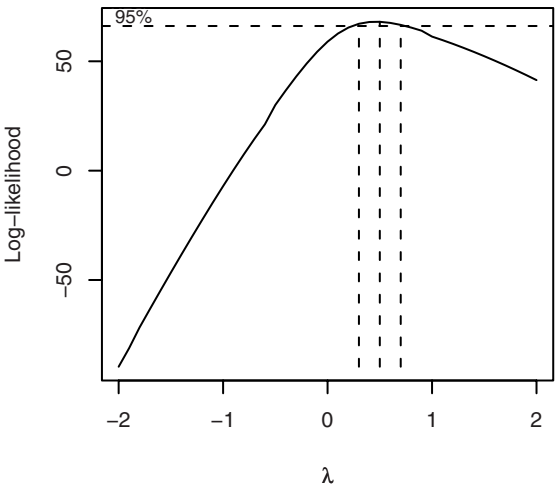



```
> pacf(color)
```

The Annual Abundance of Canadian Hare Series

The time series of annual abundance of hare of the Hudson Bay in Canada was displayed in Exhibit 1.5 on page 5, and the year-to-year dependence was demonstrated in Exhibit 1.6. It has been suggested in the literature that a transformation might be used to produce a good model for these data. Exhibit 6.27 displays the log-likelihood as a function of the power parameter, λ. The maximum occurs at $\lambda = 0.4$, but a square root transformation with $\lambda = 0.5$ is well within the confidence interval for λ. We will take the square root of the abundance values for all further analyses.

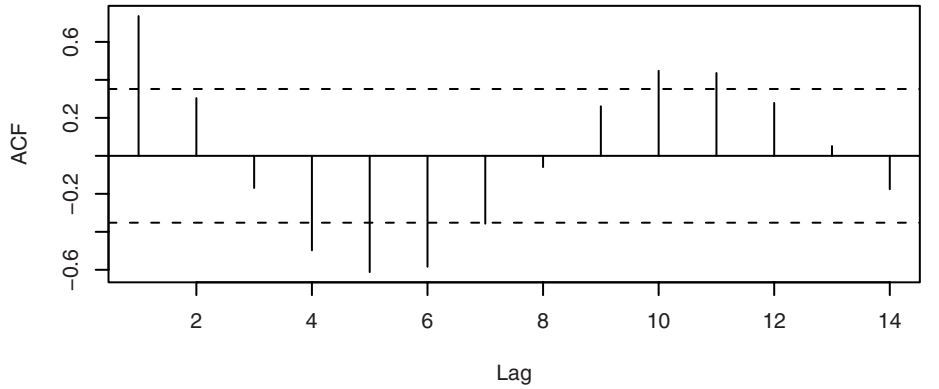
Exhibit 6.27 Box-Cox Power Transformation Results for Hare Abundance




```
> win.graph(width=3,height=3,pointsize=8)
> data(hare); BoxCox.ar(hare)
```

Exhibit 6.28 shows the sample ACF for this transformed series. The fairly strong lag 1 autocorrelation dominates but, again, there is a strong indication of damped oscillatory behavior.

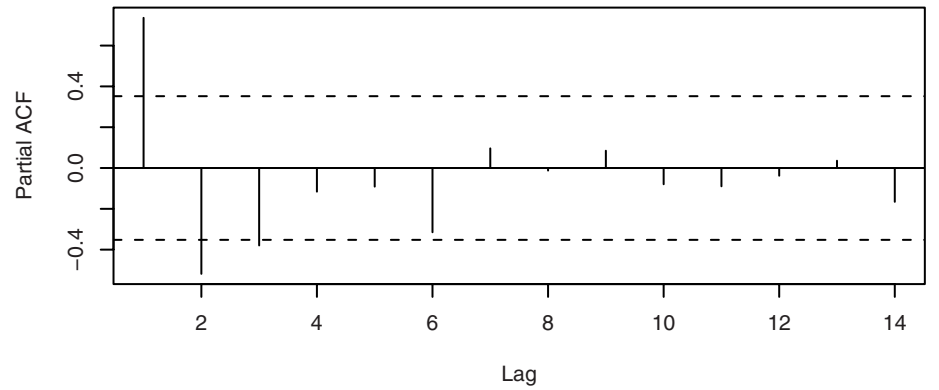
Exhibit 6.28 Sample ACF for Square Root of Hare Abundance




```
> acf(hare^.5)
```

The sample partial autocorrelation for the transformed series is shown in Exhibit 6.29. It gives strong evidence to support an AR(2) or possibly an AR(3) model for these data.

Exhibit 6.29 Sample Partial ACF for Square Root of Hare Abundance




```
> pacf(hare^.5)
```

The Oil Price Series

In Chapter 5, we began to look at the monthly oil price time series and argued graphically that the difference of the logarithms could be considered stationary—see Exhibit 5.1 on page 88. Software implementation of the Augmented Dickey-Fuller unit-root test applied to the logs of the original prices leads to a test statistic of −1.1119 and a p-value of 0.9189. With stationarity as the alternative hypothesis, this provides strong evidence of nonstationarity and the appropriateness of taking a difference of the logs. For this test, the software chose a value of $k = 6$ in Equation (6.4.1) on page 128 based on large-sample theory.

Exhibit 6.30 shows the summary EACF table for the differences of the logarithms of the oil price data. This table suggests an ARMA model with $p = 0$ and $q = 1$.

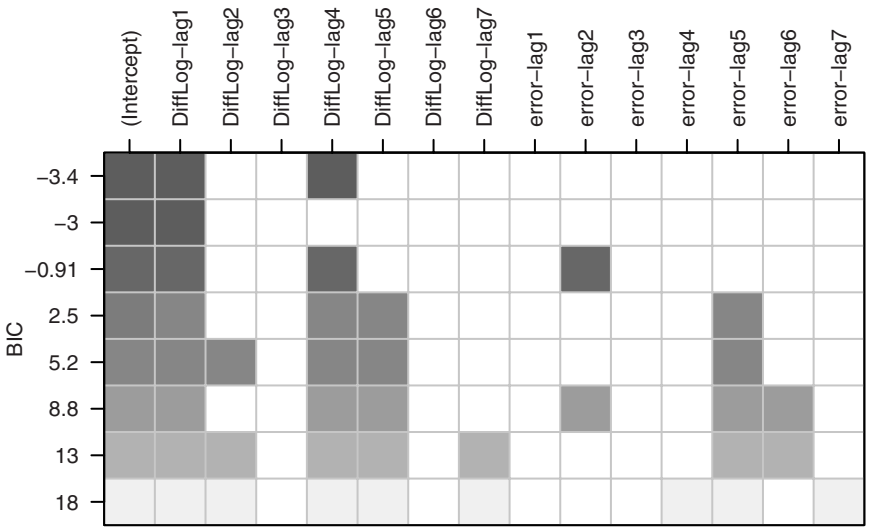
Exhibit 6.30 Extended ACF for Difference of Logarithms of Oil Price Series

AR / MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	o	o	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	o	x	o	o	o
2	o	x	o	o	o	o	o	o	o	o	o	o	o	o
3	o	x	o	o	o	o	o	o	o	o	o	o	o	o
4	o	x	x	o	o	o	o	o	o	o	o	o	o	o
5	o	x	o	x	o	o	o	o	o	o	o	o	o	o
6	o	x	o	x	o	o	o	o	o	o	o	o	o	o
7	x	x	o	x	o	o	o	o	o	o	o	o	o	o

```
> eacf(diff(log(oil.price)))
```

The results of the best subsets ARMA approach are displayed in Exhibit 6.31.

Exhibit 6.31 Best Subset ARMA Model for Difference of Log(Oil)

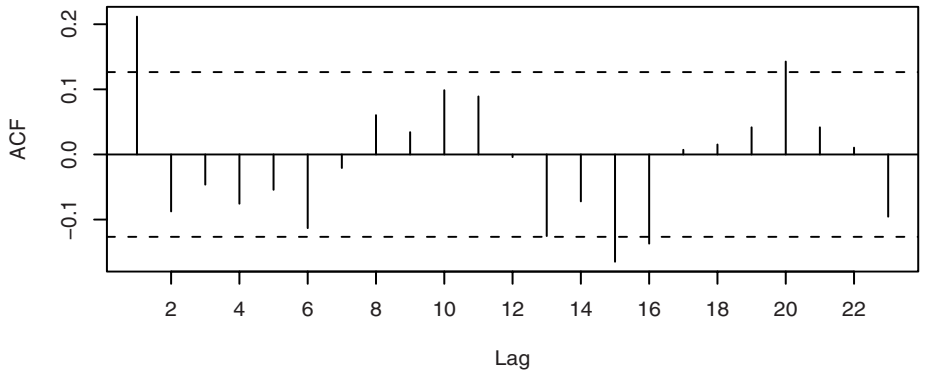



```
> res=armasubsets(y=diff(log(oil.price)),nar=7,nma=7,
    y.name='test', ar.method='ols')
> plot(res)
```

Here the suggestion is that $Y_t = \nabla \log(\text{Oil}_t)$ should be modeled in terms of Y_{t-1} and Y_{t-4} and that no lags are needed in the error terms. The second best model omits the lag 4 term so that an ARIMA(1,1,0) model on the logarithms should also be investigated further.

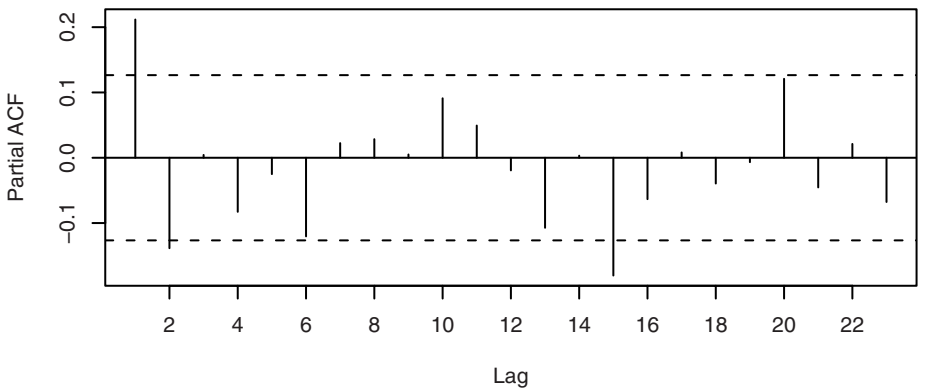
Exhibit 6.32 suggests that we specify an MA(1) model for the difference of the log oil prices, and Exhibit 6.33 says to consider an AR(2) model (ignoring some significant spikes at lags 15, 16, and 20). We will want to look at all of these models further when we estimate parameters and perform diagnostic tests in Chapters 7 and 8. (We will see later that to obtain a suitable model for the oil price series, the outliers in the series will need to be dealt with. (Can you spot the outliers in Exhibit 5.4 on page 91?)

Exhibit 6.32 Sample ACF of Difference of Logged Oil Prices




```
> acf(as.vector(diff(log(oil.price))),xaxp=c(0,22,11))
```

Exhibit 6.33 Sample PACF of Difference of Logged Oil Prices




```
> pacf(as.vector(diff(log(oil.price))),xaxp=c(0,22,11))
```

6.7 Summary

In this chapter, we considered the problem of specifying reasonable but simple models for observed times series. In particular, we investigated tools for choosing the orders (p, d, and q) for ARIMA(p,d,q) models. Three tools, the sample autocorrelation function, the sample partial autocorrelation function, and the sample extended autocorrelation function, were introduced and studied to help with this difficult task. The Dickey-Fuller unit-root test was also introduced to help distinguish between stationary and nonstationary series. These ideas were all illustrated with both simulated and actual time series.

EXERCISES

6.1 Verify Equation (6.1.3) on page 110 for the white noise process.
6.2 Verify Equation (6.1.4) on page 110 for the AR(1) process.
6.3 Verify the line in Exhibit 6.1 on page 111, for the values $\phi = \pm 0.9$.
6.4 Add new entries to Exhibit 6.1 on page 111, for the following values:
(a) $\phi = \pm 0.99$.
(b) $\phi = \pm 0.5$.
(c) $\phi = \pm 0.1$.
6.5 Verify Equation (6.1.9) on page 111 and Equation (6.1.10) for the MA(1) process.
6.6 Verify the line in Exhibit 6.2 on page 112, for the values $\theta = \pm 0.9$.
6.7 Add new entries to Exhibit 6.2 on page 112, for the following values:
(a) $\theta = \pm 0.99$.
(b) $\theta = \pm 0.8$.
(c) $\theta = \pm 0.2$.
6.8 Verify Equation (6.1.11) on page 112, for the general MA(q) process.
6.9 Use Equation (6.2.3) on page 113, to verify the value for the lag 2 partial autocorrelation function for the MA(1) process given in Equation (6.2.5) on page 114.
6.10 Show that the general expression for the partial autocorrelation function of an MA(1) process given in Equation (6.2.6) on page 114, satisfies the Yule-Walker recursion given in Equation (6.2.7).
6.11 Use Equation (6.2.8) on page 114, to find the (theoretical) partial autocorrelation function for an AR(2) model in terms of ϕ_1 and ϕ_2 and lag $k = 1, 2, 3, \ldots$.
6.12 From a time series of 100 observations, we calculate $r_1 = -0.49$, $r_2 = 0.31$, $r_3 = -0.21$, $r_4 = 0.11$, and $|r_k| < 0.09$ for $k > 4$. On this basis alone, what ARIMA model would we tentatively specify for the series?
6.13 A stationary time series of length 121 produced sample partial autocorrelation of $\hat{\phi}_{11} = 0.8$, $\hat{\phi}_{22} = -0.6$, $\hat{\phi}_{33} = 0.08$, and $\hat{\phi}_{44} = 0.00$. Based on this information alone, what model would we tentatively specify for the series?
6.14 For a series of length 169, we find that $r_1 = 0.41$, $r_2 = 0.32$, $r_3 = 0.26$, $r_4 = 0.21$, and $r_5 = 0.16$. What ARIMA model fits this pattern of autocorrelations?

6.15 The sample ACF for a series and its first difference are given in the following table. Here $n = 100$.

lag	1	2	3	4	5	6
ACF for Y_t	0.97	0.97	0.93	0.85	0.80	0.71
ACF for ∇Y_t	−0.42	0.18	−0.02	0.07	−0.10	−0.09

Based on this information alone, which ARIMA model(s) would we consider for the series?

6.16 For a series of length 64, the sample partial autocorrelations are given as:

Lag	1	2	3	4	5
PACF	0.47	−0.34	0.20	0.02	−0.06

Which models should we consider in this case?

6.17 Consider an AR(1) series of length 100 with $\phi = 0.7$.

(a) Would you be surprised if $r_1 = 0.6$?

(b) Would $r_{10} = -0.15$ be unusual?

6.18 Suppose the $\{X_t\}$ is a stationary AR(1) process with parameter ϕ but that we can only observe $Y_t = X_t + N_t$ where $\{N_t\}$ is the white noise measurement error independent of $\{X_t\}$.

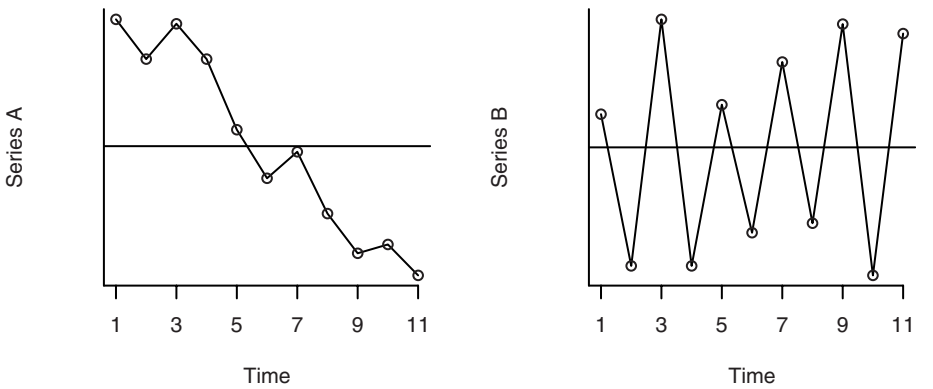
(a) Find the autocorrelation function for the observed process in terms of ϕ, σ_X^2, and σ_N^2.

(b) Which ARIMA model might we specify for $\{Y_t\}$?

6.19 The time plots of two series are shown below.

(a) For each of the series, describe r_1 using the terms strongly positive, moderately positive, near zero, moderately negative, or strongly negative. Do you need to know the scale of measurement for the series to answer this?

(b) Repeat part (a) for r_2.



6.20 Simulate an AR(1) time series with $n = 48$ and with $\phi = 0.7$.
- **(a)** Calculate the theoretical autocorrelations at lag 1 and lag 5 for this model.
- **(b)** Calculate the sample autocorrelations at lag 1 and lag 5 and compare the values with their theoretical values. Use Equations (6.1.5) and (6.1.6) page 111, to quantify the comparisons.
- **(c)** Repeat part (b) with a new simulation. Describe how the precision of the estimate varies with different samples selected under identical conditions.
- **(d)** If software permits, repeat the simulation of the series and calculation of r_1 and r_5 many times and form the sampling distributions of r_1 and r_5. Describe how the precision of the estimate varies with different samples selected under identical conditions. How well does the large-sample variance given in Equation (6.1.5) on page 111, approximate the variance in your sampling distribution?

6.21 Simulate an MA(1) time series with $n = 60$ and with $\theta = 0.5$.
- **(a)** Calculate the theoretical autocorrelation at lag 1 for this model.
- **(b)** Calculate the sample autocorrelation at lag 1, and compare the value with its theoretical value. Use Exhibit 6.2 on page 112, to quantify the comparisons.
- **(c)** Repeat part (b) with a new simulation. Describe how the precision of the estimate varies with different samples selected under identical conditions.
- **(d)** If software permits, repeat the simulation of the series and calculation of r_1 many times and form the sampling distribution of r_1. Describe how the precision of the estimate varies with different samples selected under identical conditions. How well does the large-sample variance given in Exhibit 6.2 on page 112, approximate the variance in your sampling distribution?

6.22 Simulate an AR(1) time series with $n = 48$, with
- **(a)** $\phi = 0.9$, and calculate the theoretical autocorrelations at lag 1 and lag 5;
- **(b)** $\phi = 0.6$, and calculate the theoretical autocorrelations at lag 1 and lag 5;
- **(c)** $\phi = 0.3$, and calculate the theoretical autocorrelations at lag 1 and lag 5.
- **(d)** For each of the series in parts (a), (b), and (c), calculate the sample autocorrelations at lag 1 and lag 5 and compare the values with their theoretical values. Use Equations (6.1.5) and 6.1.6, page 111, to quantify the comparisons. In general, describe how the precision of the estimate varies with the value of ϕ.

6.23 Simulate an AR(1) time series with $\phi = 0.6$, with
- **(a)** $n = 24$, and estimate $\rho_1 = \phi = 0.6$ with r_1;
- **(b)** $n = 60$, and estimate $\rho_1 = \phi = 0.6$ with r_1;
- **(c)** $n = 120$, and estimate $\rho_1 = \phi = 0.6$ with r_1.
- **(d)** For each of the series in parts (a), (b), and (c), compare the estimated values with the theoretical value. Use Equation (6.1.5) on page 111, to quantify the comparisons. In general, describe how the precision of the estimate varies with the sample size.

6.24 Simulate an MA(1) time series with $\theta = 0.7$, with
(a) $n = 24$, and estimate ρ_1 with r_1;
(b) $n = 60$, and estimate ρ_1 with r_1;
(c) $n = 120$, and estimate ρ_1 with r_1.
(d) For each of the series in parts (a), (b), and (c), compare the estimated values of ρ_1 with the theoretical value. Use Exhibit 6.2 on page 112, to quantify the comparisons. In general, describe how the precision of the estimate varies with the sample size.

6.25 Simulate an AR(1) time series of length $n = 36$ with $\phi = 0.7$.
(a) Calculate and plot the theoretical autocorrelation function for this model. Plot sufficient lags until the correlations are negligible.
(b) Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
(c) What are the theoretical partial autocorrelations for this model?
(d) Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)? Use the large-sample standard errors reported in Exhibit 6.1 on page 111, to quantify your answer.
(e) Calculate and plot the sample PACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (c)? Use the large-sample standard errors reported on page 115 to quantify your answer.

6.26 Simulate an MA(1) time series of length $n = 48$ with $\theta = 0.5$.
(a) What are the theoretical autocorrelations for this model?
(b) Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
(c) Calculate and plot the theoretical partial autocorrelation function for this model. Plot sufficient lags until the correlations are negligible. (Hint: See Equation (6.2.6) on page 114.)
(d) Calculate and plot the sample PACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (c)?

6.27 Simulate an AR(2) time series of length $n = 72$ with $\phi_1 = 0.7$ and $\phi_2 = -0.4$.
(a) Calculate and plot the theoretical autocorrelation function for this model. Plot sufficient lags until the correlations are negligible.
(b) Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
(c) What are the theoretical partial autocorrelations for this model?
(d) Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
(e) Calculate and plot the sample PACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (c)?

6.28 Simulate an MA(2) time series of length $n = 36$ with $\theta_1 = 0.7$ and $\theta_2 = -0.4$.
- **(a)** What are the theoretical autocorrelations for this model?
- **(b)** Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
- **(c)** Calculate and plot the theoretical partial autocorrelation function for this model. Plot sufficient lags until the correlations are negligible. (Hint: See Equation (6.2.6) on page 114.)
- **(d)** Calculate and plot the sample PACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (c)?

6.29 Simulate a mixed ARMA(1,1) model of length $n = 60$ with $\phi = 0.4$ and $\theta = 0.6$.
- **(a)** Calculate and plot the theoretical autocorrelation function for this model. Plot sufficient lags until the correlations are negligible.
- **(b)** Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
- **(c)** Calculate and interpret the sample EACF for this series. Does the EACF help you specify the correct orders for the model?
- **(d)** Repeat parts (b) and (c) with a new simulation using the same parameter values and sample size.
- **(e)** Repeat parts (b) and (c) with a new simulation using the same parameter values but sample size $n = 36$.
- **(f)** Repeat parts (b) and (c) with a new simulation using the same parameter values but sample size $n = 120$.

6.30 Simulate a mixed ARMA(1,1) model of length $n = 100$ with $\phi = 0.8$ and $\theta = 0.4$.
- **(a)** Calculate and plot the theoretical autocorrelation function for this model. Plot sufficient lags until the correlations are negligible.
- **(b)** Calculate and plot the sample ACF for your simulated series. How well do the values and patterns match the theoretical ACF from part (a)?
- **(c)** Calculate and interpret the sample EACF for this series. Does the EACF help you specify the correct orders for the model?
- **(d)** Repeat parts (b) and (c) with a new simulation using the same parameter values and sample size.
- **(e)** Repeat parts (b) and (c) with a new simulation using the same parameter values but sample size $n = 48$.
- **(f)** Repeat parts (b) and (c) with a new simulation using the same parameter values but sample size $n = 200$.

6.31 Simulate a nonstationary time series with $n = 60$ according to the model ARIMA(0,1,1) with $\theta = 0.8$.

(a) Perform the (augmented) Dickey-Fuller test on the series with $k = 0$ in Equation (6.4.1) on page 128. (With $k = 0$, this is the Dickey-Fuller test and is not augmented.) Comment on the results.

(b) Perform the augmented Dickey-Fuller test on the series with k chosen by the software—that is, the "best" value for k. Comment on the results.

(c) Repeat parts (a) and (b) but use the differences of the simulated series. Comment on the results. (Here, of course, you should reject the unit root hypothesis.)

6.32 Simulate a stationary time series of length $n = 36$ according to an AR(1) model with $\phi = 0.95$. This model is stationary, but just barely so. With such a series and a short history, it will be difficult if not impossible to distinguish between stationary and nonstationary with a unit root.

(a) Plot the series and calculate the sample ACF and PACF and describe what you see.

(b) Perform the (augmented) Dickey-Fuller test on the series with $k = 0$ in Equation (6.4.1) on page 128. (With $k = 0$ this is the Dickey-Fuller test and is not augmented.) Comment on the results.

(c) Perform the augmented Dickey-Fuller test on the series with k chosen by the software—that is, the "best" value for k. Comment on the results.

(d) Repeat parts (a), (b), and (c) but with a new simulation with $n = 100$.

6.33 The data file named deere1 contains 82 consecutive values for the amount of deviation (in 0.000025 inch units) from a specified target value that an industrial machining process at Deere & Co. produced under certain specified operating conditions.

(a) Display the time series plot of this series and comment on any unusual points.

(b) Calculate the sample ACF for this series and comment on the results.

(c) Now replace the unusual value by a much more typical value and recalculate the sample ACF. Comment on the change from what you saw in part (b).

(d) Calculate the sample PACF based on the revised series that you used in part (c). What model would you specify for the revised series? (Later we will investigate other ways to handle outliers in time series modeling.)

6.34 The data file named deere2 contains 102 consecutive values for the amount of deviation (in 0.0000025 inch units) from a specified target value that another industrial machining process produced at Deere & Co.

(a) Display the time series plot of this series and comment on its appearance. Would a stationary model seem to be appropriate?

(b) Display the sample ACF and PACF for this series and select tentative orders for an ARMA model for the series.

6.35 The data file named deere3 contains 57 consecutive measurements recorded from a complex machine tool at Deere & Co. The values given are deviations from a target value in units of ten millionths of an inch. The process employs a control mechanism that resets some of the parameters of the machine tool depending on the magnitude of deviation from target of the last item produced.
(a) Display the time series plot of this series and comment on its appearance. Would a stationary model be appropriate here?
(b) Display the sample ACF and PACF for this series and select tentative orders for an ARMA model for the series.

6.36 The data file named robot contains a time series obtained from an industrial robot. The robot was put through a sequence of maneuvers, and the distance from a desired ending point was recorded in inches. This was repeated 324 times to form the time series.
(a) Display the time series plot of the data. Based on this information, do these data appear to come from a stationary or nonstationary process?
(b) Calculate and plot the sample ACF and PACF for these data. Based on this additional information, do these data appear to come from a stationary or nonstationary process?
(c) Calculate and interpret the sample EACF.
(d) Use the best subsets ARMA approach to specify a model for these data. Compare these results with what you discovered in parts (a), (b), and (c).

6.37 Calculate and interpret the sample EACF for the logarithms of the Los Angeles rainfall series. The data are in the file named larain. Do the results confirm that the logs are white noise?

6.38 Calculate and interpret the sample EACF for the color property time series. The data are in the color file. Does the sample EACF suggest the same model that was specified by looking at the sample PACF?

6.39 The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.)
(a) Plot the time series, and comment on the display. Are there any unusual values?
(b) Calculate the sample ACF and PACF for this series.
(c) Now replace each of the unusual values with a value of 35 days—much more typical values—and repeat the calculation of the sample ACF and PACF. What ARMA model would you specify for this series after removing the outliers? (Later we will investigate other ways to handle outliers in time series modeling.)

CHAPTER 7

PARAMETER ESTIMATION

This chapter deals with the problem of estimating the parameters of an ARIMA model based on the observed time series $Y_1, Y_2, \ldots, Y_n$. We assume that a model has already been specified; that is, we have specified values for p, d, and q using the methods of Chapter 6. With regard to nonstationarity, since the dth difference of the observed series is assumed to be a stationary ARMA(p,q) process, we need only concern ourselves with the problem of estimating the parameters in such stationary models. In practice, then we treat the dth difference of the original time series as the time series from which we estimate the parameters of the complete model. For simplicity, we shall let $Y_1, Y_2, \ldots, Y_n$ denote our observed *stationary* process even though it may be an appropriate difference of the original series. We first discuss the method-of-moments estimators, then the least squares estimators, and finally full maximum likelihood estimators.

7.1 The Method of Moments

The method of moments is frequently one of the easiest, if not the most efficient, methods for obtaining parameter estimates. The method consists of equating sample moments to corresponding theoretical moments and solving the resulting equations to obtain estimates of any unknown parameters. The simplest example of the method is to estimate a stationary process mean by a sample mean. The properties of this estimator were studied extensively in Chapter 3.

Autoregressive Models

Consider first the AR(1) case. For this process, we have the simple relationship $\rho_1 = \phi$. In the method of moments, ρ_1 is equated to r_1, the lag 1 sample autocorrelation. Thus we can estimate ϕ by

$$\hat{\phi} = r_1 \tag{7.1.1}$$

Now consider the AR(2) case. The relationships between the parameters ϕ_1 and ϕ_2 and various moments are given by the Yule-Walker equations (4.3.13) on page 72:

$$\rho_1 = \phi_1 + \rho_1\phi_2 \quad \text{and} \quad \rho_2 = \rho_1\phi_1 + \phi_2$$

The method of moments replaces ρ_1 by r_1 and ρ_2 by r_2 to obtain

$$r_1 = \phi_1 + r_1\phi_2 \quad \text{and} \quad r_2 = r_1\phi_1 + \phi_2$$

which are then solved to obtain

$$\hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2} \quad \text{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2} \tag{7.1.2}$$

The general AR(p) case proceeds similarly. Replace ρ_k by r_k throughout the Yule-Walker equations on page 79 (or page 114) to obtain

$$\left.\begin{aligned} \phi_1 + r_1\phi_2 + r_2\phi_3 + \cdots + r_{p-1}\phi_p &= r_1 \\ r_1\phi_1 + \phi_2 + r_1\phi_3 + \cdots + r_{p-2}\phi_p &= r_2 \\ &\vdots \\ r_{p-1}\phi_1 + r_{p-2}\phi_2 + r_{p-3}\phi_3 + \cdots + \phi_p &= r_p \end{aligned}\right\} \tag{7.1.3}$$

These linear equations are then solved for $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$. The Durbin-Levinson recursion of Equation (6.2.9) on page 115 provides a convenient method of solution but is subject to substantial round-off errors if the solution is close to the boundary of the stationarity region. The estimates obtained in this way are also called **Yule-Walker estimates**.

Moving Average Models

Surprisingly, the method of moments is not nearly as convenient when applied to moving average models. Consider the simple MA(1) case. From Equations (4.2.2) on page 57, we know that

$$\rho_1 = -\frac{\theta}{1+\theta^2}$$

Equating ρ_1 to r_1, we are led to solve a quadratic equation in θ. If $|r_1| < 0.5$, then the two real roots are given by

$$-\frac{1}{2r_1} \pm \sqrt{\frac{1}{4r_1^2} - 1}$$

As can be easily checked, the product of the two solutions is always equal to 1; therefore, only one of the solutions satisfies the invertibility condition $|\theta| < 1$.

After further algebraic manipulation, we see that the invertible solution can be written as

$$\hat{\theta} = \frac{-1 + \sqrt{1-4r_1^2}}{2r_1} \tag{7.1.4}$$

If $r_1 = \pm 0.5$, unique, real solutions exist, namely ∓ 1, but neither is invertible. If $|r_1| > 0.5$ (which is certainly possible even though $|\rho_1| < 0.5$), no real solutions exist, and so the method of moments fails to yield an estimator of θ. Of course, if $|r_1| > 0.5$, the specification of an MA(1) model would be in considerable doubt.

For higher-order MA models, the method of moments quickly gets complicated. We can use Equations (4.2.5) on page 65 and replace ρ_k by r_k for $k = 1, 2, \ldots, q$, to obtain q equations in q unknowns $\theta_1, \theta_2, \ldots, \theta_q$. The resulting equations are highly nonlinear in the θ's, however, and their solution would of necessity be numerical. In addition, there will be multiple solutions, of which only one is invertible. We shall not pursue this further since we shall see in Section 7.4 that, for MA models, the method of moments generally produces poor estimates.

Mixed Models

We consider only the ARMA(1,1) case. Recall Equation (4.4.5) on page 78,

$$\rho_k = \frac{(1 - \theta\phi)(\phi - \theta)}{1 - 2\theta\phi + \theta^2}\phi^{k-1} \quad \text{for } k \geq 1$$

Noting that $\rho_2 / \rho_1 = \phi$, we can first estimate ϕ as

$$\hat{\phi} = \frac{r_2}{r_1} \tag{7.1.5}$$

Having done so, we can then use

$$r_1 = \frac{(1 - \theta\hat{\phi})(\hat{\phi} - \theta)}{1 - 2\theta\hat{\phi} + \theta^2} \tag{7.1.6}$$

to solve for $\hat{\theta}$. Note again that a quadratic equation must be solved and only the invertible solution, if any, retained.

Estimates of the Noise Variance

The final parameter to be estimated is the noise variance, σ_e^2. In all cases, we can first estimate the process variance, $\gamma_0 = Var(Y_t)$, by the sample variance

$$s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(Y_t - \bar{Y})^2 \tag{7.1.7}$$

and use known relationships from Chapter 4 among γ_0, σ_e^2, and the θ's and ϕ's to estimate σ_e^2.

For the AR(p) models, Equation (4.3.31) on page 77 yields

$$\hat{\sigma}_e^2 = (1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2 - \cdots - \hat{\phi}_p r_p)s^2 \tag{7.1.8}$$

In particular, for an AR(1) process,

$$\hat{\sigma}_e^2 = (1 - r_1^2)s^2$$

since $\hat{\phi} = r_1$.

For the MA(q) case, we have, using Equation (4.2.4) on page 65,

$$\hat{\sigma}_e^2 = \frac{s^2}{1 + \hat{\theta}_1^2 + \hat{\theta}_2^2 + \cdots + \hat{\theta}_q^2} \tag{7.1.9}$$

For the ARMA(1,1) process, Equation (4.4.4) on page 78 yields

$$\hat{\sigma}_e^2 = \frac{1 - \hat{\phi}^2}{1 - 2\hat{\phi}\hat{\theta} + \hat{\theta}^2} s^2 \tag{7.1.10}$$

Numerical Examples

The table in Exhibit 7.1 displays method-of-moments estimates for the parameters from several simulated time series. Generally speaking, the estimates for all the autoregressive models are fairly good but the estimates for the moving average models are not acceptable. It can be shown that theory confirms this observation—method-of-moments estimators are very inefficient for models containing moving average terms.

Exhibit 7.1 Method-of-Moments Parameter Estimates for Simulated Series

	True Parameters			Method-of-Moments Estimates			
Model	θ	ϕ_1	ϕ_2	θ	ϕ_1	ϕ_2	***n***
MA(1)	−0.9			−0.554			120
MA(1)	0.9			0.719			120
MA(1)	−0.9			NA†			60
MA(1)	0.5			−0.314			60
AR(1)		0.9			0.831		60
AR(1)		0.4			0.470		60
AR(2)		1.5	−0.75		1.472	−0.767	120

† No method-of-moments estimate exists since $r_1 = 0.544$ for this simulation.

```
> data(ma1.2.s); data(ma1.1.s); data(ma1.3.s); data(ma1.4.s)
> estimate.ma1.mom(ma1.2.s); estimate.ma1.mom(ma1.1.s)
> estimate.ma1.mom(ma1.3.s); estimate.ma1.mom(ma1.4.s)
> arima(ma1.4.s,order=c(0,0,1),method='CSS',include.mean=F)
> data(ar1.s); data(ar1.2.s)
> ar(ar1.s,order.max=1,AIC=F,method='yw')
> ar(ar1.2.s,order.max=1,AIC=F,method='yw')
> data(ar2.s)
> ar(ar2.s,order.max=2,AIC=F,method='yw')
```

Consider now some actual time series. We start with the Canadian hare abundance series. Since we found in Exhibit 6.27 on page 136 that a square root transformation was appropriate here, we base all modeling on the square root of the original abundance numbers. We illustrate the estimation of an AR(2) model with the hare data, even

though we shall show later that an AR(3) model provides a better fit to the data. The first two sample autocorrelations displayed in Exhibit 6.28 on page 137 are $r_1 = 0.736$ and $r_2 = 0.304$. Using Equations (7.1.2), the method-of-moments estimates of ϕ_1 and ϕ_2 are

$$\hat{\phi}_1 = \frac{r_1(1 - r_2)}{1 - r_1^2} = \frac{0.736(1 - 0.304)}{1 - (0.736)^2} = 1.1178 \tag{7.1.11}$$

and

$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2} = \frac{0.304 - (0.736)^2}{1 - (0.736)^2} = -0.519 \tag{7.1.12}$$

The sample mean and variance of this series (after taking the square root) are found to be 5.82 and 5.88, respectively. Then, using Equation (7.1.8), we estimate the noise variance as

$$\begin{aligned} \hat{\sigma}_e^2 &= (1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2)s^2 \\ &= [1 - (1.1178)(0.736) - (-0.519)(0.304)](5.88) \\ &= 1.97 \end{aligned} \tag{7.1.13}$$

The estimated model (in original terms) is then

$$\sqrt{Y_t} - 5.82 = 1.1178(\sqrt{Y_{t-1}} - 5.82) - 0.519(\sqrt{Y_{t-2}} - 5.82) + e_t \tag{7.1.14}$$

or

$$\sqrt{Y_t} = 2.335 + 1.1178\sqrt{Y_{t-1}} - 0.519\sqrt{Y_{t-2}} + e_t \tag{7.1.15}$$

with estimated noise variance of 1.97.

Consider now the oil price series. Exhibit 6.32 on page 140 suggested that we specify an MA(1) model for the first differences of the logarithms of the series. The lag 1 sample autocorrelation in that exhibit is 0.212, so the method-of-moments estimate of θ is

$$\hat{\theta} = \frac{-1 + \sqrt{1 - 4(0.212)^2}}{2(0.212)} = -0.222 \tag{7.1.16}$$

The mean of the differences of the logs is 0.004 and the variance is 0.0072. The estimated model is

$$\nabla \log(Y_t) = 0.004 + e_t + 0.222e_{t-1} \tag{7.1.17}$$

or

$$\log(Y_t) = \log(Y_{t-1}) + 0.004 + e_t + 0.222e_{t-1} \tag{7.1.18}$$

with estimated noise variance of

$$\hat{\sigma}_e^2 = \frac{s^2}{1 + \hat{\theta}^2} = \frac{0.0072}{1 + (-0.222)^2} = 0.00686 \tag{7.1.19}$$

Using Equation (3.2.3) on page 28 with estimated parameters yields a standard error of the sample mean of 0.0060. Thus, the observed sample mean of 0.004 is not significantly different from zero and we would remove the constant term from the model, giving a final model of

$$\log(Y_t) = \log(Y_{t-1}) + e_t + 0.222e_{t-1} \tag{7.1.20}$$

7.2 Least Squares Estimation

Because the method of moments is unsatisfactory for many models, we must consider other methods of estimation. We begin with least squares. For autoregressive models, the ideas are quite straightforward. At this point, we introduce a possibly nonzero mean, μ, into our stationary models and treat it as another parameter to be estimated by least squares.

Autoregressive Models

Consider the first-order case where

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t \tag{7.2.1}$$

We can view this as a regression model with predictor variable Y_{t-1} and response variable Y_t. Least squares estimation then proceeds by minimizing the sum of squares of the differences

$$(Y_t - \mu) - \phi(Y_{t-1} - \mu)$$

Since only $Y_1, Y_2, \ldots, Y_n$ are observed, we can only sum from $t = 2$ to $t = n$. Let

$$S_c(\phi, \mu) = \sum_{t=2}^{n} [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 \tag{7.2.2}$$

This is usually called the **conditional sum-of-squares function**. (The reason for the term *conditional* will become apparent later on.) According to the principle of least squares, we estimate ϕ and μ by the respective values that minimize $S_c(\phi, \mu)$ given the observed values of $Y_1, Y_2, \ldots, Y_n$.

Consider the equation $\partial S_c / \partial \mu = 0$. We have

$$\frac{\partial S_c}{\partial \mu} = \sum_{t=2}^{n} 2[(Y_t - \mu) - \phi(Y_{t-1} - \mu)](-1 + \phi) = 0$$

or, simplifying and solving for μ,

$$\mu = \frac{1}{(n-1)(1-\phi)} \left[\sum_{t=2}^{n} Y_t - \phi \sum_{t=2}^{n} Y_{t-1} \right] \tag{7.2.3}$$

Now, for large n,

$$\frac{1}{n-1}\sum_{t=2}^{n} Y_t \approx \frac{1}{n-1}\sum_{t=2}^{n} Y_{t-1} \approx \bar{Y}$$

Thus, regardless of the value of ϕ, Equation (7.2.3) reduces to

$$\hat{\mu} \approx \frac{1}{1-\phi}(\bar{Y} - \phi\bar{Y}) = \bar{Y} \tag{7.2.4}$$

We sometimes say, except for end effects, $\hat{\mu} = \bar{Y}$.

Consider now the minimization of $S_c(\phi, \bar{Y})$ with respect to ϕ. We have

$$\frac{\partial S_c(\phi, \bar{Y})}{\partial \phi} = \sum_{t=2}^{n} 2[(Y_t - \bar{Y}) - \phi(Y_{t-1} - \bar{Y})](Y_{t-1} - \bar{Y})$$

Setting this equal to zero and solving for ϕ yields

$$\hat{\phi} = \frac{\sum_{t=2}^{n} (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=2}^{n} (Y_{t-1} - \bar{Y})^2}$$

Except for one term missing in the denominator, namely $(Y_n - \bar{Y})^2$, this is the same as r_1. The lone missing term is negligible for stationary processes, and thus the least squares and method-of-moments estimators are nearly identical, especially for large samples.

For the general AR(p) process, the methods used to obtain Equations (7.2.3) and (7.2.4) can easily be extended to yield the same result, namely

$$\hat{\mu} = \bar{Y} \tag{7.2.5}$$

To generalize the estimation of the ϕ's, we consider the second-order model. In accordance with Equation (7.2.5), we replace μ by $\bar{Y}$ in the conditional sum-of-squares function, so

$$S_c(\phi_1, \phi_2, \bar{Y}) = \sum_{t=3}^{n} [(Y_t - \bar{Y}) - \phi_1(Y_{t-1} - \bar{Y}) - \phi_2(Y_{t-2} - \bar{Y})]^2 \tag{7.2.6}$$

Setting $\partial S_c / \partial \phi_1 = 0$, we have

$$-2\sum_{t=3}^{n} [(Y_t - \bar{Y}) - \phi_1(Y_{t-1} - \bar{Y}) - \phi_2(Y_{t-2} - \bar{Y})](Y_{t-1} - \bar{Y}) = 0 \tag{7.2.7}$$

which we can rewrite as

$$\sum_{t=3}^{n}(Y_t-\bar{Y})(Y_{t-1}-\bar{Y}) = \left(\sum_{t=3}^{n}(Y_{t-1}-\bar{Y})^2\right)\phi_1 + \left(\sum_{t=3}^{n}(Y_{t-1}-\bar{Y})(Y_{t-2}-\bar{Y})\right)\phi_2 \tag{7.2.8}$$

The sum of the lagged products $\sum_{t=3}^{n}(Y_t-\bar{Y})(Y_{t-1}-\bar{Y})$ is very nearly the numerator of r_1—we are missing one product, $(Y_2-\bar{Y})(Y_1-\bar{Y})$. A similar situation exists for $\sum_{t=3}^{n}(Y_{t-1}-\bar{Y})(Y_{t-2}-\bar{Y})$, but here we are missing $(Y_n-\bar{Y})(Y_{n-1}-\bar{Y})$. If we divide both sides of Equation (7.2.8) by $\sum_{t=3}^{n}(Y_t-\bar{Y})^2$, then, except for end effects, which are negligible under the stationarity assumption, we obtain

$$r_1 = \phi_1 + r_1\phi_2 \tag{7.2.9}$$

Approximating in a similar way with the equation $\partial S_c/\partial\phi_2 = 0$ leads to

$$r_2 = r_1\phi_1 + \phi_2 \tag{7.2.10}$$

But Equations (7.2.9) and (7.2.10) are just the sample Yule-Walker equations for an AR(2) model.

Entirely analogous results follow for the general stationary AR(p) case: To an excellent approximation, the conditional least squares estimates of the ϕ's are obtained by solving the sample Yule-Walker equations (7.1.3).[†]

Moving Average Models

Consider now the least-squares estimation of θ in the MA(1) model:

$$Y_t = e_t - \theta e_{t-1} \tag{7.2.11}$$

At first glance, it is not apparent how a least squares or regression method can be applied to such models. However, recall from Equation (4.4.2) on page 77 that invertible MA(1) models can be expressed as

$$Y_t = -\theta Y_{t-1} - \theta^2 Y_{t-2} - \theta^3 Y_{t-3} - \cdots + e_t$$

an autoregressive model but of infinite order. Thus least squares can be meaningfully carried out by choosing a value of θ that minimizes

[†] We note that Lai and Wei (1983) established that the conditional least squares estimators are consistent even for nonstationary autoregressive models where the Yule-Walker equations do not apply.

$$S_c(\theta) = \sum (e_t)^2 = \sum [Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \cdots]^2 \tag{7.2.12}$$

where, implicitly, $e_t = e_t(\theta)$ is a function of the observed series and the unknown parameter θ.

It is clear from Equation (7.2.12) that the least squares problem is *nonlinear* in the parameters. We will not be able to minimize $S_c(\theta)$ by taking a derivative with respect to θ, setting it to zero, and solving. Thus, even for the simple MA(1) model, we must resort to techniques of numerical optimization. Other problems exist in this case: We have not shown explicit limits on the summation in Equation (7.2.12) nor have we said how to deal with the infinite series under the summation sign.

To address these issues, consider evaluating $S_c(\theta)$ for a *single given value* of θ. The only Y's we have available are our observed series, $Y_1, Y_2, \ldots, Y_n$. Rewrite Equation (7.2.11) as

$$e_t = Y_t + \theta e_{t-1} \tag{7.2.13}$$

Using this equation, $e_1, e_2, \ldots, e_n$ can be calculated recursively if we have the initial value e_0. A common approximation is to set $e_0 = 0$—its expected value. Then, *conditional on* $e_0 = 0$, we can obtain

$$\left.\begin{aligned} e_1 &= Y_1 \\ e_2 &= Y_2 + \theta e_1 \\ e_3 &= Y_3 + \theta e_2 \\ &\vdots \\ e_n &= Y_n + \theta e_{n-1} \end{aligned}\right\} \tag{7.2.14}$$

and thus calculate $S_c(\theta) = \sum (e_t)^2$, conditional on $e_0 = 0$, for that single given value of θ.

For the simple case of one parameter, we could carry out a grid search over the invertible range $(-1,+1)$ for θ to find the minimum sum of squares. For more general MA(q) models, a numerical optimization algorithm, such as Gauss-Newton or Nelder-Mead, will be needed.

For higher-order moving average models, the ideas are analogous and no new difficulties arise. We compute $e_t = e_t(\theta_1, \theta_2, \ldots, \theta_q)$ recursively from

$$e_t = Y_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{7.2.15}$$

with $e_0 = e_{-1} = \cdots = e_{-q} = 0$. The sum of squares is minimized jointly in $\theta_1, \theta_2, \ldots, \theta_q$ using a multivariate numerical method.

Mixed Models

Consider the ARMA(1,1) case

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \tag{7.2.16}$$

As in the pure MA case, we consider $e_t = e_t(\phi,\theta)$ and wish to minimize $S_c(\phi, \theta) = \sum e_t^2$. We can rewrite Equation (7.2.16) as

$$e_t = Y_t - \phi Y_{t-1} + \theta e_{t-1} \tag{7.2.17}$$

To obtain e_1, we now have an additional "startup" problem, namely Y_0. One approach is to set $Y_0 = 0$ or to $\bar{Y}$ if our model contains a nonzero mean. However, a better approach is to begin the recursion at $t = 2$, thus avoiding Y_0 altogether, and simply minimize

$$S_c(\phi, \theta) = \sum_{t=2}^{n} e_t^2$$

For the general ARMA(p,q) model, we compute

$$\begin{aligned} e_t = {} & Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} \\ & + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \end{aligned} \tag{7.2.18}$$

with $e_p = e_{p-1} = \cdots = e_{p+1-q} = 0$ and then minimize $S_c(\phi_1,\phi_2,\ldots,\phi_p,\theta_1,\theta_2,\ldots,\theta_q)$ numerically to obtain the conditional least squares estimates of all the parameters.

For parameter sets $\theta_1, \theta_2,\ldots, \theta_q$ corresponding to invertible models, the start-up values $e_p, e_{p-1},\ldots, e_{p+1-q}$ will have very little influence on the final estimates of the parameters for large samples.

7.3 Maximum Likelihood and Unconditional Least Squares

For series of moderate length and also for stochastic seasonal models to be discussed in Chapter 10, the start-up values $e_p = e_{p-1} = \cdots = e_{p+1-q} = 0$ will have a more pronounced effect on the final estimates for the parameters. Thus we are led to consider the more difficult problem of maximum likelihood estimation.

The advantage of the method of maximum likelihood is that all of the information in the data is used rather than just the first and second moments, as is the case with least squares. Another advantage is that many large-sample results are known under very general conditions. One disadvantage is that we must for the first time work specifically with the joint probability density function of the process.

Maximum Likelihood Estimation

For any set of observations, $Y_1, Y_2,\ldots, Y_n$, time series or not, the likelihood function L is defined to be the joint probability density of obtaining the data actually observed. However, it is considered as a function of the unknown parameters in the model with the observed data held fixed. For ARIMA models, L will be a function of the ϕ's, θ's, μ, and σ_e^2 given the observations $Y_1, Y_2,\ldots, Y_n$. The maximum likelihood estimators are then defined as those values of the parameters for which the data actually observed are *most likely*, that is, the values that maximize the likelihood function.

We begin by looking in detail at the AR(1) model. The most common assumption is that the white noise terms are independent, normally distributed random variables with

zero means and common standard deviation σ_e. The probability density function (pdf) of each e_t is then

$$(2\pi\sigma_e^2)^{-1/2}\exp\left(-\frac{e_t^2}{2\sigma_e^2}\right) \quad \text{for} \quad -\infty < e_t < \infty$$

and, by independence, the joint pdf for $e_2, e_3, \ldots, e_n$ is

$$(2\pi\sigma_e^2)^{-(n-1)/2}\exp\left(-\frac{1}{2\sigma_e^2}\sum_{t=2}^{n} e_t^2\right) \tag{7.3.1}$$

Now consider

$$\left.\begin{aligned} Y_2 - \mu &= \phi(Y_1 - \mu) + e_2 \\ Y_3 - \mu &= \phi(Y_2 - \mu) + e_3 \\ &\vdots \\ Y_n - \mu &= \phi(Y_{n-1} - \mu) + e_n \end{aligned}\right\} \tag{7.3.2}$$

If we condition on $Y_1 = y_1$, Equation (7.3.2) defines a linear transformation between e_2, $e_3, \ldots, e_n$ and $Y_2, Y_3, \ldots, Y_n$ (with Jacobian equal to 1). Thus the joint pdf of $Y_2, Y_3, \ldots, Y_n$ given $Y_1 = y_1$ can be obtained by using Equation (7.3.2) to substitute for the e's in terms of the Y's in Equation (7.3.1). Thus we get

$$\begin{aligned} f(y_2, y_3, \ldots, y_n|y_1) &= (2\pi\sigma_e^2)^{-(n-1)/2} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{t=2}^{n}[(y_t - \mu) - \phi(y_{t-1} - \mu)]^2\right\} \end{aligned} \tag{7.3.3}$$

Now consider the (marginal) distribution of Y_1. It follows from the linear process representation of the AR(1) process (Equation (4.3.8) on page 70) that Y_1 will have a normal distribution with mean μ and variance $\sigma_e^2/(1-\phi^2)$. Multiplying the conditional pdf in Equation (7.3.3) by the marginal pdf of Y_1 gives us the joint pdf of $Y_1, Y_2, \ldots, Y_n$ that we require. Interpreted as a function of the parameters ϕ, μ, and σ_e^2, the likelihood function for an AR(1) model is given by

$$L(\phi, \mu, \sigma_e^2) = (2\pi\sigma_e^2)^{-n/2}(1-\phi^2)^{1/2}\exp\left[-\frac{1}{2\sigma_e^2}S(\phi, \mu)\right] \tag{7.3.4}$$

where

$$S(\phi, \mu) = \sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 + (1-\phi^2)(Y_1 - \mu) \tag{7.3.5}$$

The function $S(\phi, \mu)$ is called the **unconditional sum-of-squares function**.

As a general rule, the logarithm of the likelihood function is more convenient to

work with than the likelihood itself. For the AR(1) case, the **log-likelihood function**, denoted $\ell(\phi, \mu, \sigma_e^2)$, is given by

$$\ell(\phi, \mu, \sigma_e^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma_e^2}S(\phi, \mu) \tag{7.3.6}$$

For given values of ϕ and μ, $\ell(\phi, \mu, \sigma_e^2)$ can be maximized analytically with respect to σ_e^2 in terms of the yet-to-be-determined estimators of ϕ and μ. We obtain

$$\hat{\sigma}_e^2 = \frac{S(\hat{\phi}, \hat{\mu})}{n} \tag{7.3.7}$$

As in many other similar contexts, we usually divide by $n - 2$ rather than n (since we are estimating *two* parameters, ϕ and μ) to obtain an estimator with less bias. For typical time series sample sizes, there will be very little difference.

Consider now the estimation of ϕ and μ. A comparison of the unconditional sum-of-squares function $S(\phi,\mu)$ with the earlier conditional sum-of-squares function $S_c(\phi,\mu)$ of Equation (7.2.2) on page 154, reveals one simple difference:

$$S(\phi, \mu) = S_c(\phi, \mu) + (1-\phi^2)(Y_1 - \mu)^2 \tag{7.3.8}$$

Since $S_c(\phi,\mu)$ involves a sum of $n - 1$ components, whereas $(1-\phi^2)(Y_1 - \mu)^2$ does not involve n, we shall have $S(\phi, \mu) \approx S_c(\phi, \mu)$. Thus the values of ϕ and μ that minimize $S(\phi,\mu)$ or $S_c(\phi,\mu)$ should be very similar, at least for larger sample sizes. The effect of the rightmost term in Equation (7.3.8) will be more substantial when the minimum for ϕ occurs near the stationarity boundary of ± 1.

Unconditional Least Squares

As a compromise between conditional least squares estimates and full maximum likelihood estimates, we might consider obtaining unconditional least squares estimates; that is, estimates minimizing $S(\phi,\mu)$. Unfortunately, the term $(1-\phi^2)(Y_1 - \mu)^2$ causes the equations $\partial S/\partial\phi = 0$ and $\partial S/\partial\mu = 0$ to be nonlinear in ϕ and μ, and reparameterization to a constant term $\theta_0 = \mu(1-\phi)$ does not improve the situation substantially. Thus minimization must be carried out numerically. The resulting estimates are called **unconditional least squares estimates**.

The derivation of the likelihood function for more general ARMA models is considerably more involved. One derivation may be found in Appendix H: State Space Models on page 222. We refer the reader to Brockwell and Davis (1991) or Shumway and Stoffer (2006) for even more details.

7.4 Properties of the Estimates

The large-sample properties of the maximum likelihood and least squares (conditional or unconditional) estimators are identical and can be obtained by modifying standard maximum likelihood theory. Details can be found in Shumway and Stoffer (2006, pp. 125–129). We shall look at the results and their implications for simple ARMA models.

For large n, the estimators are approximately unbiased and normally distributed. The variances and correlations are as follows:

$$\text{AR(1): } Var(\hat{\phi}) \approx \frac{1-\phi^2}{n} \tag{7.4.9}$$

$$\text{AR(2): } \begin{cases} Var(\hat{\phi}_1) \approx Var(\hat{\phi}_2) \approx \dfrac{1-\phi_2^2}{n} \\ Corr(\hat{\phi}_1, \hat{\phi}_2) \approx -\dfrac{\phi_1}{1-\phi_2} = -\rho_1 \end{cases} \tag{7.4.10}$$

$$\text{MA(1): } Var(\hat{\theta}) \approx \frac{1-\theta^2}{n} \tag{7.4.11}$$

$$\text{MA(2): } \begin{cases} Var(\hat{\theta}_1) \approx Var(\hat{\theta}_2) \approx \dfrac{1-\theta_2^2}{n} \\ Corr(\hat{\theta}_1, \hat{\theta}_2) \approx -\dfrac{\theta_1}{1-\theta_2} \end{cases} \tag{7.4.12}$$

$$\text{ARMA(1,1): } \begin{cases} Var(\hat{\phi}) \approx \left[\dfrac{1-\phi^2}{n}\right]\left[\dfrac{1-\phi\theta}{\phi-\theta}\right]^2 \\ Var(\hat{\theta}) \approx \left[\dfrac{1-\theta^2}{n}\right]\left[\dfrac{1-\phi\theta}{\phi-\theta}\right]^2 \\ Corr(\hat{\phi}, \hat{\theta}) \approx \dfrac{\sqrt{(1-\phi^2)(1-\theta^2)}}{1-\phi\theta} \end{cases} \tag{7.4.13}$$

Notice that, in the AR(1) case, the variance of the estimator of ϕ decreases as ϕ approaches ± 1. Also notice that even though an AR(1) model is a special case of an AR(2) model, the variance of $\hat{\phi}_1$ shown in Equations (7.4.10) shows that our estimation of ϕ_1 will generally suffer if we erroneously fit an AR(2) model when, in fact, $\phi_2 = 0$. Similar comments could be made about fitting an MA(2) model when an MA(1) would suffice or fitting an ARMA(1,1) when an AR(1) or an MA(1) is adequate.

For the ARMA(1,1) case, note the denominator of $\phi - \theta$ in the variances in Equations (7.4.13). If ϕ and θ are nearly equal, the variability in the estimators of ϕ and θ can be extremely large.

Note that in all of the two-parameter models, the estimates can be highly correlated, even for very large sample sizes.

The table shown in Exhibit 7.2 gives numerical values for the large-sample approximate standard deviations of the estimates of ϕ in an AR(1) model for several values of ϕ and several sample sizes. Since the values in the table are equal to $\sqrt{(1-\phi^2)/n}$, they apply equally well to standard deviations computed according to Equations (7.4.10),

(7.4.11), and (7.4.12).

Thus, in estimating an AR(1) model with, for example, $n = 100$ and $\phi = 0.7$, we can be about 95% confident that our estimate of ϕ is in error by no more than $\pm 2(0.07) = \pm 0.14$.

Exhibit 7.2 AR(1) Model Large-Sample Standard Deviations of $\hat{\phi}$

	n		
ϕ	**50**	**100**	**200**
0.4	0.13	0.09	0.06
0.7	0.10	0.07	0.05
0.9	0.06	0.04	0.03

For stationary autoregressive models, the method of moments yields estimators equivalent to least squares and maximum likelihood, at least for large samples. For models containing moving average terms, such is not the case. For an MA(1) model, it can be shown that the large-sample variance of the method-of-moments estimator of θ is equal to

$$Var(\hat{\theta}) \approx \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{n(1 - \theta^2)^2} \tag{7.4.14}$$

Comparing Equation (7.4.14) with that of Equation (7.4.11), we see that the variance for the method-of-moments estimator is always larger than the variance of the maximum likelihood estimator. The table in Exhibit 7.3 displays the ratio of the large-sample standard deviations for the two methods for several values of θ. For example, if θ is 0.5, the method-of-moments estimator has a large-sample standard deviation that is 42% larger than the standard deviation of the estimator obtained using maximum likelihood. It is clear from these ratios that the method-of-moments estimator should not be used for the MA(1) model. This same advice applies to all models that contain moving average terms.

Exhibit 7.3 Method of Moments (MM) vs. Maximum Likelihood (MLE) in MA(1) Models

θ	SD_{MM}/SD_{MLE}
0.25	1.07
0.50	1.42
0.75	2.66
0.90	5.33

7.5 Illustrations of Parameter Estimation

Consider the simulated MA(1) series with $\theta = -0.9$. The series was displayed in Exhibit 4.2 on page 59, and we found the method-of-moments estimate of θ to be a rather poor –0.554; see Exhibit 7.1 on page 152. In contrast, the maximum likelihood estimate is –0.915, the unconditional sum-of-squares estimate is –0.923, and the conditional least squares estimate is –0.879. For this series, the maximum likelihood estimate of –0.915 is closest to the true value used in the simulation. Using Equation (7.4.11) on page 161 and replacing θ by its estimate, we have a standard error of about

$$\sqrt{\hat{Var}(\hat{\theta})} \approx \sqrt{\frac{1-\hat{\theta}^2}{n}} = \sqrt{\frac{1-(0.91)^2}{120}} \approx 0.04$$

so none of the maximum likelihood, conditional sum-of-squares, or unconditional sum-of-squares estimates are significantly far from the true value of –0.9.

The second MA(1) simulation with $\theta = 0.9$ produced the method-of-moments estimate of 0.719 shown in Exhibit 7.1. The conditional sum-of-squares estimate is 0.958, the unconditional sum-of-squares estimate is 0.983, and the maximum likelihood estimate is 1.000. These all have a standard error of about 0.04 as above. Here the maximum likelihood estimate of $\hat{\theta} = 1$ is a little disconcerting since it corresponds to a noninvertible model.

The third MA(1) simulation with $\theta = -0.9$ produced a method-of-moments estimate of –0.719 (see Exhibit 7.1). The maximum likelihood estimate here is –0.894 with a standard error of about

$$\sqrt{\hat{Var}(\hat{\theta})} \approx \sqrt{\frac{1-(0.894)^2}{60}} \approx 0.06$$

For these data, the conditional sum-of-squares estimate is –0.979 and the unconditional sum-of-squares estimate is –0.961. Of course, with a standard error of this magnitude, it is unwise to report digits in the estimates of θ beyond the tenths place.

For our simulated autoregressive models, the results are reported in Exhibits 7.4 and 7.5.

Exhibit 7.4 Parameter Estimation for Simulated AR(1) Models

Parameter ϕ	Method-of-Moments Estimate	Conditional SS Estimate	Unconditional SS Estimate	Maximum Likelihood Estimate	*n*
0.9	0.831	0.857	0.911	0.892	60
0.4	0.470	0.473	0.473	0.465	60

```
> data(ar1.s); data(ar1.2.s)
> ar(ar1.s,order.max=1,AIC=F,method='yw')
> ar(ar1.s,order.max=1,AIC=F,method='ols')
> ar(ar1.s,order.max=1,AIC=F,method='mle')
```

```
> ar(ar1.2.s,order.max=1,AIC=F,method='yw')
> ar(ar1.2.s,order.max=1,AIC=F,method='ols')
> ar(ar1.2.s,order.max=1,AIC=F,method='mle')
```

From Equation (7.4.9) on page 161, the standard errors for the estimates are

$$\sqrt{\widehat{Var}(\hat{\phi})} \approx \sqrt{\frac{1-\hat{\phi}^2}{n}} = \sqrt{\frac{1-(0.831)^2}{60}} \approx 0.07$$

and

$$\sqrt{\widehat{Var}(\hat{\phi})} = \sqrt{\frac{1-(0.470)^2}{60}} \approx 0.11$$

respectively. Considering the magnitude of these standard errors, all four methods estimate reasonably well for AR(1) models.

Exhibit 7.5 Parameter Estimation for a Simulated AR(2) Model

Parameters	Method-of-Moments Estimates	Conditional SS Estimates	Unconditional SS Estimates	Maximum Likelihood Estimate	*n*
$\phi_1 = 1.5$	1.472	1.5137	1.5183	1.5061	120
$\phi_2 = -0.75$	–0.767	–0.8050	–0.8093	–0.7965	120

```
> data(ar2.s)
> ar(ar2.s,order.max=2,AIC=F,method='yw')
> ar(ar2.s,order.max=2,AIC=F,method='ols')
> ar(ar2.s,order.max=2,AIC=F,method='mle')
```

From Equation (7.4.10) on page 161, the standard errors for the estimates are

$$\sqrt{\widehat{Var}(\hat{\phi}_1)} \approx \sqrt{\widehat{Var}(\hat{\phi}_2)} \approx \sqrt{\frac{1-\phi_2^2}{n}} = \sqrt{\frac{1-(0.75)^2}{120}} \approx 0.06$$

Again, considering the size of the standard errors, all four methods estimate reasonably well for AR(2) models.

As a final example using simulated data, consider the ARMA(1,1) shown in Exhibit 6.14 on page 123. Here $\phi = 0.6$, $\theta = -0.3$, and $n = 100$. Estimates using the various methods are shown in Exhibit 7.6.

Exhibit 7.6 Parameter Estimation for a Simulated ARMA(1,1) Model

Parameters	Method-of-Moments Estimates	Conditional SS Estimates	Unconditional SS Estimates	Maximum Likelihood Estimate	*n*
$\phi = 0.6$	0.637	0.5586	0.5691	0.5647	100
$\theta = -0.3$	–0.2066	–0.3669	–0.3618	–0.3557	100

```
> data(arma11.s)
> arima(arma11.s, order=c(1,0,1),method='CSS')
> arima(arma11.s, order=c(1,0,1),method='ML')
```

Now let's look at some real time series. The industrial chemical property time series was first shown in Exhibit 1.3 on page 3. The sample PACF displayed in Exhibit 6.26 on page 135, strongly suggested an AR(1) model for this series. Exhibit 7.7 shows the various estimates of the ϕ parameter using four different methods of estimation.

Exhibit 7.7 Parameter Estimation for the Color Property Series

Parameter	Method-of-Moments Estimate	Conditional SS Estimate	Unconditional SS Estimate	Maximum Likelihood Estimate	*n*
ϕ	0.5282	0.5549	0.5890	0.5703	35

```
> data(color)
> ar(color,order.max=1,AIC=F,method='yw')
> ar(color,order.max=1,AIC=F,method='ols')
> ar(color,order.max=1,AIC=F,method='mle')
```

Here the standard error of the estimates is about

$$\sqrt{\hat{Var}(\hat{\phi})} \approx \sqrt{\frac{1-(0.57)^2}{35}} \approx 0.14$$

so all of the estimates are comparable.

As a second example, consider again the Canadian hare abundance series. As before, we base all modeling on the square root of the original abundance numbers. Based on the partial autocorrelation function shown in Exhibit 6.29 on page 137, we will estimate an AR(3) model. For this illustration, we use maximum likelihood estimation and show the results obtained from the R software in Exhibit 7.8.

Exhibit 7.8 Maximum Likelihood Estimates from R Software: Hare Series

Coefficients:	ar1	ar2	ar3	Intercept†
	1.0519	–0.2292	–0.3931	5.6923
s.e.	0.1877	0.2942	0.1915	0.3371

```
sigma^2 estimated as 1.066: log-likelihood = -46.54, AIC = 101.08
```

† The intercept here is the estimate of the process mean μ—not of θ_0.

```
> data(hare)
> arima(sqrt(hare),order=c(3,0,0))
```

Here we see that $\hat{\phi}_1 = 1.0519$, $\hat{\phi}_2 = -0.2292$, and $\hat{\phi}_3 = -0.3930$. We also see that the estimated noise variance is $\hat{\sigma}_e^2 = 1.066$. Noting the standard errors, the estimates of the lag 1 and lag 3 autoregressive coefficients are significantly different from zero, as is the intercept term, but the lag 2 autoregressive parameter estimate is not significant.

The estimated model would be written

$$\sqrt{Y_t} - 5.6923 = 1.0519(\sqrt{Y_{t-1}} - 5.6923) - 0.2292(\sqrt{Y_{t-2}} - 5.6923) - 0.3930(\sqrt{Y_{t-3}} - 5.6923) + e_t$$

or

$$\sqrt{Y_t} = 3.25 + 1.0519\sqrt{Y_{t-1}} - 0.2292\sqrt{Y_{t-2}} - 0.3930\sqrt{Y_{t-3}} + e_t$$

where Y_t is the hare abundance in year t in original terms. Since the lag 2 autoregressive term is insignificant, we might drop that term (that is, set $\phi_2 = 0$) and obtain new estimates of ϕ_1 and ϕ_3 with this subset model.

As a last example, we return to the oil price series. The sample ACF shown in Exhibit 6.32 on page 140, suggested an MA(1) model on the differences of the logs of the prices. Exhibit 7.9 gives the estimates of θ by the various methods and, as we have seen earlier, the method-of-moments estimate differs quite a bit from the others. The others are nearly equal given their standard errors of about 0.07.

Exhibit 7.9 Estimation for the Difference of Logs of the Oil Price Series

Parameter	Method-of-Moments Estimate	Conditional SS Estimate	Unconditional SS Estimate	Maximum Likelihood Estimate	*n*
θ	–0.2225	–0.2731	–0.2954	–0.2956	241

```
> data(oil.price)
> arima(log(oil.price),order=c(0,1,1),method='CSS')
> arima(log(oil.price),order=c(0,1,1),method='ML')
```

7.6 Bootstrapping ARIMA Models

In Section 7.4, we summarized some approximate normal distribution results for the estimator $\hat{\gamma}$, where γ is the vector consisting of all the ARMA parameters. These normal approximations are accurate for large samples, and statistical software generally uses those results in calculating and reporting standard errors. The standard error of some complex function of the model parameters, for example the quasi-period of the model, if it exists, is then usually obtained by the delta method. However, the general theory provides no practical guidance on how large the sample size should be for the normal approximation to be reliable. Bootstrap methods (Efron and Tibshirani, 1993; Davison and Hinkley, 2003) provide an alternative approach to assessing the uncertainty of an estimator and may be more accurate for small samples. There are several variants of the bootstrap method for dependent data—see Politis (2003). We shall confine our discussion to the parametric bootstrap that generates the bootstrap time series $Y_1^*, Y_2^*, \ldots, Y_n^*$ by simulation from the fitted ARIMA(p,d,q) model. (The bootstrap may be done by fixing the first $p + d$ initial values of Y^* to those of the observed data. For stationary models, an alternative procedure is to simulate stationary realizations from the fitted model, which can be done approximately by simulating a long time series from the fitted model and then deleting the transient initial segment of the simulated data—the so-called burn-in.) If the errors are assumed to be normally distributed, the errors may be drawn randomly and with replacement from $N(0, \hat{\sigma}_{\hat{e}}^2)$. For the case of an unknown error distribution, the errors can be drawn randomly and with replacement from the residuals of the fitted model. For each bootstrap series, let $\hat{\gamma}^*$ be the estimator computed based on the bootstrap time series data using the method of full maximum likelihood estimation assuming stationarity. (Other estimation methods may be used.) The bootstrap is replicated, say, B times. (For example, $B = 1000$.) From the B bootstrap parameter estimates, we can form an empirical distribution and use it to calibrate the uncertainty in $\hat{\gamma}$. Suppose we are interested in estimating some function of γ, say $h(\gamma)$—for example, the AR(1) coefficient. Using the percentile method, a 95% bootstrap confidence interval for $h(\gamma)$ can be obtained as the interval from the 2.5 percentile to the 97.5 percentile of the bootstrap distribution of $h(\hat{\gamma}^*)$.

We illustrate the bootstrap method with the hare data. The bootstrap 95% confidence intervals reported in the first row of the table in Exhibit 7.10 are based on the bootstrap obtained by conditioning on the initial three observations and assuming normal errors. Those in the second row are obtained using the same method except that the errors are drawn from the residuals. The third and fourth rows report the confidence intervals based on the stationary bootstrap with a normal error distribution for the third row and the empirical residual distribution for the fourth row. The fifth row in the table shows the theoretical 95% confidence intervals based on the large-sample distribution results for the estimators. In particular, the bootstrap time series for the first bootstrap method is generated recursively using the equation

$$Y_t^* - \hat{\phi}_1 Y_{t-1}^* - \hat{\phi}_2 Y_{t-2}^* - \hat{\phi}_3 Y_{t-3}^* = \hat{\theta}_0 + e_t^* \tag{7.6.1}$$

for $t = 4, 5, \ldots, 31$, where the e_t^* are chosen independently from $N(0, \hat{\sigma}_e^2)$, $Y_1^* = Y_1$, $Y_2^* = Y_2$, $Y_3^* = Y_3$; and the parameters are set to be the estimates from the AR(3) model fitted to the (square root transformed) hare data with $\hat{\theta}_0 = \hat{\mu}(1 - \hat{\phi}_1 - \hat{\phi}_2 - \hat{\phi}_3)$. All results are based on about 1000 bootstrap replications, but full maximum likelihood estimation fails for 6.3%, 6.3%, 3.8%, and 4.8% of 1000 cases for the four bootstrap methods I, II, III, and IV, respectively.

Exhibit 7.10 Bootstrap and Theoretical Confidence Intervals for the AR(3) Model Fitted to the Hare Data

Method	ar1	ar2	ar3	intercept	noise var.
I	(0.593, 1.269)	(–0.655, 0.237)	(–0.666, –0.018)	(5.115, 6.394)	(0.551, 1.546)
II	(0.612, 1.296)	(–0.702, 0.243)	(–0.669, –0.026)	(5.004, 6.324)	(0.510, 1.510)
III	(0.699, 1.369)	(–0.746, 0.195)	(–0.666, –0.021)	(5.056, 6.379)	(0.499, 1.515)
IV	(0.674, 1.389)	(–0.769, 0.194)	(–0.665, –0.002)	(4.995, 6.312)	(0.477, 1.530)
Theoretical	(0.684, 1.42)	(–0.8058, 0.3474)	(–0.7684,–0.01776)	(5.032, 6.353)	(0.536, 1.597)

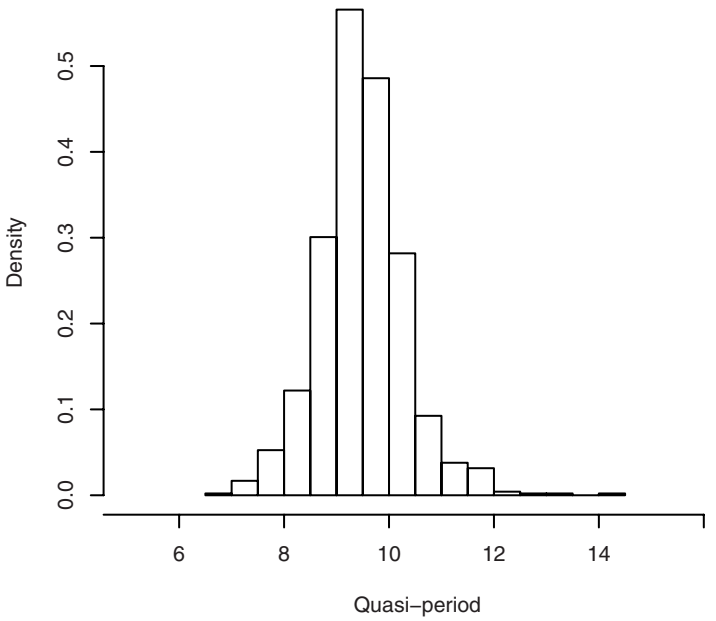
```
> See the Chapter 7 R scripts file for the extensive code
  required to generate these results.
```

All four methods yield similar bootstrap confidence intervals, although the conditional bootstrap approach generally yields slightly narrower confidence intervals. This is expected, as the conditional bootstrap time series bear more resemblance to each other because all are subject to identical initial conditions. The bootstrap confidence intervals are generally slightly wider than their theoretical counterparts that are derived from the large-sample results. Overall, we can draw the inference that the ϕ_2 coefficient estimate is insignificant, whereas both the ϕ_1 and ϕ_3 coefficient estimates are significant at the 5% significance level.

The bootstrap method has the advantage of allowing easy construction of confidence intervals for a model characteristic that is a nonlinear function of the model parameters. For example, the characteristic AR polynomial of the fitted AR(3) model for the hare data admits a pair of complex roots. Indeed, the roots are $0.84 \pm 0.647i$ and -2.26, where $i = \sqrt{-1}$. The two complex roots can be written in polar form: $1.06\exp(\pm 0.657i)$. As in the discussion of the quasi-period for the AR(2) model on page 74, the quasi-period of the fitted AR(3) model can be defined as $2\pi/0.657 = 9.57$. Thus, the fitted model suggests that the hare abundance underwent cyclical fluctuation with a period of about 9.57 years. The interesting question of constructing a 95% confidence interval for the quasi-period could be studied using the delta method. However, this will be quite complex, as the quasi-period is a complicated function of the parameters. But the bootstrap provides a simple solution: For each set of bootstrap parameter estimates, we can compute the quasi-period and hence obtain the bootstrap distribution of the quasi-period. Confidence intervals for the quasi-period can then be constructed using the percentile method, and the shape of the distribution can be explored via the histogram of the bootstrap quasi-period estimates. (Note that the quasi-period will be unde-

fined whenever the roots of the AR characteristic equation are all real numbers.) Among the 1000 stationary bootstrap time series obtained by simulating from the fitted model with the errors drawn randomly from the residuals with replacement, 952 series lead to successful full maximum likelihood estimation. All but one of the 952 series have well-defined quasi-periods, and the histogram of these is shown in Exhibit 7.11. The histogram shows that the sampling distribution of the quasi-period estimate is slightly skewed to the right.[†] The Q-Q normal plot (Exhibit 7.12) suggests that the quasi-period estimator has, furthermore, a thick-tailed distribution. Thus, the delta method and the corresponding normal distribution approximation may be inappropriate for approximating the sampling distribution of the quasi-period estimator. Finally, using the percentile method, a 95% confidence interval of the quasi-period is found to be (7.84,11.34).

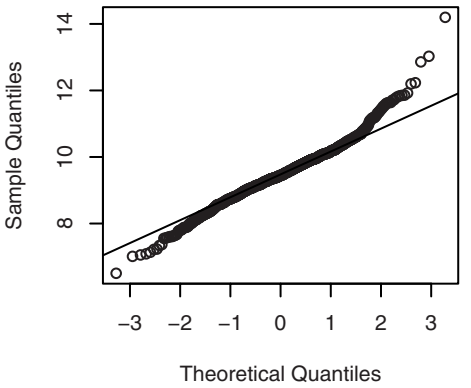
Exhibit 7.11 Histogram of Bootstrap Quasi-period Estimates




```
> win.graph(width=3.9,height=3.8,pointsize=8)
> hist(period.replace,prob=T,xlab='Quasi-period',axes=F,
    xlim=c(5,16))
> axis(2); axis(1,c(4,6,8,10,12,14,16),c(4,6,8,10,12,14,NA))
```

[†] However, see the discussion below Equation (13.5.9) on page 338 where it is argued that, from the perspective of frequency domain, there is a small parametric region corresponding to complex roots and yet the associated quasi-period may not be physically meaningful. This illustrates the subtlety of the concept of quasi-period.

Exhibit 7.12 Q-Q Normal Plot of Bootstrap Quasi-period Estimates




```
> win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(period.replace); qqline(period.replace)
```

7.7 Summary

This chapter delved into the estimation of the parameters of ARIMA models. We considered estimation criteria based on the method of moments, various types of least squares, and maximizing the likelihood function. The properties of the various estimators were given, and the estimators were illustrated both with simulated and actual time series data. Bootstrapping with ARIMA models was also discussed and illustrated.

EXERCISES

7.1 From a series of length 100, we have computed $r_1 = 0.8$, $r_2 = 0.5$, $r_3 = 0.4$, $\overline{Y} = 2$, and a sample variance of 5. If we assume that an AR(2) model with a constant term is appropriate, how can we get (simple) estimates of ϕ_1, ϕ_2, θ_0, and σ_e^2?

7.2 Assuming that the following data arise from a stationary process, calculate method-of-moments estimates of μ, γ_0, and ρ_1: 6, 5, 4, 6, 4.

7.3 If $\{Y_t\}$ satisfies an AR(1) model with ϕ of about 0.7, how long of a series do we need to estimate $\phi = \rho_1$ with 95% confidence that our estimation error is no more than ± 0.1?

7.4 Consider an MA(1) process for which it is *known* that the process mean is zero. Based on a series of length $n = 3$, we observe $Y_1 = 0$, $Y_2 = -1$, and $Y_3 = ½$.

(a) Show that the conditional least-squares estimate of θ is ½.

(b) Find an estimate of the noise variance. (Hint: Iterative methods are not needed in this simple case.)

7.5 Given the data $Y_1 = 10$, $Y_2 = 9$, and $Y_3 = 9.5$, we wish to fit an IMA(1,1) model without a constant term.
(a) Find the conditional least squares estimate of θ. (Hint: Do Exercise 7.4 first.)
(b) Estimate $\sigma_{\hat{e}}^2$.

7.6 Consider two different parameterizations of the AR(1) process with nonzero mean:

Model I. $Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$.

Model II. $Y_t = \phi Y_{t-1} + \theta_0 + e_t$.

We want to estimate ϕ and μ or ϕ and θ_0 using conditional least squares conditional on Y_1. Show that with Model I we are led to solve nonlinear equations to obtain the estimates, while with Model II we need only solve linear equations.

7.7 Verify Equation (7.1.4) on page 150.

7.8 Consider an ARMA(1,1) model with $\phi = 0.5$ and $\theta = 0.45$.
(a) For $n = 48$, evaluate the variances and correlation of the maximum likelihood estimators of ϕ and θ using Equations (7.4.13) on page 161. Comment on the results.
(b) Repeat part (a) but now with $n = 120$. Comment on the new results.

7.9 Simulate an MA(1) series with $\theta = 0.8$ and $n = 48$.
(a) Find the method-of-moments estimate of θ.
(b) Find the conditional least squares estimate of θ and compare it with part (a).
(c) Find the maximum likelihood estimate of θ and compare it with parts (a) and (b).
(d) Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare your results with your results from the first simulation.

7.10 Simulate an MA(1) series with $\theta = -0.6$ and $n = 36$.
(a) Find the method-of-moments estimate of θ.
(b) Find the conditional least squares estimate of θ and compare it with part (a).
(c) Find the maximum likelihood estimate of θ and compare it with parts (a) and (b).
(d) Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare your results with your results from the first simulation.

7.11 Simulate an MA(1) series with $\theta = -0.6$ and $n = 48$.
(a) Find the maximum likelihood estimate of θ.
(b) If your software permits, repeat part (a) many times with a new simulated series using the same parameters and same sample size.
(c) Form the sampling distribution of the maximum likelihood estimates of θ.
(d) Are the estimates (approximately) unbiased?
(e) Calculate the variance of your sampling distribution and compare it with the large-sample result in Equation (7.4.11) on page 161.

7.12 Repeat Exercise 7.11 using a sample size of $n = 120$.

7.13 Simulate an AR(1) series with $\phi = 0.8$ and $n = 48$.
- **(a)** Find the method-of-moments estimate of ϕ.
- **(b)** Find the conditional least squares estimate of ϕ and compare it with part (a).
- **(c)** Find the maximum likelihood estimate of ϕ and compare it with parts (a) and (b).
- **(d)** Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare your results with your results from the first simulation.

7.14 Simulate an AR(1) series with $\phi = -0.5$ and $n = 60$.
- **(a)** Find the method-of-moments estimate of ϕ.
- **(b)** Find the conditional least squares estimate of ϕ and compare it with part (a).
- **(c)** Find the maximum likelihood estimate of ϕ and compare it with parts (a) and (b).
- **(d)** Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare your results with your results from the first simulation.

7.15 Simulate an AR(1) series with $\phi = 0.7$ and $n = 100$.
- **(a)** Find the maximum likelihood estimate of ϕ.
- **(b)** If your software permits, repeat part (a) many times with a new simulated series using the same parameters and same sample size.
- **(c)** Form the sampling distribution of the maximum likelihood estimates of ϕ.
- **(d)** Are the estimates (approximately) unbiased?
- **(e)** Calculate the variance of your sampling distribution and compare it with the large-sample result in Equation (7.4.9) on page 161.

7.16 Simulate an AR(2) series with $\phi_1 = 0.6$, $\phi_2 = 0.3$, and $n = 60$.
- **(a)** Find the method-of-moments estimates of ϕ_1 and ϕ_2.
- **(b)** Find the conditional least squares estimates of ϕ_1 and ϕ_2 and compare them with part (a).
- **(c)** Find the maximum likelihood estimates of ϕ_1 and ϕ_2 and compare them with parts (a) and (b).
- **(d)** Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare these results to your results from the first simulation.

7.17 Simulate an ARMA(1,1) series with $\phi = 0.7$, $\theta = 0.4$, and $n = 72$.
- **(a)** Find the method-of-moments estimates of ϕ and θ.
- **(b)** Find the conditional least squares estimates of ϕ and θ and compare them with part (a).
- **(c)** Find the maximum likelihood estimates of ϕ and θ and compare them with parts (a) and (b).
- **(d)** Repeat parts (a), (b), and (c) with a new simulated series using the same parameters and same sample size. Compare your new results with your results from the first simulation.

7.18 Simulate an AR(1) series with $\phi = 0.6$, $n = 36$ but with error terms from a t-distribution with 3 degrees of freedom.

(a) Display the sample PACF of the series. Is an AR(1) model suggested?
(b) Estimate ϕ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.19 Simulate an MA(1) series with $\theta = -0.8$, $n = 60$ but with error terms from a t-distribution with 4 degrees of freedom.
(a) Display the sample ACF of the series. Is an MA(1) model suggested?
(b) Estimate θ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.20 Simulate an AR(2) series with $\phi_1 = 1.0$, $\phi_2 = -0.6$, $n = 48$ but with error terms from a t-distribution with 5 degrees of freedom.
(a) Display the sample PACF of the series. Is an AR(2) model suggested?
(b) Estimate ϕ_1 and ϕ_2 from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.21 Simulate an ARMA(1,1) series with $\phi = 0.7$, $\theta = -0.6$, $n = 48$ but with error terms from a t-distribution with 6 degrees of freedom.
(a) Display the sample EACF of the series. Is an ARMA(1,1) model suggested?
(b) Estimate ϕ and θ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.22 Simulate an AR(1) series with $\phi = 0.6$, $n = 36$ but with error terms from a chi-square distribution with 6 degrees of freedom.
(a) Display the sample PACF of the series. Is an AR(1) model suggested?
(b) Estimate ϕ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.23 Simulate an MA(1) series with $\theta = -0.8$, $n = 60$ but with error terms from a chi-square distribution with 7 degrees of freedom.
(a) Display the sample ACF of the series. Is an MA(1) model suggested?
(b) Estimate θ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.24 Simulate an AR(2) series with $\phi_1 = 1.0$, $\phi_2 = -0.6$, $n = 48$ but with error terms from a chi-square distribution with 8 degrees of freedom.
(a) Display the sample PACF of the series. Is an AR(2) model suggested?
(b) Estimate ϕ_1 and ϕ_2 from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new simulated series under the same conditions.

7.25 Simulate an ARMA(1,1) series with $\phi = 0.7$, $\theta = -0.6$, $n = 48$ but with error terms from a chi-square distribution with 9 degrees of freedom.
(a) Display the sample EACF of the series. Is an ARMA(1,1) model suggested?
(b) Estimate ϕ and θ from the series and comment on the results.
(c) Repeat parts (a) and (b) with a new series under the same conditions.

7.26 Consider the AR(1) model specified for the color property time series displayed in Exhibit 1.3 on page 3. The data are in the file named color.
(a) Find the method-of-moments estimate of ϕ.
(b) Find the maximum likelihood estimate of ϕ and compare it with part (a).

7.27 Exhibit 6.31 on page 139 suggested specifying either an AR(1) or possibly an AR(4) model for the difference of the logarithms of the oil price series. The data are in the file named oil.price.
(a) Estimate both of these models using maximum likelihood and compare it with the results using the AIC criteria.
(b) Exhibit 6.32 on page 140 suggested specifying an MA(1) model for the difference of the logs. Estimate this model by maximum likelihood and compare to your results in part (a).

7.28 The data file named deere3 contains 57 consecutive values from a complex machine tool at Deere & Co. The values given are deviations from a target value in units of ten millionths of an inch. The process employs a control mechanism that resets some of the parameters of the machine tool depending on the magnitude of deviation from target of the last item produced.
(a) Estimate the parameters of an AR(1) model for this series.
(b) Estimate the parameters of an AR(2) model for this series and compare the results with those in part (a).

7.29 The data file named robot contains a time series obtained from an industrial robot. The robot was put through a sequence of maneuvers, and the distance from a desired ending point was recorded in inches. This was repeated 324 times to form the time series.
(a) Estimate the parameters of an AR(1) model for these data.
(b) Estimate the parameters of an IMA(1,1) model for these data.
(c) Compare the results from parts (a) and (b) in terms of AIC.

7.30 The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.) The time series contains outliers that are quite obvious in the time series plot.
(a) Replace each of the unusual values with a value of 35 days, a much more typical value, and then estimate the parameters of an MA(2) model.
(b) Now assume an MA(5) model and estimate the parameters. Compare these results with those obtained in part (a).

7.31 Simulate a time series of length $n = 48$ from an AR(1) model with $\phi = 0.7$. Use that series as if it were real data. Now compare the theoretical asymptotic distribution of the estimator of ϕ with the distribution of the bootstrap estimator of ϕ.

7.32 The industrial color property time series was fitted quite well by an AR(1) model. However, the series is rather short, with $n = 35$. Compare the theoretical asymptotic distribution of the estimator of ϕ with the distribution of the bootstrap estimator of ϕ. The data are in the file named color.

CHAPTER 8

MODEL DIAGNOSTICS

We have now discussed methods for specifying models and for efficiently estimating the parameters in those models. Model diagnostics, or model criticism, is concerned with testing the goodness of fit of a model and, if the fit is poor, suggesting appropriate modifications. We shall present two complementary approaches: analysis of residuals from the fitted model and analysis of overparameterized models; that is, models that are more general than the proposed model but that contain the proposed model as a special case.

8.1 Residual Analysis

We already used the basic ideas of residual analysis in Section 3.6 on page 42 when we checked the adequacy of fitted deterministic trend models. With autoregressive models, residuals are defined in direct analogy to that earlier work. Consider in particular an AR(2) model with a constant term:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \theta_0 + e_t \tag{8.1.1}$$

Having estimated ϕ_1, ϕ_2, and θ_0, the residuals are defined as

$$\hat{e}_t = Y_t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2} - \hat{\theta}_0 \tag{8.1.2}$$

For general ARMA models containing moving average terms, we use the inverted, infinite autoregressive form of the model to define residuals. For simplicity, we assume that θ_0 is zero. From the inverted form of the model, Equation (4.5.5) on page 80, we have

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t$$

so that the residuals are defined as

$$\hat{e}_t = Y_t - \hat{\pi}_1 Y_{t-1} - \hat{\pi}_2 Y_{t-2} - \hat{\pi}_3 Y_{t-3} - \cdots \tag{8.1.3}$$

Here the π's are not estimated directly but rather implicitly as functions of the ϕ's and θ's. In fact, the residuals are not calculated using this equation but as a by-product of the estimation of the ϕ's and θ's. In Chapter 9, we shall argue, that

$$\hat{Y}_t = \hat{\pi}_1 Y_{t-1} + \hat{\pi}_2 Y_{t-2} + \hat{\pi}_3 Y_{t-3} + \cdots$$

is the best forecast of Y_t based on Y_{t-1}, Y_{t-2}, Y_{t-3},… . Thus Equation (8.1.3) can be rewritten as

$$\text{residual} = \text{actual} - \text{predicted}$$

in direct analogy with regression models. Compare this with Section 3.6 on page 42.

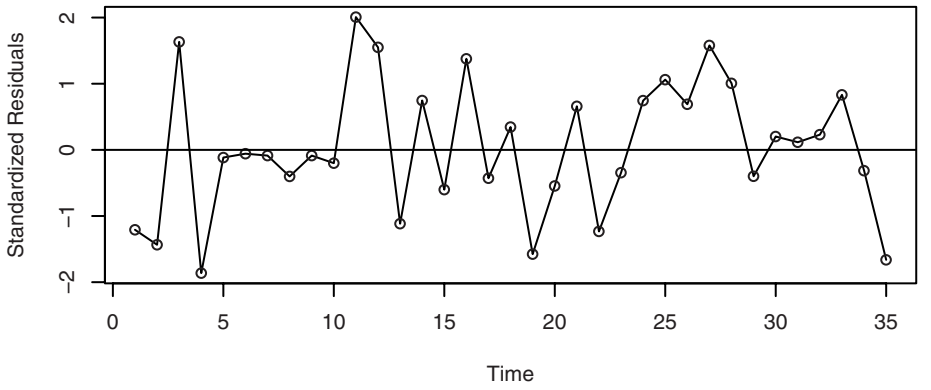
If the model is correctly specified and the parameter estimates are reasonably close to the true values, then the residuals should have nearly the properties of white noise. They should behave roughly like independent, identically distributed normal variables with zero means and common standard deviations. Deviations from these properties can help us discover a more appropriate model.

Plots of the Residuals

Our first diagnostic check is to inspect a plot of the residuals over time. If the model is adequate, we expect the plot to suggest a rectangular scatter around a zero horizontal level with no trends whatsoever.

Exhibit 8.1 shows such a plot for the standardized residuals from the AR(1) model fitted to the industrial color property series. Standardization allows us to see residuals of unusual size much more easily. The parameters were estimated using maximum likelihood. This plot supports the model, as no trends are present.

Exhibit 8.1 Standardized Residuals from AR(1) Model of Color



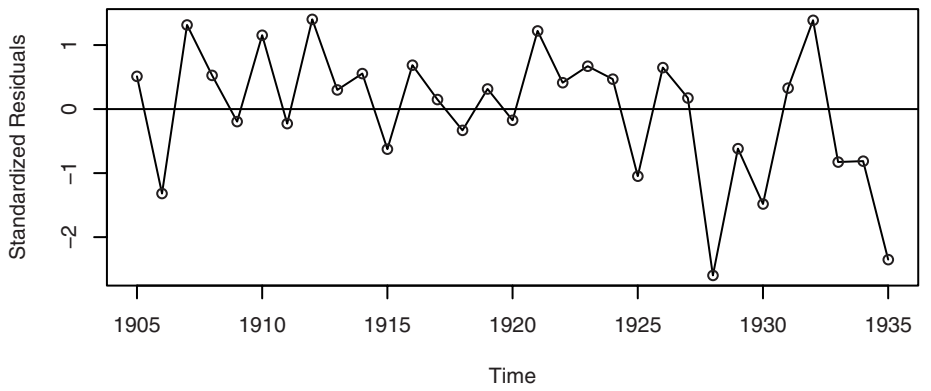

```
> win.graph(width=4.875,height=3,pointsize=8)
> data(color)
> m1.color=arima(color,order=c(1,0,0)); m1.color
> plot(rstandard(m1.color),ylab ='Standardized Residuals',
    type='o'); abline(h=0)
```

As a second example, we consider the Canadian hare abundance series. We estimate a subset AR(3) model with ϕ_2 set to zero, as suggested by the discussion following Exhibit 7.8 on page 166. The estimated model is

$$\sqrt{Y_t} = 3.483 + 0.919\sqrt{Y_{t-1}} - 0.5313\sqrt{Y_{t-3}} + e_t \tag{8.1.4}$$

and the time series plot of the standardized residuals from this model is shown in Exhibit 8.2. Here we see possible reduced variation in the middle of the series and increased variation near the end of the series—not exactly an ideal plot of residuals.[†]

Exhibit 8.2 Standardized Residuals from AR(3) Model for Sqrt(Hare)



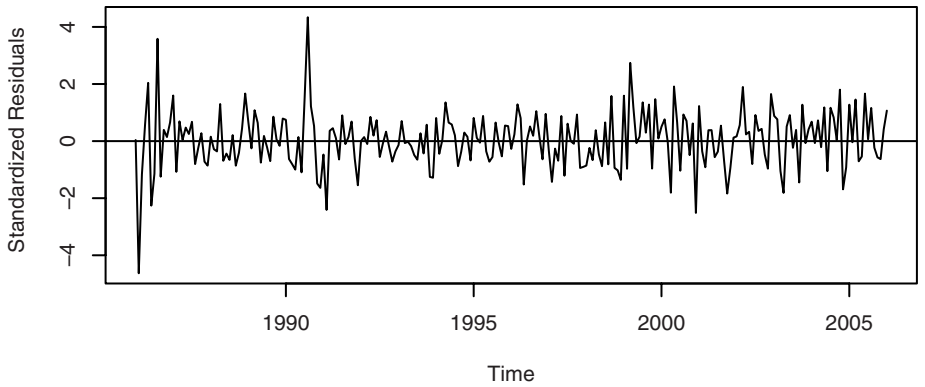

```
> data(hare)
> m1.hare=arima(sqrt(hare),order=c(3,0,0)); m1.hare
> m2.hare=arima(sqrt(hare),order=c(3,0,0),fixed=c(NA,0,NA,NA))
> m2.hare
> # Note that the intercept term given in R is actually the mean
    in the centered form of the ARMA model; that is, if
    y(t)=sqrt(hare)-intercept, then the model is
    y(t)=0.919*y(t-1)-0.5313*y(t-3)+e(t)
> # So the 'true' intercept equals 5.6889*(1-0.919+0.5313)=3.483
> plot(rstandard(m2.hare),ylab='Standardized Residuals',type='o')
> abline(h=0)
```

Exhibit 8.3 displays the time series plot of the standardized residuals from the IMA(1,1) model estimated for the logarithms of the oil price time series. The model was fitted using maximum likelihood estimation. There are at least two or three residuals early in the series with magnitudes larger than 3—very unusual in a standard normal distribution.[‡] Ideally, we should go back to those months and try to learn what outside factors may have influenced unusually large drops or unusually large increases in the price of oil.

[†] The seemingly large negative standardized residuals are not outliers according to the Bonferroni outlier criterion with critical values ±3.15.

[‡] The Bonferroni critical values with $n = 241$ and $\alpha = 0.05$ are ±3.71, so the outliers do appear to be real. We will model them in Chapter 11.

Exhibit 8.3 Standardized Residuals from Log Oil Price IMA(1,1) Model



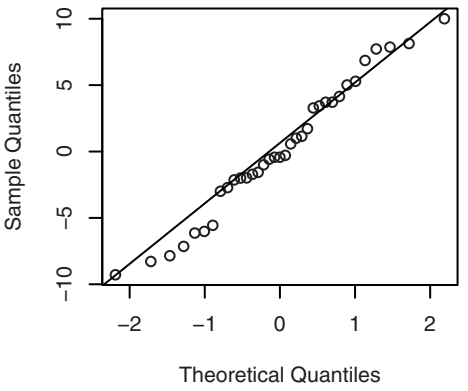

```
> data(oil.price)
> m1.oil=arima(log(oil.price),order=c(0,1,1))
> plot(rstandard(m1.oil),ylab='Standardized residuals',type='l')
> abline(h=0)
```

Normality of the Residuals

As we saw in Chapter 3, quantile-quantile plots are an effective tool for assessing normality. Here we apply them to residuals.

A quantile-quantile plot of the residuals from the AR(1) model estimated for the industrial color property series is shown in Exhibit 8.4. The points seem to follow the straight line fairly closely—especially the extreme values. This graph would not lead us to reject normality of the error terms in this model. In addition, the Shapiro-Wilk normality test applied to the residuals produces a test statistic of $W = 0.9754$, which corresponds to a p-value of 0.6057, and we would not reject normality based on this test.

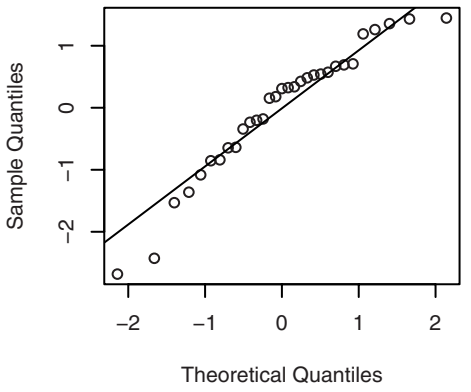
Exhibit 8.4 Quantile-Quantile Plot: Residuals from AR(1) Color Model




```
> win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(residuals(m1.color)); qqline(residuals(m1.color))
```

The quantile-quantile plot for the residuals from the AR(3) model for the square root of the hare abundance time series is displayed in Exhibit 8.5. Here the extreme values look suspect. However, the sample is small ($n = 31$) and, as stated earlier, the Bonferroni criteria for outliers do not indicate cause for alarm.

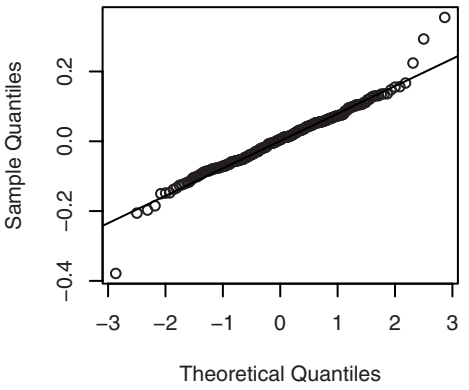
Exhibit 8.5 Quantile-Quantile Plot: Residuals from AR(3) for Hare




```
> qqnorm(residuals(m1.hare)); qqline(residuals(m1.hare))
```

Exhibit 8.6 gives the quantile-quantile plot for the residuals from the IMA(1,1) model that was used to model the logarithms of the oil price series. Here the outliers are quite prominent, and we will deal with them in Chapter 11.

Exhibit 8.6 Quantile-Quantile Plot: Residuals from IMA(1,1) Model for Oil




```
> qqnorm(residuals(m1.oil)); qqline(residuals(m1.oil))
```

Autocorrelation of the Residuals

To check on the independence of the noise terms in the model, we consider the sample autocorrelation function of the residuals, denoted $\hat{r}_k$. From Equation (6.1.3) on page 110, we know that for true white noise and large n, the sample autocorrelations are approximately uncorrelated and normally distributed with zero means and variance $1/n$. Unfortunately, even residuals from a correctly specified model with efficiently estimated parameters have somewhat different properties. This was first explored for multiple- regression models in a series of papers by Durbin and Watson (1950, 1951, 1971) and for autoregressive models in Durbin (1970). The key reference on the distribution of residual autocorrelations in ARIMA models is Box and Pierce (1970), the results of which were generalized in McLeod (1978).

Generally speaking, the residuals are approximately normally distributed with zero means; however, for small lags k and j, the variance of $\hat{r}_k$ can be substantially less than $1/n$ and the estimates $\hat{r}_k$ and $\hat{r}_j$ can be highly correlated. For larger lags, the approximate variance $1/n$ does apply, and further $\hat{r}_k$ and $\hat{r}_j$ are approximately uncorrelated.

As an example of these results, consider a correctly specified and efficiently estimated AR(1) model. It can be shown that, for large n,

$$Var(\hat{r}_1) \approx \frac{\phi^2}{n} \tag{8.1.5}$$

$$Var(\hat{r}_k) \approx \frac{1-(1-\phi^2)\phi^{2k-2}}{n} \quad \text{for} \quad k>1 \tag{8.1.6}$$

$$Corr(\hat{r}_1, \hat{r}_k) \approx -sign(\phi)\frac{(1-\phi^2)\phi^{k-2}}{1-(1-\phi^2)\phi^{2k-2}} \quad \text{for} \quad k>1 \tag{8.1.7}$$

where

$$sign(\phi) = \begin{cases} 1 & \text{if } \phi > 0 \\ 0 & \text{if } \phi = 0 \\ -1 & \text{if } \phi < 0 \end{cases}$$

The table in Exhibit 8.7 illustrates these formulas for a variety of values of ϕ and k. Notice that $Var(\hat{r}_1) \approx 1/n$ is a reasonable approximation for $k \geq 2$ over a wide range of ϕ-values.

Exhibit 8.7 Approximations for Residual Autocorrelations in AR(1) Models

ϕ	0.3	0.5	0.7	0.9	ϕ	0.3	0.5	0.7	0.9
k	**Standard deviation of $\hat{r}_k$ times $\sqrt{n}$**					**Correlation $\hat{r}_1$ with $\hat{r}_k$**			
1	0.30	0.50	0.70	0.90		1.00	1.00	1.00	1.00
2	0.96	0.90	0.87	0.92		−0.95	−0.83	−0.59	−0.21
3	1.00	0.98	0.94	0.94		−0.27	−0.38	−0.38	−0.18
4	1.00	0.99	0.97	0.95		−0.08	−0.19	−0.26	−0.16
5	1.00	1.00	0.99	0.96		−0.02	−0.09	−0.18	−0.14
6	1.00	1.00	0.99	0.97		−0.01	−0.05	−0.12	−0.13
7	1.00	1.00	1.00	0.97		−0.00	−0.02	−0.09	−0.12
8	1.00	1.00	1.00	0.98		−0.00	−0.01	−0.06	−0.10
9	1.00	1.00	1.00	0.99		−0.00	−0.00	−0.03	−0.08

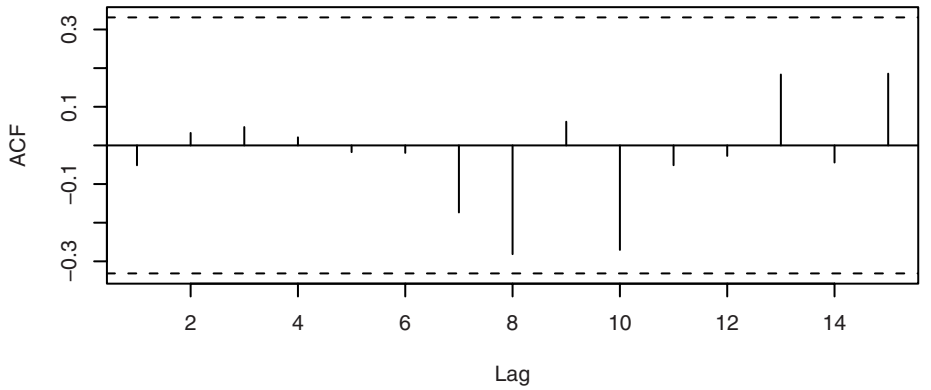
If we apply these results to the AR(1) model that was estimated for the industrial color property time series with $\hat{\phi} = 0.57$ and $n = 35$, we obtain the results shown in Exhibit 8.8.

Exhibit 8.8 Approximate Standard Deviations of Residual ACF values

Lag *k*	1	2	3	4	5	> 5
$\sqrt{Var(\hat{r}_k)}$	0.096	0.149	0.163	0.167	0.168	0.169

A graph of the sample ACF of these residuals is shown in Exhibit 8.9. The dashed horizontal lines plotted are based on the large lag standard error of $\pm 2/\sqrt{n}$. There is no evidence of autocorrelation in the residuals of this model.

Exhibit 8.9 Sample ACF of Residuals from AR(1) Model for Color




```
> win.graph(width=4.875,height=3,pointsize=8)
> acf(residuals(m1.color))
```

For an AR(2) model, it can be shown that

$$Var(\hat{r}_1) \approx \frac{\phi_2^2}{n} \tag{8.1.8}$$

and

$$Var(\hat{r}_2) \approx \frac{\phi_2^2 + \phi_1^2(1+\phi_2)^2}{n} \tag{8.1.9}$$

If the AR(2) parameters are not too close to the stationarity boundary shown in Exhibit 4.17 on page 72, then

$$Var(\hat{r}_k) \approx \frac{1}{n} \quad \text{for } k \geq 3 \tag{8.1.10}$$

If we fit an AR(2) model[†] by maximum likelihood to the square root of the hare abundance series, we find that $\hat{\phi}_1 = 1.351$ and $\hat{\phi}_2 = -0.776$. Thus we have

$$\sqrt{\widehat{Var}(\hat{r}_1)} \approx \frac{|-0.776|}{\sqrt{35}} = 0.131$$

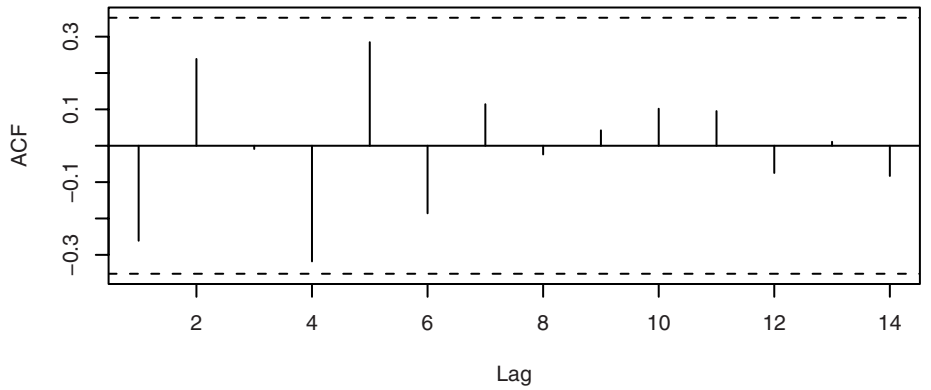
$$\sqrt{\widehat{Var}(\hat{r}_2)} \approx \sqrt{\frac{(-0.776)^2 + (1.351)^2(1+(-0.776))^2}{35}} = 0.141$$

$$\sqrt{\widehat{Var}(\hat{r}_k)} \approx 1/\sqrt{35} = 0.169 \text{ for } k \geq 3$$

[†] The AR(2) model is not quite as good as the AR(3) model that we estimated earlier, but it still fits quite well and serves as a reasonable example here.

Exhibit 8.10 displays the sample ACF of the residuals from the AR(2) model of the square root of the hare abundance. The lag 1 autocorrelation here equals –0.261, which is close to 2 standard errors below zero but not quite. The lag 4 autocorrelation equals –0.318, but its standard error is 0.169. We conclude that the graph does not show statistically significant evidence of nonzero autocorrelation in the residuals.[†]

Exhibit 8.10 Sample ACF of Residuals from AR(2) Model for Hare




```
> acf(residuals(arima(sqrt(hare),order=c(2,0,0))))
```

With monthly data, we would pay special attention to possible excessive autocorrelation in the residuals at lags 12, 24, and so forth. With quarterly series, lags 4, 8, and so forth would merit special attention. Chapter 10 contains examples of these ideas.

It can be shown that results analogous to those for AR models hold for MA models. In particular, replacing ϕ by θ in Equations (8.1.5), (8.1.6), and(8.1.7) gives the results for the MA(1) case. Similarly, results for the MA(2) case can be stated by replacing ϕ_1 and ϕ_2 by θ_1 and θ_2, respectively, in Equations (8.1.8), (8.1.9), and (8.1.10). Results for general ARMA models may be found in Box and Pierce (1970) and McLeod (1978).

The Ljung-Box Test

In addition to looking at residual correlations at individual lags, it is useful to have a test that takes into account their magnitudes as a group. For example, it may be that most of the residual autocorrelations are moderate, some even close to their critical values, but, taken together, they seem excessive. Box and Pierce (1970) proposed the statistic

$$Q = n(\hat{r}_1^2 + \hat{r}_2^2 + \cdots + \hat{r}_K^2) \tag{8.1.11}$$

to address this possibility. They showed that if the correct ARMA(p,q) model is estimated, then, for large n, Q has an approximate chi-square distribution with $K - p - q$

[†] Recall that an AR(3) model fits these data even better and has even less autocorrelation in its residuals, see Exercise 8.7.

degrees of freedom. Fitting an erroneous model would tend to inflate Q. Thus, a general "portmanteau" test would reject the ARMA(p,q) model if the observed value of Q exceeded an appropriate critical value in a chi-square distribution with $K - p - q$ degrees of freedom. (Here the maximum lag K is selected somewhat arbitrarily but large enough that the ψ-weights are negligible for $j > K$.)

The chi-square distribution for Q is based on a limit theorem as $n \to \infty$, but Ljung and Box (1978) subsequently discovered that even for $n = 100$, the approximation is not satisfactory. By modifying the Q statistic slightly, they defined a test statistic whose null distribution is much closer to chi-square for typical sample sizes. The modified Box-Pierce, or **Ljung-Box,** statistic is given by

$$Q_* = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_K^2}{n-K}\right) \tag{8.1.12}$$

Notice that since $(n + 2)/(n - k) > 1$ for every $k \geq 1$, we have $Q_* > Q$, which partly explains why the original statistic Q tended to overlook inadequate models. More details on the exact distributions of Q_* and Q for finite samples can be found in Ljung and Box (1978), see also Davies, Triggs, and Newbold (1977).

Exhibit 8.11 lists the first six autocorrelations of the residuals from the AR(1) fitted model for the color property series. Here $n = 35$.

Exhibit 8.11 Residual Autocorrelation Values from AR(1) Model for Color

Lag k	1	2	3	4	5	6
Residual ACF	–0.051	0.032	0.047	0.021	–0.017	–0.019

```
> acf(residuals(m1.color),plot=F)$acf
> signif(acf(residuals(m1.color),plot=F)$acf[1:6],2)
> # display the first 6 acf values to 2 significant digits
```

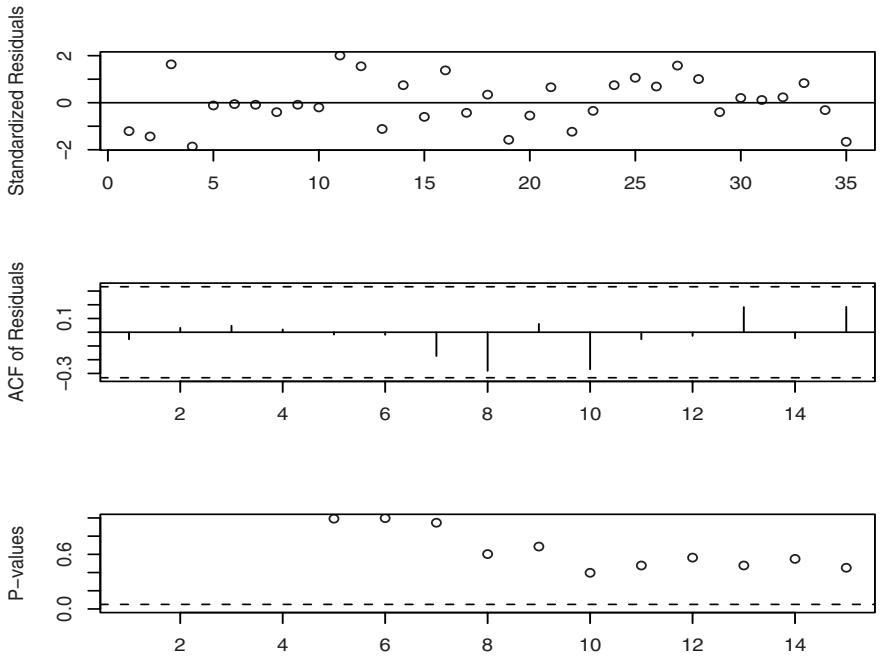
The Ljung-Box test statistic with $K = 6$ is equal to

$$Q_* = 35(35+2)\left(\frac{(-0.051)^2}{35-1} + \frac{(0.032)^2}{35-2} + \frac{(0.047)^2}{35-3} + \frac{(0.021)^2}{35-4} + \frac{(-0.017)^2}{35-5} + \frac{(-0.019)^2}{35-6}\right) \approx 0.28$$

This is referred to a chi-square distribution with $6 - 1 = 5$ degrees of freedom. This leads to a p-value of 0.998, so we have no evidence to reject the null hypothesis that the error terms are uncorrelated.

Exhibit 8.12 shows three of our diagnostic tools in one display—a sequence plot of the standardized residuals, the sample ACF of the residuals, and p-values for the Ljung-Box test statistic for a whole range of values of K from 5 to 15. The horizontal dashed line at 5% helps judge the size of the p-values. In this instance, everything looks very good. The estimated AR(1) model seems to be capturing the dependence structure of the color property time series quite well.

Exhibit 8.12 Diagnostic Display for the AR(1) Model of Color Property




```
> win.graph(width=4.875,height=4.5)
> tsdiag(m1.color,gof=15,omit.initial=F)
```

As in Chapter 3, the runs test may also be used to assess dependence in error terms via the residuals. Applying the test to the residuals from the AR(3) model for the Canadian hare abundance series, we obtain expected runs of 16.09677 versus observed runs of 18. The corresponding *p*-value is 0.602, so we do not have statistically significant evidence against independence of the error terms in this model.

8.2 Overfitting and Parameter Redundancy

Our second basic diagnostic tool is that of **overfitting**. After specifying and fitting what we believe to be an adequate model, we fit a slightly more general model; that is, a model "close by" that contains the original model as a special case. For example, if an AR(2) model seems appropriate, we might overfit with an AR(3) model. The original AR(2) model would be confirmed if:

1. the estimate of the additional parameter, ϕ_3, is not significantly different from zero, and
2. the estimates for the parameters in common, ϕ_1 and ϕ_2, do not change significantly from their original estimates.

As an example, we have specified, fitted, and examined the residuals of an AR(1) model for the industrial color property time series. Exhibit 8.13 displays the output from the R software from fitting the AR(1) model, and Exhibit 8.14 shows the results from fitting an AR(2) model to the same series. First note that, in Exhibit 8.14, the estimate of ϕ_2 is not statistically different from zero. This fact supports the choice of the AR(1) model. Secondly, we note that the two estimates of ϕ_1 are quite close—especially when we take into account the magnitude of their standard errors. Finally, note that while the AR(2) model has a slightly larger log-likelihood value, the AR(1) fit has a smaller AIC value. The penalty for fitting the more complex AR(2) model is sufficient to choose the simpler AR(1) model.

Exhibit 8.13 AR(1) Model Results for the Color Property Series

Coefficients:†	**ar1**	**Intercept**‡
	0.5705	74.3293
s.e.	0.1435	1.9151

sigma^2 estimated as 24.83: log-likelihood = -106.07, AIC = 216.15

† `m1.color` # R code to obtain table

‡ Recall that the intercept here is the estimate of the process mean μ—not θ_0.

Exhibit 8.14 AR(2) Model Results for the Color Property Series

Coefficients:	**ar1**	**ar2**	**Intercept**
	0.5173	0.1005	74.1551
s.e.	0.1717	0.1815	2.1463

sigma^2 estimated as 24.6: log-likelihood = -105.92, AIC = 217.84

```
> arima(color,order=c(2,0,0))
```

A different overfit for this series would be to try an ARMA(1,1) model. Exhibit 8.15 displays the results of this fit. Notice that the standard errors of the estimated coefficients for this fit are rather larger than what we see in Exhibits 8.13 and 8.14. Regardless, the estimate of ϕ_1 from this fit is not significantly different from the estimate in Exhibit 8.13. Furthermore, as before, the estimate of the new parameter, θ, is not significantly different from zero. This adds further support to the AR(1) model.

Exhibit 8.15 Overfit of an ARMA(1,1) Model for the Color Series

Coefficients:	ar1	ma1	Intercept
	0.6721	–0.1467	74.1730
s.e.	0.2147	0.2742	2.1357

sigma^2 estimated as 24.63: log-likelihood = -105.94, AIC = 219.88

```
> arima(color,order=c(1,0,1))
```

As we have noted, any ARMA(p,q) model can be considered as a special case of a more general ARMA model with the additional parameters equal to zero. However, when generalizing ARMA models, we must be aware of the problem of **parameter redundancy** or **lack of identifiability**.

To make these points clear, consider an ARMA(1,2) model:

$$Y_t = \phi Y_{t-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \tag{8.2.1}$$

Now replace t by $t-1$ to obtain

$$Y_{t-1} = \phi Y_{t-2} + e_{t-1} - \theta_1 e_{t-2} - \theta_2 e_{t-3} \tag{8.2.2}$$

If we multiply both sides of Equation (8.2.2) by *any constant* c and then subtract it from Equation (8.2.1), we obtain (after rearranging)

$$Y_t - (\phi + c)Y_{t-1} + \phi c Y_{t-2} = e_t - (\theta_1 + c)e_{t-1} - (\theta_2 - \theta_1 c)e_{t-2} + c\theta_2 e_{t-3}$$

This apparently defines an ARMA(2,3) process. But notice that we have the factorizations

$$1 - (\phi + c)x + \phi c x^2 = (1 - \phi x)(1 - cx)$$

and

$$1 - (\theta_1 + c)x - (\theta_2 - c\theta_1)x^2 + c\theta_2 x^3 = (1 - \theta_1 x - \theta_2 x^2)(1 - cx)$$

Thus the AR and MA characteristic polynomials in the ARMA(2,3) process have a common factor of $(1 - cx)$. Even though Y_t does satisfy the ARMA(2,3) model, clearly the parameters in that model are not unique—the constant c is completely arbitrary. We say that we have **parameter redundancy** in the ARMA(2,3) model.[†]

The implications for fitting and overfitting models are as follows:

1. Specify the original model carefully. If a simple model seems at all promising, check it out before trying a more complicated model.
2. When overfitting, do not increase the orders of both the AR and MA parts of the model simultaneously.

[†] In backshift notation, if $\phi(B)Y_t = \theta(B)e_t$ is a correct model, then so is $(1 - cB)\phi(B)Y_t = (1 - cB)\theta(B)e_t$ for *any* constant c. To have unique parameterization in an ARMA model, we must cancel any common factors in the AR and MA characteristic polynomials.

3. Extend the model in directions suggested by the analysis of the residuals. For example, if after fitting an MA(1) model, substantial correlation remains at lag 2 in the residuals, try an MA(2), not an ARMA(1,1).

As an example, consider the color property series once more. We have seen that an AR(1) model fits quite well. Suppose we try an ARMA(2,1) model. The results of this fit are shown in Exhibit 8.16. Notice that even though the estimate of σ_e^2 and the log-likelihood and AIC values are not too far from their best values, the estimates of ϕ_1, ϕ_2, and θ are way off, and none would be considered different from zero statistically.

Exhibit 8.16 Overfitted ARMA(2,1) Model for the Color Property Series

Coefficients:	ar1	ar2	ma1	Intercept
	0.2189	0.2735	0.3036	74.1653
s.e.	2.0056	1.1376	2.0650	2.1121

sigma^2 estimated as 24.58: log-likelihood = −105.91, AIC = 219.82

```
> arima(color,order=c(2,0,1))
```

8.3 Summary

The ideas of residual analysis begun in Chapter 3 were considerably expanded in this chapter. We looked at various plots of the residuals, checking the error terms for constant variance, normality, and independence. The properties of the sample autocorrelation of the residuals play a significant role in these diagnostics. The Ljung-Box statistic portmanteau test was discussed as a summary of the autocorrelation in the residuals. Lastly, the ideas of overfitting and parameter redundancy were presented.

EXERCISES

8.1 For an AR(1) model with $\phi \approx 0.5$ and $n = 100$, the lag 1 sample autocorrelation of the residuals is 0.5. Should we consider this unusual? Why or why not?

8.2 Repeat Exercise 8.1 for an MA(1) model with $\theta \approx 0.5$ and $n = 100$.

8.3 Based on a series of length $n = 200$, we fit an AR(2) model and obtain residual autocorrelations of $\hat{r}_1 = 0.13$, $\hat{r}_2 = 0.13$, and $\hat{r}_3 = 0.12$. If $\hat{\phi}_1 = 1.1$ and $\hat{\phi}_2 = -0.8$, do these residual autocorrelations support the AR(2) specification? Individually? Jointly?

8.4 Simulate an AR(1) model with $n = 30$ and $\phi = 0.5$.
- **(a)** Fit the correctly specified AR(1) model and look at a time series plot of the residuals. Does the plot support the AR(1) specification?
- **(b)** Display a normal quantile-quantile plot of the standardized residuals. Does the plot support the AR(1) specification?
- **(c)** Display the sample ACF of the residuals. Does the plot support the AR(1) specification?
- **(d)** Calculate the Ljung-Box statistic summing to $K = 8$. Does this statistic support the AR(1) specification?

8.5 Simulate an MA(1) model with $n = 36$ and $\theta = -0.5$.
- **(a)** Fit the correctly specified MA(1) model and look at a time series plot of the residuals. Does the plot support the MA(1) specification?
- **(b)** Display a normal quantile-quantile plot of the standardized residuals. Does the plot support the MA(1) specification?
- **(c)** Display the sample ACF of the residuals. Does the plot support the MA(1) specification?
- **(d)** Calculate the Ljung-Box statistic summing to $K = 6$. Does this statistic support the MA(1) specification?

8.6 Simulate an AR(2) model with $n = 48$, $\phi_1 = 1.5$, and $\phi_2 = -0.75$.
- **(a)** Fit the correctly specified AR(2) model and look at a time series plot of the residuals. Does the plot support the AR(2) specification?
- **(b)** Display a normal quantile-quantile plot of the standardized residuals. Does the plot support the AR(2) specification?
- **(c)** Display the sample ACF of the residuals. Does the plot support the AR(2) specification?
- **(d)** Calculate the Ljung-Box statistic summing to $K = 12$. Does this statistic support the AR(2) specification?

8.7 Fit an AR(3) model by maximum likelihood to the square root of the hare abundance series (filename hare).
- **(a)** Plot the sample ACF of the residuals. Comment on the size of the correlations.
- **(b)** Calculate the Ljung-Box statistic summing to $K = 9$. Does this statistic support the AR(3) specification?
- **(c)** Perform a runs test on the residuals and comment on the results.
- **(d)** Display the quantile-quantile normal plot of the residuals. Comment on the plot.
- **(e)** Perform the Shapiro-Wilk test of normality on the residuals.

8.8 Consider the oil filter sales data shown in Exhibit 1.8 on page 7. The data are in the file named oilfilters.
- **(a)** Fit an AR(1) model to this series. Is the estimate of the ϕ parameter significantly different from zero statistically?
- **(b)** Display the sample ACF of the residuals from the AR(1) fitted model. Comment on the display.

8.9 The data file named robot contains a time series obtained from an industrial robot. The robot was put through a sequence of maneuvers, and the distance from a desired ending point was recorded in inches. This was repeated 324 times to form the time series. Compare the fits of an AR(1) model and an IMA(1,1) model for these data in terms of the diagnostic tests discussed in this chapter.

8.10 The data file named deere3 contains 57 consecutive values from a complex machine tool at Deere & Co. The values given are deviations from a target value in units of ten millionths of an inch. The process employs a control mechanism that resets some of the parameters of the machine tool depending on the magnitude of deviation from target of the last item produced. Diagnose the fit of an AR(1) model for these data in terms of the tests discussed in this chapter.

8.11 Exhibit 6.31 on page 139, suggested specifying either an AR(1) or possibly an AR(4) model for the difference of the logarithms of the oil price series. (The filename is oil.price).

(a) Estimate both of these models using maximum likelihood and compare the results using the diagnostic tests considered in this chapter.

(b) Exhibit 6.32 on page 140, suggested specifying an MA(1) model for the difference of the logs. Estimate this model by maximum likelihood and perform the diagnostic tests considered in this chapter.

(c) Which of the three models AR(1), AR(4), or MA(1) would you prefer given the results of parts (a) and (b)?

CHAPTER 9

FORECASTING

One of the primary objectives of building a model for a time series is to be able to forecast the values for that series at future times. Of equal importance is the assessment of the precision of those forecasts. In this chapter, we shall consider the calculation of forecasts and their properties for both deterministic trend models and ARIMA models. Forecasts for models that combine deterministic trends with ARIMA stochastic components are considered also.

For the most part, we shall assume that the model is known *exactly*, including specific values for all the parameters. Although this is never true in practice, the use of estimated parameters for large sample sizes does not seriously affect the results.

9.1 Minimum Mean Square Error Forecasting

Based on the available history of the series up to time t, namely $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we would like to forecast the value of $Y_{t+\ell}$ that will occur ℓ time units into the future. We call time t the **forecast origin** and ℓ the **lead time** for the forecast, and denote the forecast itself as $\hat{Y}_t(\ell)$.

As shown in Appendix F, the minimum mean square error forecast is given by

$$\hat{Y}_t(\ell) = E(Y_{t+\ell} \mid Y_1, Y_2, \ldots, Y_t) \tag{9.1.1}$$

(Appendices E and F on page 218 review the properties of conditional expectation and minimum mean square error prediction.)

The computation and properties of this conditional expectation as related to forecasting will be our concern for the remainder of this chapter.

9.2 Deterministic Trends

Consider once more the deterministic trend model of Chapter 3,

$$Y_t = \mu_t + X_t \tag{9.2.1}$$

where the stochastic component, X_t, has a mean of zero. For this section, we shall assume that $\{X_t\}$ is in fact white noise with variance γ_0. For the model in Equation (9.2.1), we have

$$\begin{aligned}\hat{Y}_t(\ell) &= E(\mu_{t+\ell} + X_{t+\ell} | Y_1, Y_2, \ldots, Y_t) \\ &= E(\mu_{t+\ell} | Y_1, Y_2, \ldots, Y_t) + E(X_{t+\ell} | Y_1, Y_2, \ldots, Y_t) \\ &= \mu_{t+\ell} + E(X_{t+\ell})\end{aligned}$$

or

$$\hat{Y}_t(\ell) = \mu_{t+\ell} \tag{9.2.2}$$

since for $\ell \geq 1$, $X_{t+\ell}$ is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ and has expected value zero. Thus, in this simple case, forecasting amounts to extrapolating the deterministic time trend into the future.

For the linear trend case, $\mu_t = \beta_0 + \beta_1 t$, the forecast is

$$\hat{Y}_t(\ell) = \beta_0 + \beta_1(t + \ell) \tag{9.2.3}$$

As we emphasized in Chapter 3, this model assumes that the *same* linear time trend persists into the future, and the forecast reflects that assumption. Note that it is the lack of statistical dependence between $Y_{t+\ell}$ and $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ that prevents us from improving on $\mu_{t+\ell}$ as a forecast.

For seasonal models where, say, $\mu_t = \mu_{t+12}$, our forecast is $\hat{Y}_t(\ell) = \mu_{t+12+\ell} = \hat{Y}_t(\ell + 12)$. Thus the forecast will also be periodic, as desired.

The **forecast error**, $e_t(\ell)$, is given by

$$\begin{aligned}e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell) \\ &= \mu_{t+\ell} + X_{t+\ell} - \mu_{t+\ell} \\ &= X_{t+\ell}\end{aligned}$$

so that

$$E(e_t(\ell)) = E(X_{t+\ell}) = 0$$

That is, the forecasts are **unbiased**. Also

$$Var(e_t(\ell)) = Var(X_{t+\ell}) = \gamma_0 \tag{9.2.4}$$

is the **forecast error variance** for all lead times ℓ.

The cosine trend model for the average monthly temperature series was estimated in Chapter 3 on page 35 as

$$\hat{\mu}_t = 46.2660 + (-26.7079)\cos(2\pi t) + (-2.1697)\sin(2\pi t)$$

Here time is measured in years with a starting value of January 1964, frequency $f = 1$ per year, and the final observed value is for December 1975. To forecast the June 1976 temperature value, we use $t = 1976.41667$ as the time value[†] and obtain

[†] June is the fifth month of the year, and $5/12 \approx 0.416666666\ldots$.

$$\hat{\mu}_t = 46.2660 + (-26.7079)\cos(2\pi(1976.41667)) + (-2.1697)\sin(2\pi(1976.41667))$$
$$= 68.3\ ^\circ\text{F}$$

Forecasts for other months are obtained similarly.

9.3 ARIMA Forecasting

For ARIMA models, the forecasts can be expressed in several different ways. Each expression contributes to our understanding of the overall forecasting procedure with respect to computing, updating, assessing precision, or long-term forecasting behavior.

AR(1)

We shall first illustrate many of the ideas with the simple AR(1) process with a nonzero mean that satisfies

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t \tag{9.3.1}$$

Consider the problem of forecasting one time unit into the future. Replacing t by $t + 1$ in Equation (9.3.1), we have

$$Y_{t+1} - \mu = \phi(Y_t - \mu) + e_{t+1} \tag{9.3.2}$$

Given $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we take the conditional expectations of both sides of Equation (9.3.2) and obtain

$$\hat{Y}_t(1) - \mu = \phi[E(Y_t|Y_1, Y_2, \ldots, Y_t) - \mu] + E(e_{t+1}|Y_1, Y_2, \ldots, Y_t) \tag{9.3.3}$$

Now, from the properties of conditional expectation, we have

$$E(Y_t|Y_1, Y_2, \ldots, Y_t) = Y_t \tag{9.3.4}$$

Also, since e_{t+1} is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we obtain

$$E(e_{t+1}|Y_1, Y_2, \ldots, Y_t) = E(e_{t+1}) = 0 \tag{9.3.5}$$

Thus, Equation (9.3.3) can be written as

$$\hat{Y}_t(1) = \mu + \phi(Y_t - \mu) \tag{9.3.6}$$

In words, a proportion ϕ of the current deviation from the process mean is added to the process mean to forecast the next process value.

Now consider a general lead time ℓ. Replacing t by $t + \ell$ in Equation (9.3.1) and taking the conditional expectations of both sides produces

$$\hat{Y}_t(\ell) = \mu + \phi[\hat{Y}_t(\ell - 1) - \mu] \quad \text{for } \ell \geq 1 \tag{9.3.7}$$

since $E(Y_{t+\ell-1}|Y_1, Y_2, \ldots, Y_t) = \hat{Y}_t(\ell - 1)$ and, for $\ell \geq 1$, $e_{t+\ell}$ is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$.

Equation (9.3.7), which is recursive in the lead time ℓ, shows how the forecast for any lead time ℓ can be built up from the forecasts for shorter lead times by starting with the initial forecast $\hat{Y}_t(1)$ computed using Equation (9.3.6). The forecast $\hat{Y}_t(2)$ is then obtained from $\hat{Y}_t(2) = \mu + \phi[\hat{Y}_t(1) - \mu]$, then $\hat{Y}_t(3)$ from $\hat{Y}_t(2)$, and so on until the desired $\hat{Y}_t(\ell)$ is found. Equation (9.3.7) and its generalizations for other ARIMA models are most convenient for actually computing the forecasts. Equation (9.3.7) is sometimes called the **difference equation form** of the forecasts.

However, Equation (9.3.7) can also be solved to yield an explicit expression for the forecasts in terms of the observed history of the series. Iterating backward on ℓ in Equation (9.3.7), we have

$$\begin{aligned}\hat{Y}_t(\ell) &= \phi[\hat{Y}_t(\ell-1) - \mu] + \mu \\ &= \phi\{\phi[\hat{Y}_t(\ell-2) - \mu]\} + \mu \\ &\vdots \\ &= \phi^{\ell-1}[\hat{Y}_t(1) - \mu] + \mu\end{aligned}$$

or

$$\hat{Y}_t(\ell) = \mu + \phi^{\ell}(Y_t - \mu) \tag{9.3.8}$$

The current deviation from the mean is discounted by a factor ϕ^{ℓ}, whose magnitude decreases with increasing lead time. The discounted deviation is then added to the process mean to produce the lead ℓ forecast.

As a numerical example, consider the AR(1) model that we have fitted to the industrial color property time series. The maximum likelihood estimation results were partially shown in Exhibit 7.7 on page 165, but more complete results are shown in Exhibit 9.1.

Exhibit 9.1 Maximum Likelihood Estimation of an AR(1) Model for Color

Coefficients:	ar1	intercept[†]
	0.5705	74.3293
s.e.	0.1435	1.9151

sigma^2 estimated as 24.8: log-likelihood = –106.07, AIC = 216.15

[†]Remember that the intercept here is the estimate of the process mean μ—not θ_0.

```
> data(color)
> m1.color=arima(color,order=c(1,0,0))
> m1.color
```

For illustration purposes, we assume that the estimates $\phi = 0.5705$ and $\mu = 74.3293$ are true values. The final forecasts may then be rounded.

The last observed value of the color property is 67, so we would forecast one time period ahead as[†]

$$\begin{aligned}\hat{Y}_t(1) &= 74.3293 + (0.5705)(67 - 74.3293)\\ &= 74.3293 - 4.181366\\ &= 70.14793\end{aligned}$$

For lead time 2, we have from Equation (9.3.7)

$$\begin{aligned}\hat{Y}_t(2) &= 74.3293 + 0.5705(70.14793 - 74.3293)\\ &= 74.3293 - 2.385472\\ &= 71.94383\end{aligned}$$

Alternatively, we can use Equation (9.3.8):

$$\begin{aligned}\hat{Y}_t(2) &= 74.3293 + (0.5705)^2(67 - 74.3293)\\ &= 71.92823\end{aligned}$$

At lead 5, we have

$$\begin{aligned}\hat{Y}_t(5) &= 74.3293 + (0.5705)^5(67 - 74.3293)\\ &= 73.88636\end{aligned}$$

and by lead 10 the forecast is

$$\hat{Y}_t(10) = 74.30253$$

which is very nearly μ (= 74.3293). In reporting these forecasts we would probably round to the nearest tenth.

In general, since $|\phi| < 1$, we have simply

$$\hat{Y}_t(\ell) \approx \mu \quad \text{for large } \ell \tag{9.3.9}$$

Later we shall see that Equation (9.3.9) holds for *all stationary* ARMA models.

Consider now the **one-step-ahead forecast error**, $e_t(1)$. From Equations (9.3.2) and (9.3.6), we have

$$\begin{aligned}e_t(1) &= Y_{t+1} - \hat{Y}_t(1)\\ &= [\phi(Y_t - \mu) + \mu + e_{t+1}] - [\phi(Y_t - \mu) + \mu]\end{aligned}$$

or

$$e_t(1) = e_{t+1} \tag{9.3.10}$$

[†] As round off error will accumulate, you should use many decimal places when performing recursive calculations.

The white noise process $\{e_t\}$ can now be reinterpreted as a sequence of one-step-ahead forecast errors. We shall see that Equation (9.3.10) persists for completely general ARIMA models. Note also that Equation (9.3.10) implies that the forecast error $e_t(1)$ is independent of the history of the process $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ up to time t. If this were not so, the dependence could be exploited to improve our forecast.

Equation (9.3.10) also implies that our one-step-ahead forecast error variance is given by

$$Var(e_t(1)) = \sigma_e^2 \tag{9.3.11}$$

To investigate the properties of the forecast errors for longer leads, it is convenient to express the AR(1) model in general linear process, or MA(∞), form. From Equation (4.3.8) on page 70, we recall that

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots \tag{9.3.12}$$

Then Equations (9.3.8) and (9.3.12) together yield

$$\begin{aligned} e_t(\ell) &= Y_{t+\ell} - \mu - \phi^{\ell}(Y_t - \mu) \\ &= e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1} + \phi^{\ell} e_t \\ &\qquad + \cdots - \phi^{\ell}(e_t + \phi e_{t-1} + \cdots) \end{aligned}$$

so that

$$e_t(\ell) = e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1} \tag{9.3.13}$$

which can also be written as

$$e_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1} \tag{9.3.14}$$

Equation (9.3.14) will be shown to hold for *all* ARIMA models (see Equation (9.3.43) on page 202).

Note that $E(e_t(\ell)) = 0$; thus the forecasts are **unbiased**. Furthermore, from Equation (9.3.14), we have

$$Var(e_t(\ell)) = \sigma_e^2(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{\ell-1}^2) \tag{9.3.15}$$

We see that the forecast error variance increases as the lead ℓ increases. Contrast this with the result given in Equation (9.2.4) on page 192, for deterministic trend models.

In particular, for the AR(1) case,

$$Var(e_t(\ell)) = \sigma_e^2\left[\frac{1 - \phi^{2\ell}}{1 - \phi^2}\right] \tag{9.3.16}$$

which we obtain by summing a finite geometric series.

For long lead times, we have

$$Var(e_t(\ell)) \approx \frac{\sigma_e^2}{1-\phi^2} \text{ for large } \ell \tag{9.3.17}$$

or, by Equation (4.3.3), page 66,

$$Var(e_t(\ell)) \approx Var(Y_t) = \gamma_0 \text{ for large } \ell \tag{9.3.18}$$

Equation (9.3.18) will be shown to be valid for *all stationary* ARMA processes (see Equation (9.3.39) on page 201).

MA(1)

To illustrate how to solve the problems that arise in forecasting moving average or mixed models, consider the MA(1) case with nonzero mean:

$$Y_t = \mu + e_t - \theta e_{t-1}$$

Again replacing t by $t+1$ and taking conditional expectations of both sides, we have

$$\hat{Y}_t(1) = \mu - \theta E(e_t | Y_1, Y_2, \ldots, Y_t) \tag{9.3.19}$$

However, for an invertible model, Equation (4.5.2) on page 80 shows that e_t is a function of $Y_1, Y_2, \ldots, Y_t$ and so

$$E(e_t | Y_1, Y_2, \ldots, Y_t) = e_t \tag{9.3.20}$$

In fact, an approximation is involved in this equation since we are conditioning only on $Y_1, Y_2, \ldots, Y_t$ and not on the infinite history of the process. However, if, as in practice, t is large and the model is invertible, the error in the approximation will be very small. If the model is not invertible—for example, if we have overdifferenced the data—then Equation (9.3.20) is not even approximately valid; see Harvey (1981c, p.161).

Using Equations (9.3.19) and (9.3.20), we have the one-step-ahead forecast for an invertible MA(1) expressed as

$$\hat{Y}_t(1) = \mu - \theta e_t \tag{9.3.21}$$

The computation of e_t will be a by-product of estimating the parameters in the model.

Notice once more that the one-step-ahead forecast error is

$$\begin{aligned} e_t(1) &= Y_{t+1} - \hat{Y}_t(1) \\ &= (\mu + e_{t+1} - \theta e_t) - (\mu - \theta e_t) \\ &= e_{t+1} \end{aligned}$$

as in Equation (9.3.10), and thus Equation (9.3.11) also obtains.

For longer lead times, we have

$$\hat{Y}_t(\ell) = \mu + E(e_{t+\ell} | Y_1, Y_2, \ldots, Y_t) - \theta E(e_{t+\ell-1} | Y_1, Y_2, \ldots, Y_t)$$

But, for $\ell > 1$, both $e_{t+\ell}$ and $e_{t+\ell-1}$ are independent of $Y_1, Y_2, \ldots, Y_t$. Consequently, these conditional expected values are the unconditional expected values, namely zero, and we have

$$\hat{Y}_t(\ell) = \mu \quad \text{for } \ell > 1 \tag{9.3.22}$$

Notice here that Equation (9.3.9) on page 195 holds exactly for the MA(1) case when $\ell > 1$. Since for this model we trivially have $\psi_1 = -\theta$ and $\psi_j = 0$ for $j > 1$, Equations (9.3.14) and (9.3.15) also hold.

The Random Walk with Drift

To illustrate forecasting with nonstationary ARIMA series, consider the random walk with drift defined by

$$Y_t = Y_{t-1} + \theta_0 + e_t \tag{9.3.23}$$

Here

$$\hat{Y}_t(1) = E(Y_t | Y_1, Y_2, \ldots, Y_t) + \theta_0 + E(e_{t+1} | Y_1, Y_2, \ldots, Y_t)$$

so that

$$\hat{Y}_t(1) = Y_t + \theta_0 \tag{9.3.24}$$

Similarly, the difference equation form for the lead ℓ forecast is

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell - 1) + \theta_0 \quad \text{for } \ell \geq 1 \tag{9.3.25}$$

and iterating backward on ℓ yields the explicit expression

$$\hat{Y}_t(\ell) = Y_t + \theta_0 \ell \quad \text{for } \ell \geq 1 \tag{9.3.26}$$

In contrast to Equation (9.3.9) on page 195, if $\theta_0 \neq 0$, the forecast does not converge for long leads but rather follows a straight line with slope θ_0 for all ℓ.

Note that the presence or absence of the constant term θ_0 significantly alters the nature of the forecast. For this reason, constant terms should not be included in nonstationary ARIMA models unless the evidence is clear that the mean of the differenced series is significantly different from zero. Equation (3.2.3) on page 28 for the variance of the sample mean will help assess this significance.

However, as we have seen in the AR(1) and MA(1) cases, the one-step-ahead forecast error is

$$e_t(1) = Y_{t+1} - \hat{Y}_t(1) = e_{t+1}$$

Also

$$\begin{aligned} e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell) \\ &= (Y_t + \ell\theta_0 + e_{t+1} + \cdots + e_{t+\ell}) - (Y_t + \ell\theta_0) \\ &= e_{t+1} + e_{t+2} + \cdots + e_{t+\ell} \end{aligned}$$

which agrees with Equation (9.3.14) on page 196 since in this model $\psi_j = 1$ for all j. (See Equation (5.2.6) on page 93 with $\theta = 0$.)

So, as in Equation (9.3.15), we have

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 = \ell\sigma_e^2 \tag{9.3.27}$$

In contrast to the stationary case, here $Var(e_t(\ell))$ grows without limit as the forecast lead time ℓ increases. We shall see that this property is characteristic of the forecast error variance for all *nonstationary* ARIMA processes.

ARMA(*p*,*q*)

For the general stationary ARMA(p,q) model, the difference equation form for computing forecasts is given by

$$\begin{aligned} \hat{Y}_t(\ell) = {} & \phi_1\hat{Y}_t(\ell-1) + \phi_2\hat{Y}_t(\ell-2) + \cdots + \phi_p\hat{Y}_t(\ell-p) + \theta_0 \\ & - \theta_1 E(e_{t+\ell-1}|Y_1, Y_2, \ldots, Y_t) - \theta_2 E(e_{t+\ell-2}|Y_1, Y_2, \ldots, Y_t) \\ & - \cdots - \theta_q E(e_{t+\ell-q}|Y_1, Y_2, \ldots, Y_t) \end{aligned} \tag{9.3.28}$$

where

$$E(e_{t+j}|Y_1, Y_2, \ldots, Y_t) = \begin{cases} 0 & \text{for } j > 0 \\ e_{t+j} & \text{for } j \le 0 \end{cases} \tag{9.3.29}$$

We note that $\hat{Y}_t(j)$ is a true forecast for $j > 0$, but for $j \le 0$, $\hat{Y}_t(j) = Y_{t+j}$. As in Equation (9.3.20) on page 197, Equation (9.3.29) involves some minor approximation. For an invertible model, Equation (4.5.5) on page 80 shows that, using the π-weights, e_t can be expressed as a linear combination of the infinite sequence Y_t, Y_{t-1}, Y_{t-2},.... However, the π-weights die out exponentially fast, and the approximation assumes that π_j is negligible for $j > t - q$.

As an example, consider an ARMA(1,1) model. We have

$$\hat{Y}_t(1) = \phi Y_t + \theta_0 - \theta e_t \tag{9.3.30}$$

with

$$\hat{Y}_t(2) = \phi\hat{Y}_t(1) + \theta_0$$

and, more generally,

$$\hat{Y}_t(\ell) = \phi \hat{Y}_t(\ell - 1) + \theta_0 \quad \text{for } \ell \geq 2 \tag{9.3.31}$$

using Equation (9.3.30) to get the recursion started.

Equations (9.3.30) and (9.3.31) can be rewritten in terms of the process mean and then solved by iteration to get the alternative explicit expression

$$\hat{Y}_t(\ell) = \mu + \phi^{\ell}(Y_t - \mu) - \phi^{\ell - 1} e_t \quad \text{for } \ell \geq 1 \tag{9.3.32}$$

As Equations (9.3.28) and (9.3.29) indicate, the noise terms $e_{t-(q-1)}, \ldots, e_{t-1}, e_t$ appear directly in the computation of the forecasts for leads $\ell = 1, 2, \ldots, q$. However, for $\ell > q$, the autoregressive portion of the difference equation takes over, and we have

$$\hat{Y}_t(\ell) = \phi_1 \hat{Y}_t(\ell - 1) + \phi_2 \hat{Y}_t(\ell - 2) + \cdots + \phi_p \hat{Y}_t(\ell - p) + \theta_0 \quad \text{for } \ell > q \tag{9.3.33}$$

Thus the general nature of the forecast for long lead times will be determined by the autoregressive parameters $\phi_1, \phi_2, \ldots, \phi_p$ (and the constant term, θ_0, which is related to the mean of the process).

Recalling from Equation (5.3.17) on page 97 that $\theta_0 = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$, we can rewrite Equation (9.3.33) in terms of deviations from μ as

$$\begin{aligned} \hat{Y}_t(\ell) - \mu = {} & \phi_1[\hat{Y}_t(\ell - 1) - \mu] + \phi_2[\hat{Y}_t(\ell - 2) - \mu] + \cdots \\ & + \phi_p[\hat{Y}_t(\ell - p) - \mu] \quad \text{for } \ell > q \end{aligned} \tag{9.3.34}$$

As a function of lead time ℓ, $\hat{Y}_t(\ell) - \mu$ follows the same Yule-Walker recursion as the autocorrelation function ρ_k of the process (see Equation (4.4.8), page 79). Thus, as in Section 4.3 on page 66 and Section 4.4 on page 77, the roots of the characteristic equation will determine the general behavior of $\hat{Y}_t(\ell) - \mu$ for large lead times. In particular, $\hat{Y}_t(\ell) - \mu$ can be expressed as a linear combination of exponentially decaying terms in ℓ (corresponding to the real roots) and damped sine wave terms (corresponding to the pairs of complex roots).

Thus, for any stationary ARMA model, $\hat{Y}_t(\ell) - \mu$ decays to zero as ℓ increases, and the long-term forecast is simply the process mean μ as given in Equation (9.3.9) on page 195. This agrees with common sense since for stationary ARMA models the dependence dies out as the time span between observations increases, and this dependence is the only reason we can improve on the "naive" forecast of using μ alone.

To argue the validity of Equation (9.3.15) for $e_t(\ell)$ in the present generality, we need to consider a new representation for ARIMA processes. Appendix G shows that any ARIMA model can be written in **truncated linear process** form as

$$Y_{t+\ell} = C_t(\ell) + I_t(\ell) \quad \text{for } \ell > 1 \tag{9.3.35}$$

where, for our present purposes, we need only know that $C_t(\ell)$ is a certain function of Y_t, $Y_{t-1}, \ldots$ and

$$I_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1} \quad \text{for } \ell \geq 1 \tag{9.3.36}$$

Furthermore, for invertible models with t reasonably large, $C_t(\ell)$ is a certain function of the finite history $Y_t, Y_{t-1},\ldots, Y_1$. Thus we have

$$\begin{aligned}\hat{Y}_t(\ell) &= E(C_t(\ell)|Y_1, Y_2, \ldots, Y_t) + E(I_t(\ell)|Y_1, Y_2, \ldots, Y_t)\\ &= C_t(\ell)\end{aligned}$$

Finally,

$$\begin{aligned}e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell)\\ &= [C_t(\ell) + I_t(\ell)] - C_t(\ell)\\ &= I_t(\ell)\\ &= e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1}\end{aligned}$$

Thus, for a general invertible ARIMA process,

$$E[e_t(\ell)] = 0 \quad \text{for } \ell \geq 1 \tag{9.3.37}$$

and

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 \quad \text{for } \ell \geq 1 \tag{9.3.38}$$

From Equations (4.1.4) and (9.3.38), we see that for long lead times in stationary ARMA models, we have

$$Var(e_t(\ell)) \approx \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2$$

or

$$Var(e_t(\ell)) \approx \gamma_0 \quad \text{for large } \ell \tag{9.3.39}$$

Nonstationary Models

As the random walk shows, forecasting for nonstationary ARIMA models is quite similar to forecasting for stationary ARMA models, but there are some striking differences. Recall from Equation (5.2.2) on page 92 that an ARIMA(p,1,q) model can be written as a nonstationary ARMA(p+1,q) model, We shall write this as

$$\begin{aligned}Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \cdots + \varphi_p Y_{t-p} + \varphi_{p+1} Y_{t-p-1}\\ + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}\end{aligned} \tag{9.3.40}$$

where the script coefficients φ are directly related to the block ϕ coefficients. In particular,

$$\left.\begin{aligned} &\varphi_1 = 1 + \phi_1, \varphi_j = \phi_j - \phi_{j-1} \text{ for } j = 1, 2, \ldots, p \\ &\text{and} \\ &\varphi_{p+1} = -\phi_p \end{aligned}\right\} \tag{9.3.41}$$

For a general order of differencing d, we would have $p + d$ of the φ coefficients.

From this representation, we can immediately extend Equations (9.3.28), (9.3.29), and (9.3.30) on page 199 to cover the nonstationary cases by replacing p by $p + d$ and ϕ_j by φ_j.

As an example of the necessary calculations, consider the ARIMA(1,1,1) case. Here

$$Y_t - Y_{t-1} = \phi(Y_{t-1} - Y_{t-2}) + \theta_0 + e_t - \theta e_{t-1}$$

so that

$$Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + \theta_0 + e_t - \theta e_{t-1}$$

Thus

$$\left.\begin{aligned} &\hat{Y}_t(1) = (1 + \phi)Y_t - \phi Y_{t-1} + \theta_0 - \theta e_t \\ &\hat{Y}_t(2) = (1 + \phi)\hat{Y}_t(1) - \phi Y_t + \theta_0 \\ &\text{and} \\ &\hat{Y}_t(\ell) = (1 + \phi)\hat{Y}_t(\ell - 1) - \phi \hat{Y}_t(\ell - 2) + \theta_0 \end{aligned}\right\} \tag{9.3.42}$$

For the general invertible ARIMA model, the truncated linear process representation given in Equations (9.3.35) and (9.3.36) and the calculations following these equations show that we can write

$$e_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1} \text{ for } \ell \geq 1 \tag{9.3.43}$$

and so

$$E(e_t(\ell)) = 0 \text{ for } \ell \geq 1 \tag{9.3.44}$$

and

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 \text{ for } \ell \geq 1 \tag{9.3.45}$$

However, for nonstationary series, the ψ_j-weights do not decay to zero as j increases. For example, for the random walk model, $\psi_j = 1$ for all j; for the IMA(1,1) model, $\psi_j = 1 - \theta$ for $j \geq 1$; for the IMA(2,2) case, $\psi_j = 1 + \theta_2 + (1 - \theta_1 - \theta_2)j$ for $j \geq 1$; and for the ARI(1,1) model, $\psi_j = (1 - \phi^{j+1})/(1 - \phi)$ for $j \geq 1$ (see Chapter 5).

Thus, for any nonstationary model, Equation (9.3.45) shows that the forecast error variance will grow without bound as the lead time ℓ increases. This fact should not be too surprising since with nonstationary series the distant future is quite uncertain.

9.4 Prediction Limits

As in all statistical endeavors, in addition to forecasting or predicting the unknown $Y_{t+\ell}$, we would like to assess the precision of our predictions.

Deterministic Trends

For the deterministic trend model with a white noise stochastic component $\{X_t\}$, we recall that

$$\hat{Y}_t(\ell) = \mu_{t+\ell}$$

and

$$Var(e_t(\ell)) = Var(X_{t+\ell}) = \gamma_0$$

If the stochastic component is normally distributed, then the forecast error

$$e_t(\ell) = Y_{t+\ell} - \hat{Y}_t(\ell) = X_{t+\ell} \tag{9.4.1}$$

is also normally distributed. Thus, for a given confidence level $1 - \alpha$, we could use a standard normal percentile, $z_{1-\alpha/2}$, to claim that

$$P\left[-z_{1-\alpha/2} < \frac{Y_{t+\ell} - \hat{Y}_t(\ell)}{\sqrt{Var(e_t(\ell))}} < z_{1-\alpha/2}\right] = 1 - \alpha$$

or, equivalently,

$$P[\hat{Y}_t(\ell) - z_{1-\alpha/2}\sqrt{Var(e_t(\ell))} < Y_{t+\ell} < \hat{Y}_t(\ell) + z_{1-\alpha/2}\sqrt{Var(e_t(\ell))}] = 1 - \alpha$$

Thus we may be $(1 - \alpha)100\%$ confident that the future observation $Y_{t+\ell}$ will be contained within the prediction limits

$$\hat{Y}_t(\ell) \pm z_{1-\alpha/2}\sqrt{Var(e_t(\ell))} \tag{9.4.2}$$

As a numerical example, consider the monthly average temperature series once more. On page 192, we used the cosine model to predict the June 1976 average temperature as 68.3°F. The estimate of $\sqrt{Var(e_t(\ell))} = \sqrt{\gamma_0}$ for this model is 3.7°F. Thus 95% prediction limits for the average June 1976 temperature are

$$68.3 \pm 1.96(3.7) = 68.3 \pm 7.252 \text{ or } 61.05°\text{F to } 75.55°\text{F}$$

Readers who are familiar with standard regression analysis will recall that since the forecast involves *estimated* regression parameters, the correct forecast error variance is given by $\gamma_0[1 + (1/n) + c_{n,\ell}]$, where $c_{n,\ell}$ is a certain function of the sample size n and the lead time ℓ. However, it may be shown that for the types of trends that we are considering (namely, cosines and polynomials in time) and for large sample sizes n, the $1/n$ and $c_{n,\ell}$ are both negligible relative to 1. For example, with a cosine trend of period 12 over $N = n/12$ years, we have that $c_{n,\ell} = 2/n$; thus the correct forecast error variance is

$\gamma_0[1 + (3/n)]$ rather than our approximate γ_0. For the linear time trend model, it can be shown that $c_{n,\ell} = 3(n + 2\ell - 1)^2/[n(n^2 - 1)] \approx 3/n$ for moderate lead ℓ and large n. Thus, again our approximation seems justified.

ARIMA Models

If the white noise terms $\{e_t\}$ in a general ARIMA series each arise independently from a normal distribution, then from Equation (9.3.43) on page 202, the forecast error $e_t(\ell)$ will also have a normal distribution, and the steps leading to Equation (9.4.2) remain valid. However, in contrast to the deterministic trend model, recall that in the present case

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2$$

In practice, σ_e^2 will be unknown and must be estimated from the observed time series. The necessary ψ-weights are, of course, also unknown since they are certain functions of the unknown ϕ's and θ's. For large sample sizes, these estimations will have little effect on the actual prediction limits given above.

As a numerical example, consider the AR(1) model that we estimated for the industrial color property series. From Exhibit 9.1 on page 194, we use $\phi = 0.5705$, $\mu = 74.3293$, and $\sigma_e^2 = 24.8$. For an AR(1) model, we recall Equation (9.3.16) on page 196

$$Var(e_t(\ell)) = \sigma_e^2 \left[\frac{1 - \phi^{2\ell}}{1 - \phi^2}\right]$$

For a one-step-ahead prediction, we have

$$70.14793 \pm 1.96\sqrt{24.8} = 70.14793 \pm 9.760721 \text{ or } 60.39 \text{ to } 79.91$$

Two steps ahead, we obtain

$$71.86072 \pm 11.88343 \text{ or } 60.71 \text{ to } 83.18$$

Notice that this prediction interval is wider than the previous interval. Forecasting ten steps ahead leads to

$$74.173934 \pm 11.88451 \text{ or } 62.42 \text{ to } 86.19$$

By lead 10, both the forecast and the forecast limits have settled down to their long-lead values.

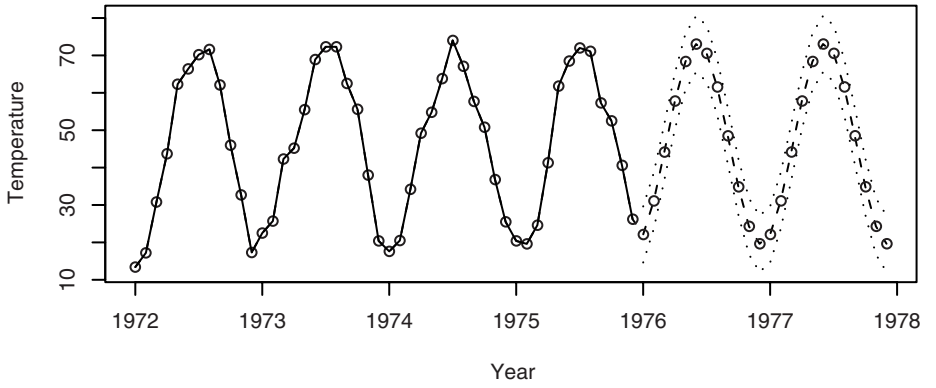
9.5 Forecasting Illustrations

Rather than showing forecast and forecast limit calculations, it is often more instructive to display appropriate plots of the forecasts and their limits.

Deterministic Trends

Exhibit 9.2 displays the last four years of the average monthly temperature time series together with forecasts and 95% forecast limits for two additional years. Since the model fits quite well with a relatively small error variance, the forecast limits are quite close to the fitted trend forecast.

Exhibit 9.2 Forecasts and Limits for the Temperature Cosine Trend

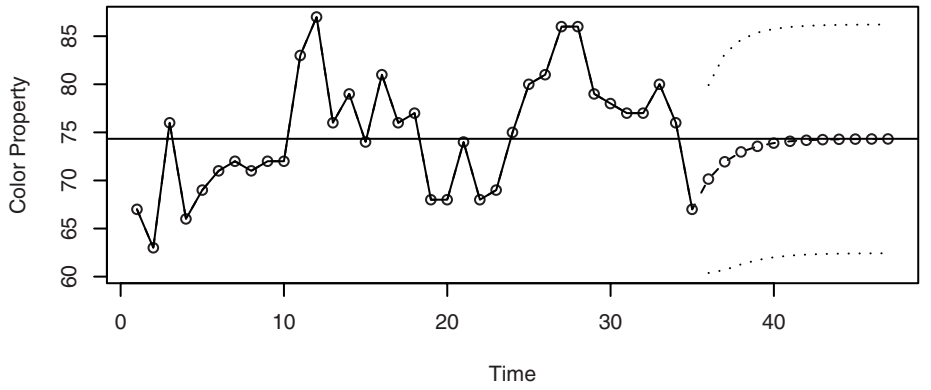



```
> data(tempdub)
> tempdub1=ts(c(tempdub,rep(NA,24)),start=start(tempdub),
    freq=frequency(tempdub))
> har.=harmonic(tempdub,1)
> m5.tempdub=arima(tempdub,order=c(0,0,0),xreg=har.)
> newhar.=harmonic(ts(rep(1,24), start=c(1976,1),freq=12),1)
> win.graph(width=4.875, height=2.5,pointsize=8)
> plot(m5.tempdub,n.ahead=24,n1=c(1972,1),newxreg=newhar.,
    type='b',ylab='Temperature',xlab='Year'))
```

ARIMA Models

We use the industrial color property series as our first illustration of ARIMA forecasting. Exhibit 9.3 displays this series together with forecasts out to lead time 12 with the upper and lower 95% prediction limits for those forecasts. In addition, a horizontal line at the estimate for the process mean is shown. Notice how the forecasts approach the mean exponentially as the lead time increases. Also note how the prediction limits increase in width.

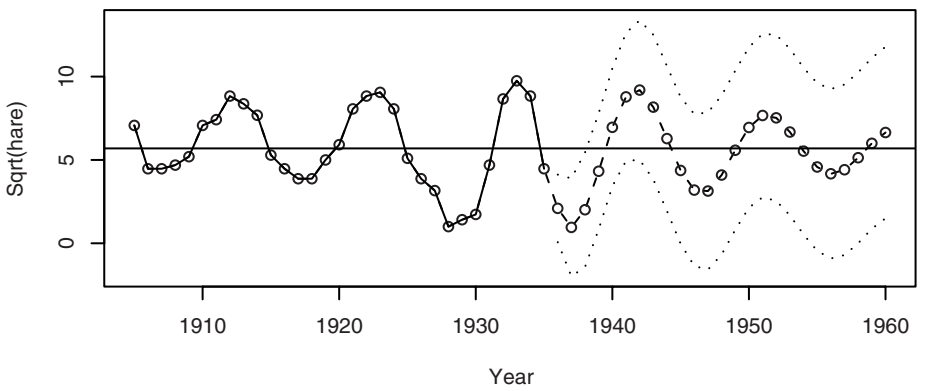
Exhibit 9.3 Forecasts and Forecast Limits for the AR(1) Model for Color




```
> data(color)
> m1.color=arima(color,order=c(1,0,0))
> plot(m1.color,n.ahead=12,type='b',xlab='Time',
    ylab='Color Property')
> abline(h=coef(m1.color)[names(coef(m1.color))=='intercept'])
```

The Canadian hare abundance series was fitted by working with the square root of the abundance numbers and then fitting an AR(3) model. Notice how the forecasts mimic the approximate cycle in the actual series even when we forecast with a lead time out to 25 years in Exhibit 9.4.

Exhibit 9.4 Forecasts from an AR(3) Model for Sqrt(Hare)




```
> data(hare)
> m1.hare=arima(sqrt(hare),order=c(3,0,0))
> plot(m1.hare, n.ahead=25,type='b',
    xlab='Year',ylab='Sqrt(hare)')
> abline(h=coef(m1.hare)[names(coef(m1.hare))=='intercept'])
```

9.6 Updating ARIMA Forecasts

Suppose we are forecasting a monthly time series. Our last observation is, say, for February, and we forecast for March, April, and May. As time goes by, the actual value for March becomes available. With this new value in hand, we would like to update or revise (and, one hopes, improve) our forecasts for April and May. Of course, we could compute new forecasts from scratch. However, there is a simpler way.

For a general forecast origin t and lead time $\ell + 1$, our original forecast is denoted $\hat{Y}_t(\ell + 1)$. Once the observation at time $t + 1$ becomes available, we would like to update our forecast as $\hat{Y}_{t+1}(\ell)$. Equations (9.3.35) and (9.3.36) on page 200 yield

$$Y_{t+\ell+1} = C_t(\ell + 1) + e_{t+\ell+1} + \psi_1 e_{t+\ell} + \psi_2 e_{t+\ell-1} + \cdots + \psi_\ell e_{t+1}$$

Since $C_t(\ell+1)$ and e_{t+1} are functions of $Y_{t+1}, Y_t, \ldots$, whereas $e_{t+\ell+1}, e_{t+\ell}, \ldots, e_{t+2}$ are independent of $Y_{t+1}, Y_t, \ldots$, we quickly obtain the expression

$$\hat{Y}_{t+1}(\ell) = C_t(\ell + 1) + \psi_\ell e_{t+1}$$

However, $\hat{Y}_t(\ell + 1) = C_t(\ell + 1)$, and, of course, $e_{t+1} = Y_{t+1} - \hat{Y}_t(1)$. Thus we have the general **updating equation**

$$\hat{Y}_{t+1}(\ell) = \hat{Y}_t(\ell + 1) + \psi_\ell[Y_{t+1} - \hat{Y}_t(1)] \tag{9.6.1}$$

Notice that $[Y_{t+1} - \hat{Y}_t(1)]$ is the actual forecast error at time $t + 1$ once Y_{t+1} has been observed.

As a numerical example, consider the industrial color property time series. Following Exhibit 9.1 on page 194, we fit an AR(1) model to forecast one step ahead as $\hat{Y}_{35}(1) = 70.096$ and two steps ahead as $\hat{Y}_{35}(2) = 71.86072$. If now the next color value becomes available as $Y_{t+1} = Y_{36} = 65$, then we update the forecast for time $t = 37$ as

$$\hat{Y}_{t+1}(1) = \hat{Y}_{36}(1) = 71.86072 + 0.5705(65 - 70.096) = 68.953452$$

9.7 Forecast Weights and Exponentially Weighted Moving Averages

For ARIMA models without moving average terms, it is clear how the forecasts are explicitly determined from the observed series $Y_t, Y_{t-1}, \ldots, Y_1$. However, for any model with $q > 0$, the noise terms appear in the forecasts, and the nature of the forecasts explicitly in terms of $Y_t, Y_{t-1}, \ldots, Y_1$ is hidden. To bring out this aspect of the forecasts, we return to the inverted form of any invertible ARIMA process, namely

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t$$

(See Equation (4.5.5) on page 80.) Thus we can also write

$$Y_{t+1} = \pi_1 Y_t + \pi_2 Y_{t-1} + \pi_3 Y_{t-2} + \cdots + e_{t+1}$$

Taking conditional expectations of both sides, given Y_t, Y_{t-1}, …, Y_1, we obtain

$$\hat{Y}_t(1) = \pi_1 Y_t + \pi_2 Y_{t-1} + \pi_3 Y_{t-2} + \cdots \tag{9.7.1}$$

(We are assuming the t is sufficiently large and/or that the π-weights die out sufficiently quickly so that π_t, π_{t+1},… are all negligible.)

For any invertible ARIMA model, the π-weights can be calculated recursively from the expressions

$$\pi_j = \begin{cases} \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} + \varphi_j \quad \text{for } 1 \le j \le p+d \\ \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} \quad \text{for } j > p+d \end{cases} \tag{9.7.2}$$

with initial value $\pi_0 = -1$. (Compare this with Equations (4.4.7) on page 79 for the ψ-weights.)

Consider in particular the nonstationary IMA(1,1) model

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1}$$

Here $p = 0$, $d = 1$, $q = 1$, with $\varphi_1 = 1$; thus

$$\pi_1 = \theta\pi_0 + 1 = 1 - \theta$$

$$\pi_2 = \theta\pi_1 = \theta(1-\theta)$$

and, generally,

$$\pi_j = \theta\pi_{j-1} \quad \text{for } j > 1$$

Thus we have explicitly

$$\pi_j = (1-\theta)\theta^{j-1} \quad \text{for } j \ge 1 \tag{9.7.3}$$

so that, from Equation (9.7.1), we can write

$$\hat{Y}_t(1) = (1-\theta)Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)\theta^2 Y_{t-2} + \cdots \tag{9.7.4}$$

In this case, the π-weights *decrease exponentially*, and furthermore,

$$\sum_{j=1}^{\infty} \pi_j = (1-\theta) \sum_{j=1}^{\infty} \theta^{j-1} = \frac{1-\theta}{1-\theta} = 1$$

Thus $\hat{Y}_t(1)$ is called an **exponentially weighted moving average** (**EWMA**).

Simple algebra shows that we can also write

$$\hat{Y}_t(1) = (1-\theta)Y_t + \theta\hat{Y}_{t-1}(1) \tag{9.7.5}$$

and

$$\hat{Y}_t(1) = \hat{Y}_{t-1}(1) + (1-\theta)[Y_t - \hat{Y}_{t-1}(1)] \tag{9.7.6}$$

Equations (9.7.5) and (9.7.6) show how to update forecasts from origin $t-1$ to origin t, and they express the result as a linear combination of the new observation and the old forecast or in terms of the old forecast and the last observed forecast error.

Using EWMA to forecast time series has been advocated, mostly on an ad hoc basis, for a number of years; see Brown (1962) and Montgomery and Johnson (1976).

The parameter $1-\theta$ is called the **smoothing constant** in EWMA literature, and its selection (estimation) is often quite arbitrary. From the ARIMA model-building approach, we let the data indicate whether an IMA(1,1) model is appropriate for the series under consideration. If so, we then estimate θ in an efficient manner and compute an EWMA forecast that we are confident is the minimum mean square error forecast. A comprehensive treatment of exponential smoothing methods and their relationships with ARIMA models is given in Abraham and Ledolter (1983).

9.8 Forecasting Transformed Series

Differencing

Suppose we are interested in forecasting a series whose model involves a first difference to achieve stationarity. Two methods of forecasting can be considered:

1. forecasting the original nonstationary series, for example by using the difference equation form of Equation (9.3.28) on page 199, with ϕ's replaced by φ's throughout, or
2. forecasting the stationary differenced series $W_t = Y_t - Y_{t-1}$ and then "undoing" the difference by summing to obtain the forecast in original terms.

We shall show that both methods lead to the same forecasts. This follows essentially because differencing is a *linear* operation and because conditional expectation of a linear combination is the same linear combination of the conditional expectations.

Consider in particular the IMA(1,1) model. Basing our work on the original nonstationary series, we forecast as

$$\hat{Y}_t(1) = Y_t - \theta e_t \tag{9.8.1}$$

and

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell-1) \quad \text{for } \ell > 1 \tag{9.8.2}$$

Consider now the differenced stationary MA(1) series $W_t = Y_t - Y_{t-1}$. We would forecast $W_{t+\ell}$ as

$$\hat{W}_t(1) = -\theta e_t \tag{9.8.3}$$

and

$$\hat{W}_t(\ell) = 0 \quad \text{for } \ell > 1 \tag{9.8.4}$$

However, $\hat{W}_t(1) = \hat{Y}_t(1) - Y_t$; thus $\hat{W}_t(1) = -\theta e_t$ is equivalent to $\hat{Y}_t(1) = Y_t - \theta e_t$ as before. Similarly, $\hat{W}_t(\ell) = \hat{Y}_t(\ell) - \hat{Y}_t(\ell - 1)$, and Equation (9.8.4) becomes Equation (9.8.2), as we have claimed.

The same result would apply to any model involving differences of any order and indeed to any type of *linear* transformation with constant coefficients. (Certain linear transformations other than differencing may be applicable to seasonal time series. See Chapter 10.)

Log Transformations

As we saw earlier, it is frequently appropriate to model the logarithms of the original series—a nonlinear transformation. Let Y_t denote the original series value and let $Z_t = \log(Y_t)$. It can be shown that we always have

$$E(Y_{t+\ell} | Y_t, Y_{t-1}, \ldots, Y_1) \geq \exp[E(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1)] \tag{9.8.5}$$

with equality holding only in trivial cases. Thus, the naive forecast $\exp[\hat{Z}_t(\ell)]$ is *not* the minimum mean square error forecast of $Y_{t+\ell}$. To evaluate the minimum mean square error forecast in original terms, we shall find the following fact useful: If X has a normal distribution with mean μ and variance σ^2, then

$$E[\exp(X)] = \exp\left[\mu + \frac{\sigma^2}{2}\right]$$

(This follows, for example, from the moment-generating function for X.) In our application

$$\mu = E(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1)$$

and

$$\begin{aligned}
\sigma^2 &= Var(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1) \\
&= Var[e_t(\ell) + C_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] + Var[C_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell)]
\end{aligned}$$

These follow from Equations (9.3.35) and (9.3.36) (applied to Z_t) and the fact that $C_t(l)$ is a function of $Z_t, Z_{t-1}, \ldots$, whereas $e_t(\ell)$ is independent of $Z_t, Z_{t-1}, \ldots$. Thus the minimum mean square error forecast in the original series is given by

$$\exp\left\{\hat{Z}_t(\ell) + \frac{1}{2} Var[e_t(\ell)]\right\} \tag{9.8.6}$$

Throughout our discussion of forecasting, we have assumed that minimum mean square forecast error is the criterion of choice. For normally distributed variables, this is an

excellent criterion. However, if Z_t has a normal distribution, then $Y_t = \exp(Z_t)$ has a log-normal distribution, for which a different criterion may be desirable. In particular, since the log-normal distribution is asymmetric and has a long right tail, a criterion based on the mean absolute error may be more appropriate. For this criterion, the optimal forecast is the **median** of the distribution of $Z_{t+\ell}$ conditional on $Z_t, Z_{t-1},\ldots, Z_1$. Since the log transformation preserves medians and since, for a normal distribution, the mean and median are identical, the naive forecast $\exp[\hat{Z}_t(\ell)]$ is the optimal forecast for $Y_{t+\ell}$ in the sense that it minimizes the mean absolute forecast error.

9.9 Summary of Forecasting with Certain ARIMA Models

Here we bring together various forecasting results for special ARIMA models.

AR(1): $Y_t = \mu + \phi(Y_{t-1} - \mu) + e_t$

$$\hat{Y}_t(\ell) = \mu + \phi[\hat{Y}_t(\ell-1) - \mu] \quad \text{for } \ell \geq 1$$

$$= \mu + \phi^{\ell}(Y_t - \mu) \quad \text{for } l \geq 1$$

$$\hat{Y}_t(\ell) \approx \mu \quad \text{for large } \ell$$

$$e_t(\ell) = e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1}$$

$$Var(e_t(\ell)) = \sigma_e^2 \left[\frac{1-\phi^{2\ell}}{1-\phi^2}\right]$$

$$Var(e_t(\ell)) \approx \frac{\sigma_e^2}{1-\phi^2} = \gamma_0 \quad \text{for large } \ell$$

$$\psi_j = \phi^j \quad \text{for } j > 0$$

MA(1): $Y_t = \mu + e_t - \theta e_{t-1}$

$$\hat{Y}_t(1) = \mu - \theta e_t$$

$$\hat{Y}_t(\ell) = \mu \quad \text{for } l > 1$$

$$e_t(1) = e_{t+1}$$

$$e_t(\ell) = e_{t+\ell} - \theta e_{t+\ell-1} \quad \text{for } \ell > 1$$

$$Var(e_t(\ell)) = \begin{cases} \sigma_e^2 & \text{for } \ell = 1 \\ \sigma_e^2(1+\theta^2) & \text{for } \ell > 1 \end{cases}$$

$$\psi_j = \begin{cases} -\theta & \text{for } j = 1 \\ 0 & \text{for } j > 1 \end{cases}$$

IMA (1,1) with Constant Term: $Y_t = Y_{t-1} + \theta_0 + e_t - \theta e_{t-1}$

$$\begin{aligned} \hat{Y}_t(\ell) &= \hat{Y}_t(\ell-1) + \theta_0 - \theta e_t \\ &= Y_t + \ell\theta_0 - \theta e_t \end{aligned}$$

$$\hat{Y}_t(1) = (1-\theta)Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)\theta^2 Y_{t-2} + \cdots (\text{the EWMA for } \theta_0 = 0)$$

$$e_t(\ell) = e_{t+\ell} + (1-\theta)e_{t+\ell-1} + (1-\theta)e_{t+\ell-2} + \cdots + (1-\theta)e_{t+1} \quad \text{for } \ell \geq 1$$

$$Var(e_t(\ell)) = \sigma_e^2[1 + (\ell-1)(1-\theta)^2]$$

$$\psi_j = 1 - \theta \quad \text{for } j > 0$$

Note that if $\theta_0 \neq 0$, the forecasts follow a straight line with slope θ_0, but if $\theta_0 = 0$, which is the usual case, then the forecast is the same for all lead times, namely

$$\hat{Y}_t(\ell) = Y_t - \theta e_t$$

IMA(2,2): $Y_t = 2Y_{t-1} - Y_{t-2} + \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$

$$\left.\begin{aligned} \hat{Y}_t(1) &= 2Y_t - Y_{t-1} + \theta_0 - \theta_1 e_t - \theta_2 e_{t-1} \\ \hat{Y}_t(2) &= 2\hat{Y}_t(1) - Y_t + \theta_0 - \theta_2 e_t \\ \hat{Y}_t(\ell) &= 2\hat{Y}_t(\ell-1) - \hat{Y}_t(\ell-2) + \theta_0 \quad \text{for } \ell > 2 \end{aligned}\right\} \tag{9.9.1}$$

$$\hat{Y}_t(\ell) = A + B\ell + \frac{\theta_0}{2}\ell^2 \tag{9.9.2}$$

where

$$A = 2\hat{Y}_t(1) - \hat{Y}_t(2) + \theta_0 \tag{9.9.3}$$

and

$$B = \hat{Y}_t(2) - \hat{Y}_t(1) - \frac{3}{2}\theta_0 \tag{9.9.4}$$

If $\theta_0 \neq 0$, the forecasts follow a quadratic curve in ℓ, but if $\theta_0 = 0$, the forecasts form a straight line with slope $\hat{Y}_t(2) - \hat{Y}_t(1)$ and will pass through the two initial forecasts $\hat{Y}_t(1)$ and $\hat{Y}_t(2)$. It can be shown that $Var(e_t(\ell))$ is a certain cubic function of ℓ; see Box, Jenkins, and Reinsel (1994, p. 156). We also have

$$\psi_j = 1 + \theta_2 + (1 - \theta_1 - \theta_2)j \quad \text{for } j > 0 \tag{9.9.5}$$

It can also be shown that forecasting the special case with $\theta_1 = 2\omega$ and $\theta_2 = -\omega^2$ is equivalent to so-called **double exponential smoothing** with smoothing constant $1 - \omega$; see Abraham and Ledolter (1983).

9.10 Summary

Forecasting or predicting future as yet unobserved values is one of the main reasons for developing time series models. Methods discussed in this chapter are all based on minimizing the mean square forecasting error. When the model is simply deterministic trend plus zero mean white noise error, forecasting amounts to extrapolating the trend. However, if the model contains autocorrelation, the forecasts exploit the correlation to produce better forecasts than would otherwise be obtained. We showed how to do this with ARIMA models and investigated the computation and properties of the forecasts. In special cases, the computation and properties of the forecasts are especially interesting and we presented them separately. Prediction limits are especially important to assess the potential accuracy (or otherwise) of the forecasts. Finally, we addressed the problem of forecasting time series for which the models involve transformation of the original series.

EXERCISES

9.1 For an AR(1) model with $Y_t = 12.2$, $\phi = -0.5$, and $\mu = 10.8$,
- **(a)** Find $\hat{Y}_t(1)$.
- **(b)** Calculate $\hat{Y}_t(2)$ in two different ways.
- **(c)** Calculate $\hat{Y}_t(10)$.

9.2 Suppose that annual sales (in millions of dollars) of the Acme Corporation follow the AR(2) model $Y_t = 5 + 1.1Y_{t-1} - 0.5Y_{t-2} + e_t$ with $\sigma_e^2 = 2$.
- **(a)** If sales for 2005, 2006, and 2007 were \$9 million, \$11 million, and \$10 million, respectively, forecast sales for 2008 and 2009.
- **(b)** Show that $\psi_1 = 1.1$ for this model.
- **(c)** Calculate 95% prediction limits for your forecast in part (a) for 2006.
- **(d)** If sales in 2006 turn out to be \$12 million, update your forecast for 2007.

9.3 Using the estimated cosine trend on page 192:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.719°F.)
- **(c)** What is the forecast for April, 1977? For April 2009?

9.4 Using the estimated cosine trend on page 192:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for May 1976.
- **(b)** Find a 95% prediction interval for that May 1976 forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.719°F.)

9.5 Using the seasonal means model *without* an intercept shown in Exhibit 3.3 on page 32:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April, 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)
- **(c)** Compare your forecast with the one obtained in Exercise 9.3.
- **(d)** What is the forecast for April 1977? April 2009?

9.6 Using the seasonal means model *with* an intercept shown in Exhibit 3.4 on page 33:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)
- **(c)** Compare your forecast with the one obtained in Exercise 9.5.

9.7 Using the seasonal means model *with* an intercept shown in Exhibit 3.4 on page 33
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for January 1976.
- **(b)** Find a 95% prediction interval for that January forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)

9.8 Consider the monthly electricity generation time series shown in Exhibit 5.8 on page 99. The data are in the file named electricity.
- **(a)** Fit a deterministic trend model containing seasonal means together with a linear time trend to the logarithms of the electricity values.
- **(b)** Plot the last five years of the series together with two years of forecasts and the 95% forecast limits. Interpret the plot.

9.9 Simulate an AR(1) process with $\phi = 0.8$ and $\mu = 100$. Simulate 48 values but set aside the last 8 values to compare forecasts to actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of ϕ and μ.
- **(b)** Using the estimated model, forecast the next eight values of the series. Plot the series together with the eight forecasts. Place a horizontal line at the estimate of the process mean.
- **(c)** Compare the eight forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and the same sample size.

9.10 Simulate an AR(2) process with $\phi_1 = 1.5$, $\phi_2 = -0.75$, and $\mu = 100$. Simulate 52 values but set aside the last 12 values to compare forecasts to actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of the ϕ's and μ.
- **(b)** Using the estimated model, forecast the next 12 values of the series. Plot the series together with the 12 forecasts. Place a horizontal line at the estimate of

the process mean.

(c) Compare the 12 forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.11 Simulate an MA(1) process with $\theta = 0.6$ and $\mu = 100$. Simulate 36 values but set aside the last 4 values to compare forecasts to actual values.

(a) Using the first 32 values of the series, find the values for the maximum likelihood estimates of the θ and μ.

(b) Using the estimated model, forecast the next four values of the series. Plot the series together with the four forecasts. Place a horizontal line at the estimate of the process mean.

(c) Compare the four forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.12 Simulate an MA(2) process with $\theta_1 = 1$, $\theta_2 = -0.6$, and $\mu = 100$. Simulate 36 values but set aside the last 4 values with compare forecasts to actual values.

(a) Using the first 32 values of the series, find the values for the maximum likelihood estimates of the θ's and μ.

(b) Using the estimated model, forecast the next four values of the series. Plot the series together with the four forecasts. Place a horizontal line at the estimate of the process mean.

(c) What is special about the forecasts at lead times 3 and 4?

(d) Compare the four forecasts with the actual values that you set aside.

(e) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(f) Repeat parts (a) through (e) with a new simulated series using the same values of the parameters and same sample size.

9.13 Simulate an ARMA(1,1) process with $\phi = 0.7$, $\theta = -0.5$, and $\mu = 100$. Simulate 50 values but set aside the last 10 values to compare forecasts with actual values.

(a) Using the first 40 values of the series, find the values for the maximum likelihood estimates of ϕ, θ, and μ.

(b) Using the estimated model, forecast the next ten values of the series. Plot the series together with the ten forecasts. Place a horizontal line at the estimate of the process mean.

(c) Compare the ten forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.14 Simulate an IMA(1,1) process with $\theta = 0.8$ and $\theta_0 = 0$. Simulate 35 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 30 values of the series, find the value for the maximum likelihood estimate of θ.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about the forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.15 Simulate an IMA(1,1) process with $\theta = 0.8$ and $\theta_0 = 10$. Simulate 35 values, but set aside the last five values to compare forecasts to actual values.
- **(a)** Using the first 30 values of the series, find the values for the maximum likelihood estimates of θ and θ_0.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about these forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.16 Simulate an IMA(2,2) process with $\theta_1 = 1$, $\theta_2 = -0.75$, and $\theta_0 = 0$. Simulate 45 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 40 values of the series, find the value for the maximum likelihood estimate of θ_1 and θ_2.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about the forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.17 Simulate an IMA(2,2) process with $\theta_1 = 1$, $\theta_2 = -0.75$, and $\theta_0 = 10$. Simulate 45 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of θ_1, θ_2, and θ_0.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about these forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.18 Consider the model $Y_t = \beta_0 + \beta_1 t + X_t$, where $X_t = \phi X_{t-1} + e_t$. We assume that β_0, β_1, and ϕ are known. Show that the minimum mean square error forecast ℓ steps ahead can be written as $\hat{Y}_t(\ell) = \beta_0 + \beta_1(t+\ell) + \phi^{\ell}(Y_t - \beta_0 - \beta_1 t)$.

9.19 Verify Equation (9.3.16) on page 196.

9.20 Verify Equation (9.3.32) on page 200.

9.21 The data file named deere3 contains 57 consecutive values from a complex machine tool process at Deere & Co. The values given are deviations from a target value in units of ten millionths of an inch. The process employs a control mechanism that resets some of the parameters of the machine tool depending on the magnitude of deviation from target of the last item produced.

(a) Using an AR(1) model for this series, forecast the next ten values.

(b) Plot the series, the forecasts, and 95% forecast limits, and interpret the results.

9.22 The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.) The time series contains outliers that are quite obvious in the time series plot. Replace each of the unusual values at "times" 63, 106, and 129 with the much more typical value of 35 days.

(a) Use an MA(2) model to forecast the next ten values of this modified series.

(b) Plot the series, the forecasts, and 95% forecast limits, and interpret the results.

9.23 The time series in the data file robot gives the final position in the "*x*-direction" after an industrial robot has finished a planned set of exercises. The measurements are expressed as deviations from a target position. The robot is put through this planned set of exercises in the hope that its behavior is repeatable and thus predictable.

(a) Use an IMA(1,1) model to forecast five values ahead. Obtain 95% forecast limits also.

(b) Display the forecasts, forecast limits, and actual values in a graph and interpret the results.

(c) Now use an ARMA(1,1) model to forecast five values ahead and obtain 95% forecast limits. Compare these results with those obtained in part (a).

9.24 Exhibit 9.4 on page 206 displayed the forecasts and 95% forecast limits for the square root of the Canadian hare abundance. The data are in the file named hare. Produce a similar plot in original terms. That is, plot the original abundance values together with the squares of the forecasts and squares of the forecast limits.

9.25 Consider the seasonal means plus linear time trend model for the logarithms of the monthly electricity generation time series in Exercise 9.8. (The data are in the file named electricity.)

(a) Find the two-year forecasts and forecast limits in original terms. That is, exponentiate (antilog) the results obtained in Exercise 9.8.

(b) Plot the last five years of the original time series together with two years of forecasts and the 95% forecast limits, all in original terms. Interpret the plot.

Appendix E: Conditional Expectation

If X and Y have joint pdf $f(x,y)$ and we denote the marginal pdf of X by $f(x)$, then the **conditional pdf** of Y given $X = x$ is given by

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

For a given value of x, the conditional pdf has all of the usual properties of a pdf. In particular, the **conditional expectation** of Y given $X = x$ is defined as

$$E(Y|X=x) = \int_{-\infty}^{\infty} yf(y|x)dy$$

As an expected value or mean, the **conditional expectation** of Y given $X = x$ has all of the usual properties. For example,

$$E(aY + bZ + c|X=x) = aE(Y|X=x) + bE(Z|X=x) + c \tag{9.E.1}$$

and

$$E[h(Y)|X = x] = \int_{-\infty}^{\infty} yf(y|x)dx \tag{9.E.2}$$

In addition, several new properties arise:

$$E[h(X)|X=x] = h(x) \tag{9.E.3}$$

That is, given $X = x$, the random variable $h(X)$ can be treated like a constant $h(x)$. More generally,

$$E[h(X, Y)|X=x] = E(h(x, Y)|X=x) \tag{9.E.4}$$

If we set $E(Y|X=x) = g(x)$, then $g(X)$ is a random variable and we can consider $E[g(X)]$. It can be shown that

$$E[g(X)] = E(Y)$$

which is often written as

$$E[E(Y|X)] = E(Y) \tag{9.E.5}$$

If Y and X are independent, then

$$E(Y|X) = E(Y) \tag{9.E.6}$$

Appendix F: Minimum Mean Square Error Prediction

Suppose Y is a random variable with mean μ_Y and variance σ_Y^2. If our object is to predict Y using only a constant c, what is the *best* choice for c? Clearly, we must first define *best*. A common (and convenient) criterion is to choose c to minimize the **mean square error of prediction**, that is, to minimize

$$g(c) = E[(Y - c)^2]$$

If we expand $g(c)$, we have

$$g(c) = E(Y^2) - 2cE(Y) + c^2$$

Since $g(c)$ is quadratic in c and opens upward, solving $g'(c) = 0$ will produce the required minimum. We have

$$g'(c) = -2E(Y) + 2c$$

so that the optimal c is

$$c = E(Y) = \mu \tag{9.F.1}$$

Note also that

$$\min_{-\infty < c < \infty} g(c) = E(Y-\mu)^2 = \sigma_Y^2 \tag{9.F.2}$$

Now consider the situation where a second random variable X is available and we wish to use the observed value of X to help predict Y. Let $\rho = Corr\,(X,Y)$. We first suppose, for simplicity, that only *linear* functions $a + bX$ can be used for the prediction. The mean square error is then given by

$$g(a, b) = E(Y - a - bX)^2$$

and expanding we gave

$$g(a, b) = E(Y^2) + a^2 + b^2E(X^2) - 2aE(Y) + 2abE(X) - 2bE(XY)$$

This is also quadratic in a and b and opens upward. Thus we can find the point of minimum by solving simultaneous linear equations $\partial g(a, b)/\partial a = 0$ and $\partial g(a, b)/\partial b = 0$. We have

$$\partial g(a, b)/\partial a = 2a - 2E(Y) + 2bE(X) = 0$$
$$\partial g(a, b)/\partial b = 2bE(X^2) + 2aE(X) - 2E(XY) = 0$$

which we rewrite as

$$a + E(X)b = E(Y)$$
$$E(X)a + E(X^2)b = EXY$$

Multiplying the first equation by $E(X)$ and subtracting yields

$$b = \frac{E(XY) - E(X)E(Y)}{E(X^2) - [E(X)]^2} = \frac{Cov(X, Y)}{Var(X)} = \rho\frac{\sigma_Y}{\sigma_X} \tag{9.F.3}$$

Then

$$a = E(Y) - bE(X) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \tag{9.F.4}$$

If we let $\hat{Y}$ be the minimum mean square error prediction of Y based on a linear function of X, then we can write

$$\hat{Y} = \left[\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right] + \left[\rho\frac{\sigma_Y}{\sigma_X}\mu_X\right]X \tag{9.F.5}$$

or

$$\left[\frac{\hat{Y}-\mu_Y}{\sigma_Y}\right] = \rho\left[\frac{X-\mu_X}{\sigma_X}\right] \tag{9.F.6}$$

In terms of standardized variables $\hat{Y}^*$ and X^*, we have simply $\hat{Y}^* = \rho X^*$.

Also, using Equations (9.F.3) and (9.F.4), we find

$$min\ g(a,b) = \sigma_Y^2(1-\rho^2) \tag{9.F.7}$$

which provides a proof that $-1 \le \rho \le +1$ since $g(a,b) \ge 0$.

If we compare Equation (9.F.7) with Equation (9.F.2), we see that the minimum mean square error obtained when we use a linear function of X to predict Y is reduced by a factor of $1-\rho^2$ compared with that obtained by ignoring X and simply using the constant μ_Y for our prediction.

Let us now consider the more general problem of predicting Y with an arbitrary function of X. Once more our criterion will be to minimize the mean square error of prediction. We need to choose the function $h(X)$, say, that minimizes

$$E[Y-h(X)]^2 \tag{9.F.8}$$

Using Equation (9.E.5), we can write this as

$$E[Y-h(X)]^2 = E(E\{[Y-h(X)]^2 \big| X\}) \tag{9.F.9}$$

Using Equation (9.E.4), the inner expectation can be written as

$$E\{[Y-h(X)]^2 \big| X=x\} = E\{[Y-h(x)]^2 \big| X=x\} \tag{9.F.10}$$

For each value of x, $h(x)$ is a constant, and we can apply the result of Equation (9.F.1) to the conditional distribution of Y given $X = x$. Thus, for each x, the best choice of $h(x)$ is

$$h(x) = E(Y|X=x) \tag{9.F.11}$$

Since this choice of $h(x)$ minimizes the inner expectation in Equation (9.F.9), it must also provide the overall minimum of Equation (9.F.8). Thus

$$h(X) = E(Y|X) \tag{9.F.12}$$

is the best predictor of Y of *all* functions of X.

If X and Y have a bivariate normal distribution, it is well-known that

$$E(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$$

so that the solutions given in Equations (9.F.12) and (9.F.5) coincide. In this case, the linear predictor is the best of all functions.

More generally, if Y is to be predicted by a function of $X_1, X_2, \ldots, X_n$, then it can be easily argued that the minimum square error predictor is given by

$$E(Y|X_1, X_2, \ldots, X_n) \tag{9.F.13}$$

Appendix G: The Truncated Linear Process

Suppose $\{Y_t\}$ satisfies the general ARIMA(p,d,q) model with AR characteristic polynomial $\phi(x)$, MA characteristic polynomial $\theta(x)$, and constant term θ_0. Then the **truncated linear process** representation for $\{Y_t\}$ is given by

$$Y_{t+l} = C_t(\ell) + I_t(\ell) \quad \text{for } \ell \geq 1 \tag{9.G.1}$$

where

$$I_t(\ell) = \sum_{j=0}^{\ell-1} \psi_j e_{t+\ell-j} \quad \text{for } \ell \geq 1 \tag{9.G.2}$$

$$C_t(\ell) = \sum_{i=0}^{d} A_i \ell^i + \sum_{i=1}^{r} \sum_{j=0}^{p_i-1} B_{ij} \ell^j (G_i)^\ell \tag{9.G.3}$$

and A_i, B_{ij}, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, p_i$, are constant in ℓ and depend only on Y_t, $Y_{t-1}, \ldots$.[†] As always, the ψ-weights are defined by the identity

$$\phi(x)(1-x)^d(1 + \psi_1 x + \psi_2 x^2 + \cdots) = \theta(x) \tag{9.G.4}$$

or

$$\varphi(x)(1 + \psi_1 x + \psi_2 x^2 + \cdots) = \theta(x) \tag{9.G.5}$$

We shall show that the representation given by Equation (9.G.1) is valid by arguing that, for fixed t, $C_t(\ell)$ is essentially the **complementary function** of the defining difference equation, that is,

$$C_t(\ell) - \varphi_1 C_t(\ell-1) - \varphi_2 C_t(\ell-2) - \cdots - \varphi_{p+d} C_t(\ell-p-d) = \theta_0 \quad \text{for } \ell \geq 0 \tag{9.G.6}$$

and that $I_t(\ell)$ is a **particular solution** (without θ_0):

$$\begin{aligned} I_t(\ell) - \varphi_1 I_t(\ell-1) - \varphi_2 I_t(\ell-2) - \cdots - \varphi_{p+d} I_t(\ell-p-d) \\ = e_{t+\ell} - \theta_1 e_{t+l-1} - \theta_2 e_{t+l-2} - \cdots - \theta_q e_{t+l-q} \quad \text{for } \ell > q \end{aligned} \tag{9.G.7}$$

Since $C_t(\ell)$ contains $p + d$ arbitrary constants (the A's and the B's), summing $C_t(\ell)$ and $I_t(\ell)$ yields the general solution of the ARIMA equation. Specific values for the A's and B's will be determined by initial conditions on the $\{Y_t\}$ process.

We note that A_d is not arbitrary. We have

$$A_d = \frac{\theta_0}{(1 - \phi_1 - \phi_2 - \cdots - \phi_p)d!} \tag{9.G.8}$$

The proof that $C_t(\ell)$ as given by Equation (9.G.2) is the complementary function and satisfies Equation (9.G.6) is a standard result from the theory of difference equations

[†] The only property of the $C_t(\ell)$ that we need is that it depends only on $Y_t, Y_{t-1}, \ldots$.

(see, for example, Goldberg, 1958). We shall show that the particular solution $I_t(\ell)$ defined by Equation (9.G.2) does satisfy Equation (9.G.7).

For convenience of notation, we let $\varphi_j = 0$ for $j > p + d$. Consider the left-hand side of Equation (9.G.7). It can be written as:

$$\left.\begin{aligned}(\psi_0 e_{t+\ell} + \psi_1 e_{t+\ell-1} + \cdots + \psi_{\ell-1} e_{t+1}) - \varphi_1(\psi_0 e_{t+\ell-1} + \psi_1 e_{t+\ell-2} + \cdots \\ + \psi_{\ell-2} e_{t+1}) - \cdots - \varphi_{p+d}(\psi_0 e_{t+\ell-p-d} \\ + \psi_1 e_{t+\ell-p-d-1} + \cdots + \psi_{\ell-p-d-1} e_{t+1})\end{aligned}\right\} \quad (9.G.9)$$

Now grouping together common e_t terms and picking off their coefficients, we obtain

$$\begin{aligned}&\text{Coefficient of } e_{t+\ell-1} : \psi_0 \\ &\text{Coefficient of } e_{t+\ell-2} : \psi_1 - \varphi_1\psi_0 \\ &\text{Coefficient of } e_{t+\ell-3} : \psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0 \\ &\qquad\qquad\vdots \\ &\text{Coefficient of } e_{t+1} \quad : \psi_{\ell-1} - \varphi_1\psi_{\ell-2} - \varphi_2\psi_{\ell-3} - \cdots - \varphi_{p+d}\psi_{\ell-p-d-1}\end{aligned}$$

If $\ell > q$, we can match these coefficients to the corresponding coefficients on the right-hand side of Equation (9.G.7) to obtain the relationships

$$\left.\begin{aligned}\psi_0 &= 1 \\ \psi_1 - \varphi_1\psi_0 &= -\theta_1 \\ \psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0 &= -\theta_2 \\ &\vdots \\ \psi_q - \varphi_1\psi_{q-1} - \varphi_2\psi_{q-2} - \cdots - \varphi_q\psi_0 &= -\theta_q \\ \psi_{\ell-1} - \varphi_1\psi_{\ell-2} - \varphi_2\psi_{\ell-3} - \cdots - \varphi_{p+d}\psi_{\ell-p-d-1} &= 0 \quad \text{for } \ell > q\end{aligned}\right\} \quad (9.G.10)$$

However, by comparing these relationships with Equation (9.G.5), we see that Equations (9.G.10) are precisely the equations defining the ψ-weights and thus Equation (9.G.7) is established as required.

Appendix H: State Space Models

Control theory engineers have developed and successfully used so-called **state space models** and **Kalman filtering** since Kalman published his seminal work in 1960. Recent references include Durbin and Koopman (2001) and Harvey et al. (2004).

Consider a general stationary and invertible ARMA(p,q) process $\{Z_t\}$. Put $m = max(p, q + 1)$ and define the **state** of the process at time t as the column vector $\boldsymbol{Z}(t)$ of length m whose jth element is the forecast $\hat{Z}(j)$ for $j = 0, 1, 2,\ldots, m - 1$, based on Z_t, $Z_{t-1},\ldots$. Note that the lead element of $\boldsymbol{Z}(t)$ is just $\hat{Z}(0) = Z_t$.

Recall the updating Equation (9.6.1) on page 207, which in the present context can

be written

$$\hat{Z}_{t+1}(\ell) = \hat{Z}_t(\ell+1) + \psi_\ell e_{t+1} \tag{9.H.1}$$

We shall use this expression directly for $\ell = 0, 1, 2, \ldots, m-2$. For $\ell = m-1$, we have

$$\begin{aligned}\hat{Z}_{t+1}(m-1) &= \hat{Z}_t(m) + \psi_{m-1} e_{t+1} \\ &= \phi_1 \hat{Z}_t(m-1) + \phi_2 \hat{Z}_t(m-2) + \ldots + \phi_p \hat{Z}_t(m-p) + \psi_{m-1} e_{t+1}\end{aligned} \tag{9.H.2}$$

where the last expression comes from Equation (9.3.34) on page 200, with $\mu = 0$.

The matrix formulation of Equations (9.H.1) and (9.H.2) relating $\boldsymbol{Z}(t+1)$ to $\boldsymbol{Z}(t)$ and e_{t+1}, called the **equations of state** (or **Akaike's Markovian representation**), is given as

$$\boldsymbol{Z}(t+1) = \boldsymbol{F}\boldsymbol{Z}(t) + \boldsymbol{G}e_{t+1} \tag{9.H.3}$$

where

$$\boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ & & & & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \\ \phi_m & \phi_{m-1} & \cdot & \cdot & \cdot & \phi_1 \end{bmatrix} \tag{9.H.4}$$

and

$$\boldsymbol{G} = \begin{bmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-1} \end{bmatrix} \tag{9.H.5}$$

with $\phi_j = 0$ for $j > p$. Note that the simplicity of Equation (9.H.3) is obtained at the expense of having to deal with vector-valued processes. Because the state space formulation also usually allows for measurement error, we do not observe Z_t directly but only observe Y_t through the **observational equation**

$$Y_t = \boldsymbol{H}\boldsymbol{Z}(t) + \varepsilon_t \tag{9.H.6}$$

where $\boldsymbol{H} = [1, 0, 0, \ldots, 0]$ and $\{\varepsilon_t\}$ is another zero-mean white noise process independent of $\{e_t\}$. The special case of *no* measurement error is obtained by setting $\varepsilon_t = 0$ in Equation (9.H.6). Equivalently, this case is obtained by taking $\sigma_\varepsilon^2 = 0$ in subsequent equations. More general state space models allow $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$ to be more general, possibly also depending on time.

Evaluation of the Likelihood Function and Kalman Filtering

First a definition: The **covariance matrix** for a vector of random variables $\boldsymbol{X}$ of dimension $n \times 1$ is defined to be the $n \times n$ matrix whose ijth element is the covariance between the ith and jth components of $\boldsymbol{X}$.

If $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{B}$, then it is easily shown that the covariance matrix for $\boldsymbol{Y}$ is $\boldsymbol{AVA}^T$, where $\boldsymbol{V}$ is the covariance matrix for $\boldsymbol{X}$ and the superscript T denotes matrix transpose.

Getting back to the Kalman filter, we let $\boldsymbol{Z}(t+1|t)$ denote the $m \times 1$ vector whose jth component is $E[\hat{Z}_{t+1}(j)|Y_t, Y_{t-1}, \ldots, Y_1]$ for $j = 0, 1, 2, \ldots, m-1$. Similarly, let $\boldsymbol{Z}(t|t)$ be the vector whose jth component is $E[\hat{Z}_t(j)|Y_t, Y_{t-1}, \ldots, Y_1]$ for $j = 0, 1, 2, \ldots, m-1$.

Then, since e_{t+1} is independent of $Z_t, Z_{t-1}, \ldots$, and hence also of $Y_t, Y_{t-1}, \ldots$, we see from Equation (9.H.3) that

$$\boldsymbol{Z}(t+1|t) = \boldsymbol{F}\boldsymbol{Z}(t|t) \tag{9.H.7}$$

Also letting $\boldsymbol{P}(t+1|t)$ be the covariance matrix for the "forecast error" $\boldsymbol{Z}(t+1) - \boldsymbol{Z}(t+1|t)$ and $\boldsymbol{P}(t|t)$ be the covariance matrix for the "forecast error" $\boldsymbol{Z}(t) - \boldsymbol{Z}(t|t)$, we have from Equation (9.H.3) that

$$\boldsymbol{P}(t+1|t) = \boldsymbol{F}[\boldsymbol{P}(t|t)]\boldsymbol{F}^T + \sigma_e^2 \boldsymbol{G}\boldsymbol{G}^T \tag{9.H.8}$$

From the observational equation (Equation (9.H.6)) and then replacing $t+1$ by t,

$$Y(t+1|t) = \boldsymbol{H}\boldsymbol{Z}(t+1|t) \tag{9.H.9}$$

where $Y(t+1|t) = E(Y_{t+1}|Y_t, Y_{t-1}, \ldots, Y_1)$.

It can now be shown that the following relationships hold (see, for example, Harvey, 1981c):

$$\boldsymbol{Z}(t+1|t+1) = \boldsymbol{Z}(t+1|t) + \boldsymbol{\kappa}(t+1)[Y_{t+1} - Y(t+1|t)] \tag{9.H.10}$$

where

$$\boldsymbol{\kappa}(t+1) = \boldsymbol{P}(t+1|t)\boldsymbol{H}^T[\boldsymbol{H}\boldsymbol{P}(t+1|t)\boldsymbol{H}^T + \sigma_\varepsilon^2]^{-1} \tag{9.H.11}$$

and

$$\boldsymbol{P}(t+1|(t+1)) = \boldsymbol{P}(t+1|t) - \boldsymbol{\kappa}(t+1)\boldsymbol{H}\boldsymbol{P}(t+1|t) \tag{9.H.12}$$

Collectively, Equations (9.H.10), (9.H.11), and (9.H.12) are referred to as the **Kalman filter equations**. The quantity

$$err_{t+1} = Y_{t+1} - Y(t+1|t) \tag{9.H.13}$$

in Equation (9.H.10) is the prediction error and is independent of (or at least uncorrelated with) the past observations $Y_t, Y_{t-1}, \ldots$. Since we are allowing for measurement error, err_{t+1} is not, in general, the same as e_{t+1}.

From Equations (9.H.13) and (9.H.6), we have

$$v_{t+1} = Var(err_{t+1}) = \boldsymbol{H}\boldsymbol{P}(t+1|t)\boldsymbol{H}^T + \sigma_\varepsilon^2 \tag{9.H.14}$$

Now consider the likelihood function for the observed series $Y_1, Y_2, \ldots, Y_n$. From the definition of the conditional probability density function, we can write

$$f(y_1, y_2, \ldots, y_n) = f(y_n|y_1, y_2, \ldots, y_{n-1})f(y_1, y_2, \ldots, y_{n-1})$$

or, by taking logs,

$$\log f(y_1, y_2, \ldots, y_n) = \log f(y_1, y_2, \ldots, y_{n-1}) + \log f(y_n|y_1, y_2, \ldots, y_{n-1}) \quad (9.H.15)$$

Assume now that we are dealing with normal distributions, that is, that $\{e_t\}$ and $\{\varepsilon_t\}$ are normal white noise processes. Then it is known that the distribution of Y_n conditional on $Y_1 = y_1, Y_2 = y_2, \ldots, Y_{n-1} = y_{n-1}$, is also normal with mean $y(n|n-1)$ and variance v_n. In the remainder of this section and the next, we write $y(n|n-1)$ for the observed value of $Y(n|n-1)$. The second term on the right-hand side of Equation (9.H.15) can then be written

$$\log f(y_n|y_1, y_2, \ldots, y_{n-1}) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log v_n - \frac{1}{2}\frac{[y_n - y(n|n-1)]^2}{v_n}$$

Furthermore, the first term on the right-hand side of Equation (9.H.15) can be decomposed similarly again and again until we have

$$\log f(y_1, y_2, \ldots, y_n) = \sum_{t=2}^{n} \log f(y_t|y_1, y_2, \ldots, y_{t-1}) + \log f(y_1) \quad (9.H.16)$$

which then becomes the prediction error decomposition of the likelihood, namely

$$\log f(y_1, y_2, \ldots, y_n) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n} v_t - \frac{1}{2}\sum_{t=1}^{n} \frac{[y_t - y(t|t-1)]^2}{v_t} \quad (9.H.17)$$

with $y(1|0) = 0$ and $v_1 = Var(Y_1)$.

The overall strategy for computing the likelihood for a given set of parameter values is to use the Kalman filter equations to generate recursively the prediction errors and their variances and then use the prediction error decomposition of the likelihood function. Only one point remains: We need initial values $\boldsymbol{Z}(0|0)$ and $\boldsymbol{P}(0|0)$ to get the recursions started.

The Initial State Covariance Matrix

The initial state vector $\boldsymbol{Z}(0|0)$ will be a vector of zeros for a zero-mean process, and $\boldsymbol{P}(0|0)$ is the covariance matrix for $\boldsymbol{Z}(0) - \boldsymbol{Z}(0|0) = \boldsymbol{Z}(0)$. Now, because $\boldsymbol{Z}(0)$ is the column vector with elements $[Z_0, \hat{Z}_0(1), \ldots, \hat{Z}_0(m-1)]$, it is necessary for us to evaluate

$$Cov[\hat{Z}_0(i), \hat{Z}_0(j)] \qquad \text{for } i, j = 0, 1, \ldots, m-1$$

From the truncated linear process form, Equation (9.3.35) on page 200 with $C_t(\ell) = \hat{Z}_t(\ell)$, we may write, for $j > 0$

$$Z_j = \hat{Z}_0(j) + \sum_{k=-j}^{-1} \psi_{j+k}\, e_{-k} \tag{9.H.18}$$

Multiplying Equation (9.H.18) by Z_0 and taking expected values yields

$$\gamma_j = E(Z_0 Z_j) = E[\hat{Z}_0(0)(\hat{Z}_0(j))] \qquad \text{for } j \geq 0 \tag{9.H.19}$$

Now multiply Equation (9.H.18) by itself with j replaced by i and take expected values. Recalling that the e's are independent of past Z's and assuming $0 < i \leq j$, we obtain

$$\gamma_{j-i} = Cov[\hat{Z}_0(i), \hat{Z}_0(j)] + \sigma_e^2 \sum_{k=0}^{i-1} \psi_k \psi_{k+j-i} \tag{9.H.20}$$

Combining Equations (9.H.19) and (9.H.20), we have as the required elements of $\boldsymbol{P}(0|0)$

$$Cov[\hat{Z}_0(i), \hat{Z}_0(j)] = \begin{cases} \gamma_i & 0 = i \leq j \leq m-1 \\ \gamma_{j-i} - \sigma_e^2 \sum_{k=0}^{i-1} \psi_k \psi_{k+j-i} & 1 \leq i \leq j \leq m-1 \end{cases} \tag{9.H.21}$$

where the ψ-weights are obtained from the recursion of Equation (4.4.7) on page 79, and γ_k, the autocovariance function for the $\{Z_t\}$ process, is obtained as in Appendix C on page 85.

The variance σ_e^2 can be removed from the problem by dividing σ_ε^2 by σ_e^2. The prediction error variance v_t is then replaced by $\sigma_e^2 v_t$ in the log-likelihood of Equation (9.H.17), and we set $\sigma_e^2 = 1$ in Equation (9.H.8). Dropping unneeded constants, we get the new log-likelihood

$$\ell = \sum_{t=1}^{n} \left\{ \log(\sigma_e^2 v_t) + \frac{[y_t - y(t|t-1)]^2}{v_t} \right\} \tag{9.H.22}$$

which can be minimized analytically with respect to σ_e^2. We obtain

$$\sigma_e^2 = \sum_{t=1}^{n} \left\{ \frac{[y_t - y(t|t-1)]^2}{\sigma_e^2 v_t} \right\} \tag{9.H.23}$$

Substituting this back into Equation (9.H.22), we now find that

$$\ell = \sum_{t=1}^{n} \log v_t + n \log \sum_{t=1}^{n} \frac{[y_t - y(t|t-1)]^2}{v_t} \tag{9.H.24}$$

which must be minimized numerically with respect to $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$, and σ_e^2. Having done so, we return to Equation (9.H.23) to estimate σ_e^2. The function defined by Equation (9.H.24) is sometimes called the **concentrated log-likelihood function**.

CHAPTER 10

SEASONAL MODELS

In Chapter 3, we saw how seasonal deterministic trends might be modeled. However, in many areas in which time series are used, particularly business and economics, the assumption of any deterministic trend is quite suspect even though cyclical tendencies are very common in such series.

Here is an example: Levels of carbon dioxide (CO_2) are monitored at several sites around the world to investigate atmospheric changes. One of the sites is at Alert, Northwest Territories, Canada, near the Arctic Circle.

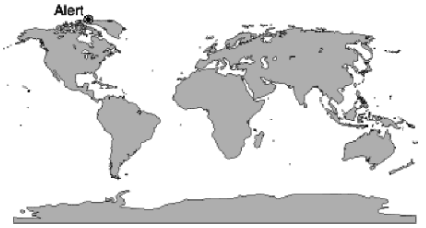
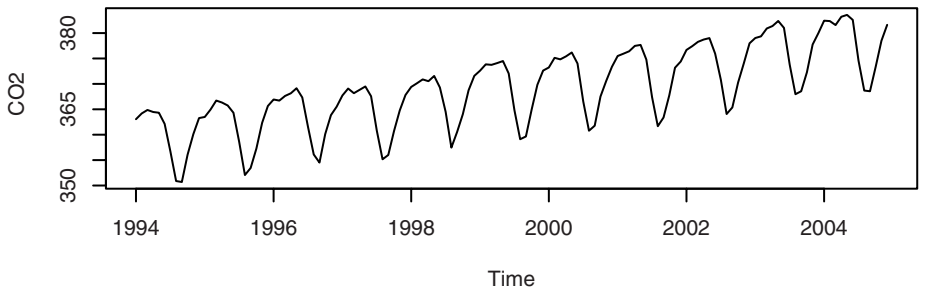


Exhibit 10.1 displays the monthly CO_2 levels from January 1994 through December 2004. There is a strong upward trend but also a seasonality that can be seen better in the more detailed Exhibit 10.2, where only the last few years are graphed using monthly plotting symbols.

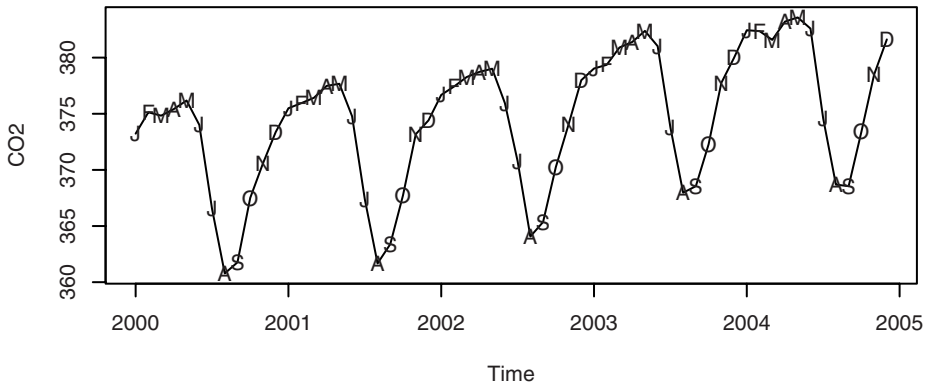
Exhibit 10.1 Monthly Carbon Dioxide Levels at Alert, NWT, Canada




```
> data(co2)
> win.graph(width=4.875,height=3,pointsize=8)
> plot(co2,ylab='CO2')
```

As we see in the displays, carbon dioxide levels are higher during the winter months and much lower in the summer. Deterministic seasonal models such as seasonal means plus linear time trend or sums of cosine curves at various frequencies plus linear time trend as we investigated in Chapter 3 could certainly be considered here. But we discover that such models do not explain the behavior of this time series. For this series and many others, it can be shown that the residuals from a seasonal means plus linear time trend model are highly autocorrelated at many lags.[†] In contrast, we will see that the **stochastic seasonal models** developed in this chapter do work well for this series.

Exhibit 10.2 Carbon Dioxide Levels with Monthly Symbols




```
> plot(window(co2,start=c(2000,1)),ylab='CO2')
> Month=c('J','F','M','A','M','J','J','A','S','O','N','D')
> points(window(co2,start=c(2000,1)),pch=Month)
```

10.1 Seasonal ARIMA Models

We begin by studying stationary models and then consider nonstationary generalizations in Section 10.3. We let s denote the known seasonal period; for monthly series $s = 12$ and for quarterly series $s = 4$.

Consider the time series generated according to

$$Y_t = e_t - \Theta e_{t-12}$$

Notice that

$$\begin{aligned} Cov(Y_t, Y_{t-1}) &= Cov(e_t - \Theta e_{t-12}, e_{t-1} - \Theta e_{t-13}) \\ &= 0 \end{aligned}$$

but that

[†] We ask you to verify this in Exercise 10.8.

$$\begin{aligned} Cov(Y_t, Y_{t-12}) &= Cov(e_t - \Theta e_{t-12}, e_{t-12} - \Theta e_{t-24}) \\ &= -\Theta\sigma_e^2 \end{aligned}$$

It is easy to see that such a series is stationary and has nonzero autocorrelations only at lag 12.

Generalizing these ideas, we define a **seasonal MA(Q) model of order Q with seasonal period s** by

$$Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - \cdots - \Theta_Q e_{t-Qs} \tag{10.1.1}$$

with **seasonal MA characteristic polynomial**

$$\Theta(x) = 1 - \Theta_1 x^s - \Theta_2 x^{2s} - \cdots - \Theta_Q x^{Qs} \tag{10.1.2}$$

It is evident that such a series is always stationary and that the autocorrelation function will be nonzero only at the seasonal lags of s, $2s$, $3s$,…, Qs. In particular,

$$\rho_{ks} = \frac{-\Theta_k + \Theta_1\Theta_{k+1} + \Theta_2\Theta_{k+2} + \cdots + \Theta_{Q-k}\Theta_Q}{1 + \Theta_1^2 + \Theta_2^2 + \cdots + \Theta_Q^2} \text{ for } k = 1, 2, \ldots, Q \tag{10.1.3}$$

(Compare this with Equation (4.2.5) on page 65 for the nonseasonal MA process.) For the model to be invertible, the roots of $\Theta(x) = 0$ must all exceed 1 in absolute value.

It is useful to note that the seasonal MA(Q) model can also be viewed as a special case of a nonseasonal MA model of order $q = Qs$ but with all θ-values zero except at the seasonal lags s, $2s$, $3s$,…, Qs.

Seasonal autoregressive models can also be defined. Consider

$$Y_t = \Phi Y_{t-12} + e_t \tag{10.1.4}$$

where $|\Phi| < 1$ and e_t is independent of Y_{t-1}, Y_{t-2},… . It can be shown that $|\Phi| < 1$ ensures stationarity. Thus it is easy to argue that $E(Y_t) = 0$; multiplying Equation (10.1.4) by Y_{t-k}, taking expectations, and dividing by γ_0 yields

$$\rho_k = \Phi\rho_{k-12} \text{ for } k \geq 1 \tag{10.1.5}$$

Clearly

$$\rho_{12} = \Phi\rho_0 = \Phi \text{ and } \rho_{24} = \Phi\rho_{12} = \Phi^2$$

More generally,

$$\rho_{12k} = \Phi^k \text{ for } k = 1, 2, \ldots \tag{10.1.6}$$

Furthermore, setting $k = 1$ and then $k = 11$ in Equation (10.1.5) and using $\rho_k = \rho_{-k}$ gives us

$$\rho_1 = \Phi\rho_{11} \text{ and } \rho_{11} = \Phi\rho_1$$

which implies that $\rho_1 = \rho_{11} = 0$. Similarly, one can show that $\rho_k = 0$ except at the seasonal lags 12, 24, 36,… . At those lags, the autocorrelation function decays exponentially like an AR(1) model.

With this example in mind, we define a **seasonal AR(P) model of order P and seasonal period s** by

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t \tag{10.1.7}$$

with **seasonal characteristic polynomial**

$$\Phi(x) = 1 - \Phi_1 x^s - \Phi_2 x^{2s} - \cdots - \Phi_P x^{Ps} \tag{10.1.8}$$

As always, we require e_t to be independent of $Y_{t-1}, Y_{t-2},\ldots$, and, for stationarity, that the roots of $\Phi(x) = 0$ be greater than 1 in absolute value. Again, Equation (10.1.7) can be seen as a special AR(p) model of order $p = Ps$ with nonzero ϕ-coefficients only at the seasonal lags s, $2s$, $3s,\ldots, Ps$.

It can be shown that the autocorrelation function is nonzero only at lags s, $2s$, $3s$, …, where it behaves like a combination of decaying exponentials and damped sine functions. In particular, Equations (10.1.4), (10.1.5), and (10.1.6) easily generalize to the general seasonal AR(1) model to give

$$\rho_{ks} = \Phi^k \text{ for } k = 1, 2, \ldots \tag{10.1.9}$$

with zero correlation at other lags.

10.2 Multiplicative Seasonal ARMA Models

Rarely shall we need models that incorporate autocorrelation *only* at the seasonal lags. By combining the ideas of seasonal and nonseasonal ARMA models, we can develop parsimonious models that contain autocorrelation for the seasonal lags but also for low lags of neighboring series values.

Consider a model whose MA characteristic polynomial is given by

$$(1 - \theta x)(1 - \Theta x^{12})$$

Multiplying out, we have $1 - \theta x - \Theta x^{12} + \theta\Theta x^{13}$. Thus the corresponding time series satisfies

$$Y_t = e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \tag{10.2.1}$$

For this model, we can check that the autocorrelation function is nonzero only at lags 1, 11, 12, and 13. We find

$$\gamma_0 = (1 + \theta^2)(1 + \Theta^2)\sigma_e^2 \tag{10.2.2}$$

$$\rho_1 = -\frac{\theta}{1 + \theta^2} \tag{10.2.3}$$

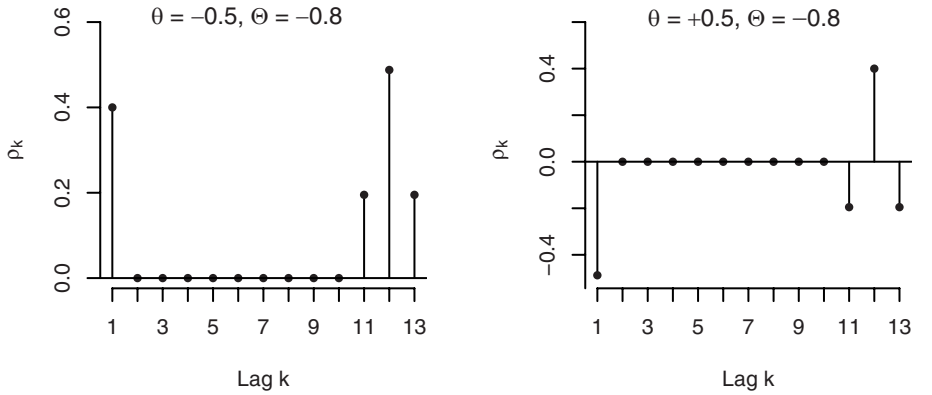
$$\rho_{11} = \rho_{13} = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)} \tag{10.2.4}$$

and

$$\rho_{12} = -\frac{\Theta}{1+\Theta^2} \tag{10.2.5}$$

Exhibit 10.3 displays the autocorrelation functions for the model of Equation (10.2.1) with $\theta = \pm 0.5$ and $\Theta = -0.8$ as given by Equations (10.2.2)–(10.2.5).

Exhibit 10.3 Autocorrelations from Equations (10.2.2)-(10.2.5)



Of course, we could also introduce both short-term and seasonal autocorrelations by defining an MA model of order 12 with only θ_1 and θ_{12} nonzero. We shall see in the next section that the "multiplicative" model arises quite naturally for nonstationary models that entail differencing.

In general, then, we define a **multiplicative seasonal ARMA(p,q)×(P,Q)$_s$ model with seasonal period s** as a model with AR characteristic polynomial $\phi(x)\Phi(x)$ and MA characteristic polynomial $\theta(x)\Theta(x)$, where

$$\left.\begin{aligned}\phi(x) &= 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \\ \Phi(x) &= 1 - \Phi_1 x^s - \Phi_2 x^{2s} - \cdots - \Phi_P x^{Ps}\end{aligned}\right\} \tag{10.2.6}$$

and

$$\left.\begin{aligned}\theta(x) &= 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q \\ \Theta(x) &= 1 - \Theta_1 x^s - \Theta_2 x^{2s} - \cdots - \Theta_Q x^{Qs}\end{aligned}\right\} \tag{10.2.7}$$

The model may also contain a constant term θ_0. Note once more that we have just a special ARMA model with AR order $p + Ps$ and MA order $q + Qs$, but the coefficients are not completely general, being determined by only $p + P + q + Q$ coefficients. If $s = 12$, $p + P + q + Q$ will be considerably smaller than $p + Ps + q + Qs$ and will allow a much more parsimonious model.

As another example, suppose $P = q = 1$ and $p = Q = 0$ with $s = 12$. The model is then

$$Y_t = \Phi Y_{t-12} + e_t - \theta e_{t-1} \tag{10.2.8}$$

Using our standard techniques, we find that

$$\gamma_1 = \Phi\gamma_{11} - \theta\sigma_e^2 \tag{10.2.9}$$

and

$$\gamma_k = \Phi\gamma_{k-12} \text{ for } k \geq 2 \tag{10.2.10}$$

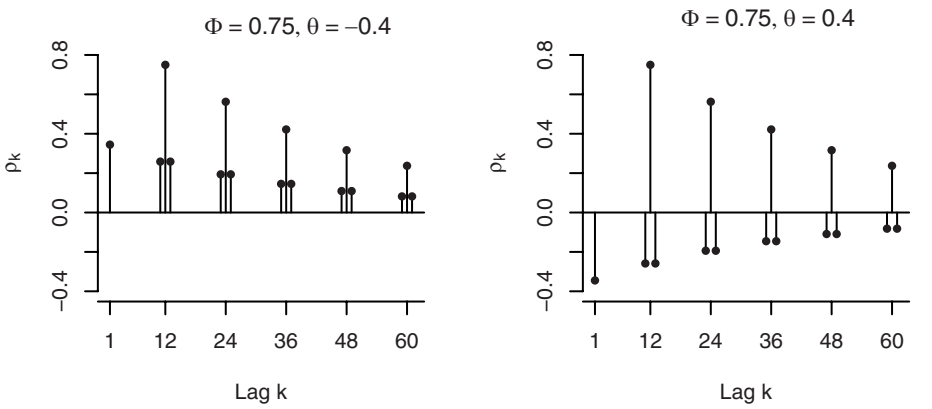
After considering the equations implied by various choices for k, we arrive at

$$\left.\begin{aligned} \gamma_0 &= \left[\frac{1+\theta^2}{1-\Phi^2}\right]\sigma_e^2 \\ \rho_{12k} &= \Phi^k \text{ for } k \geq 1 \\ \rho_{12k-1} &= \rho_{12k+1} = \left(-\frac{\theta}{1+\theta^2}\Phi^k\right) \text{ for } k = 0, 1, 2, \ldots \end{aligned}\right\} \tag{10.2.11}$$

with autocorrelations for all other lags equal to zero.

Exhibit 10.4 displays the autocorrelation functions for two of these seasonal ARIMA processes with period 12: one with $\Phi = 0.75$ and $\theta = 0.4$, the other with $\Phi = 0.75$ and $\theta = -0.4$. The shape of these autocorrelations is somewhat typical of the sample autocorrelation functions for numerous seasonal time series. The even simpler autocorrelation function given by Equations (10.2.3), (10.2.4), and (10.2.5) and displayed in Exhibit 10.3 also seems to occur frequently in practice (perhaps after differencing).

Exhibit 10.4 Autocorrelation Functions from Equation (10.2.11)



10.3 Nonstationary Seasonal ARIMA Models

An important tool in modeling nonstationary seasonal processes is the seasonal difference. The **seasonal difference** of period s for the series $\{Y_t\}$ is denoted $\nabla_s Y_t$ and is defined as

$$\nabla_s Y_t = Y_t - Y_{t-s} \tag{10.3.1}$$

For example, for monthly series we consider the changes from January to January, February to February, and so forth for successive years. Note that for a series of length n, the seasonal difference series will be of length $n - s$; that is, s data values are lost due to seasonal differencing.

As an example where seasonal differencing is appropriate, consider a process generated according to

$$Y_t = S_t + e_t \tag{10.3.2}$$

with

$$S_t = S_{t-s} + \varepsilon_t \tag{10.3.3}$$

where $\{e_t\}$ and $\{\varepsilon_t\}$ are independent white noise series. Here $\{S_t\}$ is a "seasonal random walk," and if $\sigma_\varepsilon \ll \sigma_e$, $\{S_t\}$ would model a slowly changing seasonal component.

Due to the nonstationarity of $\{S_t\}$, clearly $\{Y_t\}$ is nonstationary. However, if we seasonally difference $\{Y_t\}$, as given in Equation (10.3.1), we find

$$\begin{aligned} \nabla_s Y_t &= S_t - S_{t-s} + e_t - e_{t-s} \\ &= \varepsilon_t + e_t - e_{t-s} \end{aligned} \tag{10.3.4}$$

An easy calculation shows that $\nabla_s Y_t$ is stationary and has the autocorrelation function of an $\text{MA}(1)_s$ model.

The model described by Equations (10.2.2) and (10.2.3) could also be generalized to account for a nonseasonal, slowly changing stochastic trend. Consider

$$Y_t = M_t + S_t + e_t \tag{10.3.5}$$

with

$$S_t = S_{t-s} + \varepsilon_t \tag{10.3.6}$$

and

$$M_t = M_{t-1} + \xi_t \tag{10.3.7}$$

where $\{e_t\}$, $\{\varepsilon_t\}$, and $\{\xi_t\}$ are mutually independent white noise series. Here we take both a seasonal difference and an ordinary nonseasonal difference to obtain[†]

[†] It should be noted that $\nabla_s Y_t$ will in fact be stationary and $\nabla\nabla_s Y_t$ will be noninvertible. We use Equations (10.2.5), (10.2.6), and (10.2.7) merely to help motivate multiplicative seasonal ARIMA models.

$$\begin{aligned}\nabla\nabla_s Y_t &= \nabla(M_t - M_{t-s} + \varepsilon_t + e_t - e_{t-s}) \\ &= (\xi_t + \varepsilon_t + e_t) - (\varepsilon_{t-1} + e_{t-1}) - (\xi_{t-s} + e_{t-s}) + e_{t-s-1}\end{aligned} \tag{10.3.8}$$

The process defined here is stationary and has nonzero autocorrelation only at lags 1, $s-1$, s, and $s+1$, which agrees with the autocorrelation structure of the multiplicative seasonal model ARMA(0,1)×(0,1) with seasonal period s.

These examples lead to the definition of nonstationary seasonal models. A process $\{Y_t\}$ is said to be a **multiplicative seasonal ARIMA model** with nonseasonal (regular) orders p, d, and q, seasonal orders P, D, and Q, and seasonal period s if the differenced series

$$W_t = \nabla^d \nabla_s^D Y_t \tag{10.3.9}$$

satisfies an ARMA$(p,q)\times(P,Q)_s$ model with seasonal period s.[†] We say that $\{Y_t\}$ is an ARIMA$(p,d,q)\times(P,D,Q)_s$ model with seasonal period s.

Clearly, such models represent a broad, flexible class from which to select an appropriate model for a particular time series. It has been found empirically that many series can be adequately fit by these models, usually with a small number of parameters, say three or four.

10.4 Model Specification, Fitting, and Checking

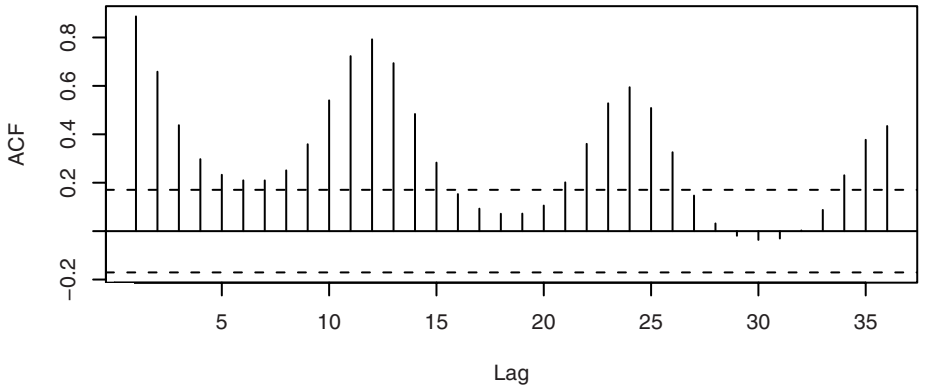
Model specification, fitting, and diagnostic checking for seasonal models follow the same general techniques developed in Chapters 6, 7, and 8. Here we shall simply highlight the application of these ideas specifically to seasonal models and pay special attention to the seasonal lags.

Model Specification

As always, a careful inspection of the time series plot is the first step. Exhibit 10.1 on page 227 displays monthly carbon dioxide levels in northern Canada. The upward trend alone would lead us to specify a nonstationary model. Exhibit 10.5 shows the sample autocorrelation function for that series. The seasonal autocorrelation relationships are shown quite prominently in this display. Notice the strong correlation at lags 12, 24, 36, and so on. In addition, there is substantial other correlation that needs to be modeled.

[†] Using the backshift operator notation of Appendix D, page 106, we may write the general ARIMA$(p,d,q)\times(P,D,Q)_s$ model as $\phi(B)\Phi(B)\nabla^d\nabla_s^D Y_t = \theta(B)\Theta(B)e_t$.

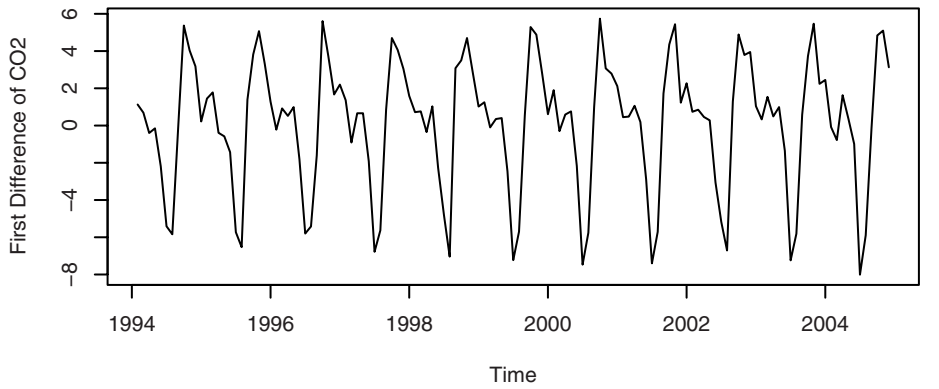
Exhibit 10.5 Sample ACF of CO_2 Levels




```
> acf(as.vector(co2),lag.max=36)
```

Exhibit 10.6 shows the time series plot of the CO_2 levels after we take a first difference.

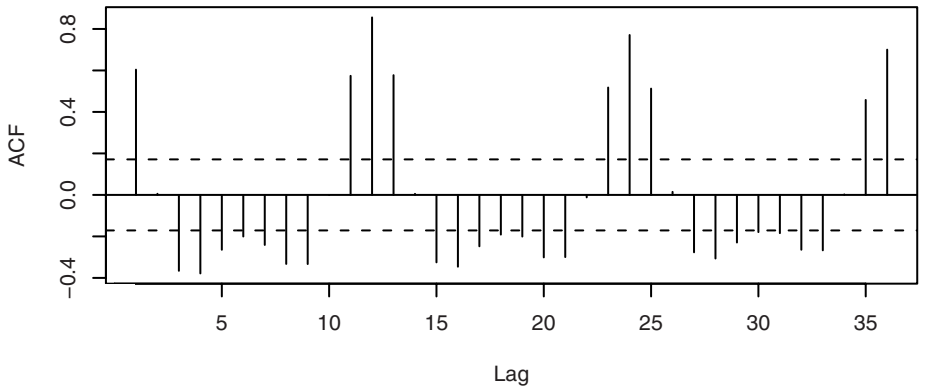
Exhibit 10.6 Time Series Plot of the First Differences of CO_2 Levels




```
> plot(diff(co2),ylab='First Difference of CO2',xlab='Time')
```

The general upward trend has now disappeared but the strong seasonality is still present, as evidenced by the behavior shown in Exhibit 10.7. Perhaps seasonal differencing will bring us to a series that may be modeled parsimoniously.

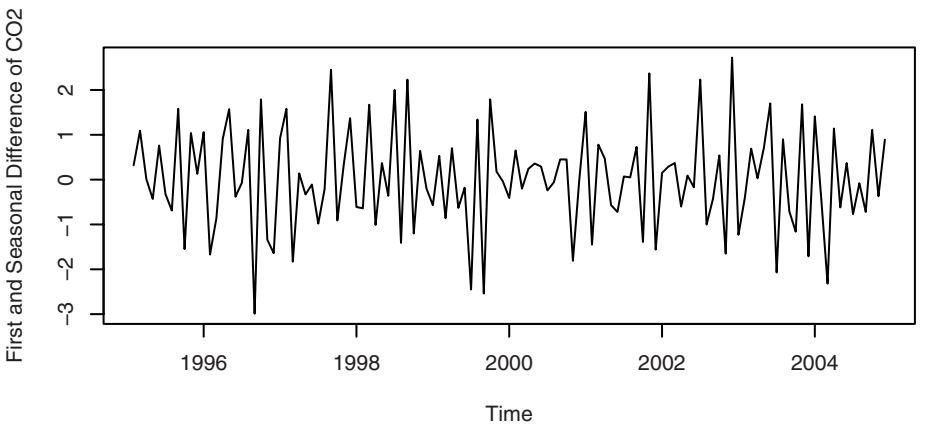
Exhibit 10.7 Sample ACF of First Differences of CO_2 Levels




```
> acf(as.vector(diff(co2)),lag.max=36)
```

Exhibit 10.8 displays the time series plot of the CO_2 levels after taking both a first difference and a seasonal difference. It appears that most, if not all, of the seasonality is gone now.

Exhibit 10.8 Time Series Plot of First and Seasonal Differences of CO_2




```
> plot(diff(diff(co2),lag=12),xlab='Time',
    ylab='First and Seasonal Difference of CO2')
```

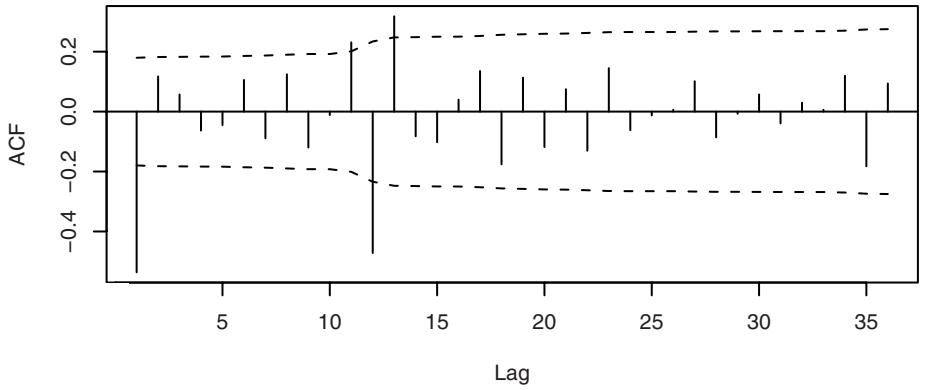
Exhibit 10.9 confirms that very little autocorrelation remains in the series after these two differences have been taken. This plot also suggests that a simple model which incorporates the lag 1 and lag 12 autocorrelations might be adequate.

We will consider specifying the multiplicative, seasonal ARIMA$(0,1,1)\times(0,1,1)_{12}$ model

$$\nabla_{12}\nabla Y_t = e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \tag{10.4.10}$$

which incorporates many of these requirements. As usual, all models are tentative and subject to revision at the diagnostics stage of model building.

Exhibit 10.9 Sample ACF of First and Seasonal Differences of CO_2




```
> acf(as.vector(diff(diff(co2),lag=12)),lag.max=36,ci.type='ma')
```

Model Fitting

Having specified a tentative seasonal model for a particular time series, we proceed to estimate the parameters of that model as efficiently as possible. As we have remarked earlier, multiplicative seasonal ARIMA models are just special cases of our general ARIMA models. As such, all of our work on parameter estimation in Chapter 7 carries over to the seasonal case.

Exhibit 10.10 gives the maximum likelihood estimates and their standard errors for the ARIMA$(0,1,1)\times(0,1,1)_{12}$ model for CO_2 levels.

Exhibit 10.10 Parameter Estimates for the CO_2 Model

Coefficient	θ	Θ
Estimate	0.5792	0.8206
Standard error	0.0791	0.1137

$\hat{\sigma}_e^2 = 0.5446$: log-likelihood = −139.54, AIC = 283.08

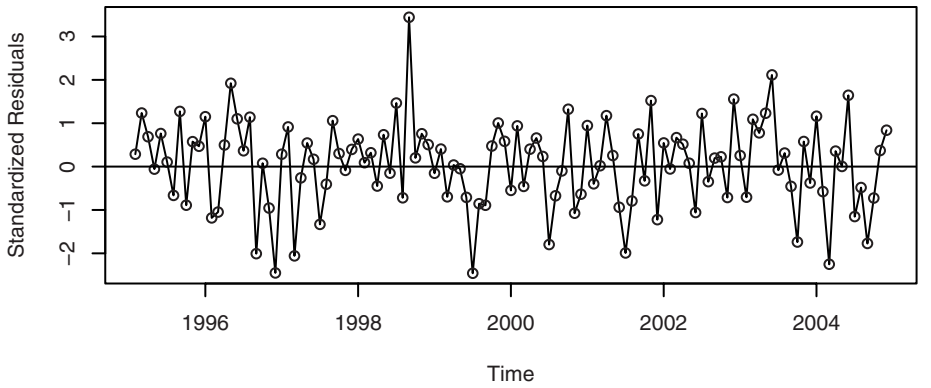
```
> m1.co2=arima(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
    period=12))
> m1.co2
```

The coefficient estimates are all highly significant, and we proceed to check further on this model.

Diagnostic Checking

To check the estimated the ARIMA$(0,1,1)\times(0,1,1)_{12}$ model, we first look at the time series plot of the residuals. Exhibit 10.11 gives this plot for standardized residuals. Other than some strange behavior in the middle of the series, this plot does not suggest any major irregularities with the model, although we may need to investigate the model further for outliers, as the standardized residual at September 1998 looks suspicious. We investigate this further in Chapter 11.

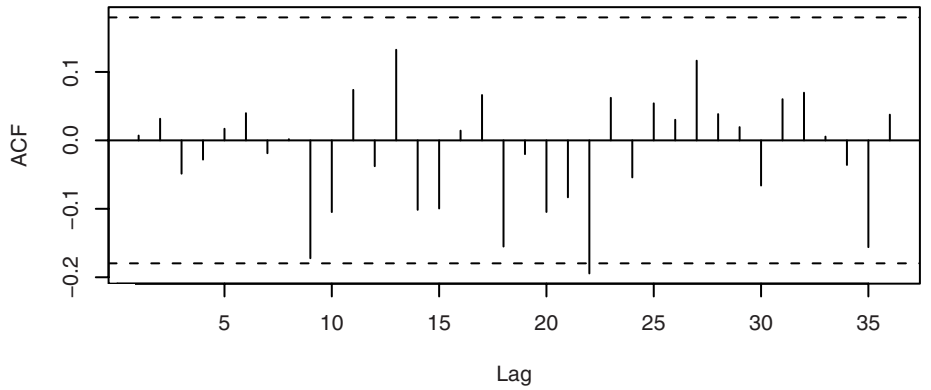
Exhibit 10.11 Residuals from the ARIMA$(0,1,1)\times(0,1,1)_{12}$ Model




```
> plot(window(rstandard(m1.co2),start=c(1995,2)),
    ylab='Standardized Residuals',type='o')
> abline(h=0)
```

To look further, we graph the sample ACF of the residuals in Exhibit 10.12. The only "statistically significant" correlation is at lag 22, and this correlation has a value of only –0.17, a very small correlation. Furthermore, we can think of no reasonable interpretation for dependence at lag 22. Finally, we should not be surprised that one autocorrelation out of the 36 displayed is statistically significant. This could easily happen by chance alone. Except for marginal significance at lag 22, the model seems to have captured the essence of the dependence in the series.

Exhibit 10.12 ACF of Residuals from the ARIMA(0,1,1)×(0,1,1)$_{12}$ Model

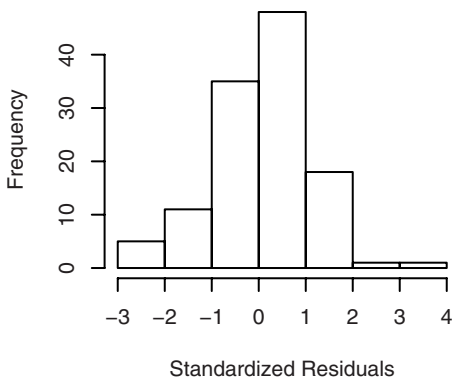



```
> acf(as.vector(window(rstandard(m1.co2),start=c(1995,2))),
    lag.max=36)
```

The Ljung-Box test for this model gives a chi-squared value of 25.59 with 22 degrees of freedom, leading to a *p*-value of 0.27—a further indication that the model has captured the dependence in the time series.

Next we investigate the question of normality of the error terms via the residuals. Exhibit 10.13 displays the histogram of the residuals. The shape is somewhat "bell-shaped" but certainly not ideal. Perhaps a quantile-quantile plot will tell us more.

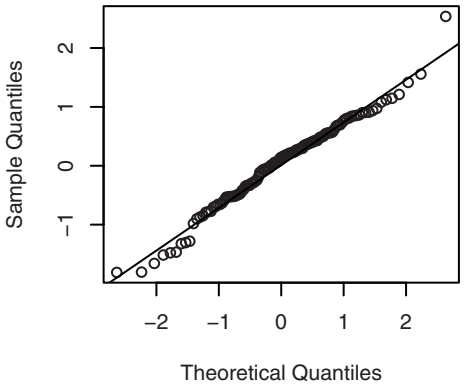
Exhibit 10.13 Residuals from the ARIMA(0,1,1)×(0,1,1)$_{12}$ Model




```
> win.graph(width=3, height=3,pointsize=8)
> hist(window(rstandard(m1.co2),start=c(1995,2)),
    xlab='Standardized Residuals')
```

Exhibit 10.14 displays the QQ-normal plot for the residuals.

Exhibit 10.14 Residuals: ARIMA(0,1,1)×(0,1,1)$_{12}$ Model




```
> win.graph(width=2.5,height=2.5,pointsize=8)
> qqnorm(window(rstandard(m1.co2),start=c(1995,2)))
> qqline(window(rstandard(m1.co2),start=c(1995,2)))
```

Here we again see the one outlier in the upper tail, but the Shapiro-Wilk test of normality has a test statistic of $W = 0.982$, leading to a p-value of 0.11, and normality is not rejected at any of the usual significance levels.

As one further check on the model, we consider overfitting with an ARIMA(0,1,2) $\times(0,1,1)_{12}$ model with the results shown in Exhibit 10.15.

Exhibit 10.15 ARIMA(0,1,2)×(0,1,1)$_{12}$ Overfitted Model

Coefficient	θ_1	θ_2	Θ
Estimate	0.5714	0.0165	0.8274
Standard error	0.0897	0.0948	0.1224

$\hat{\sigma}^2_{\tilde{e}} = 0.5427$: log-likelihood = −139.52, AIC = 285.05

```
> m2.co2=arima(co2,order=c(0,1,2),seasonal=list(order=c(0,1,1),
    period=12))
> m2.co2
```

When we compare these results with those reported in Exhibit 10.10 on page 237, we see that the estimates of θ_1 and Θ have changed very little—especially when the size of the standard errors is taken into consideration. In addition, the estimate of the new parameter, θ_2, is not statistically different from zero. Note also that the estimate $\hat{\sigma}^2_{\tilde{e}}$ and the log-likelihood have not changed much while the AIC has actually increased.

The ARIMA(0,1,1)×(0,1,1)$_{12}$ model was popularized in the first edition of the seminal book of Box and Jenkins (1976) when it was found to characterize the logarithms of

a monthly airline passenger time series. This model has come to be known as the **airline model**. We ask you to analyze the original airline data in the exercises.

10.5 Forecasting Seasonal Models

Computing forecasts with seasonal ARIMA models is, as expected, most easily carried out recursively using the difference equation form for the model, as in Equations (9.3.28), (9.3.29) on page 199 and (9.3.40) on page 201. For example, consider the model ARIMA$(0,1,1)\times(1,0,1)_{12}$.

$$Y_t - Y_{t-1} = \Phi(Y_{t-12} - Y_{t-13}) + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \tag{10.5.1}$$

which we rewrite as

$$Y_t = Y_{t-1} + \Phi Y_{t-12} - \Phi Y_{t-13} + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \tag{10.5.2}$$

The one-step-ahead forecast from origin t is then

$$\hat{Y}_t(1) = Y_t + \Phi Y_{t-11} - \Phi Y_{t-12} - \theta e_t - \Theta e_{t-11} + \theta\Theta e_{t-12} \tag{10.5.3}$$

and the next one is

$$\hat{Y}_t(2) = \hat{Y}_t(1) + \Phi Y_{t-10} - \Phi Y_{t-11} - \Theta e_{t-10} + \theta\Theta e_{t-11} \tag{10.5.4}$$

and so forth. The noise terms e_{t-13}, e_{t-12}, e_{t-11},…, e_t (as residuals) will enter into the forecasts for lead times $\ell = 1, 2,\ldots, 13$, but for $\ell > 13$ the autoregressive part of the model takes over and we have

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell-1) + \Phi\hat{Y}_t(\ell-12) - \Phi\hat{Y}_t(\ell-13) \quad \text{for } \ell > 13 \tag{10.5.5}$$

To understand the general nature of the forecasts, we consider several special cases.

Seasonal AR$(1)_{12}$

The seasonal AR$(1)_{12}$ model is

$$Y_t = \Phi Y_{t-12} + e_t \tag{10.5.6}$$

Clearly, we have

$$\hat{Y}_t(\ell) = \Phi\hat{Y}_t(\ell-12) \tag{10.5.7}$$

However, iterating back on ℓ, we can also write

$$\hat{Y}_t(\ell) = \Phi^{k+1}Y_{t+r-11} \tag{10.5.8}$$

where k and r are defined by $\ell = 12k + r + 1$ with $0 \le r < 12$ and $k = 0, 1, 2,\ldots$. In other words, k is the integer part of $(\ell-1)/12$ and $r/12$ is the fractional part of $(\ell-1)/12$. If our last observation is in December, then the next January value is forecast as Φ times the last observed January value, February is forecast as Φ times the last observed February

value, and so on. Two Januarys ahead is forecast as Φ^2 times the last observed January. Looking just at January values, the forecasts into the future will decay exponentially at a rate determined by the magnitude of Φ. All of the forecasts for each month will behave similarly but with different initial forecasts depending on the particular month under consideration.

Using Equation (9.3.38) on page 201 and the fact that the ψ-weights are nonzero only for multiple of 12, namely,

$$\psi_j = \begin{cases} \Phi^{j/12} & \text{for } j = 0, 12, 24, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{10.5.9}$$

we have that the forecast error variance can be written as

$$Var(e_t(\ell)) = \left[\frac{1 - \Phi^{2k+2}}{1 - \Phi^2}\right]\sigma_e^2 \tag{10.5.10}$$

where, as before, k is the integer part of $(\ell - 1)/12$.

Seasonal MA(1)$_{12}$

For the seasonal MA(1)$_{12}$ model, we have

$$Y_t = e_t - \Theta e_{t-12} + \theta_0 \tag{10.5.11}$$

In this case, we see that

$$\left.\begin{aligned} \hat{Y}_t(1) &= -\Theta e_{t-11} + \theta_0 \\ \hat{Y}_t(2) &= -\Theta e_{t-10} + \theta_0 \\ &\vdots \\ \hat{Y}_t(12) &= -\Theta e_t + \theta_0 \end{aligned}\right\} \tag{10.5.12}$$

and

$$\hat{Y}_t(\ell) = \theta_0 \text{ for } \ell > 12 \tag{10.5.13}$$

Here we obtain different forecasts for the months of the first year, but from then on all forecasts are given by the process mean.

For this model, $\psi_0 = 1$, $\psi_{12} = -\Theta$, and $\psi_j = 0$ otherwise. Thus, from Equation (9.3.38) on page 201,

$$Var(e_t(\ell)) = \begin{cases} \sigma_e^2 & 1 \le \ell \le 12 \\ (1 + \Theta^2)\sigma_e^2 & 12 < \ell \end{cases} \tag{10.5.14}$$

ARIMA(0,0,0)×(0,1,1)$_{12}$

The ARIMA(0,0,0)×(0,1,1)$_{12}$ model is

$$Y_t - Y_{t-12} = e_t - \Theta e_{t-12} \tag{10.5.15}$$

or

$$Y_{t+\ell} = Y_{t+\ell-12} + e_{t+\ell} - \Theta e_{t+\ell-12}$$

so that

$$\left.\begin{aligned} \hat{Y}_t(1) &= Y_{t-11} - \Theta e_{t-11} \\ \hat{Y}_t(2) &= Y_{t-10} - \Theta e_{t-10} \\ &\vdots \\ \hat{Y}_t(12) &= Y_t - \Theta e_t \end{aligned}\right\} \tag{10.5.16}$$

and then

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell - 12) \quad \text{for} \quad \ell > 12 \tag{10.5.17}$$

It follows that all Januarys will forecast identically, all Februarys identically, and so forth.

If we invert this model, we find that

$$Y_t = (1-\Theta)(Y_{t-12} + \Theta Y_{t-24} + \Theta^2 Y_{t-36} + \cdots) + e_t$$

Consequently, we can write

$$\left.\begin{aligned} \hat{Y}_t(1) &= (1-\Theta)\sum_{j=0}^{\infty} \Theta^j Y_{t-11-12j} \\ \hat{Y}_t(2) &= (1-\Theta)\sum_{j=0}^{\infty} \Theta^j Y_{t-10-12j} \\ &\vdots \\ \hat{Y}_t(12) &= (1-\Theta)\sum_{j=0}^{\infty} \Theta^j Y_{t-12j} \end{aligned}\right\} \tag{10.5.18}$$

From this representation, we see that the forecast for each January is an exponentially weighted moving average of all observed Januarys, and similarly for each of the other months.

In this case, we have $\psi_j = 1 - \Theta$ for $j = 12, 24, \ldots$, and zero otherwise. The forecast error variance is then

$$Var(e_t(\ell)) = [1 + k(1-\Theta)^2]\sigma_e^2 \tag{10.5.19}$$

where k is the integer part of $(\ell - 1)/12$.

ARIMA(0,1,1)×(0,1,1)$_{12}$

For the ARIMA$(0,1,1)\times(0,1,1)_{12}$ model

$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + e_t - \theta e_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13} \tag{10.5.20}$$

the forecasts satisfy

$$\left.\begin{aligned}
\hat{Y}_t(1) &= Y_t + Y_{t-11} - Y_{t-12} - \theta e_t - \Theta e_{t-11} + \theta\Theta e_{t-12} \\
\hat{Y}_t(2) &= \hat{Y}_t(1) + Y_{t-10} - Y_{t-11} - \Theta e_{t-10} + \theta\Theta e_{t-11} \\
&\vdots \\
\hat{Y}_t(12) &= \hat{Y}_t(11) + Y_t - Y_{t-1} - \Theta e_t + \theta\Theta e_{t-1} \\
\hat{Y}_t(13) &= \hat{Y}_t(12) + \hat{Y}_t(1) - Y_t + \theta\Theta e_t
\end{aligned}\right\} \qquad (10.5.21)$$

and

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell-1) + \hat{Y}_t(\ell-12) - \hat{Y}_t(\ell-13) \text{ for } \ell > 13 \qquad (10.5.22)$$

To understand the general pattern of these forecasts, we can use the representation

$$\hat{Y}_t(\ell) = A_1 + A_2\ell + \sum_{j=0}^{6}\left[B_{1j}\cos\left(\frac{2\pi j\ell}{12}\right) + B_{2j}\sin\left(\frac{2\pi j\ell}{12}\right)\right] \qquad (10.5.23)$$

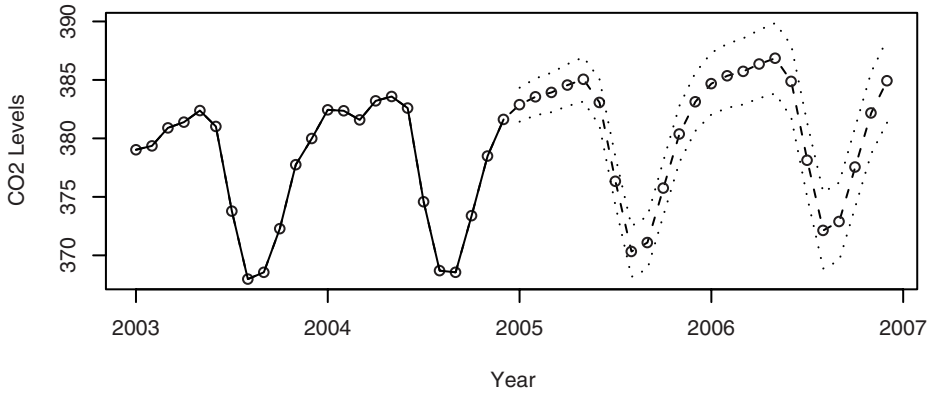
where the A's and B's are dependent on Y_t, Y_{t-1},…, or, alternatively, determined from the initial forecasts $\hat{Y}_t(1)$, $\hat{Y}_t(2)$,…, $\hat{Y}_t(13)$. This result follows from the general theory of difference equations and involves the roots of $(1-x)(1-x^{12}) = 0$.

Notice that Equation (10.5.23) reveals that the forecasts are composed of a linear trend in the lead time plus a sum of periodic components. However, the coefficients A_i and B_{ij} are more dependent on recent data than on past data and will adapt to changes in the process as our forecast origin changes and the forecasts are updated. This is in stark contrast to forecasting with deterministic time trend plus seasonal components, where the coefficients depend rather equally on both recent and past data and remain the same for all future forecasts.

Prediction Limits

Prediction limits are obtained precisely as in the nonseasonal case. We illustrate this with the carbon dioxide time series. Exhibit 10.16 shows the forecasts and 95% forecast limits for a lead time of two years for the ARIMA$(0,1,1)\times(0,1,1)_{12}$ model that we fit. The last two years of observed data are also shown. The forecasts mimic the stochastic periodicity in the data quite well, and the forecast limits give a good feeling for the precision of the forecasts.

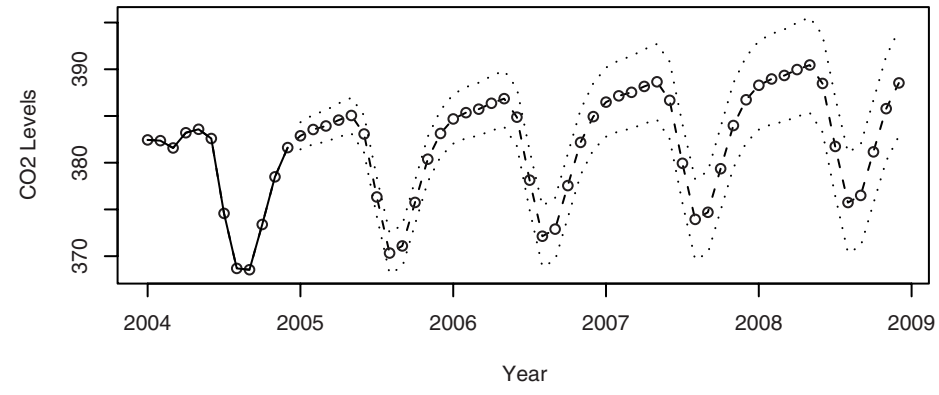
Exhibit 10.16 Forecasts and Forecast Limits for the CO_2 Model




```
> win.graph(width=4.875,height=3,pointsize=8)
> plot(m1.co2,n1=c(2003,1),n.ahead=24,xlab='Year',type='o',
    ylab='CO2 Levels')
```

Exhibit 10.17 displays the last year of observed data and forecasts out four years. At this lead time, it is easy to see that the forecast limits are getting wider, as there is more uncertainty in the forecasts.

Exhibit 10.17 Long-Term Forecasts for the CO_2 Model




```
> plot(m1.co2,n1=c(2004,1),n.ahead=48,xlab='Year',type='b',
    ylab='CO2 Levels')
```

10.6 Summary

Multiplicative seasonal ARIMA models provide an economical way to model time series whose seasonal tendencies are not as regular as we would have with a deterministic seasonal trend model which we covered in Chapter 3. Fortunately, these models are simply special ARIMA models so that no new theory is needed to investigate their properties. We illustrated the special nature of these models with a thorough modeling of an actual time series.

EXERCISES

10.1 Based on quarterly data, a seasonal model of the form

$$Y_t = Y_{t-4} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

has been fit to a certain time series.

(a) Find the first four ψ-weights for this model.

(b) Suppose that $\theta_1 = 0.5$, $\theta_2 = -0.25$, and $\sigma_e = 1$. Find forecasts for the next four quarters if data for the last four quarters are

Quarter	I	II	III	IV
Series	25	20	25	40
Residual	2	1	2	3

(c) Find 95% prediction intervals for the forecasts in part (b).

10.2 An AR model has AR characteristic polynomial

$$(1 - 1.6x + 0.7x^2)(1 - 0.8x^{12})$$

(a) Is the model stationary?

(b) Identify the model as a certain seasonal ARIMA model.

10.3 Suppose that $\{Y_t\}$ satisfies

$$Y_t = a + bt + S_t + X_t$$

where S_t is deterministic and periodic with period s and $\{X_t\}$ is a seasonal ARIMA$(p,0,q)\times(P,1,Q)_s$ series. What is the model for $W_t = Y_t - Y_{t-s}$?

10.4 For the seasonal model $Y_t = \Phi Y_{t-4} + e_e - \theta e_{t-1}$ with $|\Phi| < 1$, find γ_0 and ρ_k.

10.5 Identify the following as certain multiplicative seasonal ARIMA models:

(a) $Y_t = 0.5Y_{t-1} + Y_{t-4} - 0.5Y_{t-5} + e_t - 0.3e_{t-1}$.

(b) $Y_t = Y_{t-1} + Y_{t-12} + Y_{t-13} + e_t - 0.5e_{t-1} - e_{t-12} + 0.25e_{t-13}$.

10.6 Verify Equations (10.2.11) on page 232.

10.7 Suppose that the process $\{Y_t\}$ develops according to $Y_t = Y_{t-4} + e_t$ with $Y_t = e_t$ for $t = 1, 2, 3$, and 4.
(a) Find the variance function for $\{Y_t\}$.
(b) Find the autocorrelation function for $\{Y_t\}$.
(c) Identify the model for $\{Y_t\}$ as a certain seasonal ARIMA model.

10.8 Consider the Alert, Canada, monthly carbon dioxide time series shown in Exhibit 10.1 on page 227. The data are in the file named co2.
(a) Fit a deterministic seasonal means plus linear time trend model to these data. Are any of the regression coefficients "statistically significant"?
(b) What is the multiple R-squared for this model?
(c) Now calculate the sample autocorrelation of the residuals from this model. Interpret the results.

10.9 The monthly airline passenger time series, first investigated in Box and Jenkins (1976), is considered a classic time series. The data are in the file named airline.
(a) Display the time series plots of both the original series and the logarithms of the series. Argue that taking logs is an appropriate transformation.
(b) Display and interpret the time series plots of the first difference of the logged series.
(c) Display and interpret the time series plot of the seasonal difference of the first difference of the logged series.
(d) Calculate and interpret the sample ACF of the seasonal difference of the first difference of the logged series.
(e) Fit the "airline model" (ARIMA$(0,1,1)\times(0,1,1)_{12}$) to the logged series.
(f) Investigate diagnostics for this model, including autocorrelation and normality of the residuals.
(g) Produce forecasts for this series with a lead time of two years. Be sure to include forecast limits.

10.10 Exhibit 5.8 on page 99 displayed the monthly electricity generated in the United States. We argued there that taking logarithms was appropriate for modeling. Exhibit 5.10 on page 100 showed the time series plot of the first differences for this series. The filename is electricity.
(a) Calculate the sample ACF of the first difference of the logged series. Is the seasonality visible in this display?
(b) Plot the time series of seasonal difference and first difference of the logged series. Does a stationary model seem appropriate now?
(c) Display the sample ACF of the series after a seasonal difference and a first difference have been taken of the logged series. What model(s) might you consider for the electricity series?

10.11 The quarterly earnings per share for 1960–1980 of the U.S. company Johnson & Johnson, are saved in the file named JJ.

(a) Plot the time series and also the logarithm of the series. Argue that we should transform by logs to model this series.

(b) The series is clearly not stationary. Take first differences and plot that series. Does stationarity now seem reasonable?

(c) Calculate and graph the sample ACF of the first differences. Interpret the results.

(d) Display the plot of seasonal differences and the first differences. Interpret the plot. Recall that for quarterly data, a season is of length 4.

(e) Graph and interpret the sample ACF of seasonal differences with the first differences.

(f) Fit the model ARIMA$(0,1,1)\times(0,1,1)_4$, and assess the significance of the estimated coefficients.

(g) Perform all of the diagnostic tests on the residuals.

(h) Calculate and plot forecasts for the next two years of the series. Be sure to include forecast limits.

10.12 The file named boardings contains monthly data on the number of people who boarded transit vehicles (mostly light rail trains and city buses) in the Denver, Colorado, region for August 2000 through December 2005.

(a) Produce the time series plot for these data. Be sure to use plotting symbols that will help you assess seasonality. Does a stationary model seem reasonable?

(b) Calculate and plot the sample ACF for this series. At which lags do you have significant autocorrelation?

(c) Fit an ARMA$(0,3)\times(1,0)_{12}$ model to these data. Assess the significance of the estimated coefficients.

(d) Overfit with an ARMA$(0,4)\times(1,0)_{12}$ model. Interpret the results.

CHAPTER 11

TIME SERIES REGRESSION MODELS

In this chapter, we introduce several useful ideas that incorporate external information into time series modeling. We start with models that include the effects of interventions on time series' normal behavior. We also consider models that assimilate the effects of outliers—observations, either in the observed series or in the error terms, that are highly unusual relative to normal behavior. Lastly, we develop methods to look for and deal with spurious correlation—correlation between series that is artificial and will not help model or understand the time series of interest. We will see that prewhitening of series helps us find meaningful relationships.

11.1 Intervention Analysis

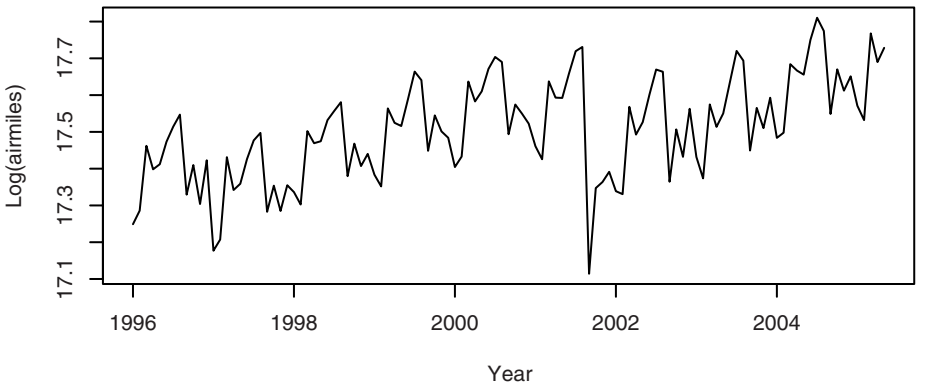
Exhibit 11.1 shows the time plot of the logarithms of monthly airline passenger-miles in the United States from January 1996 through May 2005. The time series is highly seasonal, displaying the fact that air traffic is generally higher during the summer months and the December holidays and lower in the winter months.[†] Also, air traffic was increasing somewhat linearly overall until it had a sudden drop in September 2001. The sudden drop in the number of air passengers in September 2001 and several months thereafter was triggered by the terrorist acts on September 11, 2001, when four planes were hijacked, three of which were crashed into the twin towers of the World Trade Center and the Pentagon and the fourth into a rural field in Pennsylvania. The terrorist attacks of September 2001 deeply depressed air traffic around that period, but air traffic gradually regained the losses as time went on. This is an example of an intervention that results in a change in the trend of a time series.

Intervention analysis, introduced by Box and Tiao (1975), provides a framework for assessing the effect of an intervention on a time series under study. It is assumed that the intervention affects the process by changing the mean function or trend of a time series. Interventions can be natural or man-made. For example, some animal population levels crashed to a very low level in a particular year because of extreme climate in that year. The postcrash annual population level may then be expected to be different from that in the precrash period. Another example is the increase of the speed limit from 65 miles per hour to 70 miles per hour on an interstate highway. This may make driving on

[†] In the exercises, we ask you to display the time series plot using seasonal plotting symbols on a full-screen graph, where the seasonality is quite easy to see.

the highway more dangerous. On the other hand, drivers may stay on the highway for a shorter length of time because of the faster speed, so the net effect of the increased speed limit change is unclear. The effect of the increase in speed limit may be studied by analyzing the mean function of some accident time series data; for example, the quarterly number of fatal car accidents on some segment of an interstate highway. (Note that the autocovariance function of the time series might also be changed by the intervention, but this possibility will not be pursued here.)

Exhibit 11.1 Monthly U.S. Airline Miles: January 1996 through May 2005




```
> win.graph(width=4.875,height=2.5,pointsize=8)
> data(airmiles)
> plot(log(airmiles),ylab='Log(airmiles)',xlab='Year')
```

We first consider the simple case of a single intervention. The general model for the time series $\{Y_t\}$, perhaps after suitable transformation, is given by

$$Y_t = m_t + N_t \tag{11.1.1}$$

where m_t is the change in the mean function and N_t is modeled as some ARIMA process, possibly seasonal. The process $\{N_t\}$ represents the underlying time series were there no intervention. It is referred to as the natural or unperturbed process, and it may be stationary or nonstationary, seasonal or nonseasonal. Suppose the time series is subject to an intervention that takes place at time T. Before T, m_t is assumed to be identically zero. The time series $\{Y_t, t < T\}$ is referred to as the **preintervention data** and can be used to specify the model for the unperturbed process N_t.

Based on subject matter considerations, the effect of the intervention on the mean function can often be specified up to some parameters. A useful function in this specification is the **step function**

$$S_t^{(T)} = \begin{cases} 1, \text{ if } t \geq T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.2}$$

that is 0 during the preintervention period and 1 throughout the postintervention period. The **pulse function**

$$P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} \tag{11.1.3}$$

equals 1 at $t = T$ and 0 otherwise. That is, $P_t^{(T)}$ is the indicator or dummy variable flagging the time that the intervention takes place. If the intervention results in an immediate and permanent shift in the mean function, the shift can be modeled as

$$m_t = \omega S_t^{(T)} \tag{11.1.4}$$

where ω is the unknown permanent change in the mean due to the intervention. Testing whether $\omega = 0$ or not is similar to testing whether the population means are the same with data in the form of two independent random samples from the two populations. However, the major difference here is that the pre- and postintervention data cannot generally be assumed to be independent and identically distributed. The inherent serial correlation in the data makes the problem more interesting but at the same time more difficult. If there is a **delay** of d time units before the intervention takes effect and d is known, then we can specify

$$m_t = \omega S_{t-d}^{(T)} \tag{11.1.5}$$

In practice, the intervention may affect the mean function gradually, with its full force reflected only in the long run. This can be modeled by specifying m_t as an AR(1)-type model with the error term replaced by a multiple of the lag 1 of $S_t^{(T)}$:

$$m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)} \tag{11.1.6}$$

with the initial condition $m_0 = 0$. After some algebra, it can be shown that

$$m_t = \begin{cases} \omega \dfrac{1-\delta^{t-T}}{1-\delta}, \text{ for } t > T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.7}$$

Often δ is selected in the range $1 > \delta > 0$. In that case, m_t approaches $\omega/(1-\delta)$ for large t, which is the ultimate change (gain or loss) for the mean function. Half of the ultimate change is attained when $1 - \delta^{t-T} = 0.5$; that is, when $t = T + \log(0.5)/\log(\delta)$. The duration $\log(0.5)/\log(\delta)$ is called the **half-life** of the intervention effect, and the shorter it is, the quicker the ultimate change is felt by the system. Exhibit 11.2 displays the half-life as a function of δ, which shows that the half-life increases with δ. Indeed, the half-life becomes infinitely large when δ approaches 1.

Exhibit 11.2 Half-life based on an AR(1) Process with Step Function Input

δ	0.2	0.4	0.6	0.8	0.9	1
Half-life	0.43	0.76	1.46	3.11	6.58	∞

It is interesting to note the limiting case when $\delta = 1$. Then $m_t = \omega(T - t)$ for $t \geq T$ and 0 otherwise. The time sequence plot of m_t displays the shape of a ramp with slope ω. This specification implies that the intervention changes the mean function linearly in the postintervention period. This ramp effect (with a one time unit delay) is shown in Exhibit 11.3 (c).

Short-lived intervention effects may be specified using the pulse dummy variable

$$P_t^{(T)} = \begin{cases} 1, \text{ if } t = T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.8}$$

For example, if the intervention impacts the mean function only at $t = T$, then

$$m_t = \omega P_t^{(T)} \tag{11.1.9}$$

Intervention effects that die out gradually may be specified via the AR(1)-type specification

$$m_t = \delta m_{t-1} + \omega P_t^{(T)} \tag{11.1.10}$$

That is, $m_t = \omega\delta^{T-t}$ for $t \geq T$ so that the mean changes immediately by an amount ω and subsequently the change in the mean decreases geometrically by the common factor of δ; see Exhibit 11.4 (a). Delayed changes can be incorporated by lagging the pulse function. For example, if the change in the mean takes place after a delay of one time unit and the effect dies out gradually, we can specify

$$m_t = \delta m_{t-1} + \omega P_{t-1}^{(T)} \tag{11.1.11}$$

Again, we assume the initial condition $m_0 = 0$.

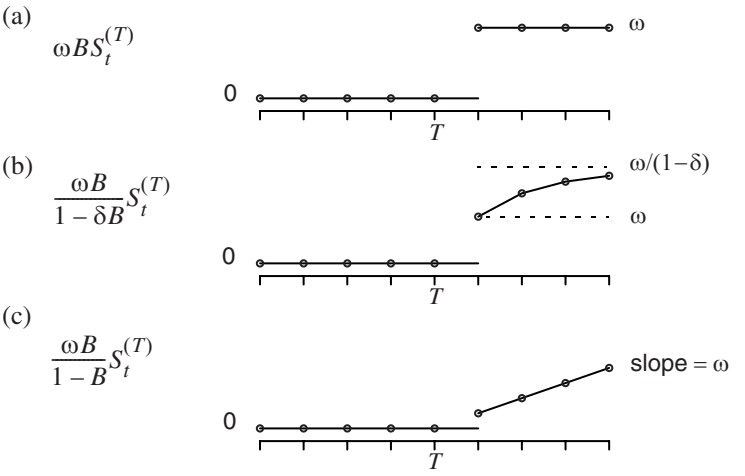
It is useful to write[†] the preceding model in terms of the backshift operator B, where $Bm_t = m_{t-1}$ and $BP_t^{(T)} = P_{t-1}^{(T)}$. Then $(1 - \delta B)m_t = \omega B P_t^{(T)}$. Or, we can write

$$m_t = \frac{\omega B}{1 - \delta B} P_t^{(T)} \tag{11.1.12}$$

Recall $(1 - B)S_t^{(T)} = P_t^{(T)}$, which can be rewritten as $S_t^{(T)} = \frac{1}{1 - B} P_t^{(T)}$.

[†] The remainder of this chapter makes use of the backshift operator introduced in Appendix D on page 106. You may want to review that appendix before proceeding further.

Exhibit 11.3 Some Common Models for Step Response Interventions (All are shown with a delay of 1 time unit)



Several specifications can be combined to model more sophisticated intervention effects.

For example,

$$m_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)} \tag{11.1.13}$$

depicts the situation displayed in Exhibit 11.4 (b) where ω_1 and ω_2 are both greater than zero, and

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)} \tag{11.1.14}$$

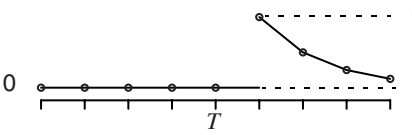
may model situations like Exhibit 11.4 (c) with ω_1 and ω_2 both negative. This last case may model the interesting situation where a special sale may cause strong rush buying, initially so much so that the sale is followed by depressed demand. More generally, we can model the change in the mean function by an ARMA-type specification

$$m_t = \frac{\omega(B)}{\delta(B)} P_t^{(T)} \tag{11.1.15}$$

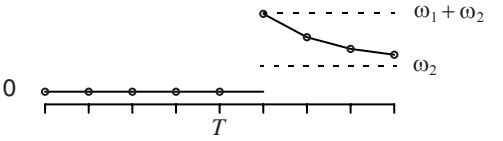
where $\omega(B)$ and $\delta(B)$ are some polynomials in B. Because $(1-B)S_t^{(T)} = P_t^{(T)}$, the model for m_t can be specified in terms of either the pulse or step dummy variable.

Exhibit 11.4 Some Common Models for Pulse Response Interventions (All are shown with a delay of 1 time unit)

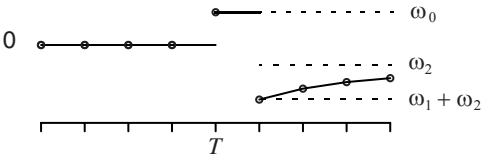
(a) $\dfrac{\omega B}{1-\delta B}P_t^{(T)}$



(b) $\left[\dfrac{\omega_1 B}{1-\delta B}+\dfrac{\omega_2 B}{1-B}\right]P_t^{(T)}$



(c) $\left[\omega_0+\dfrac{\omega_1 B}{1-\delta B}+\dfrac{\omega_2 B}{1-B}\right]P_t^{(T)}$



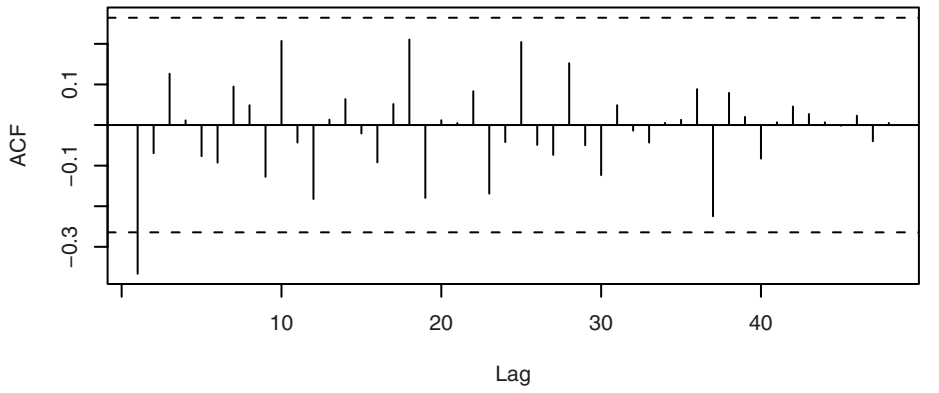
Estimation of the parameters of an intervention model may be carried out by the method of maximum likelihood estimation. Indeed, $Y_t - m_t$ is a seasonal ARIMA process so that the likelihood function equals the joint pdf of $Y_t - m_t$, $t = 1, 2,\ldots, n$, which can be computed by methods studied in Chapter 7 or else by the state space modeling methods of Appendix H on page 222.

We now revisit the monthly passenger-airmiles data. Recall that the terrorist acts in September 2001 had lingering depressing effects on air traffic. The intervention may be specified as an AR(1) process with the pulse input at September 2001. But the unexpected turn of events in September 2001 had a strong instantaneous chilling effect on air traffic. Thus, we model the intervention effect (the 9/11 effect) as

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B}P_t^{(T)}$$

where T denotes September 2001. In this specification, $\omega_0 + \omega_1$ represents the instantaneous 9/11 effect, and, for $k \geq 1$, $\omega_1(\omega_2)^k$ gives the 9/11 effect k months afterward. It remains to specify the seasonal ARIMA structure of the underlying unperturbed process. Based on the preintervention data, an ARIMA$(0,1,1)\times(0,1,0)_{12}$ model was tentatively specified for the unperturbed process; see Exhibit 11.5.

Exhibit 11.5 Sample ACF for $(1-B)(1-B^{12})$ Log(Air Passenger Miles) Over the Preintervention Period




```
> acf(as.vector(diff(diff(window(log(airmiles),end=c(2001,8)),
    12))),lag.max=48)
```

Model diagnostics of the fitted model suggested that a seasonal MA(1) coefficient was needed and the existence of some *additive* outliers occurring in December 1996, January 1997, and December 2002. (Outliers will be discussed in more detail later; here additive outliers may be regarded as interventions of unknown nature that have a pulse response function.) Hence, the model is specified as an ARIMA$(0,1,1)\times(0,1,1)_{12}$ plus the 9/11 intervention and three additive outliers. The fitted model is summarized in Exhibit 11.6.

Exhibit 11.6 Estimation of Intervention Model for Logarithms of Air Miles (Standard errors are shown below the estimates)

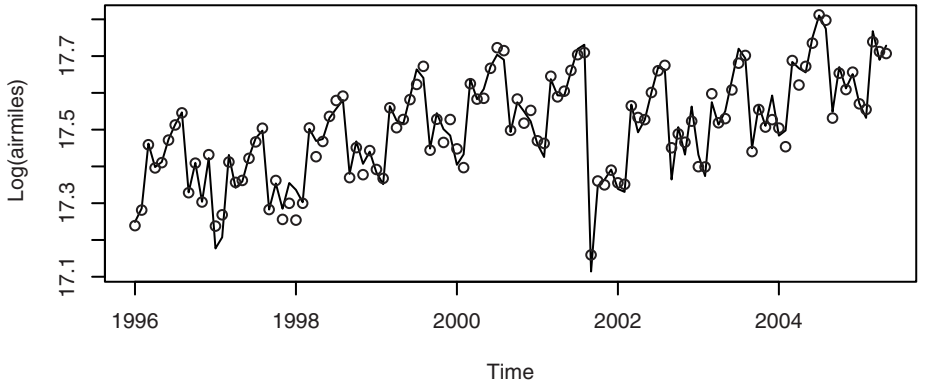
θ	Θ	Dec96	Jan97	Dec02	ω_0	ω_1	ω_2
0.383	0.650	0.099	−0.069	0.081	−0.095	−0.27	0.814
(0.093)	(0.119)	(0.023)	(0.022)	(0.020)	(0.046)	(0.044)	(0.098)

σ^2 estimated as 0.000672: log-likelihood = 219.99, AIC= −423.98

```
> air.m1=arimax(log(airmiles),order=c(0,1,1),
    seasonal=list(order=c(0,1,1),period=12),
    xtransf=data.frame(I911=1*(seq(airmiles)==69),
    I911=1*(seq(airmiles)==69)),transfer=list(c(0,0),c(1,0)),
    xreg=data.frame(Dec96=1*(seq(airmiles)==12),
    Jan97=1*(seq(airmiles)==13),Dec02=1*(seq(airmiles)==84)),
    method='ML')
> air.m1
```

Model diagnostics suggested that the fitted model above provides a good fit to the data. The open circles in the time series plot shown in Exhibit 11.7 represent the fitted values from the final estimated model. They indicate generally good agreement between the model and the data.

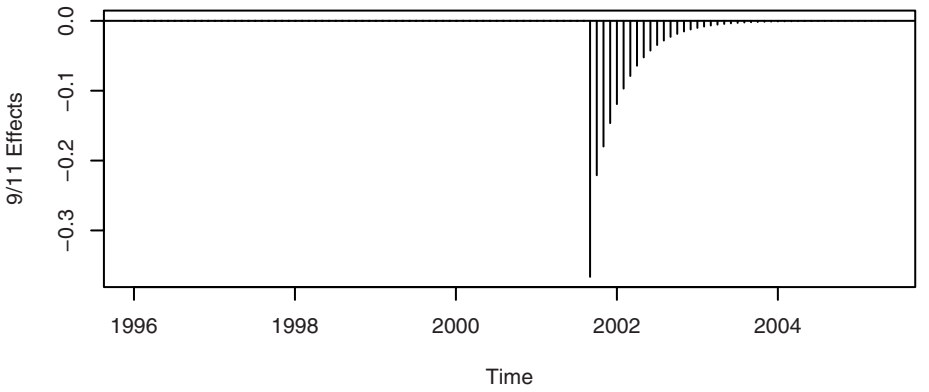
Exhibit 11.7 Logs of Air Passenger Miles and Fitted Values




```
> plot(log(airmiles),ylab='Log(airmiles)')
> points(fitted(air.m1))
```

The fitted model estimates that the 9/11 intervention reduced air traffic by 31% = $\{1 - \exp(-0.0949-0.2715)\} \times 100\%$ in September 2001, and air traffic k months later was lowered by $\{1 - \exp(-0.2715 \times 0.8139^{k})\} \times 100\%$. Exhibit 11.8 graphs the estimated 9/11 effects on air traffic, which indicate that air traffic regained its losses toward the end of 2003.

Exhibit 11.8 The Estimated 9/11 Effects for the Air Passenger Series




```
> Nine11p=1*(seq(airmiles)==69)
> plot(ts(Nine11p*(-0.0949)+
```

```
filter(Nine11p,filter=.8139,method='recursive', side=1)*
(-0.2715),frequency=12,start=1996),ylab='9/11 Effects',
type='h'); abline(h=0)
```

11.2 Outliers

Outliers refer to atypical observations that may arise because of measurement and/or copying errors or because of abrupt, short-term changes in the underlying process. For time series, two kinds of outliers can be distinguished, namely **additive** outliers and **innovative** outliers. These two kinds of outliers are often abbreviated as AO and IO, respectively. An additive outlier occurs at time T if the underlying process is perturbed additively at time T so that the data equal

$$Y_t' = Y_t + \omega_A P_t^{(T)} \tag{11.2.1}$$

where $\{Y_t\}$ is the unperturbed process. Henceforth in this section, Y' denotes the observed process that may be affected by some outliers and Y the unperturbed process should there be no outliers. Thus, $Y_T' = Y_T + \omega_A$ but $Y_t' = Y_t$ otherwise, so the time series is only affected at time T if it has an additive outlier at T. An additive outlier can also be treated as an intervention that has a pulse response at T so that $m_t = \omega_A P_t^{(T)}$.

On the other hand, an innovative outlier occurs at time t if the error (also known as an innovation) at time t is perturbed (that is, the errors equal $e_t' = e_t + \omega_I P_t^{(T)}$, where e_t is a zero-mean white noise process). So, $e_T' = e_T + \omega_I$ but $e_t' = e_t$ otherwise. Suppose that the unperturbed process is stationary and admits an MA(∞) representation

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots$$

Consequently, the perturbed process can be written

$$\begin{aligned} Y_t' &= e_t' + \psi_1 e_{t-1}' + \psi_2 e_{t-2}' + \cdots \\ &= [e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots] + \psi_{t-T}\omega_I \end{aligned}$$

or

$$Y_t' = Y_t + \psi_{t-T}\omega_I \tag{11.2.2}$$

where $\psi_0 = 1$ and $\psi_j = 0$ for negative j. Thus, an innovative outlier at T perturbs all observations on and after T, although with diminishing effect, as the observation is further away from the origin of the outlier.

To detect whether an observation is an AO or IO, we use the AR(∞) representation of the unperturbed process to define the residuals:

$$a_t = Y_t' - \pi_1 Y_{t-1}' - \pi_2 Y_{t-2}' - \cdots \tag{11.2.3}$$

For simplicity, we assume the process has zero mean and that the parameters are known. In practice, the unknown parameter values are replaced by their estimates from the possibly perturbed data. Under the null hypothesis of no outliers and for large samples, this

has a negligible effect on the properties of the test procedures described below. If the series has exactly one IO at time T, then the residual $a_T = \omega_I + e_T$ but $a_t = e_t$ otherwise. So ω_I can be estimated by $\tilde{\omega}_I = a_T$ with variance equal to σ^2. Thus, a test statistic for testing for an IO at T is

$$\lambda_{1,T} = \frac{a_T}{\sigma} \tag{11.2.4}$$

which has (approximately) a standard normal distribution under the null hypothesis that there are no outliers in the time series. When T is known beforehand, the observation in question is declared an outlier if the corresponding standardized residual exceeds 1.96 in magnitude at the 5% significance level. In practice, there is often no prior knowledge about T, and the test is applied to all observations. In addition, σ will need to be estimated. A simple conservative procedure is to use the Bonferroni rule for controlling the overall error rate of multiple tests. Let

$$\lambda_1 = \max_{1 \le t \le n} |\lambda_{1,t}| \tag{11.2.5}$$

be attained at $t = T$. Then the Tth observation is deemed an IO if λ_1 exceeds the upper $0.025/n \times 100$ percentile of the standard normal distribution. This procedure guarantees that there is at most a 5% probability of a false detection of an IO. Note that an outlier will inflate the maximum likelihood estimate of σ, so if there is no adjustment for outliers, the power of most tests is usually reduced. A robust estimate of the noise standard deviation may be used in lieu of the maximum likelihood estimate to increase the power of the test. For example, σ can be more robustly estimated by the mean absolute residual times $\sqrt{2/\pi}$.

The detection of an AO is more complex. Suppose that the process admits an AO at T and is otherwise free of outliers. Then it can be shown that

$$a_t = -\omega_A \pi_{t-T} + e_t \tag{11.2.6}$$

where $\pi_0 = -1$ and $\pi_j = 0$ for negative j. Hence, $a_t = e_t$ for $t < T$, $a_T = \omega_A + e_T$, $a_{T+1} = -\omega_A \pi_1 + e_{T+1}$, $a_{T+2} = -\omega_A \pi_2 + e_{T+2}$, and so forth. A least squares estimator of ω_A is

$$\tilde{\omega}_{T,A} = -\rho^2 \sum_{t=1}^{n} \pi_{t-T} a_t \tag{11.2.7}$$

where $\rho^2 = (1 + \pi_1^2 + \pi_2^2 + \cdots + \pi_{n-T}^2)^{-1}$, with the variance of the estimate being equal to $\rho^2\sigma^2$. We can then define

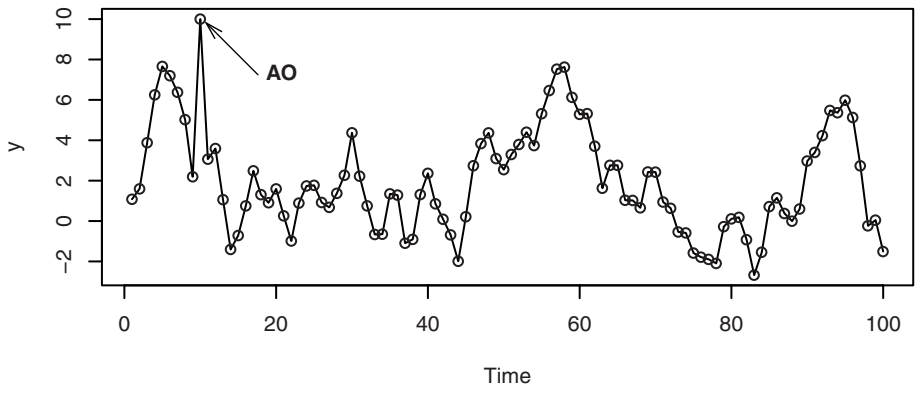
$$\lambda_{2,T} = \frac{\tilde{\omega}_{T,A}}{\rho\sigma} \tag{11.2.8}$$

as the test statistic for testing the null hypothesis that the time series has no outliers versus the alternative hypothesis of an AO at T. As before, ρ and σ will need to be estimated. The test statistic $\lambda_{2,T}$ is approximately distributed as $N(0,1)$ under the null hypothesis. Again, T is often unknown, and the test is applied repeatedly to each time point. The Bonferroni rule may again be applied to control the overall error rate. Furthermore, the nature of an outlier is not known beforehand. In the case where an outlier

is detected at T, it may be classified to be an IO if $|\lambda_{1,T}| > |\lambda_{2,T}|$ and an AO otherwise. See Chang et al. (1988) for another approach to classifying the nature of an outlier. When an outlier is found, it can be incorporated into the model, and the outlier-detection procedure can then be repeated with the refined model until no more outliers are found.

As a first example, we simulated a time series of length $n = 100$ from the ARIMA(1,0,1) model with $\phi = 0.8$ and $\theta = -0.5$. We then changed the 10th observation from −2.13 to 10 (that is, $\omega_A = 12.13$); see Exhibit 11.9. Based on the sample ACF, PACF and EACF, an AR(1) model was tentatively identified. Based on the Bonferroni rule, the 9th, 10th, and 11th observations were found to be possible additive outliers with the corresponding robustified test statistics being −3.54, 9.55, and −5.20. The test for IO revealed that the 10th and 11th observations may be IO, with the corresponding robustified test statistics being 7.11 and −6.64. Because among the tests for AO and IO the largest magnitude occurs for the test for AO at $T = 10$, the 10th observation was tentatively marked as an AO. Note that the nonrobustified test statistic for AO at $T = 10$ equals 7.49, which is substantially less than the more robust test value of 9.55, showing that robustifying the estimate of the noise standard deviation does increase the power of the test. After incorporating the AO in the model, no more outliers were found. However, the lag 1 residual ACF was significant, suggesting the need for an MA(1) component. Hence, an ARIMA(1,0,1) + AO at $T = 10$ model was fitted to the data. This model was found to have no additional outliers and passed all model diagnostic checks.

Exhibit 11.9 Simulated ARIMA(1,0,1) Process with an Additive Outlier




```
> The extensive R code for the simulation and analysis of this
    example may be found in the R code script file for Chapter 11.
```

For a real example, we return to the seasonal ARIMA$(0,1,1)\times(0,1,1)_{12}$ model that we fitted to the carbon dioxide time series in Chapter 10. The time series plot of the standardized residuals from this model, shown in Exhibit 10.11 on page 238, showed a suspiciously large standardized residual in September 1998. Calculation shows that there is no evidence of an additive outlier, as $\lambda_{2,t}$ is not significantly large for any t. However, the robustified $\lambda_1 = \max_{1\le t\le n}|\lambda_{1,t}| = 3.7527$, which is attained at $t = 57$, cor-

responding to September 1998. The Bonferroni critical value with $\alpha = 5\%$ and $n = 132$ is 3.5544. So our observed λ_1 is large enough to claim significance for an innovation outlier in September 1998. Exhibit 11.10 shows the results of fitting the ARIMA(0,1,1) $\times(0,1,1)_{12}$ model with an IO at $t = 57$ to the CO_2 time series. These results should be compared with the earlier results shown in Exhibit 10.10 on page 237, where the outlier was not taken into account. Notice that the estimates of θ and Θ have not changed very much, the AIC is better (that is, smaller), and the IO effect is highly significant. Diagnostics based on this model turn out to be excellent, no further outliers are detected, and we have a very adequate model for this seasonal time series.

Exhibit 11.10 ARIMA(0,1,1)×(0,1,1)$_{12}$ Model with IO at $t = 57$ for CO_2 Series

Coefficient	θ	Θ	IO-57
Estimate	0.5925	0.8274	2.6770
Standard Error	0.0775	0.1016	0.7246

$\hat{\sigma}^2_{\tilde{e}} = 0.4869$: log-likelihood = −133.08, AIC = 272.16

```
> m1.co2=arima(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
    period=12)); m1.co2
> detectAO(m1.co2); detectIO(m1.co2)
> m4.co2=arimax(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
    period=12),io=c(57)); m4.co2
```

11.3 Spurious Correlation

A main purpose of building a time series model is for forecasting, and the ARIMA model does this by exploiting the autocorrelation pattern in the data. Often, the time series under study may be related to, or led by, some other covariate time series. For example, Stige et al. (2006) found that pasture production in Africa is generally related to some climatic indices. In such cases, better understanding of the underlying process and/or more accurate forecasts may be achieved by incorporating relevant covariates into the time series model.

Let $Y = \{Y_t\}$ be the time series of the response variable and $X = \{X_t\}$ be a covariate time series that we hope will help explain or forecast Y. To explore the correlation structure between X and Y and their lead-led relationship, we define the cross-covariance function $\gamma_{t,s}(X,Y) = Cov(X_t,Y_s)$ for each pair of integers t and s. Stationarity of a univariate time series can be easily extended to the case of multivariate time series. For example, X and Y are jointly (weakly) stationary if their means are constant and the covariance $\gamma_{t,s}(X,Y)$ is a function of the time difference $t - s$. For jointly stationary processes, the **cross-correlation function** between X and Y at lag k can then be defined by $\rho_k(X,Y) = Corr(X_t,Y_{t-k}) = Corr(X_{t+k},Y_t)$. Note that if $Y = X$, the cross-correlation becomes the autocorrelation of Y at lag k. The coefficient $\rho_0(Y,X)$ measures the contemporaneous linear association between X and Y, whereas $\rho_k(X,Y)$ measures the linear association between X_t and that of Y_{t-k}. Recall that the autocorrelation function is an

even function, that is, $\rho_k(Y,Y) = \rho_{-k}(Y,Y)$. (This is because $Corr(Y_t,Y_{t-k}) = Corr(Y_{t-k},Y_t) = Corr(Y_t,Y_{t+k})$, by stationarity.) However, the cross-correlation function is generally not an even function since $Corr(X_t,Y_{t-k})$ need not equal $Corr(X_t,Y_{t+k})$.

As an illustration, consider the regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + e_t \tag{11.3.1}$$

where the X's are independent, identically distributed random variables with variance σ_X^2 and the e's are also white noise with variance σ_e^2 and are independent of the X's. It can be checked that the cross-correlation function (CCF) $\rho_k(X,Y)$ is identically zero except for lag $k = -d$, where

$$\rho_{-d}(X, Y) = \frac{\beta_1 \sigma_X}{\sqrt{\beta_1^2 \sigma_X^2 + \sigma_e^2}} \tag{11.3.2}$$

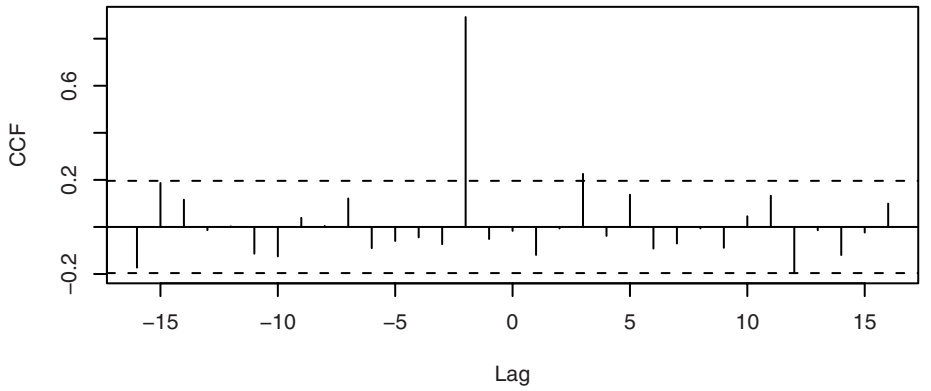
In this case, the theoretical CCF is nonzero only at lag $-d$, reflecting the fact that X is "leading" Y by d units of time. The CCF can be estimated by the **sample cross-correlation function** (sample CCF) defined by

$$r_k(X, Y) = \frac{\sum (X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum (X_t - \bar{X})^2} \sqrt{\sum (Y_t - \bar{Y})^2}} \tag{11.3.3}$$

where the summations are done over all data where the summands are available. The sample CCF becomes the sample ACF when $Y = X$. The covariate X is independent of Y if and only if $\beta_1 = 0$, in which case the sample autocorrelation $r_k(X,Y)$ is approximately normally distributed with zero mean and variance $1/n$, where n is the sample size—the number of pairs of (X_t,Y_t) available. Sample cross-correlations that are larger than $1.96/\sqrt{n}$ in magnitude are then deemed significantly different from zero.

We have simulated 100 pairs of (X_t,Y_t) from the model of Equation (11.3.1) with $d = 2$, $\beta_0 = 0$, and $\beta_1 = 1$. The X's and e's are generated as normal random variables distributed as $N(0,1)$ and $N(0,0.25)$, respectively. Theoretically, the CCF should then be zero except at lag -2, where it equals $\rho_{-2}(X, Y) = 1/\sqrt{1 + 0.25} = 0.8944$. Exhibit 11.11 shows the sample CCF of the simulated data, which is significant at lags -2 and 3. But the sample CCF at lag 3 is quite small and only marginally significant. Such a false alarm is not unexpected as the exhibit displays a total of 33 sample CCF values out of which we may expect $33 \times 0.05 = 1.65$ false alarms on average.

Exhibit 11.11 Sample Cross-Correlation from Equation (11.3.1) with $d = 2$




```
> win.graph(width=4.875,height=2.5,pointsize=8)
> set.seed(12345); X=rnorm(105); Y=zlag(X,2)+.5*rnorm(105)
> X=ts(X[-(1:5)],start=1,freq=1); Y=ts(Y[-(1:5)],start=1,freq=1)
> ccf(X,Y,ylab='CCF')
```

Even though X_{t-2} correlates with Y_t, the regression model considered above is rather restrictive, as X and Y are each white noise series. For stationary time series, the response variable and the covariate are each generally autocorrelated, and the error term of the regression model is also generally autocorrelated. Hence a more useful regression model is given by

$$Y_t = \beta_0 + \beta_1 X_{t-d} + Z_t \tag{11.3.4}$$

where Z_t may follow some ARIMA(p,d,q) model. Even if the processes X and Y are independent of each other ($\beta_1 = 0$), the autocorrelations in Y and X have the unfortunate consequence of implying that the sample CCF is no longer approximately $N(0,1/n)$. Under the assumption that both X and Y are stationary and that they are independent of each other, it turns out that the sample variance tends to be different from $1/n$. Indeed, it may be shown that the variance of $\sqrt{n}r_k(X, Y)$ is approximately

$$1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y) \tag{11.3.5}$$

where $\rho_k(X)$ is the autocorrelation of X at lag k and $\rho_k(Y)$ is similarly defined for the Y-process. For refinement of this asymptotic result, see Box et al. (1994, p. 413). Suppose X and Y are both AR(1) processes with AR(1) coefficients ϕ_X and ϕ_Y, respectively. Then $r_k(X,Y)$ is approximately normally distributed with zero mean, but the variance is now approximately equal to

$$\frac{1 + \phi_X\phi_Y}{n(1 - \phi_X\phi_Y)} \tag{11.3.6}$$

When both AR(1) coefficients are close to 1, the ratio of the sampling variance of $r_k(X,Y)$ to the nominal value of $1/n$ approaches infinity. Thus, the unquestioned use of the $1/n$ rule in deciding the significance of the sample CCF may lead to many more false positives than the nominal 5% error rate, even though the response and covariate time series are independent of each other. Exhibit 11.12 shows some numerical results for the case where $\phi_X = \phi_Y = \phi$.

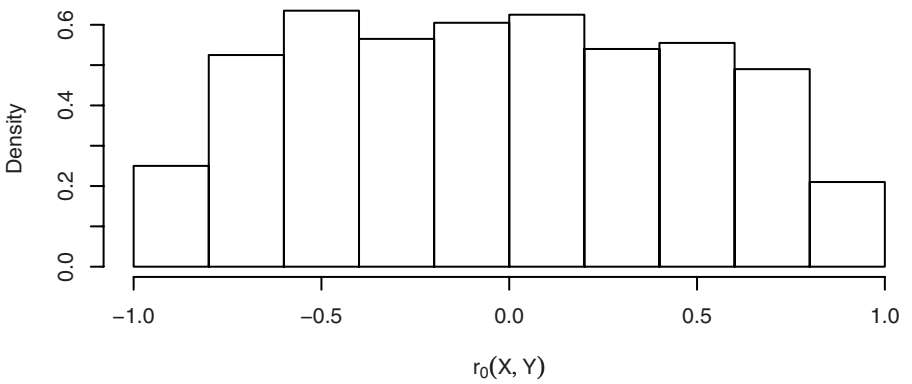
Exhibit 11.12 Asymptotic Error Rates of a Nominal 5% Test of Independence for a Pair of AR(1) Processes

$\phi = \phi_X = \phi_Y$	0.00	0.15	0.30	0.45	0.60	0.75	0.90
Error Rate	5%	6%	7%	11%	18%	30%	53%

```
> phi=seq(0,.95,.15)
> rejection=2*(1-pnorm(1.96*sqrt((1-phi^2)/(1+phi^2))))
> M=signif(rbind(phi,rejection),2)
> rownames(M)=c('phi', 'Error Rate')
> M
```

The problem of inflated variance of the sample cross-correlation coefficients becomes more acute for nonstationary data. In fact, the sample cross-correlation coefficients may no longer be approximately normally distributed even with a large sample size. Exhibit 11.13 displays the histogram of 1000 simulated lag zero cross-correlations between two independent IMA(1,1) processes each of size 500. An MA(1) coefficient of $\theta = 0.8$ was used for both simulated processes. Note that the distribution of $r_0(X,Y)$ is far from normal and widely dispersed between −1 and 1. See Phillips (1998) for a relevant theoretical discussion.

Exhibit 11.13 Histogram of 1000 Sample Lag Zero Cross-Correlations of Two Independent IMA(1,1) Processes Each of Size 500

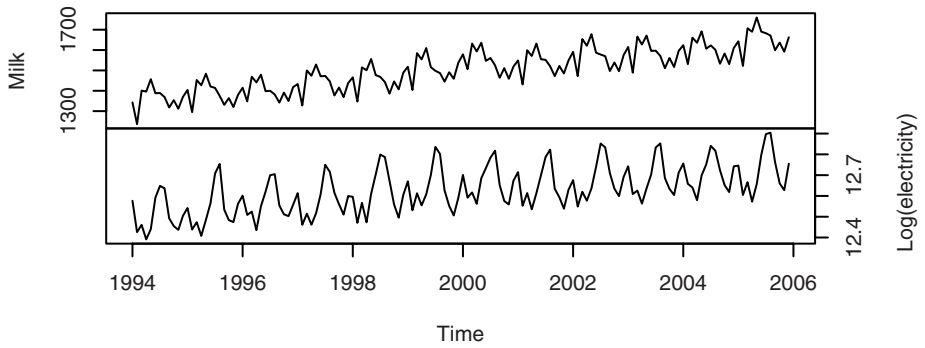


```
> set.seed(23457)
> correlation.v=NULL; B=1000; n=500
> for (i in 1:B) {x=cumsum(arima.sim(model=list(ma=.8),n=n))
> y=cumsum(arima.sim(model=list(ma=.8),n=n))
> correlation.v=c(correlation.v,ccf(x,y,lag.max=1,
   plot=F)$acf[2])}
> hist(correlation.v,prob=T,xlab=expression(r[0](X,Y)))
```

These results provide insight into why we sometimes obtain nonsense (spurious) correlation between time series variables. The phenomenon of spurious correlation was first studied systematically by Yule (1926).

As an example, the monthly milk production and the logarithms of monthly electricity production in the United States from January 1994 to December 2005 are shown in Exhibit 11.14. Both series have an upward trend and are highly seasonal.

Exhibit 11.14 Monthly Milk Production and Logarithms of Monthly Electricity Production in the U.S.

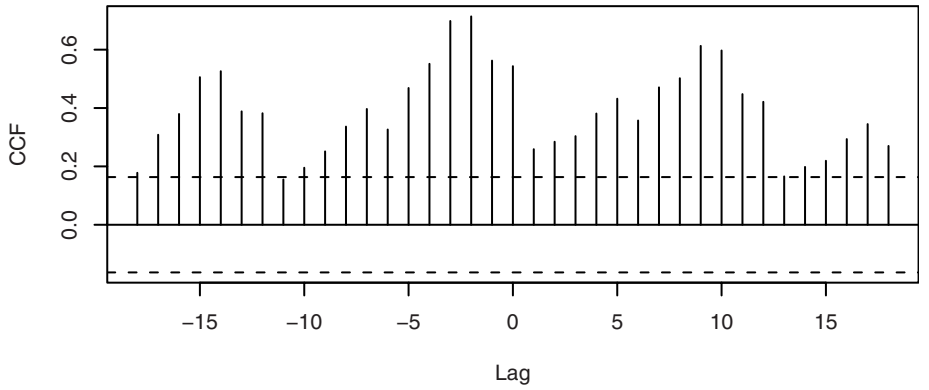



```
> data(milk); data(electricity)
> milk.electricity=ts.intersect(milk,log(electricity))
> plot(milk.electricity,yax.flip=T)
```

Calculation shows that these series have a cross-correlation coefficient at lag zero of 0.54, which is "statistically significantly different from zero" as judged against the standard error criterion of $1.96/\sqrt{n} = 0.16$. Exhibit 11.15 displays the strong cross-correlations between these two variables at a large number of lags.

Needless to say, it is difficult to come up with a plausible reason for the relationship between monthly electricity production and monthly milk production. The nonstationarity in the milk production series and in the electricity series is more likely the cause of the spurious correlations found between the two series. The following section contains further discussion of this example.

Exhibit 11.15 Sample Cross-Correlation Between Monthly Milk Production and Logarithm of Monthly Electricity Production in the U.S.




```
> ccf(as.vector(milk.electricity[,1]),
    as.vector(milk.electricity[,2]),ylab='CCF')
```

11.4 Prewhitening and Stochastic Regression

In the preceding section, we found that with strongly autocorrelated data it is difficult to assess the dependence between the two processes. Thus, it is pertinent to disentangle the linear association between X and Y, say, from their autocorrelation. A useful device for doing this is prewhitening. Recall that, for the case of stationary X and Y that are independent of each other, the variance of $r_k(X, Y)$ is approximately

$$\frac{1}{n}\left[1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)\right] \tag{11.4.1}$$

An examination of this formula reveals that the approximate variance is $1/n$ if either one (or both) of X or Y is a white noise process. In practice, the data may be nonstationary, but they may be transformed to approximately white noise by replacing the data by the residuals from a fitted ARIMA model. For example, if X follows an ARIMA(1,1,0) model with no intercept term, then

$$\tilde{X}_t = X_t - X_{t-1} - \phi(X_{t-1} - X_{t-2}) = 1 - (1 + \phi B) + \phi B^2]X_t \tag{11.4.2}$$

is white noise. More generally, if X_t follows some invertible ARIMA(p,d,q) model, then it admits an AR(∞) representation

$$\tilde{X}_t = (1 - \pi_1 B - \pi_2 B^2 - \cdots)X_t = \pi(B)X_t$$

where the $\tilde{X}$'s are white noise. The process of transforming the X's to the $\tilde{X}$'s via the *filter* $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$ is known as **whitening** or **prewhitening**. We now can

study the CCF between X and Y by prewhitening the Y and X using the *same* filter based on the X process and then computing the CCF of $\tilde{Y}$ and $\tilde{X}$; that is, the prewhitened Y and X. Since prewhitening is a *linear* operation, any linear relationships between the original series will be preserved after prewhitening. Note that we have abused the terminology, as $\tilde{Y}$ need not be white noise because the filter $\pi(B)$ is tailor-made only to transform X to a white noise process—not Y. We assume, furthermore, that $\tilde{Y}$ is stationary. This approach has two advantages: (i) the statistical significance of the sample CCF of the prewhitened data can be assessed using the cutoff $1.96/\sqrt{n}$, and (ii) the theoretical counterpart of the CCF so estimated is proportional to certain regression coefficients.

To see (ii), consider a more general regression model relating X to Y and, without loss of generality, assume both processes have zero mean:

$$Y_t = \sum_{j=-\infty}^{\infty} \beta_j X_{t-j} + Z_t \tag{11.4.3}$$

where X is independent of Z and the coefficients β are such that the process is well-defined. In this model, the coefficients β_k could be nonzero for any integer k. However, in real applications, the doubly infinite sum is often a finite sum so that the model simplifies to

$$Y_t = \sum_{j=m_1}^{m_2} \beta_j X_{t-j} + Z_t, \tag{11.4.4}$$

which will be assumed below even though we retain the doubly infinite summation notation for ease of exposition. If the summation ranges only over a finite set of *positive* indices, then X *leads* Y and the covariate X serves as a useful **leading indicator** for future Y's. Applying the filter $\pi(B)$ to both sides of this model, we get

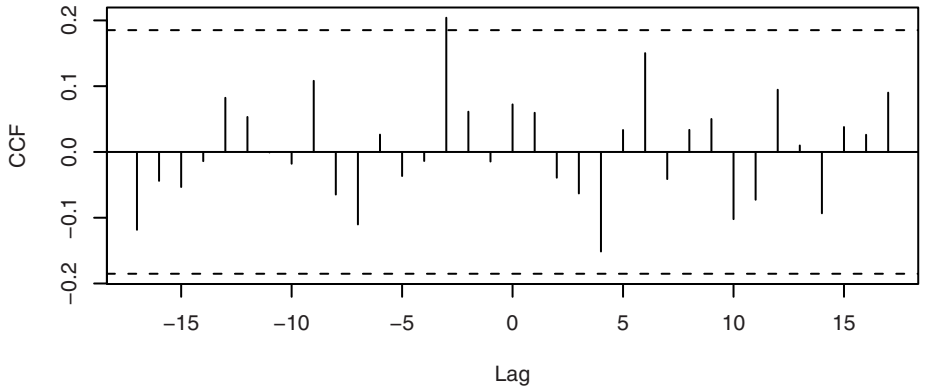
$$\tilde{Y}_t = \sum_{k=-\infty}^{\infty} \beta_k \tilde{X}_{t-k} + \tilde{Z}_t \tag{11.4.5}$$

where $\tilde{Z}_t = Z_t - \pi_1 Z_{t-1} - \pi_2 Z_{t-2} - \cdots$. The prewhitening procedure thus orthogonalizes the various lags of X in the original regression model. Because $\tilde{X}$ is a white noise sequence and $\tilde{X}$ is independent of $\tilde{Z}$, the theoretical cross-correlation coefficient between $\tilde{X}$ and $\tilde{Y}$ at lag k equals $\beta_{-k}(\sigma_{\tilde{X}}/\sigma_{\tilde{Y}})$. In other words, the theoretical cross-correlation of the prewhitened processes at lag k is proportional to the regression coefficient β_{-k}.

For a quick preliminary analysis, an approximate prewhitening can be done easily by first differencing the data (if needed) and then fitting an approximate AR model with the order determined by minimizing the AIC. For example, for the milk production and electricity consumption data, both are highly seasonal and contain trends. Consequently, they can be differenced with both regular differencing and seasonal differencing, and then the prewhitening can be carried out by filtering both differenced series by an AR model fitted to the differenced milk data. Exhibit 11.16 shows the sample CCF between the prewhitened series. None of the cross-correlations are now significant except for lag −3, which is just marginally significant. The lone significant cross-correlation is likely a false alarm since we expect about 1.75 false alarms out of the 35 sample cross-correla-

tions examined. Thus, it seems that milk production and electricity consumption are in fact largely uncorrelated, and the strong cross-correlation pattern found between the raw data series is indeed spurious.

Exhibit 11.16 Sample CCF of Prewhitened Milk and Electricity Production

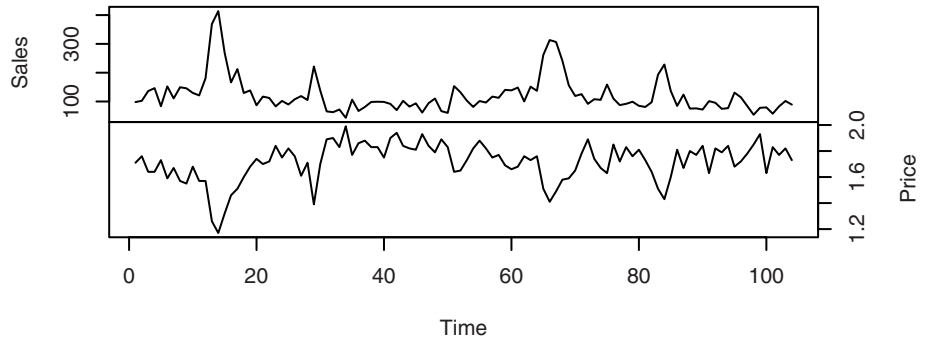



```
> me.dif=ts.intersect(diff(diff(milk,12)),
    diff(diff(log(electricity),12)))
> prewhiten(as.vector(me.dif[,1]),as.vector(me.dif[,2]),
    ylab='CCF')
```

The model defined by Equation (11.3.4) on page 262 is known variously as the **transfer-function model**, the distributed-lag model, or the dynamic regression model. The specification of which lags of the covariate enter into the model is often done by inspecting the sample cross-correlation function based on the prewhitened data. When the model appears to require a fair number of lags of the covariate, the regression coefficients may be parsimoniously specified via an ARMA specification similar to the case of intervention analysis; see Box et al. (1994, Chapter 11) for some details. We illustrate the method below with two examples where only one lag of the covariate appears to be needed. The specification of the stochastic noise process Z_t can be done by examining the residuals from an ordinary least squares (OLS) fit of Y on X using the techniques learned in earlier chapters.

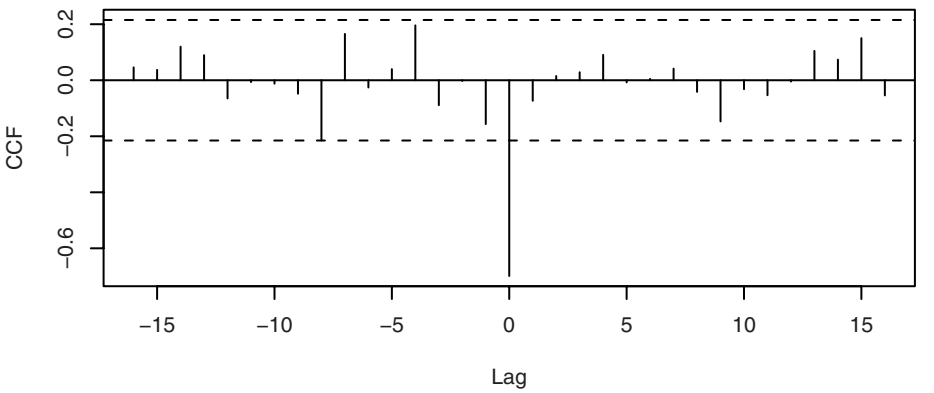
Our first example of this section is a sales and price dataset of a certain potato chip from Bluebird Foods Ltd., New Zealand. The data consist of the log-transformed weekly unit sales of large packages of standard potato chips sold and the weekly average price over a period of 104 weeks from September 20, 1998 through September 10, 2000; see Exhibit 11.17. The logarithmic transformation is needed because the sales data are highly skewed to the right. These data are clearly nonstationary. Exhibit 11.18 shows that, after differencing and using prewhitened data, the CCF is significant only at lag 0, suggesting a strong contemporaneous negative relationship between lag 1 of price and sales. Higher prices are associated with lower sales.

Exhibit 11.17 Weekly Log(Sales) and Price for Bluebird Potato Chips




```
> data(bluebird)
> plot(bluebird,yax.flip=T)
```

Exhibit 11.18 Sample Cross Correlation Between Prewhitened Differenced Log(Sales) and Price of Bluebird Potato Chips




```
> prewhiten(y=diff(bluebird)[,1],x=diff(bluebird)[,2],ylab='CCF')
```

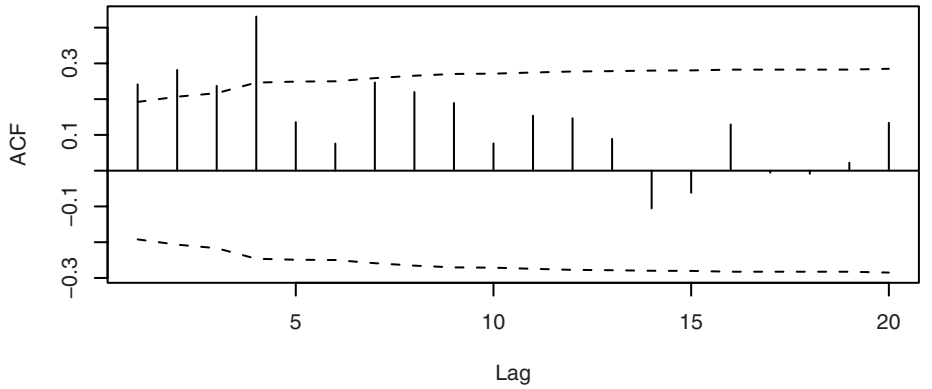
Exhibit 11.19 reports the estimates from the OLS regression of log(sales) on price. The residuals are, however, autocorrelated, as can be seen from their sample ACF and PACF displayed in Exhibits 11.20 and 11.21, respectively. Indeed, the sample autocorrelations of the residuals are significant for the first four lags, whereas the sample partial autocorrelations are significant at lags 1, 2, 4, and 14.

Exhibit 11.19 OLS Regression Estimates of Log(Sales) on Price

	Estimate	Std. Error	*t* value	*Pr*(>)
Intercept	15.90	0.2170	73.22	< 0.0001
Price	−2.489	0.1260	−19.75	< 0.0001

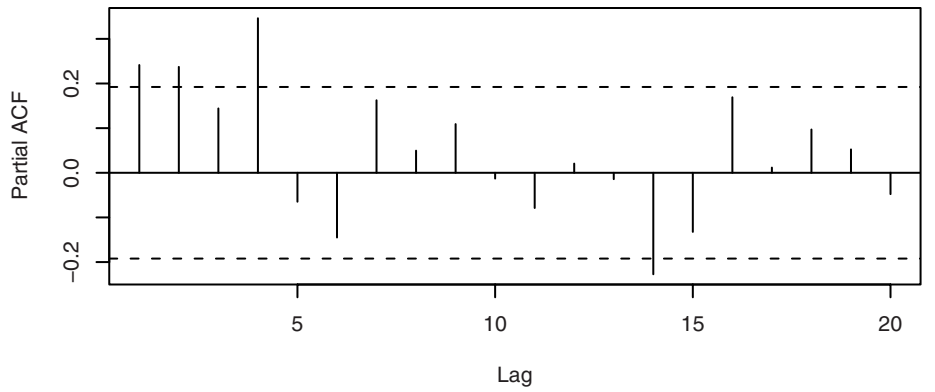
```
> sales=bluebird[,1]; price=bluebird[,2]
> chip.m1=lm(sales~price,data=bluebird)
> summary(chip.m1)
```

Exhibit 11.20 Sample ACF of Residuals from OLS Regression of Log(Sales) on Price




```
> acf(residuals(chip.m1),ci.type='ma')
```

Exhibit 11.21 Sample PACF of Residuals from OLS Regression of Log(Sales) on Price




```
> pacf(residuals(chip.m1))
```

The sample EACF of the residuals, shown in Exhibit 11.22, contains a triangle of zeros with a vertex at (1,4), thereby suggesting an ARMA(1,4) model. Hence, we fit a regression model of log(sales) on price with an ARMA(1,4) error.

Exhibit 11.22 The Sample EACF of the Residuals from the OLS Regression of Log(Sales) on Price

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	0	0	x	x	0	0	0	0	0	0
1	x	0	0	x	0	0	0	0	0	0	0	0	0	0
2	x	x	0	x	0	0	0	0	0	0	0	0	0	0
3	x	x	0	x	0	0	0	0	0	0	0	0	0	0
4	0	x	x	0	0	0	0	0	0	0	0	0	0	0
5	x	x	x	0	x	0	0	0	0	0	0	0	0	0
6	x	x	0	x	x	x	0	0	0	0	0	0	0	0
7	x	0	x	0	0	0	0	0	0	0	0	0	0	0

```
> eacf(residuals(chip.m1))
```

It turns out that the estimates of the AR(1) coefficient and the MA coefficients θ_1 and θ_3 are not significant, and hence a model fixing these coefficients to be zero was subsequently fitted and reported in Exhibit 11.23.

Exhibit 11.23 Maximum Likelihood Estimates of a Regression Model of Log(sales) on Price with a Subset MA(4) for the Errors

Parameter	θ_1	θ_2	θ_3	θ_4	Intercept	Price
Estimate	0	–0.2884	0	–0.5416	15.86	–2.468
Standard Error	0	0.0794	0	0 0.1167	0.1909	0.1100

σ^2 estimated as 0.02623: log likelihood = 41.02, AIC = –70.05

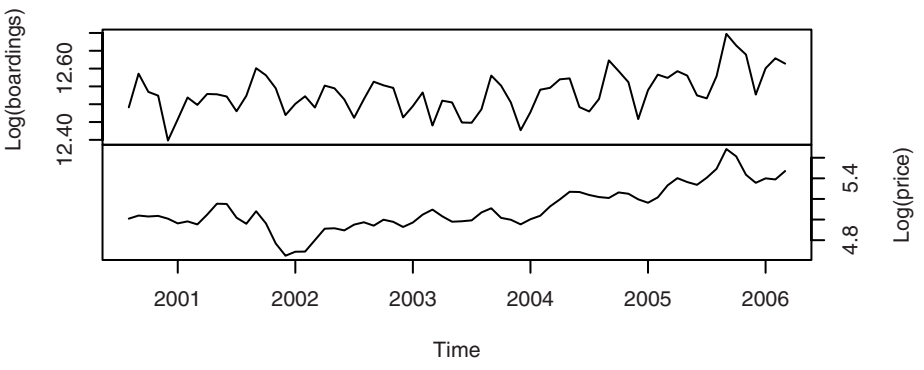
```
> chip.m2=arima(sales,order=c(1,0,4),xreg=data.frame(price))
> chip.m2
> chip.m3=arima(sales,order=c(1,0,4),xreg=data.frame(price),
    fixed=c(NA,0,NA,0,NA,NA,NA)); chip.m3
> chip.m4=arima(sales,order=c(0,0,4),xreg=data.frame(price),
    fixed=c(0,NA,0,NA,NA,NA)); chip.m4
```

Note that the regression coefficient estimate on Price is similar to that from the OLS regression fit earlier, but the standard error of the estimate is about 10% lower than that from the simple OLS regression. This illustrates the general result that the simple OLS estimator is consistent but the associated standard error is generally not trustworthy.

The residuals from this fitted model by and large pass various model diagnostic tests except that the residual ACF is significant at lag 14. As a result, some Box-Ljung test statistics have *p*-values bordering on 0.05 when 14 or more lags of the residual autocorrelations are included in the test. Even though the significant ACF at lag 14 may suggest a quarterly effect, we do not report a more complex model including lag 14 because (1) 14 weeks do not exactly make a quarter and (2) adding a seasonal MA(1) component of period 14 only results in marginal improvement in terms of model diagnostics.

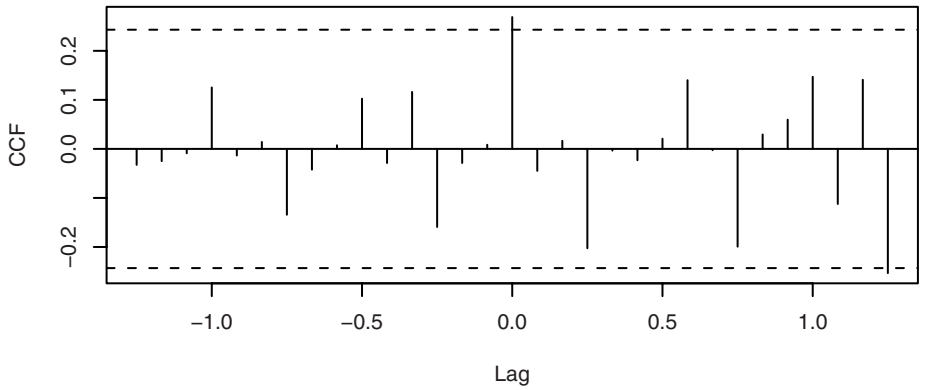
For a second example, we study the impact of higher gasoline price on public transportation usage. The dataset consists of the monthly number of boardings on public transportation in the Denver, Colorado, region together with the average monthly gasoline prices in Denver from August 2000 through March 2006. Both variables are skewed to the right and hence are log-transformed. As we shall see below, the logarithmic transformation also makes the final fitted model more interpretable. The time series plots, shown in Exhibit 11.24, display the increasing trends for both variables and the seasonal fluctuation in the number of boardings. Based on the sample ACF and PACF, an ARIMA(2,1,0) model was fitted to the gasoline price data. This fitted model was then used to filter the boardings data before computing their sample CCF which is shown in Exhibit 11.25. The sample CCF is significant at lags 0 and 15, suggesting positive contemporaneous correlation between gasoline price and public transportation usage. The significant CCF at lag 15, however, is unlikely to be real, as it is hard to imagine why the number of boardings might lead the gasoline price with a lag of 15 months. In this case, the quick preliminary approach of prewhitening the series by fitting a long AR model, however, showed that none of the CCFs are significant. It turns out that even after differencing the data, the AIC selects an AR(16) model. The higher order selected coupled with the relatively short time span may substantially weaken the power to detect correlations between the two variables. Incidentally, this example warns against simply relying on the AIC to select a high-order AR model to do prewhitening, especially with relatively short time series data.

Exhibit 11.24 Logarithms of Monthly Public Transit Boardings and Gasoline Prices in Denver, August 2000 through March 2006




```
> data(boardings)
> plot(boardings,yax.flip=T)
```

Exhibit 11.25 Sample CCF of Prewhitened Log(Boardings) and Log(Price)




```
> m1=arima(boardings[,2],order=c(2,1,0))
> prewhiten(x=boardings[,2],y=boardings[,1],x.model=m1)
```

Based on the sample ACF, PACF, and EACF of the residuals from a linear model of boardings on gasoline price, a seasonal ARIMA$(2,0,0)\times(1,0,0)_{12}$ model was tentatively specified for the error process in the regression model. However, the ϕ_2 coefficient estimate was not significant, and hence the AR order was reduced to $p = 1$. Using the outlier detection techniques discussed in Section 11.2, we found an additive outlier for March 2003 and an innovative outlier for March 2004. Because the test statistic for the additive outlier had a larger magnitude than that of the innovative outlier (–4.09 vs. 3.65), we incorporated the additive outlier in the model.[†] Diagnostics of the subsequent fitted model reveals that the residual ACF was significant at lag 3, which suggests the error process is a seasonal ARIMA$(1,0,3)\times(1,0,0)_{12}$ + outlier process. As the estimates of the coefficients θ_1 and θ_2 were found to be insignificant, they were suppressed from the final fitted model that is reported in Exhibit 11.26.

Diagnostics of the final fitted model suggest a good fit to the data. Also, no further outliers were detected. A 95% confidence interval for the regression coefficient on Log(Price) is (0.0249, 0.139). Note the interpretation of the fitted model: a 100% increase in the price of gasoline will lead to about an 8.2% increase in public transportation usage.

[†] Subsequent investigation revealed that a 30 inch snowstorm in March 2003 completely shut down Denver for one full day. It remained partially shut down for a few more days.

Exhibit 11.26 Maximum Likelihood Estimates of the Regression Model of Log(Boardings) on Log(Price) with ARMA Errors

Parameter	ϕ_1	θ_3	Φ_1	Intercept	Log(Price)	Outlier
Estimate	0.8782	0.3836	0.8987	12.12	0.0819	–0.0643
Standard Error	0.0645	0.1475	0.0395	0.1638	0.0291	0.0109

σ^2 estimated as 0.0004094: log-likelihood = 158.02, AIC = –304.05

```
> log.boardings=boardings[,1]
> log.price=boardings[,2]
> boardings.m1=arima(log.boardings,order=c(1,0,0),
    seasonal=list(order=c(1,0,0),period=12),
    xreg=data.frame(log.price))
> boardings.m1
> detectAO(boardings.m1); detectIO(boardings.m1)
> boardings.m2=arima(log.boardings,order=c(1,0,3),
    seasonal=list(order=c(1,0,0),period=12),
    xreg=data.frame(log.price,outlier=c(rep(0,31),1,rep(0,36))),
    fixed=c(NA,0,0,rep(NA,5)))
> boardings.m2
> detectAO(boardings.m2); detectIO(boardings.m2)
> tsdiag(boardings.m2,tol=.15,gof.lag=24)
```

It is also of interest to note that dropping the outlier term from the model results in a new regression estimate on Log(Price) of 0.0619 with a standard error of 0.0372. Thus, when the outlier is not properly modeled, the regression coefficient ceases to be significant at the 5% level. As demonstrated by this example, the presence of an outlier can adversely affect inference in time series modeling.

11.5 Summary

In this chapter, we used information from other events or other time series to help model the time series of main interest. We began with the so-called intervention models, which attempt to incorporate known external events that we believe have a significant effect on the time series of interest. Various simple but useful ways of modeling the effects of these interventions were discussed. Outliers are observations that deviate rather substantially from the general pattern of the data. Models were developed to detect and incorporate outliers in time series. The material in the section on spurious correlation illustrates how difficult it is to assess relationships between two time series, but methods involving prewhitening were shown to help in this regard. Several substantial examples were used to illustrate the methods and techniques discussed.

EXERCISES

11.1 Produce a time series plot of the air passenger miles over the period January 1996 through May 2005 using seasonal plotting symbols. Display the graph full-screen and discuss the seasonality that is displayed. The data are in the file named airmiles.

11.2 Show that the expression given for m_t in Equation (11.1.7) on page 251 satisfies the "AR(1)" recursion given in Equation (11.1.6) with the initial condition $m_0 = 0$.

11.3 Find the "half-life" for the intervention effect specified in Equation (11.1.6) on page 251 when $\delta = 0.7$.

11.4 Show that the "half-life" for the intervention effect specified in Equation (11.1.6) on page 251 increases without bound as δ increases to 1.

11.5 Show that for the intervention effect specified by Equation (11.1.6) on page 251

$$\lim_{\delta \to 1} m_t = \begin{cases} \omega(T - t), \text{ for } t \geq T \\ 0, \text{ otherwise} \end{cases}$$

11.6 Consider the intervention effect displayed in Exhibit 11.3, (b), page 253.
(a) Show that the jump at time $T + 1$ is of height ω as displayed.
(b) Show that, as displayed, the intervention effect tends to $\omega/(1 - \delta)$ as t increases without bound.

11.7 Consider the intervention effect displayed in Exhibit 11.3, (c), page 253. Show that the effect increases linearly starting at time $T + 1$ with slope ω as displayed.

11.8 Consider the intervention effect displayed in Exhibit 11.4, (a), page 254.
(a) Show that the jump at time $T + 1$ is of height ω as displayed.
(b) Show that, as displayed, the intervention effect tends to go back to 0 as t increases without bound.

11.9 Consider the intervention effect displayed in Exhibit 11.4, (b), page 254.
(a) Show that the jump at time $T + 1$ is of height $\omega_1 + \omega_2$ as displayed.
(b) Show that, as displayed, the intervention effect tends to ω_2 as t increases without bound.

11.10 Consider the intervention effect displayed in Exhibit 11.4, (c), page 254.
(a) Show that the jump at time T is of height ω_0 as displayed.
(a) Show that the jump at time $T + 1$ is of height $\omega_1 + \omega_2$ as displayed.
(b) Show that, as displayed, the intervention effect tends to ω_2 as t increases without bound.

11.11 Simulate 100 pairs of (X_t, Y_t) from the model of Equation (11.3.1) on page 261 with $d = 3$, $\beta_0 = 0$, and $\beta_1 = 1$. Use $\sigma_X = 2$ and $\sigma_e = 1$. Display and interpret the sample CCF between these two series.

11.12 Show that when the X and Y are independent AR(1) time series with parameters ϕ_X and ϕ_Y, respectively, Equation (11.3.5) on page 262 reduces to give Equation (11.3.6).

11.13 Show that for the process defined by Equation (11.4.5) on page 266, the cross-correlation between $\tilde{X}$ and $\tilde{Y}$ at lag k is given by $\beta_{-k}(\sigma_{\tilde{X}}/\sigma_{\tilde{Y}})$.

11.14 Simulate an AR time series with $\phi = 0.7$, $\mu = 0$, $\sigma_e = 1$, and of length $n = 48$. Plot the time series, and inspect the sample ACF and PACF of the series.

(a) Now add a step function response of $\omega = 1$ unit height at time $t = 36$ to the simulated series. The series now has a theoretical mean of zero from $t = 1$ to 35 and a mean of 1 from $t = 36$ on. Plot the new time series and calculate the sample ACF and PACF for the new series. Compare these with the results for the original series.

(b) Repeat part (a) but with an impulse response at time $t = 36$ of unit height, $\omega = 1$. Plot the new time series, and calculate the sample ACF and PACF for the new series. Compare these with the results for the original series. See if you can detect the additive outlier at time $t = 36$ assuming that you do not know where the outlier might occur.

11.15 Consider the air passenger miles time series discussed in this chapter. The file is named airmiles. Use only the *preintervention* data (that is, data prior to September 2001) for this exercise.

(a) Verify that the sample ACF for the twice differenced series of the logarithms of the preintervention data is as shown in Exhibit 11.5 on page 255.

(b) The plot created in part (a) suggests an ARIMA$(0,1,1)\times(0,1,0)_{12}$. Fit this model and assess its adequacy. In particular, verify that additive outliers are detected in December 1996, January 1997, and December 2002.

(c) Now fit an ARIMA$(0,1,1)\times(0,1,0)_{12}$ + three outliers model and assess its adequacy.

(d) Finally, fit an ARIMA$(0,1,1)\times(0,1,1)_{12}$ + three outliers model and assess its adequacy.

11.16 Use the logarithms of the Denver region public transportation boardings and Denver gasoline price series. The data are in the file named boardings.

(a) Display the time series plot of the monthly boardings using seasonal plotting symbols. Interpret the plot.

(b) Display the time series plot of the monthly average gasoline prices using seasonal plotting symbols. Interpret the plot.

11.17 The data file named deere1 contains 82 consecutive values for the amount of deviation (in 0.000025 inch units) from a specified target value that an industrial machining process at Deere & Co. produced under certain specified operating conditions. These data were first used in Exercise 6.33, page 146, where we observed an obvious outlier at time $t = 27$.

(a) Fit an AR(2) model using the original data including the outlier.

(b) Test the fitted AR(2) model of part (a) for both AO and IO outliers.

(c) Now fit the AR(2) model incorporating a term in the model for the outlier.

(d) Assess the fit of the model in part (c) using all of our diagnostic tools. In particular, compare the properties of this model with the one obtained in part (a).

11.18 The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.) These data were first investigated in Exercise 6.39, page 147, but several outliers were observed. When the observed outliers were replaced by more typical values, an MA(2) model was suggested.

(a) Fit an MA(2) model to the original data, and test the fitted model for both AO and IO outliers.

(b) Now fit the MA(2) model incorporating the outliers into the model.

(c) Assess the fit of the model obtained in part (b). In particular, are any more outliers indicated?

(d) Fit another MA(2) model incorporating any additional outliers found in part (c), and assess the fit of this model.

11.19 The data file named bluebirdlite contains weekly sales and price data for Bluebird Lite potato chips. Carry out an analysis similar to that for Bluebird Standard potato chips that was begun on page 267.

11.20 The file named units contains annual unit sales of a certain product from a widely known international company over the years 1983 through 2005. (The name of the company must remain anonymous for proprietary reasons.)

(a) Plot the time series of units and describe the general features of the plot.

(b) Use ordinary least squares regression to fit a straight line in time to the series.

(c) Display the sample PACF of the residuals from this model, and specify an ARIMA model for the residuals.

(d) Now fit the model unit sales = AR(2) + time. Interpret the output. In particular, compare the estimated regression coefficient on the time variable obtained here with the one you obtained in part (b).

(e) Perform a thorough analysis of the residuals from this last model.

(f) Repeat parts (d) and (e) using the logarithms of unit sales as the response variable. Compare these results witjh those obtained in parts (d) and (e).

11.21 In Chapters 5–8, we investigated an IMA(1,1) model for the logarithms of monthly oil prices. Exhibit 8.3 on page 178 suggested that there may be several outliers in this series. Investigate the IMA(1,1) model for this series for outliers using the techniques developed in this chapter. Be sure to compare your results with those obtained earlier that ignored the outliers. The data are in the file named oil.