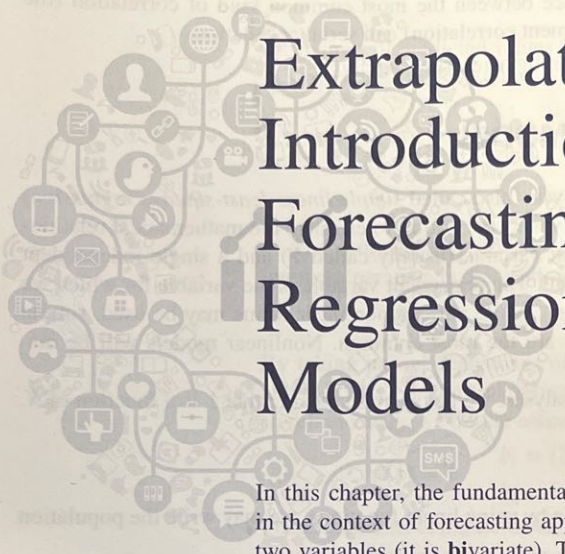
Chapter **Four**

Extrapolation 2. Introduction to Forecasting with Regression Trend Models



©VLADGRIN/Getty Images

In this chapter, the fundamentals of bivariate regression analysis are presented in the context of forecasting applications. A bivariate regression model has just two variables (it is **bi**variate). The variables are commonly designated as Y and X. The Y variable is called the **dependent variable**. The X variable is called the **independent** variable. So, Y **depends on the value of** X. Or, Y is a function of X.

In this chapter, regression models are developed for jewelry store sales and disposable personal income), based on quarterly data. These regression models are then used to make forecasts of each series. As you might expect, jewelry store sales are quite seasonal. You will see how seasonal data can be forecast with a bivariate regression.

At the end of the chapter, we return to our continuing examples of forecasting total new houses sold and to the continuing The Gap case study. In both these situations, the variables being forecast have a seasonal pattern. You will see again how the seasonal pattern can be handled in a bivariate regression. In Chapter 5, you will learn another way to deal with seasonality using regression analysis.

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

1. Explain why it is important to look at data in a graph rather than only in a table.
2. Describe the type of data patterns for which a linear regression trend forecast would be appropriate.

3. Explain how a seasonal data set can be forecast with a linear regression trend.
4. Discuss the four steps that should be used to evaluate a linear regression model.
5. Explain the difference between a trend model and a causal model..
6. Explain the difference between the most common kind of correlation (the Pearson product moment correlation) and serial correlation.
7. Explain what is meant by heteroscedasticity.

THE BIVARIATE REGRESSION MODEL

Bivariate regression analysis (also called *simple linear least-squares regression*) is a statistical tool that gives us the ability to estimate the mathematical relationship between a dependent variable (usually called Y) and a single independent variable (usually called X).[1] The dependent variable is the variable for which we want to develop a forecast. While various nonlinear forms may be used, simple linear regression models are the most common. Nonlinear models will be discussed in Chapter 5.

In using regression analyses, we begin by supposing that Y is a function of X. That is:

$$Y = f(X)$$

Since we most often begin by using linear functions, we may write the population regression model as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 represents the intercept of the regression line on the vertical (or Y) axis and β_1 is the slope of the regression line. Thus, β_1 tells us the rate of change in Y per unit change in X. The intercept (β_0) is the value that the dependent variable would have if $X = 0$. While this is a correct interpretation from an algebraic perspective, such an interpretation is often not valid in applications, since a value of $X = 0$ is frequently not in the relevant range of observations on X. The ε in this model represents an error term. That is, every Y is not likely to be predicted exactly from the values of β_0 and $\beta_1 X$. The resulting error is ε.

We would like to estimate values of β_0 and β_1 such that the resulting equation best fits the data. To do so, we need to decide on a criterion against which the fit of the estimated model can be evaluated. The most common such rule is called the *ordinary least-squares* (OLS) criterion. This rule says that the best model is the one that minimizes the sum of the squared error terms.

The unobserved model that describes the whole population of data is expressed as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

[1] For a more detailed discussion of the regression model, including underlying assumptions, see Bruce Bowerman, Richard T. O'Connell, and Emily Murphree, *Business Statistics in Practice*, 8th ed. (New York, NY. McGraw-Hill Education, 2017).

These values of the intercept (β_0) and slope (β_1) are population parameters that are typically estimated using sample data. The corresponding sample statistics are b_0 and b_1. The estimated regression model is expressed as

$$\hat{Y} = b_0 + b_1 X$$

Deviations of predicted values ($\hat{Y}$) from the actual values of Y are called *residuals* or *errors* and are denoted by e, where

$$e = Y - \hat{Y}$$

or,

$$e = Y - b_0 - b_1 X$$

The ordinary least-squares method seeks to find estimates of the slope and intercept parameters that minimize the sum of squared residuals:

$$Minimize \Sigma e^2 = \Sigma (Y - b_0 - b_1 X)^2$$

By taking partial derivatives of the sum of squared residuals with respect to b_0 and b_1, setting the partial derivatives equal to zero, and solving the two equations simultaneously, we obtain estimating formulas:

$$b_1 = (\Sigma XY - n\bar{X}\bar{Y})/(\Sigma X^2 - n\bar{X}^2)$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

These formulas could be used to calculate b_0 and b_1 by hand. However, even for simple regression, a computer program is normally used for such calculations.

VISUALIZATION OF DATA: AN IMPORTANT STEP IN REGRESSION ANALYSIS

There was a time when regression lines were estimated in a rather *ad hoc* manner, based solely on an analyst's visual interpretation of the data. The analyst would plot the data by hand and would "eyeball" the resulting scatter of points to determine the position of a straight line that was believed to "best" represent the general relationship between Y and X. Such a straight line was then drawn through the scatterplot, and by selecting two points from the line, its algebraic equation was calculated (i.e., values for b_0 and b_1 were estimated). One obvious problem with such a procedure is that different analysts would almost surely come up with differing estimates of b_0 and b_1.

Today, it is doubtful that anyone would take this approach to estimating a regression equation. Modern computer technology makes it very easy to obtain the OLS equation without ever looking at the data. This equation is best, according to the ordinary least-squares criterion, and numerous evaluative statistics can be simultaneously determined. Every analyst obtains precisely the same results, and those results are easily replicated. Thus, it may appear that computer-based regression analysis is a clearly superior method. However, something is lost. Analysts may

TABLE 4.1 Four Dissimilar Data Sets with Similar Regression Results (c4t1)

Source: Anscombe, F. J. "Graphs in Statistical Analysis," *American Statistician* 27, February 1973, 17–21, as reported in Edward R. Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press, 1983, 13.

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

just enter data into Excel or some statistical software, issue appropriate commands, get the corresponding statistical results, and run off to apply the model in some decision-based context such as forecasting. In the process, they would never have *looked* at the data. Such blind attention to statistical estimates can be dangerous.

To illustrate this point, consider the four data sets in Table 4.1. For all four of the data sets in Table 4.1, the calculated regression results show an OLS equation of:

$$\hat{Y} = 3 + 0.5X$$

It might also be noted that the mean of the X's is 9.0 and the mean of the Y's is 7.5 in all four cases. The standard deviation is 3.32 for all of the X variables and 2.03 for all of the Y variables. Similarly, the correlation for each pair of X and Y variables is 0.82.[2]

From these results, an analyst who looks only at these summary statistics would likely conclude that the four data sets are identical or, at the very least, quite similar. But, oh, how wrong this conclusion would be. If you take the time to prepare a scattergram of each of the four data sets, dramatic differences become apparent. In Figure 4.1, we have plotted each XY pair in a separate plot, along with the corresponding OLS regression lines (all four of the regression lines have the same equation: $\hat{Y} = 3 + 0.5X$).

It is important to *look* at the data before plunging into data analysis and the selection of an appropriate set of forecasting techniques.

Visualization of these data allows us to see stark differences that would not be apparent from the descriptive statistics we have reviewed. The regression line is most clearly inappropriate for the data in the lower-right plot. The lower-left plot has, with the exception of one outlier, a perfectly linear relationship between Y and X, which is not so clear without visual inspection of the data. The upper-right plot of data suggests that a nonlinear model would fit the data better than a linear

[2] Many statistical diagnostics on the regression equations, which we will cover later in this chapter, are also equal. These include standard errors of the regression, t-ratios for the coefficients, R-squared, and the regression sum of squares. Statistics related to the evaluation of residuals, such as the Durbin-Watson statistic, show some differences.

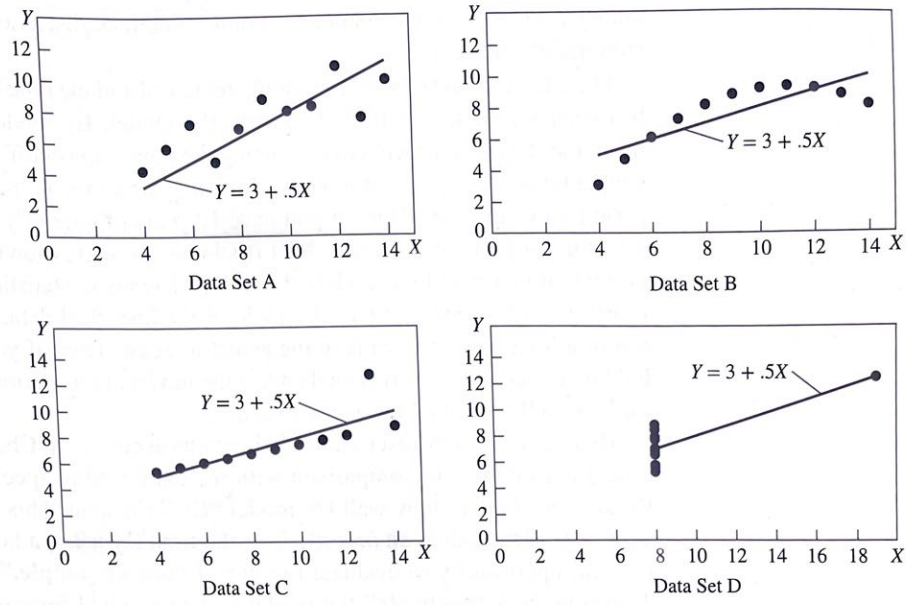


FIGURE 4.1
Scatterplots of Four *XY* Data Sets That Have Very Similar Statistical Properties but Are Visually Quite Different (c4f1)
For Each of the Data Sets, the OLS Regression Equation Is $Y = 3 + 0.5X$.

function. Only the upper-left plot suggests a data set that is a good candidate for a linear regression model. Visually, these data sets are quite dissimilar, even though they have some very similar statistical properties.

Forecasters can benefit from this example.

A PROCESS FOR REGRESSION FORECASTING

It is useful to have a plan at hand when approaching any task. And so it is with developing a regression-based forecast. In this section, we suggest one such plan, or process, that helps to organize the task of preparing a regression forecast. What we say here is not separate from the forecast process discussed in Chapter 2. Rather, it complements that process, especially data considerations, model selection, model evaluation, and forecast preparation.

The forecaster should utilize graphic techniques to inspect the data, looking especially for trend, seasonal, and cyclical components, as well as for outliers.

We begin with data considerations, which become somewhat more complex for regression models. Not only do we need to pay attention to the dependent variable, the series to be forecasted, but we must also consider the independent variable(s) that will drive the regression forecast. We should utilize graphic techniques to inspect the data, looking especially for trend, seasonal, and cyclical components, as well as for outliers. This will help in determining what type of regression model may be most appropriate (e.g., linear versus nonlinear, or trend versus causal).

Next we must make a forecast of the independent variable(s). This becomes a separate, yet related, forecasting effort. Each potential independent variable should be forecast using a method that is appropriate to that particular series,

taking into account the model-selection guidelines discussed in Chapter 2 and summarized in Table 2.1.

Once the data have been thoroughly reviewed and the type of regression model has been selected, it is time to specify the model. By model specification, we mean the statistical process of estimating the regression coefficients (b_0 and b_1, in simple bivariate regression models). In doing so, we recommend using a holdout period for evaluation. Thus, if you have 10 years of quarterly data ($n = 40$), you might use 9 years of data ($n = 36$) to estimate the regression coefficients. Initial evaluation of regression models (based on diagnostic statistics we will discuss shortly) can be done on this subsample of the historical data. However, the real test of a forecasting model is in the actual forecast. Thus, if you have set aside a holdout period of data, you can then test the model in this period to get a truer feel for how well the model meets your needs.

This relates to our discussion of fit versus accuracy in Chapter 2. When the model is evaluated in comparison with the data used in specifying the model, we are determining how well the model **"fits"** the data. This is a retrospective approach, often called an ***in-sample*** evaluation. By using a holdout period, we have an opportunity to evaluate the model **"out of sample."** That is, we can determine how **"accurate"** the model is for an actual forecast horizon. After an evaluation of fit and accuracy, a forecaster should respecify the best of the models using the entire span of data that are available. The newly specified model is then used to forecast beyond the frontier of what is known at the time of the forecast.

FORECASTING WITH A SIMPLE LINEAR TREND[3]

It is sometimes possible to make reasonably good forecasts on the basis of a simple linear time trend. To do so, we set up a time index (T) to use as the independent or X variable in the basic regression model, where T is usually set equal to 1 for the first observation and increased by 1 for each subsequent observation. The regression model is then:

$$\hat{Y} = b_0 + b_1(T)$$

where Y is the series we wish to forecast.

To illustrate this process, consider the data in Table 4.2. DPI is disposable personal income in billions of dollars on a quarterly basis. In the "Date" column, the months represent the end month of each calendar quarter. The data are for 2010 through 2016. Only data though December 2015 will be used to develop a forecast so that we can evaluate it against actual data for the four quarters of 2016.

[3] Throughout this chapter, you may find some situations in which the standard calculations that we show do not match exactly with the ForecastX results. This is because, at times, they invoke proprietary alterations from the standard calculations. The results are usually very close but may not match perfectly with "hand" calculations or those done in Excel.

TABLE 4.2
Disposable Personal Income In Billions Of Dollars. Quarterly From 2010 Through 2016 (c4t2&f2)

Date	DPI (B$)	Time Index
Mar-10	11,041	1
Jun-10	11,198	2
Sep-10	11,287	3
Dec-10	11,426	4
Mar-11	11,652	5
Jun-11	11,752	6
Sep-11	11,877	7
Dec-11	11,925	8
Mar-12	12,190	9
Jun-12	12,321	10
Sep-12	12,355	11
Dec-12	12,748	12
Mar-13	12,259	13
Jun-13	12,336	14
Sep-13	12,454	15
Dec-13	12,534	16
Mar-14	12,736	17
Jun-14	12,962	18
Sep-14	13,127	19
Dec-14	13,265	20
Mar-15	13,277	21
Jun-15	13,465	22
Sep-15	13,612	23
Dec-15	13,726	24
Mar-16	13,807	25
Jun-16	13,977	26
Sep-16	14,129	27
Dec-16	14,270	28

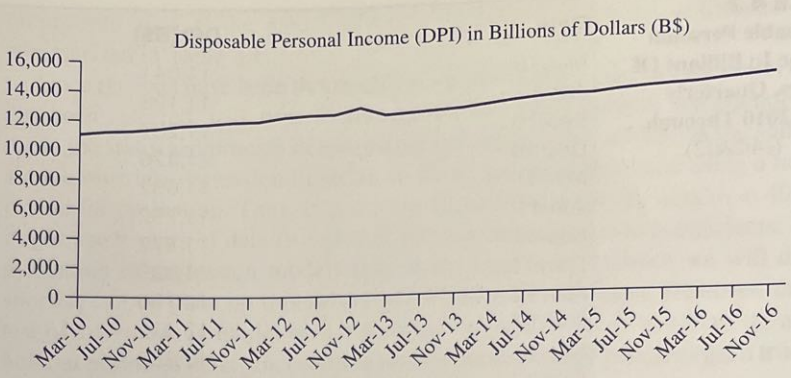
Disposable personal income is an important economic series, since income is an important determinant for many kinds of sales. The linear time-trend model for DPI is:

$$\widehat{\text{DPI}} = b_0 + b_1(T)$$

You see in Table 4.2 that T (time) equals 1 for the first quarter of 2010 and 28 for the fourth quarter of 2016.

It is usually a good idea to look at data such as those given in Table 4.2 in graphic form before beginning to do any regression analysis. A visual inspection of the data can be helpful in deciding whether a linear or nonlinear model would be most appropriate. A graph of DPI versus T is shown in Figure 4.2. From this graph, you can get a good feel for how this important measure of income has increased over the period presented. All observations do not fall on a single straight line. However, it does appear that a linear trend line would fit the data well. The positive trend to DPI is more easily seen in the graphic form of Figure 4.2 than in the tabular form of Table 4.2.

FIGURE 4.2
Graph of Disposable Personal Income (DPI) Over Time While DPI Does not Follow A Perfectly Linear Path, it Does Follow a Trend That is Very Close to Linear. (c4t2&f2)



Suppose that you are asked to forecast DPI for the four quarters of 2016, using a simple linear trend, based only on data from 2010 through 2015. The first thing you would do is to use the linear regression part of your regression software to provide the estimates of b_0 and b_1 for the following model:

$$\text{DPI} = b_0 + b_1(T)$$

The regression results from ForecastX and from Excel are shown at the bottom of Figure 4.3. From those results we see that the intercept (b_0) is 11,042.86 and that the coefficient for T (b_1, or the slope) is 108.32. Thus, the regression forecast model may be written as:

$$\widehat{\text{DPI}} = 11{,}042.86 + 108.32\left(T\right)$$

The slope term in this model tells us that, on average, disposable personal income increased by $108.32 billion per quarter. The other statistical results shown Table 4.3 are helpful in evaluating the usefulness of the model. Most of these will

FIGURE 4.3 Disposable personal income (DPI) with a linear trend line forecast The linear trend follows the actual DPI quite well and provides a forecast for 2016 that looks very reasonable. The trend equation is: DPI = 11,042.86 + 108.32(T) (c4f3)

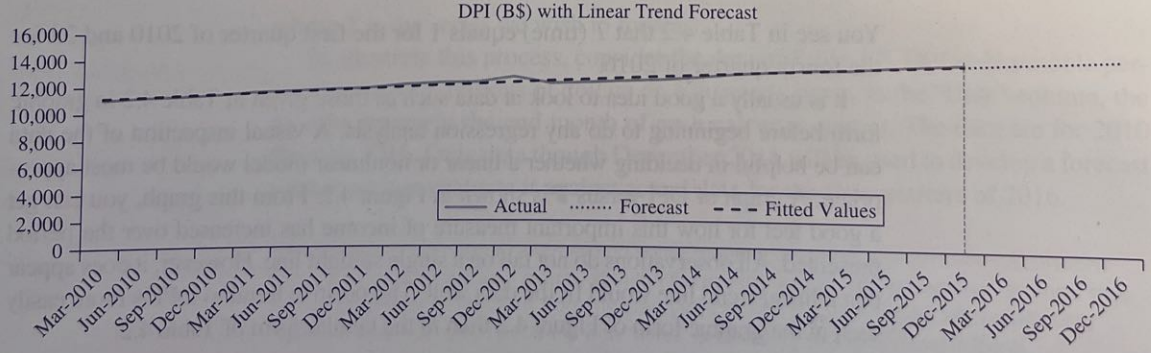


TABLE 4.3
Regression Trend Statistical Results from ForecastX (Audit Report).

Audit Trail—ANOVA Table (Trend (Linear) Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	1,34,93,902.63	1	1,34,93,902.63		571.28
Error	5,19,653.32	22	23,620.61	153.69	
Total	1,40,13,555.95	23			

Audit Trail—Coefficient Table (Trend (Linear) Regression Selected)

Name	Value	Standard Error	T-test	P-value
Intercept	11,042.86	64.76	170.53	0.00
Slope	108.32	4.53	23.90	0.00

Audit Trail—Statistics

Accuracy Measures	Value
MAPE	0.94%
R-Square	96.29%

Forecast Statistics	Value
Durbin Watson (1)	1.03

Method Statistics	Value
Method Selected	Trend (Linear) Regression

be discussed in detail in the section "Statistical Evaluation of Regression Models" in this chapter. Our discussion of others will be held in abeyance until Chapter 5. For now, we will just comment that statistical evaluation suggests that this linear equation provides a very good fit to the data.

The results in Table 4.3 are from the ForecastX™ Audit Report (the F value has been added) when "Trend (Linear) Regression" is selected as the forecast method.

The results in Table 4.4 are from Excel. You should look carefully at Tables 4.3 and 4.4. Notice that while the organization of information is different, the results are the same.

To use this equation to make a forecast for the four quarters of 2016, we need only substitute the appropriate values for time (T). These are 25 through 28, as seen in Table 4.2. The trend estimates of DPI for four representative quarters follow:

$$\text{2016 Quarter 1: DPI} = 11{,}042.86 + 108.32(25) = 13{,}750.86$$
$$\text{2016 Quarter 2: DPI} = 11{,}042.86 + 108.32(26) = 13{,}859.18$$
$$\text{2016 Quarter 3: DPI} = 11{,}042.86 + 108.32(27) = 13{,}967.50$$
$$\text{2016 Quarter 4: DPI} = 11{,}042.86 + 108.32(28) = 14{,}075.82$$

track the past time trend and project it forward for the forecast horizon of interest. Note that we do not imply any sense of causality in such a model. Time does not cause income to rise. Income has increased over time at a reasonably steady rate for reasons not explained by our model.

USING A CAUSAL REGRESSION MODEL TO FORECAST

Trend models, such as the one we looked at in the previous section for disposable personal income, use the power of regression analysis to determine the best linear trend line. However, such uses do not exploit the full potential of this powerful statistical tool. Regression analysis is especially useful for developing causal models.

In a causal model, expressed as $Y = f(X)$, a change in the independent variable (X) is assumed to cause a change in the dependent variable (Y). The selection of an appropriate causal variable (X) should be based on some insight that suggests that a causal relationship is reasonable. A forecaster does not arbitrarily select an X variable but rather looks to past experience and understanding to identify potential causal factors. For example, suppose that you were attempting to develop a bivariate regression model that might be helpful in explaining and predicting the level of jewelry sales in the United States. What factors do you think might have an impact on jewelry sales? Some potential causal variables that might come to mind could include income, some measure of the level of interest rates, and the unemployment rate, among others.

Discussions with knowledgeable people in the jewelry industry would help you determine other variables and would be helpful in prioritizing those that are identified. Library research in areas related to jewelry sales and to consumer behavior may turn up yet other potential X variables. One thing you would learn quickly is that there is a substantial seasonal aspect to jewelry sales.

It is important that the independent variable be selected on the basis of a logical construct that relates it to the dependent variable. Otherwise, you might find a variable through an arbitrary search process that works well enough in a given historical period, more or less by accident, but then breaks down severely out of sample. Consider, for example, William Stanley Jevons' sunspot theory of business cycles. For a certain historical period, a reasonably strong correlation appeared to support such a notion. Outside that period, however, the relationship was quite weak. In this case, it is difficult to develop a strong conceptual theory tying business cycles to sunspot activity.

To illustrate the use of a causal model, we will consider how well jewelry sales (JS) can be forecast on the basis of disposable personal income, as a measure of overall purchasing power.

Before we start to develop a forecast of jewelry sales, we should take a look at a time-series plot of the series. In this example, we will use quarterly data for jewelry sales from February 2010 through November 2015, and we want to forecast JS for each of the four quarters of 2016. A time-series plot of JS is found in Figure 4.4, and the raw data are in Table 4.6.

FIGURE 4.4
Jewelry Store Sales in Millions of Dollars Here We See Clearly the Seasonality of Jewelry Sales in the Raw Data (Solid Line). the Deseasonalized Data (Dashed Line) Help us See the Upward Trend More Clearly. (c4t6&f4)

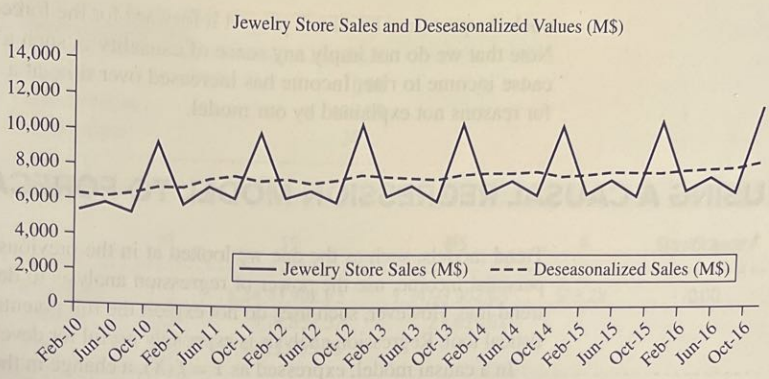


TABLE 4.6
Jewelry Store Sales in Millions of Dollars with the Deseasonalized Values and the Seasonal Indices

Date	Jewelry Store Sales (M$)	Deseasonalized Sales (M$)	Seasonal Indices
Feb-10	5,372	6,304.7	0.85
May-10	5,841	6,201.2	**0.94**
Aug-10	5,311	6,434.3	0.83
Nov-10	9,385	6,797.8	1.38
Feb-11	5,823	6,834.0	0.85
May-11	6,911	7,337.2	**0.94**
Aug-11	6,243	7,563.4	0.83
Nov-11	10,073	7,296.1	1.38
Feb-12	6,337	7,437.2	0.85
May-12	6,772	7,189.6	**0.94**
Aug-12	6,102	7,392.6	0.83
Nov-12	10,651	7,714.8	1.38
Feb-13	6,484	7,609.8	0.85
May-13	7,106	7,544.2	**0.94**
Aug-13	6,175	7,481.0	0.83
Nov-13	10,686	7,740.1	1.38
Feb-14	6,681	7,841.0	0.85
May-14	7,397	7,853.2	**0.94**
Aug-14	6,548	7,932.9	0.83
Nov-14	10,586	7,667.7	1.38
Feb-15	6,617	7,765.9	0.85
May-15	7,477	7,938.1	**0.94**
Aug-15	6,529	7,909.9	0.83
Nov-15	10,882	7,882.1	1.38
Feb-16	6,851	8,040.5	0.85
May-16	7,648	8,119.6	**0.94**
Aug-16	6,735	8,159.4	0.83
Nov-16	11,684	8,463.0	1.38

In Table 4.6, the values for the four quarters of 2016 are separated. We will hold these four quarters out when we make forecasts and use them to evaluate accuracy using the MAPE for just those four quarters.

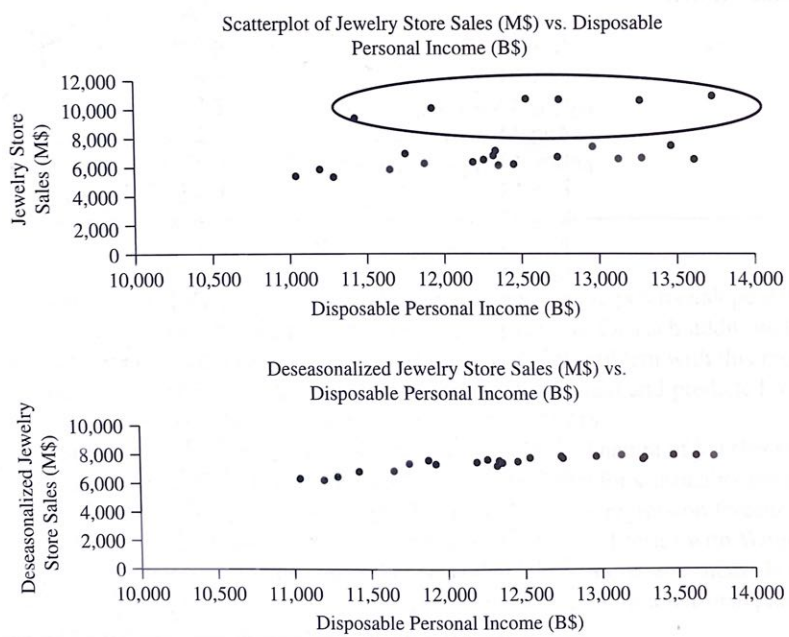
Note also that the seasonal indices repeat year after year. The May (second quarter) seasonal indices have been put in bold type to help illustrate this. For this example, the seasonal indices were calculated using time series decomposition. This method will be covered in detail in Chapter 6.

EWELRY SALES FORECAST BASED I DISPOSABLE PERSONAL INCOME

If we hypothesize that disposable personal income (DPI) is influential in determining jewelry store sales (JS), we might initially want to look at a scattergram of these two variables. This is shown in Figure 4.5, where JS is plotted on the vertical axis and DPI is on the horizontal axis. Note that the horizontal axis does not start at zero. This helps us see the data more clearly. You can see that higher values of JS appear to be associated with higher incomes.

In the top graph, you can see the effect of seasonality in a dramatic way. Look at the six values that are circled. These six points are all fourth-quarter data points due to high holiday season sales.

In the lower graph, the JS data are shown after the seasonality has been removed. In Chapter 6, you will learn how to deseasonalize data and how to find



GURE 4.5
atterplots of welry Store Sales op Graph) and seasonalized welry Store Sales ottom Graph) with sposable Personal come. (c4f5)

seasonal indices. In the bottom graph, you can see that a straight line through those points could provide a reasonably good fit to the data. You also can see that all of these observations are well away from the origin. If you do not look closely in the graph, you might think the observations are not so far from the vertical axis. But note that again in this case, the horizontal axis starts at 10,000, not at zero. This was done to better display the deseasonalized data points. It is important to look at graphs carefully.

The bivariate regression model for seasonally adjusted jewelry sales (which we will call SAJS) as a function of DPI may be written as:

$$\text{SAJS} = b_0 + b_1(\text{DPI})$$

The JS data used to estimate values for b_0 and b_1 are given in Table 4.7, along with the data for DPI.

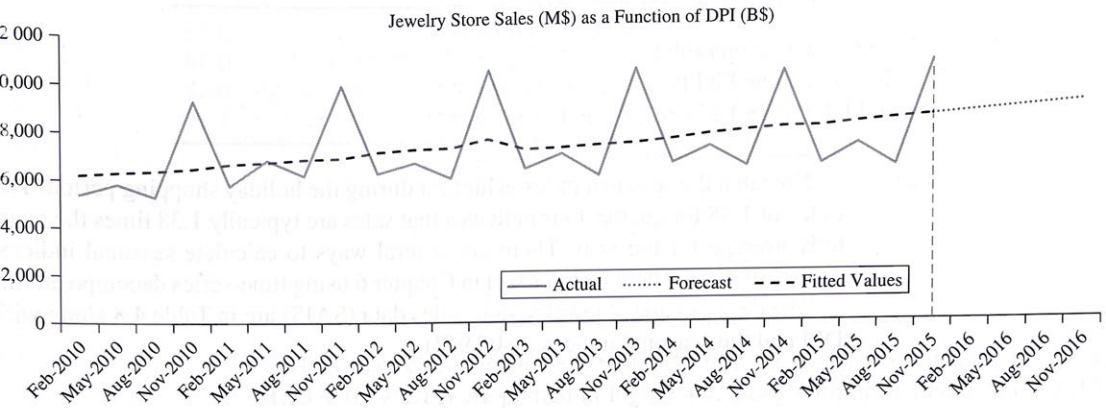
The basic regression results are shown in Figure 4.6, along with a graph of the actual and predicted values based on this model. To use this model to forecast for 2016, a Holt's exponential smoothing forecast of DPI was used. On the basis of these results, the forecast model (equation) for jewelry sales as a function of disposable personal income per capita is:

$$\text{JS} = -3{,}812.18 + 0.91(\text{DPI})$$

TABLE 4.7
Jewelry Sales and Disposable Personal Income (c4t7f6)

Date	Jewelry Store Sales (M$)	DPI
Feb-10	5,372	11041.47
May-10	5,841	11197.63
Aug-10	5,311	11286.63
Nov-10	9,385	11425.73
Feb-11	5,823	11652.23
May-11	6,911	11751.63
Aug-11	6,243	11876.67
Nov-11	10,073	11924.93
Feb-12	6,337	12189.97
May-12	6,772	12321.33
Aug-12	6,102	12355.43
Nov-12	10,651	12748.13
Feb-13	6,484	12259.27
May-13	7,106	12335.9
Aug-13	6,175	12453.83
Nov-13	10,686	12534.33
Feb-14	6,681	12735.83
May-14	7,397	12962.43
Aug-14	6,548	13127.43
Nov-14	10,586	13265.27
Feb-15	6,617	13276.5
May-15	7,477	13464.7
Aug-15	6,529	13611.7
Nov-15	10,882	13726.37

FIGURE 4.6 Jewelry Sales Forecast as a Function of DPI we See That the Upward Trend in Jewelry ales is Accounted for by the Regression Model but the Seasonality is not Taken Into Account. Thus, for ny Given Quarter the Forecast is Likely to be Substantially Incorrect. The Quarter Represented by the ecember 2016 Forecast is Surely Much Too Low. (c4t7&f6)



Audit Trail—Coefficient Table (Multiple Regression Selected)

	Coefficient	Standard error	T-test	P-value
Intercept	−3,812.18	5,766.40	−0.66	0.52
DPI	0.91	0.46	1.95	0.06

Audit Trail - Statistics

Accuracy	Value	Forecast Statistics	Value
MAPE	17.78%	Durbin Watson (4)	0.04
R-Square	14.75%		

Data for 2010–2015 were used to estimate this model. The positive slope (0.91) indicates that, on average, JS increases by $0.91 million for each additional $1 billion increase in disposable personal income. A major problem with this model is apparent in Figure 4.6. It is clear from the graph of actual and predicted retail sales that this model fails to deal with the seasonality in JS.

The failure of this model to deal well with the seasonal nature of jewelry sales suggests that either we should use a model that can account for seasonality directly or we should deseasonalize the data before developing the regression forecasting model. In Chapter 3, you learned how to forecast a seasonal series with Winters' exponential smoothing. In Chapter 5, you will see how regression methods can also incorporate seasonality, and in Chapter 6, you will see how a seasonal pattern can be modeled using time-series decomposition.

We will now develop a model based on seasonally adjusted jewelry sales dat (SAJS) and then reintroduce the seasonality to the forecast. To seasonally adjus jewelry sales, the following seasonal indices were used:

Quarter 1 (February)	0.85
Quarter 2 (May)	0.94
Quarter 3 (August)	0.82
Quarter 4 (November)	1.38

Note that the seasonal index is highest during the holiday shopping period. The index of 1.38 for quarter four indicates that sales are typically 1.38 times the quarterly average for the year. There are several ways to calculate seasonal indices. The method used here is described in Chapter 6 using time-series decomposition.[4]

The deseasonalized jewelry store sales data (SAJS) are in Table 4.8 along with DPI and the seasonal indices. (c4t8&f7)

TABLE 4.8. Deseasonalized Jewelry Sales, Disposable Personal Income, and the Seasonal Indices

Date	Deseasonalized Sales (M$)	DPI (B$)	Seasonal Indices
Feb-10	6,313.6	11,041.5	0.85
May-10	6,202.7	11,197.6	0.94
Aug-10	6,438.1	11,286.6	0.82
Nov-10	6,788.4	11,425.7	1.38
Feb-11	6,843.6	11,652.2	0.85
May-11	7,339.0	11,751.6	0.94
Aug-11	7,567.8	11,876.7	0.82
Nov-11	7,286.0	11,924.9	1.38
Feb-12	7,447.7	12,190.0	0.85
May-12	7,191.3	12,321.3	0.94
Aug-12	7,396.9	12,355.4	0.82
Nov-12	7,704.1	12,748.1	1.38
Feb-13	7,620.5	12,259.3	0.85
May-13	7,546.0	12,335.9	0.94
Aug-13	7,485.4	12,453.8	0.82
Nov-13	7,729.4	12,534.3	1.38
Feb-14	7,852.0	12,735.8	0.85
May-14	7,855.1	12,962.4	0.94
Aug-14	7,937.6	13,127.4	0.82
Nov-14	7,657.1	13,265.3	1.38
Feb-15	7,776.8	13,276.5	0.85
May-15	7,940.0	13,464.7	0.94
Aug-15	7,914.5	13,611.7	0.82
Nov-15	7,871.2	13,726.4	1.38

[4] In ForecastX™, use the "Decomposition" forecast method. Then for "Type" select "Multiplicative" and for "Decomposed Data" select "Trend (Linear) Regression."

When we regress the seasonally adjusted values of jewelry sales (SAJS) as a ıction of disposable personal income using data for 2010–2015, we get the re- ts shown at the bottom of Figure 4.7 and summarized by the following equation:

$$\text{SAJS} = 61.53 + 0.59(\text{DPI})$$

We can substitute values of DPI into this equation to get predictions for season- y adjusted jewelry sales (SAJS). To get DPI values for 2016, a Holt's exponen- l smoothing forecast was used. The results for 2016 DPI are:

Date	DPI Forecast for 2016
Feb-16	13,832.2
May-16	13,943.8
Aug-16	14,055.5
Nov-16	14,167.1

The values calculated for SAJS in 2016 from the regression equation (SAJS = .53 + 0.59*DPI) are plotted in Figure 4.7, along with the actual data for 2010 rough 2015. During the historical period, actual values of DPI were used to lculate SAJS, while in the forecast period (2016), forecast values of DPI using olt's exponential smoothing were used.

Multiplying SAJS by the seasonal index for each quarter, we obtain a predic- on of the unadjusted jewelry sales for each quarter. This process is illustrated in able 4.9 and graphed in Figure 4.8.

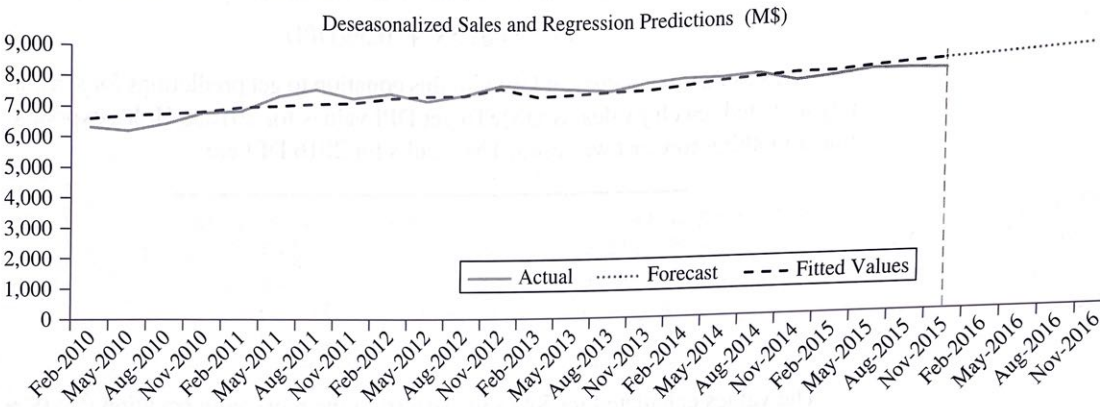
Let's summarize how this forecast was done.

1. We first found the seasonal indices using a process you will learn in Chapter 6 (although you already know a way to get seasonal indices using Winters' method).
2. We then calculated the deseasonalized (sometimes called seasonally adjusted) jewelry store sales (SAJS) by dividing the actual values by the seasonal indices.
3. The deseasonalized values were forecast using a simple regression with SAJS as a function of disposable personal income. In Figure 4.7, you see that the model fit fairly well.
4. The final forecast was found by multiplying the SAJS fitted and forecast values by the seasonal indices. This put the seasonality back into the final forecast. In Figure 4.8, you see that the process appears to have worked well.

Ve do have the real data for 2016, so let's use that to calculate the MAPE for the our quarters of 2016.

Date	Actual Jewelry Store Sales	Forecast Jewelry Store Sales	Error	Absolute Error	Absolute % Error
Feb-16	6,851	7,023.49	-172.49	172.49	2.52
May-16	7,648	7,835.44	-187.44	187.44	2.45
Aug-16	6,735	6,918.56	-183.56	183.56	2.73
Nov-16	11,684	11,686.17	-2.17	2.17	0.02
				MAPE =	1.93

FIGURE 4.7 Seasonally Adjusted Jewelry Sales as a Function of DPI The Upward Trend in SAJS is Seen Clearly in This Graph. The Forecast Values Now Need to be Readjusted to Put the Seasonality Back into the Forecast. This is Done in Table 4.8. (C4t8&f7)



Date	Forecast
Feb-2016	8,254.53
May-2016	8,320.64
Aug-2016	8,386.76
Nov-2016	8,452.88

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source	SS	df	MS	SEE	F-ratio
Regression	49,16,436.99	1	49,16,436.99		79.43
Error	13,61,670.56	22	61,894.12	248.79	
Total	62,78,107.55	23			

Audit Trail—Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value
Intercept	61.53	825.44	0.07	0.94
DPI (B$)	0.59	0.07	8.91	0.00

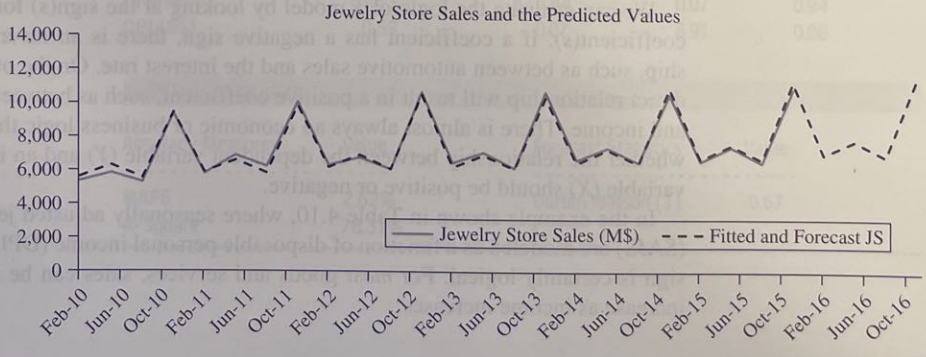
Audit Trail - Statistics

Accuracy Measures	Value	Forecast Statistics	Value
MAPE	2.82%	Durbin Watson (1)	0.67
R-Square	78.31%		

TABLE 4.9
Calculation of a Final Forecast for Jewelry Store Sales (c4t9&f8)

Date	Fitted and Forecast SAJS	Seasonal Indices	Fitted and Forecast JS
Feb-10	6,601.53	0.85	5,617.02
May-10	6,694.03	0.94	6,303.68
Aug-10	6,746.75	0.82	5,565.65
Nov-10	6,829.14	1.38	9,441.34
Feb-11	6,963.30	0.85	5,924.83
May-11	7,022.17	0.94	6,612.69
Aug-11	7,096.23	0.82	5,853.96
Nov-11	7,124.82	1.38	9,850.12
Feb-12	7,281.80	0.85	6,195.84
May-12	7,359.61	0.94	6,930.45
Aug-12	7,379.81	0.82	6,087.89
Nov-12	7,612.41	1.38	10,524.22
Feb-13	7,322.85	0.85	6,230.76
May-13	7,368.24	0.94	6,938.58
Aug-13	7,438.09	0.82	6,135.97
Nov-13	7,485.78	1.38	10,349.15
Feb-14	7,605.13	0.85	6,470.94
May-14	7,739.34	0.94	7,288.04
Aug-14	7,837.08	0.82	6,465.11
Nov-14	7,918.72	1.38	10,947.69
Feb-15	7,925.37	0.85	6,743.43
May-15	8,036.84	0.94	7,568.19
Aug-15	8,123.91	0.82	6,701.73
Nov-15	8,191.83	1.38	11,325.28
Feb-16	8,254.53	0.85	7,023.49
May-16	8,320.64	0.94	7,835.44
Aug-16	8,386.76	0.82	6,918.56
Nov-16	8,452.88	1.38	11,686.17

FIGURE 4.8 **Jewelry Sales Final Forecast** Actual and Forecast Values are Shown in the Graph for Each Quarter from 2010 Through 2015. The Forecast Values for 2016 are Shown by the Dotted Line. (c4t9&f8)



We see that the MAPE is 1.93 percent. This confirms our visual inspection of Figure 4.8, which suggests that this model does a good job of forecasting jewelry store sales.

STATISTICAL EVALUATION OF REGRESSION MODELS

Now that you have a basic understanding of how simple bivariate regression models ($Y = fX$) can be applied to forecasting, let us look more closely at some things that should be considered in evaluating regression models. The regression model developed for the deseasonalized jewelry store sales (SAJS) as a function of disposable personal income (DPI) will be used as an example. Parts of the ForecastX Audit Report are shown in Table 4.10. The graphic result is also shown to help you remember the nature of the original SAJS data and the values predicted by the regression model.

Basic Diagnostic Checks for Evaluating Regression Results

First, ask yourself whether the sign on the slope term makes sense.

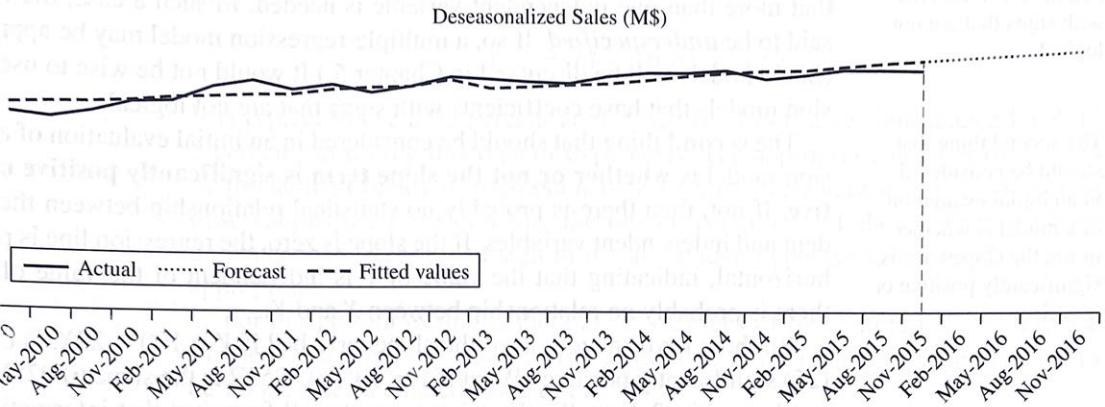
As we progress, we will develop a five-step process for evaluating a regression model. Right now, we will consider just the first three steps, then in the "Serial Correlation" section of this chapter, we will introduce the fourth step. The fifth step will not become relevant until we get to Chapter 5.

To start with, there are several things you should consider when you look at regression results. **First**, ask yourself whether the model you estimate is **logical**. The answer to this question depends on you business/economic knowledge about the situation. There is no statistical test involved. You need to have a basic understanding of the data being analyzed. For example, if you were trying to model automobile sales in units and were asked whether you would expect sales to go up or down if the interest rate were to decline, you would surely say sales would increase. There is certainly an abundance of evidence that this is the belief within the auto industry. Otherwise, car companies would not offer zero- or low-interest financing from time to time. But consider the affect of a decline in personal income. Again your answer would be obvious: sales would decline if people have less money to spend.

We can evaluate the logic of a model by looking at the sign(s) for regression coefficient(s). If a coefficient has a negative sign, there is an inverse relationship, such as between automotive sales and the interest rate. On the other hand, a direct relationship will result in a positive coefficient, such as between auto sales and income. There is almost always an economic or business logic that indicates whether the relationship between the dependent variable (Y) and an independent variable (X) should be positive or negative.

In the example shown in Table 4.10, where seasonally adjusted jewelry sales (SAJS) are modeled as a function of disposable personal income (DPI), a positive sign is certainly logical. For *most* goods and services, sales can be expected to increase as income increases.

4.10 Basic Regression Results. Seasonally adjusted jewelry store sales as a function of disposable income.



Multiple Regression—Result Formula

Deseasonalized Sales (M$) = 61.53 + ((DPI (B$)) * 0.592313)

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	F-test
Regression	49,16,436.99	1	49,16,436.99		79.43
Error	13,61,670.56	22	61,894.12	248.79	
Total	62,78,107.55	23			

Audit Trail—Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value
Intercept	61.53	825.44	0.07	0.94
DPI (B$)	0.59	0.07	8.91	0.00

Audit Trail - Statistics

Accuracy Measures	Value	Forecast Statistics	Value
MAPE	2.82%	Durbin Watson (1)	0.67
R-Square	78.31%		

It would not be wise to use regression models that have coefficients with signs that are not logical.

What if the signs do not make sense? This is a clear indication that something is wrong with the regression model. It may be that the model is incomplete and that more than one independent variable is needed. In such a case, the model is said to be ***underspecified***. If so, a multiple-regression model may be appropriate. (Such models will be discussed in Chapter 5.) It would not be wise to use regression models that have coefficients with signs that are not logical.

The second thing that should be considered in an initial evaluation of a model is whether or not the slope term is significantly positive or negative.

The **second** thing that should be considered in an initial evaluation of a regression model is **whether or not the slope term is significantly positive or negative**. If not, then there is probably no statistical relationship between the dependent and independent variables. If the slope is zero, the regression line is perfectly horizontal, indicating that the value of Y is independent of the value of X (i.e., there is probably no relationship between X and Y).

But how far from zero need the slope term be? If $Y = 100 + 25X$, is the slope (25) significantly positive? What if $Y = 10 + 0.025X$? Is the slope (0.025) significantly positive? Actually. there is no way to tell from just that information. The size of the slope is dependent on how the data are scaled. In our jewelry store sales examples, we used disposable personal income in billions of dollars. If we had used the income data in trillions of dollars, the slope term in the regression would have been larger. Because 1,500 billion is 1.5 trillion, the slope using billions would be smaller than the slope using trillions. The resulting predictions, however, would be the same either way income is measured. To determine if the slope is significantly greater or less than zero, we must test a hypothesis concerning the true slope. Remember that our basic regression model is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

If $\beta_1 = 0$, then $Y = \beta_0$ regardless of the value of X.

When we have a predisposition about whether the coefficient should be positive or negative based on our knowledge of the relationship, a one-tailed hypothesis test is appropriate. If our belief suggests a positive coefficient, the hypothesis would be set up as follows:

$$H_0{:}\beta \leq 0$$
$$H_1{:}\beta > 0$$

This form would be correct for the case in Table 4.10, since a direct (positive) relationship is expected.

When our belief suggests a negative coefficient, the hypothesis would be set up as follows:

$$H_0{:}\beta \geq 0$$
$$H_1{:}\beta < 0$$

This form would be correct when an inverse (negative) relationship is expected.

In some situations, we may not have a specific expectation about the direction of causality, in which case a two-tailed hypothesis test is used. The hypothesis would be set up as follows:

$$H_0{:}\beta = 0$$
$$H_1{:}\beta = 0$$

The appropriate statistical test is a *t*-test, where the calculated value of t (t_{calc}) is equal to the slope term minus zero, divided by the standard error of the slope.[5] That is:

$$t_{calc} = (b_1 - 0) / (s.e.\ of b_1)$$

It is typical to use a 95 percent confidence level (an α, or significance level, of 5 percent) in testing this type of hypothesis. The appropriate number of degrees of freedom in bivariate regression is always $n - 2$, where n is the number of observations used in estimating the model. As described above, when we have a greater-than or less-than sign in the alternative hypothesis, a one-tailed test is appropriate.

For our present example, there are 22 degrees of freedom (24 – 2). From the *t*-table on page 66, we find the critical value of t (such that 0.05 is in one tail) to be 1.717. The calculated value of t *is*:

For the SAJS = *f*(DPI) Causal Model

$t_{calc} = (0.59 - 0)/0.07$

$= 8.91$

The *t*-value shown here is from Table 4.10. If you do this calculation by hand, the results may differ from the value shown here and in Table 4.10 due to rounding.

For our example, the calculated value is larger (more positive) than the critical, or table, value, so we can reject H_0 and conclude that the regression coefficient is significantly greater than zero. If this statistical evaluation of the coefficients in a regression analysis results in failure to reject the null hypothesis, then it is probably not wise to use the model as a forecasting tool.[6] However, it is not uncommon to relax the criterion for evaluation of the hypothesis test to a 90 percent confidence level (a 10 percent significance level).

In determining whether or not to reject H_0, an alternative to comparing *t*-values is to consider the significance level (often called the *P*-value) given in most computer output. Let us assume that we desire a 95 percent confidence level. This is the equivalent of saying that we desire a 5 percent significance level.[7]

[5] The standard error of the estimated regression coefficient measures the sampling variability of b_1 about its expected value β_1, the true population parameter.

[6] A phenomenon known as *serial correlation* (which we will discuss shortly) may cause coefficients to appear significantly different from zero (as measured by the t-test) when in fact they are not.

[7] Remember that the confidence level and the significance level add to 1. Thus, if we know one of these, we can easily determine the other.

For a two-tailed hypothesis test (H_1: $\beta_1 \neq 0$), we can then reject H_0 if the reported two-tailed significance level[8] in the output is less than 0.05. For a one-tailed hypothesis test (H_1: $\beta_1 < 0$ or H_1: $\beta_1 > 0$), we can reject H_0 if one-half of the reported two-tailed significance level is less than 0.05.

In our example in Table 4.10, the two-tailed significance level associated with the calculated t-ratio is 0.00. Clearly, one-half of 0.00 is less than 0.05, so it is appropriate to reject H_0. Note that we reach the same conclusion whether we evaluate a hypotheses by comparing the calculated and table t-ratios or by looking at the significance level.

The third check of regression results is to evaluate what percent of the variation (i.e., up-and-down movement) in the dependent variable is explained by variation in the independent variable.

The **third** check of regression results is to evaluate **what percent of the variation** (i.e., up-and-down movement) **in the dependent variable is explained by variation in the independent variable.** This is evaluated by interpreting the R-squared value that is reported in regression output. R-squared is the **coefficient of determination,** which tells us the fraction of the variation in the dependent variable that is explained by variation in the independent variable. Thus, R-squared can range between zero and one. Zero would indicate no explanatory power, while one would indicate that all of the variation in Y is explained by the variation in X. (A related statistic, adjusted R-squared, will be discussed in Chapter 5.)

The model for seasonally adjusted jewelry sales as a function of DPI has an R-squared of .7831. Thus, variations in the DPI explain 78.31 percent of the variation in seasonally adjusted jewelry sales.

It is possible to perform a statistical test to determine whether the coefficient of determination (R^2) is significantly different from zero. The hypothesis test may be stated as:

$$H_0: R^2 = 0$$
$$H_1: R^2 \neq 0$$

The appropriate statistical test is an F-test, which will be presented in Chapter 5. With bivariate regression, it turns out that the t-test for the slope term in the regression equation is equivalent to the F-test for R-squared. Thus, we will wait until we explore multiple-regression models to discuss the application of the F-test. But you might note that if you square the t-ratio in our example, you get the F-ratio that is reported (you may get a slightly different answer due to rounding).

Before considering other statistical diagnostics, let us summarize these three initial evaluative steps for bivariate regression models:

1. **Logic**: Ask whether the sign for the slope term makes sense.
2. **Statistical significance**: Check to see whether the slope term is statistically positive or negative at the desired significance level by using a t-test.
3. **Explanatory power**: Evaluate how much of the variation in the dependent variable is explained by the regression model using the R-squared (R^2) value.

These three items can be evaluated from the results presented in standard computer output, such as shown in Table 4.10.

[8] In ForecastX™, as well as most other statistical packages, two-tailed significance levels are reported. These are frequently referred to as *P-values*, as is the case in ForecastX™.

USING THE STANDARD ERROR OF THE ESTIMATE

The forecasts we made in the preceding pages—using a simple linear trend model and the two causal regression models—were point estimates. In each case, we substituted a value for the independent variable into the regression equation to obtain a single number representing our best estimate (forecast) of the dependent variable. It is sometimes useful to provide an interval estimate rather than a point estimate.

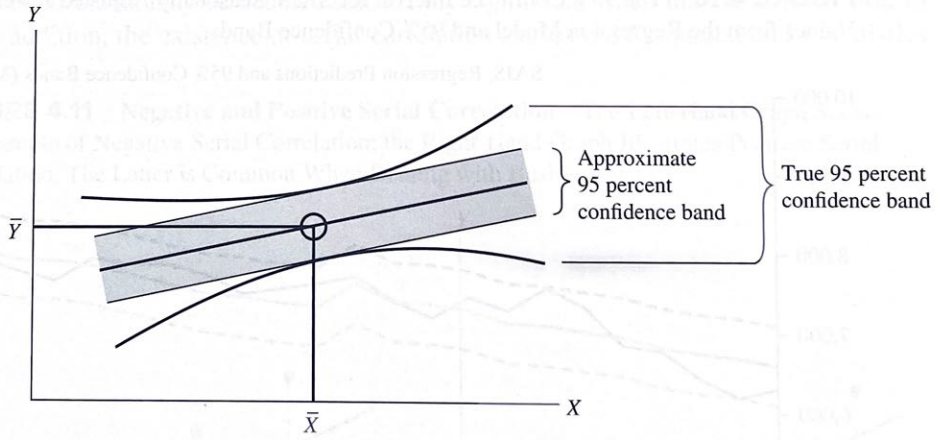
The standard error of the estimate (SEE) can be used to generate *approximate* confidence intervals with relative ease. The SEE is often also called the *standard error of the regression.* The confidence intervals we present here are approximate because the true confidence band is not parallel to the regression line but rather bows away from the regression line at values of Y and X far from the means. This is illustrated in Figure 4.9. The approximate 95 percent confidence interval can be calculated as follows:[9]

$$\text{Point estimate} \pm 2 \text{ (standard error of the estimate)}$$

The value of 2 is used as an easy approximation for the correct t-value. Recall that if there are a large number of degrees of freedom, $t = 1.96$.

Representative calculations of approximate 95 percent confidence bands for the regression forecasts developed for seasonally adjusted jewelry sales (SAJS)

FIGURE 4.9 Confidence Bands Around a Regression Line The true Confidence Band Bows away From the Regression Line. An Approximate 95 Percent Confidence Band can be Calculated by Taking the Point Estimate for Each X, Plus or Minus 2 Times the Standard Error of the Estimate.



[9] The true 95 percent confidence band for predicting Y for a given value of X (X_0) can be found as follows:

$$\hat{Y} \pm t(\text{SEE}) \sqrt{1 + (1/n) + [(X_0 + \bar{X})^2 / \Sigma(X - \bar{X})^2]}$$

where t is the appropriate value from the t-distribution at $n - 2$ degrees of freedom and the desired significance level, SEE is the standard error of the estimate, and $\hat{Y}$ is the point estimate determined from the estimated regression equation.

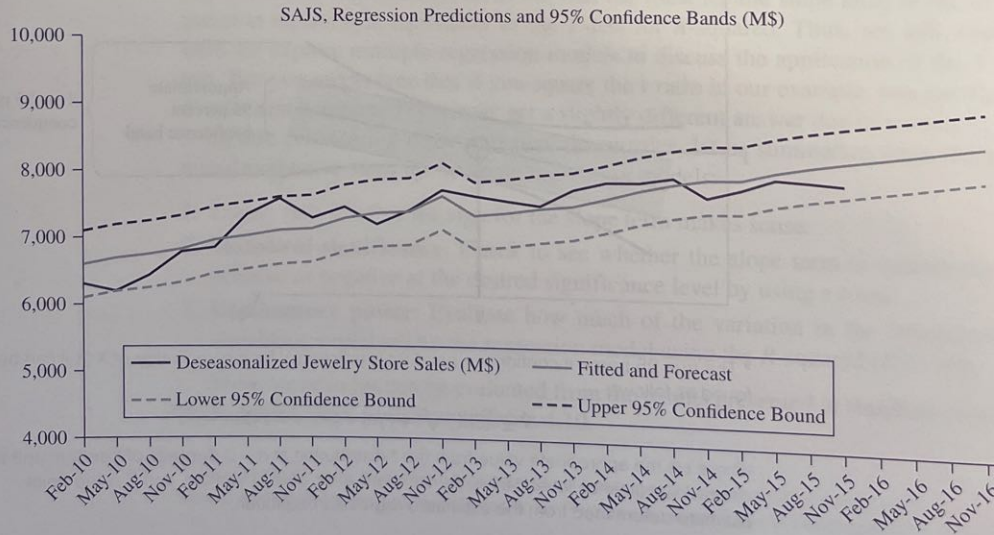
TABLE 4.11 Representative Calculations of Approximate 95 Percent Confidence Intervals: Point Estimate ±2 × Standard Error of the Estimate (SEE). The SEE is taken from the ANOVA part of Figure 4.7.

For SAJS: 2 × SEE = 2 × 248.79 = 497.58		
Period	**95 Percent Confidence Interval**	**Actual SAJS†**
February 2015	7,925.37 ± 497.58 = 7,427.79 to 8,422.95	7,776.8
May 2015	8,036.84 ± 497.58 = 7,539.26 to 8,534.42	7,940.0
August 2015	8,123.91 ± 497.58 = 7,626.33 to 8,621.49	7,914.5
November 2015	8,191.83 ± 497.58 = 7,694.25 to 8,689.41	7,871.2

† Note that this is for jewelry sales seasonally adjusted.

are shown in Table 4.11. The standard errors of the regressions are taken from Table 4.10, while the point estimates for each model and each quarter are those that were found in the "Using a Causal Regression Model to Forecast" section. Figure 4.10 shows the original SAJS along with the point estimates and the upper and lower bounds of the approximate 95 percent confidence interval. All of the actual SAJS data happen to fall within 95 percent confidence bounds in this case, although the May 2010 and August 2011 values are close to the lower and upper bounds, respectively. It would not be uncommon for a few actual values to fall outside of the 95 percent confidence bands. A 90 percent confidence band would be more narrow, and there would be a higher likelihood of values falling outside.

FIGURE 4.10 The 95% Confidnce Interval for SAJS. Seasonally Adjusted Jewelry Store Sales, Predicted Values from the Regression Model and 95% Confidence Bands.



RRELATION

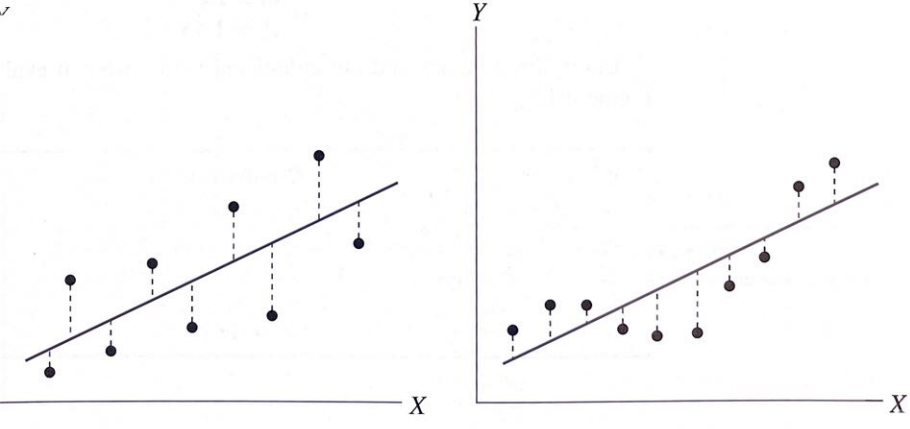
Business and economic data used in forecasting are most often time-series data. The disposable personal income data and the jewelry store sales data used in this chapter are typical of such time series. In using regression analysis with time-series data, the problem known as *serial correlation* can cause some difficulty.

One of the assumptions of the ordinary least-squares regression model is that the error terms are independent and normally distributed, with a mean of zero and a constant variance. If this is true for a particular case, we would not expect to find any regular pattern in the error terms. When a significant time pattern in the error terms occurs, it violates the independence assumption, and therefore serial correlation is a potential problem.

Figure 4.11 illustrates the two possible cases of serial correlation. In the left-hand graph, the case of negative serial correlation is apparent. Negative serial correlation exists when a negative error is followed by a positive error, then another negative error, and so on. The error terms alternate in sign. Positive serial correlation is shown in the right-hand graph in Figure 4.11. In positive serial correlation, positive errors tend to be followed by other positive errors, while negative errors are followed by other negative errors.

When serial correlation exists, problems can develop in using and interpreting the OLS regression function. The existence of **serial correlation does not bias the coefficients** that are estimated, but it **does make the estimates of the standard errors smaller** than the true standard errors. This means that the ***t*-ratios calculated for each coefficient will be overstated**, which in turn may lead to the rejection of null hypotheses that should not have been rejected. That is, regression coefficients may be deemed statistically significant when indeed they are not. In addition, the existence of serial correlation causes the R-squared and F-statistics

FIGURE 4.11 Negative and Positive Serial Correlation The Left-Hand Graph Shows an Example of Negative Serial Correlation; the Right-Hand Graph Illustrates Positive Serial Correlation. The Latter is Common When Dealing with Business Data.



to be unreliable in evaluating the overall significance of the regression function (the F-statistic will be discussed in Chapter 5).

There are a number of ways to test statistically for the existence of serial correlation. The method most frequently used is the evaluation of the Durbin-Watson statistic (DW). This statistic is calculated as follows:

$$\text{DW} = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2}$$

where e_t is the residual for the time period t, and e_{t-1} is the residual for the preceding time period $(t - 1)$. Almost all computer programs for regression analysis include the Durbin-Watson statistic, so you are not likely to have to calculate it directly. Excel is an exception to this. Therefore, in Excel one needs to do the calculations from the residuals that are provided when requested.

The DW statistic will always be in the range of 0 to 4. As a rule of thumb, a value close to 2 (e.g., between 1.50 and 2.50) indicates that there is no serial correlation. When the DW statistic approaches 4, the degree of negative serial correlation increases. When positive serial correlation exists, the value of DW approaches 0.

To be more precise in evaluating the significance and meaning of the calculated DW statistic, we must refer to a Durbin-Watson table, such as Table 4.12. Note that for each number of independent variables (k), two columns of values labeled d_l and d_u are given. The values in these columns for the appropriate number of observations (N) are used in evaluating the calculated value of DW according to the criteria shown in Figure 4.12.

To illustrate, let us consider the simple regression for seasonally adjusted jewelry sales as a function of disposable personal income (DPI). From Table 4.10 (see page 179), you see that the calculated Durbin-Watson statistic is 0.67. This value is well below 2 and relatively close to zero. Therefore, without doing a formal evaluation, we would suspect positive serial correlation could be a problem. However, it is always wise to do the formal DW test.

Using Table 4.12, we find for $k = 1$ and $N = 24$ that:

$$d_l = 1.27$$
$$d_u = 1.45$$

Using these values and our calculated value, we can evaluate the criteria in Figure 4.12:

Region	Comparison	Result
A	$4 > 0.67 > (4 - 1.27)$	False
B	$(4 - 1.27) > 0.67 > (4 - 1.45)$	False
C	$(4 - 1.69) > 0.67 > 1.69$	False
D	$1.69 > 0.67 > 1.45$	False
E	$1.27 > 0.67 > 0$	True

TABLE 4.12
The Durbin-Watson Statistic

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
N	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.53	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

k = the number of independent variables; N = the number of observations used in the regression.

Source: Durbin, J. and Watson, G. S., "Testing for Serial Correlation in Least Squares Regression," *Biometrika* 38, June 1951, 173.

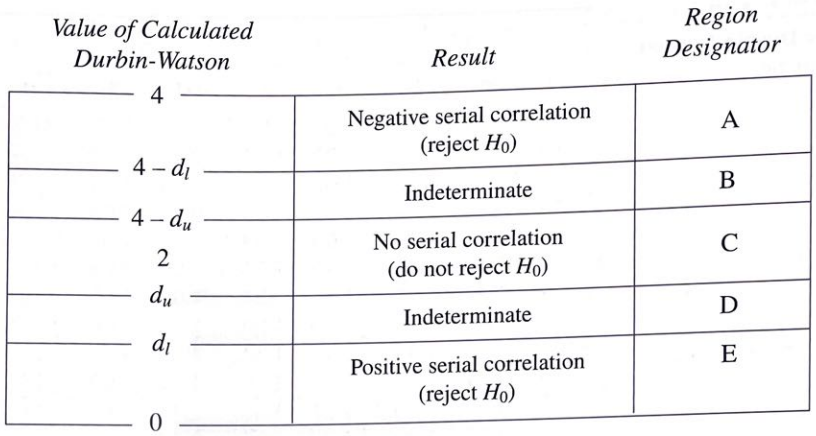


FIGURE 4.12 A Schematic for Evaluating Serial Correlation Using the Durbin-Watson Statistic

d_u = Upper value of Durbin-Watson from Table 4.12

d_l = Lower value of Durbin-Watson from Table 4.12

H_0: $\rho = 0$ (i.e., no serial correlation)

H_1: $\rho \neq 0$ (i.e., serial correlation exists)

Since our result is in region E, we can conclude that positive serial correlation exists in this case. You can see evidence of this positive serial correlation if you look at Figure 4.10 (page 184) at how the regression line (fitted) is at first above, then below, then above, then below, and finally above the actual data in a recurring pattern. This is a classic case of positive serial correlation. Positive serial correlation is more common with business/economic data than is negative serial correlation.

A primary cause of positive serial correlation is the existence of long-term cycles and trends in economic and business data.

You might well ask: What causes positive serial correlation and what can be done about it? A primary cause of serial correlation is the existence of long-term cycles and trends in economic and business data. Such trends and cycles are likely to produce positive serial correlation. Serial correlation can also be caused by a misspecification of the model. Either leaving out one or more important variables or failing to include a nonlinear term when one is called for can be a cause.

We can try several relatively simple things to reduce serial correlation. One is to first use differences of the variables rather than the actual values when performing the regression analysis. That is, use the change in each variable from period to period in the regression. For example, we could try the following:

$$\Delta Y = b_0 + b_1(\Delta X)$$

where Δ means "change in" and is calculated as follows:

$$\Delta Y_t = Y_t - Y_{t-1}$$
$$\Delta X_t = X_t - X_{t-1}$$

This process of "first-differencing" will be seen again in Chapter 7, when we discuss ARIMA forecasting models.

Other approaches to solving the serial correlation problem often involve moving into the realm of multiple regression, where there is more than one independent variable in the regression model. For example, it may be that other causal factors account for the differences between the actual and predicted values. For example,

in the jewelry sales regression, we might add the interest rate and the unemployment rate as additional independent variables.

A third, and somewhat related, approach to dealing with serial correlation is to introduce the square of an existing causal variable as another independent variable. The model might look as follows:

$$Y_t = b_0 + b_1 X_t + b_2 Y^2$$

Also, we might introduce a lag of the dependent variable as an independent variable. Such a model might look as follows:

$$Y_t = b_0 + b_1 X_t + b_2 Y_{t-1}$$

where t represents the current time period and $t - 1$ represents the previous time period.

There are other procedures, based on more sophisticated statistical models, that are helpful in dealing with the problems created by serial correlation. These are typically based on an extension of the use of first differences in that they involve the use of generalized differencing to alter the basic linear regression model into one for which the error terms are independent of one another (i.e., $\rho = 0$, where ρ [rho] is the correlation between successive error terms).

The basic regression model is:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

and since this is true for all time periods, it follows that:

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1}$$

Multiplying the second of these equations by ρ and subtracting the result from the first yields the following generalized-differencing transformed equation:

$$Y_t^* = (1 - \rho)\beta_0 + \beta_1 X_t^* + v_t$$

where:

$$\begin{aligned} Y_t^* &= Y_t - \rho Y_{t-1} \\ X_t^* &= X_t - \rho X_{t-1} \\ v_t &= \varepsilon_t - \rho \varepsilon_{t-1} \end{aligned}$$

It can be shown that the resulting error term, v_t, is independently distributed with a mean of zero and a constant variance.[10] The problem with this generalized-differencing model is that we do not know the correct value for ρ. Two common methods for estimating ρ and the corresponding regression model are the Cochrane-Orcutt procedure and the Hildreth-Lu procedures.[11]

[10] Most econometrics books describe the underlying statistical theory as well as the two correction procedures we include herein. For example, see Robert S. Pindyck and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 3rd ed. (New York: McGraw-Hill, 1991), pp. 137–47.

[11] While these methods help solve the serial-correlation problem, they are not often used in practice for forecasting, largely due to their added complexity and inability to produce forecasts beyond a very short time frame.

HETEROSCEDASTICITY

One of the assumptions of regression analysis is that the error terms in the population regression (εi) have a constant variance across all values of the independent variable (X). When this is true, the model is said to be *homoscedastic,* and if this assumption is violated, the model is termed *heteroscedastic.* With heteroscedasticity, the standard errors of the regression coefficients may be underestimated, causing the calculated t-ratios to be larger than they should be, which may lead us to conclude incorrectly that a variable is statistically significant. Heteroscedasticity is more common with cross-sectional data than with time-series data.

We can evaluate a regression model for heteroscedasticity by looking at a scatterplot of the residuals (on the vertical axis) versus the independent variable (on the horizontal axis). In an ideal model, the plot of the residuals would fall within a horizontal band, as shown in the top graph of Figure 4.13. This graph illustrates a residual pattern representative of homoscedasticity. A typical heteroscedastic situation is shown by the funnel-shaped pattern of residuals in the lower graph of Figure 4.13.

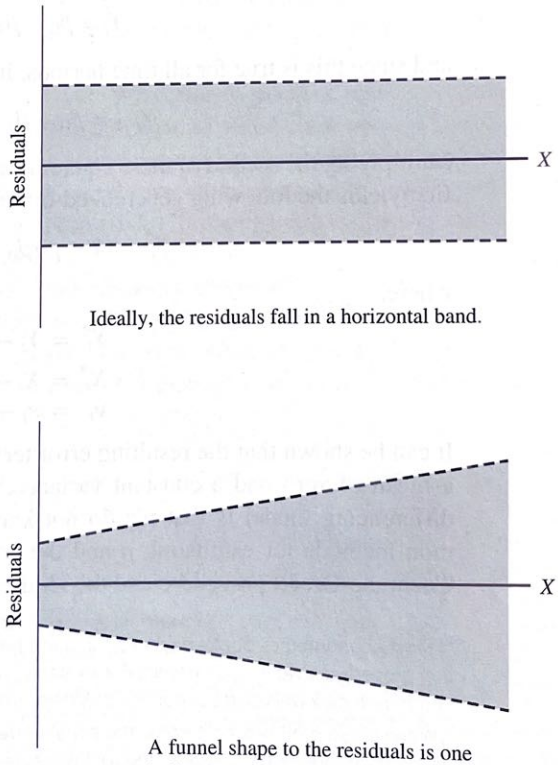


FIGURE 4.13
Residual Patterns Indicative of Homoscedasticity (Top Graph) and Heteroscedasticity (Bottom Graph).

One common way to reduce or eliminate a problem of heteroscedasticity is to use the logarithm of the dependent variable in the estimation of the regression model. This often works because the logarithms will have less overall variability than the raw data. A second possible solution would be to use a form of regression analysis other than the ordinary least-squares method. Discussion of such methods is beyond the scope of this text but can be found in many econometric texts.

CROSS-SECTIONAL FORECASTING

While most forecasting is based on time-series data, there are situations in which cross-sectional analysis is useful. In cross-sectional analysis, the data all pertain to one time period rather than a sequence of periods. Suppose, for example, that you are the sales manager for a firm that sells small specialty sandwiches through convenience stores. You currently operate in eight cities and are considering expanding into another. You have the data shown at the top of Table 4.13 for the most recent year's sales and the population of each city. You may try to predict sales based on population by using a bivariate regression model. The model may be written as:

$$\text{Sales} = b_0 + b_1(\text{POP})$$

Regression results for this model, given the eight data points just shown, are presented in the lower part of Table 4.13.

TABLE 4.13 Regression Results for Sales as a Function of Population

Population (000)	Sales (000)
505	372
351	275
186	214
175	135
132	81
115	144
108	90
79	97

Regression Statistic	
R-Square	0.914
Standard error	32.72
Observations	8

	Coefficient	Standard Error	*T*-test	*P*-value
Intercept	37.02	20.86	1.77	0.126
Population (000)	0.67	0.08	8.00	0.000

The Durbin-Watson statistic is not shown because it is not relevant for cross-sectional data. Indeed, the order in which the data are placed will change the Durbin-Watson statistics.

While most forecasting is based on time-series data, there are situations in which cross-sectional analysis is useful. In cross-sectional analysis, the data all pertain to one time period rather than a sequence of periods.

The statistical results show the expected positive sign for the coefficient of population. The critical value of t from the t-table at six degrees of freedom ($n - 2 = 6$) and a 5 percent significance level (one-tailed test) is 1.943. Since the calculated value for population is greater ($8.005 > 1.943$), we conclude that there is a statistically significant positive relationship between sales and population. The coefficient of determination (R-squared) is 0.914, which tells us that 91.4 percent of the variation in sales is explained by the variation in population.

The actual and predicted sales for each city are shown in the top graph of Figure 4.14. Visually, this graph helps you see that population size is a pretty good predictor of sales. In the lower graph of Figure 4.14, you see a scatter-gram depicting the relationship between sales (vertical axis) and population (horizontal axis).

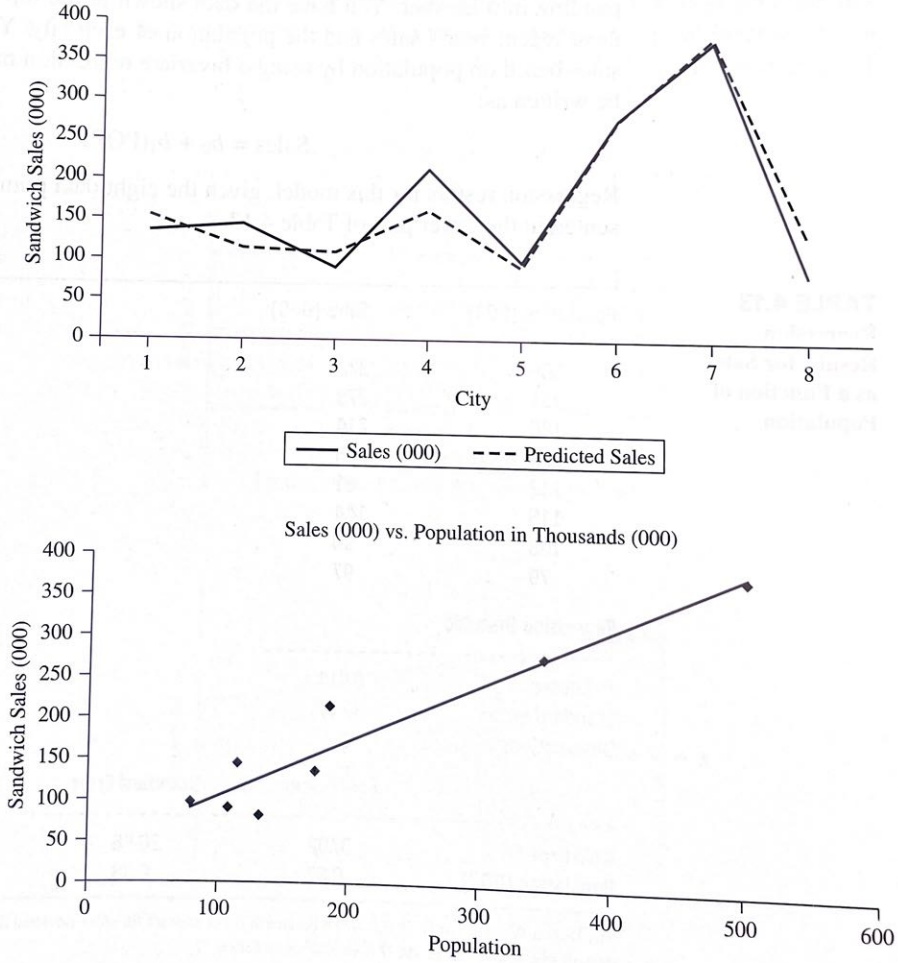


FIGURE 4.14
Sandwich Sales in Eight Cities The Upper Graph Shows How Well Population can Help Predict Sales. In the Lower Graph You See the Relationship Between Sales and Population in a Scatter-Gram.

Now, suppose that the city that you are considering expanding into has a population of 155,000. You can use the regression results to forecast sales as follows:

$$\begin{aligned} \text{Sales} &= 37.02 + 0.67(\text{POP}) \\ &= 37.02 + 0.67(155) \\ &= 140.87 \end{aligned}$$

Remember that sales are in thousands, so this is a point estimate of 140,870 sandwiches. An approximate 95 percent confidence interval estimate could be constructed as follows:

$$\begin{aligned} \text{Point estimate} \pm 2(\text{standard error of regression}) &= 140.870 \pm 2(32.72) \\ &= 140.870 \pm 65.44 \\ &= 75.43 \text{ to } 206.31 \end{aligned}$$

That is, about 75,430 to 206,310 sandwiches.

FORECASTING TOTAL HOUSES SOLD WITH TWO BIVARIATE REGRESSION MODELS

You may recall that the total houses sold (THS) series that we forecast in Chapters 1 and 3 showed quite a bit of variability, including a substantial seasonal component. Therefore, you might expect that it would be difficult to forecast such a series based on a simple regression equation with one causal variable. One thing that would make the process more workable would be to deseasonalize the THS data prior to attempting to build a regression model. In this section, we will look at THS data and on a quarterly basis using quarterly seasonal indices. The seasonal indices are shown in Table 4.14 along with the data.

In this section, we will first prepare a forecast of THS based solely on a simple linear trend; then we will do a second forecast using disposable personal income as a causal variable.

When seasonally adjusted monthly data for total houses sold (SATHS) are regressed as a function of a time index, where $t = 1$ for the first quater of 2010, the results are as shown in Figure 4.15. Data used to develop the model and forecast were from 2010 through 2015. The forecast was made 2016 on a quarterly basis. The equation for seasonally adjusted total houses sold is:

$$\text{SATHS} = 68.69 + 2.40(\text{Time})$$

The positive slope for time of 2.40 is logical, and from the t-ratio (11.21), we see that the slope is quite statistically significant in this model (the significance level, or p-value, is .000—even at a two-tailed level). The R-squared (R^2) tells us that 85.10 percent of the variation in seasonally adjusted total houses sold is explained by this model. We see that the Durbin-Watson test for serial correlation indicates positive serial correlation (DW = 1.11, where $1.11 < 1.27$).

To make a forecast of SATHS for 2016 with this model, we use time index values of 25, 26. 27, and 28. Doing so gives us the the dotted portion of the straight

TABLE 4.14 **Data and Seasonal Indices for Forecasting Total Houses Sold in Thousands and the Corresponding Seasonally Adjusted Total Houses Sold (SATHS). Data for 2010 through 2015 are used to make a forecast for 2016. The trend regression results for seasonally adjusted THS are shown in Figure 4.15 along with the forecast for 2016.** (c4t14&f15)

Date	THS (000)	Seasonal Indices	SATHS (000)
Feb-10	87	0.999	87.091
May-10	95	1.128	84.229
Aug-10	74	0.994	74.477
Nov-10	66	0.880	75.035
Feb-11	71	0.999	71.075
May-11	86	1.128	76.250
Aug-11	76	0.994	76.490
Nov-11	72	0.880	81.857
Feb-12	87	0.999	87.091
May-12	103	1.128	91.322
Aug-12	94	0.994	94.606
Nov-12	85	0.880	96.637
Feb-13	109	0.999	109.114
May-13	126	1.128	111.715
Aug-13	95	0.994	95.613
Nov-13	99	0.880	112.553
Feb-14	107	0.999	107.112
May-14	120	1.128	106.395
Aug-14	108	0.994	108.696
Nov-14	104	0.880	118.238
Feb-15	130	0.999	130.137
May-15	139	1.128	123.241
Aug-15	119	0.994	119.767
Nov-15	113	0.880	128.470
Feb-16	134	0.999	134.141
May-16	158	1.128	140.087
Aug-16	144	0.994	144.929
Nov-16	125	0.880	142.113

FIGURE 4.15 **Trend Forecast for Seasonally Adjusted Total Houses Sold (000)** The Straight Line in This Graph Represents the Forecast Values. It is Dotted in the 2016 Forecast Horizon and Dashed in the Historic Period. Data From 2010 Through 2015 Were Used to Develop the Forecast. (c4t14&f15)

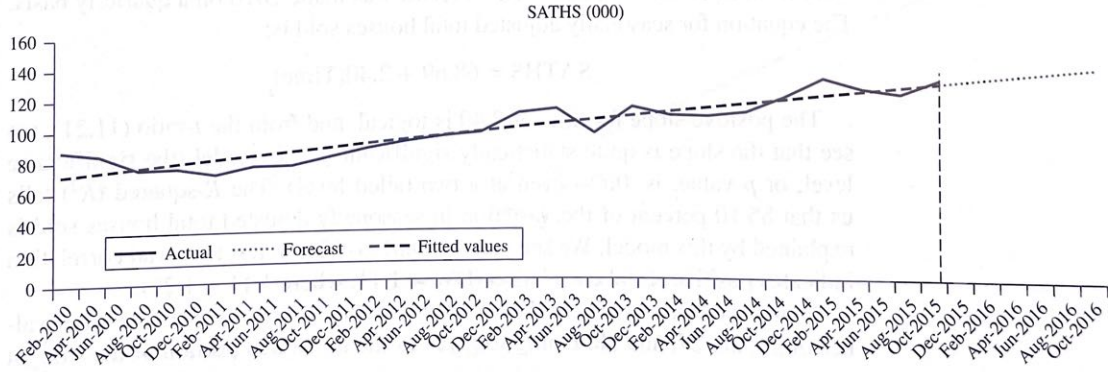


TABLE 4.15
Statistical Results for the Regression Time Trend for SATHS
(c4t14&f15)

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	F-test
Regression	6,600.56	1	6,600.56		125.60
Error	1,156.11	22	52.55	7.25	
Total	7,756.67	23			

Audit Trail— Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value
Intercept	68.69	3.05	22.49	0.00
Time Index (T)	2.40	0.21	11.21	0.00

Audit Trail—Statistics

Accuracy Measures	Value	Forecast Statistics	Value
MAPE	6.06%	Durbin Watson (1)	1.11
R-Square	85.10%		

line in Figure 4.15. Below are values for the seasonally adjusted total houses sold in the four quarters of 2016.

Feb-2016	128.54
May-2016	130.89
Aug-2016	133.24
Nov-2016	135.59

To get the forecast of nonseasonally adjusted values we multiply the seasonally adjusted forecast values by the corresponding seasonal index for each quarter.[12] This is shown below:

Date	SATHS Forecast	SI	THS Forecast
Feb-2016	128.54	0.999	128.402
May-2016	130.89	1.128	147.625
Aug-2016	133.24	0.994	132.386
Nov-2016	135.59	0.880	119.264

[12] The seasonal indices used are from a time-series decomposition of the data using ForecastX™. This will be discussed in Chapter 6.

The MAPE for the forecast period is:

Date	Actual 2016 THS	THS Forecast	Error	Absolute Error	Absolute % Error
Feb-2016	134	128.40	5.60	5.60	4.18
May-2016	158	147.63	10.37	10.37	6.57
Aug-2016	144	132.39	11.61	11.61	8.06
Nov-2016	125	119.26	5.74	5.74	4.59
				MAPE =	5.85%

What are the causal factors that you think would influence the sales of houses? You might come up with a fairly long list. Some of the variables that might be on such a list are:

Income
Unemployment rate
Interest or mortgage rates
Consumer attitudes[13]
Housing prices

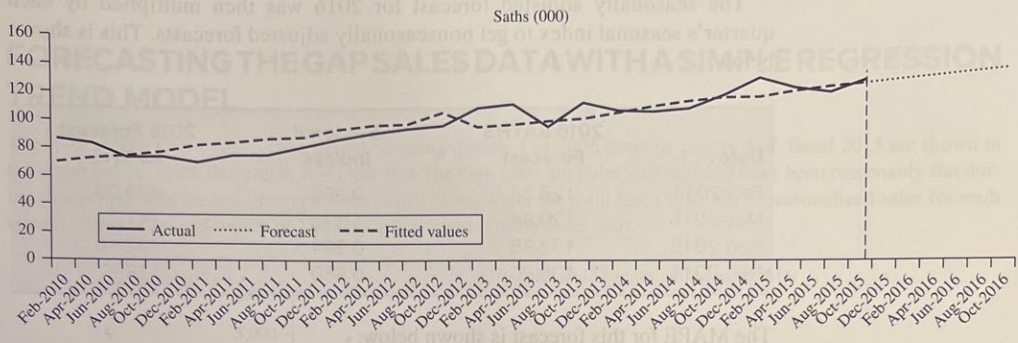
To develop a forecast of THS as a function of disposable personal income (DPI), the THS data were again deseasonalized; then those values were regressed as a function of DPI. These results are shown in Figure 4.16. The slope of 0.02 is logical since you would expect that more new houses would be sold as income increases. The *t*-value of 8.66 is very significant, as indicated by the two-tailed

[13] Consumer attitudes are often measured by the University of Michigan's Index of Consumer Sentiment. This is an index that is released each month by the University of Michigan Survey Research Center. Each month, 500 respondents in a national survey are interviewed about a variety of topics. There are five specific questions in the survey that go into the calculation of the Index of Consumer Sentiment, which has been adjusted to a base of 100 for 1966. Those five questions are:

1. We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?
2. Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or about the same as now?
3. Now turning to business conditions in the country as a whole—do you think that during the next 12 months we'll have good times financially, or bad times, or what?
4. Looking ahead, which would you say is more likely—that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?
5. About the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or bad time for people to buy major household items?

The way in which the index is computed makes it higher when people's responses to these questions are more positive.

FIGURE 4.16 Forecast of Seasonally Adjusted Total Houses Sold as a Function of Disposable Personal Income (DPI) Data for 2010 Through 2015 Were Used to Develop the Forecast. For 2016, Holt's Exponential Smoothing was Used to Forecast DPI. (c4t16&f16)



P-value of 0.00. The R^2 indicates that 77.33 percent of the variation in new houses sold is explained by this model. The equation for SATHS is:

$$\text{Seasonally adjusted total houses sold (000)} = 6.29 + 0.01(\text{DPI})$$

as shown in Figure 4.16.

The regression equation is: SATHS = −157.85 + 0.02*(DPI). The statistical results are in Table 4.16. In this case, the MAPE is 7.72 percent, and the explanatory

TABLE 4.16 Statistical Results for the Regression of SATHS as a Function of Disposable personal Income (DPI). (c4t15)

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	F-test
Regression	5,998.47	1	5,998.47		75.06
Error	1,758.20	22	79.92	8.94	
Total	7,756.67	23			

Audit Trail—Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value
Intercept	−157.85	29.66	−5.32	0.00
DPI (B$)	0.02	0.00	8.66	0.00

Audit Trail—Statistics

Accuracy Measures	Value	Forecast Statistics	Value
MAPE	7.72%	Durbin Watson (1)	1.09
R-Square	77.33%		

power (R2) is 77.33 percent. Again, there is positive serial correlation, as evidenced by the DW of 1.09. However, since the *t*-ratio (8.66) is so large, any bias may not be too bad.

The seasonally adjusted forecast for 2016 was then multiplied by each quarter's seasonal index to get nonseasonally adjusted forecasts. This is shown below:

Date	2016 SATHS Forecast	Seasonal Indices	2016 Forecast for THS
Feb-2016	128.33	0.999	128.20
May-2016	130.64	1.128	147.34
Aug-2016	132.95	0.994	132.10
Nov-2016	135.26	0.880	118.97

The MAPE for this forecast is shown below:

Date	Actual 2016 THS	THS Forecast	Error	Absolute Error	Absolute % Error
Feb-2016	134	128.200	5.800	5.800	4.328
May-2016	158	147.340	10.660	10.660	6.747
Aug-2016	144	132.100	11.900	11.900	8.264
Nov-2016	125	118.970	6.030	6.030	4.824
				MAPE =	6.04

Comments from the Field 1

While working for Dow Plastics, a business group of the Dow Chemical Company, Jan Neuenfeldt received on-the-job training while assisting others in developing forecasts. This led her to enroll in an MBA forecasting class in which she obtained formal training in quantitative forecasting methods.

The methodology that Jan uses most is regression analysis. On occasion, she also uses exponential smoothing models, such as Winters'. However, the marketing and product managers who use the forecasts usually are interested in *why* as well as in the forecast values. Most of the forecasts Jan prepares are on a quarterly basis. It is fairly typical for annual forecasts one year out to be within a 5 percent margin of error. For large-volume items in mature market segments, the annual margin for error is frequently only about 2 percent.

Each quarter, Jan reports forecast results to management, using a newsletter format. She begins with an exposition of the results, followed by the supporting statistical information and a graphic presentation of the forecast. She finds that graphics are extremely useful as she prepares forecasts, as well as when she communicates results to end users.

Source: This comment is based on an interview with Jan Neuenfeldt.

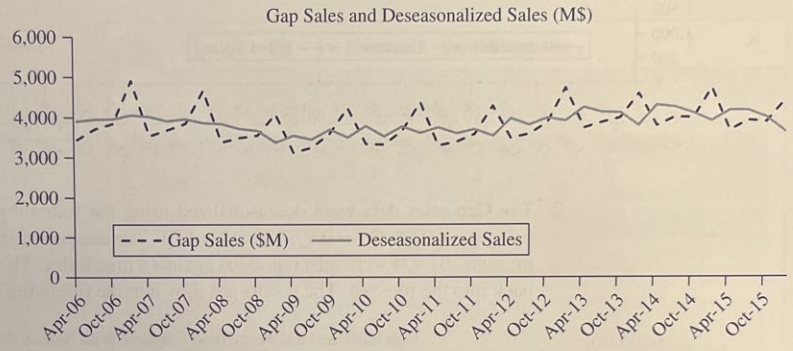
Integrative Case

The Gap

FORECASTING THE GAP SALES DATA WITH A SIMPLE REGRESSION TREND MODEL

The sales of The Gap stores for the period covering quarter 1 of 2006 through quarter 4 of fiscal 2015 are shown in the graph below. From this graph, it is clear that The Gap sales are quite seasonal and have been reasonably flat during this period. The dashed line represents actual sales, while the solid line shows the deseasonalized sales for each quarter. April is the end month of The Gap's first quarter of its fiscal year.

(c4Gap)



Case Questions

1. Do you think that the general growth path of The Gap sales has followed a linear path over the period shown? As part of your answer, show a graph of the deseasonalized The Gap sales along with a linear trend line. What does this graph suggest to you about the results you might expect from using a linear trend as the basis of a forecast of The Gap sales for 2017? The deseasonalized sales can be calculated using the following seasonal indices:

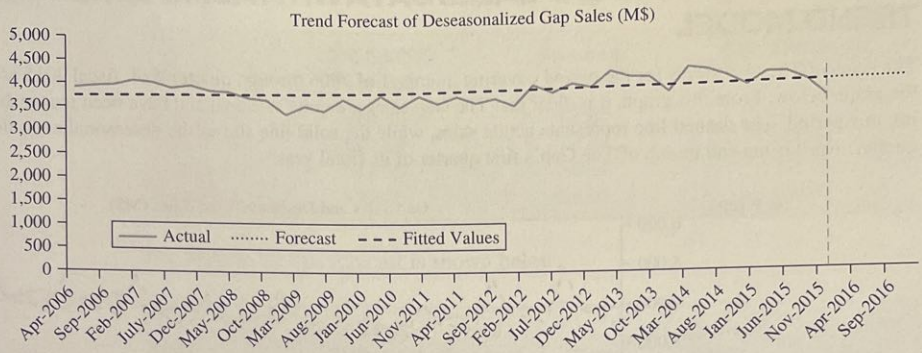
Seasonal Indexes	
Index 1 April	0.88
Index 2 July	0.94
Index 3 October	0.97
Index 4 January	1.21

2. Use a regression of deseasonalized The Gap sales as a function of a linear regression time trend as the basis for a forecast of The Gap sales for 2017. Be sure to reseasonalize your forecast; then graph the actual The Gap sales along with your forecast. What do you think about this forecast based on your graph?
3. Calculate the MAPE for both the 2016 fiscal year using the quarterly data.

Solutions to Case Questions

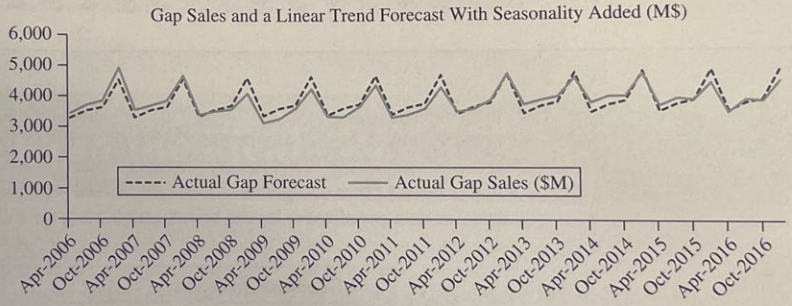
1. When The Gap sales data are deseasonalized and a linear trend is plotted through the deseasonalized series, it becomes clear that the trend in sales was essentailly stationary. This can be seen in the graph below, in which actual sales (seasonally adjusted) are graphed along with the trend line.

(c4Gap)



2. The Gap sales data were deseasonalized using the following seasonal indices: Q1 = 0.88, Q2 = 0.94, Q3 = 0.97, and Q4 = 1.21. The deseasonalized sales data for 2006Q1 through 2015Q4 were then regressed against a time index. The seasonality was then put back into the forecast. The results are shown in the following graph.

(c4Gap)



3. The MAPE calculation for the forecast year (for quarters) is as follows:

Date	Actual Gap Sales ($M)	Actual Gap Forecast (M$)	Error	Absolute Error	Absolute % Error
Apr-2016	3438	3,496.29	-58.29	58.29	1.70
Jul-2016	3851	3,740.19	110.81	110.81	2.88
Oct-2016	3798	3,865.25	-67.25	67.25	1.77
Jan-2017	4429	4,828.69	-399.69	399.69	9.02
				MAPE =	3.84

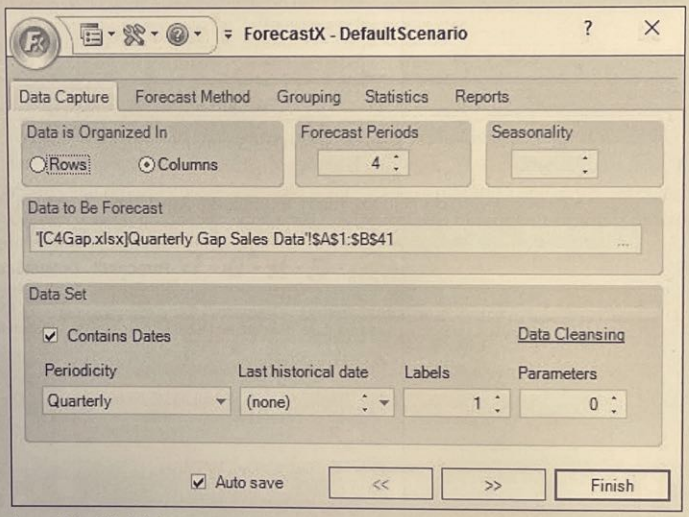
USING FORECASTX™ TO MAKE REGRESSION FORECASTS

What follows is a brief discussion of how to use ForecastX™ for making a forecast based on a regression model. This will increase your familiarity with the of use of ForecastX™. The illustration used here is for a trend forecast.

First, put your data into an Excel spreadsheet in column format, such as the sample of The Gap data shown in the table below. Once you have your data in this format, while in Excel place your cursor in any data cell such as B6. Then go to Add-ins to open ForecastX™. The following dialog box appears.

A Sample of The Gap Data in Column Format

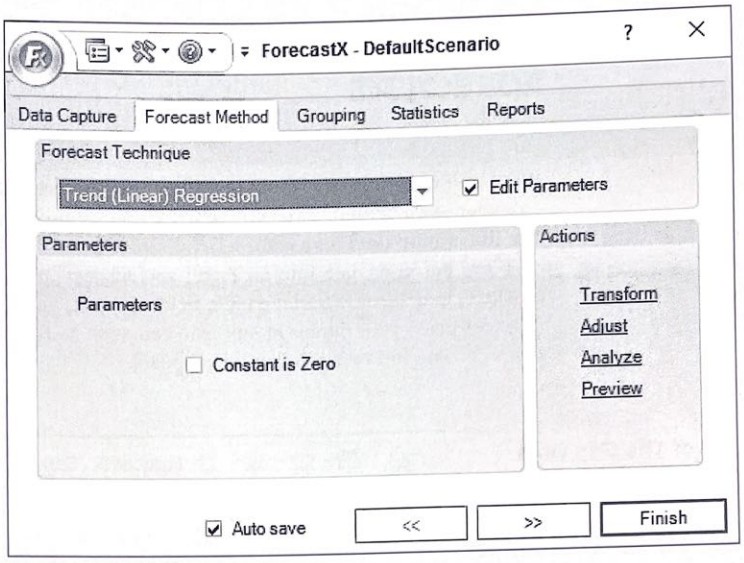
Date	Gap Sales ($M)
Apr-06	3441
Jul-06	3,714
Oct-06	3,851
Jan-07	4,919
Apr-07	3,549
Jul-07	3,685
Oct-07	3,854
Jan-08	4,675
Apr-08	3,384
Jul-08	3,499
Oct-08	3,561
Jan-09	4,082
Apr-09	3,127
Jul-09	3,245
Oct-09	3,589
Jan-10	4,236
Apr-10	3,329
Jul-10	3,317
Oct-10	3,654
Jan-11	4,364
Apr-11	3,295
Jul-11	3,386
Oct-11	3,585
Jan-12	4,283
Apr-12	3,487
Jul-12	3,575
Oct-12	3,864



Source: John Galt Solutions

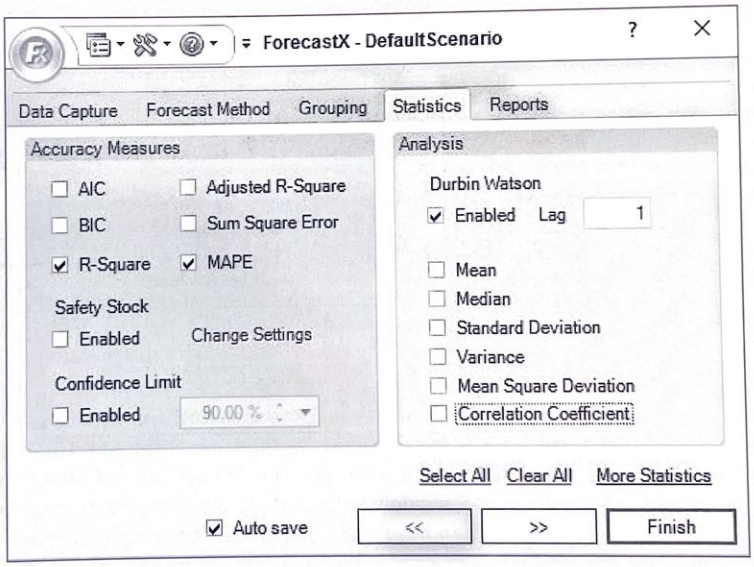
Verify the periodicity of your data (**Quarterly** for this example); then click the **Forecast Method** tab at the top. Click the down arrow in the **Forecasting Technique** window, and select **Trend (Linear) Regression.** The following window will result. In the terminology used by ForecastX™, *linear regression* refers to a method that makes a regression trend forecast. If you want to develop a causal regression model, select **Multiple Regression.** (See the "Further Comments on Regression Models" section on page 210.)

After selecting **Trend (Linear) Regression,** the dialog box will then look like the one below.



Source: John Galt Solutions

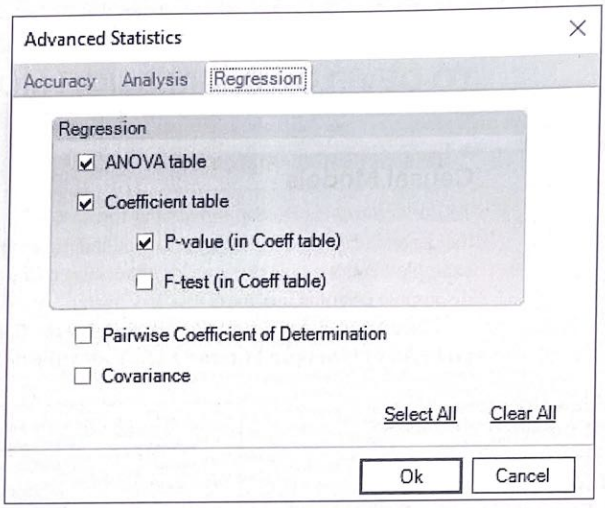
Now you are ready to click the **Statistics** tab, which will take you to the next dialog box.



Source: John Galt Solutions

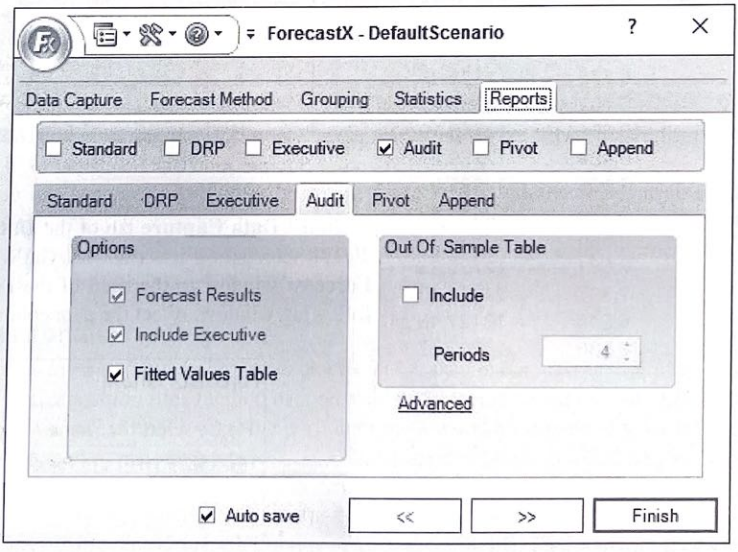
Here you want to select the desired statistics. Often the ones selected in this example would be what you would want for simple regression models.

In addition, you will want to click the **More Statistics** button at the bottom and check the box for **P-value** (in **Coeff table**) under the **Regression** tab. Look at the other tabs in this box and select desired statistics. Then click **OK** and you return to the **Statistics** box.



Source: John Galt Solutions

lext click the **Reports** tab to get the **Reports** dialog box. This is where you select the cular reports and report contents that you want. Some exploration and experimentation these options will help you see what each option leads to in terms of results. Clicking e Audit report yields the following:



Source: John Galt Solutions

ter you click **Finish** in the lower right corner, reports will be put in new Excel work- —Book 2, Book 3, and so forth. The book numbers will vary depending on what you een doing in Excel up to that point.

FURTHER COMMENTS ON USING FORECASTX™ TO DEVELOP REGRESSION MODELS

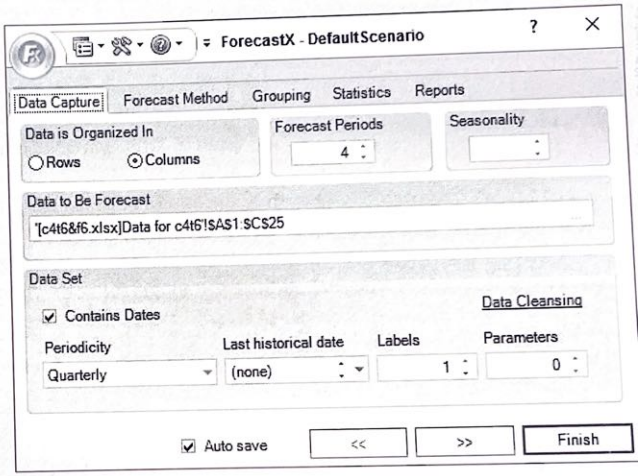
Causal Models

To do a *causal regression model* and forecast, select the data sheet with dates in column A, the dependent variable in column B, and independent variables in columns C, D, E, etc. In our example here, we will use jewelry store sales as a function of only one independent variable, disposable personal income (DPI). In Chapter 5, we will expand to more independent variables.

Place your cursor in a cell with data for the dependent variable, such as cell B4. Then go to Add-ins to open ForecastX™. You will get the following dialog box:

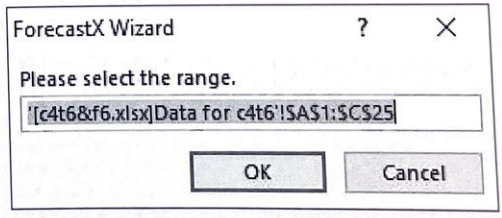
A sample of jewelry store sales data and DPI follows:

Date	Jewelry Store Sales (M$)	DPI
Feb-10	5,372	11041.47
May-10	5,841	11197.63
Aug-10	5,311	11286.63
Nov-10	9,385	11425.73
Feb-11	5,823	11652.23
May-11	6,911	11751.63
Aug-11	6,243	11876.67
Nov-11	10,073	11924.93
Feb-12	6,337	12189.97
May-12	6,772	12321.33
Aug-12	6,102	12355.43
Nov-12	10,651	12748.13
Feb-13	6,484	12259.27
May-13	7,106	12335.9
Aug-13	6,175	12453.83
Nov-13	10,686	12534.33
Feb-14	6,681	12735.83
May-14	7,397	12962.43
Aug-14	6,548	13127.43
Nov-14	10,586	13265.27



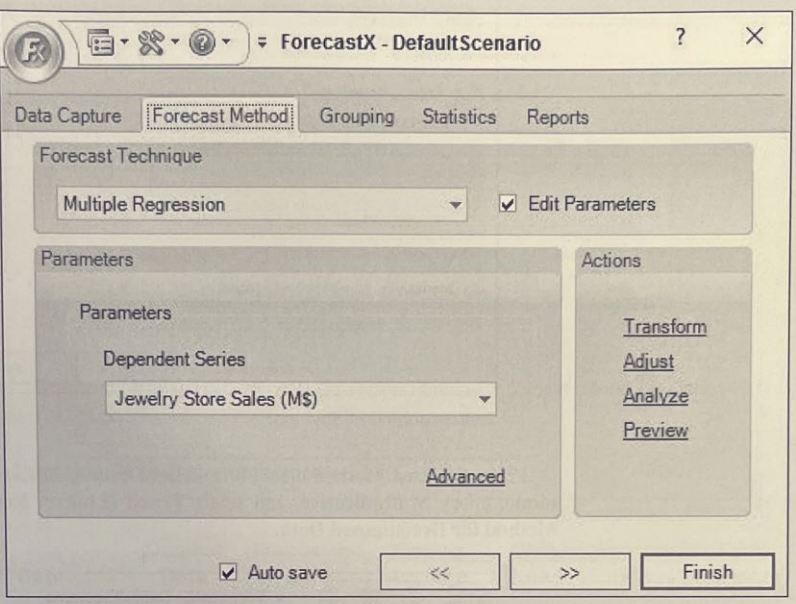
Source: John Galt Solutions

In the **Data Capture** tab of the **Data Capture** window, look at th default selection. If it is not what you want, click inside the **Data To B Forecast** window to the right of the **Data to Be Forecast** box. In th following window, select the data columns you want, then click **OK.**



Source: John Galt Solutions

Next click the **Forecast Method** tab and select **Multiple Regression** in the **Forecasting Technique** window. Under parameters, in the **Dependent Series** window select the variable you want to forecast (**Jewelry Store Sales** in this example).



Source: John Galt Solutions

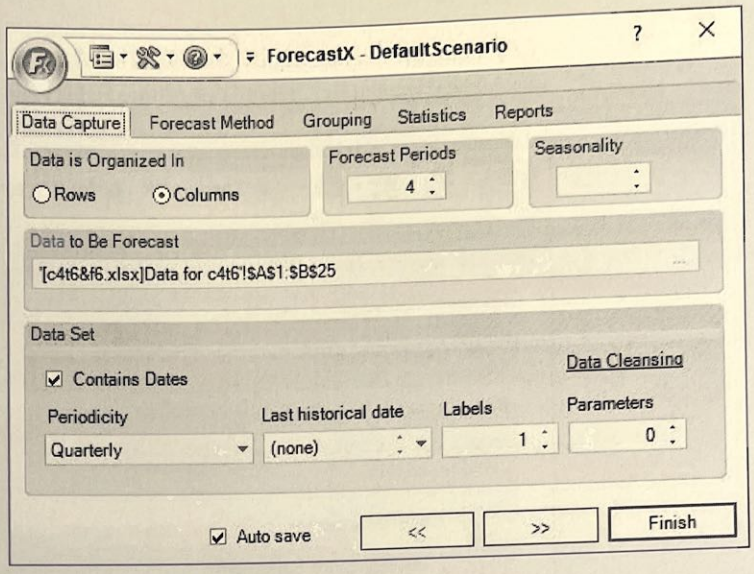
From this point on, you follow the same selections as described above for regression trend forecasts.

You are probably wondering how you forecast the independent variable into the future and unknown forecast horizon. You can use any acceptable method to do this, but ForecastX™ makes it easy by doing an automated forecast using a procedure called ProCast™.

Deseasonalizing Data

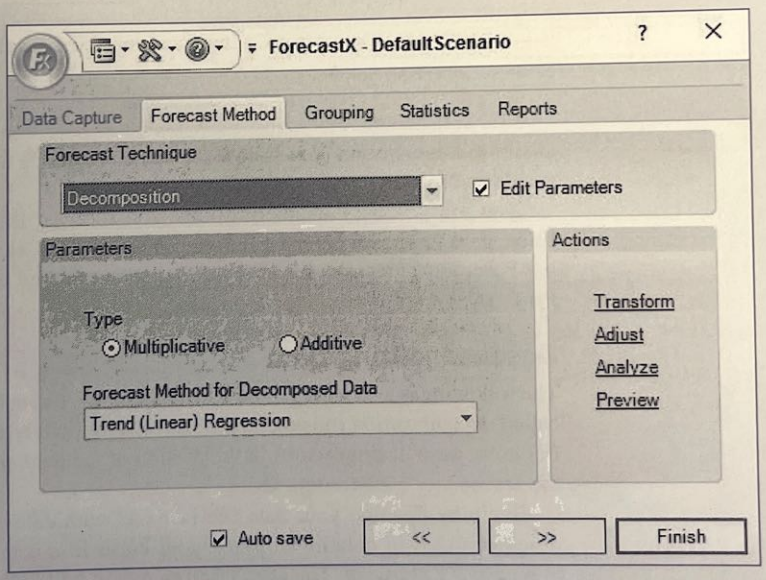
The following is a summary of how to deseasonalize data in ForecastX™. We use a method called *decomposition* (this method of forecasting will be discussed in detail in Chapter 6). For now, we will simply look at the portion of the method and results that we need to take the seasonality out of a data series.

Begin by opening your data file in Excel with dates in column A and the data to be deseasonalized in column B. Place your curor in a data cell such as B5. Then start the ForecastX™ software. The Data Capture screen will look like the following (just check to be sure everything is correct; it is almost always correct):



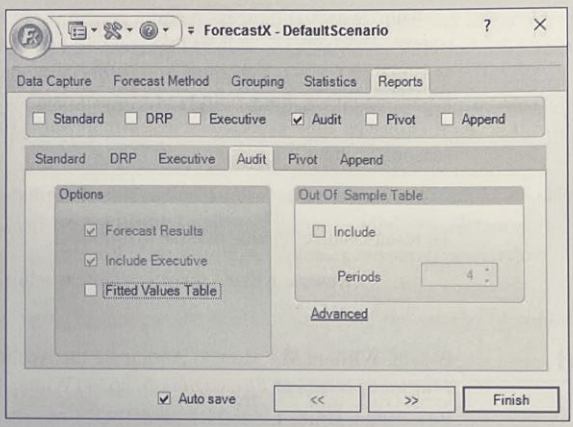
Source: John Galt Solutions

In the **Forecast Method** dialog box, select **Decomposition** as the **Forecasting Technique,** check **Multiplicative,** and select **Trend (Linear) Regression** as the **Forecast Method for Decomposed Data.**



Source: John Galt Solutions

For this application, we do not care about statistics, so skip that tab. Click the **Reports** tab, and select only the **Audit** report. There is no need now to ask for the Fitted Values Table.



Source: John Galt Solutions

Now click **Finish,** and you will get results that will include the following near the bottom of the page:

Components of Decomposition

Date	Original Data	Forecasted Data	Centered Moving Average	CMA Trend	Seasonal Indices	Cycle Factors
Feb-2010	5,372.00				0.85	
May-2010	5,841.00				0.94	
Aug-2010	5,311.00	5,389.84	6,533.63	6,913.58	0.82	0.95
Nov-2010	9,385.00	9,295.64	6,723.75	6,971.31	1.38	0.96
Feb-2011	5,823.00	5,933.94	6,974.00	7,029.04	0.85	0.99
May-2011	6,911.00	6,758.02	7,176.50	7,086.76	0.94	1.01
Aug-2011	6,243.00	6,044.12	7,326.75	7,144.49	0.82	1.03
Nov-2011	10,073.00	10,194.10	7,373.63	7,202.22	1.38	1.02
Feb-2012	6,337.00	6,244.19	7,338.63	7,259.94	0.85	1.01
May-2012	6,772.00	6,962.13	7,393.25	7,317.67	0.94	1.01
Aug-2012	6,102.00	6,173.74	7,483.88	7,375.40	0.82	1.01
Nov-2012	10,651.00	10,429.64	7,544.00	7,433.12	1.38	1.01
Feb-2013	6,484.00	6,462.22	7,594.88	7,490.85	0.85	1.01
May-2013	7,106.00	7,164.71	7,608.38	7,548.58	0.94	1.01
Aug-2013	6,175.00	6,300.37	7,637.38	7,606.30	0.82	1.00
Nov-2013	10,686.00	10,643.07	7,698.38	7,664.03	1.38	1.00
Feb-2014	6,681.00	6,620.91	7,781.38	7,721.76	0.85	1.01
May-2014	7,397.00	7,359.75	7,815.50	7,779.48	0.94	1.00
Aug-2014	6,548.00	6,430.40	7,795.00	7,837.21	0.82	0.99
Nov-2014	10,586.00	10,779.42	7,797.00	7,894.94	1.38	0.99
Feb-2015	6,617.00	6,640.69	7,804.63	7,952.66	0.85	0.98
May-2015	7,477.00	7,382.12	7,839.25	8,010.39	0.94	0.98
Aug-2015	6,529.00	6,559.58		8,068.12	0.82	0.99
Nov-2015	10,882.00	11,046.36		8,125.84	1.38	0.98

The "Seasonal Indices" column heading is in bold here but will not be in bold in your output. These are the indices that you will use to deseasonalize the original data and to reseasonalize results. You should copy this column of seasonal indices and paste it into your Excel workbook along with your original data.

You can now calculate a deseasonalized series by dividing the original data by the seasonal indices.

$$\text{Deseasonalized series} = \text{Original series} \div \text{Seasonal indices}$$

To reseasonalize results, reverse the process.

$$\text{Reseasonalized results} = \text{Deseasonalized results} \times \text{Seasonal indices}$$

Suggested Readings

Bassin, William M. "How to Anticipate the Accuracy of a Regression Based Model." *Journal of Business Forecasting* 6, no. 4 (Winter 1987–88), pp. 26–28.

Bowerman, Bruce L.; and Richard T. O' Connell. *Applied Statistics: Improving Business Processes.* Chicago: Irwin, 1997.

Bowerman, Bruce L.; Richard T. O' Connell; and J. B. Orris. *Essentials of Business Statistics.* Boston: Irwin/McGraw-Hill, 2004.

Dalrymple, Douglas J.; William M. Strahle; and Douglas B. Bock. "How Many Observations Should Be Used in Trend Regression Forecasts?" *Journal of Business Forecasting* 8, no. 1 (Spring 1989), pp. 7–9.

Harris, John L.; and Lon-Mu Liu. "GNP as a Predictor of Electricity Consumption." *Journal of Business Forecasting,* Winter 1990–91, pp. 24–27.

Lapide, Larry. "Do You Need to Use Causal Forecasting?" *Journal of Business Forecasting,* Summer 1999, pp. 13–14.

Lind, Douglas A.; Robert D. Mason; and William G. Marchal. *Basic Statistics for Business and Economics,* 3rd ed. New York: Irwin/McGraw-Hill, 2000.

Meade, Nigel; and Towhidul Islam. "Forecasting with Growth Curves: An Empirical Comparison." *International Journal of Forecasting* 11, no. 2 (June 1995), pp. 199–215.

Monaco, Ralph M. "MEXVAL: A Simple Regression Diagnostic Tool." *Journal of Business Forecasting,* Winter 1989–90, pp. 23–27.

Morrison, Jeffrey S. "Target Marketing with Logit Regression." *Journal of Business Forecasting* 14, no. 4 (Winter 1995–96), pp. 10–12.

Pindyck, Robert S.; and Daniel L. Rubinfeld. *Econometric Models and Economic Forecasts.* 3rd ed. New York: McGraw-Hill, 1991.

Wang, George C. S.; and Charles K. Akabay. "Heteroscedasticity: How to Handle in Regression Modeling." *Journal of Business Forecasting* 13, no. 2 (Summer 1992), pp. 11–17.

West, Kenneth D.; et al. "Regression-Based Tests of Predictive Ability." *International Economic Review* 39, no. 4 (November 1998), pp. 817–40.

Wooldridge, Jeffrey M. *Introductory Econometrics.* 2nd ed. Mason, OH: Thompson/South-Western, 2003.

Exercises

1. Why is it useful to look at data in a graph as well as in a table? What is the main advantage of seeing a graph of the data?
2. For what kind of data pattern is a linear regression model most applicable? Give an example based on data used in this chapter.

3. How can seasonal data be forecast with a simple bivariate linear regression model? Explain the deseasonalize-forecast-reseasonalize process. How does the material in this chapter suggest that you find seasonal indices?
4. In this chapter, you learned four steps that should be used to evaluate a regression model. What is the first step and why is it so important? Explain the other three steps, indicating what you learn from each of those three steps.
5. Explain the difference between a simple trend model and a causal model.
6. Explain the difference between the most common kind of correlation (the Pearson product moment correlation, discussed in Chapter 2) and serial correlation.
7. Explain what is meant by heteroscedasticity.
8. The following regression results relate to a study of the salaries of public school teachers in a midwestern city (the sample size was 450 teachers):

Variable	Coefficient	Standard Error	*t*-ratio
Constant	20,720	6,820	3.04
EXP	805	258	

R-squared = 0.684; $n = 105$.
Standard error of the estimate = 2,000.
EXP is the experience of teachers in years of full-time teaching.

   a. What is the *t*-ratio for EXP? Does it indicate that experience is a statistically significant determinant of salary if a 95 percent confidence level is desired?
   b. What percentage of the variation in salary is explained by this model?
   c. Determine the point estimate of a salary for a teacher with 20 years of experience.
   d. What is the approximate 95 percent confidence interval for your point estimate from part (*c*)?
9. Nelson Industries manufactures a part for a type of aircraft engine that is becoming obsolete. The sales history for the last 10 years is as follows:

(c4p9)

Year	Sales
2008	945
2009	875
2010	760
2011	690
2012	545
2033	420
2014	305
2015	285
2016	250
2017	210

   a. Plot sales versus time.
   b. Estimate the regression model for a linear time trend of sales.
   c. What is the mean absolute percent error of the linear regression estimates for these 10 years?
   d. Using this model, estimate sales for year 11.

10. Mid-Valley Travel Agency (MVTA) has offices in 12 cities. The company believes that its monthly airline bookings are related to the mean income in those cities and has collected the following data:

(c4p10)

Location	Bookings	Income
1	1,098	$43,299
2	1,131	45,021
3	1,120	40,290
4	1,142	41,893
5	971	30,620
6	1,403	48,105
7	855	27,482
8	1,054	33,025
9	1,081	34,687
10	982	28,725
11	1,098	37,892
12	1,387	46,198

a. Develop a linear regression model of monthly airline bookings as a function of income.

b. Use the process described in the chapter to evaluate your results.

c. Make the point and approximate 95 percent confidence interval estimates of monthly airline bookings for another city in which MVTA is considering opening a branch, given that income in that city is $39,020.

11. Barbara Lynch is the product manager for a line of skiwear produced by HeathCo Industries and privately branded for sale under several different names, including Northern Slopes and Jacque Monri. A new part of Ms. Lynch's job is to provide a quarterly forecast of sales for the northern United States, a region composed of 27 states stretching from Maine to Washington. A 10-year sales history is shown:

(c4p11)

	Sales ($000)			
Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2007	$ 72,962	$ 81,921	$ 97,729	$ 142,161
2008	145,592	117,129	114,159	151,402
2009	153,907	100,144	123,242	128,497
2010	176,076	180,440	162,665	220,818
2011	202,415	211,780	163,710	200,135
2012	174,200	182,556	198,990	243,700
2013	253,142	218,755	225,422	253,653
2014	257,156	202,568	224,482	229,879
2015	289,321	266,095	262,938	322,052
2016	313,769	315,011	264,939	301,479

a. Because Ms. Lynch has so many other job responsibilities, she has hired you to help with the forecasting effort. First, she would like you to prepare a time-series plot of the data and to write her a memo indicating what the plot appears to show and whether it seems likely that a simple linear trend would be useful in preparing forecasts.

b. In addition to plotting the data over time, you should estimate the least-squares trend line in the form:

$$\text{SALES} = a + b(\text{TIME})$$

Set TIME = 1 for 2007Q1 through TIME = 40 for 2016Q4. Write the trend equation:

$$\text{SALES} = \_\_\_\_\_ +/- \_\_\_\_\_(\text{TIME})$$

(Circle + or − as appropriate)

c. Do your regression results indicate to you that there is a significant trend to the data? Explain why or why not.

d. On the basis of your results, prepare a forecast for the four quarters of 2017.

Period	TIME	Sales Forecast (F1)
2017Q1	41	
2017Q2	42	
2017Q3	43	
2017Q4	44	

e. A year later, Barbara gives you a call and tells you that the actual sales for the four quarters of 2017 were: Q1 = 334,271, Q2 = 328,982, Q3 = 317,921, and Q4 = 350,118. How accurate was your model? What was the mean absolute percentage error (MAPE)?

12. Dick Staples, another product manager with HeathCo (see Exercise 11), has mentioned to Barbara Lynch that he has found both the unemployment rate and the level of income to be useful predictors for some of the products under his responsibility.

a. Suppose that Ms. Lynch provides you with the following unemployment data for the northern region she is concerned with:

(c4p12)

	Unemployment Rate (%)			
Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2007	8.4%	8.2%	8.4%	8.4%
2008	8.1	7.7	7.5	7.2
2009	6.9	6.5	6.5	6.4
2010	6.3	6.2	6.3	6.5
2011	6.8	7.9	8.3	8.0
2012	8.0	8.0	8.0	8.9
2013	9.6	10.2	10.7	11.5
2014	11.2	11.0	10.1	9.2
2015	8.5	8.0	8.0	7.9
2016	7.9	7.9	7.8	7.6

b. Using Excel, plot a scattergram of SALES versus northern-region unemployment rate (NRUR). Does there appear to be a relationship? Explain.

c. Prepare a bivariate regression model of sales as a function of NRUR in the following form:

$$\text{SALES} = a + b(\text{NRUR})$$

Write your answer in the following equation:

$$\text{SALES} = \_\_\_\_\_ +/- \_\_\_\_\_(\text{NRUR})$$

(Circle + or − as appropriate)

d. Write a memo to Ms. Lynch in which you evaluate these results and indicate how well you think this model would work in forecasting her sales series.

e. Use the model to make a forecast of sales for each quarter of 2017, given the forecast for unemployment (FNRUR) that HeathCo has purchased from a macroeconomic consulting firm (MacroCast):

Period	FNRUR	Sales Forecast (F2)
2017Q1	7.6%	_____
2017Q2	7.7	_____
2017Q3	7.5	_____
2017Q4	7.4	_____

f. For the actual sales given in Exercise 11(c), calculate the MAPE for this model. How does it compare with what you found in Exercise 11(*c*)?

g. Barbara Lynch also has data on income (INC), in billions of dollars, for the region as follows:

(c4p12)

	Income ($ Billions)			
Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2007	$ 218	$ 237	$ 263	$ 293
2008	318	359	404	436
2009	475	534	574	622
2010	667	702	753	796
2011	858	870	934	1,010
2012	1,066	1,096	1,162	1,187
2013	1,207	1,242	1,279	1,318
2014	1,346	1,395	1,443	1,528
2015	1,613	1,646	1,694	1,730
2016	1,755	1,842	1,832	1,882

Using Excel, plot a scattergram of SALES with INC. Does there appear to be a relationship? Explain.

h. Prepare a bivariate regression model of SALES as a function of income (INC) and write your results in the equation:

$$\text{SALES} = a + b(\text{INC})$$

$$\text{SALES} = \_\_\_\_\_ +/- \_\_\_\_\_(\text{INC})$$

(Circle + or − as appropriate)

i. Write a memo to Ms. Lynch in which you explain and evaluate this model, indicating how well you think it would work in forecasting sales.

j. HeathCo has also purchased a forecast of income from MacroCast. Use the following income forecast (INCF) to make your own forecast of SALES for 2017:

Period	INCF	Sales Forecast (F3)
2017Q1	$ 1,928	_____
2017Q2	1,972	_____
2017Q3	2,017	_____
2017Q4	2,062	_____

k. On the basis of the actual sales given in Exercise 11(*c*), calculate the MAPE for this model. How does it compare with the other two models you have used to forecast sales?

l. Prepare a time-series plot with actual sales for 2007Q1 through 2016Q4 along with the sales forecast you found in part (*j*) of this exercise. To accompany this plot, write a brief memo to Ms. Lynch in which you comment on the strengths and weaknesses of the forecasting model.

13. Carolina Wood Products, Inc., a major manufacturer of household furniture, is interested in predicting expenditures on furniture (FURN) for the entire United States. It has the following data by quarter for 2007 through 2016:

(c4p13)

	FURN (in $ Billions)			
Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2007	$ 98.1	$ 96.8	$ 96.0	$ 95.0
2008	93.2	95.1	96.2	98.4
2009	100.7	104.4	108.1	111.1
2010	114.3	117.2	119.4	122.7
2011	125.9	129.3	132.2	136.6
2012	137.4	141.4	145.3	147.7
2013	148.8	150.2	153.4	154.2
2014	159.8	164.4	166.2	169.7
2015	173.7	175.5	175.0	175.7
2016	181.4	180.0	179.7	176.3

a. Prepare a naive forecast for 2017Q1 based on the following model (see Chapter 1):

$$\text{NFURN}_t = \text{FURN}_{t-1}$$

Period	Naive Forecast
2017Q1	

b. Estimate the bivariate linear trend model for the data where TIME = 1 for 2007Q1 through TIME = 40 for 2016Q4.

$$\text{FURN} = a + b(\text{TIME})$$

$$\text{FURN} = \_\_\_\_\_ +/- \_\_\_\_\_(\text{TIME})$$

(Circle + or − as appropriate)

c. Write a paragraph in which you evaluate this model, with particular emphasis on its usefulness in forecasting.

d. Prepare a time-trend forecast of furniture and household equipment expenditures for 2017 based on the model in part (*b*).

Period	TIME	Trend Forecast
2017Q1	41	_____
2017Q2	42	_____
2017Q3	43	_____
2017Q4	44	_____

e. Suppose that the actual values of FURN for 2017 were as shown in the following table. Calculate the MAPE for both of your forecasts and interpret the results. (For the naive forecast, there will be only one observation, for 2017Q1.)

Period	Actual FURN ($ Billions)
2017Q1	177.6
2017Q2	180.5
2017Q3	182.8
2017Q4	178.7

14. Fifteen midwestern and mountain states have united in an effort to promote and forecast tourism. One aspect of their work has been related to the dollar amount spent per year on domestic travel (DTE) in each state. They have the following estimates for disposable personal income per capita (DPI) and DTE:

(c4p14)

State	DPI	DTE ($ Millions)
Minnesota	$ 17,907	$ 4,933
Iowa	15,782	1,766
Missouri	17,158	4,692
North Dakota	15,688	628
South Dakota	15,981	551
Nebraska	17,416	1,250
Kansas	17,635	1,729
Montana	15,128	725
Idaho	15,974	934
Wyoming	17,504	778
Colorado	18,628	4,628
New Mexico	14,587	1,724
Arizona	15,921	3,836
Utah	14,066	1,757
Nevada	19,781	6,455

a. From these data, estimate a bivariate linear regression equation for domestic travel expenditures (DTE) as a function of disposable income per capita (DPI):

$$\text{DTE} = a + b(\text{DPI})$$

$$\text{DTE} = \_\_\_\_\_ +/- \_\_\_\_\_(\text{DPI})$$

(Circle + or − as appropriate)

Evaluate the statistical significance of this model.

b. Illinois, a bordering state, has asked that this model be used to forecast DTE for Illinois under the assumption that DPI will be \$19,648.

c. Given that actual DTE turned out to be \$7,754 (million), calculate the percentage error in your forecast.

15. Collect data on population for your state (http://www.economagic.com may be a good source for these data) over the past 20 years and use a bivariate regression trend line to forecast population for the next five years. Prepare a time-series plot that shows both actual and forecast values. Do you think the model looks as though it will provide reasonably accurate forecasts for the five-year horizon? (c4p15)

16. AmerPlas, Inc., produces 16-ounce plastic drinking cups that are embossed with the names of prominent beers and soft drinks. It has been observed that sales of the cups match closely the seasonal pattern associated with beer production but that, unlike beer production, there has been a positive trend over time. The sales data, by month, for 2013 through 2016 are as follows:

(c4p16)

Period	T	Sales	Period	T	Sales
2013M01	1	857	2015M01	25	1,604
2013M02	2	921	2015M02	26	1,643
2013M03	3	1,071	2015M03	27	1,795
2013M04	4	1,133	2015M04	28	1,868
2013M05	5	1,209	2015M05	29	1,920
2013M06	6	1,234	2015M06	30	1,953
2013M07	7	1,262	2015M07	31	1,980
2013M08	8	1,258	2015M08	32	1,989
2013M09	9	1,175	2015M09	33	1,897
2013M10	10	1,174	2015M10	34	1,910
2013M11	11	1,123	2015M11	35	1,854
2013M12	12	1,159	2015M12	36	1,957
2014M01	13	1,250	2016M01	37	1,955
2014M02	14	1,289	2016M02	38	2,008
2014M03	15	1,448	2016M03	39	2,171
2014M04	16	1,497	2016M04	40	2,202
2014M05	17	1,560	2016M05	41	2,288
2014M06	18	1,586	2016M06	42	2,314
2014M07	19	1,597	2016M07	43	2,343
2014M08	20	1,615	2016M08	44	2,339
2014M09	21	1,535	2016M09	45	2,239
2014M10	22	1,543	2016M10	46	2,267
2014M11	23	1,493	2016M11	47	2,206
2014M12	24	1,510	2016M12	48	2,226

a. Use these data to estimate a linear time trend as follows:

$$\text{SALES} = a + b(T)$$

$$\text{SALES} = \_\_\_\_\_ +/- \_\_\_\_\_(T)$$

(Circle + or − as appropriate)

Do your regression results support the notion that there has been a positive time trend in the SALES data? Explain.

b. Use your equation to forecast SALES for the 12 months of 2017:

Period	SALES Forecast
2017M01	_____
M02	_____
M03	_____
M04	_____
M05	_____
M06	_____
M07	_____
M08	_____
M09	_____
M10	_____
M11	_____
M12	_____

c. Actual SALES for 2017 are:

Period	Actual SALES
2017M01	2,318
M02	2,367
M03	2,523
M04	2,577
M05	2,646
M06	2,674
M07	2,697
M08	2,702
M09	2,613
M10	2,626
M11	2,570
M12	2,590

On the basis of your results in part (*b*) in comparison with these actual sales, how well do you think your model works? What is the MAPE for 2017?

d. Prepare a time-series plot of the actual sales and the forecast of sales for 2013M01 through 2017M12. Do the same for just the last two years (2016M01 to 2017M12). Do your plots show any evidence of seasonality in the data? If so, how might you account for it in preparing a forecast?

17. Alexander Enterprises manufactures plastic parts for the automotive industry. Its sales (in thousands of dollars) for 2012Q1 through 2016Q4 are as follows:

(c4p17)

Period	Sales	Period	Sales
2012Q1	3,816.5	Q2	4,169.4
Q2	3,816.7	Q3	4,193.0
Q3	3,978.8	Q4	4,216.4
Q4	4,046.6	2014Q1	4,238.1
2013Q1	4,119.1	Q2	4,270.5

(continued on next page)

(continued)

(c4p17)

Period	Sales	Period	Sales
Q3	4,321.8	Q2	4,517.8
Q4	4,349.5	Q3	4,563.6
2015Q1	4,406.4	Q4	4,633.0
Q2	4,394.6	2017Q1	NA
Q3	4,422.3	Q2	NA
Q4	4,430.8	Q3	NA
2016Q1	4,463.9	Q4	NA

a. Begin by preparing a time-series plot of sales. Does it appear from this plot that a linear trend model might be appropriate? Explain.

b. Use a bivariate linear regression trend model to estimate the following trend equation:

$$\text{SALES} = a + b(\text{TIME})$$

Is the sign for b what you would expect? Is b significantly different from zero? What is the coefficient of determination for this model? Is there a potential problem with serial correlation? Explain.

c. Based on this model, make a trend forecast of sales for the four quarters of 2017.

d. Given that actual sales for the four quarters of 2017 are:

2017Q1	4,667.1
2017Q2	4,710.3
2017Q3	4,738.7
2017Q4	4,789.0

calculate the MAPE for this forecast model in the historical period (2012Q1–2016Q4) as well as for the forecast horizon (2017Q1–2017Q4). Which of these measures accuracy and which measures fit?

18. The following data are for shoe store sales in the United States in millions of dollars after being seasonally adjusted (SASSS). (c4p18)

Date	SASSS	Date	SASSS	Date	SASSS	Date	SASSS
Jan-02	1627	Mar-03	1524	May-04	1623	Jul-05	1692
Feb-02	1588	Apr-03	1560	Jun-04	1619	Aug-05	1695
Mar-02	1567	May-03	1575	Jul-04	1667	Sep-05	1721
Apr-02	1578	Jun-03	1588	Aug-04	1660	Oct-05	1698
May-02	1515	Jul-03	1567	Sep-04	1681	Nov-05	1770
Jun-02	1520	Aug-03	1602	Oct-04	1696	Dec-05	1703
Jul-02	1498	Sep-03	1624	Nov-04	1710	Jan-06	1745
Aug-02	1522	Oct-03	1597	Dec-04	1694	Feb-06	1728
Sep-02	1560	Nov-03	1614	Jan-05	1663	Mar-06	1776
Oct-02	1569	Dec-03	1644	Feb-05	1531	Apr-06	1807
Nov-02	1528	Jan-04	1637	Mar-05	1707	May-06	1800
Dec-02	1556	Feb-04	1617	Apr-05	1707	Jun-06	1758
Jan-03	1593	Mar-04	1679	May-05	1715	Jul-06	1784
Feb-03	1527	Apr-04	1607	Jun-05	1735	Aug-06	1791

(continued on next page)

(continued)

Date	SASSS	Date	SASSS	Date	SASSS	Date	SASSS
Sep-06	1743	Jul-09	1905	May-12	1940	Mar-15	2002
Oct-06	1785	Aug-09	1892	Jun-12	1963	Apr-15	2090
Nov-06	1765	Sep-09	1893	Jul-12	1920	May-15	2104
Dec-06	1753	Oct-09	1869	Aug-12	1937	Jun-15	2114
Jan-07	1753	Nov-09	1867	Sep-12	1867	Jul-15	2124
Feb-07	1790	Dec-09	1887	Oct-12	1918	Aug-15	2098
Mar-07	1830	Jan-10	1885	Nov-12	1914	Sep-15	2105
Apr-07	1702	Feb-10	1885	Dec-12	1931	Oct-15	2206
May-07	1769	Mar-10	1925	Jan-13	1867	Nov-15	2232
Jun-07	1793	Apr-10	1891	Feb-13	1887	Dec-15	2194
Jul-07	1801	May-10	1900	Mar-13	1939	Jan-16	2218
Aug-07	1789	Jun-10	1888	Apr-13	1860	Feb-16	2271
Sep-07	1791	Jul-10	1865	May-13	1898	Mar-16	2165
Oct-07	1799	Aug-10	1921	Jun-13	1924	Apr-16	2253
Nov-07	1811	Sep-10	1949	Jul-13	1967	May-16	2232
Dec-07	1849	Oct-10	1923	Aug-13	1994	Jun-16	2237
Jan-08	1824	Nov-10	1922	Sep-13	1966	Jul-16	2231
Feb-08	1882	Dec-10	1894	Oct-13	1943	Aug-16	2278
Mar-08	1859	Jan-11	1908	Nov-13	1973	Sep-16	2259
Apr-08	1831	Feb-11	1855	Dec-13	1976	Oct-16	2231
May-08	1832	Mar-11	1858	Jan-14	1969	Nov-16	2217
Jun-08	1842	Apr-11	1941	Feb-14	1989	Dec-16	2197
Jul-08	1874	May-11	1938	Mar-14	2040	Jan-17	
Aug-08	1845	Jun-11	1901	Apr-14	1976	Feb-17	
Sep-08	1811	Jul-11	1964	May-14	1964	Mar-17	
Oct-08	1898	Aug-11	1963	Jun-14	1947	Apr-17	
Nov-08	1878	Sep-11	1838	Jul-14	1961	May-17	
Dec-08	1901	Oct-11	1877	Aug-14	1931	Jun-17	
Jan-09	1916	Nov-11	1927	Sep-14	1960	Jul-17	
Feb-09	1894	Dec-11	1911	Oct-14	1980		
Mar-09	1883	Jan-12	1962	Nov-14	1944		
Apr-09	1871	Feb-12	1980	Dec-14	2014		
May-09	1918	Mar-12	1955	Jan-15	2013		
Jun-09	1943	Apr-12	1967	Feb-15	2143		

a. Make a linear trend forecast for SASSS though the first seven months of 2017. Given that the actual seasonally adjusted values for 2017 were as shown below, calculate the MAPE for those seven months of 2017.

Date	SASSS
Jan-17	2,422
Feb-17	2,112
Mar-17	2,290
Apr-17	2,354
May-17	2,013
Jun-17	2,156
Jul-17	2,425

b. Reseasonalize the 2017 forecast and the 2017 actual sales using the following seasonal indices:

Month	SI
Jan	0.74
Feb	0.81
Mar	1.00
Apr	1.03
May	1.04
Jun	0.98
Jul	0.98
Aug	1.23
Sep	0.96
Oct	0.94
Nov	0.98
Dec	1.31

c. Plot the final forecast along with the actual sales data. Does the forecast appear reasonable? Explain.

d. Why do you think the April, May, August, and December seasonal indices are greater than 1?

TABLE 4.4
Regression Trend Statistical Results from Excel.

Regression Statistics	
R Square	0.96
Standard Error	153.69
Observations	24

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1,34,93,902.63	1,34,93,902.63	571.28	0.00
Residual	22	5,19,653.32	23,620.61		
Total	23	1,40,13,555.95			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	11,042.86	64.76	170.53	0.00
Time Index	108.32	4.53	23.90	0.00

You can see in Figure 4.3 that the simple linear trend line does fit the actual data quite well and provides a reasonable forecast for the first four quarters of 2016. It is useful to provide a metric to evaluate the goodness of fit and accuracy of a forecast model. Goodness of fit refers to how well the model predicts values in sample (within the historic data set). Accuracy refers to how well the model works in the forecast period (here, the four quarters of 2016). We will use the MAPE as the metric.

In Table 4.3, you can see that the historic MAPE is 0.94 percent, indicating very low errors in the historic period (a measure of fit). This is consistent with the visual evaluation we see in Figure 4.3. The calculation of the MAPE for the holdout quarters of 2016 is shown in Table 4.5. The fit and accuracy MAPEs are the same, and less than 1 percent, suggesting that this linear regression trend forecast is quite good.

Trend models such as this can sometimes be very helpful in forecasting, and, as you see, they are easy to develop and to implement. In such models, we simply