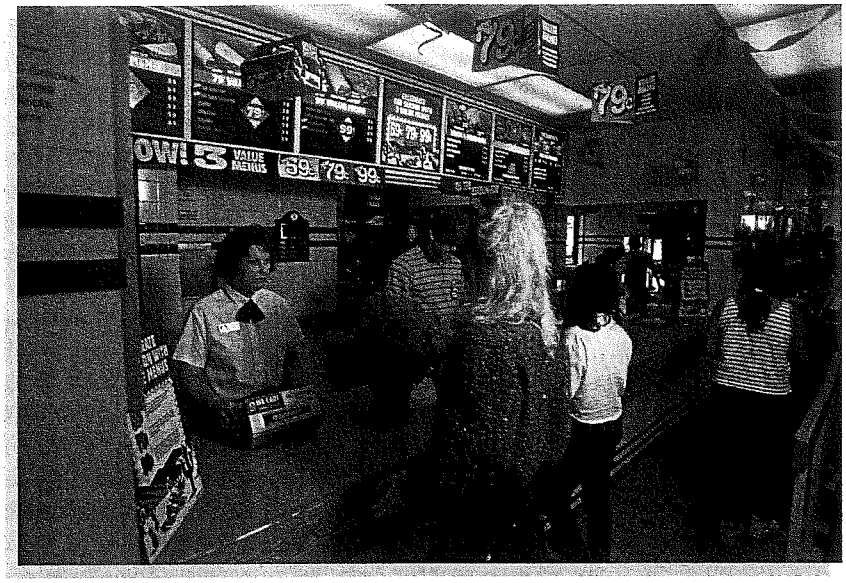
CHAPTER

13 Time Series Analysis and Forecasting



©Mark Richards/Photo Edit, Inc.

FORECASTING LABOR REQUIREMENTS AT TACO BELL

How much quantitative analysis occurs at fast-food restaurants? At Taco Bell, a lot! An article by Huerter and Swart (1998) explains the approach to labor management that has occurred at Taco Bell restaurants over the past decade. Labor is a large component of costs at Taco Bell. Approximately 30% of every sales dollar goes to labor. However, the unique characteristics of fast-food restaurants make it difficult to plan labor utilization efficiently. In particular, the Taco Bell product—food—cannot be inventoried; it must be made fresh at the time the customer orders it. Because of shifting demand throughout any given day, where the lunch period accounts for approximately 52% of a day's sales and as much as 25% of a day's sales can occur during the busiest hour, labor requirements vary greatly throughout the day. If too many workers are on hand during slack times, they are paid for doing practically nothing. Worse than that, however, are the lost sales (and unhappy customers) that occur if too few workers are on hand during peak times. Before 1988, Taco Bell made very little effort to manage the labor problem in an efficient, centralized manner. The company simply allocated about 30% of each store's sales to the store managers and let them allocate it as best they could—not always with good results.

In 1988 Taco Bell initiated its "value meal" deals, where certain meals were priced as low as 59 cents. This increased demand to the point where management could no longer ignore the labor allocation problem. Therefore,

in-store computers were installed, data from all stores were collected, and a team of analysts was assigned the task of developing a cost-efficient labor allocation system. This system, which has now been fully integrated into all Taco Bell stores since 1993, is composed of three subsystems: (1) a forecasting subsystem that, for each store, forecasts the arrival rate of customers by 15-minute interval by day of week; (2) a simulation subsystem that, for each store, simulates the congestion and number of lost customers that will occur for any customer arrival rate, given a specific number (and deployment) of workers; and (3) an optimization subsystem that, for each store, indicates the minimum cost allocation of workers, subject to various constraints, such as a minimum service level and a minimum shift length for workers. Although all three of these subsystems are important, the forecasting subsystem is where it all starts. Each store must have a reasonably accurate forecast of future customer arrival rates, broken down by small time intervals (such as 11:15 A.M. to 11:30 A.M. on Friday), before labor requirements can be predicted and labor allocations can be made in an intelligent manner. Like many real-world forecasting systems, Taco Bell's has two important characteristics: (1) it requires extensive data, which have been made available by the in-store computer systems, and (2) the eventual forecasting method used is mathematically a fairly simple one, namely, 6-week moving averages, which we study in this chapter.

Simple or not, the forecasts, as well as the other system components, have enabled Taco Bell to cut costs and increase profits considerably. In its first 4 years, 1993 to 1996, the labor management system is estimated to have saved Taco Bell approximately $40.34 million in labor costs. Because the number of Taco Bell stores is constantly increasing, the annual company-wide savings from the system will certainly grow in the future. In addition, the focus on quantitative analysis has produced other side benefits for Taco Bell. Its service is now better and more consistent across stores, with many fewer customers leaving because of slow service. Also, the quantitative models developed have enabled Taco Bell to evaluate the effectiveness of various potential productivity enhancements, including self-service drink islands, customer-activated touch screens for ordering, and smaller kitchen areas. So the next time you order food from Taco Bell, you can be assured that there is definitely a method to the madness! ■

13.1 INTRODUCTION

Many decision-making applications depend on a forecast of some quantity. Here are several examples.

Examples of Forecasting Applications

- When a service organization, such as a fast-food restaurant, plans its staffing over some time period, it must forecast the customer demand as a function of time. This might be done at a very detailed level, such as the demand in successive 15-minute periods, or at a more aggregate level, such as the demand in successive weeks.
- When a company plans its ordering or production schedule for a product it sells to the public, it must forecast the customer demand for this product so that it can stock appropriate quantities—neither too much nor too little.
- When an organization plans to invest in stocks, bonds, or other financial instruments, it typically attempts to forecast movements in stock prices and interest rates.
- When government representatives plan policy, they attempt to forecast movements in macroeconomic variables such as inflation, interest rates, and unemployment.

Unfortunately, forecasting is a very difficult task, both in the short run and in the long run. Typically, we base forecasts on historical data. We search for patterns or relationships in

the historical data, and then we make forecasts. There are two problems with this approach. The first is that it is not always easy to uncover historical patterns or relationships. In particular, it is often difficult to separate the noise, or random behavior, from the underlying patterns. Some forecasts can even overdo it, by attributing importance to patterns that are in fact random variations and are unlikely to repeat themselves.

The second problem is that there are no guarantees that past patterns will continue in the future. A new war could break out somewhere in the world, a company's competitor could introduce a new product into the market, the bottom could fall out of the stock market, and so on. Each of these shocks to the system being studied could drastically alter the future in a highly unpredictable way. This partly explains why forecasts are almost always wrong. Unless they have inside information to the contrary, forecasters must assume that history will repeat itself. But we all know that history does *not* always repeat itself. Therefore, there are many famous forecasts that turned out to be way off the mark, even though the forecasters made reasonable assumptions and used standard forecasting techniques. Nevertheless, forecasts are required constantly, so fear of failure is no excuse for not giving it our best effort.

13.2 FORECASTING METHODS: AN OVERVIEW

There are many forecasting methods available, and all practitioners have their favorites. To say the least, there is little agreement among practitioners or theoreticians as to the best forecasting method. The methods can generally be divided into three groups: (1) *judgmental* methods, (2) *extrapolation* (or *time series*) methods, and (3) *econometric* (or *causal*) methods. The first of these is basically nonquantitative and will not be discussed here; the last two are quantitative. In this section we describe extrapolation and econometric methods in some generality. In the rest of the chapter, we go into more detail, particularly about the extrapolation methods.

13.2.1 Extrapolation Methods

Extrapolation methods are quantitative methods that use past data of a time series variable—and nothing else, except possibly time itself—to forecast future values of the variable. The idea is that we can use past movements of a variable, such as company sales or U.S. exports to Japan, to forecast its future values. Many extrapolation methods are available, including trend-based regression, exponential smoothing, moving averages, and autoregression. Some of these methods are relatively simple, both conceptually and in terms of the calculations required, whereas others are quite complex. Also, as the names imply, some of these methods use the same regression methods we discussed in the previous two chapters, whereas others do not.

All of these extrapolation methods search for *patterns* in the historical series and then extrapolate these patterns into the future. Some try to track long-term upward or downward trends and then project these. Some try to track the seasonal patterns (sales up in November and December, down in other months, for example) and then project these. Basically, the more complex the method, the more closely it tries to track historical patterns. Researchers have long believed that good forecasting methods should be able to track the ups and downs—the zigzags on a graph—of a time series. This has led to voluminous research and increasingly complex methods. But is complexity always better?

Surprisingly, empirical evidence shows that it is sometimes worse. This is documented in the quarter-century review article by Armstrong (1986) and the article by Schnarrs and Bavuso (1986). They document a number of empirical studies on literally thousands of time series forecasts where complex methods fared no better, and often worse, than simple

methods. In fact, the Schnarrs and Bavuso article presents evidence that a naive forecast from a "random walk" model often outperforms all of the more sophisticated extrapolation methods. With this naive model we forecast that next period's value will be the same as this period's value. So if today's closing stock price is 51.375, we forecast that tomorrow's closing stock price will be 51.375. This method is certainly simple, and it sometimes works quite well. We discuss random walks in more detail in Section 13.5.

The evidence in favor of simpler models is not accepted by everyone, particularly not those who have spent years investigating complex models, and complex models continue to be studied and used. However, there is a very plausible reason why simple models might provide better forecasts. The whole idea behind extrapolation methods is to extrapolate historical patterns into the future. But it is often difficult to determine which patterns are real and which represent noise—random ups and downs that are not likely to repeat themselves. Also, if something important changes (a competitor introduces a new product or interest rates increase, for example), it is certainly possible that the historical patterns will change. A potential problem with complex methods is that they often track a historical series *too* closely. That is, they often track patterns that are really noise. Simpler methods, on the other hand, track only the most basic underlying patterns and therefore can be more flexible and accurate in forecasting the future.

13.2.2 Econometric Models

Econometric models, also called **causal** models, use regression to forecast a time series variable by using other explanatory time series variables. For example, a company might use a causal model to regress future sales on its advertising level, the population income level, the interest rate, and possibly others. In one sense, regression analysis involving time series variables is similar to the regression analysis discussed in the previous two chapters. We can still use the same least squares approach and the same multiple regression software in many time series regression models. In fact, several examples and problems in the previous two chapters used time series data.

However, causal regression models for time series data present new mathematical challenges that go well beyond the level of this book. To get a glimpse of the potential difficulties, suppose a company wants to use a causal regression model to forecast its monthly sales for some product, based on two other time series variables: its monthly advertising levels for the product and its main competitor's monthly advertising levels for a competing product. We could simply estimate a regression equation of the form

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} \tag{13.1}$$

Here, Y_t is the company's sales in month t, and X_{1t} and X_{2t} are the company's and the competitor's advertising levels in month t. We might learn something useful from this regression model, but we should be aware of the following problems.

One problem is that we must decide on the appropriate "lags" for the regression equation. Do sales this month depend only on advertising levels *this* month, as specified in equation (13.1), or also on advertising levels in the previous month, the previous two months, and so on? A second problem is whether to include lags of the *sales* variable in the regression equation as explanatory variables. Presumably, sales in one month might depend on the level of sales in previous months (as well as on advertising levels). A third problem is that the two advertising variables can be *autocorrelated* and *cross-correlated.* Autocorrelation means, for example, that the company's advertising level in one month can depend on its advertising levels in previous months. Cross-correlation means, for example, that the company's advertising level in one month can be related to the competitor's advertising levels in previous months, or that the competitor's advertising in one month can be related to the company's advertising levels in previous months.

These are difficult issues, and the way in which they are addressed can make a big difference in the usefulness of the resulting regression model. We examine several regression-based models in this chapter, but we avoid situations such as the one just described, where one time series variable Y is regressed on one or more time series X's. [Pankratz (1991) is a good reference for these latter types of models.]

13.2.3 Combining Forecasts

There is one other general forecasting method that is worth mentioning. In fact, it has attracted a lot of attention in recent years, and many researchers believe that it has great potential for increasing forecast accuracy. The method is simple—we combine two or more forecasts to obtain the final forecast. The reasoning behind this method is also simple—the forecast errors from different forecasting methods might cancel one another. The forecasts that are combined can be of the same general type—extrapolation forecasts, for example—or they can be of different types, such as judgmental and extrapolation. The *number* of forecasts to combine and the *weights* to use in combining them have been the subject of several research studies.

Although the findings are not entirely consistent, it appears that the marginal benefit from each individual forecast after the first two or three is minor. Also, there is not much evidence to suggest that the simplest weighting scheme—weight each forecast equally, that is, average them—is any less accurate than more complex weighting schemes.

13.2.4 Components of Time Series Data

In Chapter 2 we discussed time series graphs, a useful graphical means of depicting time series data. We now use these time series graphs to help explain and identify four important components of a time series. These components are called the *trend* component, the *seasonal* component, the *cyclic* component, and the *random* (or *noise*) component.

Let's start by looking at a very simple time series. This is a time series where every observation is the same. Such a series is shown in Figure 13.1. The graph in this figure shows time (t) on the horizontal axis and the observation values (Y) on the vertical axis. We assume that Y is measured at regularly spaced intervals, usually days, weeks, months, quarters, or years, with Y_t being the value of the observation at time period t. As indicated in Figure 13.1, the individual observation points are usually joined by straight lines to make any patterns in the time series more apparent. Because all observations in this time series are equal, the resulting time series graph is a horizontal line. We refer to this time series as the *base* series. We will now illustrate more interesting time series built from this base series.

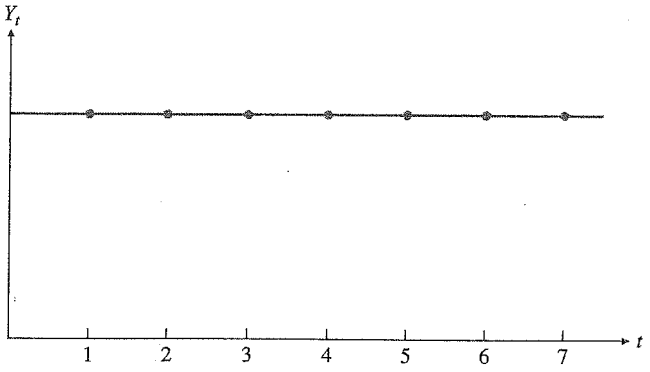
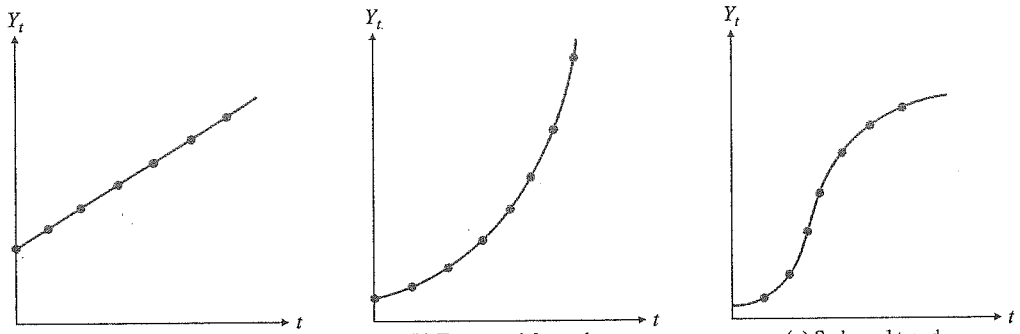


Figure 13.1
The Base Series

If the observations increase or decrease regularly through time, we say that the time series has a **trend**. The graphs in Figure 13.2 illustrate several possible trends. The *linear* trend in Figure 13.2a occurs if a company's sales, for example, increase by the same amount from period to period. This constant per period change is then the slope of the linear trend line. The curve in Figure 13.2b is an *exponential* trend curve. It occurs in a business such as the personal computer business, where sales have increased at a tremendous rate (at least in the 1990s, the boom years). For this type of curve, the *percentage* increase in Y_t from period to period remains constant. The curve in Figure 13.2c is an *S-shaped* trend curve. For example, this type of trend curve is appropriate for a new product that takes a while to catch on, then exhibits a rapid increase in sales as the public becomes aware of it, and finally tapers off to a fairly constant level. The curves in Figure 13.2 all represent *upward* trends. Of course, we could just as well have *downward* trends of the same types.

Figure 13.2 Series with Trends

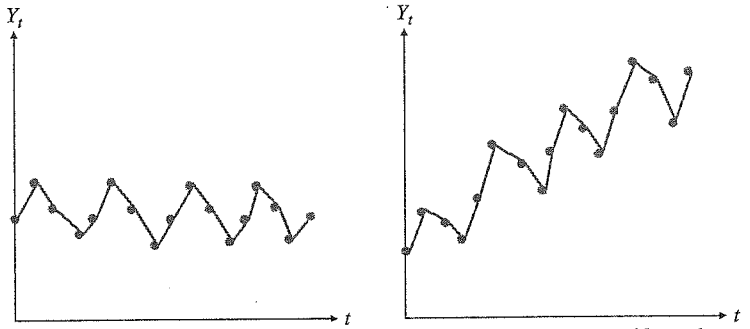


(a) Linear trend (b) Exponential trend (c) S-shaped trend

Many time series have a **seasonal** component. For example, a company's sales of swimming pool equipment increase every spring, then stay relatively high during the summer, and then drop off until next spring, at which time the yearly pattern repeats itself. An important aspect of the seasonal component is that it tends to be predictable from one year to the next. That is, the *same* seasonal pattern tends to repeat itself every year.

In Figure 13.3 we show two possible seasonal patterns. In Figure 13.3a there is nothing but the seasonal component. That is, if there were no seasonal variation, we would have the base series in Figure 13.1. In Figure 13.3b we show a seasonal pattern superimposed on a linear trend line.

Figure 13.3
Series with Seasonality



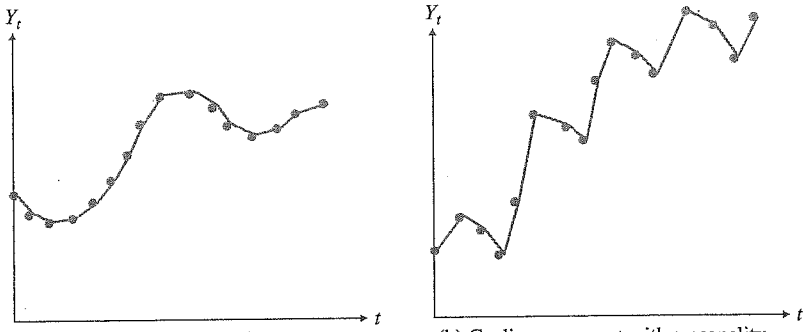
(a) Seasonal component only (b) Seasonal component with trend

The third component of a time series is the **cyclic** component. By studying past movements of many business and economic variables, it becomes apparent that there are business cycles that affect many variables in similar ways. For example, during a recession housing starts generally go down, unemployment goes up, stock prices go down, and so

on. But when the recession is over, all of these variables tend to move in the opposite direction. Unfortunately, the cyclic component is harder to predict than the seasonal component. The reason is that seasonal variation is much more regular. For example, swimming pool supplies sales *always* start to increase during the spring. Cyclic variation, on the other hand, is more irregular for the simple reason that the "business cycle" does not always have the same length. A further distinction is that the length of a seasonal cycle is generally one year; the length of a business cycle is generally much longer than one year.

The graphs in Figure 13.4 illustrate the cyclic component of a time series. In Figure 13.4a cyclic variation is superimposed on the base series in Figure 13.1. In Figure 13.4b this same cyclic variation is superimposed on the series in Figure 13.3b. The resulting graph has trend, seasonal variation, and cyclic variation.

Figure 13.4
Series with Cyclic Component



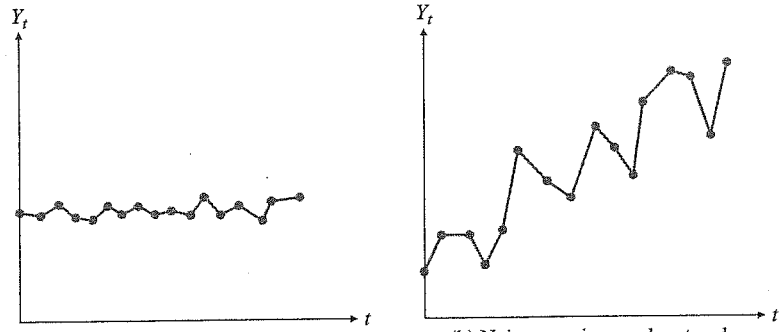
(a) Cyclic component only

(b) Cyclic component with seasonality and trend

The final component in a time series is called the **random** component, or simply **noise**. This unpredictable component gives most time series graphs their irregular, jagged-edge appearances. Usually, a time series can be determined only to a certain extent by its trend, seasonal, and cyclic components. Then other factors determine the rest. These other factors may be inherent randomness, unpredictable "shocks" to the system, the unpredictable behavior of human beings who interact with the system, and possibly others. These factors combine to create a certain amount of unpredictability in almost all time series.

Figures 13.5 and 13.6 show the effect that noise can have on a time series graph. The graph on the left of each figure shows the random component only, superimposed on the base series. Then on the right of each figure, the random component is superimposed on the trend-with-seasonal-component graph from Figure 13.3b. The difference between Figures 13.5 and 13.6 is the relative magnitude of the noise. When it is small, as in Figure 13.5, the other components emerge fairly clearly; they are not disguised by the noise. But if the noise is large in magnitude, as in Figure 13.6, the noise makes it very difficult to distinguish the other components.

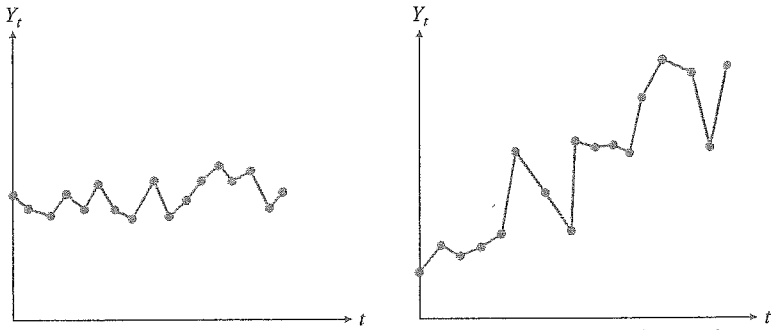
Figure 13.5
Series with Noise



(a) Noise only

(b) Noise superimposed on trend and seasonal components

Figure 13.6
Series with More Noise



(a) Noise only

(b) Noise superimposed on trend and seasonal components

13.2.5 General Notation and Formulas

We now introduce a bit of notation and discuss some aspects common to most forecasting methods. In general, we let Y denote the variable we want to forecast. Then Y_t denotes the observed value of Y at time t. Typically, the first observation (the most distant one) corresponds to period $t = 1$, and the last observation (the most recent one) corresponds to period $t = T$, so that T denotes the number of historical observations of Y. The periods themselves might be weeks, months, quarters, years, or any other convenient unit of time.

Suppose we have just observed Y_{t-k} and want to make a "k-period-ahead" forecast; that is, we want to use the information through time $t - k$ to forecast Y_t. Then we denote the resulting forecast by $F_{t-k,t}$. The first subscript denotes the period in which the forecast is made, and the second subscript denotes the period being forecasted. As an example, if the data are monthly and September 2004 corresponds to $t = 67$, then a forecast of Y_{69}, the value in November 2004, would be labeled $F_{67,69}$. The **forecast error** is the difference between the actual value and the forecast. It is denoted by E with appropriate subscripts. Specifically, the forecast error associated with $F_{t-k,t}$ is

$$E_{t-k,t} = Y_t - F_{t-k,t}$$

This double-subscript notation is necessary to specify when the forecast is being made and which period is being forecasted. However, the former is often clear from context. Therefore, to simplify the notation, we usually drop the first subscript and write F_t and E_t to denote the forecast of Y_t and the error in this forecast.

We first develop a model to fit the historical data. Then we use this model to forecast the future.

There are actually two steps in any forecasting procedure. The first step is to build a model that fits the historical data well. The second step is to use this model to forecast the future. Most of the work goes into the first step. For any trial model we see how well it "tracks" the known values of the time series. Specifically, we calculate the one-period-ahead forecasts, F_t (or more precisely, $F_{t-1,t}$), from the model and compare these to the known values, Y_t, for each t in the historical time period. We attempt to find a model that produces small forecast errors, E_t. We expect that if the model forecasts the *historical* data well, it will also forecast *future* data well.

Forecasting software packages typically report several summary measures of the forecast errors. The most important of these are MAE (mean absolute error), RMSE (root mean square error), and MAPE (mean absolute percentage error). These are defined in equations (13.2), (13.3), and (13.4). Fortunately, models that make any one of these measures small tend to make the others small, so that we can choose whichever measure we want to minimize. In the following formulas, N denotes the number of terms in each sum. This value is typically slightly less than T, the number of historical observations, because it is usually not possible to provide a forecast for each historical period.

Mean Absolute Error

$$\text{MAE} = \left(\sum_{t=1}^{N} |E_t|\right)/N \tag{13.2}$$

Root Mean Square Error

$$\text{RMSE} = \sqrt{\left(\sum_{t=1}^{N} E_t^2\right)/N} \tag{13.3}$$

Mean Absolute Percentage Error

$$\text{MAPE} = 100\% \times \left(\sum_{t=1}^{N} |E_t/Y_t|\right)/N \tag{13.4}$$

A model that makes any one of these error measures small tends to make the other two small as well.

RMSE is similar to a standard deviation in that the errors are squared; because of the square root, its units are the same as those of the forecasted variable. The MAE is similar to the RMSE, except that absolute values of errors are used instead of squared errors. The MAPE is probably the most easily understood measure because it does not depend on the units of the forecasted variable; it is always stated as a percentage. For example, the statement that the forecasts are off on average by 2% has a clear meaning, even if you do not know the units of the variable being forecasted.

Some forecasting software packages choose the best model from a given class (such as the best exponential smoothing model) by minimizing MAE, RMSE, or MAPE. However, small values of these measures guarantee only that the model forecasts the *historical* observations well. There is still no guarantee that the model will forecast *future* values accurately.

We now examine a number of useful forecasting models. You should be aware that more than one of these models can be appropriate for any particular time series data. For example, a random walk model and an autoregression model could be equally effective for forecasting stock price data. (Remember also that we can combine forecasts from more than one model to obtain a possibly better forecast.) We try to give some insights into choosing the best type of model for various types of time series data, but ultimately the choice depends on the experience of the forecaster.

13.3 TESTING FOR RANDOMNESS

All forecasting models we build have the general form shown in equation (13.5). The fitted value in this equation is the part we calculate from past data and any other available information (such as the season of the year), and it is used as a forecast for Y. The residual is the forecast error, the difference between the observed value of Y and its forecast:

$$Y_t = \text{Fitted Value} + \text{Residual} \tag{13.5}$$

In a time series context the terms residual *and* forecast error *are used interchangeably.*

For time series data, there is a residual for each historical period, that is, for each value of t. We want this time series of residuals to be random "noise," as discussed in Section 13.2.4. The reason is that if this series of residuals is not noise, then it can be modeled further. For example, if the residuals trend upwardly, then we can refine our model to include this trend component in the *fitted* value. The point is that we want the fitted value to include all

components of the original series that can possibly be forecasted, and we want the leftover residuals to be noise.

We now discuss ways to determine whether a time series of residuals is random noise (which we usually abbreviate to "random"). The simplest method, but not always a reliable one, is to examine time series graphs of residuals visually. This often enables us to detect nonrandom patterns. For example, the time series graphs in Figures 13.7 through 13.11 illustrate some common nonrandom patterns. In Figure 13.7, there is an upward trend. In Figure 13.8, the variance increases through time (larger zigzags to the right). Figure 13.9 exhibits seasonality, where observations in certain months are consistently larger than those in other months. There is a "meandering" pattern in Figure 13.10, where large observations tend to be followed by other large observations, and small observations tend to be followed by other small observations. Finally, the opposite behavior of Figure 13.10 is illustrated in Figure 13.11. Here, there are *too many* zigzags—large observations tend to follow small observations and vice versa. None of the time series in these figures can be considered random.

Figure 13.7
A Series with Trend

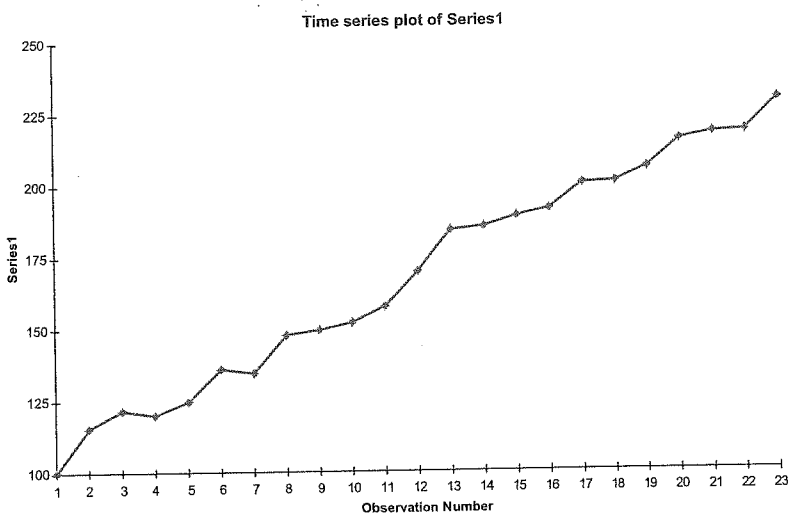


Figure 13.8
A Series with Increasing Variance Through Time

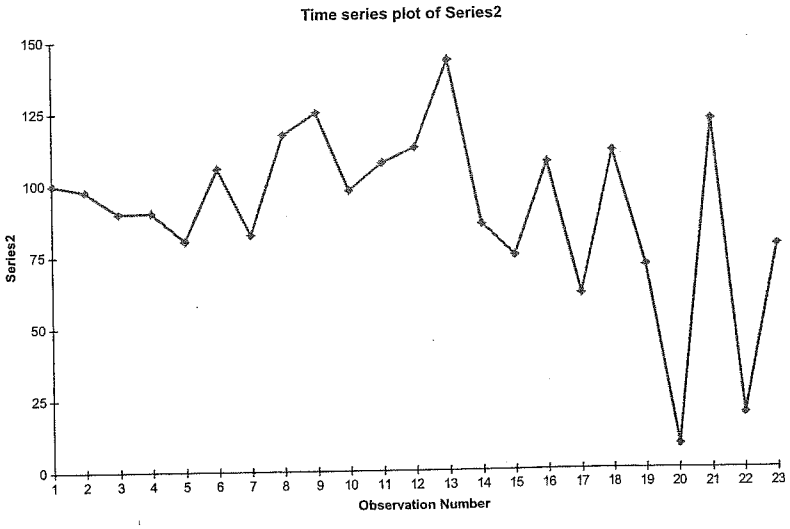


Figure 13.9
A Series with Seasonality

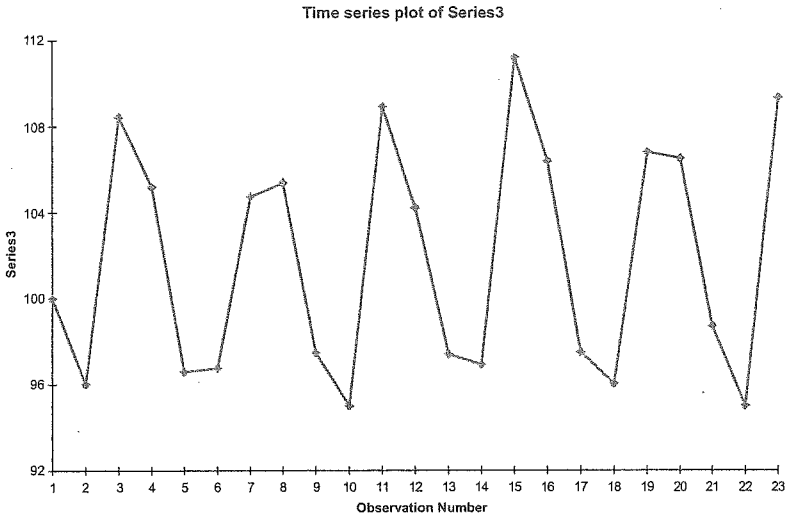


Figure 13.10
A Series That Meanders

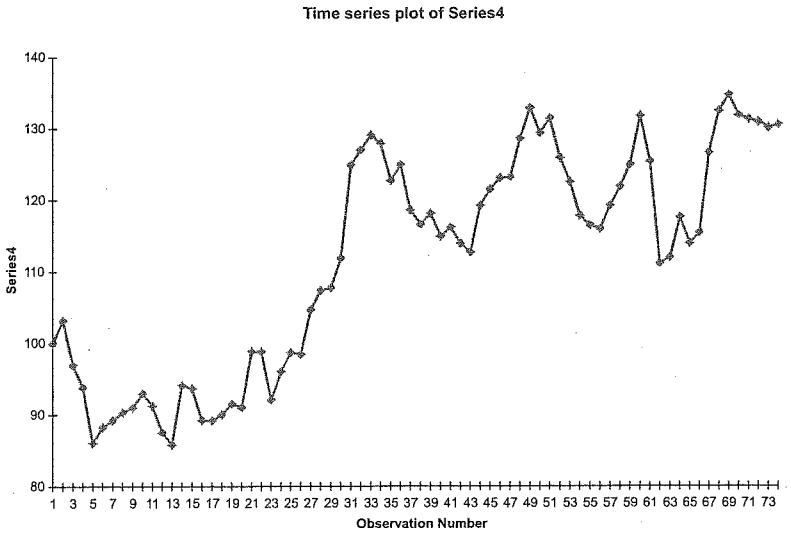
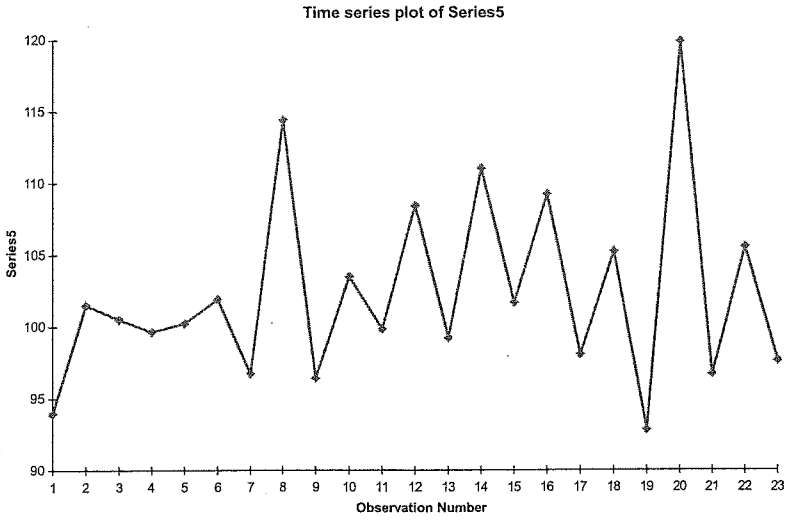


Figure 13.11
A Series That Oscillates Frequently



13.3.1 The Runs Test

It is not always easy to detect randomness or the lack of it from the visual inspection of a graph. Therefore, we discuss two quantitative methods of that test for randomness. The first is called the *runs test.* We first choose a base value, which could be the average value of the series, the median value, or even some other value. Then we define a **run** as a consecutive series of observations that remain on one side of this base level. For example, if the base level is 0 and we observe the series 1, 5, 3, −3, −2, −4, −1, 3, 2, then there are three runs: 1, 5, 3; −3, −2, −4, −1; and 3, 2. The idea behind the runs test is that a random series should have a number of runs that is neither too large nor too small. If the series has too few runs, then it could be trending (as in Figure 13.7) or it could be meandering (as in Figure 13.10). If the series has too many runs, then it is zigzagging too often (as in Figure 13.11).

This runs test can be used on any time series, not just a series of residuals.

The **runs test** is a formal test of the null hypothesis of randomness. If there are too many or too few runs in the series, then we conclude that the series is not random.

We do not provide the mathematical details of the runs test, but we illustrate how it is implemented in StatTools in the following example.

EXAMPLE 13.1 FORECASTING MONTHLY STEREO SALES

Monthly sales for a chain of stereo retailers are listed in the file **Stereo Sales.xlsx**. They cover the period from the beginning of 2004 to the end of 2007, during which there was no upward or downward trend in sales and no clear seasonal peaks or valleys. This behavior is apparent in the time series graph of sales in Figure 13.12. Therefore, a simple forecast model of sales is to use the *average* of the series, 182.67, as a forecast of sales for each month. Do the resulting residuals represent random noise?

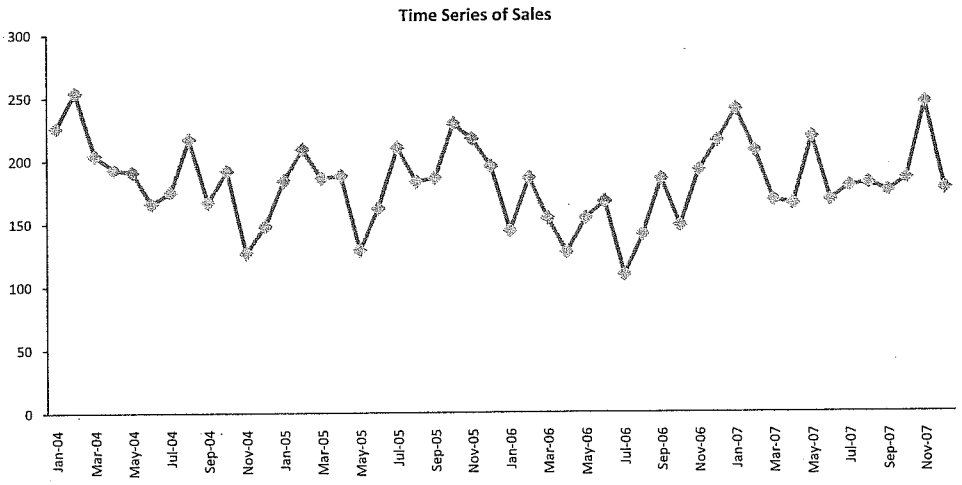


Figure 13.12
Time Series Graph of Stereo Sales

Objective To use StatTools's Runs Test procedure to check whether the residuals from this simple forecasting model represent random noise.

Solution

To obtain the residuals for this forecasting model, we subtract the average, 182.67, from each observation. Therefore, the plot of the residuals, shown in Figure 13.13, has exactly the same shape as the plot of sales. The only difference is that it is shifted down by 182.67 and has mean 0. We now use the runs test to check whether there are too many or too few runs around the base value of 0 in this residual plot. To do so, we select Runs Test for Randomness from the StatTools Time Series and Forecasting dropdown, choose Residual as the variable to analyze, and choose Mean of Series as the cutoff value. (This corresponds to the horizontal line at 0 in Figure 13.13.) This produces the output in shown in Figure 13.14.

Figure 13.13
Time Series Graph of Residuals

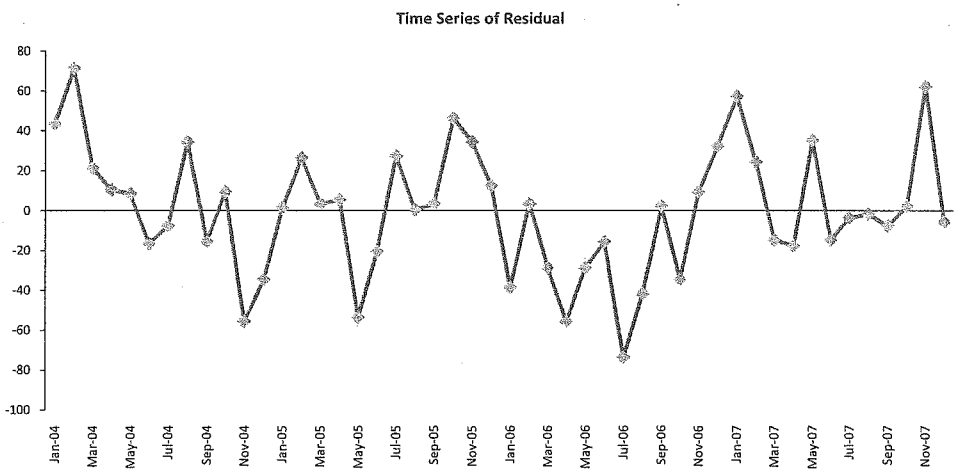


Figure 13.14
Runs Test for Randomness

	I	J
7		Residual
8	*Runs Test for Randomness*	Data Set #1
9	Observations	48
10	Below Mean	22
11	Above Mean	26
12	Number of Runs	20
13	Mean	0.00
14	E(R)	24.8333
15	StdDev(R)	3.4027
16	Z-Value	-1.4204
17	P-Value (two-tailed)	0.1555

A small p-value in the runs test provides evidence of nonrandomness.

The important elements of this output are the following:

- The number of observed runs is 20, in cell J12.
- The number of runs *expected* under an assumption of randomness is 24.833, in cell J14. Therefore, the series of residuals has too *few* runs. Positive values tend to follow positive values, and negative values tend to follow negative values.
- The z-value in cell J16, –1.42, indicates how many standard errors the observed number of runs is below the expected number of runs. The corresponding p-value

indicates how extreme this z-value is. It can be interpreted just like other p-values for hypothesis tests. If it is small, say, less than 0.05, then we can reject the null hypothesis of randomness and conclude that the series of residuals is *not* random noise. However, the p-value for this example is only 0.1555. Therefore, there is not convincing evidence of nonrandomness in the residuals, and we can conclude that the residuals represent noise. ■

13.3.2 Autocorrelation

Like the runs test, autocorrelations can be calculated for any time series, not just a series of residuals.

In this section we discuss another way to check for randomness of a time series of residuals—we examine its *autocorrelations*. The "auto" means that successive observations are correlated with one another. For example, in the most common form of autocorrelation, *positive* autocorrelation, large observations tend to follow large observations, and small observations tend to follow small observations. In this case the runs test is likely to pick it up because there will be fewer runs than expected. Another way to check for the same nonrandomness property is to calculate the autocorrelations of the time series.

An **autocorrelation** is a type of correlation used to measure whether values of a time series are related to their own past values.

To understand autocorrelations it is first necessary to understand what it means to *lag* a time series. This concept is easy to illustrate in a spreadsheet. We again use the monthly stereo sales data in the **Stereo Sales.xlsx** file. To lag by 1 month, we simply "push down" the series by one row. See column D of Figure 13.15. Note that there is a blank cell at the top of the lagged series (in cell D2). We can continue to push the series down one row at a time to obtain other lags. For example, the lag 3 version of the series appears in column F. Now there are three missing observations at the top. Note that in December 2004, say, the first, second, and third lags correspond to the observations in November 2004, October 2004, and September 2004, respectively. That is, lags are simply previous observations, removed by a certain number of periods from the present time. These lagged columns can be obtained by copying and pasting the original series or by using Lag from the StatTools Utilities dropdown.

Figure 13.15
Lags for Stereo Sales

	A	B	C	D	E	F
1	Month	Sales	Residual	Lag1(Residual)	Lag2(Residual)	Lag3(Residual)
2	Jan-04	226	43.333			
3	Feb-04	254	71.333	43.333		
4	Mar-04	204	21.333	71.333	43.333	
5	Apr-04	193	10.333	21.333	71.333	43.333
6	May-04	191	8.333	10.333	21.333	71.333
7	Jun-04	166	-16.667	8.333	10.333	21.333
8	Jul-04	175	-7.667	-16.667	8.333	10.333
9	Aug-04	217	34.333	-7.667	-16.667	8.333
10	Sep-04	167	-15.667	34.333	-7.667	-16.667
11	Oct-04	192	9.333	-15.667	34.333	-7.667
12	Nov-04	127	-55.667	9.333	-15.667	34.333
13	Dec-04	148	-34.667	-55.667	9.333	-15.667
14	Jan-05	184	1.333	-34.667	-55.667	9.333
15	Feb-05	209	26.333	1.333	-34.667	-55.667
16	Mar-05	186	3.333	26.333	1.333	-34.667

Then the autocorrelation of lag k, for any integer k, is essentially the correlation between the original series and the lag k version of the series. For example, in Figure 13.15 the lag 1 autocorrelation is the correlation between the observations in columns C and D. Similarly, the lag 2 autocorrelation is the correlation between the observations in columns C and E.[1]

We have shown the lagged versions of Sales in Figure 13.15, and we have explained autocorrelations in terms of these lagged variables, to help motivate the concept of autocorrelation. However, we can use StatTools's Autocorrelation procedure directly, *without* forming the lagged variables, to calculate autocorrelations. This is illustrated in the following continuation of Example 13.1.

EXAMPLE 13.1 FORECASTING MONTHLY STEREO SALES (CONTINUED)

The runs test on the stereo sales data suggests that the pattern of sales is not random. Large values tend to follow large values, and small values tend to follow small values. Do autocorrelations support this conclusion?

Objective To examine the autocorrelations of the residuals from the forecasting model for evidence of nonrandomness.

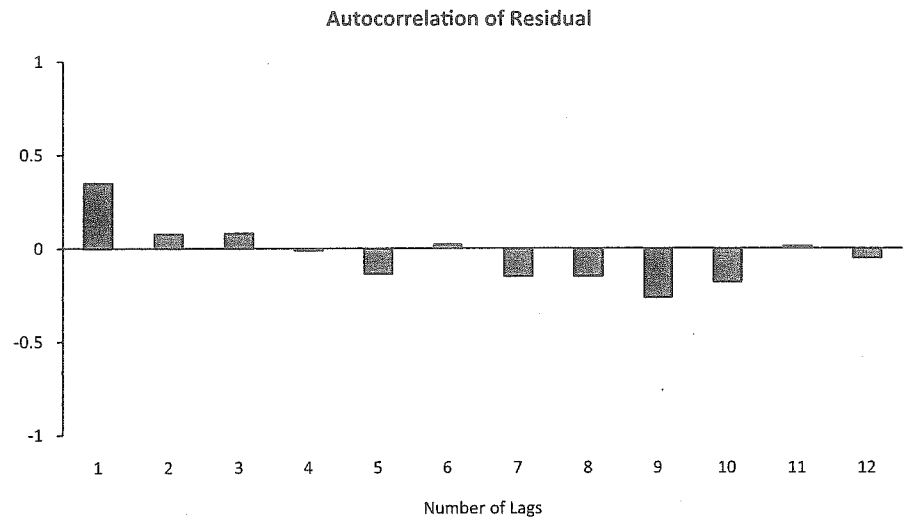
Solution

We use StatTools's Autocorrelation procedure, found under the StatTools Time Series and Forecasting dropdown. It requires us to specify the time series variable (Residual), the number of lags we want (we chose the StatTools default value), and whether we want a chart of the autocorrelations. This chart is called a **correlogram**. The resulting autocorrelations and correlogram appear in Figure 13.16. A typical autocorrelation of lag k indicates the relationship between observations k periods apart. For example, the autocorrelation of lag 3, 0.0814, indicates that there is very little relationship between residuals separated by 3 months.

How large is a "large" autocorrelation? Under the assumption of randomness, it can be shown that the standard error of any autocorrelation is approximately $1/\sqrt{T}$, in this case $1/\sqrt{48} = 0.1443$. (Recall that T denotes the number of observations in the series.) If the series is truly random, then only an occasional autocorrelation will be larger than 2 standard errors in magnitude. Therefore, any autocorrelation that *is* larger than 2 standard errors in magnitude is worth our attention. These significantly nonzero autocorrelations are boldfaced in the StatTools output. The only "large" autocorrelation for the residuals is the first, or lag 1, autocorrelation of 0.3492. The fact that it is *positive* indicates once again that there is some tendency for large residuals to follow large residuals and for small to follow small. The autocorrelations for other lags are less than two standard errors in magnitude and can be ignored.

[1]We ignore the exact details of the calculations here. Just be aware that the formula for autocorrelations that is usually used differs slightly from the correlation formula in Chapter 3. However, the difference is very slight and of little practical importance.

Figure 13.16
Correlogram and Autocorrelations of Residuals



Autocorrelation Table	Residual Data Set #1
Number of Values	48
Standard Error	0.1443
Lag #1	0.3492
Lag #2	0.0772
Lag #3	0.0814
Lag #4	-0.0095
Lag #5	-0.1353
Lag #6	0.0206
Lag #7	-0.1494
Lag #8	-0.1492
Lag #9	-0.2626
Lag #10	-0.1792
Lag #11	0.0121
Lag #12	-0.0516

Typically, we can ask for autocorrelations up to as many lags as we like. However, there are several practical considerations to keep in mind. First, it is common practice to ask for no more lags than 25% of the number of observations. For example, if there are 48 observations, we should ask for no more than 12 autocorrelations (lags 1–12). (StatTools chooses this number of lags if you accept its "auto" setting.)

Second, the first few lags are typically the most important. Intuitively, if there is any relationship between successive observations, it is likely to be between nearby observations. The June 2007 observation is more likely to be related to the May 2007 observation than to the October 2006 observation. Sometimes there is a fairly large spike in the correlogram at some large lag, such as lag 9. However, this can often be ignored as a random "blip" unless there is some obvious reason for its occurrence. A similarly large autocorrelation at lag 1 or 2 is usually taken more seriously. The one exception to this is a *seasonal* lag. For example, for monthly data an autocorrelation at lag 12 corresponds to a relationship between observations a year apart, such as May 2007 and May 2006. If this autocorrelation is significantly large, it probably should not be ignored.

Autocorrelation analysis is somewhat advanced. However, it is the basis for many important forecasting methods.

We do not examine autocorrelations much further in this book. However, many advanced forecasting techniques are based largely on the examination of the autocorrelation structure of time series. This autocorrelation structure tells us how a series is related to its own past values through time, which can be very valuable information for forecasting *future* values.

PROBLEMS

Level A

1. The file **P13_01.xlsx** contains the monthly number of airline tickets sold by the CareFree Travel Agency. Is this time series *random*? Perform a runs test and compute a few autocorrelations to support your answer.
2. The file **P13_02.xlsx** contains the weekly sales at a local bookstore for each of the past 25 weeks. Is this time series *random*? Perform a runs test and compute a few autocorrelations to support your answer.
3. The number of employees on the payroll at a food-processing plant is recorded at the start of each month. These data are provided in the file **P13_03.xlsx**. Perform a runs test and compute a few autocorrelations to determine whether this time series is random.
4. The quarterly numbers of applications for home mortgage loans at a branch office of Northern Central Bank are recorded in the file **P13_04.xlsx**. Perform a runs test and compute a few autocorrelations to determine whether this time series is random.
5. The number of reported accidents at a manufacturing plant located in Flint, Michigan, was recorded at the start of each month. These data are provided in the file **P13_05.xlsx**. Is this time series *random*? Perform a runs test and compute a few autocorrelations to support your answer.
6. The file **P13_06.xlsx** contains the weekly sales at the local outlet of WestCoast Video Rentals for each of the past 36 weeks. Perform a runs test and compute a few autocorrelations to determine whether this time series is random.

Level B

7. Determine whether the RAND() function in Excel actually generates a random stream of numbers. Generate at least 100 random numbers to test their randomness with a runs test and with autocorrelations. Summarize your findings.
8. Use a runs test and calculate autorrelations to decide whether the random series explained in each part (a–c) are random. For each part, generate at least 100 random numbers in the series.
   a. A series of independent normally distributed values, each with mean 70 and standard deviation 5.
   b. A series where the first value is normally distributed with mean 70 and standard deviation 5, and each succeeding value is normally distributed with mean equal to the *previous* value and standard deviation 5. (For example, if the fourth value is 67.32, then the fifth value will be normally distributed with mean 67.32.)
   c. A series where the first value, Y_1, is normally distributed with mean 70 and standard deviation 5, and each succeeding value, Y_t, is normally distributed with mean $(1 + a_t)Y_{t-1}$ and standard deviation $5(1 + a_t)$, where the a_t's are independent, normally distributed values with mean 0 and standard deviation 0.2. (For example, if $Y_{t-1} = 67.32$ and $a_t = -0.2$, then Y_t will be normally distributed with mean $0.8(67.32) = 53.856$ and standard deviation $0.8(5) = 4$.)

13.4 REGRESSION-BASED TREND MODELS

Many time series follow a long-term trend except for random variation. This trend can be upward or downward. A straightforward way to model this trend is to estimate a regression equation for Y_t, using time t as the *single* explanatory variable. In this section we discuss the two most frequently used trend models, *linear* trend and *exponential* trend.

13.4.1 Linear Trend

A linear trend means that the time series variable changes by a constant *amount* each time period. The relevant equation is equation (13.6), where, as in previous regression equations, a is the intercept, b is the slope, and ϵ_t is an error term.

Linear Trend Model

$$Y_t = a + bt + \epsilon_t \tag{13.6}$$

The interpretation of b is that it represents the expected change in the series from one period to the next. If b is positive, the trend is upward; if b is negative, the trend is downward. The intercept term a is less important. It literally represents the expected value of the series at time $t = 0$. If time t is coded so that the first observation corresponds to $t = 1$, then a is where we expect the series to have been one period before we started observing. However, it is possible that time is coded in another way. For example, we might have annual data that start in 1997. Then the first value of t might be entered as 1997, which means that the intercept a corresponds to a period 1997 years earlier! Clearly, we would not take its value literally in this case.

As always, a graph of the time series is a good place to start. It indicates whether a linear trend model is likely to provide a good fit. Generally, the graph should rise or fall at approximately a constant rate through time, without too much random variation. But even if there is a lot of random variation—a lot of zigzags—a linear trend to the data might be a good starting point. Then the *residuals* from this trend line, which should have no remaining trend, could possibly be modeled by some other method in this chapter.

EXAMPLE 13.2 QUARTERLY PHARMACEUTICAL SALES

The file **Pharmaceutical Sales.xlsx** contains quarterly sales data for a large pharmaceutical company from first quarter 1998 through fourth quarter 2007 (in millions of dollars). The time series graph of these data appears in Figure 13.17. Sales increase from \$3062 million in the initial quarter to \$8307 million in the final quarter. How well does a linear trend fit these data? Are the residuals from this fit random?

Figure 13.17 Time Series Graph of Pharmaceutical Sales

Objective To fit a linear trend line to quarterly sales and examine its residuals for randomness.

Solution

The graph in Figure 13.17 indicates a clear upward trend with little or no curvature. Therefore, a linear trend is certainly plausible. To estimate it with regression, we first need

a *numeric* time variable—labels such as Q1-98 will not do. We construct this time variable in column C of the data set, using the consecutive values 1 through 40. We then run a simple regression of Sales versus Time, with the results shown in Figure 13.18. The estimated linear trend line is

$$\text{Forecasted Sales} = 2686.7 + 131.991\text{Time}$$

Figure 13.18
Regression Output for Linear Trend

Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate
	0.9806	0.9615	0.9605	312.87

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value
Explained	1	92856588.48	92856588.48	948.5801	< 0.0001
Unexplained	38	3719823.497	97890.09201		

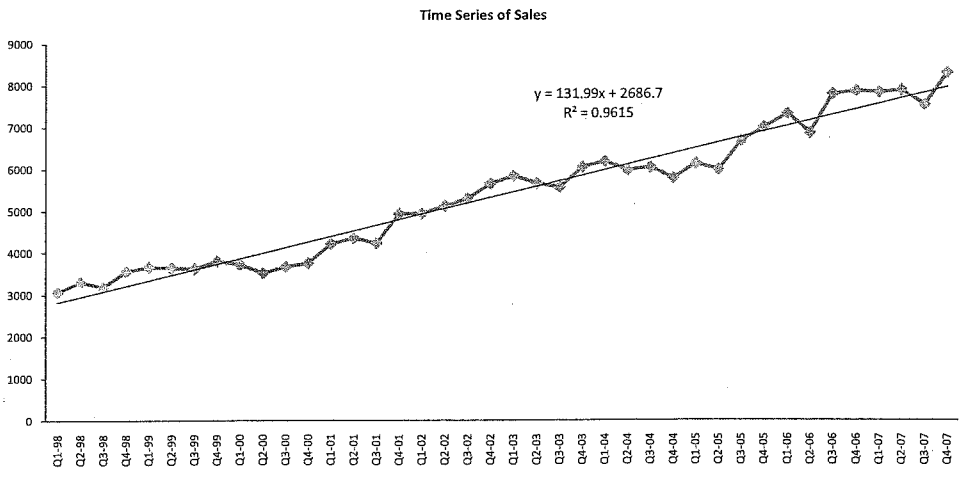
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	2686.72	100.82	26.6476	< 0.0001	2482.61	2890.83
Time	131.991	4.29	30.7990	< 0.0001	123.31	140.67

This equation implies that we expect sales to increase by $131.991 million per quarter. (The 2686.7 value in this equation is what we would predict sales to be at time 0—quarter 4 of 1997.) To use this equation to forecast future sales, we substitute later values of Time into the regression equation, so that each future prediction is $131.991 larger than the previous prediction. For example, the forecast for Q4 of 2008 is

$$\text{Forecasted Sales Q4-08} = 2686.7 + 131.991(44) = 8494.3$$

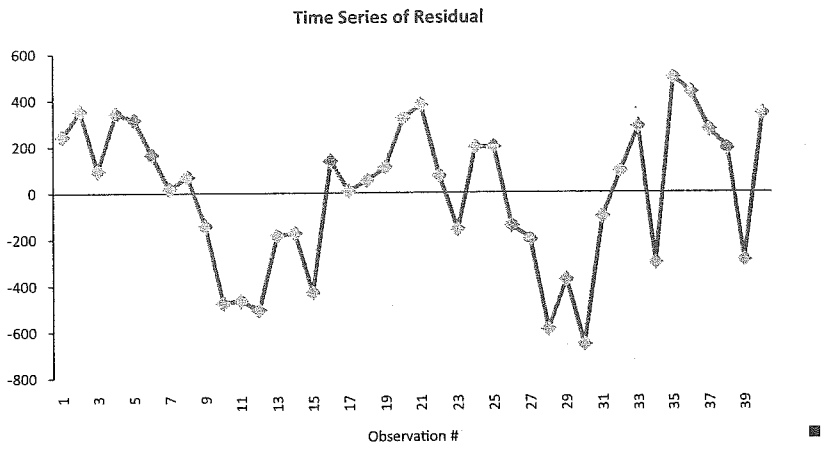
Excel provides an easier way to obtain this trend line. Once the graph in Figure 13.17 is constructed, we can use Excel's Trendline tool. (Select the chart and then select More Trendline Options from the Trendline dropdown on the Chart Tools Layout ribbon.) This gives us several types of trend lines to choose from, and we select the linear option for this example. We can also check the options to show the regression equation and its R^2 value on the chart, as we have done in Figure 13.19. This superimposed trend line indicates a reasonably good fit.

Figure 13.19
Time Series Graph with Linear Trend Superimposed



However, the fit is not perfect, as the plot of the residuals in Figure 13.20 indicates. These residuals tend to "meander," staying positive for a while, then negative, then positive, and so on. You can check that the runs test for these residuals produces a z-value of -3.074, with a corresponding p-value of 0.002, and that its first two autocorrelations are significantly positive. In short, these residuals are *not* random noise, and they could be modeled further. However, we do not pursue this analysis here.

Figure 13.20
Time Series Graph of Residuals



13.4.2 Exponential Trend

In contrast to a linear trend, an exponential trend is appropriate when the time series changes by a constant *percentage* (as opposed to a constant dollar amount) each period. Then the appropriate regression equation is equation (13.7), where c and b are constants, and u_t represents a *multiplicative* error term.

Exponential Trend Model

$$Y_t = ce^{bt}u_t \tag{13.7}$$

An exponential trend for Y is equivalent to a linear trend for the logarithm of Y.

Equation (13.7) is useful for understanding how an exponential trend works, as we will discuss, but it is not useful for estimation. For that, we require a *linear* equation. Fortunately, we can achieve linearity by taking natural logarithms of both sides of equation (13.7). (The key, as usual, is that the logarithm of a product is the sum of the logarithms.) The result appears in equation (13.8), where $a = \ln(c)$ and $\epsilon_t = \ln(u_t)$. This equation represents a *linear* trend, but the dependent variable is now the logarithm of the original Y_t. This implies the following important fact: If a time series exhibits an exponential trend, then a plot of its logarithm should be approximately linear.

***Equivalent Linear Trend for Logarithm of* Y**

$$\ln(Y_t) = a + bt + \epsilon_t \tag{13.8}$$

Because the computer does the calculations, our main responsibility is to interpret the final result. This is not too difficult. It can be shown that the coefficient b (expressed as a percentage) is approximately the percentage change per period. For example, if $b = 0.05$, then the series is increasing by approximately 5% per period.[2] On the other hand, if $b = -0.05$, then the series is decreasing by approximately 5% per period.

An exponential trend can be estimated with StatTools's Regression procedure, but only after the log transformation has been made on Y_t. We illustrate this in the following example.

EXAMPLE 13.3 QUARTERLY PC DEVICE SALES

The file **PC Device Sales.xlsx** contains quarterly sales data (in millions of dollars) for a large PC device manufacturer from the first quarter of 1993 through the fourth quarter of 2007. Are the company's sales growing exponentially through this entire period?

Objective To estimate the company's exponential growth and to see whether it has been maintained during the entire period from 1993 until the end of 2007.

Solution

We first estimate and interpret an exponential trend for the years 1993 through 2003. Then we see how well the projection of this trend into the future fits the data after 2003. The time series graph through 2003 appears in Figure 13.21. We have used Excel's Trendline tool, with the Exponential option, to superimpose an exponential trend line and the corresponding equation on this plot. The fit is evidently quite good. Equivalently, Figure 13.22 illustrates the time series of log sales for this same period, with a *linear* trend line superimposed. Its fit is equally good.

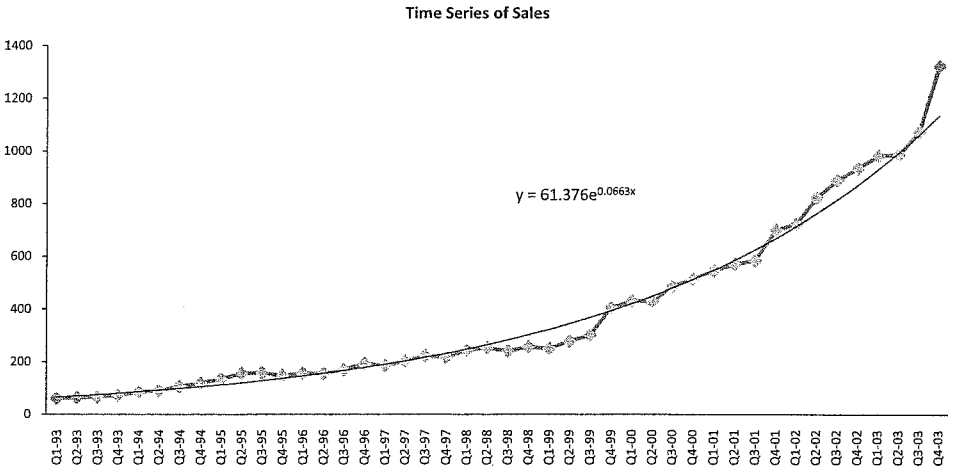
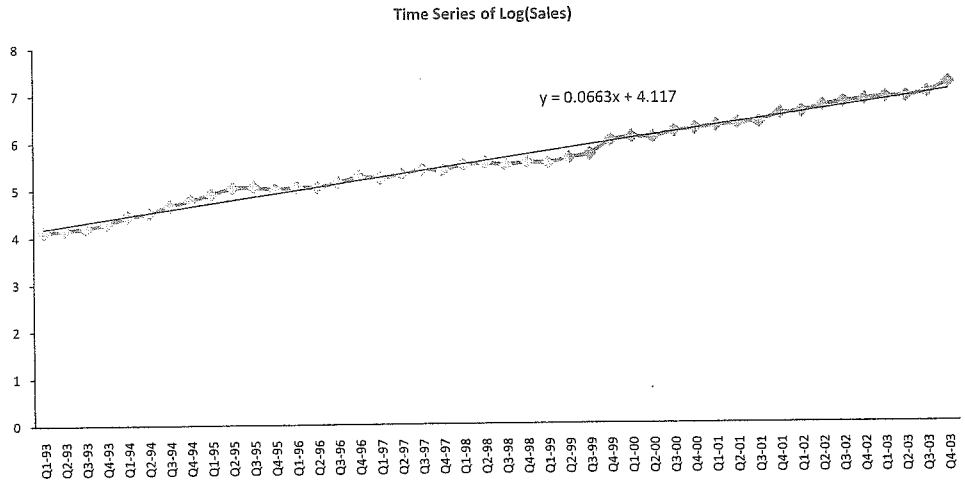


Figure 13.21 Time Series Graph of Sales with Exponential Trend Superimposed

[2]More precisely, this percentage change is $e^b - 1$. For example, when $b = 0.05$, this is $e^b - 1 = 5.13\%$.

Figure 13.22
Time Series Graph of Log Sales with Linear Trend Superimposed



We can also use StatTools's Regression procedure to estimate this exponential trend, as shown in Figure 13.23. To produce this output, we must first add a time variable in column C (with values 1 through 44) and make a logarithmic transformation of Sales in column D. Then we regress Log(Sales) on Time (using the data through 2003 only) to obtain the regression output. Note that its two coefficients in cells B18 and B19 are the same as those shown for the linear trend in Figure 13.22. If we take the antilog of the constant 4.117 (with the formula = EXP(B18)), we obtain the constant *multiple* shown in Figure 13.21. It corresponds to the constant c in equation (13.7).

Figure 13.23
Regression Output for Estimating Exponential Trend

	A	B	C	D	E	F	G
7							
8	*Summary*	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
9		0.9922	0.9844	0.9840	0.1086		
10							
11							
12	*ANOVA Table*	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
13	Explained	1	31.21992793	31.21992793	2645.6403	< 0.0001	
14	Unexplained	42	0.495621782	0.011800519			
15							
16						Confidence Interval 95%	
17	*Regression Table*	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
18	Constant	4.1170	0.0333	123.5616	< 0.0001	4.0498	4.1843
19	Time	0.0663	0.0013	51.4358	< 0.0001	0.0637	0.0689

What does it all mean? The estimated equation (13.7) is

$$\text{Forecasted Sales} = 61.376e^{0.0663t}$$

The most important constant in this equation is the coefficient of Time, $b = 0.0663$. Expressed as a percentage, this coefficient implies that the company's sales increased by approximately 6.63% per quarter throughout this 11-year period. (The constant multiple, $c = 61.376$, is our forecast of sales at time 0—in quarter 4 of 1992.) To use this equation for forecasting the future, we substitute later values of Time into the regression equation, so that each future forecast is about 6.63% larger than the previous forecast. For example, the forecast of the second quarter of 2004 is

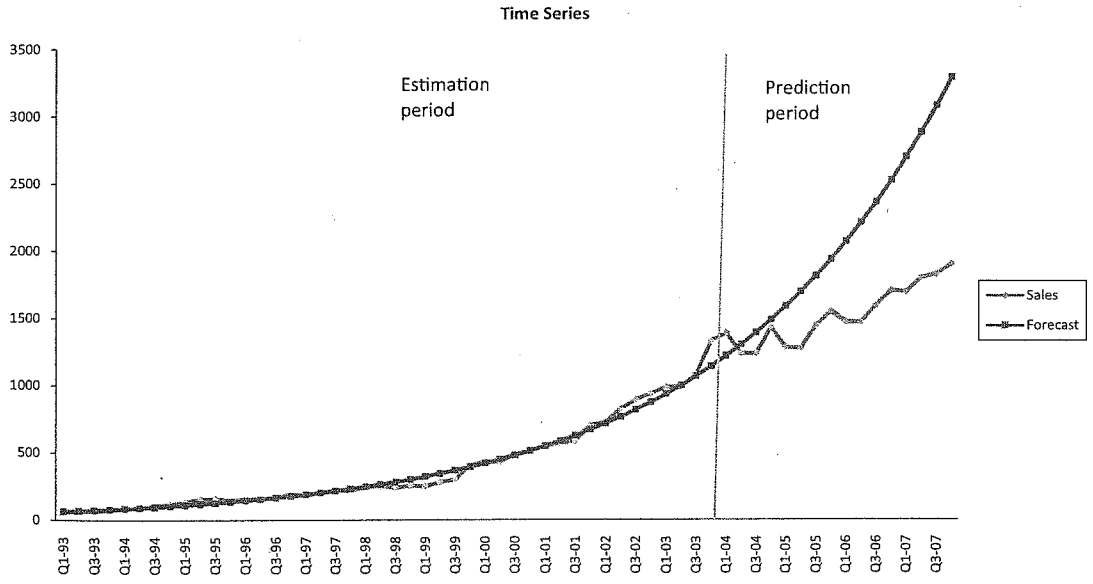
$$\text{Forecasted Sales in Q2-04} = 61.376e^{0.0663(46)} = 1295.72$$

Has this exponential growth continued beyond 2003? It has *not,* due possibly to slumping sales in the computer industry or increased competition from other manufacturers. We checked this by creating the Forecast column in Figure 13.24 (by substituting into the regression equation for the entire period though Q4–07). We then used StatTools to create a time series graph of the two series Sales and Forecast, shown in Figure 13.25. It is clear that sales in the forecast period did not exhibit nearly the 6.63% growth observed in the estimation period. As the company clearly realizes, nothing that good lasts forever.

Figure 13.24
Creating Forecasts of Sales

	A	B	C	D	E
1	Quarter	Sales	Time	Log(Sales)	Forecast
2	Q1-93	61.14	1	4.1131663	65.58583
3	Q2-93	64.07	2	4.1599762	70.08398
4	Q3-93	66.18	3	4.1923783	74.89063
5	Q4-93	72.76	4	4.2871664	80.02694
6	Q1-94	84.70	5	4.4391156	85.51552
7	Q2-94	90.05	6	4.5003651	91.38053
8	Q3-94	106.06	7	4.664005	97.64778
9	Q4-94	118.21	8	4.7724627	104.3449
10	Q1-95	134.38	9	4.9006716	111.5013
11	Q2-95	154.67	10	5.0412938	119.1485
12	Q3-95	157.41	11	5.0588539	127.3202
13	Q4-95	147.16	12	4.9915204	136.0523

Figure 13.25 Time Series Graph of Forecasts Superimposed on Sales for the Entire Period



Before leaving this example, we comment briefly on the standard error of estimate shown in cell E9 of Figure 13.23. This value, 0.1086, is in *log* units, not original dollar units. Therefore, it is a totally misleading indicator of the forecast errors we might make from the exponential trend equation. To obtain more meaningful measures, we first obtain the forecasts of sales, as explained previously. Then we can easily obtain any of the three forecast error measures discussed previously in equations (13.2), (13.3), and (13.4). The results appear in Figure 13.26. The squared errors, absolute errors, and absolute percentage errors are first calculated with the formulas =(B2-E2)^2, =ABS(B2-E2), and =G2/B2 in cells F2, G2, and H2, which are then copied down. The error measures (for the data through 2003 only) then appear in cells K2, K3, and K4. The corresponding formulas for RMSE, MAE, and MAPE are straightforward. RMSE is the square root of the average of the squared errors in column F, and MAE and MAPE are the averages of the values in columns G and H, respectively. The latter is particularly simple to interpret. Forecasts for the 11-year estimate period were off, on average, by 7.86%. (Of course, as you can check, forecasts for the quarters *after* 2003 were off by much more!)

Figure 13.26
Measures of Forecast Errors

	A	B	C	D	E	F	G	H	I	J	K	L
1	Quarter	Sales	Time	Log(Sales)	Forecast	SqError	AbsError	AbsPctError		Measures of forecast error		
2	Q1-93	61.14	1	4.1131663	65.58583	19.76541	4.445831	0.07271559		RMSE	41.86	
3	Q2-93	64.07	2	4.1599762	70.08398	36.16795	6.013979	0.09386576		MAE	25.44	
4	Q3-93	66.18	3	4.1923783	74.89063	75.87506	8.710629	0.13162027		MAPE	7.86%	
5	Q4-93	72.76	4	4.2871664	80.02694	52.8084	7.266939	0.09987547				
6	Q1-94	84.70	5	4.4391156	85.51552	0.66507	0.815518	0.00962831				
7	Q2-94	90.05	6	4.5003651	91.38053	1.770302	1.330527	0.01477542				
8	Q3-94	106.06	7	4.664005	97.64778	70.7654	8.412218	0.07931565				

Whenever we observe a time series that is increasing at an increasing rate (or decreasing at a decreasing rate), an exponential trend model is worth trying. The key to the analysis is to regress the *logarithm* of the time series variable versus time (or use Excel's Trendline tool). The coefficient of time, written as a percentage, is then the approximate percentage increase (if positive) or decrease (if negative) per period.

PROBLEMS

Level A

9. The file **P13_01.xlsx** contains the monthly number of airline tickets sold by the CareFree Travel Agency.
   a. Does a linear trend appear to fit these data well? If so, estimate and interpret the linear-trend model for this time series. Also, interpret the R^2 and s_e values.
   b. Provide an indication of the typical forecast error generated by the estimated model in part **a**.
   c. Is there evidence of some seasonal pattern in these sales data? If so, characterize the seasonal pattern.

10. The file **P13_10.xlsx** contains the daily closing prices of Wal-Mart stock for a 1-year period. Does a linear or exponential trend fit these data well? If so, estimate and interpret the best trend model for this time series. Also, interpret the R^2 and s_e values.

11. The file **P13_11.xlsx** contains annual data on the amount of life insurance in force in the United States. Fit an exponential growth curve to these data. Write a short report to summarize your findings.

12. The file **P13_12.xlsx** contains 5 years of monthly data on sales (number of units sold) for a particular company. The company suspects that except for random noise, its sales are growing by a constant *percentage* each month and that they will continue to do so for at least the near future.
   a. Explain briefly whether the plot of the series visually supports the company's suspicion.
   b. Fit the appropriate regression model to the data. Report the resulting equation and state explicitly what it says about the percentage growth per month.
   c. What are the RMSE and MAPE for the forecast model in part **b**? In words, what do they measure? Considering their magnitudes, does the model seem to be doing a good job?
   d. In words, how does the model make forecasts for future months? Specifically, given the forecast

value for the last month in the data set, what simple arithmetic could you use to obtain forecasts for the next few months?

13. The file **P13_13.xlsx** contains quarterly data on GDP. (The data are expressed in billions of current dollars, they are seasonally adjusted, and they represent annualized rates.)
    a. Look at a time series plot of GDP. Does it suggest a linear relationship; an exponential relationship?
    b. Use regression to estimate a linear relationship between GDP and Time. Interpret the associated "constant" term and the "slope" term. Would you say that the fit is good?

Level B

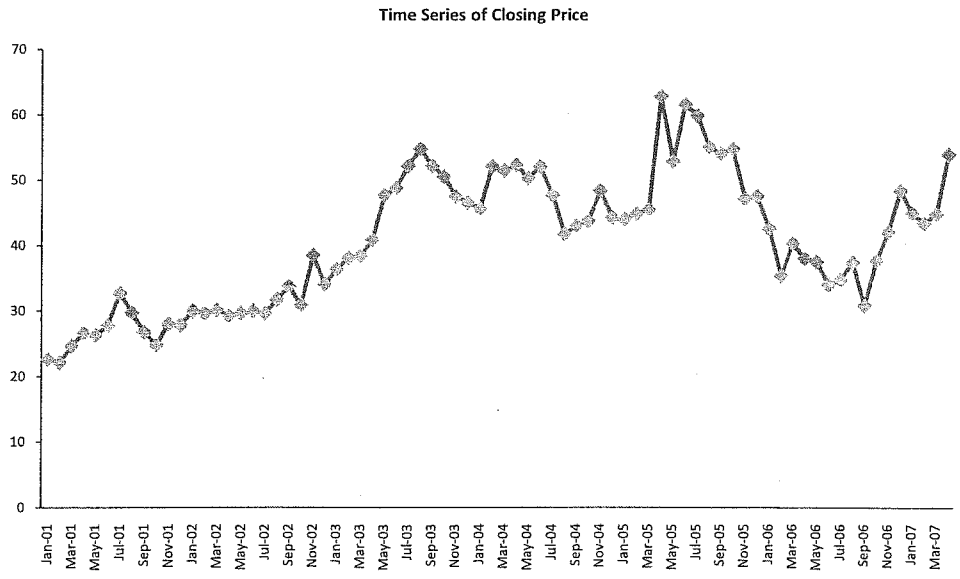
14. The file **P13_14.xlsx** gives monthly exchange rates (dollars per unit of local currency) for 25 countries. Technical analysts believe that by charting past changes in exchange rates, it is possible to predict future changes of exchange rates. After analyzing the autocorrelations for these data, do you believe that technical analysis has potential?

15. The unit sales of a new drug for the first 25 months after its introduction to the marketplace are recorded in the file **P13_15.xlsx**.
    a. Estimate a linear trend equation using the given data. How well does the linear trend fit these data? Are the residuals from this linear trend model *random*?
    b. If the residuals from this linear trend model are *not* random, propose another regression-based trend model that more adequately explains the long-term trend in this time series. Estimate the alternative model(s) using the given data. Check the residuals from the model(s) for randomness. Summarize your findings.
    c. Given the best estimated model of the trend in this time series, interpret R^2 and s_e.

13.5 THE RANDOM WALK MODEL

Random series are sometimes building blocks for other time series models. The model we now discuss, the **random walk** model, is an example of this. In a random walk model the series itself is not random. However, its *differences*—that is, the changes from one period to the next—are random. This type of behavior is typical of stock price data (as well as various other time series data). For example, the graph in Figure 13.27 shows monthly closing prices for a tractor manufactor's stock from January 2001 through April 2007. (See the file **Tractor Closing Prices.xlsx**.) This series is not random, as can be seen from its gradual upward trend at the beginning and the general meandering behavior throughout. (Although the runs test and autocorrelations are not shown for the series itself, they confirm that the series is not random. There are significantly *fewer* runs than expected, and the autocorrelations are significantly *positive* for many lags.)

Figure 13.27
Time Series Graph of Tractor Stock Prices



If we were standing in April 2007 and were asked to forecast the company's prices for the next few months, it is intuitive that we would not use the average of the historical values as our forecast. This forecast would tend to be too low because of the upward trend. Instead, we might base our forecast on the most recent observation. This is exactly what the random walk model does.

Equation (13.9) for the random walk model is given in the box, where μ is a constant and ϵ_t is a random series (noise) with mean 0 and some standard deviation σ that remains *constant* through time.

Random Walk Model

$$Y_t = Y_{t-1} + \mu + \epsilon_t \tag{13.9}$$

If we let $DY_t = Y_t - Y_{t-1}$, the change in the series from time t to time $t-1$ (where D stands for difference), then we can write the random walk model as in equation (13.10). This implies that the differences form a random series with mean μ and standard deviation σ. An estimate of μ is the average of the differences, labeled $\bar{Y}_D$, and an estimate of σ is the sample standard deviation of the differences, labeled s_D.

Difference Form of Random Walk Model

$$DY_t = \mu + \epsilon_t \tag{13.10}$$

In words, a series that behaves according to this random walk model has random differences, and the series tends to trend upward (if $\mu > 0$) or downward (if $\mu < 0$) by an amount μ each period. If we are standing in period t and want to forecast Y_{t+1}, then a reasonable forecast is given by equation (13.11). That is, we add the estimated trend to the current observation to forecast the next observation.

One-Step-Ahead Forecast for Random Walk Model

$$F_{t+1} = Y_t + \bar{Y}_D \tag{13.11}$$

We illustrate this method in the following example.

EXAMPLE 13.4 Random Walk Model of Stock Prices

The monthly closing prices of the tractor company's stock from January 2001 through April 2007, shown in Figure 13.27, indicate some upward trend. Does this series follow a random walk model with an upward trend? If so, how should future values of these stock prices be forecasted?

Objective To check whether the company's monthly closing prices follow a random walk model with an upward trend, and to see how future prices can be forecasted.

Solution

We have already seen that the closing price series itself is not random, due to the upward trend. To check for the adequacy of a random walk model, we need the *differenced* series. Each value in the differenced series is that month's closing price minus the previous

month's closing price. This series can be calculated easily with an Excel formula, or it can be generated automatically with the Difference item on the StatTools Utilities dropdown. (When asked for the *number* of difference variables, we accept the default value of 1.) This differenced series appears in column C of Figure 13.28. This figure also shows the mean and standard deviation of the differences, 0.418 and 4.245, which are used in forecasting. Finally, Figure 13.28 shows several autocorrelations of the differences, only one of which is (barely) significant. A runs test for the differences, not shown here, has a large p-value, which supports the conclusion that the differences are random.

Figure 13.28
Differences of Closing Prices

	A	B	C	D	E	F
1	Month	Closing Price	Diff1(Closing Price)			Diff1(Closing Price)
2	Jan-01	22.595			*One Variable Summary*	Data Set #1
3	Feb-01	22.134	-0.461		Mean	0.418
4	Mar-01	24.655	2.521		Std. Dev.	4.245
5	Apr-01	26.649	1.994		Count	75
6	May-01	26.303	-0.346			
7	Jun-01	27.787	1.484			Diff1(Closing Price)
8	Jul-01	32.705	4.918		*Autocorrelation Table*	Data Set #1
9	Aug-01	29.745	-2.96		Number of Values	75
10	Sep-01	26.741	-3.004		Standard Error	0.1155
11	Oct-01	24.852	-1.889		Lag #1	**-0.2435**
12	Nov-01	28.050	3.198		Lag #2	0.1348
13	Dec-01	27.847	-0.203		Lag #3	-0.0049
14	Jan-02	30.040	2.193		Lag #4	-0.0507
15	Feb-02	29.680	-0.36		Lag #5	0.0696
16	Mar-02	30.139	0.459		Lag #6	0.0009
17	Apr-02	29.276	-0.863		Lag #7	-0.0630
18	May-02	29.703	0.427		Lag #8	-0.0295
19	Jun-02	30.017	0.314		Lag #9	0.0496
20	Jul-02	29.687	-0.33		Lag #10	-0.1728
21	Aug-02	31.765	2.078		Lag #11	-0.0334
22	Sep-02	33.788	2.023		Lag #12	-0.0554
23	Oct-02	30.942	-2.846			
24	Nov-02	38.526	7.584			
25	Dec-02	34.099	-4.427			

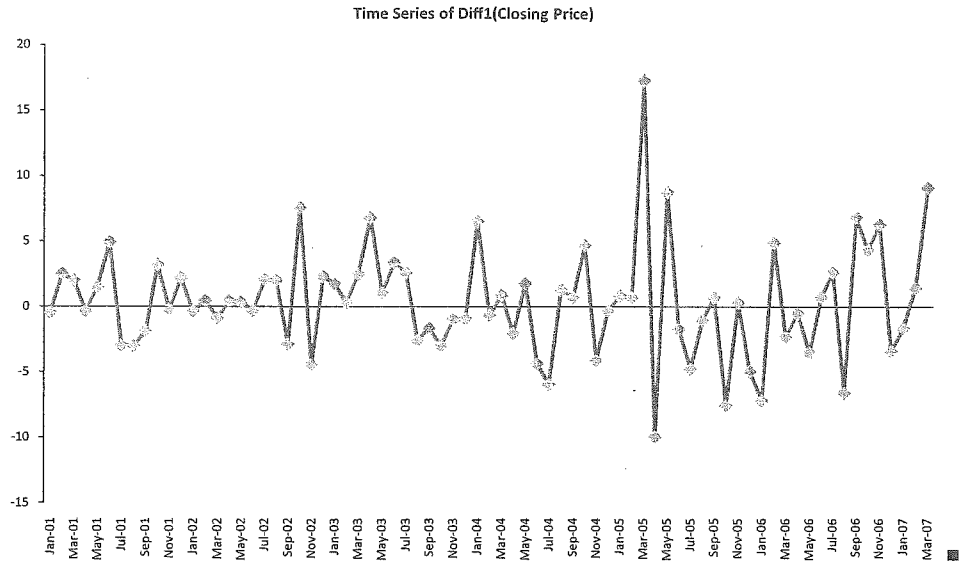
The plot of the differences appears in Figure 13.29. A visual inspection of the plot also supports the conclusion of random differences, although these differences do not vary around a mean of 0. Rather, they vary around a mean of 0.418. This positive value measures the upward trend—the closing prices increase, on average, by 0.418 per month. Finally, the variability in this figure is fairly constant (except for the two wide swings in 2005). Specifically, the zigzags do not tend to get appreciably wider through time. Therefore, we can conclude that the random walk model with an upward drift fits these data quite well.

To forecast future closing prices, we add the number of months ahead being forecasted times the mean difference to the final closing price (53.947 in April 2007). For example, a forecast of the closing price for September 2007 is as follows:

$$\text{Forecasted Closing Price for 9/07} = 53.947 + 0.418(5) = 56.037$$

As a rough measure of the accuracy of this forecast, we can use the standard deviation of the differences, 4.245. Specifically, it can be shown that the standard error for forecasting k periods ahead is the standard deviation of the differences multiplied by the square root of k. In this case, the standard error is 9.492. As usual, we can be 95% confident that the actual closing price in September will be no more than 2 standard errors from the forecast. Unfortunately, this results in a wide interval—from about 37 to 75. This reflects the fact that it is very difficult to make accurate forecasts for a series with this much variability.

Figure 13.29
Time Series Graph of Differences



PROBLEMS

Level A

16. The file **P13_16.xlsx** contains the daily closing prices of American Express stock for a 1-year period.
- **a.** Use the random walk model to forecast the closing price of this stock on the next trading day.
- **b.** We can be about 95% certain that the forecast made in part **a** is off by no more than how many dollars?

17. The closing value of the AMEX Airline Index for each trading day during a 1-year period, is given in the file **P13_17.xlsx**.
- **a.** Use the random walk model to forecast the closing price of this stock on the next trading day.
- **b.** We can be about 68% certain that the forecast made in part **a** is off by no more than how many dollars?

18. The file **P13_18.xlsx** contains the daily closing prices of ChevronTexaco stock for a 1-year period.
- **a.** Use the random walk model to forecast the closing price of this stock on the next trading day.
- **b.** We can be about 99.7% certain that the forecast made in part **a** is off by no more than how many dollars?

19. The closing value of the Dow Jones Industrial Average for each trading day for a 1-year period is provided in the file **P13_19.xlsx**.
- **a.** Use the random walk model to forecast the closing price of this index on the next trading day.
- **b.** Would it be wise to use the random walk model to forecast the closing price of this index for a trading day approximately *one month* after the next trading day? Explain why or why not.

20. Continuing the previous problem, consider the differences between consecutive closing values of the Dow Jones Industrial Average for the given set of trading days. Do these differences form a random series? Demonstrate why or why not.

21. The closing price of a share of JPMorgan's stock for each trading day during a 1-year period is recorded in the file **P13_21.xlsx**.
- **a.** Use the random walk model to forecast the closing price of this stock on the next trading day.
- **b.** We can be about 68% certain that the forecast made in part **a** is off by no more than how many dollars?

22. The purpose of this problem is to get you used to the concept of autocorrelation in a time series. You could do it with any time series, but here you should use the series of Wal-Mart daily stock prices in the file **P13_10.xlsx**.
- **a.** First, do it the "easy" way. Use the Autocorrelation procedure in StatTools to get a list of autocorrelations and a corresponding correlogram of the closing prices. You can choose the number of lags.
- **b.** Now do it the "hard" way. Create columns of lagged versions of the Close variable—3 or 4 lags will suffice. Next, look at scatterplots of Close versus its first few lags. If the autocorrelations are large, you should see fairly tight scatters—that's

what autocorrelation is all about. Also, generate a correlation matrix to see the correlations between Close and its first few lags. These should be approximately the same as the autocorrelations from part a. (Autocorrelations are calculated slightly differently than regular correlations, which accounts for any slight discrepancies you might notice, but these discrepancies should be minor.)

c. Create the first differences of Close in a new column. (You can do this manually with formulas, or you can use StatTools's Difference procedure under Data Utilities.) Now repeat parts a and b with the differences instead of the original closing prices—that is, examine the autocorrelations of the differences. They should be small, and the scatterplots of the differences versus lags of the differences should be "swarms." This illustrates what happens when the differences of a time series variable have "insignificant" autocorrelations.

d. Write a short report of your findings.

Level B

23. Consider a random walk model with the following equation: $Y_t = Y_{t-1} + 500 + \epsilon_t$, where ϵ_t is a normally distributed random series with mean 0 and standard deviation 10.
    a. Use Excel to simulate a time series that behaves according to this random walk model.
    b. Use the time series you constructed in part a to forecast the next observation.

24. The file **P13_24.xlsx** contains the daily closing prices of Procter & Gamble stock for a one-year period. Use only the data from 2003 to estimate the trend component of the random walk model. Next, use the estimated random walk model to forecast the behavior of the time series for the 2004 dates in the series. Comment on the accuracy of the generated forecasts over this period. How could you improve the forecasts as you progress through the 2004 trading days?

13.6 AUTOREGRESSION MODELS[3]

We now discuss a regression-based extrapolation method that regresses the current value of the time series on past (lagged) values. This is called **autoregression**, where the "auto" means that the explanatory variables in the equation are lagged values of the dependent variable, so that we are regressing the dependent variable on lagged versions of itself. This procedure is fairly straightforward on a spreadsheet. We first create lags of the dependent variable and then use a regression procedure to regress the original column on the lagged columns. Some trial and error is generally required to see how many lags are useful in the regression equation. The following example illustrates the procedure.

EXAMPLE 13.5 FORECASTING HAMMER SALES

A retailer has recorded its weekly sales of hammers (units purchased) for the past 42 weeks. (See the file **Hammer Sales.xlsx**.) A graph of this time series appears in Figure 13.30. It reveals a "meandering" behavior. The values begin high and stay high awhile, then get lower and stay lower awhile, then get higher again. (This behavior could be caused by any number of things, including the weather, increases and decreases in building projects, and possibly others.) How useful is autoregression for modeling these data and how can it be used for forecasting?

[3]This section can be omitted without any loss of continuity.

Objective To use autoregression, with an appropriate number of lagged terms, to forecast hammer sales.

Solution

It is generally best to begin with plenty of lags and then delete the higher numbered lags that aren't necessary.

A good place to start is with the autocorrelations of the series. These indicate whether the Sales variable is linearly related to any of its lags. The first six autocorrelations are shown in Figure 13.31. The first three of them are significantly positive, and then they decrease. Based on this information, we create three lags of Sales and run a regression of Sales versus these three lags. The output from this regression appears in Figure 13.32. We see that R^2 is fairly high, about 57%, and that s_e is about 15.7. However, the p-values for lags 2 and 3 are both quite large. It appears that once the first lag is included in the regression equation, the other two are not really needed.

Figure 13.30 Time Series Graph of Sales of Hammers

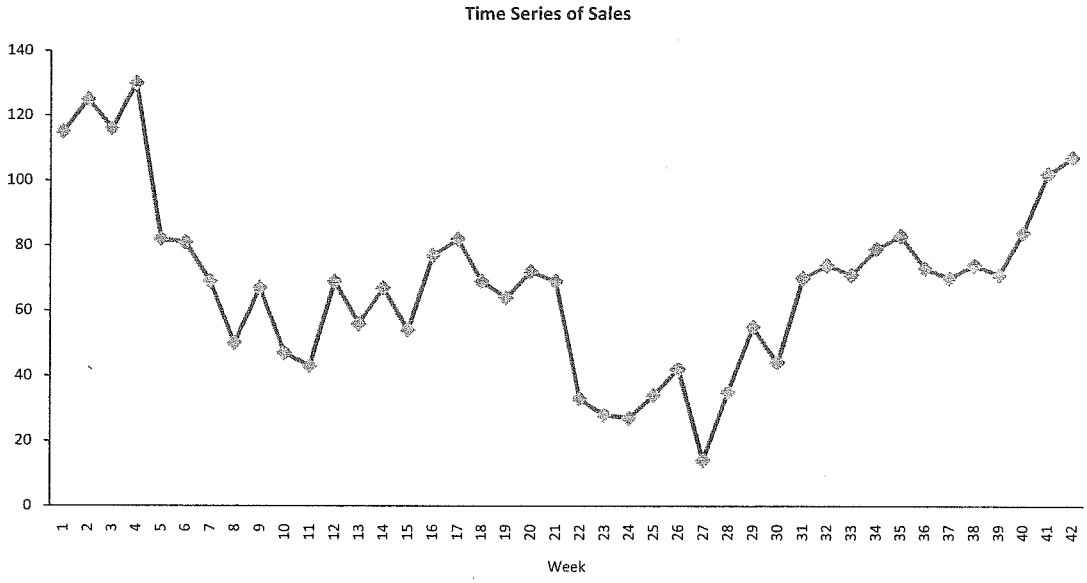


Figure 13.31 Autocorrelations for Hammer Sales Data

	A	B
27		Sales
28	*Autocorrelation Table*	Data Set #1
29	Number of Values	42
30	Standard Error	0.1543
31	Lag #1	**0.7523**
32	Lag #2	**0.5780**
33	Lag #3	**0.4328**
34	Lag #4	0.2042
35	Lag #5	0.1093
36	Lag #6	-0.0502

Figure 13.32 Autoregression Output with Three Lagged Variables

	A	B	C	D	E	F	G
7		Multiple		Adjusted	StErr of		
8	*Summary*	R	R-Square	R-Square	Estimate		
9		0.7573	0.5736	0.5370	15.7202		
10							
11		Degrees of	Sum of	Mean of			
12	*ANOVA Table*	Freedom	Squares	Squares	F-Ratio	p-Value	
13	Explained	3	11634.19978	3878.066594	15.6927	< 0.0001	
14	Unexplained	35	8649.38996	247.1254274			
15							
16			Standard			Confidence Interval 95%	
17	*Regression Table*	Coefficient	Error	t-Value	p-Value	Lower	Upper
18	Constant	15.4986	7.8820	1.9663	0.0572	-0.5027	31.5000
19	Lag1(Sales)	0.6398	0.1712	3.7364	0.0007	0.2922	0.9874
20	Lag2(Sales)	0.1523	0.1987	0.7665	0.4485	-0.2510	0.5556
21	Lag3(Sales)	-0.0354	0.1641	-0.2159	0.8303	-0.3686	0.2977

The two curves in this figure look pretty close to one another. However, a comparison of the vertical distances between pairs of points indicates that they are not so close after all.

Therefore, we reran the regression with only the first lag included. (Actually, we first omitted only the third lag. But the resulting output showed that the second lag was still insignificant.) The regression output with only the first lag included appears in Figure 13.33. In addition, a graph of the dependent and fitted variables, that is, the original Sales variable and its forecasts, appears in Figure 13.34. (This latter graph was formed from the Week, Sales, and Fitted columns.) The estimated regression equation is

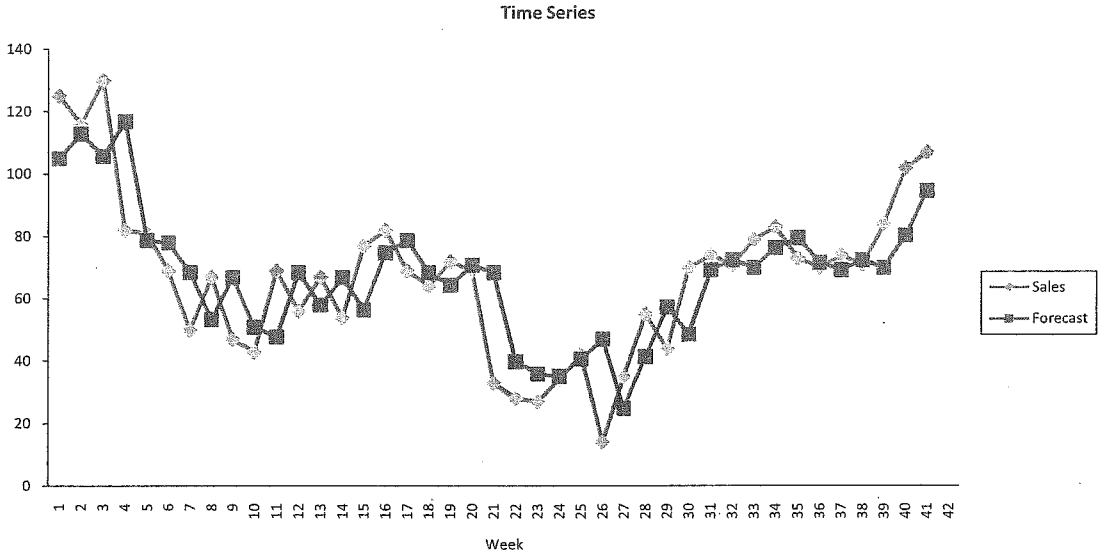
$$\text{Forecasted Sales}_t = 13.763 + 0.793\text{Sales}_{t-1}$$

The associated R^2 and s_e values are approximately 65% and 15.4. The R^2 value is a measure of the reasonably good fit we see in Figure 13.34, whereas s_e is a measure of the likely forecast error for short-term forecasts. It implies that a short-term forecast could easily be off by as much as 2 standard errors, or about 31 hammers.

Figure 13.33 Autoregression Output with a Single Lagged Variable

	A	B	C	D	E	F	G
7		Multiple		Adjusted	StErr of		
8	*Summary*	R	R-Square	R-Square	Estimate		
9		0.8036	0.6458	0.6367	15.4476		
10							
11		Degrees of	Sum of	Mean of			
12	*ANOVA Table*	Freedom	Squares	Squares	F-Ratio	p-Value	
13	Explained	1	16969.97657	16969.97657	71.1146	< 0.0001	
14	Unexplained	39	9306.511237	238.6284932			
15							
16			Standard			Confidence Interval 95%	
17	*Regression Table*	Coefficient	Error	t-Value	p-Value	Lower	Upper
18	Constant	13.7634	6.7906	2.0268	0.0496	0.0281	27.4988
19	Lag1(Sales)	0.7932	0.0941	8.4329	< 0.0001	0.6029	0.9834

Figure 13.34 Forecasts from Autoregression



To forecast, substitute known values of Y into the regression equation if they are available. Otherwise, substitute forecasted values.

To use the regression equation for forecasting *future* sales values, we substitute known or forecasted sales values in the right-hand side of the equation. Specifically, the forecast for week 43, the first week after the data period, is

$$\text{Forecasted Sales}_{43} = 13.763 + 0.793\text{Sales}_{42} = 13.763 + 0.793(107) \simeq 98.6$$

Here we use the *known* value of sales in week 42. However, the forecast for week 44 requires the *forecasted* value of sales in week 43:

$$\text{Forecasted Sales}_{44} = 13.763 + 0.793\text{Forecasted Sales}_{43}$$

$$\text{Forecasted Sales}_{44} = 13.763 + 0.793(98.6) \simeq 92.0$$

Perhaps these two forecasts of future sales values are on the mark, and perhaps they are not. The only way we will know for certain is by observing future sales values. However, it is interesting that in spite of the *upward* movement in the series in the last 3 weeks, the forecasts for weeks 43 and 44 are for *downward* movements. This is a combination of two properties of the regression equation. First, the coefficient of Sales_{t-1}, 0.793, is positive. Therefore, the equation forecasts that large sales will be followed by large sales, that is, positive autocorrelation. Second, however, this coefficient is less than 1, and this provides a dampening effect. The equation forecasts that a large will follow a large, but not *that* large. ■

Sometimes an autoregression model can be virtually equivalent to another forecasting model. As an example, suppose we find that the following equation adequately models a time series variable Y:

$$Y_t = 75.65 + 0.976Y_{t-1}$$

The coefficient of the lagged term, 0.976, is nearly equal to 1. If this coefficient were 1, we could subtract the lagged term from both sides of the equation and write that the *difference* series is a constant—that is, we would have a random walk model. As you can see, a random walk model is a special case of an autoregression model. However, autoregression models are much more general. Unfortunately, a more thorough study of them would take us into the realm of econometrics, which is well beyond the level of this book.

PROBLEMS

Level A

25. Consider the Consumer Price Index (CPI), which provides the annual percentage change in consumer prices. The data are in the file **P02_26.xlsx**.

a. Compute the first six autocorrelations of this time series.

b. Use the results of part **a** to specify one or more "promising" autoregression models. Estimate each model with the available data. Which model provides the best fit to the given data?

c. Use the best autoregression model from part **b** to produce a forecast of the CPI in the next year. Also, provide a measure of the likely forecast error.

26. The Consumer Confidence Index (CCI) attempts to measure people's feelings about general business conditions, employment opportunities, and their own income prospects. The file **P02_28.xlsx** contains the annual average values of the CCI.

a. Compute the first six autocorrelations of this time series.

b. Use the results of part **a** to specify one or more "promising" autoregression models. Estimate each model with the available data. Which model provides the best fit to the given data?

c. Use the best autoregression model from part **b** to produce a forecast of the CCI in the next year. Also, provide a measure of the likely forecast error.

27. Consider the proportion of Americans under the age of 18 living below the poverty level. The data are in the file **P02_29.xlsx**.

a. Compute the first six autocorrelations of this time series.

b. Use the results of part **a** to specify one or more "promising" autoregression models. Estimate each model with the available data. Which model provides the best fit to the given data?

c. Use the best autoregression model from part **b** to produce a forecast of the proportion of American children living below the poverty level in the next year. Also, provide a measure of the likely forecast error.

28. Examine the trend in the annual average values of the discount rate. The data are in the file **P02_30.xlsx**.

a. Specify one or more "promising" autoregression models based on autocorrelations of this time series. Estimate each model with the available data. Which model provides the best fit to given data?

b. Use the best autoregression model from part **a** to produce forecasts of the discount rate in the next 2 years.

29. The file **P02_34.xlsx** contains time series data on the percentage of the resident population in the United States who completed four or more years of college.

a. Specify one or more "promising" autoregression models based on autocorrelations of this time series. Estimate each model with the available data. Which model provides the best fit to the given data?

b. Use the best autoregression model from part **a** to produce forecasts of higher education attainment (i.e., completion of four or more years of college) in the United States in the next 3 years.

30. Consider the average annual interest rates on 30-year fixed mortgages in the United States. The data are recorded in the file **P02_35.xlsx**.

a. Specify one or more "promising" autoregression models based on autocorrelations of this time series. Estimate each model with the available data. Which model provides the best fit to the given data?

b. Use the best autoregression model from part **a** to produce forecasts of the average annual interest rates on 30-year fixed mortgages in the next 3 years.

31. The file **P13_31.xlsx** lists the monthly unemployment rates for several years. A common way to forecast time series is by using regression with lagged variables.

a. Predict future monthly unemployment rates using some combination of the unemployment rates for the last 4 months. For example, you might use last month's unemployment rate and the unemployment rate from 3 months ago as explanatory variables. Make sure all variables that you finally decide to keep in your equation are significant at the 0.15 level.
b. Do the residuals in your equation exhibit any autocorrelation?
c. Predict the next month's unemployment rate.
d. There is a 5% chance that the next month's unemployment rate will be less than what value?
e. What is the probability the next month's unemployment rate will be less than 6%?

Level B

32. The unit sales of a new drug for the first 25 months after its introduction to the marketplace are recorded in the file **P13_15.xlsx**. Specify one or more "promising" autoregression models based on autocorrelations of this time series. Estimate each model with the available data. Which model provides the best fit to the given data? Use the best autoregression model you found to forecast the sales of this new drug in the 26th month.

33. The file **P13_02.xlsx** contains the weekly sales at a local bookstore for each of the past 25 weeks.
a. Specify one or more "promising" autoregression models based on autocorrelations of this time series. Estimate each model with the available data. Which model provides the best fit to the given data?
b. What general result emerges from your analysis in part a? In other words, what is the most appropriate autoregression model for any given *random* time series?
c. Use the best autoregression model from part a to produce forecasts of the weekly sales at this bookstore for the next 3 weeks.

34. The file **P13_24.xlsx** contains the daily closing prices of Procter & Gamble stock for a one-year period.
a. Use only the data from 2003 to estimate an appropriate autoregression model.
b. Next, use the estimated autoregression model from part a to forecast the behavior of this time series for the 2004 dates of the series. Comment on the accuracy of the generated forecasts over this period.
c. How well does the autoregression model perform in comparison to the random walk model with respect to the accuracy of these forecasts? Explain any significant differences between the forecasting abilities of the two models.

13.7 MOVING AVERAGES

Perhaps the simplest and one of the most frequently used extrapolation methods is the method of **moving averages**. To implement the moving averages method, we first choose a **span**, the number of terms in each moving average. Let's say the data are monthly and we choose a span of 6 months. Then the forecast of next month's value is the average of the values of the last 6 months. For example, we average January to June to forecast July, we average February to July to forecast August, and so on. This procedure is the reason for the term *moving* averages.

> A **moving average** is the average of the observations in the past few periods, where the number of terms in the average is the **span**.

A moving averages model with a span of 1 is a random walk model with a mean trend of 0.

The role of the span is important. If the span is large—say, 12 months—then many observations go into each average, and extreme values have relatively little effect on the forecasts. The resulting series of forecasts will be much smoother than the original series. (For this reason, the moving average method is called a *smoothing* method.) In contrast, if the span is small—say, 3 months—then extreme observations have a larger effect on the forecasts, and the forecast series will be much less smooth. In the extreme, if the span is 1, there is no smoothing effect at all. The method simply forecasts next month's value to be the same as the current month's value. This is often called the **naive** forecasting model. It is a special case of the random walk model we discussed previously, with the mean difference equal to 0.

What span should we use? This requires some judgment. If we believe the ups and downs in the series are random noise, then we don't want future forecasts to react too

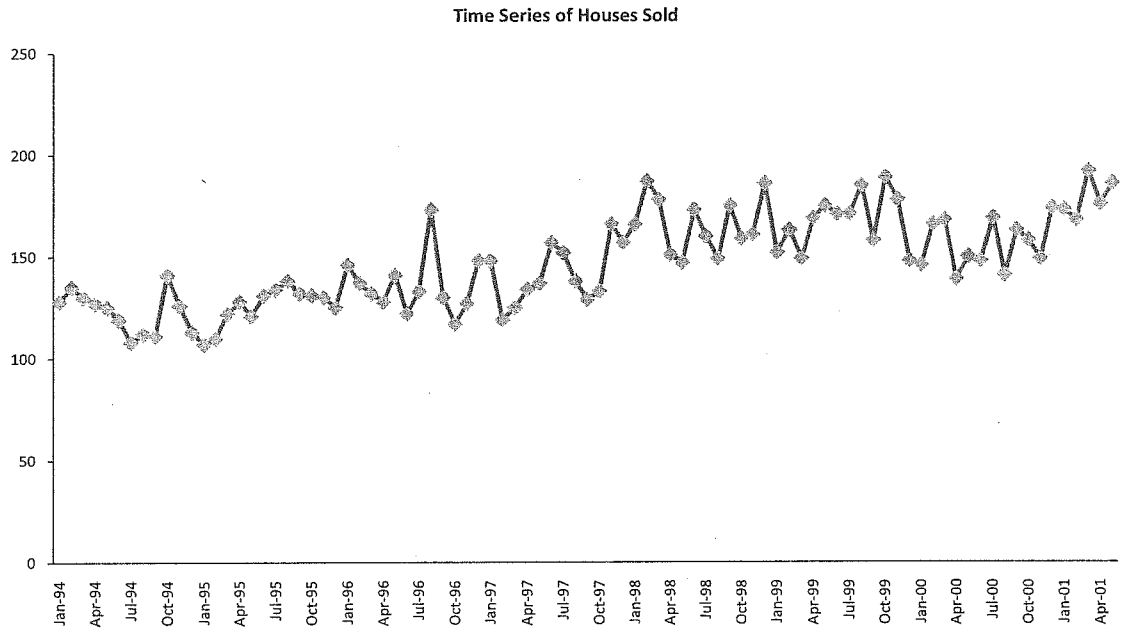
quickly to these ups and downs, and we should use a relatively large span. But if we want to track every little zigzag—under the belief that each up or down is predictable—then we should use a smaller span. We shouldn't be fooled, however, by a plot of the (smoothed) forecast series superimposed on the original series. This graph will almost always look better when a small span is used, because the forecast series will appear to track the original series better. Does this mean it will always provide better future forecasts? Not necessarily. There is little point in tracking random ups and downs closely if they represent unpredictable noise.

The following example illustrates the use of moving averages.

EXAMPLE 13.6 HOUSES SOLD IN THE MIDWEST

The file **House Sales.xlsx** contains monthly data on the number of houses sold in the Midwest (in thousands) from January 1994 through May 2001. (These data are seasonally adjusted.)[4] A time series graph of the data appears in Figure 13.35. Does a moving averages model fit this data set well? What span should be used?

Figure 13.35 Time Series Plot of Monthly House Sales



Objective To see whether a moving averages model with an appropriate span fits the housing sales data, and to see how StatTools implements this method.

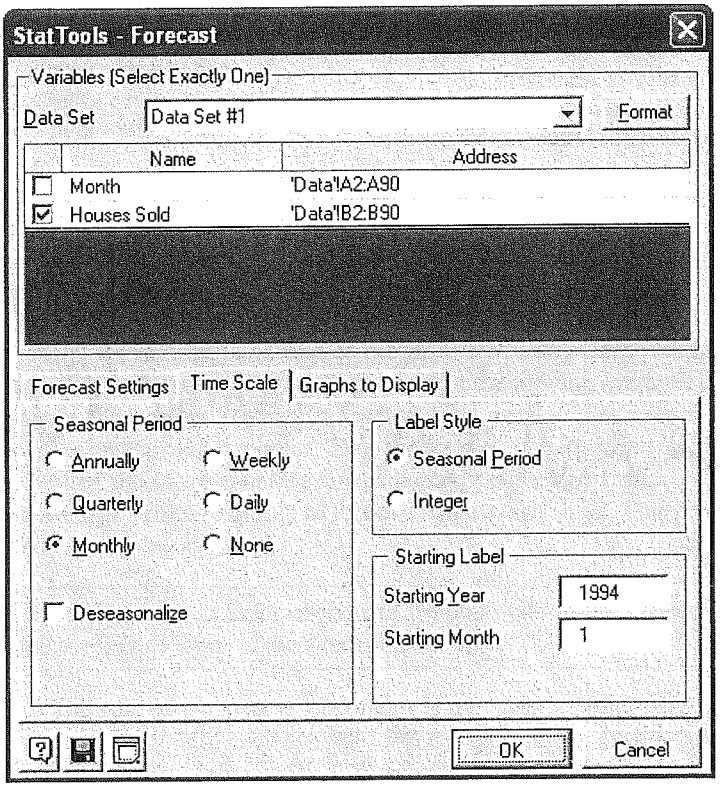
[4]We discuss seasonal adjustment in Section 13.9. Government data are often reported in seasonally adjusted form, with the seasonality removed, to make any trends more apparent.

Solution

Although the moving averages method is quite easy to implement in Excel—we just form an average of the appropriate span and copy it down—it can be tedious. Therefore, we call on the forecasting procedure of StatTools. Actually, this procedure is fairly general in that it allows us to forecast with several methods, either with or without taking seasonality into account. Because this is our first exposure to this procedure, we go through it in some detail in this example. In later examples, we mention some of its other capabilities.

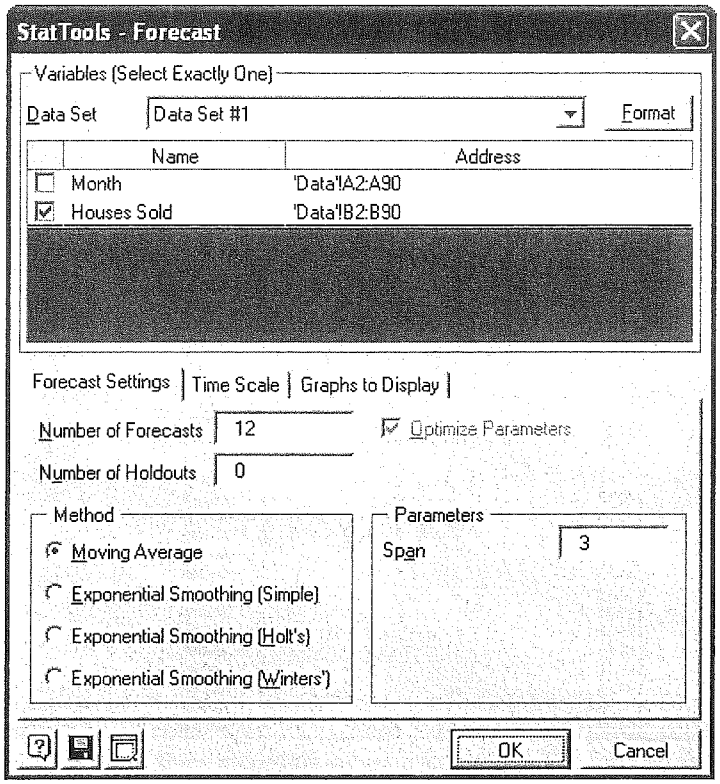
To use the StatTools Forecasting procedure, select Forecast from the StatTools Time Series and Forecasting dropdown. This brings up the dialog box in Figure 13.36, which has three tabs in its bottom section. The Time Scale tab, shown in Figure 13.36, allows us to select the time period. The Forecast Settings tab, shown in Figure 13.37, allows us to select a forecasting method. Finally, the Graphs to Display tab, not shown here, allows us to select several optional time series graphs. For now, fill out the dialog box sections as shown and select the Forecast Overlay option in the Graphs to Display tab. In particular, note from Figure 13.37 that we are using the moving averages method with a span of 3, and we are asking for forecasts of the next 12 months.

Figure 13.36
Forecast Dialog Box with Time Scale Tab Visible



Another option in Figure 13.37 is that we can elect to "hold out" a subset of the data for validation purposes. If we hold out several periods at the end of the data set for validation, then any model that is built is estimated only for the nonholdout observations, and summary measures are reported for the nonholdout and holdout subsets separately. For now, we are not using a holdout period.

Figure 13.37
Forecast Dialog Box with Forecast Settings Tab Visible



The output consists of several parts, as shown in Figures 13.38 through 13.41. We actually ran the analysis twice, once for a span of 3 and once for a span of 12. These figures show the comparison. (We also obtained output for a span of 6, with results similar to those for a span of 12.) First, the summary measures MAE, RMSE, and MAPE of the forecast errors are shown in Figure 13.38. As we see, both spans produce similar magnitudes of forecast errors. For example, they are both off, on average, by slightly more than 8%.

Figure 13.38 Moving Averages Summary Output

	A	B	C	D	E	F	G	H
8	*Forecasting Constant*						*Forecasting Constant*	
9	Span	3					Span	12
10								
11					Moving averages method with spans 3 and 12			
12	*Moving Averages*						*Moving Averages*	
13	Mean Abs Err	12.26					Mean Abs Err	12.43
14	Root Mean Sq Err	14.96					Root Mean Sq Err	15.62
15	Mean Abs Per% Err	8.37%					Mean Abs Per% Err	8.07%

Figure 13.39 Moving Averages Detailed Output

	A	B	C	D	E	F	G	H	I	J
40	*Forecasting Data*	Houses Sold	Forecast	Error			*Forecasting Data*	Houses Sold	Forecast	Error
41	Jan-1994	128.0000					Jan-1994	128.0000		
42	Feb-1994	135.0000					Feb-1994	135.0000		
43	Mar-1994	130.0000					Mar-1994	130.0000		
44	Apr-1994	127.0000	131.00	-4.00			Apr-1994	127.0000		
45	May-1994	125.0000	130.67	-5.67			May-1994	125.0000		
46	Jun-1994	119.0000	127.33	-8.33			Jun-1994	119.0000		
47	Jul-1994	108.0000	123.67	-15.67			Jul-1994	108.0000		
48	Aug-1994	112.0000	117.33	-5.33			Aug-1994	112.0000		
49	Sep-1994	111.0000	113.00	-2.00			Sep-1994	111.0000		
50	Oct-1994	141.0000	110.33	30.67			Oct-1994	141.0000		
51	Nov-1994	126.0000	121.33	4.67			Nov-1994	126.0000		
52	Dec-1994	113.0000	126.00	-13.00			Dec-1994	113.0000		
53	Jan-1995	107.0000	126.67	-19.67			Jan-1995	107.0000	122.92	-15.92
54	Feb-1995	110.0000	115.33	-5.33			Feb-1995	110.0000	121.17	-11.17
55	Mar-1995	122.0000	110.00	12.00			Mar-1995	122.0000	119.08	2.92
56	Apr-1995	128.0000	113.00	15.00			Apr-1995	128.0000	118.42	9.58
126	Feb-2001	168.0000	165.33	2.67			Feb-2001	168.0000	158.17	9.83
127	Mar-2001	192.0000	171.67	20.33			Mar-2001	192.0000	158.33	33.67
128	Apr-2001	176.0000	177.67	-1.67			Apr-2001	176.0000	160.33	15.67
129	May-2001	186.0000	178.67	7.33			May-2001	186.0000	163.42	22.58
130	Jun-2001		184.67				Jun-2001		166.42	
131	Jul-2001		182.22				Jul-2001		167.95	
132	Aug-2001		184.30				Aug-2001		167.86	
133	Sep-2001		183.73				Sep-2001		170.10	
134	Oct-2001		183.42				Oct-2001		170.69	
135	Nov-2001		183.81				Nov-2001		171.75	
136	Dec-2001		183.65				Dec-2001		173.65	
137	Jan-2002		183.63				Jan-2002		173.62	
138	Feb-2002		183.70				Feb-2002		173.67	
139	Mar-2002		183.66				Mar-2002		174.14	
140	Apr-2002		183.66				Apr-2002		172.66	
141	May-2002		183.67				May-2002		172.38	

Figure 13.40
Moving Averages Forecasts with Span 3

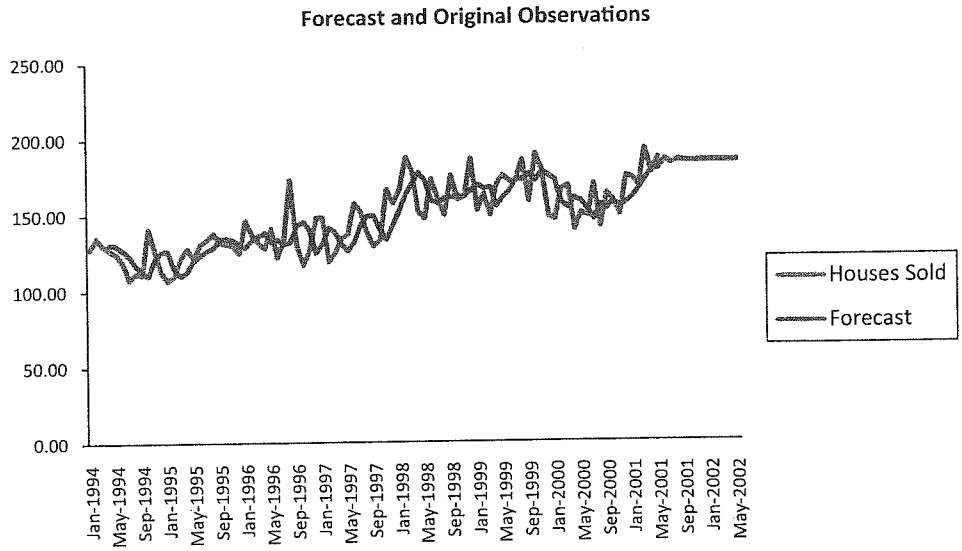
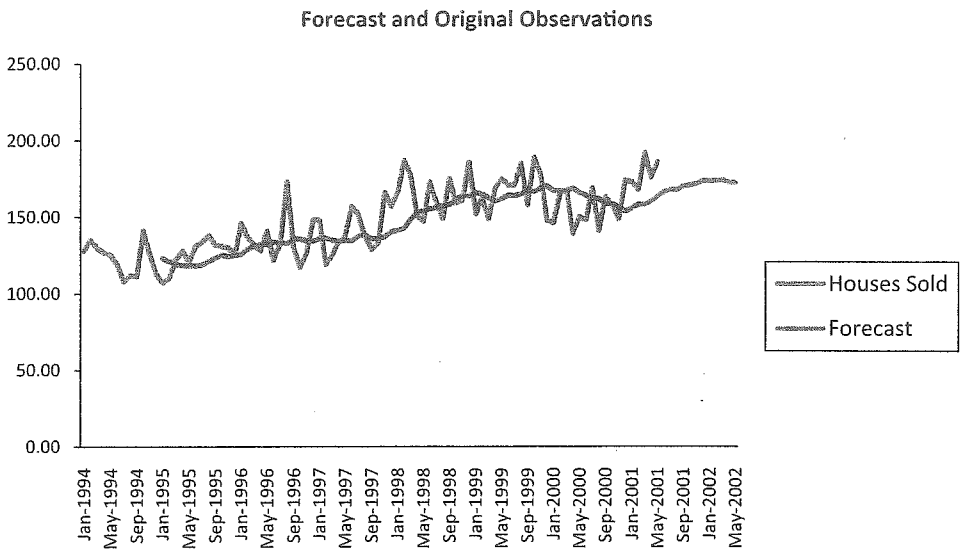


Figure 13.41
Moving Averages Forecasts with Span 12



The essence of the forecasting method is very simple and is captured in column C of Figure 13.39 (for a span of 3). Each value in the historical period in this column is an average of the three preceding values in column B. The forecast errors are then just the differences between columns B and C. For the future periods, the forecast formulas in column C use observations when they are available. If they are not available, previous forecasts are used. For example, the value in cell C131, the forecast for July 2001, is the average of the *observed* values in April and May and the *forecasted* value in June.

The graphs in Figures 13.40 and 13.41 show the behavior of the forecasts. The forecasted series with span 3 follows the ups and downs of the actual series fairly closely, whereas the forecasted series with span 12 is much smoother and doesn't react nearly as much to these ups and downs. Which of these is better? The error summary measures indicate that it is a virtual toss-up. The MAPE with span 12 is slightly lower, but the RMSE with span 3 is slightly lower. Note that the *future* forecasts are considerably lower with span 12 than with span 3.

At this point, how to proceed is up to the judgment of the forecaster, who presumably has some knowledge of the housing sales market in the Midwest. If she believes that the ups and downs in the original series are largely unpredictable noise, then she will probably trust the smooth forecasts from a span of 12. Otherwise, she might use a span of 3 (or some other intermediate span, such as 6). ■

The moving average method we have presented is the simplest of a group of moving average methods used by professional forecasters. We *smoothed* exactly once; that is, we took moving averages of several observations at a time and used these as forecasts. More complex methods smooth more than once, basically to get rid of random noise. They take moving averages, then moving averages of these moving averages, and so on for several stages. This can become quite complex, but the objective is quite simple—to smooth the data so that we can see underlying patterns.

PROBLEMS

Level A

35. The file **P13_16.xlsx** contains the daily closing prices of American Express stock for a 1-year period.
    a. Using a span of 3 days, forecast the price of this stock for the next trading day with the moving average method. How well does this method with span 3 forecast the known observations in this data set?
    b. Repeat part a with a span of 10.
    c. Which of these two spans appears to be more appropriate? Explain your choice.

36. The closing value of the AMEX Airline Index for each trading day during a 1-year period is given in the file **P13_17.xlsx**.
    a. How well does the moving average method track this series when the span is 4 days; when the span is 12 days?
    b. Using the more appropriate span, forecast the closing value of this index on the next trading day with the moving average method.

37. The closing value of the Dow Jones Industrial Average for each trading day during a 1-year period is provided in the file **P13_19.xlsx**.
    a. Using a span of 2 days, forecast the price of this index on the next trading day with the moving average method. How well does the moving average method with span 2 forecast the known observations in this data set?
    b. Repeat part a with a span of 5 days; with a span of 15 days.
    c. Which of these three spans appears to be most appropriate? Explain your choice.

38. The file **P13_10.xlsx** contains the daily closing prices of Wal-Mart stock during a 1-year period. Use the moving average method with a carefully chosen span to forecast this time series for the next 3 trading days. Defend your choice of the span used.

39. The Consumer Confidence Index (CCI) attempts to measure people's feelings about general business conditions, employment opportunities, and their own income prospects. The file **P02_28.xlsx** contains the annual average values of the CCI. Use the moving average method with a carefully chosen span to forecast this time series in the next 2 years. Defend your choice of the span used here.

Level B

40. Consider the file **P02_37.xlsx**, which contains total monthly U.S. retail sales data. While retaining the final 6 months of observations for validation purposes, use the method of moving averages with a carefully chosen span to forecast U.S. retail sales in the next year. Comment on the performance of your model. What makes this time series more challenging to forecast?

41. Consider a random walk model with the following equation: $Y_t = Y_{t-1} + \epsilon_t$, where ϵ_t is a random series with mean 0 and standard deviation 1. Specify a moving average model that is equivalent to this random walk model. In particular, what is the appropriate size of the span in the equivalent moving average model? Describe the smoothing effect of this span choice.

13.8 EXPONENTIAL SMOOTHING

There are two possible criticisms of the moving averages method. First, it puts equal weight on each value in a typical moving average when making a forecast. Many people would argue that if next month's forecast is to be based on the previous 12 months' observations, then more weight ought to be placed on the more recent observations. The second criticism is that the moving averages method requires a lot of data storage. This is particularly true for companies that routinely make forecasts of hundreds or even thousands of items. If 12-month moving averages are used for 1000 items, then 12,000 values are needed for next month's forecasts. This may or may not be a concern considering today's inexpensive computer storage capabilities.

Exponential smoothing is a method that addresses both of these criticisms. It bases its forecasts on a weighted average of past observations, with more weight put on the more recent observations, and it requires very little data storage. In addition, it is not difficult for most business people to understand, at least conceptually. Therefore, this method finds widespread use in the business world, particularly when frequent and automatic forecasts of many items are required.

There are many versions of exponential smoothing. The simplest is appropriately called *simple* exponential smoothing. It is relevant when there is no pronounced trend or seasonality in the series. If there is a trend but no seasonality, then *Holt's* method is applicable. If, in addition, there is seasonality, then *Winters'* method can be used. This does not exhaust the list of exponential smoothing models—researchers have invented many other variations—but these three models will suffice for us.

Exponential Smoothing Models

Simple exponential smoothing is appropriate for a series with no pronounced trend or seasonality. **Holt's** method is appropriate for a series with trend but no seasonality. **Winters'** method is appropriate for a series with seasonality (and possibly trend).

In this section we examine simple exponential smoothing and Holt's model for trend. Then in the next section we examine Winters' model when we focus on seasonal models in general.

13.8.1 Simple Exponential Smoothing

The level is where we think the series would be if it were not for random noise.

We now examine simple exponential smoothing in some detail. We first introduce two new terms. Every exponential model has at least one **smoothing constant**, which is always between 0 and 1. Simple exponential smoothing has a single smoothing constant denoted by α. (Its role is discussed shortly.) The second new term is L_t, called the **level** of the series at time t. This value is not observable but can only be estimated. Essentially, it is where we think the series would be at time t if there were no random noise. Then the simple exponential smoothing method is defined by the following two equations, where F_{t+k} is the forecast of Y_{t+k} made at time t:

Simple Exponential Smoothing Formulas

$$L_t = \alpha Y_t + (1 - \alpha)L_{t-1} \quad (13.12)$$

$$F_{t+k} = L_t \quad (13.13)$$

Even though you usually won't have to substitute into these equations manually, you should understand what they say. Equation (13.12) shows how to update the estimate of the level. It is a weighted average of the current observation, Y_t, and the previous level, L_{t-1}, with respective weights α and $1 - \alpha$. Equation (13.13) shows how forecasts are made. It says that the k-period-ahead forecast, F_{t+k}, made of Y_{t+k} in period t is the most recently estimated level, L_t. This is the *same* for any value of $k \geq 1$. The idea is that in simple exponential smoothing, we believe that the series is not really going anywhere. So as soon as we estimate where the series ought to be in period t (if it weren't for random noise), we forecast that this is where it will also be in any future period.

The smoothing constant α is analogous to the span in moving averages. There are two ways to see this. The first way is to rewrite equation (13.12), using the fact that the forecast error, E_t, made in forecasting Y_t at time $t - 1$ is $Y_t - F_t = Y_t - L_{t-1}$. A bit of algebra then gives equation (13.14).

Equivalent Formula for Simple Exponential Smoothing

$$L_t = L_{t-1} + \alpha E_t \quad (13.14)$$

This equation says that the next estimate of the level is adjusted from the previous estimate by adding a multiple of the most recent forecast error. This makes sense. If our previous forecast was too high, then E_t is negative, and we adjust the estimate of the level downward. The opposite is true if our previous forecast was too low. However, equation (13.14) says that we do not adjust by the entire magnitude of E_t, but only by a fraction of it. If α is small, say, $\alpha = 0.1$, then the adjustment is minor; if α is close to 1, the adjustment is large. So if we want to react quickly to movements in the series, we choose a large α; otherwise, we choose a small α.

Another way to see the effect of α is to substitute recursively into the equation for L_t. If you are willing to go through some algebra, you can verify that L_t satisfies equation (13.15), where the sum extends back to the first observation at time $t = 1$.

Another Equivalent Formula for Simple Exponential Smoothing

$$L_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \alpha(1-\alpha)^3 Y_{t-3} + \cdots \quad (13.15)$$

Equation (13.15) shows how the exponentially smoothed forecast is a weighted average of previous observations. Furthermore, because $1 - \alpha$ is less than 1, the weights on the Y's decrease from time t backward. Therefore, if α is close to 0, then $1 - \alpha$ is close to 1 and the weights decrease very slowly. In other words, observations from the distant past continue to have a large influence on the next forecast. This means that the graph of the forecasts will be relatively smooth, just as with a large span in the moving averages method. But when α is close to 1, the weights decrease rapidly, and only very recent observations have much influence on the next forecast. In this case forecasts react quickly to sudden changes in the series.

Small smoothing constants provide forecasts that respond slowly to changes in the data. Large smoothing constants do the opposite.

What value of α should we use? There is no universally accepted answer to this question. Some practitioners recommend always using a value around 0.1 or 0.2. Others recommend experimenting with different values of α until a measure such as RMSE or MAPE is minimized. Some packages even have an optimization feature to find this optimal value of α. (This is the case with StatTools.) But just as we discussed in the moving averages section, the value of α that tracks the historical series most closely does not necessarily guarantee the most accurate *future* forecasts.

EXAMPLE 13.6 HOUSES SOLD IN THE MIDWEST (CONTINUED)

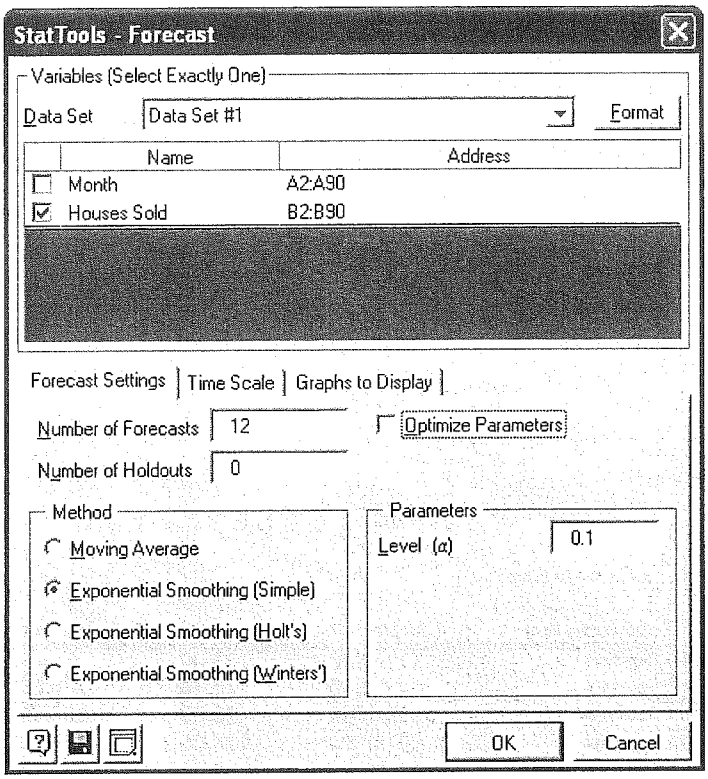
Previously, we used the moving averages method to forecast monthly housing sales in the Midwest. (See the **House Sales.xlsx** file.) How well does simple exponential smoothing work with this data set? What smoothing constant should we use?

Objective To see how well a simple exponential smoothing model, with an appropriate smoothing constant, fits the housing sales data, and to see how StatTools implements this method.

Solution

We use StatTools to implement the simple exponential smoothing model, specifically equations (13.12) and (13.13). We do this again with the Forecast item from StatTools Time Series and Forecasting dropdown. We then fill in the forecast dialog box essentially like we did with moving averages, except that we select the exponential smoothing options in the Forecast Settings tab (see Figure 13.42). That is, we select the simple exponential smoothing option, choose a smoothing constant (0.1 was chosen here, but any other value could be chosen), and elect not to optimize.

Figure 13.42
Forecast Settings for Exponential Smoothing



The results appear in Figures 13.43 and 13.44. The heart of the method takes place in columns C, D, and E of Figure 13.43. Column C calculates the smoothed levels (L_t) from equation (13.12), column D calculates the forecasts (F_t) from equation (13.13), and column E calculates the forecast errors (E_t) as the observed values minus the forecasts. Although we do not list the Excel formulas here, you can examine them in the StatTools output.

Every exponential smoothing method requires *initial* values, in this case the initial smoothed level in cell C41. There is no way to calculate this value, L_1, from equation (13.12) because the *previous* value, L_0, is unknown. Different implementations of exponential smoothing initialize in different ways. We have simply set L_1 equal to Y_1 (in cell B41). The effect of initializing in different ways is usually minimal because any effect of early data is usually washed out as we forecast into the future. In the present example, data from 1994 have little effect on forecasts of 2001 and beyond.

Note that the 12 future forecasts (rows 130 down) are all equal to the last calculated smoothed level, the one for May 2001 in cell C129. The fact that these remain constant is a consequence of the assumption behind *simple* exponential smoothing, namely, that the series is not really going anywhere. Therefore, the last smoothed level is the best indication of future values of the series we have.

In the next subsection we use Holt's method on this series to see whether it captures the trend better than simple exponential smoothing.

Figure 13.44 shows the forecast series superimposed on the original series. We see the obvious smoothing effect of a relatively small α level. The forecasts don't track the series very well, but if the various zigzags in the original series are really random noise, then perhaps we don't want the forecasts to track these random ups and downs too closely. Perhaps we instead prefer a forecast series that emphasizes the basic underlying pattern.

We see several summary measures of the forecast errors in Figure 13.43. The RMSE and MAE indicate that the forecasts from this model are typically off by a magnitude of about 12 to 15 thousand, and the MAPE indicates that they are off by about 7.9%. (These

are similar to the errors we obtained with moving averages.) These imply fairly sizable errors. One way to reduce the errors is to use a different smoothing method. We try this in the next subsection with Holt's method. Another way to reduce the errors is to use a different smoothing constant. There are two methods you can use. First, you can simply enter different values in the smoothing constant cell in the Forecast sheet. All formulas, including those for MAE, RMSE, and MAPE, will update automatically.

Figure 13.43
Simple Exponential Smoothing Output

	A	B	C	D	E
7					
8	*Forecasting Constant*				
9	Level (Alpha)	0.100			
10					
11					
12	*Simple Exponential*				
13	Mean Abs Err	11.93			
14	Root Mean Sq Err	15.08			
15	Mean Abs Per% Err	7.91%			
39					
40	*Forecasting Data*	Houses Sold	Level	Forecast	Error
41	Jan-1994	128.0000	128.00		
42	Feb-1994	135.0000	128.70	128.00	7.00
43	Mar-1994	130.0000	128.83	128.70	1.30
44	Apr-1994	127.0000	128.65	128.83	-1.83
123	Nov-2000	149.0000	157.82	158.80	-9.80
124	Dec-2000	174.0000	159.44	157.82	16.18
125	Jan-2001	173.0000	160.80	159.44	13.56
126	Feb-2001	168.0000	161.52	160.80	7.20
127	Mar-2001	192.0000	164.56	161.52	30.48
128	Apr-2001	176.0000	165.71	164.56	11.44
129	May-2001	186.0000	167.74	165.71	20.29
130	Jun-2001			167.74	
131	Jul-2001			167.74	
132	Aug-2001			167.74	
133	Sep-2001			167.74	
134	Oct-2001			167.74	
135	Nov-2001			167.74	
136	Dec-2001			167.74	
137	Jan-2002			167.74	
138	Feb-2002			167.74	
139	Mar-2002			167.74	
140	Apr-2002			167.74	
141	May-2002			167.74	

Second, you can check the Optimize Parameters option in Figure 13.42. This automatically runs an optimization algorithm (not Solver, by the way) to find the smoothing constant that minimizes RMSE. (StatTools is programmed to minimize RMSE. However, you could try minimizing MAPE, say, by using Excel's Solver add-in.) We did this for the housing data and obtained the forecasts in Figure 13.45 (from a smoothing constant of 0.295). The corresponding MAE, RMSE, and MAPE are 11.4, 14.1, and 7.7%, respectively—slightly better than before. This larger smoothing constant produces a less smooth forecast curve and slightly better error measures. However, there is no guarantee that *future* forecasts made with this optimal smoothing constant will be any better than with a smoothing constant of 0.1.

Figure 13.44
Graph of Forecasts from Simple Exponential Smoothing

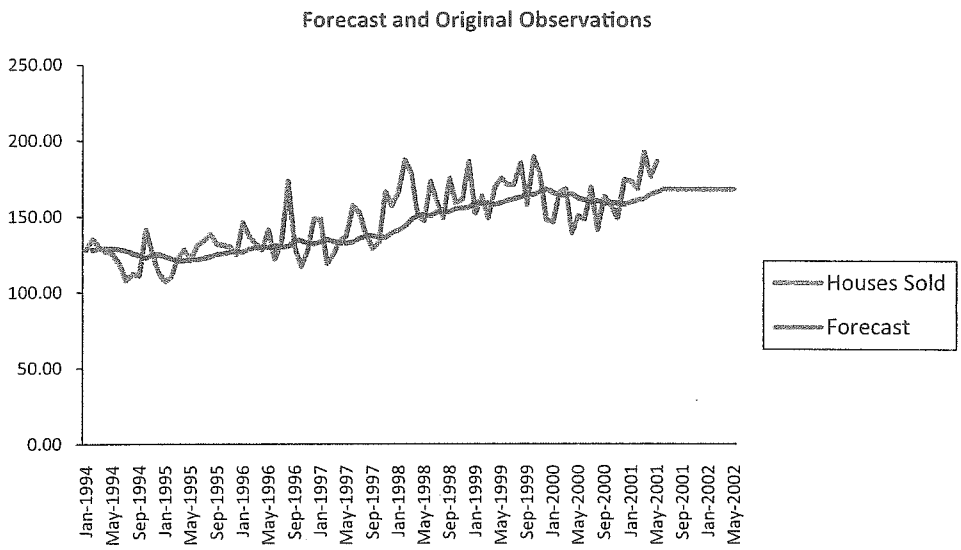
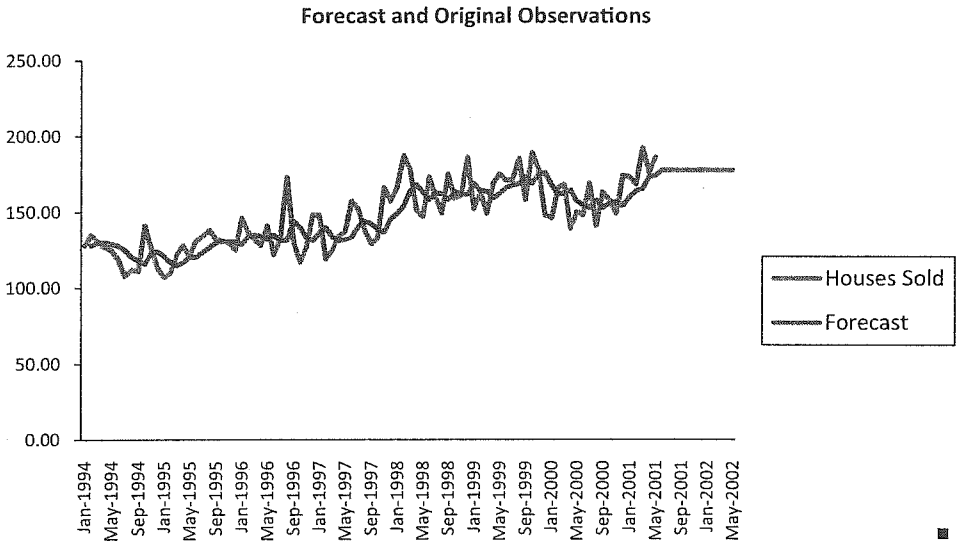


Figure 13.45
Graph of Forecasts with an Optimal Smoothing Constant



13.8.2 Holt's Model for Trend

The trend term in Holt's method estimates the change from one period to the next.

The simple exponential smoothing model generally works well if there is no obvious trend in the series. But if there is a trend, then this method consistently lags behind it. For example, if the series is constantly increasing, simple exponential smoothing forecasts will be consistently low. Holt's method rectifies this by dealing with trend explicitly. In addition to the level of the series, L_t, Holt's method includes a trend term, T_t, and a corresponding smoothing constant β. The interpretation of L_t is exactly as before. The interpretation of T_t is that it represents an estimate of the change in the series from one period to the next. The equations for Holt's model are as follows.

Formula's for Holt's Exponential Smoothing Method

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad (13.16)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad (13.17)$$

$$F_{t+k} = L_t + kT_t \quad (13.18)$$

These equations are not as bad as they look. (And don't forget that the computer typically does all of the calculations for you.) Equation (13.16) says that the updated level is a weighted average of the current observation and the previous level plus the estimated change. Equation (13.17) says that the updated trend term is a weighted average of the difference between two consecutive levels and the previous trend term. Finally, equation (13.18) says that the k-period-ahead forecast made in period t is the estimated level plus k times the estimated change per period.

Everything we said about α for simple exponential smoothing applies to both α and β in Holt's model. The new smoothing constant β controls how quickly the method reacts to perceived changes in the trend. If β is small, the method reacts slowly. If it is large, the method reacts more quickly. Of course, there are now two smoothing constants to select. Some practitioners suggest using a small value of α (0.1 to 0.2, say) and setting β equal to α. Others suggest using an optimization option (available in StatTools) to select the "best" smoothing constants. We illustrate the possibilities in the following continuation of the housing sales example.

EXAMPLE 13.6 HOUSES SOLD IN THE MIDWEST (CONTINUED)

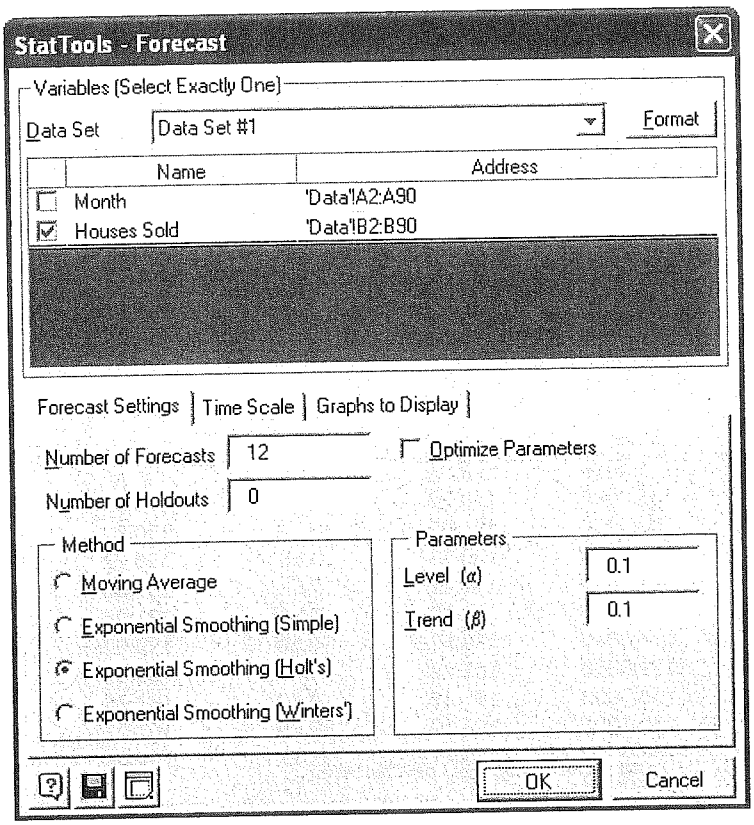
We again examine the monthly data on housing sales in the Midwest. In the previous subsection, we saw that simple exponential smoothing, even with an optimal smoothing constant, does only a "fair" job of forecasting housing sales. Given that there is an upward trend in housing sales over this period, we might expect Holt's method to perform better. Does it? What smoothing constants are appropriate?

Objective To see whether Holt's method, with appropriate smoothing constants, captures the trend in the housing sales data better than simple exponential smoothing (or moving averages).

Solution

We implement Holt's method in StatTools almost exactly like we did for simple exponential smoothing. The only difference is that we can now choose *two* smoothing constants, as shown in Figure 13.46. They can have different values, although we have chosen them to be their default values of 0.1.

Figure 13.46
Dialog Box for Holt's Method



The StatTools outputs in Figures 13.47 and 13.48 are also very similar to the simple exponential smoothing outputs. The only difference is that there is now a trend column, column D, in the numerical output. You can check that the formulas in columns C, D, and E implement equations (13.16), (13.17), and (13.18). As before, there is an initialization problem in row 42. These require values of L_1 and T_1 to get the method started. Different implementations of Holt's method obtain these initial values in slightly different ways, but the effect is fairly minimal in most cases. (You can check cells C42 and D42 to see how StatTools does it.)

Somewhat surprisingly, the error measures for this implementation of Holt's method are no better than for simple exponential smoothing, even though this series exhibits a gradual upward trend. Perhaps this is because 0.1 and 0.1 are not the *optimal* smoothing constants. Therefore, we ran Holt's method a second time, checking the Optimize Parameters option. This resulted in somewhat better results and the forecasts shown in Figure 13.49. The optimal smoothing constants are $\alpha = 0.251$ and $\beta = 0.000$, and the MAE, RMSE, and MAPE values were 11.2, 13.9, and 7.7%—almost identical to simple exponential smoothing with an optimal smoothing constant.[5]

[5]The fact that β is 0 does not mean there is no trend. It simply means that we never update our initial estimate of trend, which is positive.

Figure 13.47 Output from Holt's Method

	A	B	C	D	E	F
7						
8	*Forecasting Constants*					
9	Level (Alpha)	0.100				
10	Trend (Beta)	0.100				
11						
12						
13	*Holt's Exponential*					
14	Mean Abs Err	12.05				
15	Root Mean Sq Err	15.08				
16	Mean Abs Per% Err	8.23%				
17						
41	*Forecasting Data*	Houses Sold	Level	Trend	Forecast	Error
42	Jan-1994	128.0000	128.00	0.65		
43	Feb-1994	135.0000	129.29	0.72	128.65	6.35
44	Mar-1994	130.0000	130.00	0.72	130.00	0.00
45	Apr-1994	127.0000	130.35	0.68	130.72	-3.72
46	May-1994	125.0000	130.42	0.62	131.02	-6.02
47	Jun-1994	119.0000	129.83	0.50	131.04	-12.04
125	Dec-2000	174.0000	160.52	-0.60	159.02	14.98
126	Jan-2001	173.0000	161.22	-0.47	159.91	13.09
127	Feb-2001	168.0000	161.47	-0.40	160.75	7.25
128	Mar-2001	192.0000	164.17	-0.09	161.07	30.93
129	Apr-2001	176.0000	165.27	0.03	164.08	11.92
130	May-2001	186.0000	167.37	0.23	165.30	20.70
131	Jun-2001				167.60	
132	Jul-2001				167.84	
133	Aug-2001				168.07	
134	Sep-2001				168.31	
135	Oct-2001				168.54	
136	Nov-2001				168.78	
137	Dec-2001				169.01	
138	Jan-2002				169.25	
139	Feb-2002				169.48	
140	Mar-2002				169.72	
141	Apr-2002				169.95	
142	May-2002				170.19	

You should not conclude from this example that Holt's method is never superior to simple exponential smoothing. Holt's method is often able to react quickly to a sudden upswing or downswing in the data, whereas simple exponential smoothing typically has a delayed reaction to such a change. It just happened in this example that the trend was gradual, so that both methods were able to react to it in equivalent ways.

Figure 13.48
Forecasts from Holt's Method with Nonoptimal Smoothing Constants

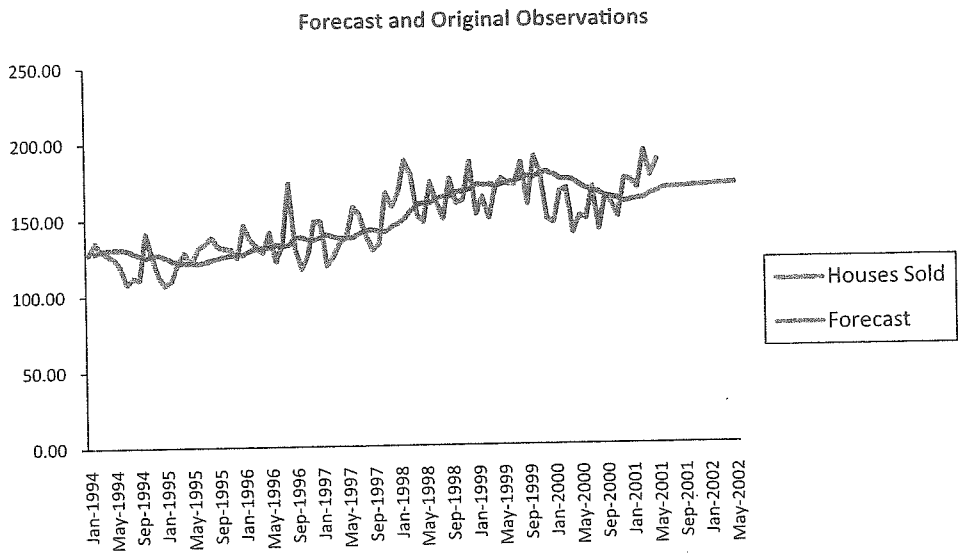
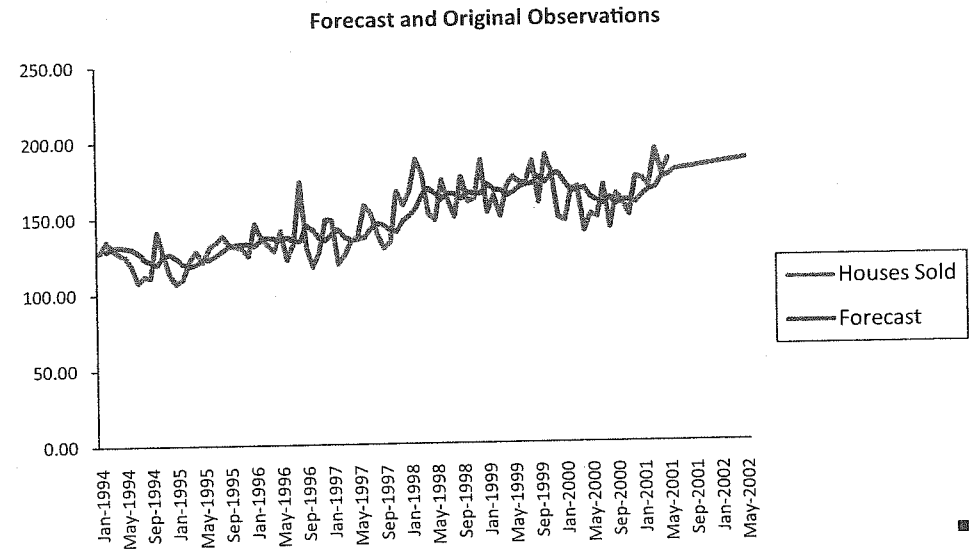


Figure 13.49
Forecasts from Holt's Method with Optimal Smoothing Constants



PROBLEMS

Level A

42. Consider the airline ticket data in the file **P13_01.xlsx**.
    a. Create a time series chart of the data. Based on what you see, which of the exponential smoothing models do you think should be used for forecasting? Why?
    b. Use simple exponential smoothing to forecast these data, using no holdout period and requesting 12 months of future forecasts. Use the default smoothing constant of 0.1.
    c. Repeat part **b**, optimizing the smoothing constant. Does it make much of an improvement?
    d. Write a short report to summarize your results.

43. Consider the applications for home mortgages data in the file **P13_04.xlsx**.
    a. Create a time series chart of the data. Based on what you see, which of the exponential smoothing models do you think should be used for forecasting? Why?
    b. Use simple exponential smoothing to forecast these data, using no holdout period and requesting 4 quarters of future forecasts. Use the default smoothing constant of 0.1.
    c. Repeat part **b**, optimizing the smoothing constant. Does it make much of an improvement?
    d. Write a short report to summarize your results.

44. Consider the American Express closing price data in the file **P13_16.xlsx**. Focus only on the closing prices.
    a. Create a time series chart of the data. Based on what you see, which of the exponential smoothing models do you think should be used for forecasting? Why? (*Note*: The data are currently sorted from most recent to most distant in the past. Sort them in the opposite order first.)
    b. Use Holt's exponential smoothing to forecast these data, using no holdout period and requesting 20 days of future forecasts. Use the default smoothing constants of 0.1.
    c. Repeat part **b**, optimizing the smoothing constant. Does it make much of an improvement?
    d. Repeat parts **a** and **b**, this time using a holdout period of 50 days.
    e. Write a short report to summarize your results.

45. Consider the poverty level data in the file **P02_29.xlsx**. Focus only on the Percent variable.
    a. Create a time series chart of the data. Based on what you see, which of the exponential smoothing models do you think should be used for forecasting? Why?
    b. Use simple exponential smoothing to forecast these data, using no holdout period and requesting 3 years of future forecasts. Use the default smoothing constants of 0.1.
    c. Repeat part **b**, optimizing the smoothing constant. Make sure you request a chart of the series with the forecasts superimposed. Does the optimal smoothing constant make much of an improvement?
    d. Write a short report to summarize your results. Considering the chart in part **c**, would you say the forecasts are "good"?

Problems 46 through 48 ask you to apply the exponential smoothing formulas. These do not require StatTools. In fact, they do not even require Excel. You can do them with a hand calculator (or with Excel).

46. TOD Chevy is using Holt's method to forecast weekly car sales. Currently, the level is estimated to be 50 cars per week, and the trend is estimated to be 6 cars per week. During the current week, 30 cars are sold. After observing the current week's sales, forecast the number of cars 3 weeks from now. Use $\alpha = \beta = 0.3$.

47. You have been assigned to forecast the number of aircraft engines ordered each month by Commins Engine Company. At the end of February, the forecast is that 100 engines will be ordered during April. Then during March, 120 engines are actually ordered.
    a. Using $\alpha = 0.3$, determine a forecast (at the end of March) for the number of orders placed during April; during May. Use simple exponential smoothing.
    b. Suppose MAE = 16 at the end of March. At the end of March, Commins can be 68% sure that April orders will be between what two values, assuming normally distributed forecast errors? (*Hint*: It can be shown that the standard deviation of forecast errors is approximately 1.25 times MAE.)

48. Simple exponential smoothing with $\alpha = 0.3$ is being used to forecast sales of radios at Lowland Appliance. Forecasts are made on a monthly basis. After August radio sales are observed, the forecast for September is 100 radios.
    a. During September, 120 radios are sold. After observing September sales, what do we forecast for October radio sales? For November radio sales?
    b. It turns out that June sales were recorded as 10 radios. Actually, however, 100 radios were sold in June. After correcting for this error, develop a forecast for October radio sales.

Level B

49. Holt's method assumes an additive trend. For example, a trend of 5 means that the level will increase by 5 units per period. Suppose there is actually a **multiplicative trend**. For example, if the current estimate of the level is 50 and the current estimate of the trend is 1.2, we would predict demand to increase by 20% per period. So we would forecast the next period's demand to be 50(1.2) and forecast the demand 2 periods in the future to be $50(1.2)^2$. If we want to use a multiplicative trend in Holt's method, we should use the following equations:

$$L_t = \alpha Y_t + (1 - \alpha)(I)$$
$$T_t = \beta(II) + (1 - \beta)T_{t-1}$$

   a. What should (*I*) and (*II*) be?
   b. Suppose we are working with monthly data and month 12 is December, month 13 is January, and so on. Also suppose that $L_{12} = 100$ and $T_{12} = 1.2$. Suppose $Y_{13} = 200$. At the end of month 13, what is the prediction for Y_{15}? Assume $\alpha = \beta = 0.5$ and a multiplicative trend.

50. A version of simple exponential smoothing can be used to predict the outcome of sporting events. To illustrate, consider pro football. We first assume that all games are played on a neutral field. Before each day of play, we assume that each team has a rating. For example, if the rating for the Bears is +10 and the rating for the Bengals is +6, we predict the Bears to beat the Bengals by 10 − 6 = 4 points. Suppose that the Bears play the Bengals and win by 20 points. For this game, we "underpredicted" the Bears' performance by 20 − 4 = 16 points. Assuming that the best α for pro football is $\alpha = 0.10$, we would increase the Bears' rating by 16(0.1) = 1.6 and decrease the Bengals' rating by 1.6 points. In a rematch, the Bears would then be favored by (10 + 1.6) − (6 − 1.6) = 7.2 points.
   a. How does this approach relate to the equation $L_t = L_{t-1} + \alpha e_t$?
   b. Suppose that the home field advantage in pro football is 3 points; that is, home teams tend to outscore visiting teams by an average of 3 points a game. How could the home field advantage be incorporated into this system?
   c. How might we determine the *best* α for pro football?
   d. How might we determine ratings for each team at the beginning of the season?
   e. Suppose we apply this method to predict pro football (16-game schedule), college football (11-game schedule), college basketball (30-game schedule), and pro basketball (82-game schedule). Which sport do you think would have the smallest optimal α; the largest optimal α? Why?
   f. Why might this approach yield poor forecasts for major league baseball?

13.9 SEASONAL MODELS

So far we have said practically nothing about seasonality. Seasonality is the consistent month-to-month (or quarter-to-quarter) differences that occur each year. For example, there is seasonality in beer sales—high in the summer months, lower in other months. Toy sales are also seasonal, with a huge peak in the months preceding Christmas. In fact, if you start thinking about time series variables that you are familiar with, the majority of them probably have some degree of seasonality.

Some time series software packages have special types of graphs for spotting seasonality, but we don't discuss these here.

How do we know whether there is seasonality in a time series? The easiest way is to check whether a graph of the time series has a *regular* pattern of ups and/or downs in particular months or quarters. Although random noise can sometimes obscure such a pattern, the seasonal pattern is usually fairly obvious.

There are basically three methods for dealing with seasonality. First, we can use Winters' exponential smoothing model. It is similar to simple exponential smoothing and Holt's method, except that it includes another component (and smoothing constant) to capture seasonality. Second, we can *deseasonalize* the data, then use any of our forecasting methods to model the deseasonalized data, and finally "reseasonalize" these forecasts. Finally, we can use multiple regression with dummy variables for the seasons. We discuss all three of these methods in this section.

As we saw with the housing sales data, government agencies often perform part of the second method for us—that is, they deseasonalize the data.

Seasonal models are usually classified as *additive* or *multiplicative*. Suppose that we have monthly data, and that the average of the 12 monthly values for a typical year is 150. An **additive** model finds seasonal indexes, one for each month, that we *add* to the monthly

average, 150, to get a particular month's value. For example, if the index for March is 22, then we expect a typical March value to be 150 + 22 = 172. If the seasonal index for September is −12, then we expect a typical September value to be 150 − 12 = 138. A **multiplicative** model also finds seasonal indexes, but we *multiply* the monthly average by these indexes to get a particular month's value. Now if the index for March is 1.3, we expect a typical March value to be 150(1.3) = 195. If the index for September is 0.9, then we expect a typical September value to be 150(0.9) = 135. These models are summarized here.

In an **additive** seasonal model, we add an appropriate seasonal index to a "base" forecast. These indexes, one for each season, typically average to 0.

In a **multiplicative** seasonal model, we multiply a "base" forecast by an appropriate seasonal index. These indexes, one for each season, typically average to 1.

Either an additive or a multiplicative model can be used to forecast seasonal data. However, because multiplicative models are somewhat easier to interpret (and have worked well in applications), we focus on them. Note that the seasonal index in a multiplicative model can be interpreted as a percentage. Using the figures in the previous paragraph as an example, March tends to be 30% above the monthly average, whereas September tends to be 10% below it. Also, the seasonal indexes in a multiplicative model should average to 1. Computer packages typically ensure that this happens.

13.9.1 Winters' Exponential Smoothing Model

We now turn to Winters' exponential smoothing model. It is very similar to Holt's model—it again has level and trend terms and corresponding smoothing constants α and β—but it also has seasonal indexes and a corresponding smoothing constant γ (gamma). This new smoothing constant γ controls how quickly the method reacts to perceived changes in the pattern of seasonality. If γ is small, the method reacts slowly. If it is large, the method reacts more quickly. As with Holt's model, there are equations for updating the level and trend terms, and there is one extra equation for updating the seasonal indexes. For completeness, we list these equations in the accompanying box, but they are clearly too complex for hand calculation and are best left to the computer. In equation (13.21), S_t refers to the multiplicative seasonal index for period t. In equations (13.19), (13.21), and (13.22), M refers to the number of seasons ($M = 4$ for quarterly data, $M = 12$ for monthly data).

Formulas for Winters' Exponential Smoothing Model

$$L_t = \alpha \frac{Y_t}{S_{t-M}} + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad \textbf{(13.19)}$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad \textbf{(13.20)}$$

$$S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-M} \quad \textbf{(13.21)}$$

$$F_{t+k} = (L_t + kT_t)S_{t+k-M} \quad \textbf{(13.22)}$$

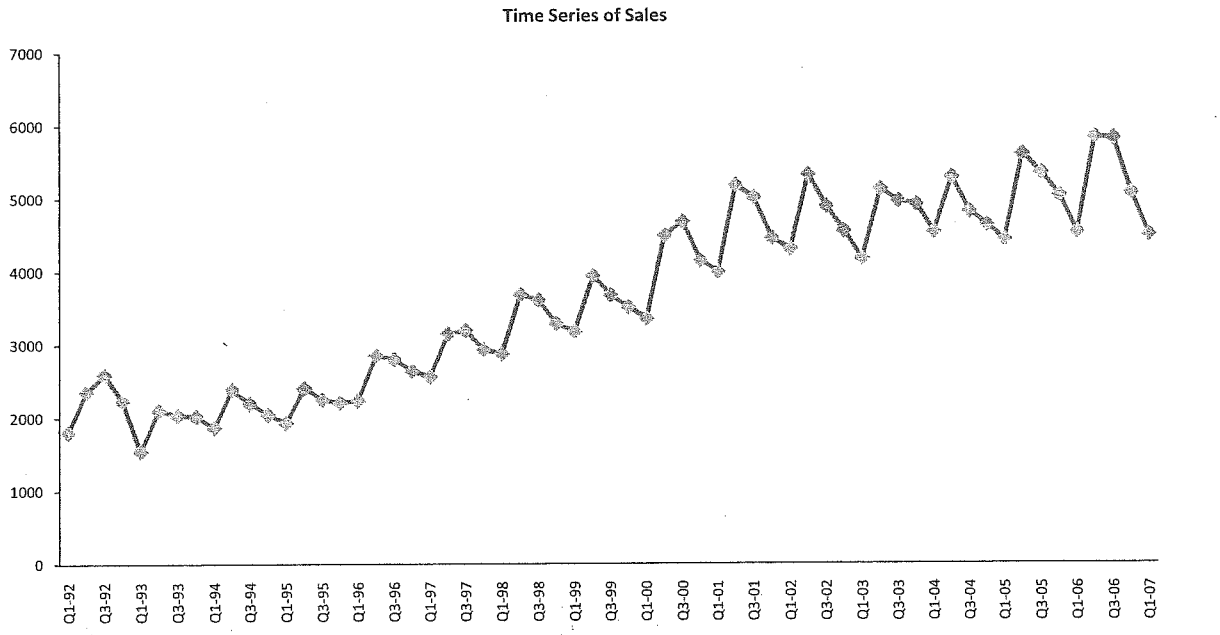
To see how the forecasting in equation (13.22) works, suppose we have observed data through June and want a forecast for the coming September, that is, a 3-month-ahead forecast. (In this case t refers to June and $t + k = t + 3$ refers to September.) Then we first add 3 times the current trend term to the current level. This gives a forecast for September that would be appropriate if there were no seasonality. Next, we multiply this forecast by the

most recent estimate of September's seasonal index (the one from the previous September) to get the forecast for September. Of course, the computer does all of the arithmetic, but this is basically what it is doing. We illustrate the method in the following example.

EXAMPLE 13.7 QUARTERLY SOFT DRINK SALES

The data in the **Soft Drink Sales.xlsx** file represent quarterly sales (in millions of dollars) for a large soft drink company from quarter 1 of 1992 through quarter 1 of 2007. As we might expect, there has been an upward trend in sales during this period, and there is also a fairly regular seasonal pattern, as shown in Figure 13.50. Sales in the warmer quarters, 2 and 3, are consistently higher than in the colder quarters, 1 and 4. How well can Winters' method track this upward trend and seasonal pattern?

Figure 13.50 Time Series Graph of Soft Drink Sales



Objective To see how well Winters' method, with appropriate smoothing constants, can forecast the company's seasonal soft drink sales.

Solution

To use Winters' method with StatTools, we proceed exactly as with any of the other exponential smoothing methods. However, for a change (and because we have so many years of data), we use StatTools's option of holding out some of the data for validation. Specifically, we fill out the Time Scale tab in the Forecast dialog box as shown in Figure 13.51. Then we fill in the Forecast Settings tab of this dialog box as shown in Figure 13.52, selecting Winters' method, basing the model on the data through quarter 1, 2003, holding out 8 quarters of data (quarter 2,

2003, through quarter 1, 2005), and forecasting 4 quarters into the future. Note that when we choose Winters' method in Figure 13.52, the Deseasonalize option in Figure 13.51 is automatically disabled. It wouldn't make sense to deseasonalize *and* use Winters' method; we do one or the other. Also, we have elected to optimize the smoothing constants, but this is optional.

Figure 13.51
Time Scale Settings for Soft Drink Sales

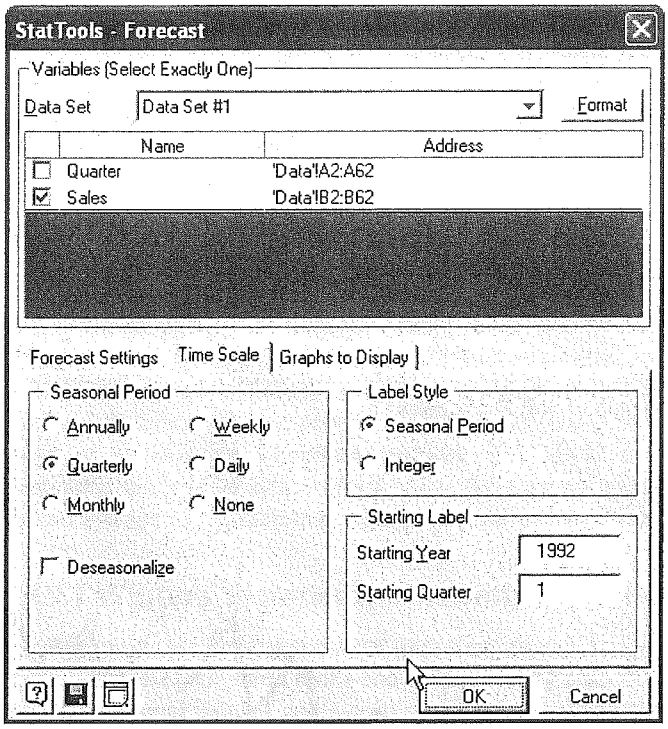
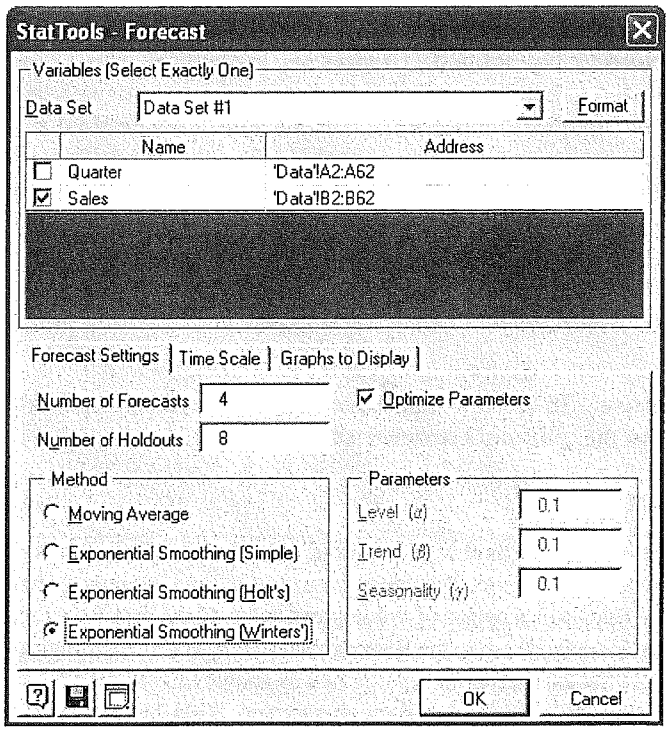


Figure 13.52
Forecast Settings for Soft Drink Sales



You can check that if we hold out 3 years of data, the MAPE for the holdout period increases quite a lot. It is common for the fit to be considerably better in the estimation period than in the holdout period.

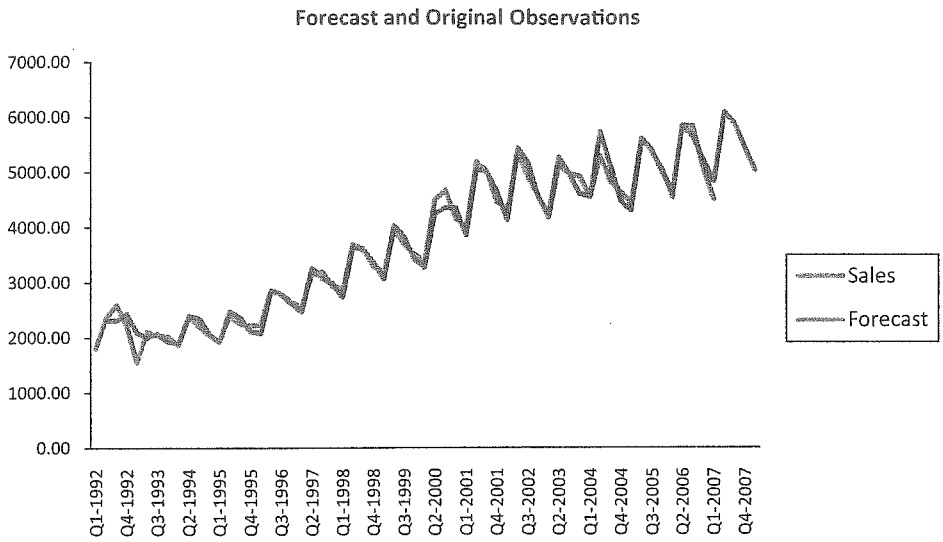
Parts of the output are shown in Figure 13.53. The following points are worth noting: (1) The optimal smoothing constants (those that minimize RMSE) are $\alpha = 1.0$, $\beta = 0.0$, and $\gamma = 0.0$. Intuitively, these mean that we react right away to changes in level, but we never react to changes in trend or the seasonal pattern. (2) If we ignore seasonality, the series is trending upward at a rate of 56.05 per quarter (see column D). This is our initial estimate of trend and, because $\beta = 0$, it never changes. (3) The seasonal pattern stays constant throughout this 10-year period. The seasonal indexes, shown in column E, are 0.88, 1.10, 1.05, and 0.96. For example, quarter 1 is 12% below the yearly average, and quarter 2 is 10% above the yearly average. (4) The forecast series tracks the actual series quite well during the nonholdout period. For example, MAPE is 3.98%, meaning that on average our forecasts are off by about 4% on average. Surprisingly, MAPE for the holdout period is even lower, at 2.20%.

Figure 13.53 Output from Winters' Method for Soft Drink Sales

	A	B	C	D	E	F	G
7							
8	*Forecasting Constants (Optimized)*						
9	Level (Alpha)	1.000					
10	Trend (Beta)	0.000					
11	Season (Gamma)	0.000					
12							
13		Estimation	Holdouts				
14	*Winters' Exponential*	Period	Period				
15	Mean Abs Err	125.59	108.31				
16	Root Mean Sq Err	168.81	148.94				
17	Mean Abs Per% Err	3.98%	2.20%				
18							
41							
42	*Forecasting Data*	Sales	Level	Trend	Season	Forecast	Error
43	Q1-1992	1807.3700	2046.27	56.05	0.88		
44	Q2-1992	2355.3200	2140.57	56.05	1.10	2313.24	42.08
45	Q3-1992	2591.8300	2463.29	56.05	1.05	2311.24	280.59
46	Q4-1992	2236.3900	2319.32	56.05	0.96	2429.27	-192.88
47	Q1-1993	1549.1400	1753.91	56.05	0.88	2098.05	-548.91
48	Q2-1993	2105.7900	1913.79	56.05	1.10	1991.54	114.25
49	Q3-1993	2041.3200	1940.09	56.05	1.05	2072.63	-31.31
50	Q4-1993	2021.0100	2095.95	56.05	0.96	1924.76	96.25
92	Q2-2004	5284.7100	4802.87	56.05	1.10	5720.69	-435.98
93	Q3-2004	4817.4300	4578.52	56.05	1.05	5112.46	-295.03
94	Q4-2004	4634.5000	4806.36	56.05	0.96	4468.86	165.64
95	Q1-2005	4431.3600	5017.10	56.05	0.88	4294.73	136.63
96	Q2-2005	5602.2100				5582.11	20.10
97	Q3-2005	5349.8500				5396.85	-47.00
98	Q4-2005	5036.0000				4999.85	36.15
99	Q1-2006	4534.6100				4629.40	-94.79
100	Q2-2006	5836.1700				5828.82	7.35
101	Q3-2006	5818.2800				5632.76	185.52
102	Q4-2006	5070.4200				5216.05	-145.63
103	Q1-2007	4497.4700				4827.43	-329.96
104	Q2-2007					6075.52	
105	Q3-2007					5868.67	
106	Q4-2007					5432.24	
107	Q1-2008					5025.47	

The plot of the forecasts superimposed on the original series, shown in Figure 13.54, indicates that Winters' method clearly picks up the seasonal pattern and the upward trend and projects both of these into the future. In later examples, we investigate whether other seasonal forecasting methods can do this well.

Figure 13.54
Graph of Forecasts from Winters' Method



One final comment is that we are not obligated to find the *optimal* smoothing constants. Some analysts might suggest using more "typical" values such as $\alpha = \beta = 0.2$ and $\gamma = 0.5$. (We often choose γ larger than α and β because each season's seasonal index gets updated only once per year.) To see how these smoothing constants affect the results, we can simply substitute their values in the range B9:B11 of Figure 13.53. As we would expect, MAE, RMSE, and MAPE all get somewhat worse (they increase to 184, 237, and 5.89%, respectively, for the estimation period), but a plot of the forecasts superimposed on the original sales data still indicates a very good fit. ■

The three exponential smoothing methods we have examined are not the only ones available. For example, there are linear and quadratic models available in some software packages. These are somewhat similar to Holt's model except that they use only a single smoothing constant. There are also adaptive exponential smoothing models, where the smoothing constants themselves are allowed to change through time. Although these more complex models have been studied thoroughly in the academic literature and are used by some practitioners, they typically offer only marginal gains in forecast accuracy over the models we have examined.

13.9.2 Deseasonalizing: The Ratio-to-Moving-Averages Method

You have probably seen references to time series data that have been *deseasonalized*. In this section we discuss why this is done and how it is done. We also see how it can be used to forecast seasonal time series. First, data are often published in deseasonalized form so that readers can spot trends more easily. For example, if we see a time series of sales that has not been deseasonalized, and it shows a large increase from November to December, we might not be sure whether this represents a real increase in sales or a seasonal phenomenon (Christmas sales). However, if this increase is really just a seasonal effect, then the deseasonalized version of the series will show no such increase in sales.

Government economists and statisticians have a variety of sophisticated methods for deseasonalizing time series data, but they are typically variations of the **ratio-to-moving-averages** method described here. This method is applicable when we believe that seasonality

is multiplicative, as described in the previous section. Our job is to find the seasonal indexes, which can then be used to deseasonalize the data. For example, if we estimate the index for June to be 1.3, this means that June's values are typically about 30% larger than the average for all months. Therefore, to deseasonalize a June value, we *divide* it by 1.3 (to make it smaller). Similarly, if February's index is 0.85, then February's values are 15% below the average for all months. So to deseasonalize a February value, we divide it by 0.85 (to make it larger).

To **deseasonalize** an observation (assuming a multiplicative model of seasonality), *divide* it by the appropriate seasonal index.

To find the seasonal index for June 2005 (or any other month) in the first place, we essentially divide June's observation by the average of the 12 observations surrounding June. (This is the reason for the term "ratio" in the name of the method.) There is one minor problem with this approach. June 2005 is not exactly in the middle of any 12-month sequence. If we use the 12 months from January 2005 to December 2005, June 2005 is in the *first* half of the sequence; if we use the 12 months from December 2004 to November 2005, June 2005 is in the *last* half of the sequence. Therefore, we compromise by averaging the January-to-December and December-to-November averages. This is called a **centered** average. Then the seasonal index for June is June's observation divided by this centered average. The following equation shows more specifically how it works.

$$\text{Jun2005 index} = \frac{\text{Jun2005}}{\left(\frac{\text{Dec2004}+\cdots+\text{Nov2005}}{12} + \frac{\text{Jan2005}+\cdots+\text{Dec2005}}{12}\right)/2}$$

The only remaining question is how to combine all of the indexes for any specific month such as June. After all, if we have data for several years, the above procedure produces several June indexes, one for each year. The usual way to combine them is simply to average them. This single average index for June is then used to deseasonalize *all* of the June observations.

Once the seasonal indexes are obtained, we divide each observation by its seasonal index to deseasonalize the data. The deseasonalized data can then be forecasted by *any* of the methods we have described (other than Winters' method, which wouldn't make much sense). For example, we could use Holt's method or the moving averages method to forecast the deseasonalized data. Finally, we "reseasonalize" the forecasts by *multiplying* them by the seasonal indexes.

As this description suggests, the method is not meant for hand calculations! However, it is straightforward to implement in StatTools, as we illustrate in the following example.

EXAMPLE 13.7 QUARTERLY SOFT DRINK SALES (CONTINUED)

We return to the soft drink sales data. (See the file **Soft Drink Sales.xlsx**.) Is it possible to obtain the same forecast accuracy with the ratio-to-moving-averages method as we obtained with Winters' method?

Objective To use the ratio-to-moving-averages method to deseasonalize the soft drink data and then forecast the deseasonalized data.

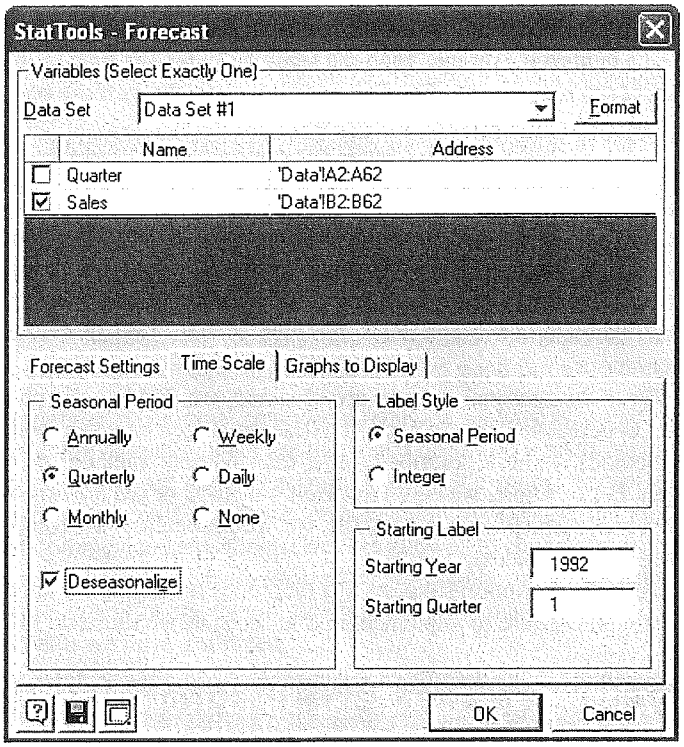
Solution

The answer to this question depends on which forecasting method we use to forecast the *deseasonalized* data. The ratio-to-moving-averages method only provides a means for

deseasonalizing the data and providing seasonal indexes. Beyond this, any method can be used to forecast the deseasonalized data, and some methods obviously work better than others. For this example, we compared two possibilities: the moving averages method with a span of 4 quarters, and Holt's exponential smoothing method optimized. However, we show the results only for the latter. Because the deseasonalized series still has a clear upward trend, we would expect Holt's method to do well, and we would expect the moving averages forecasts to lag behind the trend. This is exactly what occurred. For example, the values of MAPE for the two methods are 8.33% (moving averages) and 3.98% (Holt's). (To make a fair comparison with the Winters' method output for these data, we again held out 8 quarters. The MAPE values reported are for the nonholdout period.)

To implement this latter method in StatTools, we proceed exactly as before, but this time we check the Deseasonalize option in the Time Scale tab of the Forecast dialog box. (See Figure 13.55.) Note that when the Holt's option is checked, this Deseasonalize option is enabled. When we check this option, we get a larger selection of optional charts in the Graphs to Display tab. We can see charts of the deseasonalized data and/or the original "reseasonalized" data.

Figure 13.55
Checking the Deseasonalizing Option



Selected outputs are shown in Figures 13.56 through 13.59. Figures 13.56 and 13.57 show the numerical output. In particular, Figure 13.57 shows the seasonal indexes from the ratio-to-moving averages method in column C. These are virtually identical to the seasonal indexes we found using Winters' method, although the methods are mathematically different. Column D contains the deseasonalized sales (column B divided by column C), columns E through H implement Holt's method on the deseasonalized data, and columns I and J are the "reseasonalized" forecasts and errors.

Figure 13.56
Summary Measures for Forecast Errors

	A	B	C	D	E
8	*Forecasting Constants (Optimized)*				
9	Level (Alpha)	1.000			
10	Trend (Beta)	0.000			
11					
12		Estimation	Holdouts	Deseason	Deseason
13	*Holt's Exponential*	Period	Period	Estimate	Holdouts
14	Mean Abs Err	125.59	108.31	126.66	114.42
15	Root Mean Sq Err	168.81	148.94	171.57	161.54
16	Mean Abs Per% Err	3.98%	2.20%	3.98%	2.20%

Figure 13.57 Ratio-to-Moving-Averages Output

	A	B	C	D	E	F	G	H	I	J
61			Season	Deseason	Deseason	Deseason	Deseason	Deseason	Season	Season
62	*Forecasting Data*	Sales	Index	Sales	Level	Trend	Forecast	Errors	Forecast	Errors
63	Q1-1992	1807.3700	0.88	2046.27	2046.27	56.05				
64	Q2-1992	2355.3200	1.10	2140.57	2140.57	56.05	2102.32	38.25	2313.24	42.08
65	Q3-1992	2591.8300	1.05	2463.29	2463.29	56.05	2196.62	266.67	2311.24	280.59
66	Q4-1992	2236.3900	0.96	2319.32	2319.32	56.05	2519.35	-200.03	2429.27	-192.88
67	Q1-1993	1549.1400	0.88	1753.91	1753.91	56.05	2375.37	-621.47	2098.05	-548.91
68	Q2-1993	2105.7900	1.10	1913.79	1913.79	56.05	1809.96	103.83	1991.54	114.25
69	Q3-1993	2041.3200	1.05	1940.09	1940.09	56.05	1969.84	-29.76	2072.63	-31.31
70	Q4-1993	2021.0100	0.96	2095.95	2095.95	56.05	1996.14	99.81	1924.76	96.25
112	Q2-2004	5284.7100	1.10	4802.87	4802.87	56.05	5199.09	-396.23	5720.69	-435.98
113	Q3-2004	4817.4300	1.05	4578.52	4578.52	56.05	4858.92	-280.40	5112.46	-295.03
114	Q4-2004	4634.5000	0.96	4806.36	4806.36	56.05	4634.58	171.78	4468.86	165.64
115	Q1-2005	4431.3600	0.88	5017.10	5017.10	56.05	4862.41	154.69	4294.73	136.63
116	Q2-2005	5602.2100	1.10	5091.42			5073.15	18.27	5582.11	20.10
117	Q3-2005	5349.8500	1.05	5084.54			5129.20	-44.67	5396.85	-47.00
118	Q4-2005	5036.0000	0.96	5222.75			5185.26	37.49	4999.85	36.15
119	Q1-2006	4534.6100	0.88	5133.99			5241.31	-107.32	4629.40	-94.79
120	Q2-2006	5836.1700	1.10	5304.05			5297.36	6.68	5828.82	7.35
121	Q3-2006	5818.2800	1.05	5529.74			5353.42	176.32	5632.76	185.52
122	Q4-2006	5070.4200	0.96	5258.44			5409.47	-151.03	5216.05	-145.63
123	Q1-2007	4497.4700	0.88	5091.95			5465.52	-373.58	4827.43	-329.96
124	Q2-2007		1.10				5521.58		6075.52	
125	Q3-2007		1.05				5577.63		5868.67	
126	Q4-2007		0.96				5633.68		5432.24	
127	Q1-2008		0.88				5689.74		5025.47	

Figure 13.58
Forecast Graph of Deseasonalized Series

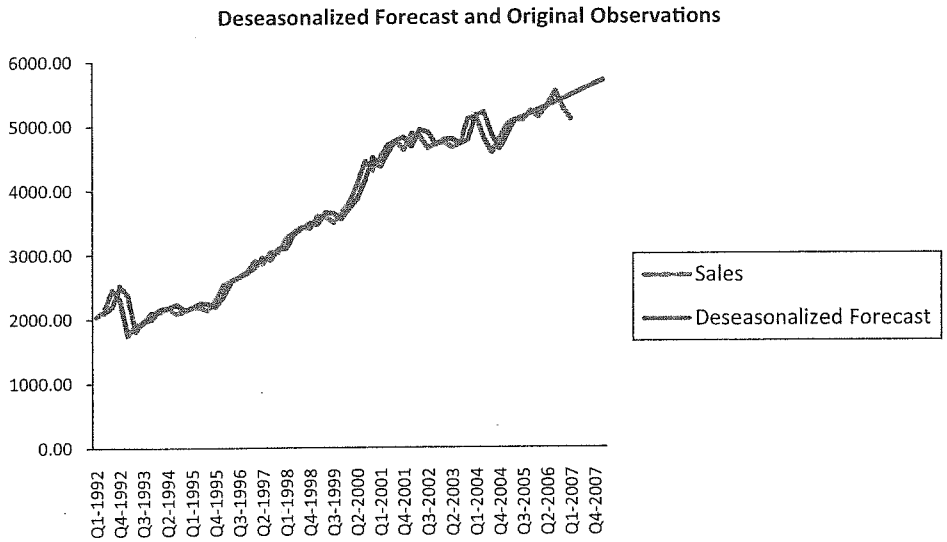
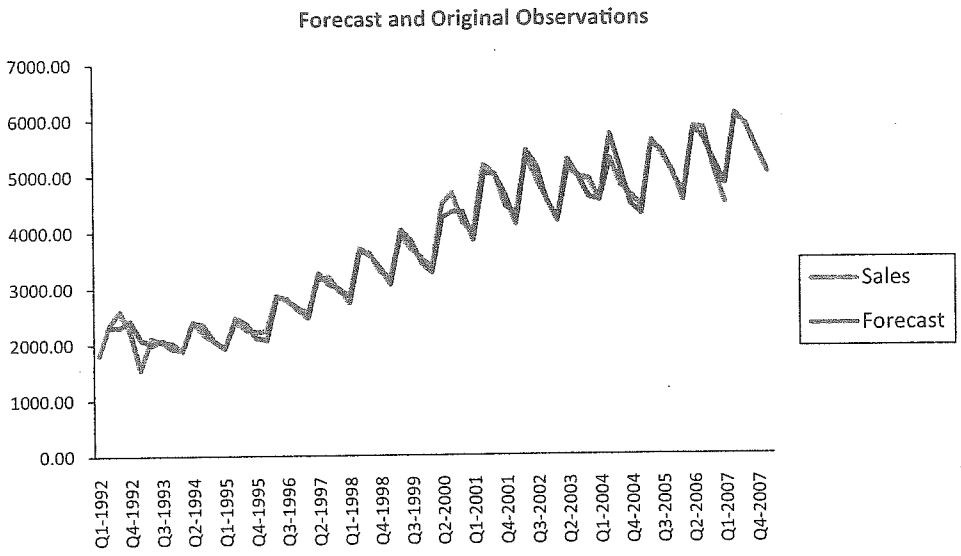


Figure 13.59
Forecast Graph of Reseasonalized (Original) Series



The deseasonalized data, with forecasts superimposed, appear in Figure 13.58. Here we see only the smooth upward trend with no seasonality, which Holt's method is able to track very well. Then Figure 13.59 shows the results of reseasonalizing. Again, the forecasts track the actual sales data very well. In fact, we see that the summary measures of forecast errors (in Figure 13.56, range B14:B16) are quite comparable to those from Winters' method. The reason is that both arrive at virtually the same seasonal pattern. ■

13.9.3 Estimating Seasonality with Regression

We now examine a regression approach to forecasting seasonal data that uses dummy variables for the seasons. Depending on how we write the regression equation, we can create either an additive or a multiplicative seasonal model.

As an example, suppose that the data are quarterly data with a possible linear trend. Then we can introduce dummy variables Q_1, Q_2, and Q_3 for the first three quarters (using quarter 4 as the reference quarter) and estimate the additive equation

$$\hat{Y}_t = a + bt + b_1Q_1 + b_2Q_2 + b_3Q_3$$

Then the coefficients of the dummy variables, b_1, b_2 and b_3, indicate how much each quarter differs from the reference quarter, quarter 4, and the coefficient b represent the trend.

For example, if the estimated equation is

$$\hat{Y}_t = 130 + 25t + 15Q_1 + 5Q_2 - 20Q_3$$

then the average increase from one quarter to the next is 25 (the coefficient of t). This is the trend effect. However, quarter 1 averages 15 units higher than quarter 4, quarter 2 averages 5 units higher than quarter 4, and quarter 3 averages 20 units lower than quarter 4. These coefficients indicate the seasonality effect.

As discussed in Chapter 11, it is also possible to estimate a *multiplicative* model using dummy variables for seasonality (and possibly time for trend). Then we would estimate the equation

$$\hat{Y}_t = ae^{bt}e^{b_1Q_1}e^{b_2Q_2}e^{b_3Q_3}$$

or, after taking logs,

$$\ln \hat{Y}_t = \ln a + bt + b_1Q_1 + b_2Q_2 + b_3Q_3$$

One advantage of this approach is that it provides a model with *multiplicative* seasonal factors. It is also fairly easy to interpret the regression output, as illustrated in the following continuation of the soft drink sales example.

EXAMPLE 13.7 QUARTERLY SOFT DRINK SALES (CONTINUED)

Returning to the soft drink sales data (see the file **Soft Drink Sales.xlsx**), does a regression approach provide forecasts that are as accurate as those provided by the other seasonal methods in this chapter?

Objective To use a multiplicative regression equation, with dummy variables for seasons and a time variable for trend, to model soft drink sales.

Solution

We illustrate the multiplicative approach, although an additive approach is also possible. Figure 13.60 illustrates the data setup. Besides the Sales and Time variables, we need dummy variables for three of the four quarters (we created these manually), and a Log(Sales) variable. We then use multiple regression, with Log(Sales) as the dependent variable, and Time, Q1, Q2, and Q3 as the explanatory variables.

Figure 13.60
Data Setup for Multiplicative Model with Dummies

	A	B	C	D	E	F	G
1	Quarter	Sales	Time	Q1	Q2	Q3	Log(Sales)
2	Q1-92	1807.37	1	1	0	0	7.499628
3	Q2-92	2355.32	2	0	1	0	7.7644319
4	Q3-92	2591.83	3	0	0	1	7.8601195
5	Q4-92	2236.39	4	0	0	0	7.7126182
6	Q1-93	1549.14	5	1	0	0	7.3454552
7	Q2-93	2105.79	6	0	1	0	7.652446
8	Q3-93	2041.32	7	0	0	1	7.6213519
9	Q4-93	2021.01	8	0	0	0	7.6113527
10	Q1-94	1870.46	9	1	0	0	7.5339397
11	Q2-94	2390.56	10	0	1	0	7.7792829
12	Q3-94	2198.03	11	0	0	1	7.6953168
13	Q4-94	2046.83	12	0	0	0	7.6240475
14	Q1-95	1934.19	13	1	0	0	7.5674439
15	Q2-95	2406.41	14	0	1	0	7.7858913

The regression output appears in Figure 13.61. (Again, to make a fair comparison with previous methods, we base the regression only on the data through quarter 1 of 2005. That is, we again hold out the last 8 quarters. This means that the StatTools data set should be redefined so that it extends only through row 54.) Of particular interest are the coefficients of the explanatory variables. Recall that for a log dependent variable, these coefficients can be interpreted as *percentage* changes in the original sales variable. Specifically, the coefficient of Time means that deseasonalized sales increase by about 2.1% per quarter. Also, the coefficients of Q1, Q2, and Q3 mean that sales in quarters 1, 2, and 3 are, respectively, about 8.4% below, 14.0% above, and 8.9% above sales in the reference quarter, quarter 4. This pattern is quite comparable to the pattern of seasonal indexes we saw in previous models for these data.

Figure 13.61 Regression Output for Multiplicative Model

	A	B	C	D	E	F	G
7		Multiple		Adjusted	StErr of		
8	*Summary*	R	R-Square	R-Square	Estimate		
9		0.9660	0.9332	0.9276	0.0945		
10							
11		Degrees of	Sum of	Mean of			
12	*ANOVA Table*	Freedom	Squares	Squares	F-Ratio	p-Value	
13	Explained	4	5.9813	1.4953	167.6184	< 0.0001	
14	Unexplained	48	0.4282	0.0089			
15							
16			Standard			Confidence Interval 95%	
17	*Regression Table*	Coefficient	Error	t-Value	p-Value	Lower	Upper
18	Constant	7.4689	0.0354	211.1359	< 0.0001	7.3977	7.5400
19	Time	0.0214	0.0008	25.1710	< 0.0001	0.0197	0.0231
20	Q1	-0.0836	0.0364	-2.2973	0.0260	-0.1568	-0.0104
21	Q2	0.1402	0.0371	3.7806	0.0004	0.0656	0.2148
22	Q3	0.0894	0.0371	2.4122	0.0197	0.0149	0.1639

To compare the forecast accuracy of this method with earlier models, we must go through several steps manually. (See Figure 13.62 for reference.) We first calculate the forecasts in column H by entering the formula

=EXP(Regression!$B$18+MMULT(Data!C2:F2,Regression!$B$19:$B$22))

in cell H2 and copying it down. (This formula assumes the regression output is in a sheet named Regression. It uses Excel's MMULT function to sum the products of explanatory values and regression coefficients. You can replace this by "writing out" the sum of products if you like. The formula then takes EXP of the resulting sum to convert the log sales value back to the original sales units.) Next, we calculate the absolute errors, squared errors, and absolute percentage errors in columns I, J, and K, and we summarize them in the usual way, both for the estimation period and the holdout period, in columns N and O.

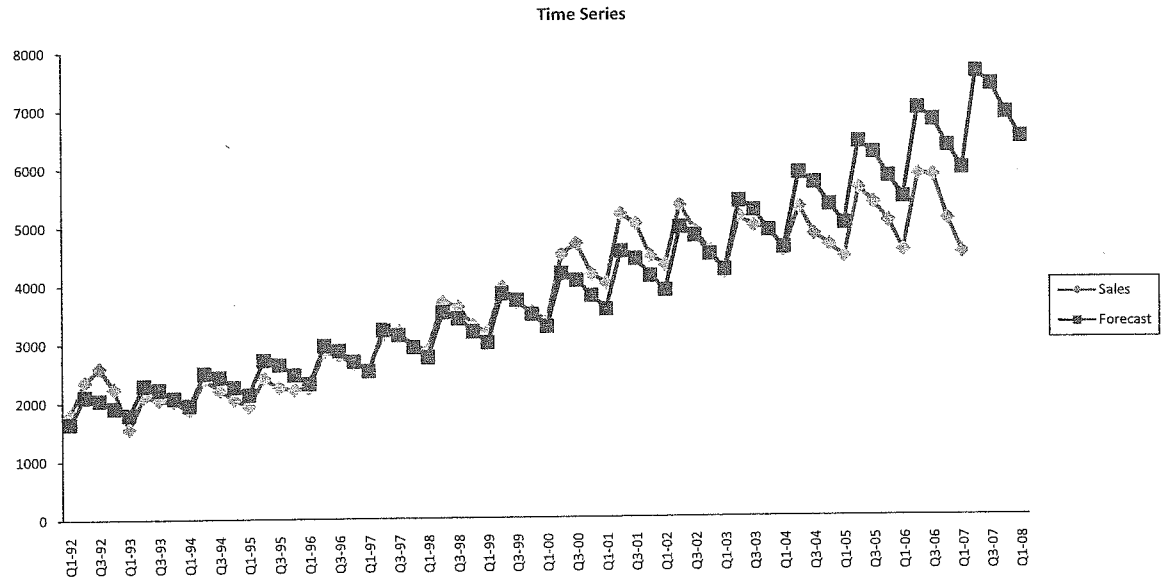
Note that these summary measures are considerably larger for this regression model than for the previous seasonality models, especially in the holdout period. We can get some idea why the holdout period does so poorly by looking at the plot of observations versus forecasts in Figure 13.63. The multiplicative regression model with Time included really implies *exponential* growth (as in Section 13.4.2), with seasonality superimposed. However, this company's sales growth tapered off in the last couple of years and did not keep up with the exponential growth curve. In short, the dummy variables do a good job of

tracking seasonality, but the underlying exponential trend curve outpaces actual sales. We conclude that this regression model is *not* as good for forecasting this company's sales as Winters' method or Holt's method on the deseasonalized data.

Figure 13.62 Forecast Errors and Summary Measures

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Quarter	Sales	Time	Q1	Q2	Q3	Log(Sales)	Forecast	SqError	AbsError	PctAbsError		Error measures		
2	Q1-92	1807.37	1	1	0	0	7.499628	1646.863	25762.59	160.5073	0.0888071			Estimation	Holdout
3	Q2-92	2355.32	2	0	1	0	7.7644319	2104.437	62942.35	250.8831	0.10651764		RMSE	319.38	1035.32
4	Q3-92	2591.83	3	0	0	1	7.8601195	2043.366	300812.4	548.4637	0.21161252		MAE	244.81	1010.96
5	Q4-92	2236.39	4	0	0	0	7.7126182	1909.004	107181.4	327.3857	0.14639024		MAPE	7.28%	19.71%
6	Q1-93	1549.14	5	1	0	0	7.3454552	1793.833	59874.7	244.6931	0.15795414				
7	Q2-93	2105.79	6	0	1	0	7.652446	2292.242	34764.48	186.4524	0.08854271				
8	Q3-93	2041.32	7	0	0	1	7.6213519	2225.722	34004	184.4017	0.09033455				
9	Q4-93	2021.01	8	0	0	0	7.6113527	2079.369	3405.764	58.35892	0.02887612				
10	Q1-94	1870.46	9	1	0	0	7.5339397	1953.919	6965.484	83.45947	0.04461976				
11	Q2-94	2390.56	10	0	1	0	7.7792829	2496.808	11288.66	106.2481	0.04444486				
12	Q3-94	2198.03	11	0	0	1	7.6953168	2424.351	51221.2	226.321	0.10296538				
13	Q4-94	2046.83	12	0	0	0	7.6240475	2264.937	47570.79	218.1073	0.10655858				
14	Q1-95	1934.19	13	1	0	0	7.5674439	2128.292	37675.74	194.1024	0.10035333				

Figure 13.63 Graph of Forecasts for Multiplicative Model



This method of detecting seasonality by using dummy variables in a regression equation is always an option. The other variables included in the regression equation could be time t, lagged versions of Y_t, and/or current or lagged versions of other independent variables. These variables would capture any time series behavior other than seasonality. Just remember that there is always one less dummy variable than the number of seasons. If the

data are quarterly, then three dummies are needed; if the data are monthly, then 11 dummies are needed. If the coefficients of any of these dummies turn out to be statistically insignificant, they can be omitted from the equation. Then the omitted terms are effectively combined with the reference season. For example, if the Q_1 term were omitted, then quarters 1 and 4 would essentially be combined and treated as the reference season, and the other two seasons would be compared to them through their dummy variable coefficients. ■

PROBLEMS

Level A

51. The University Credit Union is open Monday through Saturday. Winters' method is being used (with $\alpha = \beta = \gamma = 0.5$) to predict the number of customers entering the bank each day. After incorporating the arrivals of 16 October, $L_t = 200$, $T_t = 1$, and the "seasonalities" are as follows: Monday, 0.90; Tuesday, 0.70; Wednesday, 0.80; Thursday, 1.1; Friday, 1.2; Saturday, 1.3. For example, the number of customers entering the bank on a typical Monday is 90% of the number of customers entering the bank on an average day. On Tuesday, 17 October, 182 customers enter the bank. At the close of business on 17 October, forecast the number of customers who will enter the bank on 25 October.

52. Last National Bank is using Winters' method (with $\alpha = 0.2$, $\beta = 0.1$, and $\gamma = 0.5$) to forecast the number of customers served each day. The bank is open Monday through Friday. At present, the following "seasonalities" have been estimated: Monday, 0.80; Tuesday, 0.90; Wednesday, 0.95; Thursday, 1.10; Friday, 1.25. A seasonality of 0.80 for Monday means that on a Monday, the number of customers served by the bank tends to be 80% of the average daily value. Currently, the level is estimated to be 20 customers, and the trend is estimated to equal 1 customer. After observing that 30 customers are served by the bank on Monday, forecast the number of customers who will be served by the bank on Wednesday.

53. Suppose that Winters' method is used to forecast quarterly U.S. retail sales (in billions of dollars). At the end of the first quarter of 2006, $L_t = 300$, $T_t = 30$, and the seasonal indexes are as follows: quarter 1, 0.90; quarter 2, 0.95; quarter 3, 0.95; quarter 4, 1.20. During the second quarter of 2006, retail sales are $360 billion. Assume $\alpha = 0.2$, $\beta = 0.4$, and $\gamma = 0.5$.
    a. At the end of the second quarter of 2006, develop a forecast for retail sales during the fourth quarter of 2006.
    b. At the end of the second quarter of 2006, develop a forecast for the second quarter of 2007.

54. The file **P02_38.xlsx** contains monthly retail sales of U.S. liquor stores.
    a. Is seasonality present in these data? If so, characterize the seasonality pattern and then deseasonalize this time series using the ratio-to-moving-average method.
    b. If you decided to deseasonalize this time series in part **a**, forecast the deseasonalized data for each month of the next year using the moving average method with an appropriate span.
    c. Does Holt's exponential smoothing method, with optimal smoothing constants, outperform the moving average method employed in part **b**? Demonstrate why or why not.

55. Continuing the previous problem, how do your responses to the questions change when you employ Winters' method to handle seasonality in this time series? Explain. Which forecasting method do you prefer, Winters' method or a method used in the previous problem? Defend your choice.

56. The file **P02_39.xlsx** contains monthly time series data for total U.S. retail sales of building materials (which includes retail sales of building materials, hardware and garden supply stores, and mobile home dealers).
    a. Is seasonality present in these data? If so, characterize the seasonality pattern and then deseasonalize this time series using the ratio-to-moving-average method.
    b. If you decided to deseasonalize this time series in part **a**, forecast the deseasonalized data for each month of the next year using the moving average method with an appropriate span.
    c. Does Holt's exponential smoothing method, with optimal smoothing constants, outperform the moving average method employed in part **b**? Demonstrate why or why not.

57. The file **P02_40.xlsx** consists of the monthly retail sales levels of U.S. gasoline service stations.
    a. Is there a seasonal pattern in these data? If so, how do you explain this seasonal pattern? Also, if necessary, deseasonalize these data using the ratio-to-moving-average method.
    b. Forecast this time series for the first 4 months of the next year using the most appropriate method for these data. Defend your choice of forecasting method.

58. The number of employees on the payroll at a food processing plant is recorded at the start of each month. These data are provided in the file **P13_03.xlsx**.
   a. Is there a seasonal pattern in these data? If so, how do you explain this seasonal pattern? Also, if necessary, deseasonalize these data using the ratio-to-moving-average method.
   b. Forecast this time series for the first 4 months of the next year using the most appropriate method. Defend your choice of forecasting method.

59. Consider the file **P02_37.xlsx**, which contains total monthly U.S. retail sales data. Compare the effectiveness of Winters' method with that of the ratio-to-moving-average method in deseasonalizing this time series. Using the deseasonalized time series generated by each of these two methods, forecast U.S. retail sales with the most appropriate method. Defend your choice of forecasting method.

60. Suppose that a time series consisting of 6 years (2000–2005) of quarterly data exhibits definite seasonality. In fact, assume that the seasonal indexes turn out to be 0.75, 1.45, 1.25, and 0.55.
   a. If the last four observations of the series (the four quarters of 2005) are 2502, 4872, 4269, and 1924, calculate the deseasonalized values for the four quarters of 2005.
   b. Suppose that a plot of the deseasonalized series shows an upward linear trend, except for some random noise. Therefore, a linear regression of this series versus time is estimated, and it produces the equation

$$\text{Predicted deseasonalized value} = 2250 + 51\text{Quarter}$$

   Here the time variable "Quarter" is coded so that Quarter = 1 corresponds to first quarter 2000, Quarter = 24 corresponds to fourth quarter 2005, and the others fall in between. Forecast the actual (not deseasonalized) values for the four quarters of 2006.

61. The file **P13_61.xlsx** contains monthly data on federal receipts of taxes. There are two variables: IndTax (taxes from individuals) and CorpTax (corporate taxes). For this problem, work only with IndTax.
   a. What evidence is there that seasonality is important in this series? Find seasonal indexes (by any method you like) and state briefly what they mean.
   b. Forecast the next 12 months by using a linear trend on the seasonally adjusted data. State briefly the steps you use to get this type of forecast, give the final RMSE, MAPE, and forecast for the next month. Then show numerically how you could replicate this forecast (i.e., show on paper how the package uses its estimated model to get the next month forecast). (Substitute in specific numbers, and do the arithmetic.)

62. Quarterly sales for a department store over a 6-year period are given in the file **P13_62.xlsx**.
   a. Use multiple regression to develop a model that can be used to predict future quarterly sales. (*Hint*: Use dummy variables and an explanatory variable for the quarter number, 1–24.)
   b. Letting Y_t be the sales during quarter number t, discuss how to fit the following model to these data.

$$Y_t = \alpha\beta^t{}_1\beta_2{}^{X_1}\beta_3{}^{X_2}\beta_4{}^{X_3}$$

   Here $X_1 = 1$ if t is a first quarter, 0 otherwise; $X_2 = 1$ if t is a second quarter, 0 otherwise; and $X_3 = 1$ if t is a third quarter, 0 otherwise.
   a. Interpret the answer to part **b**.
   b. Which model appears to yield better predictions for sales, the one in part **a** or the one in part **b**?

63. Confederate Express Service is attempting to determine how its shipping costs for a month depend on the number of units shipped during a month. The number of units shipped and total shipping cost for the last 15 months are given in the file **P13_63.xlsx**.
   a. Determine a relationship between units shipped and monthly shipping cost.
   b. Plot the errors for the predictions in order of time sequence. Is there any unusual pattern?
   c. We have been told that there was a trucking strike during months 11 through 15, and we believe that this might have influenced shipping costs. How could the answer to part **a** be modified to account for the effect of the strike? After accounting for this effect, does the unusual pattern in part **b** disappear?

Level B

64. In our discussion of Winters' method, a monthly seasonality of 0.80 for January, say, means that during January, air conditioner (AC) sales are expected to be 80% of the sales during an average month. An alternative approach to modeling seasonality, called an **additive model**, is to let the seasonality factor for each month represent how far above average AC sales will be during the current month. For instance, if $S_{\text{Jan}} = -50$, then AC sales during January are expected to be 50 fewer than AC sales during an average month. (This is 50 ACs, not 50%.) If $S_{\text{July}} = 90$, then AC sales during July are expected to be 90 more than AC sales during an average month. Let

   S_t = Seasonality for month t after observing month t demand

   L_t = Estimate of level after observing month t demand

   T_t = Estimate of trend after observing month t demand

   Then the Winters' method equations given in the text should be modified as follows:

$$L_t = \alpha(I) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$

$S_t = \gamma(II) + (1 - \gamma)S_{t-12}$

a. What should (*I*) and (*II*) be?

b. Suppose that month 13 is January, $L_{12} = 30$, $T_{12} = -3$, $S_1 = -50$, and $S_2 = -20$. Let $\alpha = \gamma = \beta = 0.5$. Suppose 12 ACs are sold during month 13. At the end of month 13, what is the prediction for AC sales during month 14 using this additive model?

65. Winters' method assumes a multiplicative seasonality but an additive trend. For example, a trend of 5 means that the level will increase by 5 units per period. Suppose that there is actually a *multiplicative* trend. Then (ignoring seasonality) if the current estimate of the level is 50 and the current estimate of the trend is 1.2, we would predict demand to increase by 20% per period. So we would forecast the next period's demand to be 50(1.2) and forecast the demand 2 periods in the future to be $50(1.2)^2$. If we want to use a multiplicative trend in Winters' method, we should use the following equations (assuming a period is a month):

$$L_t = \alpha\left(\frac{Y_t}{S_{t-12}}\right) + (1 - \alpha)(I)$$

$$T_t = \beta(II) + (1 - \beta)\,T_{t-1}$$

$$S_t = \gamma\left(\frac{Y_t}{L_t}\right) + (1 - \gamma)S_{t-12}$$

a. What should (*I*) and (*II*) be?

b. Suppose that we are working with monthly data and month 12 is December, month 13 is January, and so on. Also suppose that $L_{12} = 100$, $T_{12} = 1.2$, $S_1 = 0.90$, $S_2 = 0.70$, and $S_3 = 0.95$. Also, suppose $Y_{13} = 200$. At the end of month 13, what is the prediction for Y_{15} using $\alpha = \beta = \gamma = 0.5$ and a multiplicative trend?

66. Consider the file **P02_37.xlsx**, which contains total monthly U.S. retail sales data. Does a regression approach for estimating seasonality provide forecasts that are as accurate as those provided by (a) Winters' method and (b) the ratio-to-moving-average method? Compare the summary measures of forecast errors associated with each method for deseasonalizing the given time series. Summarize your findings after performing these comparisons.

67 The file **P02_39.xlsx** contains monthly time series data for total U.S. retail sales of building materials (which includes retail sales of building materials, hardware and garden supply stores, and mobile home dealers). Does a regression approach for estimating seasonality provide forecasts that are as accurate as those provided by (a) Winters' method and (b) the ratio-to-moving-average method? Compare the summary measures of forecast errors associated with each method for deseasonalizing the given time series. Summarize your findings after performing these comparisons.

13.10 CONCLUSION

We have covered a lot of ground in this chapter. Because forecasting is such an important activity in business, it has received a tremendous amount of attention by both academicians and practitioners. All of the methods discussed in this chapter—and more—are actually used, often on a day-to-day basis. There is really no point in arguing which of these methods is best. All of them have their strengths and weaknesses. The most important point is that when they are applied properly, they have all been found to be useful in real business situations.

Summary of Key Terms

Term	Explanation	Excel	Page	Equation Number
Extrapolation methods	Forecasting methods where only past values of a variable (and possibly time itself) are used to forecast future values		715	
Causal (or econometric) methods	Forecasting methods based on regression, where other time series variables are used as explanatory variables		716	
Trend	A systematic increase or decrease of a time series variable through time		718	

(*continued*)

Term	Explanation	Excel	Page	Equation Number
Seasonality	A regular pattern of ups and downs based on the season of the year, typically months or quarters		718	
Cyclic variation	An irregular pattern of ups and downs caused by business cycles		718	
Noise (or random variation)	The unpredictable ups and downs of a time series variable		719	
Mean absolute error (MAE)	The average of the absolute forecast errors	StatTools/ Time Series & Forecasting/Forecast	721	13.2
Root mean square error (RMSE)	The square root of the average of the squared forecast errors	StatTools/ Time Series & Forecasting/Forecast	721	13.3
Mean absolute percentage error (MAPE)	The average of the absolute percentage forecast errors	StatTools/ Time Series & Forecasting/Forecast	721	13.4
Runs test	A test of whether the forecast errors are random noise	StatTools/ Time Series & Forecasting/ Runs Test for Randomness	724	
Autocorrelations of residuals	Correlations of forecast errors with themselves, used to check whether they are random noise	StatTools/ Time Series & Forecasting/ Autocorrelation	726	
Correlogram	A bar chart of autocorrelations at different lags	StatTools/ Time Series & Forecasting/ Autocorrelation	727	
Linear trend model	A regression model where a time series variable changes by a constant amount each time period	StatTools/ Regression & Classification/ Regression	729	13.6
Exponential trend model	A regression model where a time series variable changes by a constant percentage each time period	StatTools/ Regression & Classification/ Regression	732	13.7
Random walk model	A model indicating that the differences between adjacent observations of a time series variable are constant except for random noise		738	13.9–13.11
Autoregression model	A regression model where the only explanatory variables are lagged values of the dependent variable	StatTools/ Regression & Classification/ Regression	741	
Moving averages model	A forecasting model where the average of several past observations is used to forecast the next observation	StatTools/ Time Series & Forecasting/ Forecast	746	

(*continued*)

Term	Explanation	Excel	Page	Equation Number
Span	The number of observations in each average of a moving averages model	StatTools/ Time Series & Forecasting/ Forecast	746	
Exponential smoothing models	A class of forecasting models where forecasts are based on weighted averages of previous observations, giving more weight to more recent observations	StatTools/ Time Series & Forecasting/ Forecast	753	
Smoothing constants	Constants between 0 and 1 that prescribe the weight attached to previous observations and hence the smoothness of the series of forecasts	StatTools/ Time Series & Forecasting/ Forecast	753	
Simple exponential smoothing	An exponential smoothing model useful for time series with no prominent trend or seasonality	StatTools/ Time Series & Forecasting/ Forecast	753	13.12–13.15
Holt's method	An exponential smoothing model useful for time series with trend but no seasonality	StatTools/ Time Series & Forecasting/ Forecast	753	13.16–13.18
Winters' method	An exponential smoothing model useful for time series with seasonality (and possibly trend)	StatTools/ Time Series & Forecasting/ Forecast	753	13.19–13.22
Deseasonalizing	A method for removing the seasonal component from time series data	StatTools/ Time Series & Forecasting/ Forecast	768	
Ratio-to-moving-averages method	A method for deseasonalizing a time series, so that some other method can then be used to forecast the deseasonalized series	StatTools/ Time Series & Forecasting/ Forecast	768	
Dummy variables for seasonality	A regression-based method for forecasting seasonality, where dummy variables are used for the seasons	StatTools/ Regression & Classification/ Regression	772	

PROBLEMS

Conceptual Exercises

C.1. "A truly random series will likely have a very small number of runs." Is this statement true or false? Explain your choice.

C.2. Distinguish between a *correlation* and an *autocorrelation*. How are these measures similar? How are these measures different?

C.3. What is the relationship between the random walk model and the autoregression model?

C.4. Under what conditions would you prefer a simple exponential smoothing model to the moving averages method for forecasting a time series?

C.5. Is it more appropriate to use an *additive* or a *multiplicative* model to forecast seasonal data? Summarize the difference(s) between these two types of seasonal models.

Level A

68. The file **P13_68.xlsx** contains quarterly revenues of Toys "R" Us. Discuss the seasonal and trend components of the growth of Toys "R" Us revenues. Also, use any reasonable forecasting method to forecast quarterly revenues for the next year. Explain your choice of forecasting method.

69. The file **P13_69.xlsx** contains quarterly revenues and earnings per share (EPS) for the following companies: Mattel, McDonald's, Eli Lilly, General Motors, Microsoft, AT&T, Nike, GE, Coca-Cola, and Ford.
    a. For each company, use a regression model with trend and seasonal components to forecast revenues and EPS.
    b. For each company, use Winters' method to forecast revenues and EPS.
    c. For each company, which method appears to be more accurate?

70. The file **P13_70.xlsx** contains the sales in (millions of dollars) for Sun Microsystems.
    a. Use these data to predict the company's sales for the next 2 years. You need consider only a linear and exponential trend, but you should justify the equation you choose.
    b. In words, how do your predictions of sales increase from year to year?
    c. Are there any outliers?

71. The file **P13_71.xlsx** contains the sales in (millions of dollars) for Procter & Gamble.
    a. Use these data to predict Procter & Gamble sales for the next 2 years. You need consider only a linear and exponential trend, but you should justify the equation you choose.
    b. Use your answer from part **a** to explain how your predictions of Procter & Gamble sales increase from year to year.
    c. Are there any outliers?
    d. We can be approximately 95% sure that Procter & Gamble sales in the year following next year will be between what two values?

72. The file **P13_72.xlsx** lists the sales of Nike. Forecast sales in the next 2 years with a linear or exponential trend. Are there any outliers in your predictions for the observed period?

73. The file **P12_44.xlsx** contains data on pork sales. Price is in dollars per hundred pounds sold, quantity sold is in billions of pounds, per capita income is in dollars, U.S. population is in millions, and GNP is in billions of dollars.
    a. Use these data to develop a regression equation that could be used to predict the quantity of pork sold during future periods. Is autocorrelation, heteroscedasticity, or multicollinearity a problem?
    b. Suppose that during each of the next two quarters, price is \$45, U.S. population is 240, GNP is 2620, and per capita income is \$10,000. (All of these are expressed in the units described above.) Predict the quantity of pork sold during each of the next 2 quarters.
    c. We expect our prediction of pork sales to be accurate within what value 68% of the time?
    d. Use Winters' method to develop a forecast of pork sales during the next 2 quarters.

74. The file **P13_74.xlsx** contains data on a motel chain's revenue and advertising.
    a. Use these data and multiple regression to make predictions of the motel chain's revenues during the next 4 quarters. Assume that advertising during each of the next 4 quarters is \$50,000. (*Hint*: Try using advertising, lagged by 1 quarter, as an explanatory variable.)
    b. Use simple exponential smoothing to make predictions for the motel chain's revenues during the next 4 quarters.
    c. Use Holt's method to make forecasts for the motel chain's revenues during the next 4 quarters.
    d. Use Winters' method to determine predictions for the motel chain's revenues during the next 4 quarters.
    e. Which of these forecasting methods would you expect to be the most reliable for these data?

75. The file **P13_75.xlsx** contains data on monthly U.S. housing sales (in thousands of houses).
    a. Using Winters' method, find values of α, β, and γ that yield an RMSE as small as possible.
    b. Although we have not discussed autocorrelation for smoothing methods, good forecasts derived from smoothing methods should exhibit no autocorrelation. Do the forecast errors for this problem exhibit autocorrelation?
    c. At the end of the observed period, what is the forecast of housing sales during the next few months?

76. Let Y_t be the sales during month t (in thousands of dollars) for a photography studio, and let P_t be the price charged for portraits during month t. The data are in the file **P12_45.xlsx**. Use regression to fit the following model to these data:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 P_t + \epsilon_t$$

    This equation indicates that last month's sales and the current month's price are explanatory variables. The last term, ϵ_t, is an error term.
    a. If the price of a portrait during month 21 is \$10, what would we predict for sales in month 21?
    b. Does there appear to be a problem with autocorrelation, heteroscedasticity, or multicollinearity?

77. The file **P13_77.xlsx** gives quarterly auto sales, GNP, interest rates and unemployment rates.

a. With all but the most recent 2 years of data, use regression to forecast auto sales. Interpret the coefficients in your final equation.
b. Use all but the most recent 2 years of data to develop an exponential smoothing model to forecast future auto sales.
c. To *validate* your model, determine which model does the best job of forecasting for the most recent 2 years of data. It is usually recommended to hold out some of your data to validate any forecast model. This helps avoid "overfitting."

Level B

78. The file **P13_78.xlsx** contains monthly time series data on corporate bond yields. These are averages of daily figures, and each is expressed as an annual rate. The variables are:
- YieldAAA: average yield on AAA bonds
- YieldBAA: average yield on BAA bonds

If you examine either Yield variable, you will notice that the autocorrelations of the series are not only large for many lags, but that the lag 1 autocorrelation of the *differences* is significant. This is very common. It means that the series is not a random walk and that it is probably possible to provide a better forecast than the "naive" forecast from the random walk model. Here is the idea. The large lag 1 autocorrelation of the differences means that the differences are related to the first lag of the differences. This relationship can be estimated by creating the difference variable and a lag of it, then regressing the former on the latter, and finally using this information to forecast the original Yield variable.

a. Verify that the autocorrelations are as described, and form the difference variable and the first lag of it. Call these DYield and L1DYield (where D is for difference, L1 is for first lag).
b. Run a regression with DYield as the dependent variable and L1DYield as the single explanatory variable. In terms of the original variable Yield, this equation can be written as

$$\text{Yield}_t - \text{Yield}_{t-1} = a + b(\text{Yield}_{t-1} - \text{Yield}_{t-2})$$

Solving for Yield_t is equivalent to the following equation that can be used for forecasting:

$$\text{Yield}_t = a + (1 + b)\text{Yield}_{t-1} - b\text{Yield}_{t-2}$$

Try it—that is, try forecasting the next month from the known last 2 months' values. How might you forecast values 2 or 3 months from the last observed month? (*Hint*: If you do not have an *observed* value to use in the right side of the equation, use a forecasted value.)
c. The autocorrelation structure led us to the equation in part **b**. That is, the autocorrelations of the original series took a long time to die down, so we looked at the autocorrelations of the differences, and the large spike at lag 1 led to regressing DYield on L1DYield. In turn, this led ultimately to an equation for Yield_t in terms of its first two lags. Now see what you would have obtained if you had tried regressing Yield_t on its first two lags in the first place—that is, if you had used regression to estimate the equation

$$\text{Yield}_t = a + b_1\text{Yield}_{t-1} + b_2\text{Yield}_{t-2}$$

When you use multiple regression to estimate this equation, do you get the same equation as in part **b**?

79. The file **P13_79.xlsx** contains 5 years of monthly data for a particular company. The first variable is Time (1–60). The second variable, Sales1, has data on sales of a product. Note that Sales1 increases linearly throughout the period, with only a minor amount of "noise." (The third variable, Sales2, is discussed and used in the next problem.) For this problem use the Sales1 variable to see how the following forecasting methods are able to track a linear trend.
a. Forecast this series with the moving average method with various spans such as 3, 6, and 12. What can you conclude?
b. Forecast this series with simple exponential smoothing with various smoothing constants such as 0.1, 0.3, 0.5, and 0.7. What can you conclude?
c. Now repeat part **b** with Holt's exponential smoothing method, again for various smoothing constants. Can you do significantly better than in parts **a** and **b**?
d. What can you conclude from your findings in parts **a**, **b**, and **c** about forecasting this type of series?

80. The Sales2 variable in the file from the previous problem was created from the Sales1 variable by multiplying by monthly seasonal factors. Basically, the summer months are high and the winter months are low. This might represent the sales of a product that has a linear trend and seasonality.
a. Repeat parts **a**, **b**, and **c** from the previous problem to see how well these forecasting methods can deal with trend *and* seasonality.
b. Now use Winters' method, with various values of the three smoothing constants, to forecast the series. Can you do much better? Which smoothing constants work well?
c. Use the ratio-to-moving-average method, where you first do the seasonal decomposition and then forecast (by any appropriate method) the deseasonalized series. Does this do as well as, or better than, Winters' method?
d. What can you conclude from your findings in parts **a**, **b**, and **c** about forecasting this type of series?

81. The file **P13_81.xlsx** contains monthly time series data on federal expenditures in various categories. All values are in billions of current dollars. The variables are:

- Defense: expenditures on national defense
- Science: expenditures on science, space, and technology
- Energy: expenditures on energy
- Environ: expenditures on natural resources and environment
- Trans: expenditures on transportation

Analyze the Science variable by (a) simple exponential smoothing, (b) Holt's method, (c) simple exponential smoothing on the trend-adjusted data (the residuals from regressing linearly versus time), and (d) moving averages on the adjusted or unadjusted data. Experiment with the smoothing constants [or span in (d)], or use the optimize feature. Do any of these methods produce significantly better fits than the others as measured by RMSE or MAPE?

82. The data in the file **P13_82.xlsx** represent annual changes in the average surface air temperature of the earth. (The source doesn't say exactly how this was measured.) A look at the time series shows a gradual upward trend, starting with negative values and ending with (mostly) positive values. This might be used to support the theory of global warming.

a. Is this series a random walk? Explain.
b. Regardless of your answer in part a, use a random walk model to forecast the next value of the series. What is your forecast, and what is an approximate 95% forecast interval?
c. Forecast the series in three ways: (i) simple exponential smoothing ($\alpha = 0.35$), (ii) Holt's method ($\alpha = 0.5$, $\beta = 0.1$), and (iii) simple exponential smoothing ($\alpha = 0.3$) on trend-adjusted data, that is, the residuals from regressing linearly versus time. (These smoothing constants are close to "optimal.") For each of these, list the MAPE, the RMSE, and the forecast for next year. Also, comment on any "problems" with forecast errors from any of these three approaches. Finally, compare the "qualitative" features of the three forecasts (for example, how do their short-run or longer-run forecasts differ?). Is any one of the methods clearly superior to the others?
d. Does your analysis predict convincingly that global warming would occur during the observed years? Explain.

CASE 13.1 Arrivals at the Credit Union

The Indiana University Credit Union Eastland Plaza Branch was having trouble getting the correct staffing levels to match customer arrival patterns. On some days, the number of tellers was too high relative to the customer traffic, so that tellers were often idle. On other days, the opposite occurred. Long customer waiting lines formed because the relatively few tellers could not keep up with the number of customers. The credit union manager, James Chilton, knew that there was a problem, but he had little of the quantitative training he believed would be necessary to find a better staffing solution. James figured that the problem could be broken down into three parts. First, he needed a reliable forecast of each day's number of customer arrivals. Second, he needed to translate these forecasts into staffing levels that would make an adequate trade-off between teller idleness and customer waiting. Third, he needed to translate these staffing levels into individual teller work assignments—who should come to work when.

The last two parts of the problem require analysis tools (queueing and scheduling) that we have not covered. However, you can help James with the first part—forecasting. The file **Credit Union Arrivals.xlsx** lists the number of customers entering this credit union branch each day of the past year. It also lists other information: the day of the week, whether the day was a staff or faculty payday, and whether the day was the day before or after a holiday. Use this data set to develop one or more forecasting models that James could use to help solve his problem. Based on your model(s), make any recommendations about staffing that appear reasonable. ■

CASE 13.2 Forecasting Weekly Sales at Amanta

Amanta Appliances sells two styles of refrigerators at more than 50 locations in the Midwest. The first style is a relatively expensive model, whereas the second is a standard, less expensive model. Although weekly demand for these two products is fairly stable from week to week, there is enough variation to concern management at Amanta. There have been relatively unsophisticated attempts to forecast weekly demand, but they haven't been very successful. Sometimes demand (and the corresponding sales) are lower than forecasted, so that inventory costs are high. Other times the forecasts are too low. When this happens and on-hand inventory is not sufficient to meet customer demand, Amanta requires expedited shipments to keep customers happy—and this nearly wipes out Amanta's profit margin on the expedited units.[6] Profits at Amanta would almost certainly increase if demand could be forecasted more accurately.

Data on weekly sales of both products appear in the file **Amanta Sales.xlsx**. A time series chart of the two sales variables indicates what Amanta management expected—namely, there is no evidence of any upward or downward trends or of any seasonality. In fact, it might appear that each series is an unpredictable sequence of random ups and downs. But is this really true? Is it possible to forecast either series, with some degree of accuracy, with an extrapolation method (where only past values of *that* series are used to forecast current and future values)? What method appears to be best? How accurate is it? Also, is it possible, when trying to forecast sales of one product, to somehow incorporate current or past sales of the *other* product in the forecast model? After all, these products might be "substitute" products, where high sales of one go with low sales of the other, or they might be complementary products, where sales of the two products tend to move in the *same* direction. ■

[6]Because Amanta uses expediting when necessary, its sales each week are equal to its customer demands. Therefore, we use the terms *demand* and *sales* interchangeably.