9 Simple Linear Regression and Correlation

Often the marketing analyst needs to determine how variables are related, and much of the rest of this book is devoted to determining the nature of relationships between variables of interest. Some commonly important marketing questions that require analyzing the relationships between two variables of interest include:

- How does price affect demand?
- How does advertising affect sales?
- How does shelf space devoted to a product affect product sales?

This chapter introduces the simplest tools you can use to model relationships between variables. It first covers finding the line that best fits the hypothesized causal relationship between two variables. You then learn to use correlations to analyze the nature of non-causal relationships between two or more variables.

Simple Linear Regression

Every business analyst should have the ability to estimate the relationship between important business variables. In Microsoft Office Excel, the Trendline feature can help you determine the relationship between two variables. The variable you want to predict is the *dependent variable*. The variable used for prediction is the *independent variable*. Table 9-1 shows some examples of business relationships you might want to estimate.

Table 9-1: Examples of Relationships

Independent Variable	Dependent Variable
Units produced by a plant in 1 month	Monthly cost of operating a plant
Dollars spent on advertising in 1 month	Monthly sales
Number of employees	Annual travel expenses
Daily sales of cereal	Daily sales of bananas
Shelf space devoted to chocolate	Sales of chocolate
Price of bananas sold	Pounds of bananas sold

The first step to determine how two variables are related is to graph the data points so that the independent variable is on the x-axis and the dependent variable is on the y-axis. You can do this by using the Scatter Chart option in Microsoft Excel and performing the following steps:

1. With the Scatter Chart option selected, click a data point (displayed in blue) and click Trendline in the Analysis group on the Chart Tools Layout tab.
2. Next click More Trendline Options..., or right-click and select Add Trendline... You'll see the Format Trendline dialog box, which is shown in Figure 9-1.
3. If your graph indicates that a straight line can be drawn that provides a reasonable fit (a reasonable fit will be discussed in the "Defining R^2" section of this chapter) to the points, choose the Linear option. Nonlinear relationships are discussed in the "Modeling Nonlinearities and Interactions" section of Chapter 10, "Using Multiple Regression to Forecast Sales."

Analyzing Sales at Mao's Palace Restaurant

To illustrate how to model a linear relationship between two variables, take a look at the daily sales of products at Mao's Palace, a local Chinese restaurant (see Figure 9-2). Mao's main product is bowls filled with rice, vegetables, and meat made to the customer's order. The file `Maospalace.xlsx` gives daily unit sales of bowl price, bowls, soda, and beer.

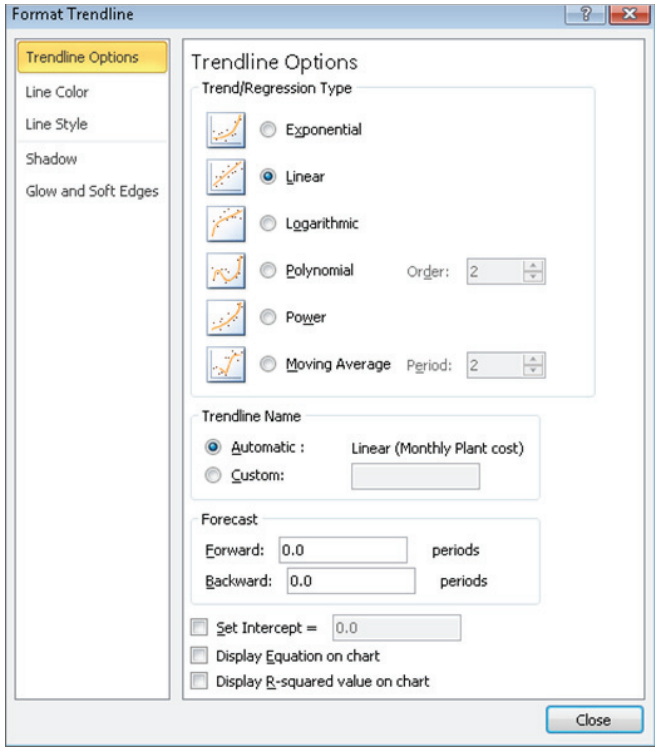

Figure 9-1: Trendline dialog box

Now suppose you want to determine how the price of the bowls affects daily sales. To do this you create an XY chart (or a scatter plot) that displays the independent variable (price) on the x-axis and the dependent variable (bowl sales) on the y-axis. The column of data that you want to display on the x-axis must be located to the left of the column of data you want to display on the y-axis. To create the graph, you perform two steps:

1. Select the data in the range E4:F190 (including the labels in cells E4 and F4).
2. Click Scatter in the Charts group on the Insert tab of the Ribbon, and select the first option (Scatter with only Markers) as the chart type. Figure 9-3 shows the graph.

	E	F	G	H
1				
2				
3				
4	Bowl Price	Bowls	Soda	Beer
5	$9.30	391	313	90
6	$9.10	418	326	100
7	$8.50	459	358	115
8	$9.50	424	331	81
9	$8.70	447	380	89
10	$9.70	383	291	92
11	$9.80	399	307	96
12	$8.80	440	361	66
13	$8.60	436	344	74
14	$9.60	413	351	62
15	$8.20	428	338	64
16	$8.00	479	374	101
17	$8.10	462	388	69
18	$9.80	387	325	77
19	$8.90	454	341	114
20	$9.40	418	314	88
21	$8.30	447	375	107
22	$9.60	442	376	102
23	$9.90	381	312	95
24	$9.30	401	301	68
25	$8.10	468	370	70
26	$8.70	428	321	64
27	$8.10	480	374	115

Figure 9-2: Sales at Mao's Palace

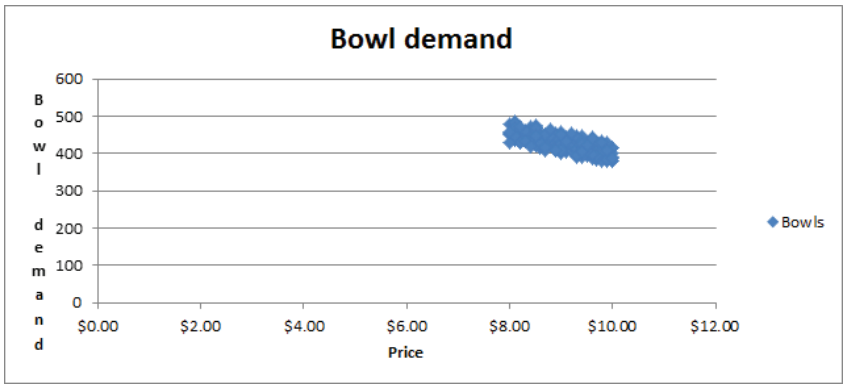

Figure 9-3: Scatterplot of Bowl demand versus Price

If you want to modify this chart, you can click anywhere inside the chart to display the Chart Tools contextual tab. Using the commands on the Chart Tools Design tab, you can do the following:

- Change the chart type.
- Change the source data.
- Change the style of the chart.
- Move the chart.

Using the commands on the Chart Tools Layout tab, you can do the following:

- Add a chart title.
- Add axis labels.
- Add labels to each point that gives the x and y coordinate of each point.
- Add gridlines to the chart.

Looking at the scatter plot, it seems reasonable that there is a straight line (or linear relationship) between the price and bowl sales. You can see the straight line that "best fits" the points by adding a trend line to the chart. To do so, perform the following steps:

1. Click within the chart to select it, and then click a data point. All the data points display in blue with an X covering each point.
2. Right-click and then click Add Trendline...
3. In the Format Trendline dialog box, select the Linear option, and then check the Display Equation on chart and the Display R-squared value on chart boxes, as shown in Figure 9-4. The R-Squared Value on the chart is defined in the "Defining R^2" section of this chapter.
4. Click Close to see the results shown in Figure 9-5. To add a title to the chart and labels for the x-and y-axes, select Chart Tools, click Chart Title, and then click Axis Titles in the Labels group on the Layout tab.
5. To add more decimal points to the equation, select the trend-line equation; after selecting Layout from Chart Tools, choose Format Selection.
6. Select Number and choose the number of decimal places to display.

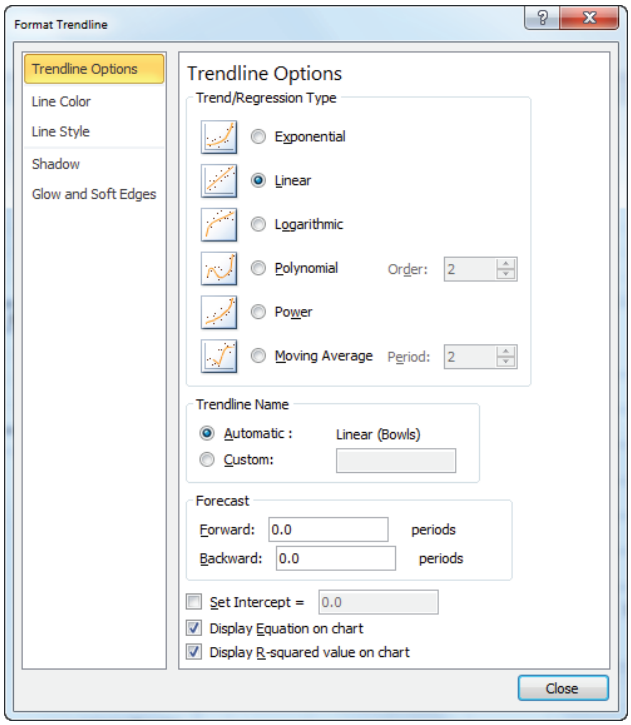

Figure 9-4: Trendline settings for Bowl demand

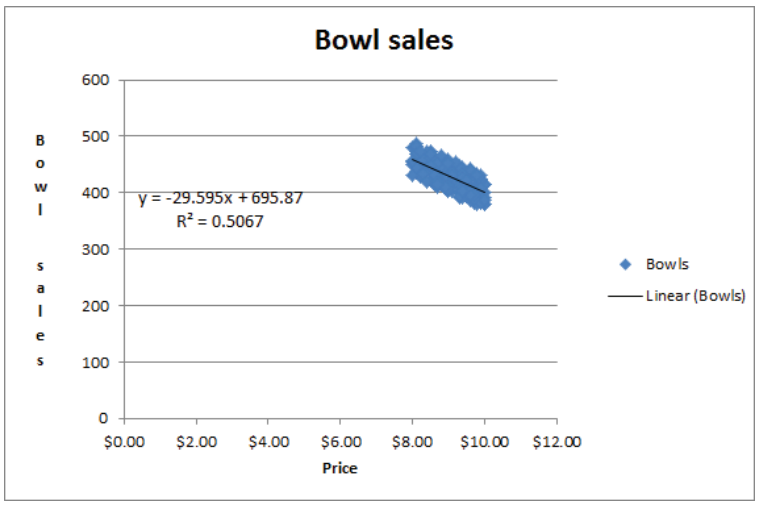

Figure 9-5: Trendline for Bowl demand

How Excel Determines the Best-Fitting Line

When you create a scatter chart and plot a trend line using the Trendline feature, it chooses the line that minimizes (over all lines that could be drawn) the sum of the squared vertical distance from each point to the line. The vertical distance from each point to the line is an *error*, or *residual*. The line created by Excel is called the *least-squares line*. You minimize the sum of squared errors rather than the sum of the errors because in simply summing the errors, positive and negative errors can cancel each other out. For example, a point 100 units above the line and a point 100 units below the line cancel each other if you add errors. If you square errors, however, the fact that your predictions for each point are wrong will be used by Excel to find the best-fitting line. Another way to see that minimizing the sum of squared errors is reasonable is to look at a situation in which all points lie on one line. Then minimizing the least squares line would yield this line and a sum of squared errors equal to 0.

Thus, Excel calculates that the best-fitting straight line for predicting daily bowl sales from the price by using the equation `Daily Bowl Sales=-29.595*Price + 695.87`. The -29.595 slope of this line indicates that the best guess is that a $1 increase in the price of a bowl reduces demand by 29.595 bowls.

WARNING You should not use a least-squares line to predict values of an independent variable that lies outside the range for which you have data. Your line should be used only to predict daily bowl sales for days in which the bowl price is between $8 and $10.

Computing Errors or Residuals

Referring back to the Mao's Palace example, you can compute predicted bowl sales for each day by copying the formula `=-29.595*E5+695.87` from C5 to C6:C190. Then copy the formula `=F5-C5` from D5 to D6:D190. This computes the errors (or residuals). These errors are shown in Figure 9-6. For each data point, you can define the error by the amount by which the point varies from the least-squares line. For each day, the error equals the observed demand minus the predicted demand. A positive error indicates a point is above the least-squares line, and a negative error indicates that the point is below the least-squares line. In cell D2, the sum of the errors is computed, which obtained 1.54. In reality, for any least-squares line, the sum of the errors should equal 0. 1.54 is obtained because the equation is rounded

to three decimal points.) The fact that errors sum to 0 implies that the least-squares line has the intuitively satisfying property of splitting the points in half.

B	C	D	E	F	G	H
		sum errors				
		1.537				
	Predicted Bowl Sales	Error	Bowl Price	Bowls	Soda	Beer
	420.6365	-29.6365	$9.30	391	313	90
	426.5555	-8.5555	$9.10	418	326	100
	444.3125	14.6875	$8.50	459	358	115
	414.7175	9.2825	$9.50	424	331	81
	438.3935	8.6065	$8.70	447	380	89
	408.7985	-25.7985	$9.70	383	291	92
	405.839	-6.839	$9.80	399	307	96
	435.434	4.566	$8.80	440	361	66
	441.353	-5.353	$8.60	436	344	74
	411.758	1.242	$9.60	413	351	62
	453.191	-25.191	$8.20	428	338	64
	459.11	19.89	$8.00	479	374	101

Figure 9-6: Errors in predicting Bowl demand

Defining R^2

As you can see in the Mao's Palace example, each day both the bowl price and bowl sales vary. Therefore it is reasonable to ask what percentage of the monthly variation in sales is explained by the daily variation in price. In general the percentage of the variation in the dependent variable explained by the least squares line is known as R^2. For this regression the R^2 value is 0.51, which is shown in Figure 9-5. You can state that the linear relationship explains 51 percent of the variation in monthly operating costs.

Once you determine the R^2 value, your next question might be what causes the other 49 percent of the variation in daily bowl sales costs. This value is explained by various other factors. For example, the day of the week and month of the year might affect bowl sales. Chapter 10, "Using Multiple Regression to Forecast Sales" explains how to use *multiple regression* to determine other factors that influence operating costs. In most cases, finding factors that increase R^2 increases prediction accuracy. If a factor only results in a slight increase in R^2, however, using that factor to predict the dependent variable can actually decrease forecast accuracy. (See Chapter 10 for further discussion of this idea.)

Another question that comes up a lot in reference to R^2 values is what is a good R^2 value? There is no definitive answer to this question. As shown in Exercise 5 toward the end of the chapter, a high R^2 can occur even when

a trend line is not a good predictor of *y*. With one independent variable, of course, a larger R^2 value indicates a better fit of the data than a smaller R^2 value. A better measure of the accuracy of your predictions is the *standard error of the regression*, described in the next section.

Accuracy of Predictions from a Trend Line

When you fit a line to points, you obtain a standard error of the regression that measures the *spread* of the points around the least-squares line. You can compute the standard error associated with a least-squares line with the `STEYX` function. The syntax of this function is `STEYX(known_y's, known_x's)`, where *yrange* contains the values of the dependent variable, and *xrange* contains the values of the independent variable. To use this function, select the range E4:F190 and use `FORMULAS CREATE FROM SELECTION` to name your price data `Bowl_Price` and your sales data `Bowls`. Then in cell K1, compute the standard error of your cost estimate line with the formula `=STEYX(Bowls,Bowl_Price)`. Figure 9-7 shows the result.

	J	K	L
1			
2	**STD ERROR**	**17.41867**	
3	**SLOPE**	**-29.5945**	
4	**INTERCEPT**	**695.8741**	
5	**RSQ**	**0.506748**	

Figure 9-7: Computing standard error of the regression

Approximately 68 percent of your points should be within one standard error of regression (SER) of the least-squares line, and approximately 95 percent of your points should be within two SER of the least-squares line. These measures are reminiscent of the descriptive statistics rule of thumb described in Chapter 2, "Using Excel Charts to Summarize Marketing Data." In your example, the absolute value of approximately 68 percent of the errors should be 17.42 or smaller, and the absolute value of approximately 95 percent of the errors should be 34.84, or 2 * *17.42*, or smaller. You can find that 57 percent of your points are within one SER of the least-squares line, and all (100 percent) of the points are within two standard SER of the least-squares line. Any point that is more than two SER from the least-squares line is called an *outlier.*

Looking for causes of outliers can often help you to improve the operation of your business. For example, a day in which actual demand was 34.84 higher than anticipated would be a demand outlier on the high side. If you ascertain the cause of this high sales outlier and make it recur, you would clearly improve profitability.

Similarly, consider a month in which actual sales are over 34.84 less than expected. If you can ascertain the cause of this low demand outlier and ensure it occurred less often, you would improve profitability. Chapters 10 and 11 explain how to use outliers to improve forecasting.

The Excel Slope, Intercept, and RSQ Functions

You have learned how to use the Trendline feature to find the line that best fits a linear relationship and to compute the associated R^2 value. Sometimes it is more convenient to use Excel functions to compute these quantities. In this section, you learn how to use the Excel `SLOPE` and `INTERCEPT` functions to find the line that best fits a set of data. You also see how to use the RSQ function to determine the associated R^2 value.

The Excel `SLOPE(known_y's, known_x's)` and `INTERCEPT(known_y's, known_x's)` functions return the slope and intercept, respectively, of the least-squares line. Thus, if you enter the formula `SLOPE(Bowls, Bowl_Price)` in cell K3 (see Figure 9-7) it returns the slope (–29.59) of the least-squares line. Entering the formula `INTERCEPT(Bowls, Bowl_Price)` in cell K4 returns the intercept (695.87) of the least-squares line. By the way, the `RSQ(known_y's, known_x's)` function returns the R^2 value associated with a least-squares line. So, entering the formula `RSQ(Bowls, Bowl_Price)` in cell K5 returns the R^2 value of 0.507 for your least-squares line. Of course this R^2 value is identical to the RSQ value obtained from the Trendline.

Using Correlations to Summarize Linear Relationships

Trendlines are a great way to understand how two variables are related. Often, however, you need to understand how more than two variables are related. Looking at the correlation between any pair of variables can provide insights into how multiple variables move up and down in value together. Correlation measures linear association, not causation.

The correlation (usually denoted by *r*) between two variables (call them *x* and *y*) is a unit-free measure of the strength of the linear relationship between *x* and *y*. The correlation between any two variables is always between –1 and +1. Although the exact formula used to compute the correlation between two variables isn't very important, interpreting the correlation between the variables is.

A correlation near +1 means that *x* and *y* have a strong positive linear relationship. That is, when *x* is larger than average, *y* is almost always larger than average, and when

x is smaller than average, y is almost always smaller than average. For example, for the data shown in Figure 9-8, (x = units produced and y = monthly production cost), x and y have a correlation of +0.95. You can see that in Figure 9-8 the least squares line fits the points very well and has a positive slope which is consistent with large values of x usually occurring with large values of y.

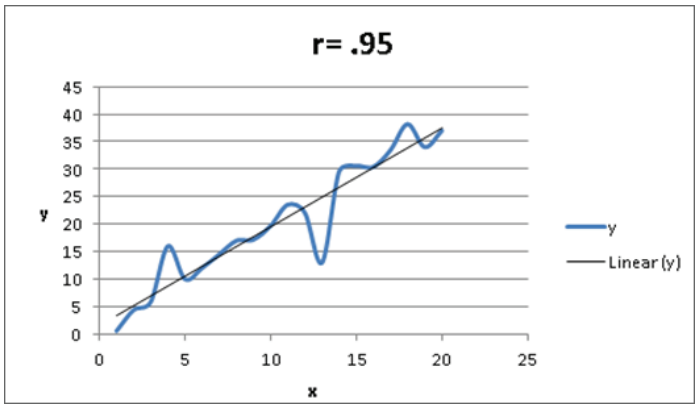

Figure 9-8: Correlation = +0.95

If x and y have a correlation near –1, this means that there is a strong negative linear association between x and y. That is, when x is larger than average, y is usually be smaller than average, and when x is smaller than average, y is usually larger than average. For example, for the data shown in Figure 9-9, x and y have a correlation of –0.90. You can see that in Figure 9-9 the least squares line fits the points very well and has a negative slope which is consistent with large values of x usually occurring with small values of y.

A correlation near 0 means that x and y have a weak linear association. That is, knowing whether x is larger or smaller than its mean tells you little about whether y will be larger or smaller than its mean. Figure 9-10 shows a graph of the dependence of unit sales (y) on years of sales experience (x). Years of experience and unit sales have a correlation of 0.003. In the data set, the average experience is 10 years. You can see that when a person has more than 10 years of sales experience, sales can be either low or high. You also see that when a person has fewer than 10 years of sales experience, sales can be low or high. Although experience and sales have little or no linear relationship, there is a strong nonlinear relationship (see the fitted curve in Figure 9-10) between years of experience and sales. Correlation does not measure the strength of nonlinear associations.

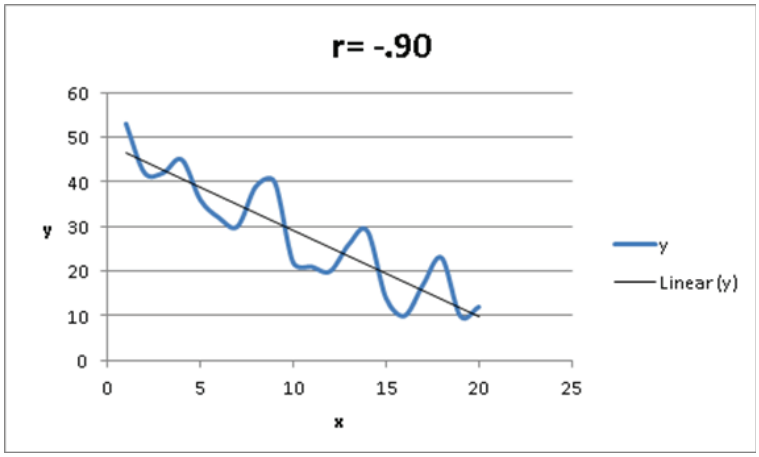

Figure 9-9: Correlation = -0.90

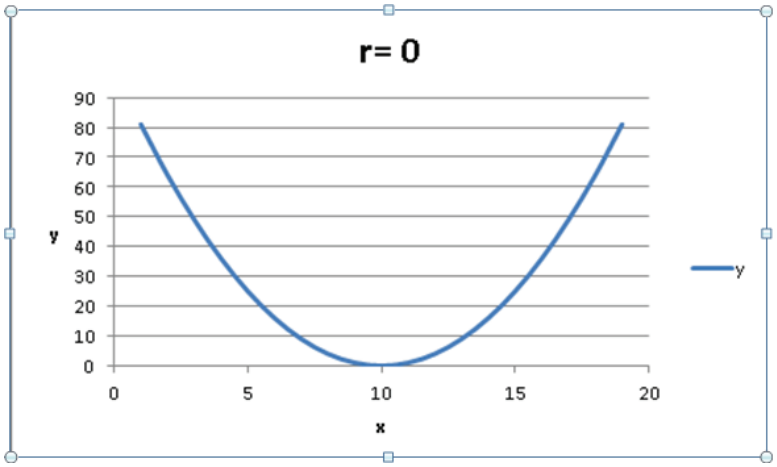

Figure 9-10: Correlation near 0

Finding a Correlation with the Data Analysis Add-In

You will now learn how Excel's Data Analysis Add-in and the Excel Correlation function can be used to compute correlations. The Data Analysis Add-In makes it easy to find correlations between many variables. To install the Data Analysis Add-in, perform the following steps:

1. Click the File tab and select Options.
2. In the Manage box click Excel Add-Ins, and choose Go.
3. In the Add-Ins dialog box, select Analysis ToolPak and then click OK.

Now you can access the Analysis ToolPak functions by clicking Data Analysis in the Analysis group on the Data tab.

You can use this functionality to find the correlations between each pair of variables in the Mao's Palace data set. To begin select the Data Analysis Add-In, and choose Correlation. Then fill in the dialog box, as shown in Figure 9-11.

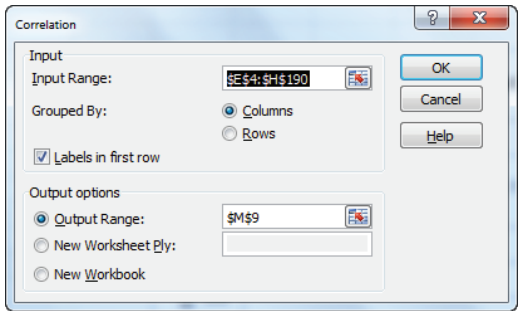

Figure 9-11: Correlation dialog box

To compute correlations with the Data Analysis Add-in proceed as follows:

1. Select the range which contains the relevant data and data labels. The easiest way to accomplish this is to select the upper-left cell of the data range (E5) and then press Ctrl+Shift+Right Arrow, followed by Ctrl+Shift+Down Arrow.
2. Check the Labels In First Row option because the first row of the input range contains labels. Enter cell M9 as the upper-left cell of the output range.
3. After clicking OK, you see the results, as shown in Figure 9-12.

	M	N	O	P	Q
3					
4					
5		**Price Bowl Correlation**			
6		**-0.71186**			
7					
8					
9		***Bowl Price***	***Bowls***	***Soda***	***Beer***
10	**Bowl Price**	**1**			
11	**Bowls**	**-0.71186**	**1**		
12	**Soda**	**-0.58095**	**0.831008**	**1**	
13	**Beer**	**-0.19367**	**0.338691**	**0.246803**	**1**

Figure 9-12: Correlation matrix

From Figure 9-12, you find there is a –0.71 correlation between Bowl Price and Bowl Sales, indicating a strong negative linear association The .0.83 correlation

between Soda Sales and Bowl Sales indicates a strong positive linear association. The +0.25 correlation between beer and soda sales indicates a slight positive linear association between beer and soda sales.

Using the CORREL Function

As an alternative to using the Correlation option of the Analysis Toolpak, you can use the CORREL function. For example, enter the formula `=CORREL(Bowl_Price,F5:F190)` in cell N6 and you can confirm that the correlation between price and bowl sales is -0.71.

Relationship Between Correlation and R^2

The correlation between two sets of data is simply $-\sqrt{R^2}$ for the trend line, where you choose the sign for the square root to be the same as the sign of the slope of the trend line. Thus the correlation between bowl price and bowl sales is $-\sqrt{.507} = -0.711$.

Correlation and Regression Toward the Mean

You have probably heard the phrase "regression toward the mean." Essentially, this means that the predicted value of a dependent variable will be in some sense closer to its average value than the independent variable. More precisely, suppose you try to predict a dependent variable y from an independent variable x. If x is k standard deviations above average, then your prediction for y will be $r \times k$ standard deviations above average. (Here, r = correlation between x and y.) Because r is between –1 and +1, this means that y is fewer standard deviations away from the mean than x. This is the real definition of "regression toward the mean." See Exercise 9 for an interesting application of the concept of regression toward the mean.

Summary

Here is a summary of what you have learned in this chapter:

- The Excel Trendline can be used to find the line that best fits data.
- The R^2 value is the fraction of variation in the dependent variable explained by variation in the independent variable.
- Approximately 95 percent of the forecasts from a least-squares line are accurate within two standard errors of the regression.
- Given two variables x and y, the correlation r (always between –1 and +1) between x and y is a measure of the strength of the linear association between x and y.

- Correlation may be computed with the Analysis ToolPak or the CORREL function.
- If x is k standard deviations above the mean, you can predict y to be rk standard deviations above the mean.

Exercises

1. The file `Delldata.xlsx` (available on the companion website) contains monthly returns for the Standard & Poor's stock index and for Dell stock. The *beta* of a stock is defined as the slope of the least-squares line used to predict the monthly return for a stock from the monthly return for the market. Use this file to perform the following exercises:
    a. Estimate the beta of Dell.
    b. Interpret the meaning of Dell's beta.
    c. If you believe a recession is coming, would you rather invest in a high-beta or low-beta stock?
    d. During a month in which the market goes up 5 percent, you are 95 percent sure that Dell's stock price will increase between which range of values?
2. The file `Housedata.xlsx` (available on the companion website) gives the square footage and sales prices for several houses in Bellevue, Washington. Use this file to answer the following questions:
    a. You plan to build a 500-square-foot addition to your house. How much do you think your home value will increase as a result?
    b. What percentage of the variation in home value is explained by the variation in the house size?
    c. A 3,000-square-foot house is listed for $500,000. Is this price out of line with typical real estate values in Bellevue? What might cause this discrepancy?
3. You know that 32 degrees Fahrenheit is equivalent to 0 degrees Celsius, and that 212 degrees Fahrenheit is equivalent to 100 degrees Celsius. Use the trend curve to determine the relationship between Fahrenheit and Celsius temperatures. When you create your initial chart, before clicking Finish, you must indicate (using Switch Rows and Columns from the Design Tab on Chart Tools) that data is in columns and not rows because with only two data points, Excel assumes different variables are in different rows.

4. The file `Electiondata.xlsx` (available on the companion website) contains, for several elections, the percentage of votes Republicans gained from voting machines (counted on election day) and the percentage Republicans gained from absentee ballots (counted after election day). Suppose that during an election, Republicans obtained 49 percent of the votes on election day and 62 percent of the absentee ballot votes. The Democratic candidate cried "Fraud." What do you think?
5. The file `GNP.xls` (available on the companion website) contains quarterly GNP data for the United States in the years 1970–2012. Try to predict next quarter's GNP from last quarter's GNP. What is the R^2? Does this mean you are good at predicting next quarter's GNP?
6. Find the trend line to predict soda sales from daily bowl sales.
7. The file `Parking.xlsx` contains the number of cars parked each day both in the outdoor lot and in the parking garage near the Indiana University Kelley School of Business. Find and interpret the correlation between the number of cars parked in the outdoor lot and in the parking garage.
8. The file `Printers.xlsx` contains daily sales volume (in dollars) of laser printers, printer cartridges, and school supplies. Find and interpret the correlations between these quantities.
9. NFL teams play 16 games during the regular season. Suppose the standard deviation of the number of games won by all teams is 2, and the correlation between the number of games a team wins in two consecutive seasons is 0.5. If a team goes 12 and 4 during a season, what is your best prediction for how many games they will win next season?

10 Using Multiple Regression to Forecast Sales

A common need in marketing analytics is forecasting the sales of a product. This chapter continues the discussion of *causal forecasting* as it pertains to this need. In causal forecasting, you try and predict a dependent variable (usually called Y) from one or more independent variables (usually referred to as X_1, X_2, …, X_n). In this chapter the dependent variable Y usually equals the sales of a product during a given time period.

Due to its simplicity, univariate regression (as discussed in Chapter 9, "Simple Linear Regression and Correlation") may not explain all or even most of the variance in Y. Therefore, to gain better and more accurate insights about the often complex relationships between a variable of interest and its predictors, as well as to better forecast, one needs to move towards multiple regression in which more than one independent variable is used to forecast Y. Utilizing multiple regression may lead to improved forecasting accuracy along with a better understanding of the variables that actually cause Y.

For example, a multiple regression model can tell you how a price cut increases sales or how a reduction in advertising decreases sales. This chapter uses multiple regression in the following situations:

- Setting sales quotas for computer sales in Europe
- Predicting quarterly U.S. auto sales
- Understanding how predicting sales from price and advertising requires knowledge of nonlinearities and interaction
- Understanding how to test whether the assumptions needed for multiple regression are satisfied
- How multicollinearity and/or autocorrelation can disturb a regression model

Introducing Multiple Linear Regression

In a multiple linear regression model, you can try to predict a dependent variable Y from independent variables $X_1, X_2, \ldots X_n$. The assumed model is as follows:

$$(1)\ Y = B_0 + B_1X_1 + B_2X_2 + \ldots B_nX_n + \text{error term}$$

In Equation 1:

- B_0 is called the *intercept* or *constant term*.
- B_i is called the *regression coefficient* for the independent variable X_i.

The *error term* is a random variable that captures the fact that regression models typically do not fit the data perfectly; rather they approximate the relationships in the data. A positive value of the error term occurs if the actual value of the dependent variable exceeds your predicted value $(B_0 + B_1X_1 + B_2X_2 + \ldots B_nX_n)$. A negative value of the error term occurs when the actual value of the dependent variable is less than the predicted value.

The error term is required to satisfy the following assumptions:

- The error term is normally distributed.
- The variability or spread of the error term is assumed not to depend on the value of the dependent variable.
- For time series data successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.

In the "Testing Validity of Multiple Regression Assumptions," section of this chapter you will learn how to determine if the assumptions of regression analysis are satisfied, and what to do if the assumptions are not satisfied.

To best illustrate how to use multiple regression, the remainder of the chapter presents examples of its use based on a fictional computer sales company, HAL Computer. HAL sets sales quotas for all salespeople based on their territory. To set fair quotas, HAL needs a way to accurately forecast computer sales in each person's territory. From the 2011 *Pocket World in Figures* by *The Economist*, you can obtain the following data from 2007 (as shown in Figure 10-1 and file `Europe.xlsx`) for European countries:

- Population (in millions)
- Computer sales (in millions of U.S. dollars)

- Sales per capita (in U.S. dollars)
- GNP per head
- Average Unemployment Rate 2002–2007
- Percentage of GNP spent on education

	F	G	H	I	J	K	L
3				Source Economist Pocket World in Figures 2011			
4	Country	Pop (millions)	Computer Sales	Sales/Capita	GNP per head	Unemployment rate	%age spend on education
5	Austria	8.4	941.2	$112.05	$49,600	4.2	5.8
6	Belgium	10.5	1681.9	$160.18	$47,090	8.1	5.9
7	Bulgaria	7.6	154	$20.26	$6,550	13.5	3.5
8	Czech Rep.	10.2	1028.7	$100.85	$20,670	6.6	4.4
9	Denmark	5.5	935.4	$170.07	$62,120	5.2	8.4
10	Finland	5.3	1971	$371.89	$51,320	9.9	6.3
11	France	61.9	5928.9	$95.78	$44,510	10	5.7
12	Germany	82.5	6824.3	$82.72	$44,450	9.1	4.6
13	Greece	11.2	813	$72.59	$31,670	9.9	3.9
14	Hungary	10	449	$44.90	$15,410	7.3	5.1
15	Ireland	4.4	576.9	$131.11	$60,460	6.3	4.3
16	Italy	58.9	3858.2	$65.50	$38,490	9.3	5
17	Netherlands	16.5	2168.5	$131.42	$52,960	4.4	5
18	Poland	38	2847	$74.92	$13,850	14.4	5.6
19	Portugal	10.7	728.6	$68.09	$22,920	6.3	5.9
20	Romania	21.3	687.2	$32.26	$9,300	7	3.3
21	Spain	44.8	4745.8	$105.93	$35,220	14.2	4.4
22	Switzerland	7.5	1130.4	$150.72	$64,430	3.6	5.6
23	Sweden	9.2	2113.4	$229.72	$51,950	6.3	7.6
24	Turkey	75.8	2879	$37.98	$9,940	8.6	3.7

Figure 10-1: HAL computer data

This data is *cross-sectional data* because the same dependent variable is measured in different locations at the same point in time. In *time series data*, the same dependent variable is measured at different times.

In order to apply the multiple linear regression model to the example, Y = Per Capital Computer spending, $n = 3$, X_1 = Per Capita GNP, X_2 = Unemployment Rate, and X_3 = Percentage of GNP spent on education.

Running a Regression with the Data Analysis Add-In

You can use the Excel Data Analysis Add-In to determine the best-fitting multiple linear regression equation to a given set of data. See Chapter 9 for a refresher on installation instructions for the Data Analysis Add-In.

To run a regression, select Data Analysis in the Analysis Group on the Data tab, and then select Regression. When the Regression dialog box appears, fill it in, as shown in Figure 10-2.

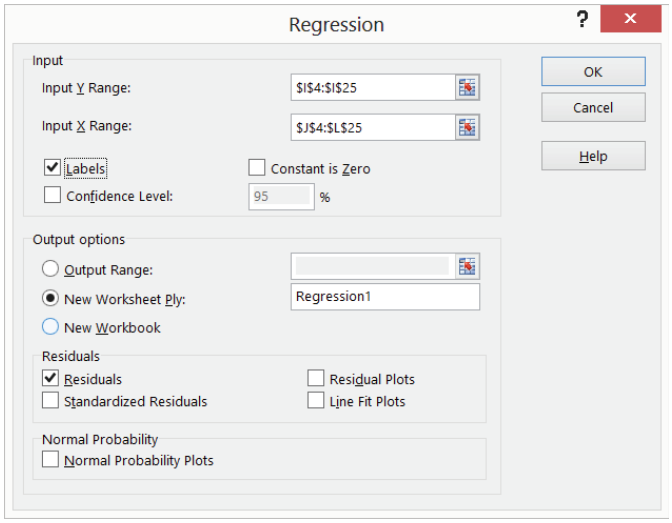

Figure 10-2: Regression dialog box

- The *Y* Range (I4:I25) includes the data you want to predict (computer per capita sales), including the column label.
- The *X* Range (J4:L25) includes those values of the independent variables for each country, including the column label.
- Check the Labels box because your X range and Y range include labels. If you do not include labels in the X and Y range, then Excel will use generic labels like Y, X_1, X_2,…,X_n which are hard to interpret.
- The worksheet name `Regression1` is the location where the output is placed.
- By checking the Residuals box, you can ensure Excel will generate the error (for each observation error = actual value of Y – predicted value for Y).

After selecting OK, Excel generates the output shown in Figures 10-3 and 10-4. For Figure 10-4, the highlighted text indicates data that is thrown out later in the chapter.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.731106465							
5	R Square	0.534516664							
6	Adjusted R Square	0.452372545							
7	Standard Error	58.42625704							
8	Observations	21							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	66638.03186	22212.68	6.507059	0.003940222			
13	Residual	17	58031.66769	3413.628					
14	Total	20	124669.6996						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-114.8351503	78.28996449	-1.46679	0.160688	-280.012537	50.34224	-280.0125369	50.3422364
18	GNP per head	0.002297712	0.00095193	2.413741	0.027355	0.000289316	0.004306	0.000289316	0.004306108
19	Unemployment rate	4.219524573	4.840005896	0.871802	0.395463	-5.99199526	14.43104	-5.991995264	14.43104441
20	%age spend on education	21.4226983	12.73611957	1.682043	0.110837	-5.44816518	48.29356	-5.448165177	48.29356178

Figure 10-3: First multiple regression output

	A	B	C	D
24	RESIDUAL OUTPUT			
25				
26	*Observation*	*Predicted Sales/Capita*	*Residuals*	
27	1	141.105011	-29.0574	
28	2	153.9361699	6.244782	
29	3	32.15788817	-11.8947	
30	4	54.76728846	46.08565	
31	5	229.7909036	-59.7182	
32	6	179.8197146	192.0671	Finland
33	7	151.7406304	-55.9587	
34	8	124.2402274	-41.5214	
35	9	83.25520078	-10.6659	
36	10	60.63088012	-15.7309	
37	11	142.7851159	-11.6715	
38	12	119.958849	-54.4546	
39	13	132.5310691	-1.10683	
40	14	97.71642325	-22.7954	
41	15	90.80533021	-22.7119	
42	16	6.765146383	25.49776	
43	17	120.2673827	-14.3343	
44	18	168.3638234	-17.6438	
45	19	193.9264924	35.7909	
46	20	23.55600061	14.42553	
47	21	112.9313707	49.15388	

Figure 10-4: Residuals from first regression

Interpreting the Regression Output

After you run a regression, you next must interpret the output. To do this you must analyze a variety of elements listed in the output. Each element of the output affects the output in a unique manner. The following sections explain how to interpret the important elements of the regression output.

Coefficients

The Coefficients column of the output (cells B17:B20) gives the best fitting estimate of the multiple regression equation. Excel returns the following equation:

(2) Predicted Computer Sales / Capita = –114.84 + .002298 * (Per Capita GNP) + 4.22 * (Unemployment Rate) + 21.42(Percentage Spent on Education)

Excel found this equation by considering all values of B_0, B_1, B_2, and B_3 and choosing the values that minimize the sum over all observations of (Actual Dependent Variable – Predicted Value)^2. The coefficients are called the *least squares estimates* of B_0, B_1,…,B_n. You square the errors so positive and negative values do not cancel. Note that if the equation perfectly fits each observation, then the sum of squared errors is equal to 0.

F Test for Hypothesis of No Linear Regression

Just because you throw an independent variable into a regression does not mean it is a helpful predictor. If you used the number of games each country's national soccer team won during 2007 as an independent variable, it would probably be irrelevant and have no effect on computer sales. The ANOVA section of the regression output (shown in Figure 10-3) in cells A10:F14 enables you to test the following hypotheses:

- Null Hypothesis: The *Hypothesis of No Linear Regression*: Together all the independent variables are not useful (or significant) in predicting *Y*.
- Alternative Hypothesis: Together all the independent variables are useful (or significant).

To decide between these hypotheses, you must examine the *Significance F Value* in cell F12. The Significance F value of .004 tells you that the data indicates that there are only 4 chances in 1000 that your independent variables are not useful in predicting *Y*, so you would reject the null hypothesis. Most statisticians agree that a Significance F (often called *p-value*) of .05 or less should cause rejection of the Null Hypothesis.

Accuracy and Goodness of Fit of Regression Forecasts

After you conclude that the independent variables together are significant, a natural question is, how well does your regression equation fit the data? The R^2 value in B5 and *Standard Error* in B7 (see Figure 10-3) answer this question.

- The R^2 value of .53 indicates that 53 percent of the variation in *Y* is explained by Equation 1. Therefore, 47 percent of the variation in *Y* is unexplained by the multiple linear regression model.
- The Standard Error of 58.43 indicates that approximately 68 percent of the predictions for *Y* made from Equation 2 are accurate within one standard error ($58.43) and 95 percent of your predictions for *Y* made from Equation 2 are accurate within two standard errors ($116.86.)

Determining the Significant Independent Variables

Because you concluded that together your independent variables are useful in predicting *Y*, you now must determine which independent variables are useful. To do this look at the *p-values* in E17:E20. A p-value of .05 or less for an independent variable indicates that the independent variable is (after including the effects of all other independent variables in the equation) a significant predictor for *Y*. It appears that only GNP per head (p-value .027) is a significant predictor. At this point you want to see if there are any *outliers* or unusual data points. Outliers in regression are data points where the absolute value of the error (actual value of *y* – predicted value of *y*) exceeds two standard errors. Outliers can have a drastic effect on regression coefficients, and the analyst must decide whether to rerun the regression without the outliers.

The Residual Output and Outliers

For each data point or observation, the Residual portion of the regression output, as shown in Figure 10-4, gives you two pieces of information.

- The Predicted Value of *Y* from Equation 2. For example, Austria predicted per capita expenditures are given by the following:

  ($116.86) + (0.00229) * (49,600) + (4.22) * (4.2) + 21.52 (5.8)
  = $141.10

- The Residuals section of the output gives for each observation the error = Actual value of Y – Predicted Value of Y. For Austria you find the residual is $112.05 – $141.10 = $–29.05. The regression equation found by least squares

has the intuitively pleasing property that the sum of the residuals equals 0. This implies that overestimates and underestimates of *Y* cancel each other out.

Dealing with Insignificant Independent Variables

In the last section you learned that GNP per head was the only significant independent variable and the other two independent variables were insignificant. When an independent variable is insignificant (has a p-value greater than .05) you can usually drop it and run the regression again. Before doing this though, you must decide what to do with your outlier(s). Because the standard error or the regression is 58.4, any error exceeding 116.8 in absolute value is an outlier. Refer to Figure 10-4 and you can see that Finland (which is highlighted) is a huge outlier. Finland's spending on computers is more than three standard errors greater than expected. When you delete Finland as an outlier, and then rerun the analysis, the result is in the worksheet `Regression2` of file `Europe.xlsx`, as shown in Figure 10-5.

Checking the residuals you find that Switzerland is an outlier. (You under predict expenditures by slightly more than two standard errors.) Because Switzerland is not an outrageous outlier, you can choose to leave it in the data set in this instance. Unemployment Rate is insignificant (p-value of .84 > .05) so you can delete it from the model and run the regression again. The resulting regression is in worksheet `Regression 3`, of file `Europe.xlsx` as shown in Figure 10-6.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.860805637					
5	R Square	0.740986344					
6	Adjusted R Square	0.692421283					
7	Standard Error	29.9835813					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	41150.44477	13716.81	15.2576	5.93265E-05	
13	Residual	16	14384.24236	899.0151			
14	Total	19	55534.68713				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-32.20876114	41.89082135	-0.76887	0.453169	-121.0133353	56.595813
18	GNP per head	0.001678416	0.000496537	3.380244	0.003816	0.000625805	0.002731
19	Unemployment rate	-0.527867146	2.575579641	-0.20495	0.840195	-5.987852075	4.9321178
20	%age spend on education	15.22764461	6.596202921	2.308547	0.034658	1.244319079	29.21097

Figure 10-5: Regression results: Finland outlier removed

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.860410573					
5	R Square	0.740306355					
6	Adjusted R Square	0.709754161					
7	Standard Error	29.12650434					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	41112.68179	20556.34	24.23087	1.0542E-05	
13	Residual	17	14422.00534	848.3533			
14	Total	19	55534.68713				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-38.48026121	27.79129164	-1.38462	0.184076	-97.1147612	20.154239
18	GNP per head	0.001723168	0.000433202	3.977751	0.000973	0.000809193	0.0026371
19	%age spend on education	15.30973984	6.395825812	2.393708	0.028487	1.815726905	28.803753

Figure 10-6: Regression output: unemployment rate removed

Both independent variables are significant, so use the following equation to predict Per Capita Computer Spending:

(3) -38.48 + 0.001723 * (GNP Per Capita) + 15.30974 * (Percentage GNP Spent on Education)

Because $R^2 = 0.74$, the equation explains 74 percent of the variation in Computer Spending. Because the Standard error is 29.13, you can expect 95 percent of your forecasts to be accurate within $58.26. From the Residuals portion of the output, you can see that Switzerland (error of $62.32) is the only outlier.

Interpreting Regression Coefficients

The regression coefficient of a variable estimates the effect (after adjusting for all other independent variables used to estimate the regression equation) of a unit increase in the independent variable. Therefore Equation 3 may be interpreted as follows:

- After adjusting for a fraction of GNP spent on education, a $1,000 increase in Per Capita GNP yields a $1.72 increase in Per Capital Computer spending.
- After adjusting for Per Capita GNP, a 1 percent increase in the fraction of GNP spent on education yields a $15.31 increase in Per Capita Computer spending.

Setting Sales Quotas

Often part of a salesperson's compensation is a commission based on whether a salesperson's sales quota is met. For commission payments to be fair, the company needs to ensure that a salesperson with a "good" territory has a higher quota than a salesperson with a "bad" territory. You'll now see how to use the multiple regression model to set fair sales quotas. Using the multiple regression, a reasonable annual sales quota for a territory equals the population * company market share * regression prediction for per capita spending.

Assume that a province in France has a per capita GNP of $50,000 and spends 10 percent of its GNP on education. If your company has a 30 percent market share, then a reasonable per capita annual quota for your sales force would be the following:

$$0.30(-38.48 + 0.001723 * (50{,}000) + 15.30974 * (10)) = \$60.23$$

Therefore, a reasonable sales quota would be $60.23 per capita.

Beware of Blind Extrapolation

While you can use regressions to portray a lot of valuable information, you must be wary of using them to predict values of the independent variables that differ greatly from the values of the independent variables that fit the regression equation. For example, the Ivory Coast has a Per Capita GNP of $1,140, which is far less than any country in your European data set, so you could not expect Equation 3 to give a reasonable prediction for Per Capita Computer spending in the Ivory Coast.

Using Qualitative Independent Variables in Regression

In the previous example of multiple regression, you forecasted Per Capita Computer sales using Per Capita GNP and Fraction of GNP spent on education. Independent variables can also be quantified with an exact numerical value and are referred to as *quantitative independent variables*. In many situations, however, independent variables can't be easily quantified. This section looks at ways to incorporate a qualitative factor, such as seasonality, into a multiple regression analysis.

Suppose you want to predict quarterly U.S. auto sales to determine whether the quarter of the year impacts auto sales. Use the data in the file `Autos.xlsx`, as shown in Figure 10-7. Sales are listed in thousands of cars, and GNP is in billions of dollars.

You might be tempted to define an independent variable that equals 1 during the first quarter, 2 during the second quarter, and so on. Unfortunately, this approach

would force the fourth quarter to have four times the effect of the first quarter, which might not be true. The quarter of the year is a qualitative independent variable. To model a qualitative independent variable, create an independent variable (called a *dummy variable*) for all but one of the qualitative variable's possible values. (It is arbitrary which value you leave out. This example omits Quarter 4.) The dummy variables tell you which value of the qualitative variable occurs. Thus, you have a dummy variable for Quarter 1, Quarter 2, and Quarter 3 with the following properties:

- Quarter 1 dummy variable equals 1 if the quarter is Quarter 1 and 0 if otherwise.
- Quarter 2 dummy variable equals 1 if the quarter is Quarter 2 and 0 if otherwise.
- Quarter 3 dummy variable equals 1 if the quarter is Quarter 3 and 0 if otherwise.

	A	B	C	D	E	F
9	Historical data					
10	Year	Quarter	Sales	GNP	Unemp	Int
11	79	1	Sales	2541	5.9	9.4
12	79	2	2910	2640	5.7	9.4
13	79	3	2562	2595	5.9	9.7
14	79	4	2385	2701	6	12
15	80	1	2520	2785	6.2	13
16	80	2	2142	2509	7.3	9.6
17	80	3	2130	2570	7.7	9.2
18	80	4	2190	2667	7.4	14
19	81	1	2370	2878	7.4	14
20	81	2	2208	2835	7.4	15
21	81	3	2196	2897	7.4	15
22	81	4	1758	2744	8.3	12
23	82	1	1944	2582	8.8	13
24	82	2	2094	2613	9.4	12
25	82	3	1911	2529	10	9.3
26	82	4	2031	2544	10.7	7.9
27	83	1	2046	2633	10.4	7.8
28	83	2	2502	2878	10.1	8.4
29	83	3	2238	3051	9.4	9.1
30	83	4	2394	3274	8.5	8.8
31	84	1	2586	3594	7.9	9.2
32	84	2	2898	3774	7.5	9.8
33	84	3	2448	3861	7.5	10
34	84	4	2460	3919	7.2	8.8
35	85	1	2646	4040	7.4	8.2
36	85	2	2988	4133	7.3	7.5
37	85	3	2967	4303	7.1	7.1
38	85	4	2439	4393	7	7.2
39	86	1	2598	4560	7.1	8.9
40	86	2	3045	4587	7.1	7.7
41	86	3	3213	4716	6.9	7.4
42	86	4	2685	4796	6.8	7.4

Figure 10-7: Auto sales data

A Quarter 4 observation can be identified because the dummy variables for Quarter 1 through Quarter 3 equal 0. It turns out you don't need a dummy variable for Quarter 4. In fact, if you include a dummy variable for Quarter 4 as an independent variable in your regression, Microsoft Office Excel returns an error message. The reason you get an error is because if an exact linear relationship exists between any set of independent variables, Excel must perform the mathematical equivalent of dividing by 0 (an impossibility) when running a multiple regression. In this situation, if you include a Quarter 4 dummy variable, every data point satisfies the following exact linear relationship:

```
(Quarter 1 Dummy)+(Quarter 2 Dummy)+(Quarter 3 Dummy)
+(Quarter 4 Dummy)=1
```

NOTE An exact linear relationship occurs if there exists constants c_0, c_1, … c_N, such that for each data point $c_0 + c_1x_1 + c_2x_2 + \ldots c_Nx_N = 0$. Here x_1, … x_N are the values of the independent variables.

You can interpret the "omitted" dummy variable as a "baseline" scenario; this is reflected in the "regular" intercept. Therefore, you can think of dummies as changes in the intercept.

To create your dummy variable for Quarter 1, copy the formula `IF(B12=1,1,0)` from G12 to G13:G42. This formula places a 1 in column G whenever a quarter is the first quarter, and places a 0 in column G whenever the quarter is not the first quarter. In a similar fashion, you can create dummy variables for Quarter 2 (in H12:H42) and Quarter 3 (in I12:I42). Figure 10-8 shows the results of the formulas.

In addition to seasonality, you'd like to use macroeconomic variables such as gross national product (GNP, in billions of 1986 dollars), interest rates, and unemployment rates to predict car sales. Suppose, for example, that you want to estimate sales for the second quarter of 1979. Because values for GNP, interest rate, and unemployment rate aren't known at the beginning of the second quarter 1979, you can't use the second quarter 1979 GNP, interest rate, and unemployment rate to predict Quarter 2 1979 auto sales. Instead, you use the values for the GNP, interest rate, and unemployment rate lagged one quarter to forecast auto sales. By copying the formula `=D11` from J12 to J12:L42, you can create the lagged value for GNP, the first of your macroeconomic-independent variables. For example, the range J12:L12 contains GNP, unemployment rate, and interest rate for the first quarter of 1979.

You can now run your multiple regression by clicking Data Analysis on the Data tab and then selecting Regression in the Data Analysis dialog box. Use C11:C42 as the Input *Y* Range and G11:L42 as the Input *X* Range; check the Labels box (row 11 contains labels), and also check the Residuals box. After clicking OK, you can

obtain the output, which you can see in the `Regression` worksheet of the file `Autos.xlsx` and in Figure 10-9.

	G	H	I	J	K	L
9						
10	Q1	Q2	Q3	LagGNP	LagUnemp	LagInt
11	Q1	Q2	Q3	LagGNP	LagUnemp	LagInt
12	0	1	0	2541	5.9	9.4
13	0	0	1	2640	5.7	9.4
14	0	0	0	2595	5.9	9.7
15	1	0	0	2701	6	11.9
16	0	1	0	2785	6.2	13.4
17	0	0	1	2509	7.3	9.6
18	0	0	0	2570	7.7	9.2
19	1	0	0	2667	7.4	13.6
20	0	1	0	2878	7.4	14.4
21	0	0	1	2835	7.4	15.3
22	0	0	0	2897	7.4	15.1
23	1	0	0	2744	8.3	11.8
24	0	1	0	2582	8.8	12.8
25	0	0	1	2613	9.4	12.4
26	0	0	0	2529	10	9.3
27	1	0	0	2544	10.7	7.9
28	0	1	0	2633	10.4	7.8
29	0	0	1	2878	10.1	8.4
30	0	0	0	3051	9.4	9.1
31	1	0	0	3274	8.5	8.8
32	0	1	0	3594	7.9	9.2
33	0	0	1	3774	7.5	9.8
34	0	0	0	3861	7.5	10.3
35	1	0	0	3919	7.2	8.8
36	0	1	0	4040	7.4	8.2
37	0	0	1	4133	7.3	7.5
38	0	0	0	4303	7.1	7.1
39	1	0	0	4393	7	7.2
40	0	1	0	4560	7.1	8.9
41	0	0	1	4587	7.1	7.7
42	0	0	0	4716	6.9	7.4

Figure 10-8: Dummy and lagged variables

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.884139126					
5	R Square	0.781701994					
6	Adjusted R Square	0.727127492					
7	Standard Error	190.5240756					
8	Observations	31					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	6	3119625.193	519937.5322	14.32357552	6.79746E-07	
13	Residual	24	871186.1616	36299.4234			
14	Total	30	3990811.355				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	3154.700285	462.6530922	6.818716525	4.7214E-07	2199.83143	4109.56914
18	Q1	156.833091	98.87110703	1.586237838	0.125774521	-47.22680256	360.8929846
19	Q2	379.7835116	96.08921514	3.95240518	0.000594196	181.4651595	578.1018637
20	Q3	203.035501	95.40891864	2.128055783	0.043800625	6.12121161	399.9497905
21	LagGNP	0.174156906	0.05842	2.981117865	0.006490201	0.053583977	0.294729835
22	LagUnemp	-93.83233214	28.32328716	-3.312904029	0.002918487	-152.2887117	-35.37595254
23	LagInt	-73.9167147	17.78851573	-4.155305356	0.000355622	-110.6303992	-37.20303022

Figure 10-9: Summary regression output for auto example

In Figure 10-9, you can see that Equation 1 is used to predict quarterly auto sales as follows:

```
Predicted quarterly sales=3154.7+156.833Q1+379.784Q2+203.03
6Q3+.174(LAGGNP in billions)-93.83(LAGUNEMP)-73.91(LAGINT)
```

Also in Figure 10-9, you see that each independent variable except Q1 has a p-value less than or equal to 0.05. The previous discussion would indicate that you should drop the Q1 variable and rerun the regression. Because Q2 and Q3 are significant, you know there is significant seasonality, so leave Q1 as an independent variable because this treats the seasonality indicator variables as a "package deal." You can therefore conclude that all independent variables have a significant effect on quarterly auto sales. You interpret all coefficients in your regression equation *ceteris paribus* (which means that each coefficient gives the effect of the independent variable after adjusting for the effects of all other variables in the regression).

Each regression coefficient is interpreted as follows:

- A $1 billion increase in last quarter's GNP increases quarterly car sales by 174.
- An increase of 1 percent in last quarter's unemployment rate decreases quarterly car sales by 93,832.
- An increase of 1 percent in last quarter's interest rate decreases quarterly car sales by 73,917.

To interpret the coefficients of the dummy variables, you must realize that they tell you the effect of seasonality relative to the value left out of the qualitative variables. Therefore

- In Quarter 1, car sales exceed Quarter 4 car sales by 156,833.
- In Quarter 2, car sales exceed Quarter 4 car sales by 379,784.
- In Quarter 3, car sales exceed Quarter 4 car sales by 203,036.

Car sales are highest during the second quarter (April through June; tax refunds and summer are coming) and lowest during the third quarter. (October through December; why buy a new car when winter salting will ruin it?)

You should note that each regression coefficient is computed after adjusting for all other independent variables in the equation (this is often referred to as ceteris paribus, or all other things held equal).

From the Summary output shown in Figure 10-9, you can learn the following:

- The variation in your independent variables (macroeconomic factors and seasonality) explains 78 percent of the variation in your dependent variable (quarterly car sales).

- The standard error of your regression is 190,524 cars. You can expect approximately 68 percent of your forecasts to be accurate within 190,524 cars and about 95 percent of your forecasts to be accurate within 381,048 cars (2 * 190,524).
- There are 31 observations used to fit the regression.

The only quantity of interest in the ANOVA portion of Figure 10-9 is the significance (0.00000068). This measure implies that there are only 6.8 chances in 10,000,000, that when taken together, all your independent variables are useless in forecasting car sales. Thus, you can be quite sure that your independent variables are useful in predicting quarterly auto sales.

Figure 10-10 shows for each observation the predicted sales and residual. For example, for the second quarter of 1979 (observation 1), predicted sales from Equation 1 are 2728.6 thousand, and your residual is 181,400 cars (2910 – 2728.6). Note that no residual exceeds 381,000 in absolute value, so you have no outliers.

	A	B	C
27	RESIDUAL OUTPUT		
28			
29	*Observation*	*Predicted Sales*	*Residuals*
30	1	2728.588616	181.4113836
31	2	2587.848606	-25.84860587
32	3	2336.034563	48.96543676
33	4	2339.328281	180.6717193
34	5	2447.266343	-305.2663429
35	6	2400.118977	-270.1189769
36	7	2199.7408	-9.740800106
37	8	2076.383266	293.6167341
38	9	2276.947422	-68.94742189
39	10	2026.185621	169.8143789
40	11	1848.731191	-90.73119119
41	12	2138.394335	-194.3943352
42	13	2212.298456	-118.2984563
43	14	2014.216596	-103.2165964
44	15	1969.394332	61.60566841
45	16	2166.640544	-120.6405443
46	17	2440.632301	61.3676994
47	18	2290.352403	-52.35240272
48	19	2131.386979	262.6130214
49	20	2433.681173	152.3188271
50	21	2739.094517	158.9054833
51	22	2586.877653	-138.8776532
52	23	2362.035446	97.96455439
53	24	2667.994409	-21.99440885
54	25	2937.601377	50.39862257
55	26	2838.174893	128.8251074
56	27	2713.079218	-274.0792178
57	28	2887.577992	-289.5779921
58	29	3004.570968	40.42903226
59	30	2921.225251	291.7747488
60	31	2781.597472	-96.59747188

Figure 10-10: Residual output for Auto example

Modeling Interactions and Nonlinearities

Equation 1 assumes that each independent variable affects *Y* in a linear fashion. This means, for example, that a unit increase in X^1 will increase *Y* by B_1 for any values of $X_1, X_2, \ldots, X_n$. In many marketing situations this assumption of linearity is unrealistic. In this section, you learn how to model situations in which an independent variable can interact with or influence *Y* in a nonlinear fashion.

Nonlinear Relationship

An independent variable can often influence a dependent variable through a nonlinear relationship. For example, if you try to predict product sales using an equation such as the following, price influences sales linearly.

$$\text{Sales} = 500 - 10 * \textit{Price}$$

This equation indicates that a unit increase in price can (at any price level) reduce sales by 10 units. If the relationship between sales and price were governed by an equation such as the following, price and sales would be related nonlinearly.

$$\text{Sales} = 500 + 4 * \textit{Price} - .40 * \textit{Price}^2$$

As shown in Figure 10-11, larger increases in price result in larger decreases in demand. In short, if the change in the dependent variable caused by a unit change in the independent variable is not constant, there is a nonlinear relationship between the independent and dependent variables.

	A	B
4	Price	Demand
5	1	495.6
6	2	490.4
7	3	484.4
8	4	477.6
9	5	470
10	6	461.6
11	7	452.4
12	8	442.4
13	9	431.6
14	10	420
15	11	407.6
16	12	394.4
17	13	380.4
18	14	365.6
19	15	350
20	16	333.6
21	17	316.4
22	18	298.4
23	19	279.6
24	20	260

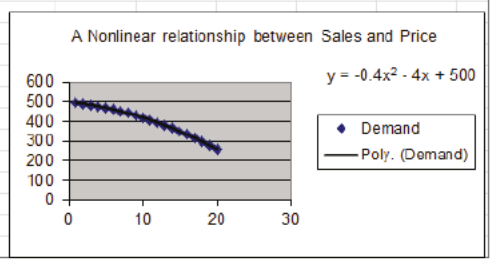

Figure 10-11: Nonlinear relationship between Sales and Price

Interaction

If the effect of one independent variable on a dependent variable depends on the value of another independent variable, you can say that the two independent variables exhibit *interaction*. For example, suppose you try to predict sales using the price and the amount spent on advertising. If the effect to change the level of advertising dollars is large when the price is low and small when the price is high, price and advertising exhibit interaction. If the effect to change the level of advertising dollars is the same for any price level, sales and price do not exhibit any interaction. You will encounter interactions again in Chapter 41, "Analysis of Variance: Two-way ANOVA."

Testing for Nonlinearities and Interactions

To see whether an independent variable has a nonlinear effect on a dependent variable, simply add an independent variable to the regression that equals the square of the independent variable. If the squared term has a low p-value (less than 0.05), you have evidence of a nonlinear relationship.

To check whether two independent variables exhibit interaction, simply add a term to the regression that equals the product of the independent variables. If the term has a low p-value (less than 0.05), you have evidence of interaction. The file `Priceandads.xlsx` illustrates this procedure. In worksheet `data` from this file (see Figure 10-12), you have the weekly unit sales of a product, weekly price, and weekly ad expenditures (in thousands of dollars).

With this example, you'll want to predict weekly sales from the price and advertising. To determine whether the relationship is nonlinear or exhibits any interactions, perform the following steps:

1. Add in Column H *Advertising*Price*, in Column I $Price^2$, and in Column J Ad^2.
2. Next, run a regression with *Y* Range E4:E169 and *X* Range F4:J169. You can obtain the regression output, as shown in the worksheet `nonlinear` and Figure 10-13.
3. All independent variables except for $Price^2$ have significant p-values (less than .05). Therefore, drop $Price^2$ as an independent variable and rerun the regression. The result is in Figure 10-14 and the worksheet `final`.

	E	F	G	H	I	J
1						
2						
3						
4	Sales	Price	Ad	A*P	Price^2	Ad^2
5	22845	8	1	8	64	1
6	20417	9	8	72	81	64
7	23761	5	3	15	25	9
8	22674	4	12	48	16	144
9	22782	7	5	35	49	25
10	23807	5	3	15	25	9
11	18924	10	9	90	100	81
12	21855	9	5	45	81	25
13	21749	10	4	40	100	16
14	22683	4	12	48	16	144
15	20968	6	11	66	36	121
16	22202	10	2	20	100	4
17	23241	6	5	30	36	25
18	19004	10	9	90	100	81
19	23978	4	1	4	16	1
20	20497	8	9	72	64	81
21	22322	9	3	27	81	9
22	22628	8	4	32	64	16
23	21051	10	6	60	100	36
24	24515	3	3	9	9	9
25	22126	10	2	20	100	4
26	22141	5	11	55	25	121
27	21151	9	7	63	81	49
28	22558	5	10	50	25	100

Figure 10-12: Nonlinearity and interaction data

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.996924531							
5	R Square	0.99385852							
6	Adjusted R Square	0.993665392		Price^2 has high p value so delete it and rerun analysis					
7	Standard Error	135.2764087							
8	Observations	165							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	5	470861057.7	94172212	5146.105	7.9582E-174			
13	Residual	159	2909653.375	18299.71					
14	Total	164	473770711.1						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	24005.74767	111.4951345	215.3076	5.5E-198	23785.54521	24225.95	23785.55	24225.95
18	Price	-135.6678621	32.18950019	-4.21466	4.18E-05	-199.242002	-72.0937	-199.242	-72.0937
19	Ad	660.0035108	16.15110952	40.86428	3.07E-86	628.1051313	691.9019	628.1051	691.9019
20	A*P	-74.12725368	1.425595543	-51.9974	1E-101	-76.94279942	-71.3117	-76.9428	-71.3117
21	Price^2	-0.178202781	2.349205511	-0.07586	0.939629	-4.817874684	4.461469	-4.81787	4.461469
22	Ad^2	-37.37381917	1.019418942	-36.6619	1.84E-79	-39.38716769	-35.3605	-39.3872	-35.3605

Figure 10-13: First regression output for Nonlinearity and Interaction example

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.996924419							
5	R Square	0.993858298		All ind variables have low p value so use this equation to predict sales					
6	Adjusted R Squ	0.993704755		Ads have nonlinear effect and Price and ads interact.					
7	Standard Error	134.8554475		At higher price ads have less effect on sales					
8	Observations	165							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	470860952.4	117715238.1	6472.852287	9.3099E-176			
13	Residual	160	2909758.675	18185.99172					
14	Total	164	473770711.1						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	24012.24758	71.11479957	337.6547179	3.3475E-230	23871.80286	24152.69231	23871.80286	24152.69231
18	Price	-137.997013	9.633696108	-14.32441001	1.78044E-30	-157.0226141	-118.9714118	-157.0226141	-118.9714118
19	Ad	660.0418883	16.09294845	41.01435424	8.35145E-87	628.2598999	691.8238767	628.2598999	691.8238767
20	A*P	-74.12897476	1.420979292	-52.16752641	2.3559E-102	-76.93526893	-71.32268059	-76.93526893	-71.32268059
21	Ad^2	-37.37288222	1.016172056	-36.77810466	5.94521E-80	-39.37972197	-35.36604248	-39.37972197	-35.36604248

Figure 10-14: Final regression output for Nonlinearity and Interaction example

The Significance F Value is small, so the regression model has significant predictive values. All independent variables have extremely small p-values, so you can predict the weekly unit sales with the equation

$$\text{Predicted Unit Sales} = 24{,}012 - 138 * \text{Price} + 660.04 * \text{Ad} - 74.13 * \text{Ad} * \text{P} - 37.33\text{AD}^2$$

The –37.33 Ad^2 term implies that each additional $1,000 in advertising can generate fewer sales (diminishing returns). The –74.13*Ad*P term implies that at higher prices additional advertising has a smaller effect on sales.

The R^2 value of 99.4 percent implies your model explains 99.4 percent of the variation in weekly sales. The Standard Error of 134.86 implies that roughly 95 percent of your forecasts should be accurate within 269.71. Interactions and nonlinear effects are likely to cause multicollinearity, which is covered in the section "Multicollinearity" later in this chapter.

Testing Validity of Regression Assumptions

Recall earlier in the chapter you learned the regression assumptions that should be satisfied by the error term in a multiple linear regression. For ease of presentation, these assumptions are repeated here:

- The error term is normally distributed.
- The variability or spread of the error term is assumed not to depend on the value of the dependent variable.

- For time series data, successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.

This section further discusses how to determine if these assumptions are satisfied, the consequences of violating the assumptions, and how to resolve violation of these assumptions.

Normally Distributed Error Term

You can infer the nature of an unknown error term through examination of the residuals. If the residuals come from a normal random variable, the normal random variable should have a symmetric density. Then the skewness (as measured by Excel SKEW function described in Chapter 2) should be near 0.

Kurtosis, which may sound like a disease but isn't, can also help you identify if the residuals are likely to have come from a normal random variable. Kurtosis near 0 means a data set exhibits "peakedness" close to the normal. Positive kurtosis means that a data set is more peaked than a normal random variable, whereas negative kurtosis means that data is less peaked than a normal random variable. The kurtosis of a data set may be computed with the Excel KURT function.

For different size data sets, Figure 10-15 gives 95 percent confidence intervals for the skewness and kurtosis of data drawn from a normal random variable.

	Kurtosis	Kurtosis	Skewness	Skewness
Sample Size	2.5	97.5	2.5	97.5
10	-1.74	3.41	-1.37	1.36
20	-1.27	2.46	-1.02	1.03
30	-1.09	2.06	-0.86	0.85
40	-0.99	1.77	-0.73	0.75
50	-0.91	1.62	-0.66	0.67
60	-0.85	1.49	-0.61	0.62
70	-0.80	1.36	-0.57	0.57
80	-0.77	1.27	-0.53	0.54
90	-0.73	1.20	-0.51	0.51
100	-0.71	1.13	-0.48	0.48

Figure 10-15: 95 percent confidence interval for skewness and kurtosis for sample from a normal distribution

For example, it is 95 percent certain that in a sample of size 50 from a normal random variable, kurtosis is between –0.91 and 1.62. It is also 95 percent certain that

in a sample of size 50 from a normal random variable, skewness is between –0.66 and 0.67. If your residuals yield a skewness or kurtosis outside the range shown in Figure 10-15, then you have reason to doubt the normality assumption.

In the computer spending example for European countries, you obtained a skewness of 0.83 and a kurtosis of 0.18. Both these numbers are inside the ranges specified in Figure 10-15, so you have no reason to doubt the normality of the residuals.

Non-normality of the residuals invalidates the p-values that you used to determine significance of independent variables or the entire regression. The most common solution to the problem of non-normal random variables is to transform the dependent variable. Often replacing y by Ln y, $\sqrt{y}$, or $\frac{1}{y}$ can resolve the non-normality of the errors.

Heteroscedasticity: A Nonconstant Variance Error Term

If larger values of an independent variable lead to a larger variance in the errors, you have violated the constant variance of the error term assumption, and *heteroscedasticity* is present. Heteroscedasticity, like non-normal residuals, invalidates the p-values used earlier in the chapter to test for significance. In most cases you can identify heteroscedasticity by graphing the predicted value on the x-axis and the absolute value of the residual on the y-axis. To see an illustration of this, look at the file `Heteroscedasticity.xlsx`. A sample of the data is shown in Figure 10-16.

In this file, you are using the data in `Heteroscedasticity.xlsx` and trying to predict the amount a family spends annually on food from their annual income. After running a regression, you can graph the absolute value of the residuals against predicted food spending. Figure 10-17 shows the resulting graph.

The upward slope of the line that best fits the graph indicates that your forecast accuracy decreases for families with more income, and heteroscedasticity is clearly present. Usually heteroscedasticity is resolved by replacing the dependent variable Y by Ln Y or $\sqrt{Y}$. The reason why these transformations often resolve heteroscedasticity is that these transformations reduce the spread in the dependent variable. For example, if three data points have Y = 1, Y = 10,000 and Y = 1,000,000 then after using the $\sqrt{Y}$ transformation the three points now have a dependent variable with values 1, 100, and 1000 respectively.

	I	J
4	Income	Food spending
5	$74,201.00	$9,646.13
6	$41,659.00	$8,331.80
7	$44,085.00	$9,698.70
8	$63,529.00	$10,799.93
9	$48,436.00	$9,202.84
10	$82,481.00	$13,196.96
11	$35,243.00	$4,934.02
12	$57,563.00	$9,210.08
13	$39,589.00	$5,938.35
14	$53,826.00	$10,226.94
15	$78,861.00	$14,194.98
16	$87,406.00	$11,362.78
17	$74,020.00	$15,544.20
18	$82,290.00	$9,874.80
19	$38,921.00	$4,670.52
20	$80,960.00	$17,001.60
21	$37,107.00	$8,163.54
22	$80,531.00	$14,495.58
23	$79,760.00	$13,559.20
24	$57,427.00	$12,633.94
25	$67,657.00	$9,471.98
26	$75,449.00	$14,335.31
27	$71,390.00	$10,708.50

Figure 10-16: Heteroscedasticity data

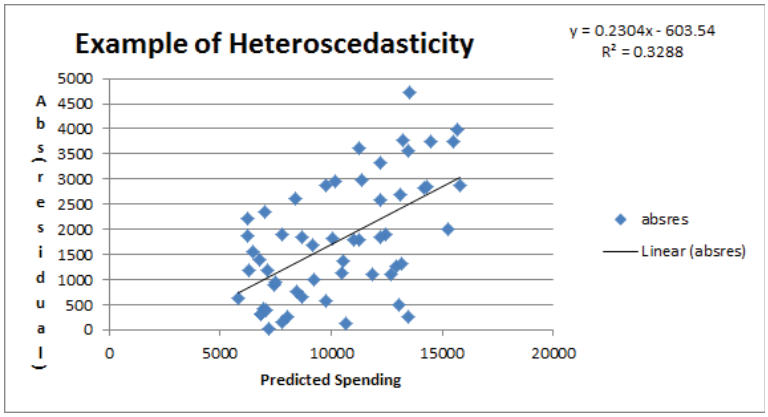

Figure 10-17: Example of Heteroscedasticity

Autocorrelation: The Nonindependence of Errors

Suppose your data is times series data. This implies the data is listed in chronological order. The `auto` data is a good example. The p-values used to test the hypothesis

of no linear regression and the significance of an independent variable are not valid if your error terms appear to be dependent (nonindependent). Also, if your error terms are nonindependent, you can say that *autocorrelation* is present. If autocorrelation is present, you can no longer be sure that 95 percent of your forecasts will be accurate within two standard errors. Probably fewer than 95 percent of your forecasts will be accurate within two standard errors. This means that in the presence of autocorrelation, your forecasts can give a false sense of security. Because the residuals mirror the theoretical value of the error terms in Equation 1, the easiest way to see if autocorrelation is present is to look at a plot of residuals in chronological order. Recall the residuals sum to 0, so approximately half are positive and half are negative. If your residuals are independent, you would expect sequences of the form ++, + –, – +, and – – to be equally likely. Here + is a positive residual and – is a negative residual.

Graphical Interpretation of Autocorrelation

You can use a simple time series plot of residuals to determine if the error terms exhibit autocorrelation, and if so, the type of autocorrelation that is present.

Figure 10-18 shows an illustration of independent residuals exhibiting no autocorrelation.

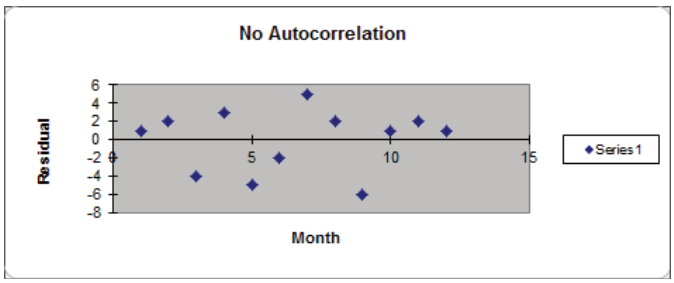

Figure 10-18: Residuals indicate no autocorrelation

Here you can see 6 changes in sign out of 11 possible changes.

Figure 10-19, however, is indicative of *positive autocorrelation*. Figure 10-19 shows only one sign change out of 11 possible changes. Positive residuals are followed by positive residuals, and negative residuals are followed by negative residuals. Thus, successive residuals are positively correlated. When residuals exhibit few sign changes (relative to half the possible number of sign changes), positive autocorrelation is suspected. Unfortunately, positive autocorrelation is common in business and economic data.

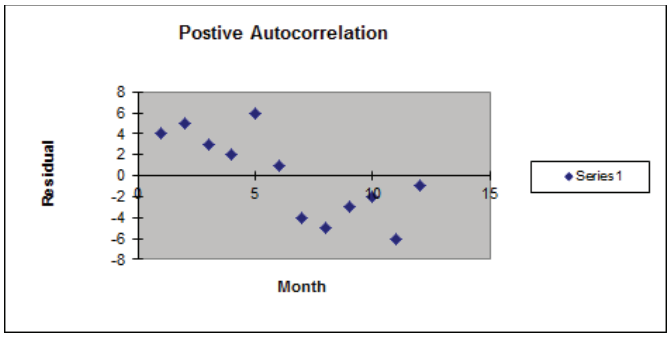

Figure 10-19: Residuals indicate positive autocorrelation

Figure 10-20 is indicative of *negative autocorrelation*. Figure 10-20 shows 11 sign changes out of a possible 11. This indicates that a small residual tends to be followed by a large residual, and a large residual tends to be followed a small residual. Thus, successive residuals are negatively correlated. This shows that many sign changes (relative to half the number of possible sign changes) are indicative of negative autocorrelation.

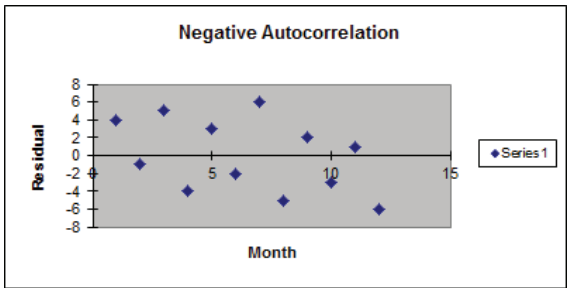

Figure 10-20: Residuals indicate negative autocorrelation

To help clarify these three different types of graphical interpretation, suppose you have n observations. If your residuals exhibit no correlation, then the chance of seeing either less than $\frac{n-1}{2} - \sqrt{n-1}$ or more than $\frac{n-1}{2} + \sqrt{n-1}$ sign changes is approximately 5 percent. Thus you can conclude the following:

- If you observe less than or equal to $\frac{n-1}{2} - \sqrt{n-1}$ sign changes, conclude that positive autocorrelation is present.
- If you observe at least $\frac{n-1}{2} + \sqrt{n-1}$ sign changes, conclude that negative autocorrelation is present.
- Otherwise you can conclude that no autocorrelation is present.

Detecting and Correcting for Autocorrelation

The simplest method to correct for autocorrelation is presented in the following steps. To simplify the presentation, assume there is only one independent variable (Call it *X*):

1. Determine the correlation between the following two time series: your residuals and your residuals lagged one period. Call this correlation *p*.
2. Run a regression with the dependent variable for time *t* being $Y_t - pY_{t-1}$ and independent variable $X_t - pX_{t-1}$.
3. Check the number of sign changes in the new regression's residuals. Usually, autocorrelation is no longer a problem, and you can rearrange your equation to predict Y_t from Y_{t-1}, X_t, and X_{t-1}.

To illustrate this procedure, you can try and predict consumer spending (in billions of $) during a year as a function of the money supply (in billions of $). Twenty years of data are given in Figure 10-21 and are available for download from the file `autocorr.xls`.

Now complete the following steps:

1. Run a regression with *X* Range B1:B21 and *Y* Range A1:A21, and check the Labels and Residuals box. Figure 10-22 shows the residuals.
2. Observe that a sign change in the residuals occurs if, and only if, the product of two successive residuals is <0. Therefore, copying the formula `=IF(I27*I26<0,1,0)` from J27 to J28:J45 counts the number of sign changes. Compute the total number of sign changes (4) in cell J24 with the formula `=SUM(J27:J45)`.

	A	B
1	Exp	Money stock
2	214.6	159.3
3	217.7	161.2
4	219.6	162.8
5	227.2	164.6
6	230.9	165.9
7	233.3	167.9
8	234.1	168.3
9	232.3	169.7
10	233.7	170.5
11	236.5	171.6
12	238.7	173.9
13	243.2	176.1
14	249.4	178
15	254.3	179.1
16	260.9	180.2
17	263.3	181.2
18	265.6	181.6
19	268.2	182.5
20	270.4	183.3
21	275.6	184.3

Figure 10-21: Data for Autocorrelation example

	E	F	G	H	I	J	K
23			RESIDUAL OUTPUT			sign changes	
24						4	
25			*Observation*	*Predicted Exp*	*Residuals*		*Residual(t-1)*
26	correlation		1	211.7298848	2.87012	Sign change	
27	Residual(t)		2	216.1005891	1.59941	0	2.87012
28	and Residual(t-1)		3	219.7811822	-0.18118	1	1.59941
29	0.8227		4	223.9218494	3.27815	1	-0.18118
30			5	226.9123312	3.98767	0	3.27815
31			6	231.5130725	1.78693	0	3.98767
32			7	232.4332208	1.66678	0	1.78693
33			8	235.6537397	-3.35374	1	1.66678
34			9	237.4940363	-3.79404	0	-3.35374
35			10	240.024444	-3.52444	0	-3.79404
36			11	245.3152965	-6.6153	0	-3.52444
37			12	250.376112	-7.17611	0	-6.6153
38			13	254.7468163	-5.34682	0	-7.17611
39			14	257.277224	-2.97722	0	-5.34682
40			15	259.8076317	1.09237	1	-2.97722
41			16	262.1080024	1.192	0	1.09237
42			17	263.0281506	2.57185	0	1.192
43			18	265.0984842	3.10152	0	2.57185
44			19	266.9387808	3.46122	0	3.10152
45			20	269.2391514	6.36085	0	3.46122

Figure 10-22: Residuals for Autocorrelation example

3. In cell J22 compute the "cutoff" for the number of sign changes that indicates the presence of positive autocorrelation. If the number of sign changes is <5.41, then you can suspect the positive autocorrelation is present: `=9.5-SQRT(19)`.
4. Because you have only four sign changes, you can conclude that positive autocorrelation is present.
5. To correct for autocorrelation, find the correlation between the residuals and lagged residuals. Create the lagged residuals in K27:K45 by copying the formula `=I26` from K27 to K28:K45.
6. Find the correlation between the residuals and lagged residuals (0.82) in cell L26 using the formula `=CORREL(I27:I45, K27:K45)`.
7. To correct for autocorrelation run a regression with dependent variable $Expenditures_t$ – .82 $Expenditures_{t-1}$ and independent variable $Money\ Supply_t$ – .82 $Money\ Supply_{t-1}$. See Figure 10-23.

	A	B	C	D
1	Exp	Money stock	Exp(t)-.82Exp(t-	MS(t)-.82(MS(t-1)
2	214.6	159.3		
3	217.7	161.2	41.728	30.574
4	219.6	162.8	41.086	30.616
5	227.2	164.6	47.128	31.104
6	230.9	165.9	44.596	30.928
7	233.3	167.9	43.962	31.862
8	234.1	168.3	42.794	30.622
9	232.3	169.7	40.338	31.694
10	233.7	170.5	43.214	31.346
11	236.5	171.6	44.866	31.79
12	238.7	173.9	44.77	33.188
13	243.2	176.1	47.466	33.502
14	249.4	178	49.976	33.598
15	254.3	179.1	49.792	33.14
16	260.9	180.2	52.374	33.338
17	263.3	181.2	49.362	33.436
18	265.6	181.6	49.694	33.016
19	268.2	182.5	50.408	33.588
20	270.4	183.3	50.476	33.65
21	275.6	184.3	53.872	33.994

Figure 10-23: Transformed data to correct for autocorrelation

8. In Column C create your transformed dependent variable by copying the formula `=A3-0.82*A2` from C3 to C4:C21.
9. Copy this same formula from D3 to D4:D21 to create the transformed independent variable *Money Supply*$_t$ – .82*Money Supply*$_{t-1}$.
10. Now run a regression with the *Y* Range as C3:C21 and *X* Range as D3:D21. Figure 10-24 shows the results.

Because the p-value for your independent variable is less than .15, you can conclude that your transformed independent variable is useful for predicting your transformed independent variable. You can find the residuals from your new regression change sign seven times. This exceeds the positive autocorrelation cutoff of 4.37 sign changes. Therefore you can conclude that you have successfully removed the positive autocorrelation. You can predict period *t* expenditures with the following equation:

Period *t* expenditures – 0.82Period(t – 1) Expenditures = -41.97 + 2.74(Period(t) Money Supply – .82Period(t – 1) Money Supply)

	H	I	J	K	L	M	N
1							
2	SUMMARY OUTPUT						
3			Exp(t)-.82Exp(t-1)=-41.98+2.74*(MS(t)-.82MS(t-1))				
4	*Regression Statistics*		or				
5	Multiple R	0.82491	Exp(t)=.82Exp(t-1)-41.98+2.74*(MS(t)-.82MS(t-1))				
6	R Square	0.68048					
7	Adjusted R Square	0.66051					
8	Standard Error	2.30236					
9	Observations	18					
10							
11	ANOVA						
12		*df*	*SS*	*MS*	*F*	*Significance F*	
13	Regression	1	180.6254254	180.625	34.0748	2.5233E-05	
14	Residual	16	84.81364059	5.30085			
15	Total	17	265.439066				
16							
17		*Coefficient*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
18	Intercept	-41.9766	15.25392438	-2.75186	0.01418	-74.313504	-9.6397687
19	Var 1	2.74079	0.469526345	5.83736	2.5E-05	1.74544379	3.7361461

Figure 10-24: Regression output for transformed data

You can rewrite this equation as the following:

Period t expenditures = .82Period(t – 1) Expenditures – 41.97 + 2.74(Period(t) Money Supply – .82Period(t – 1) Money Supply)

Because everything on the right hand side of the last equation is known at Period t, you can use this equation to predict Period t expenditures.

Multicollinearity

If two or more independent variables in a regression analysis are highly correlated, a regression analysis may yield strange results. Whenever two or more independent variables are highly correlated and the regression coefficients do not make sense, you can say that *multicollinearity* exists.

Figure 10-25 (see file `housing.xls`) gives the following data for the years 1963–1985: the number of housing starts (in thousands), U.S. population (in millions), and mortgage rate. You can use this data to develop an equation that can forecast housing starts by performing the following steps:

1. It seems logical that housing starts should increase over time, so include the year as an independent variable to account for an upward trend. The more people in the United States, the more housing starts you would expect, so include Housing Starts as an independent variable. Clearly, an increase in mortgage rates decreases housing starts, so include the mortgage rate as an independent variable.

	A	B	C	D
1	Housing data			
2	thousand	millions		
3	Starts	Pop	Mort Rate	Year
4	1635	189	5.89	1963
5	1561	192	5.82	1964
6	1510	194	5.81	1965
7	1196	197	6.25	1966
8	1322	199	6.46	1967
9	1545	201	6.97	1968
10	1500	203	7.8	1969
11	1434	205	8.45	1970
12	2085	208	7.74	1971
13	2379	210	7.6	1972
14	2057	212	7.96	1973
15	1353	214	8.92	1974
16	1171	216	9	1975
17	1548	218	9	1976
18	2002	220	9.02	1977
19	2036	223	9.56	1978
20	1760	225	10.78	1979
21	1312	228	12.66	1980
22	1100	230	14.7	1981
23	1072	232	15.14	1982
24	1712	234	12.57	1983
25	1756	236	12.38	1984
26	1745	238	11.55	1985

Figure 10-25: Multicollinearity data

2. Now run a multiple regression with the *Y* range being A3:A26 and the *X* Range being B3:D26 to obtain the results shown in Figure 10-26.
3. Observe that neither `POP` nor `YEAR` is significant. (They have p-values of .59 and .74, respectively.) Also, the negative coefficient of `YEAR` indicates that there is a downward trend in housing starts. This doesn't make sense though. The problem is that `POP` and `YEAR` are highly correlated. To see this, use the `DATA ANALYSIS TOOLS CORRELATION` command to find the correlations between the independent variables.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.660983152					
5	R Square	0.436898728					
6	Adjusted R Squ	0.347988001					
7	Standard Error	279.5855911					
8	Observations	23					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	1152331.526	384110.5087	4.913903437	0.010796446	
13	Residual	19	1485193.952	78168.10274			
14	Total	22	2637525.478				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	221794.5255	661959.9507	0.335057318	0.741252754	-1163704.005	1607293.056
18	Pop	90.39255757	162.9146358	0.554846145	0.585473195	-250.5917998	431.376915
19	Mort Rate	-206.8964396	55.49611616	-3.728124668	0.00142535	-323.0511818	-90.74169748
20	Year	-120.3847361	352.8922741	-0.341137352	0.736743719	-858.9969839	618.2275117

Figure 10-26: First regression output: Multicoillinearity example

4. Select Input Range B3:D26.
5. Check the labels box.
6. Put the output on the new sheet Correlation.

You should obtain the output in Figure 10-27.

	A	B	C	D
1		*Pop*	*Mort Rate*	*Year*
2	Pop	1		
3	Mort Rate	0.913679	1	
4	Year	0.999655	0.90995	1

Figure 10-27: Correlation matrix for Multicollinearity example

The .999 correlation between POP and YEAR occurs because both POP and YEAR increase linearly over time. Also note that the correlation between Mort Rate and the other two independent variables exceeds .9. Due to this, *multicollinearity* exists. What has happened is that the high correlation between the independent variables has confused the computer about which independent variables are important. The solution to this problem is to drop one or more of the highly correlated independent variables and hope that the independent variables remaining in the regression will be significant. If you decide to drop YEAR, change your *X* Range to B3:C26 to obtain the output shown in Figure 10-28. If you have access to a statistical package, such as SAS or SPSS, you can identify the presence of multicollinearity by looking at the Variance Inflation Factor (VIF) of each independent variable. A general rule of thumb is that any independent variable with a variance inflation factor exceeding 5 is evidence of multicollinearity.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.658369001					
5	R Square	0.433449742					
6	Adjusted R Squa	0.376794716					
7	Standard Error	273.3396002					
8	Observations	23					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	1143234.737	571617.3685	7.650684739	0.003407096	
13	Residual	20	1494290.741	74714.53706			
14	Total	22	2637525.478				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-4024.025797	1627.113433	-2.473107108	0.022486161	-7418.123366	-629.9282276
18	Pop	34.91659242	9.564839902	3.650515092	0.001590215	14.96469527	54.86848956
19	Mort Rate	-200.8475581	51.41238103	-3.906599034	0.000875172	-308.0918558	-93.60326032

Figure 10-28: Final regression output for Multicollinearity example

POP is now highly significant (p-value = .001). Also, by dropping YEAR you actually decreased s_e from 280 to 273. This decrease is because dropping YEAR reduced the

confusion the computer had due to the strong correlation between POP and YEAR. The final predictive equation is as follows:

$$\text{Housing Starts} = -4024.03 + 34.92\text{POP} - 200.85\text{MORT RAT}$$

The interpretation of this equation is that after adjusting for interest rates, an increase in U.S. population of one million people results in $34,920 in housing starts. After adjusting for Population, an increase in interest rates of 1 percent can reduce housing starts by $200,850. This is valuable information that could be used to forecast the future cash flows of construction-related industries.

NOTE After correcting for multicollinearity, the independent variables now have signs that agree with common sense. This is a common by-product of correcting for multicollinearity.

Validation of a Regression

The ultimate goal of regression analysis is for the estimated models to be used for accurate forecasting. When using a regression equation to make forecasts for the future, you must avoid over fitting a set of data. For example, if you had seven data points and only one independent variable, you could obtain an $R^2 = 1$ by fitting a sixth degree polynomial to the data. Unfortunately, such an equation would probably work poorly in fitting future data. Whenever you have a reasonable amount of data, you should hold back approximately 20 percent of your data (called the *Validation Set*) to validate your forecasts. To do this, simply fit regression to 80 percent of your data (called the *Test Set*). Compute the standard deviation of the errors for this data. Now use the equation generated from the Test Set to compute forecasts and the standard deviation of the errors for the Validation Set. Hopefully, the standard deviation for the Validation Set will be fairly close to the standard deviation for the Test Set. If this is the case, you can use the regression equation for future forecasts and be fairly confident that the accuracy of future forecasts will be approximated by the s_e for the Test Set. You can illustrate the important idea of validation with the data from your housing example.

Using the years 1963–1980 as your Test Set and the years 1981–1985 as the Validation Set, you can determine the suitability of the regression with independent variables *POP* and *MORT RAT* for future forecasting using the powerful TREND function. The syntax of the TREND function is TREND(known_y's,[known_x's],[new_x's],[const]). This function fits a multiple regression using the known *y*'s and known *x*'s and then uses this regression to make forecasts for the dependent variable using

the new *x*'s data. `[Constant]` is an optional argument. Setting `[Constant]=False` causes Excel to fit the regression with the constant term set equal to 0. Setting `[Constant]=True` or omitting `[Constant]` causes Excel to fit a regression in the normal fashion.

The `TREND` function is an array function (see Chapter 2) so you need to select the cell range populated by the `TREND` function and finally press Ctrl+Shift+Enter to enable `TREND` to calculate the desired results. As shown in Figure 10-29 and worksheet `Data`, you will now use the `TREND` function to compare the accuracy of regression predictions for the 1981-1985 validation period to the accuracy of regression predictions for the fitted data using the following steps.

	A	B	C	D	E	F	G	H
1	Housing data						Std Dev	
2	thousand millions						1963-1980	285.701
3	Starts	Pop	Mort Rate	Year	Predictions	Error	1981-1985	255.886
4	1635	189	5.89	1963	1329.1723	305.828		
5	1561	192	5.82	1964	1490.6514	70.3486		
6	1510	194	5.81	1965	1587.9926	-77.9926		
7	1196	197	6.25	1966	1606.047	-410.047		
8	1322	199	6.46	1967	1641.5187	-319.519		
9	1545	201	6.97	1968	1592.6229	-47.6229		
10	1500	203	7.8	1969	1453.7352	46.2648		
11	1434	205	8.45	1970	1365.468	68.532		
12	2085	208	7.74	1971	1706.931	378.069		
13	2379	210	7.6	1972	1840.8314	538.169		
14	2057	212	7.96	1973	1834.1193	222.881		
15	1353	214	8.92	1974	1658.6724	-305.672		
16	1171	216	9	1975	1730.7033	-559.703		
17	1548	218	9	1976	1825.2322	-277.232		
18	2002	220	9.02	1977	1914.1366	87.8634		
19	2036	223	9.56	1978	1904.0686	131.931		
20	1760	225	10.78	1979	1655.5031	104.497		
21	1312	228	12.66	1980	1268.5938	43.4062		
22	1100	230	14.7	1981	789.42395	310.576		
23	1072	232	15.14	1982	760.21392	311.786		
24	1712	234	12.57	1983	1577.4907	134.509		
25	1756	236	12.38	1984	1725.4524	30.5476		
26	1745	238	11.55	1985	2053.3979	-308.398		

Figure 10-29: Use of Trend function to validate regression

1. To generate forecasts for the years 1963–1985 using the 1963–1980 data, simply select the range E4:E26 and enter in E4 the array formula `=TREND(A4:A21, B4:C21,B4:C26)` (refer to Figure 10-29). Rows 4-21 contain the data for the years 1963-1980 and Rows 4-26 contain the data for the years 1963-1985.
2. Compute the error for each year's forecast in Column F. The error for 1963 is computed in F4 with the formula `=A4-F4`.
3. Copy this formula down to row 26 to compute the errors for the years 1964–1985.
4. In cell H2 compute the standard deviation (285.70) of the errors for the years 1963–1980 with the formula `=STDEV(F4:F21)`.

5. In cell H3 compute the standard deviation (255.89) of the forecast errors for the years 1981–1985 with the formula `=AVERAGE(F22:F26)`.

The forecasts are actually more accurate for the Validation Set! This is unusual, but it gives you confidence that 95 percent of all future forecasts should be accurate within $2s_e$ = 546,700 housing starts.

Summary

In this chapter you learned the following:

- The multiple linear regression model models a dependent variable Y as $B_0 + B_1X_1 + B_2X_2 + \ldots B_nX_n + \textit{error term}$.
- The error term is required to satisfy the following assumptions:
  - The error term is normally distributed.
  - The variability or spread of the error term is assumed not to depend on the value of the dependent variable.
  - For time series data, successive values of the error term must be independent. This means, for example, that if for one observation the error term is a large positive number, then this tells you nothing about the value of successive error terms.
- Violation of these assumptions can invalidate the p-values in the Excel output.
- You can run a regression analysis using the Data Analysis Tool.
- The Coefficients portion of the output gives the least squares estimates of $B_0, B_1, \ldots, B_n$.
- A Significance F in the ANOVA section of the output less than .05 causes you to reject the hypothesis of no linear regression and conclude that your independent variables have significant predictive value.
- Independent variables with p-value greater than .05 should be deleted, and the regression should be rerun until all independent variables have p-values of .05 or less.
- Approximately 68 percent of predictions from a regression should be accurate within one standard error and approximately 95 percent of predictions from a regression should be accurate within two standard errors.
- Qualitative independent variables are modeled using indicator variables.
- By adding the square of an independent variable as a new independent variable, you can test whether the independent variable has a nonlinear effect on Y.

- By adding the product of two independent variables (say X_1 and X_2) as a new independent variable, you can test whether X_1 and X_2 interact in their effect on Y.
- You can check for the presence of autocorrelation in a regression based on time series data by examining the number of sign changes in the residuals; too few sign changes indicate positive autocorrelation and too many sign changes indicate negative autocorrelation.
- If independent variables are highly correlated, then their coefficients in a regression may be misleading. This is known as multicollinearity.

Exercises

1. Fizzy Drugs wants to optimize the yield from an important chemical process. The company thinks that the number of pounds produced each time the process runs depends on the size of the container used, the pressure, and the temperature. The scientists involved believe the effect to change one variable might depend on the values of other variables. The size of the process container must be between 1.3 and 1.5 cubic meters; pressure must be between 4 and 4.5 mm; and temperature must be between 22 and 30 degrees Celsius. The scientists patiently set up experiments at the lower and upper levels of the three control variables and obtain the data shown in the file `Fizzy.xlsx`.

   a. Determine the relationship between yield, size, temperature, and pressure.
   b. Discuss the interactions between pressure, size, and temperature.
   c. What settings for temperature, size, and pressure would you recommend?

2. For 12 straight weeks, you have observed the sales (in number of cases) of canned tomatoes at Mr. D's Supermarket. (See the file `Grocery.xlsx`.) Each week, you keep track of the following:

   a. Was a promotional notice for canned tomatoes placed in all shopping carts?
   b. Was a coupon for canned tomatoes given to each customer?
   c. Was a price reduction (none, 1, or 2 cents off) given?

Use this data to determine how the preceding factors influence sales. Predict sales of canned tomatoes during a week in which you use a shopping cart notice, a coupon, and reduce price by 1 cent.

3. The file `Countryregion.xlsx` contains the following data for several underdeveloped countries:

   - Infant mortality rate
   - Adult literacy rate
   - Percentage of students finishing primary school
   - Per capita GNP

   Use this data to develop an equation that can be used to predict infant mortality. Are there any outliers in this set of data? Interpret the coefficients in your equation. Within what value should 95 percent of your predictions for infant mortality be accurate?

4. The file `Baseball96.xlsx` gives runs scored, singles, doubles, triples, home runs, and bases stolen for each major league baseball team during the 1996 season. Use this data to determine the effects of singles, doubles, and other activities on run production.
5. The file `Cardata.xlsx` provides the following information for 392 different car models:

   - Cylinders
   - Displacement
   - Horsepower
   - Weight
   - Acceleration
   - Miles per gallon (MPG)

   Determine an equation that can predict MPG. Why do you think all the independent variables are not significant?

6. Determine for your regression predicting computer sales whether the residuals exhibit non-normality or heteroscedasticity.
7. The file `Oreos.xlsx` gives daily sales of Oreos at a supermarket and whether Oreos were placed 7" from the floor, 6" from the floor, or 5" from the floor. How does shelf position influence Oreo sales?
8. The file `USmacrodata.xlsx` contains U.S. quarterly GNP, Inflation rates, and Unemployment rates. Use this file to perform the following exercises:

a. Develop a regression to predict quarterly GNP growth from the last four quarters of growth. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.
b. Develop a regression to predict quarterly inflation rate from the last four quarters of inflation. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.
c. Develop a regression to predict quarterly unemployment rate from the unemployment rates of the last four quarters. Check for non-normality of residuals, heteroscedasticity, autocorrelation, and multicollinearity.

9. Does our regression model for predicting auto sales exhibit autocorrelation, non-normality of errors, or heteroscedasticity?

12 Modeling Trend and Seasonality

Whether the marketing analyst works for a car manufacturer, airline, or consumer packaged goods company, she often must forecast sales of her company's product. Whatever the product, it is important to understand the trends (either upward or downward) and seasonal aspects of the product's sales. This chapter discusses how to determine the trends and seasonality of product sales. Using monthly data on U.S. air passenger miles (2003–2012) you will learn how to do the following:

- Use moving averages to eliminate seasonality to easily see trends in sales.
- Use the Solver to develop an additive or multiplicative model to estimate trends and seasonality.

Using Moving Averages to Smooth Data and Eliminate Seasonality

Moving averages smooth out noise in the data. For instance, suppose you work for Amazon.com and you are wondering whether sales are trending upward. For each January sales are less than the previous month (December sales are always high because of Christmas), so the unsuspecting marketing analyst might think there is a downward trend in sales during January because sales have dropped. This conclusion is incorrect, though, because it ignores the fact that seasonal influences tend to drop January sales below December sales. You can use moving averages to smooth out seasonal data and better understand the trend and seasonality characteristics of your data.

NOTE All work in this chapter uses the file `airlinemiles.xlsx`, which contains monthly airlines miles (in thousands) traveled in the United States during the period from January 2003 through April 2012. A sample of this data is shown in Figure 12-1.

	D	E	F
8	MonthNumber	Month	AirlineMiles (000'S)
9	1	Jan2003	32,854,790.00
10	2	Feb2003	30,814,269.00
11	3	Mar2003	37,586,654.00
12	4	Apr2003	35,226,398.00
13	5	May2003	36,569,670.00
14	6	Jun2003	39,750,216.00
15	7	Jul2003	43,367,508.00
16	8	Aug2003	42,092,669.00
17	9	Sep2003	32,549,732.00
18	10	Oct2003	36,442,428.00
19	11	Nov2003	34,350,366.00
20	12	Dec2003	37,389,382.00
21	13	Jan2004	33,537,392.00
22	14	Feb2004	33,909,139.00
23	15	Mar2004	40,805,211.00
24	16	Apr2004	40,172,829.00
25	17	May2004	39,671,007.00
26	18	Jun2004	43,652,277.00
27	19	Jul2004	46,262,249.00
28	20	Aug2004	44,701,691.00
29	21	Sep2004	35,470,844.00
30	22	Oct2004	39,627,851.00
31	23	Nov2004	37,567,116.00
32	24	Dec2004	39,117,678.00

Figure 12-1: US airline miles

To further illustrate the concept of moving averages, take a look at the graph of United States airline miles shown in Figure 12-2. To obtain this graph select the data from the `Moving average` worksheet of the `airlinemiles.xlsx` file in the range E8:F120 and select Insert ➢ Charts ➢ Scatter and choose the second option (Scatter with Smooth Lines and Markers). You obtain the graph shown in Figure 12-2.

Due to seasonality (primarily because people travel more in the summer), miles traveled usually increase during the summer and then decrease during the winter. This makes it difficult to ascertain the trend in airline travel. Graphing the *moving average* of airline miles can help to better understand the trend in this data. A 12-month moving average, for example, graphs the average of the current month's miles and the last 11 months. Because moving averages smooth out noise in the data, you can use a 12-month moving average to eliminate the influence of seasonality. This is because a 12-month moving average includes one data point for each month. When analyzing a trend in quarterly data, you should plot a four-quarter moving average.

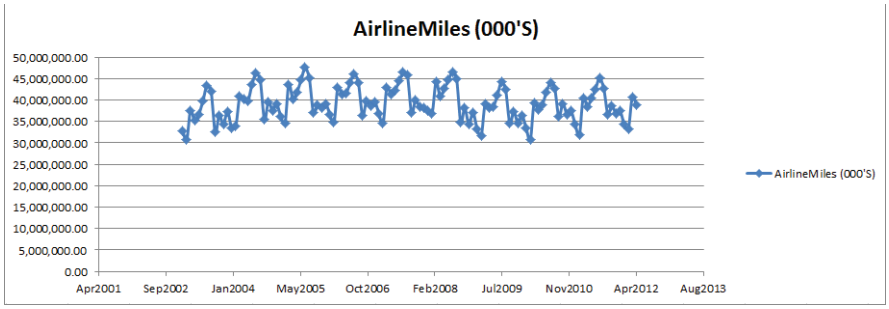

Figure 12-2: Graph of US airline miles

To overlay a 12-month moving average on the scatterplot, you return to an old friend, the Excel Trendline. Right-click the data series and select Add Trendline... Choose Moving Average and select 12 periods. Then you can obtain the trendline, as shown in Figure 12-3.

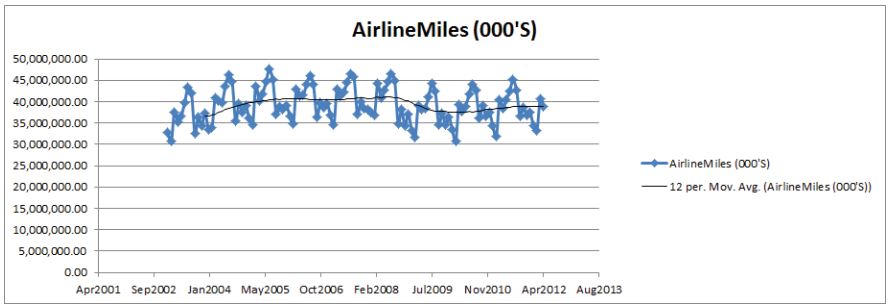

Figure 12-3: Moving average trendline

The moving average trendline makes it easy to see how airline travel trended between 2003 and 2012. You can now see the following:

- In 2003 and 2004 there was a sharp upward trend in airline travel (perhaps a rebound from 9/11).
- In 2005–2008 airline travel appeared to stagnate.
- In late 2008 there was a sharp drop in airline travel, likely due to the financial crisis.
- In 2010 a slight upward trend in air travel occurred.

The next section uses the Excel Solver to quantify the exact nature of the trend in airline miles and also to learn how to determine how seasonality influences demand for air travel.

An Additive Model with Trends and Seasonality

Based on the previous section's discussion it should be clear that to accurately forecast sales when the data has seasonality and trends, you need to identify and separate these from the data series. In this section you learn how this process can be modeled using Excel's Solver. These analyses enable you to identify and separate between the baseline, seasonality, and trend components of a data series.

When predicting product sales, the following additive model is often used to estimate the trend and seasonal influence of sales:

(1) Predicted Period *t* Sales = Base + Trend*Period Number + Seasonal Index for Month *t*

In Equation 1 you need to estimate the base, trend, and seasonal index for each month of the year. The work for this appears in the `Additive trend` worksheet (see Figure 12-4). To simplify matters the data is rescaled in billions of miles. The base, trend, and seasonal index may be described as follows:

- Base: The base is the best estimate of the level (without seasonality) of monthly airline miles at the beginning of the observed time period.
- Trend: The trend is the best estimate of the monthly rate of increase in airline miles traveled. A trend of 5, for example, would mean that the level of airline travel is increasing at a rate of 5 billion miles per month.
- Seasonal Index: Each month of the year has a seasonal index to reflect if travel during the month tends to be higher or lower than average. A seasonal index of +5 for June would mean, for example, that June airline travel tends to be 5 billion miles higher than an average month.

NOTE The seasonal indices must average to 0.

	A	B	C	D	E	F	G	H	I	J	K
1											
2	baseadd	37.37856									
3	trendadd	0.059026									
4										stddeverr	0.386323
5	1	-4.45733								RSQ	0.988934
6	2	-6.62334								SSE	4.9251
7	3	1.601041									
8	4	-0.319		MonthNumber	Month	Month	Airline Miles (billion	Forecast	Error	Sq Error	
9	5	1.274636		1	7/1/2009	7	44.22	43.7288	0.49	0.236896	
10	6	3.795057		2	8/1/2009	8	42.40	41.95583	0.44	0.194662	
11	7	6.291206		3	9/1/2009	9	34.68	35.12698	-0.45	0.203932	
12	8	4.459215		4	10/1/2009	10	37.32	37.69339	-0.38	0.140881	
13	9	-2.42866		5	11/1/2009	11	34.58	35.31697	-0.74	0.54817	
14	10	0.078726		6	12/1/2009	12	36.46	36.41789	0.04	0.001696	
15	11	-2.35673		7	1/1/2010	1	33.49	33.33441	0.15	0.023327	
16	12	-1.31482		8	2/1/2010	2	30.72	31.22743	-0.51	0.259421	
17				9	3/1/2010	3	39.37	39.51084	-0.14	0.019948	
18	mean	0		10	4/1/2010	4	37.76	37.64982	0.11	0.012653	
19				11	5/1/2010	5	38.88	39.30248	-0.42	0.175395	
20				12	6/1/2010	6	41.90	41.88193	0.02	0.000401	
21				13	7/1/2010	7	44.02	44.43711	-0.42	0.172428	
22				14	8/1/2010	8	42.81	42.66414	0.15	0.02222	
23				15	9/1/2010	9	36.13	35.8353	0.30	0.087799	
24				16	10/1/2010	10	39.18	38.4017	0.78	0.611145	
25				17	11/1/2010	11	36.67	36.02528	0.65	0.41766	

Figure 12-4: Additive trend model

To estimate base, trend, and seasonal indices, you need to create formulas based on trial values of the parameters in Column H. Then in Column I, you will determine the error for each month's forecast, and in Column J, you compute the squared error for each forecast. Finally, you use the Solver to determine the parameter values that minimize squared errors. To execute this estimation process, perform the following steps:

1. Enter trial values of the base and trend in cells B2 and B3. Name cell B2 `baseadd` and cell B3 `trend`.
2. Enter trial seasonal indices in the range B5:B16.
3. In cell B18, average the seasonal indices with the formula `=AVERAGE(B5:B16)`. The Solver model can set this average to 0 to ensure the seasonal indices average to 0.
4. Copy the formula `=baseadd+trend*D9+VLOOKUP(F9,$A$5:$B$16,2)` from H9 to H10:H42 to compute the forecast for each month.
5. Copy the formula `=G9-H9` from I9 to I10:I42 to compute each month's forecast error.
6. Copy the formula `=(I9^2)` from J9 to J10:J42 to compute each month's squared error.
7. In cell K6, compute the Sum of Squared Errors (SSE) using the formula `=SUM(J9:J42)`.

8. Now set up the Solver model, as shown in Figure 12-5. Change the parameters to minimize SSE and constrain the average of the seasonal indices to 0. Do not check the non-negative box because some seasonal indices must be negative. The forecasting model of Equation 1 is a *linear forecasting model* because each unknown parameter is multiplied by a constant. When the forecasts are created by adding together terms that multiply changing cells by constants, the GRG Solver Engine always finds a unique solution to the least square minimizing parameter estimates for a forecasting model.

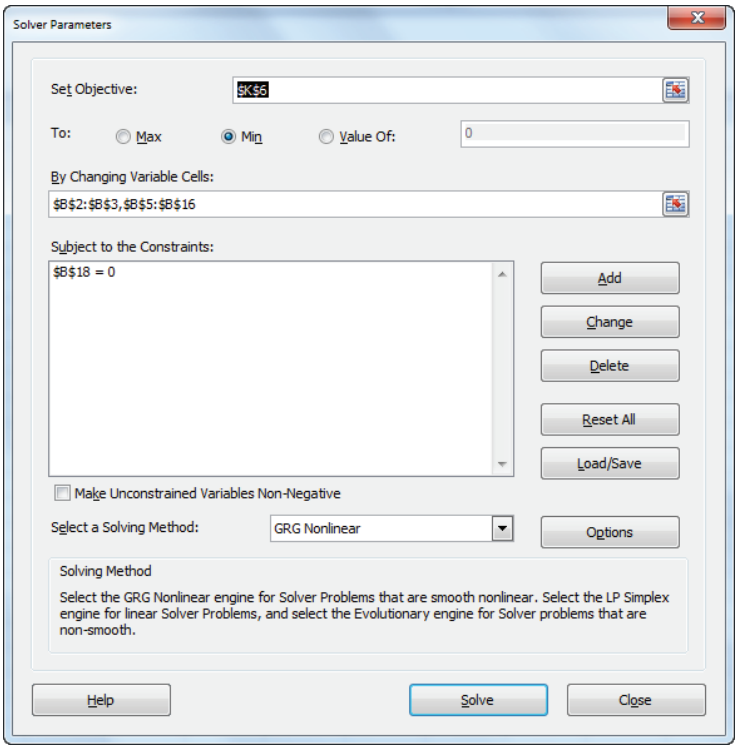

Figure 12-5: Additive trend Solver model

Refer to the data shown in Figure 12-4 and you can make the following estimates:

- At the beginning of July 2009, the base level of airline miles is 37.38 billion.
- An upward trend in airline miles is 59 billion miles per month.
- The busiest month is July (6.29 billion miles above average) and the slowest month is February with 6.62 billion miles below average.

Cell K5 uses the formula `=RSQ(G9:G42,H9:H42)` to show that the model explains 98.9 percent of the variation in miles traveled. Cell K4 also computes the standard deviation of the errors (989 billion) with the formula `=STDEV(I9:I42)`. You should expect 95 percent of the predictions to be accurate within 2 * 0.386 = 0.772 billion miles. Looking at Column I, no outliers are found.

A Multiplicative Model with Trend and Seasonality

When predicting product sales, the following multiplicative model is often used to estimate the trend and seasonal influence of sales:

(2) Predicted Period t Sales = Base * (Trend^t) * (Seasonal Index for Month t)

As in the additive model, you need to estimate the base, trend, and seasonal indices. In Equation 2 the trend and seasonal index have different meanings than in the additive model.

- Trend: The trend now represents the percentage monthly increase in the level of airline miles. For example, a trend value of 1.03 means monthly air travel is increasing 3 percent per month, and a trend value of .95 means monthly air travel is decreasing at a rate of 5 percent per month. If per period growth is independent of the current sales value, the additive trend model will probably outperform the multiplicative trend model. On the other hand, if per period growth is an increasing function of current sales, the multiplicative trend model will probably outperform the additive trend model.
- Seasonal Index: The seasonal index for a month now represents the percentage by which airline travel for the month is above or below an average month. For example, a seasonal index for July of 1.16 means July has 16 percent more air travel than an average month, whereas a seasonal index for February of .83 means February has 17 percent less air travel than an average month. Of course, multiplicative seasonal indices must average to 1. This is because months with above average sales are indicated by a seasonal index exceeding 1, while months with below average sales are indicated by a seasonal index less than 1.

The work for this equation appears in the `Multiplicative trend` worksheet. All the formulas are the same as the additive model with the exception of the monthly forecasts in Column H. You can implement Equation 2 by copying the formula `=base*(trend^D9)*VLOOKUP(F9,$A$5:$B$16,2)` from H9 to H10:H42.

The forecasting model in Equation 2 is a nonlinear forecasting model because you can raise the trend to a power and multiply, rather than add terms involving the seasonal indices. For nonlinear forecasting models, the GRG Solver Engine often fails to find an optimal solution unless the starting values for the changing cells are close to the optimal solution. The remedy to this issue is as follows:

1. In Solver select Options, and from the GRG tab, select Multistart. This ensures the Solver will try many (between 50 and 200) starting solutions and find the optimal solution from each starting solution. Then the Solver reports the "best of the best" solutions.
2. To use the Multistart option, input lower and upper bounds on the changing cells. To speed up solutions, these bounds should approximate sensible values for the estimated parameters. For example, a seasonal index will probably be between 0 and 3, so an upper bound of 100 would be unreasonable. As shown in Figure 12-6, you can choose an upper bound of 3 for each seasonal index and an upper bound of 2 for the trend. For this example, choose an upper bound of 100 for the base.

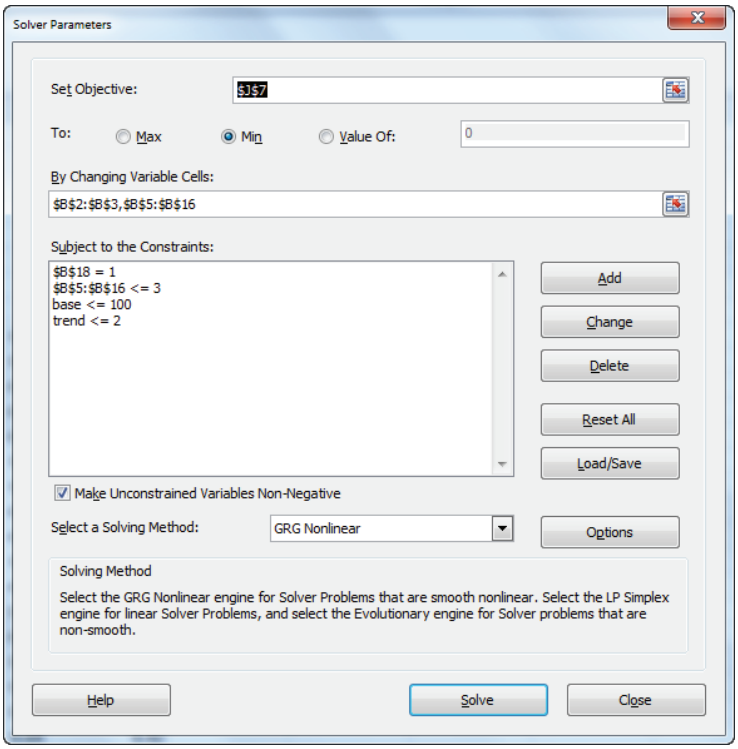

Figure 12-6: Solver window for multiplicative trend model

3. Cell B18 averages the seasonal indices, so in the Solver window add the constraint $B$18 =1 to ensure that the seasonal indices average to 1.
4. Select Solve, and the Solver will then find the optimal solution (refer to Figure 12-7).

	A	B	C	D	E	F	G	H	I	J	K
1											
2	base	3.74E+01									
3	trend	1.001493569					1.00E+06				
4										stddeverrors	0.411756002
5	1	0.884049011								RSQ	0.987429707
6	2	0.82837254								SSE	
7	3	1.041400111								5.59	
8	4	0.991684904		MonthNumber	Month	Month	AirlineMiles (b	Forecast	Error	Sq Error	
9	5	1.03315296		1	7/1/2009	7	44.22	43.63945105	0.58	0.33	
10	6	1.098599337		2	8/1/2009	8	42.40	41.89570886	0.50	0.25	
11	7	1.164327334		3	9/1/2009	9	34.68	35.19980697	-0.52	0.28	
12	8	1.116136195		4	10/1/2009	10	37.32	37.71609694	-0.40	0.16	
13	9	0.936353344		5	11/1/2009	11	34.58	35.38768125	-0.81	0.66	
14	10	1.00179316		6	12/1/2009	12	36.46	36.46161088	0.00	0.00	
15	11	0.938545346		7	1/1/2010	1	33.49	33.43255124	0.05	0.00	
16	12	0.965585734		8	2/1/2010	2	30.72	31.37379361	-0.66	0.43	
17				9	3/1/2010	3	39.37	39.50091231	-0.13	0.02	
18	mean	0.999999998		10	4/1/2010	4	37.76	37.67136647	0.09	0.01	
19				11	5/1/2010	5	38.88	39.30524076	-0.42	0.18	
20				12	6/1/2010	6	41.90	41.85750467	0.04	0.00	
21				13	7/1/2010	7	44.02	44.4280508	-0.41	0.16	
22				14	8/1/2010	8	42.81	42.65279779	0.16	0.03	
23				15	9/1/2010	9	36.13	35.8358956	0.30	0.09	
24				16	10/1/2010	10	39.18	38.39765695	0.79	0.62	
25				17	11/1/2010	11	36.67	36.02716493	0.64	0.42	

Figure 12-7: Multiplicative trend model

NOTE If the Solver assigns a changing cell, a value near its lower or upper bound should be relaxed. For example, if you set the upper bound for the base to 30, the Solver will find a value near 30, thereby indicating the bound should be relaxed.

From the optimal Solver solution you find the following:

- The estimated base level of airline miles is 37.4 billion.
- You can estimate airline miles increase at a rate of 0.15 percent per month or $1.00149^{12} - 1 = 1.8$ percent per year.
- The busiest month for the airlines is July, when miles traveled are 16 percent above average, and the least busy month is February, during which miles traveled are 17 percent below average.

A natural question is whether the additive or multiplicative model should be used to predict airline miles for future months. Because the additive model has a lower standard deviation of residuals, you should use the additive model to forecast future airline miles traveled.

Summary

In this chapter you learned the following:

- Using a 12-month or 4-quarter moving average chart enables you to easily see the trend in a product's sales.
- You can often use seasonality and trend to predict sales by using the following equation:

$$\text{Predicted Period } t \text{ Sales} = \text{Base} + \text{Trend} * \text{Period Number} + \text{Seasonal Index for Month } t$$

- You can often use the following equation to predict sales of a product:

$$\text{Predicted period } t \text{ Sales} = \text{Base} * (\text{Trend}^t) * (\text{Seasonal Index for Month } t)$$

Exercises

The following exercises use the file `airlinedata.xlsx`, which contains monthly U.S. domestic air miles traveled during the years 1970–2004.

1. Determine the trend and seasonality for the years 1970–1980.
2. Determine the trend and seasonality for the years 1981–1990.
3. Determine the trend and seasonality for the years 1995–2004.

13 Ratio to Moving Average Forecasting Method

In Chapter 12, "Modeling Trend and Seasonality," you learned how to estimate trend and seasonal indices. Naturally you would like to use your knowledge of trend and seasonality to make accurate forecasts of future sales. The *Ratio to Moving Average Method* provides an accurate, easy-to-use forecasting method for future monthly or quarterly sales. This chapter shows how to use this method to easily estimate seasonal indices and forecast future sales.

Using the Ratio to Moving Average Method

The simple Ratio to Moving Average Forecasting Method is described in this section via examples using data from the `Ratioma.xlsx` file, which includes sales of a product during 20 quarters (as shown in Figure 13-1 in rows 5 through 24). This technique enables you to perform two tasks:

- Easily estimate a time series' trend and seasonal indices.
- Generate forecasts of future values of the time series.

Using the first 20 quarters for the data exemplified in this chapter, you will be able to forecast sales for the following four quarters (Quarters 21 through 24). Similar to the one in Chapter 12, this time series data has both trend and seasonality.

The Ratio to Moving Average Method has four main steps:

- Estimate the deseasonalized level of the series during each period (using centered moving averages).
- Fit a trend line to your deseasonalized estimates (in Column G).
- Determine the seasonal index for each quarter and estimate the future level of the series by extrapolating the trend line.
- Predict future sales by reseasonalizing the trend line estimate.

	B	C	D	E	F	G	H	I	J	K	L
1				slope	6.9387868						
2				intercept	30.166176				quarter	seasonal index	normalized
3									1	0.818547	0.81373678
4	Quarter#	Year	Quarter	Sales	4 period MA	Centered MA	Actual/CMA	Forecast	2	0.93934	0.9338196
5	1	1	1	24					3	1.067364	1.06109143
6	2	1	2	44	52				4	1.198394	1.19135219
7	3	1	3	61	58	55.00	1.11				
8	4	1	4	79	63.5	60.75	1.30				
9	5	2	1	48	71	67.25	0.71				
10	6	2	2	66	77.5	74.25	0.89				
11	7	2	3	91	82.5	80.00	1.14				
12	8	2	4	105	87.25	84.88	1.24				
13	9	3	1	68	89.5	88.38	0.77				
14	10	3	2	85	94.5	92.00	0.92				
15	11	3	3	100	104.25	99.38	1.01				
16	12	3	4	125	114.25	109.25	1.14				
17	13	4	1	107	123.75	119.00	0.90				
18	14	4	2	125	132.25	128.00	0.98				
19	15	4	3	138	139.25	135.75	1.02				
20	16	4	4	159	146.75	143.00	1.11				
21	17	5	1	135	156	151.38	0.89				
22	18	5	2	155	164.25	160.13	0.97				
23	19	5	3	175							
24	20	5	4	192							
25	21	6	1			175.880699		143.121			
26	22	6	2			182.819485		170.72			
27	23	6	3			189.758272		201.351			
28	24	6	4			196.697059		234.335			

Figure 13-1: Example of Ratio to Moving Average Method

The following sections walk you through each main part of this process.

Calculating Moving Averages and Centered Moving Averages

To begin, you compute a four-quarter (four quarters eliminates seasonality) moving average for each quarter by averaging the prior quarter, current quarter, and next two quarters. To do this you copy the formula `=AVERAGE(E5:E8)` down from cell F6 to F7:F22. For example, for Quarter 2, the moving average is (24 + 44 + 61 + 79) / 4 = 52.

Because the moving average for Quarter 2 averages Quarters 1 through 4 and the numbers 1–4 average to 2.5, the moving average for Quarter 2 is centered at Quarter 2.5. Similarly, the moving average for Quarter 3 is centered at Quarter 3.5. Therefore, averaging these two moving averages gives a centered moving average that estimates the level of the process at the end of Quarter 3. To estimate the level of the series during each series (without seasonality), copy the formula `=AVERAGE(F6:F7)` down from cell G7.

Fitting a Trend Line to the Centered Moving Averages

You can use the centered moving averages to fit a trend line that can be used to estimate the future level of the series. To do so, follow these steps:

1. In cell F1 use the formula `=SLOPE(G7:G22,B7:B22)` to find the slope of the trend line.
2. In cell F2 use the formula `=INTERCEPT(G7:G22,B7:B22)` to find the intercept of the trend line.
3. Estimate the level of the series during Quarter *t* to be 6.94*t* *c* + 30.17.
4. Copy the formula `=intercept + slope*B25` down from cell G25 to G26:G28 to compute the estimated level (excluding seasonality) of the series from Quarter 21 onward.

Compute the Seasonal Indexes

Recall that a seasonal index of 2 for a quarter means sales in that quarter are twice the sales during an average quarter, whereas a seasonal index of .5 for a quarter would mean that sales during that quarter were one-half of an average quarter. Therefore, to determine the seasonal indices, begin by determining for each quarter for which you have sales *(Actual Sales) / Centered Moving Average*. To do this, copy the formula `=E7/G7` down from cell H7 to H8:H22. You find, for example, that during Quarter 1 sales were 77 percent, 71 percent, 90 percent and 89 percent of average, so you could estimate the seasonal index for Quarter 1 as the average of these four numbers (82 percent). To calculate the initial seasonal index estimates, you can copy the formula `=AVERAGEIF($D$7:$D$22,J3,$H$7:$H$22)` from cell K3 to K4:K6. This formula averages the four estimates you have for Q1 seasonality.

Unfortunately, the seasonal indices do not average exactly to 1. To ensure that your final seasonal indices average to 1, copy the formula `=K3/AVERAGE($K$3:$K$6)` from cell L3 to L4:L6.

Forecasting Sales during Quarters 21–24

To create your sales forecast for each future quarter, simply multiply the trend line estimate for the quarter's level (from Column G) by the appropriate seasonal index. Copy the formula `=VLOOKUP(D25,season,3)*G25` from cell G25 to G26:G28 to compute the final forecast for Quarters 21–24. This forecast includes estimates of trend and seasonality.

If you think the trend of the series has changed recently, you can estimate the series' trend based on more recent data. For example, you could use the centered moving averages for Quarters 13–18 to get a more recent trend estimate by using the formula `=SLOPE(G17:G22,B17:B22)`. This yields an estimated trend of 8.09 units per quarter. If you want to forecast Quarter 22 sales, for example, you take the last centered moving average you have (from Quarter 18) of 160.13 and add 4 (8.09) to estimate the level of the series in Quarter 22. Then multiply the estimate of the Quarter 22 level by the Quarter 2 seasonal index of .933 to yield a final forecast for Quarter 22 sales of (160.13 + 4(8.09)) * (.933) = 179.6 units.

Applying the Ratio to Moving Average Method to Monthly Data

Often the Ratio to Moving Average Method is used to forecast monthly sales as well as quarterly sales. To illustrate the application of this method to monthly data, let's look at U.S. housing starts.

The `Housingstarts.xlsx` file gives monthly U.S. housing starts (in thousands) for the period January 2000 through May 2011. Based on the data through November 2010, you can apply the Ratio to Moving Average Method to forecast monthly U.S. housing starts for the period December 2010 through May 2011. You can forecast a total of 3.5 million housing starts, and in reality there were 3.374 million housing starts. The key difference between applying the method to monthly and quarterly data is that for monthly data you need to use 12-month moving averages to eliminate seasonality.

Summary

In this chapter you learned the following:

- Applying the Ratio to Moving Average Method involves the following tasks:
  - Compute four-quarter moving averages and then determine the centered moving averages.
  - Fit a trend line to the centered moving averages.
  - Compute seasonal indices.
  - Compute forecasts for future periods.
- You can apply the Ratio to Moving Average Method to monthly data as well by following the same process but use 12-month moving averages to eliminate seasonality.

Exercises

1. The file `Walmartdata.xls` contains quarterly revenues of Wal-Mart during the years 1994–2009. Use the Ratio to Moving Average Method to forecast revenues for Quarters 3 and 4 in 2009 and Quarters 1 and 2 in 2010. Use Quarters 53–60 to create a trend estimate that you use in your forecasts.
2. Based on the data in the file `airlinemiles.xlsx` from Chapter 12, use the Ratio to Moving Average Method to forecast airline miles for the remaining months in 2012.

14 Winter's Method

Predicting future values of a time series is usually difficult because the characteristics of any time series are constantly changing. For instance, as you saw in Chapter 12, "Modeling Trend and Seasonality," the trend in U.S. airline passenger miles changed several times during the 2000–2012 period. *Smoothing* or *adaptive* methods are usually best suited for forecasting future values of a time series. Essentially, smoothing methods create forecasts by combining information from a current observation with your prior view of a parameter, such as trend or a seasonal index. Unlike many other smoothing methods, Winter's Method incorporates both trend and seasonal factors. This makes it useful in situations where trend and seasonality are important. Because in an actual situation (think U.S. monthly housing starts) trend and seasonality are constantly changing, a method such as Winter's Method that changes trend and seasonal index estimates during each period has a better chance of keeping up with changes than methods like the trend and seasonality approaches based on curve fitting discussed in Chapter 12, which use constant estimates of trend and seasonal indices.

To help you understand how Winter's Method works, this chapter uses it to forecast airline passenger miles for April through December 2012 based on the data studied in Chapter 12. This chapter describes the three key characteristics of a time series (level, trend, and seasonality) and explains the initialization process, notation, and key formulas needed to implement Winter's Method. Finally, you explore forecasting with Winter's Method and the concept of Mean Absolute Percentage Error (MAPE).

Parameter Definitions for Winter's Method

In this chapter you will develop Winter's exponential smoothing method using the three time series characteristics, level (also called base), trend, and seasonal index,

discussed in Chapter 12 in the "Multiplicative Model with Trend and Seasonality" section. After observing data through the end of month t you can estimate the following quantities of interest:

- L_t = Level of series
- T_t = Trend of series
- S_t = Seasonal index for current month

The key to Winter's Method is the use of the following three equations, which are used to update L_t, T_t, and S_t. In the following equations, *alp*, *bet*, and *gam* are called *smoothing parameters*. The values of these parameters will be chosen to optimize your forecasts. In the following equations, c equals the number of periods in a seasonal cycle (c = 12 months for example) and x_t equals the observed value of the time series at time t.

(1) $L_t = alp(x_t) / (s_{t-c}) + (1 - alp)(L_{t-1} * T_{t-1})$

(2) $T_t = bet(L_t / L_{t-1}) + (1 - bet)\ T_{t-1}$

(3) $S_t = gam(x_t / L_t) + (1 - gam)s_{(t-c)}$

Equation 1 indicates that the new base estimate is a weighted average of the current observation (deseasonalized) and last period's base is updated by the last trend estimate. Equation 2 indicates that the new trend estimate is a weighted average of the ratio of the current base to last period's base (this is a current estimate of trend) and last period's trend. Equation 3 indicates that you update the seasonal index estimate as a weighted average of the estimate of the seasonal index based on the current period and the previous estimate. In equations 1–3 the first term uses an estimate of the desired quantity based on the current observation and the second term uses a past estimate of the desired quantity.

NOTE Note that larger values of the smoothing parameters correspond to putting more weight on the current observation.

You can define $F_{t,k}$ as your forecast (F) after period t for the period $t + k$. This results in the following equation:

(4) $F_{t,k} = L_t*(T_t)^k s_{t+k-c}$

Equation 4 first uses the current trend estimate to update the base k periods forward. Then the resulting base estimate for period $t + k$ is adjusted by the appropriate seasonal index.

Initializing Winter's Method

To start Winter's Method, you must have initial estimates for the series base, trend, and seasonal indices. You can use the data from the `airline winters.xls` file, which contains monthly U.S. airline passenger miles for the years 2003 and 2004 to obtain initial estimates of level, trend, and seasonality. See Figure 14-1.

	A	B	C	D	E	F	G	H	I
1									
2	base	3.51E+01							SSE
3	trend	1.006491							5.63
4			MonthNu	Month	Month	AirlineMiles (billions)	Forecast	Error	Sq Error
5	1	0.90305	1	Jan2003	1	32.85	31.87289	0.98	0.96
6	2	0.875947	2	Feb2003	2	30.81	31.11699	-0.30	0.09
7	3	1.053815	3	Mar2003	3	37.59	37.6785	-0.09	0.01
8	4	1.008006	4	Apr2003	4	35.23	36.27456	-1.05	1.10
9	5	1.011706	5	May2003	5	36.57	36.644	-0.07	0.01
10	6	1.099865	6	Jun2003	6	39.75	40.09572	-0.35	0.12
11	7	1.173706	7	Jul2003	7	43.37	43.06529	0.30	0.09
12	8	1.129149	8	Aug2003	8	42.09	41.69932	0.39	0.15
13	9	0.879644	9	Sep2003	9	32.55	32.696	-0.15	0.02
14	10	0.977359	10	Oct2003	10	36.44	36.5638	-0.12	0.01
15	11	0.918146	11	Nov2003	11	34.35	34.57155	-0.22	0.05
16	12	0.969609	12	Dec2003	12	37.39	36.74627	0.64	0.41
17			13	Jan2004	1	33.54	34.44595	-0.91	0.83
18	mean	1	14	Feb2004	2	33.91	33.62903	0.28	0.08
19			15	Mar2004	3	40.81	40.72023	0.08	0.01
20			16	Apr2004	4	40.17	39.20296	0.97	0.94
21			17	May2004	5	39.67	39.60222	0.07	0.00
22			18	Jun2004	6	43.65	43.33259	0.32	0.10
23			19	Jul2004	7	46.26	46.54189	-0.28	0.08
24			20	Aug2004	8	44.70	45.06565	-0.36	0.13
25			21	Sep2004	9	35.47	35.3355	0.14	0.02
26			22	Oct2004	10	39.63	39.51555	0.11	0.01
27			23	Nov2004	11	37.57	37.36247	0.20	0.04
28			24	Dec2004	12	39.12	39.71275	-0.60	0.35

Figure 14-1: Data for Winter's Method

In the `Initial` worksheet you can fit the Multiplicative Trend Model from Chapter 12 to the 2003–2004 data. As shown in Figure 14-2, you use the trend and seasonal index from this fit as the original seasonal index and the December 2004 trend. Cell C25 determines an estimate of the base for December 2004 by deseasonalizing the observed December 2004 miles. This is accomplished with the formula `=(B25/H25)`.

	A	B	C	D	E	F	G	H	I	J	K
1	DATE	Airline Miles(billions)									
2	Jan2003	32.85									
3	Feb2003	30.81									
4	Mar2003	37.59									
5	Apr2003	35.23									
6	May2003	36.57									
7	Jun2003	39.75									
8	Jul2003	43.37									
9	Aug2003	42.09									
10	Sep2003	32.55						alp	bet	gam	
11	Oct2003	36.44						0.548512014	0.049142	0.58877279	
12	Nov2003	34.35									
13	Dec2003	37.39									
14	Jan2004	33.54						0.903049602			
15	Feb2004	33.91						0.875947455			
16	Mar2004	40.81						1.053814727			
17	Apr2004	40.17						1.00800602			
18	May2004	39.67						1.011705575			
19	Jun2004	43.65						1.099865309			
20	Jul2004	46.26						1.173705527			
21	Aug2004	44.70				SSE	77.8196	1.129148564			
22	Sep2004	35.47				stdeverro	0.9369659	0.879643964		34 sign changes of 87	
23	Oct2004	39.63						0.977358616		MAPE	
24	Nov2004	37.57	Base	Trend	Forecast	Error	Sq Error	0.918146041	34	0.02062096	
25	Dec2004	39.12	40.34378	1.006491				0.969608616	Sign char	APE	
26	Jan2005	36.12	40.27083	1.006083	36.6689	-0.55	0.3038	0.899411078		0.01526167	
27	Feb2005	34.56	39.93414	1.005373	35.4897	-0.93	0.8628	0.869764432	0	0.02687615	
28	Mar2005	43.64	40.8425	1.006227	42.3093	1.33	1.7767	1.062490091	1	0.03054212	
29	Apr2005	40.24	40.45404	1.005453	41.42583	-1.18	1.3953	1.000244124	1	0.02935135	
30	May2005	41.80	41.02748	1.005882	41.15077	0.65	0.4235	1.015922148	1	0.01556855	
31	Jun2005	44.68	40.91303	1.005456	45.39013	-0.71	0.5089	1.095230183	1	0.0159679	
32	Jul2005	47.56	40.80036	1.005052	48.28184	-0.72	0.5166	1.169022868	0	0.01511097	

Figure 14-2: Initialization of Winter's Method

The next part of Winter's Method includes choosing the smoothing parameters to optimize the one-month-ahead forecasts for the years 2005 through 2012.

Estimating the Smoothing Constants

After observing each month's airline miles (in billions), you are now ready to update the smoothing constants. In Column C, you will update the series base; in Column D, the series trend; and in Column H, the seasonal indices. In Column E, you compute the forecast for next month, and in Column G, you compute the squared error for each month. Finally, you'll use Solver to choose smoothing constant values that minimize the sum of the squared errors. To enact this process, perform the following steps:

1. In H11:J11, enter trial values (between 0 and 1) for the smoothing constants.
2. In C26:C113, compute the updated series level with Equation 1 by copying the formula `=alp*(B26/H14)+(1-alp)*(C25*D25)` from cell C26 to C27:C113.
3. In D26:D113, use Equation 2 to update the series trend. Copy the formula `=bet*(C26/C25)+(1-bet)*D25` cell from D26 to D27:D113.

4. In H26:H113, use Equation 3 to update the seasonal indices. Copy the formula `=gam*(B26/C26)+(1-gam)*H14` from cell H26 to H27:H113.
5. In E26:E113, use Equation 4 to compute the forecast for the current month by copying the formula `=(C25*D25)*H14` from cell E26 to E27:E113.
6. In F26:F113 compute each month's error by copying the formula `=(B26-E26)` from cell E26 to E27:E113.
7. In G26:G113, compute the squared error for each month by copying the formula `=F26^2` from cell F26 to F27:F113. In cell G21 compute the Sum of Squared Errors (SSE) using the formula `=SUM(G26:G113)`.
8. Now use the Solver to determine smoothing parameter values that minimize SSE. The Solver Parameters dialog box is shown in Figure 14-3.

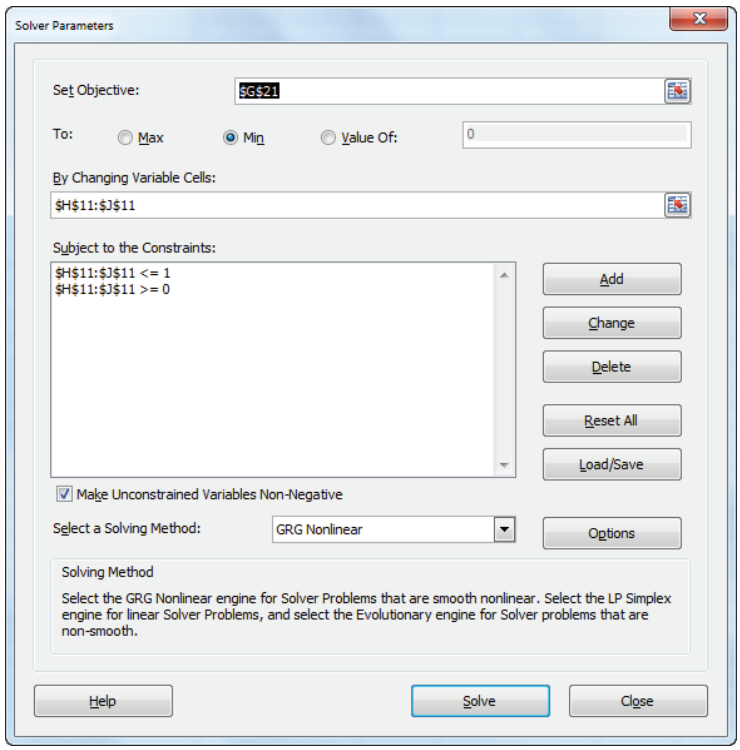

Figure 14-3: Solver Window for optimizing smoothing constants

9. Choose the smoothing parameters (H11:J11) to minimize SSE (cell G21). The Excel Solver ensures you can find the best combination of smoothing constants. Smoothing constants must be α. The Solver finds that alp = 0.55, bet = 0.05, and gamma = 0.59.

Forecasting Future Months

Now that you have estimated the Winter's Method smoothing constants (α, β, γ, etc.), you are ready to use these estimates to forecast future airline miles. This can be accomplished using the formula in cell D116. Copying this formula down to cells D117:D123 enables you to forecast sales for the months of May through December of 2012. Figure 14-4 offers a visual summary of the forecasted sales.

	A	B	C	D	E	F	G	H
101	Apr2011	38.51	38.09331	0.999196	38.63321	-0.13	0.0162	1.011448481
102	May2011	40.43	38.7357	1.000064	39.16698	1.26	1.5942	1.03767775
103	Jun2011	42.57	38.84163	1.000195	42.36397	0.21	0.0425	1.095009094
104	Jul2011	45.07	39.00168	1.000388	44.75388	0.32	0.1025	1.154171797
105	Aug2011	42.78	38.66139	0.99994	43.50483	-0.72	0.5220	1.110060016
106	Sep2011	36.70	39.65425	1.001205	35.05387	1.65	2.7064	0.917771587
107	Oct2011	38.70	39.72586	1.001235	38.66143	0.04	0.0018	0.974072493
108	Nov2011	36.83	40.16433	1.001716	36.182	0.65	0.4171	0.913943328
109	Dec2011	37.49	39.82658	1.001219	38.1972	-0.70	0.4955	0.944695094
110	Jan2012	34.31	39.32558	1.000541	35.19792	-0.88	0.7821	0.876725748
111	Feb2012	33.26	39.53299	1.000773	32.97971	0.28	0.0809	0.840092107
112	Mar2012	40.78	38.94334	1.000002	41.9811	-1.20	1.4396	1.05291499
113	Apr2012	38.81	38.6274	0.999604	39.38927	-0.58	0.3396	1.00743814
114			Base	Trend	Forecast	Error	Sq Error	Seasonal Indices
115			forecasts					
116		1	May-12	40.06691				
117		2	Jun-12	42.26383				
118		3	Jul-12	44.52966				
119		4	Aug-12	42.81079				
120		5	Sep-12	35.38093				
121		6	Oct-12	37.53649				
122		7	Nov-12	35.20542				
123		8	Dec-12	36.37556				
124			total	314.1696				

Figure 14-4: Forecasting with Winter's Method

Figure 14-4 shows the forecasted sales for May through December 2012 by copying the formula `=($C$113*$D$113^B116)*H102` from cell D116 to D117:D123. Cell D124 adds up these forecasts and predicts the rest of 2012 to see 314.17 billion airline miles traveled.

Cell G22 computes the standard deviation (0.94 billion) of the one-month-ahead forecast errors. This implies that approximately 95 percent of the forecast errors should be at most 1.88 billion. From Column F you see none of the one-month-ahead forecasts are outliers.

Mean Absolute Percentage Error (MAPE)

Statisticians like to estimate parameters for a forecast model by minimizing squared errors. In reality, however, most people are more interested in measuring forecast accuracy by looking at the Mean of Absolute Percentage Error (MAPE). This is probably because MAPE, unlike SSE, is measured in the same units as the data. Figure 14-5 shows that the one-month-ahead forecasts are off by an average of 2.1 percent. To compute the Absolute Percentage Error (APE) for each month, copy the formula `=ABS(B26-E26)/B26` from cell G26 to J26:J113. In cell J24 the formula `=AVERAGE(J26:J113)` computes the MAPE.

	C	D	E	F	G	H	I	J	K	L
22				stdeverrors	0.9369659	0.879643964		34 sign changes of 87		
23						0.977358616		MAPE		
24	Base	Trend	Forecast	Error	Sq Eerror	0.918146041	34	0.0206		
25	40.34378	1.006491				0.969608616	Sign change	APE		
26	40.27083	1.006083	36.6689	-0.55	0.3038	0.899411078		0.0153		
27	39.93414	1.005373	35.4897	-0.93	0.8628	0.869764432	0	0.0269		
28	40.8425	1.006227	42.3093	1.33	1.7767	1.062490091	1	0.0305		
29	40.45404	1.005453	41.42583	-1.18	1.3953	1.000244124	1	0.0294		
30	41.02748	1.005882	41.15077	0.65	0.4235	1.015922148	1	0.0156		
31	40.91303	1.005456	45.39013	-0.71	0.5089	1.095230183	1	0.016		
32	40.80036	1.005052	48.28184	-0.72	0.5166	1.169022868	0	0.0151		
33	40.43957	1.00437	46.30243	-1.17	1.3621	1.121476998	0	0.0259		
34	41.43753	1.005368	35.72786	1.32	1.7346	0.888092908	1	0.0356		
35	40.61218	1.004125	40.71671	-1.87	3.4855	0.965138567	1	0.0481		
36	41.20776	1.004643	37.44173	0.72	0.5134	0.922768159	1	0.0188		
37	40.85333	1.003992	40.14091	-0.96	0.9307	0.963331236	1	0.0246		
38	40.88625	1.003835	36.89062	-0.21	0.0456	0.898023391	0	0.0058		
39	40.44253	1.003114	35.6978	-0.95	0.9068	0.863505332	0	0.0274		
40	40.45961	1.002981	43.10358	-0.21	0.0445	1.061104844	0	0.0049		

Figure 14-5: Computation of MAPE

Winter's Method is an attractive forecasting method for several reasons:

- Given past data, the method can easily be programmed to provide quick forecasts for thousands of products.
- Winter's Method catches changes in trend or seasonality.
- Smoothing methods "adapt" to the data. That is, if you underforecast you raise parameter estimates and if you overforecast you lower parameter estimates.

Summary

In this chapter you learned the following:

- Exponential smoothing methods update time series parameters by computing a weighted average of the estimate of the parameter from the current observation with the prior estimate of the parameter.
- Winter's Method is an exponential smoothing method that updates the base, trend, and seasonal indices after each equation:

  (1) $L_t = alp(x_t) / (s_{t-c}) + (1-alp)(L_{t-1} * T_{t-1})$

  (2) $T_t = bet(L_t / L_{t-1}) + (1-bet)T_{t-1}$

  (3) $S_t = gam(x_t / L_t) + (1-gam)s_{(t-c)}$
- Forecasts for k periods ahead at the end of period t are made with Winter's Method using Equation 4:

  (4) $F_{t,k} = L_t * (T_t)^k s_{t+k-c}$

Exercises

All the data for the following exercises can be found in the file `Quarterly.xlsx`.

1. Use Winter's Method to forecast one-quarter-ahead revenues for Wal-Mart.
2. Use Winter's Method to forecast one-quarter-ahead revenues for Coca-Cola.
3. Use Winter's Method to forecast one-quarter-ahead revenues for Home Depot.
4. Use Winter's Method to forecast one-quarter-ahead revenues for Apple.
5. Use Winter's Method to forecast one-quarter-ahead revenues for Amazon.com.
6. Suppose at the end of 2007 you were predicting housing starts in Los Angeles for the years 2008 and 2009. Why do you think Winter's Method would provide better forecasts than multiple regression?