

Chapter 9

Regression Analysis

Ragsdale, Cliff T.
Spreadsheet Modeling & Decision Analysis, 5e
South-Western, Mason, OH, 2008:409-458
ISBN 978-0-324-65663-3

9.0 Introduction

Regression analysis is a modeling technique for analyzing the relationship between a *continuous* (real-valued) dependent variable Y and one or more independent variables X_1, X_2, \dots, X_k . The goal in regression analysis is to identify a function that describes, as closely as possible, the relationship between these variables so that we can predict what value the dependent variable will assume given specific values for the independent variables. This chapter shows how to estimate these functions and how to use them to make predictions in a business environment.

9.1 An Example

As a simple example of how regression analysis might be used, consider the relationship between sales for a company and the amount of money it spends on advertising. Few would question that the level of sales for a company will depend on or be influenced by advertising. Thus, we could view sales as the dependent variable Y and advertising as the independent variable X_1 . Although some relationship exists between sales and advertising, we might not know the exact functional form of this relationship. Indeed, there probably is not an exact functional relationship between these variables.

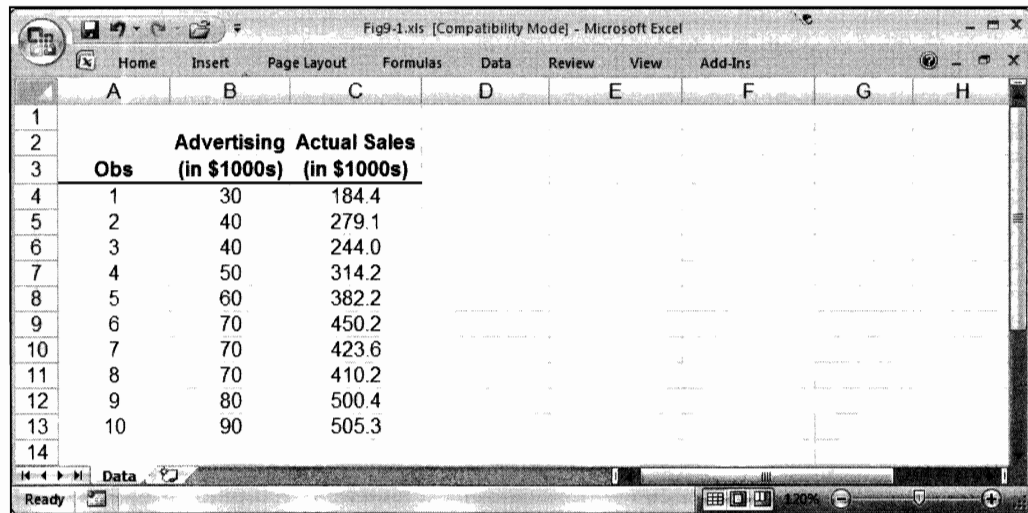
We expect that sales for a company depend to some degree on the amount of money the company spends on advertising. But many other factors also might affect a company's sales, such as general economic conditions, the level of competition in the marketplace, product quality, and so on. Nevertheless, we might be interested in studying the relationship between the dependent variable sales (Y) and the independent variable advertising (X_1) and predicting the average level of sales expected for a given level of advertising. Regression analysis provides the tool for making such predictions.

To identify a function that describes the relationship between advertising and sales for a company, we first need to collect sample data to analyze. Suppose that we obtain the data shown in Figure 9.1 (and in the file Fig9-1.xls on your data disk) for a company on the level of sales observed for various levels of advertising expenditures in 10 different test markets around the country. We will assume that the different test markets are similar in terms of size and other demographic and economic characteristics. The main difference in each market is the level of advertising expenditure.

The data from Figure 9.1 are displayed graphically in Figure 9.2. This graph suggests a strong linear relationship between advertising expenditures and sales. Note that as advertising expenditures increase, sales increase proportionately. However, the relationship between advertising and sales is not perfect. For example, advertising expenditures

FIGURE 9.1

Sample data for
advertising
expenditures and
observed sales

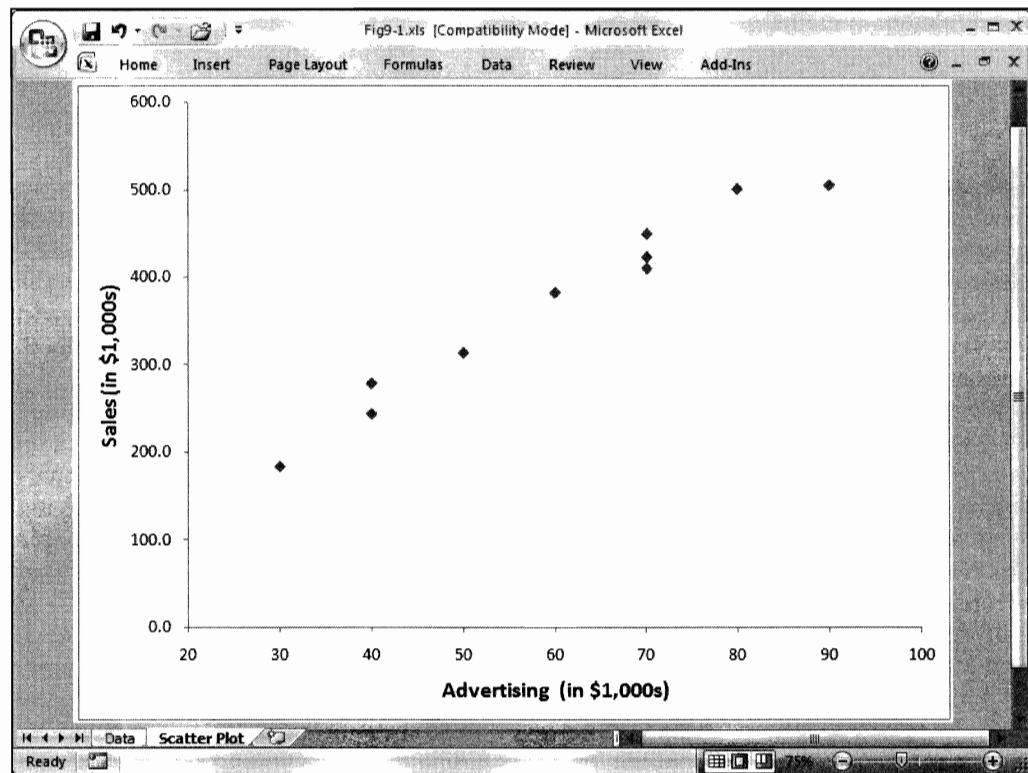


The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H
1								
2		Advertising Actual Sales						
3		Obs	(in \$1000s)	(in \$1000s)				
4		1	30	184.4				
5		2	40	279.1				
6		3	40	244.0				
7		4	50	314.2				
8		5	60	382.2				
9		6	70	450.2				
10		7	70	423.6				
11		8	70	410.2				
12		9	80	500.4				
13		10	90	505.3				
14								

FIGURE 9.2

Scatter diagram
for sales and
advertising data



of \$70,000 were used in three different test markets and resulted in three different levels of sales. Thus, the level of sales that occurs for a given level of advertising is subject to random fluctuation.

The random fluctuation, or scattering, of the points in Figure 9.2 suggests that some of the variation in sales is not accounted for by advertising expenditures. Because of the

scattering of points, this type of graph is called a **scatter diagram** or **scatter plot**. So although there is not a perfect *functional* relationship between sales and advertising (where each level of advertising yields one unique level of sales), there does seem to be a *statistical* relationship between these variables (where each level of sales is associated with a range or distribution of possible sales values).

Creating a Scatter Plot

To create a scatter plot like the one shown in Figure 9.2:

1. Select cells B4 through C13 shown in Figure 9.1.
2. Click the Insert command.
3. Click Scatter on the Charts menu.
4. Click Scatter with only Markers.

Excel's Chart Tools command then appears at the top of the screen, allowing you to make several selections concerning the type of chart you want and how it should be labeled and formatted. After Excel creates a basic chart, you can customize it in many ways. Double-click a chart element to display a dialog box with options for modifying the appearance of the element.

9.2 Regression Models

We will formalize the somewhat imprecise nature of a statistical relationship by adding an *error term* to what is otherwise a functional relationship. That is, in regression analysis, we consider models of the form:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon \quad 9.1$$

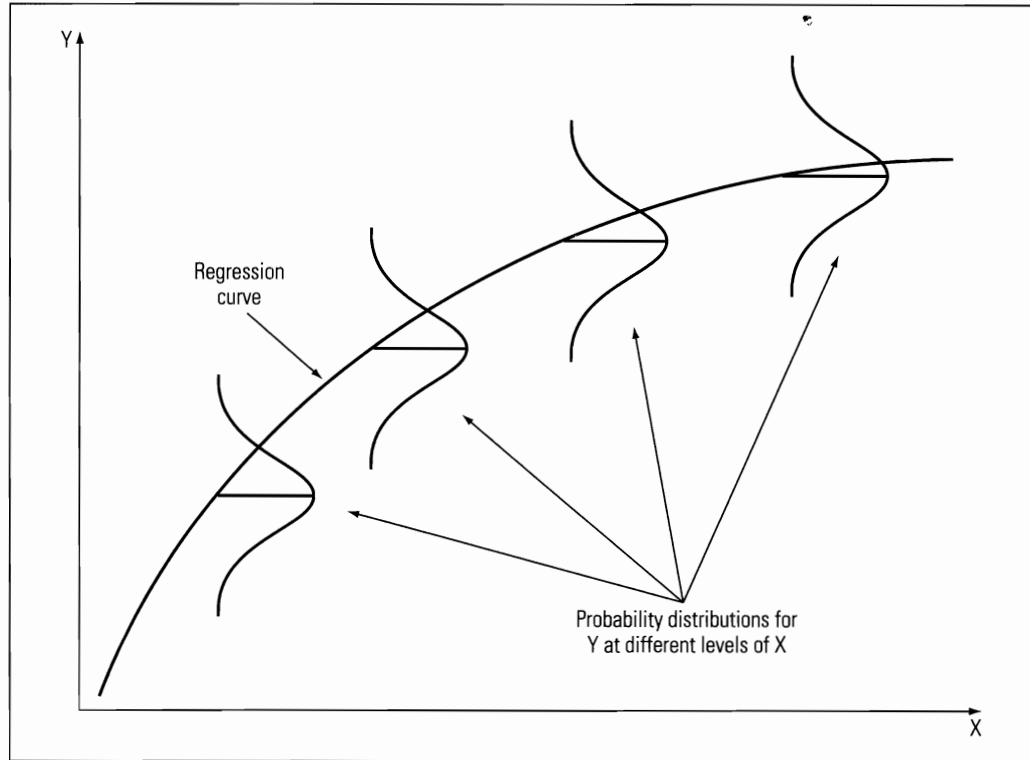
where ε represents a random disturbance, or error, term. Equation 9.1 is a **regression model**. The number of independent variables in a regression model differs from one application to another. Similarly, the form of $f(\cdot)$ varies from simple linear functions to more complex polynomial and nonlinear forms. In any case, the model in equation 9.1 conveys the two essential elements of a statistical relationship:

1. A tendency for the dependent variable Y to vary with the independent variable(s) in a systematic way, as expressed by $f(X_1, X_2, \dots, X_k)$ in equation 9.1.
2. An element of *unsystematic* or random variation in the dependent variable, as expressed by ε in equation 9.1.

The regression model in equation 9.1 indicates that for any values assumed by the independent variables X_1, \dots, X_k there is a probability distribution that describes the possible values that can be assumed by the dependent variable Y . This is portrayed graphically in Figure 9.3 for the case of a single independent variable. The curve drawn in Figure 9.3 represents the regression line (or regression function). It denotes the *systematic* variation between the dependent and independent variables (represented by $f(X_1, X_2, \dots, X_k)$ in equation 9.1). The probability distributions in Figure 9.3 denote the *unsystematic* variation in the dependent variable Y at different levels of the independent variable. This represents random variation in the dependent variable (represented by ε in equation 9.1) that cannot be accounted for by the independent variable.

FIGURE 9.3

Diagram of the distribution of Y values at various levels of X



Notice that the regression function in Figure 9.3 passes through the mean, or average, value for each probability distribution. Therefore, the regression function indicates what value, on average, the dependent variable is expected to assume at various levels of the independent variable. If we want to predict what value the dependent variable Y would assume at some level of the independent variable, the best estimate we could make is given by the regression function. That is, our best estimate of the value that Y will assume at a given level of the independent variable X_1 is the mean (or average) of the distribution of values for Y at that level of X_1 .

The actual value assumed by the dependent variable is likely to be somewhat different from our estimate because there is some random, unsystematic variation in the dependent variable that cannot be accounted for by our regression function. If we could repeatedly sample and observe actual values of Y at a given level of X_1 , sometimes the actual value of Y would be higher than our estimated (mean) value and sometimes it would be lower. So, the difference between the actual value of Y and our predicted value of Y would, on average, tend toward 0. For this reason, we can assume that the error term ε in equation 9.1 has an average, or expected, value of 0 if the probability distributions for the dependent variable Y at the various levels of the independent variable are normally distributed (bell-shaped) as shown in Figure 9.3.

9.3 Simple Linear Regression Analysis

As mentioned earlier, the function $f(\bullet)$ in equation 9.1 can assume many forms. However, the scatter plot in Figure 9.2 suggests that a strong linear relationship exists between the independent variable in our example (advertising expenditures) and the

dependent variable (sales). That is, we could draw a straight line through the data in Figure 9.2 that would fit the data fairly well. So, the formula of a straight line might account for the systematic variation between advertising and sales. Therefore, the following simple linear regression model might be an appropriate choice for describing the relationship between advertising and sales:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad 9.2$$

In equation 9.2, Y_i denotes the *actual* sales value for the i th observation, X_{1i} denotes the advertising expenditures associated with Y_i , and ε_i is an error term indicating that when X_{1i} dollars are spent on advertising, sales might not always equal $\beta_0 + \beta_1 X_{1i}$. The parameter β_0 represents a constant value (sometimes referred to as the Y-intercept because it represents the point where the line goes through the Y-axis) and β_1 represents the slope of the line (that is, the amount by which the line rises or falls per unit increase in X_1). Assuming that a straight line accounts for the systematic variation between Y and X_1 , the error terms ε_i represent the amounts by which the actual levels of sales are scattered around the regression line. Again, if the errors are scattered randomly around the regression line, they should average out to 0 or have an expected value of 0.

The model in equation 9.2 is a simple model because it contains only one independent variable. It is linear because none of the parameters (β_0 and β_1) appear as an exponent in the model or are multiplied or divided by one another.

Conceptually, it is important to understand that we are assuming that a large population of Y values occurs at each level of X_1 . The parameters β_0 and β_1 represent, respectively, the intercept and slope of the *true* regression line relating these populations. For this reason, β_0 and β_1 are sometimes referred to as **population parameters**. Usually we never know the exact numeric values for the population parameters in a given regression problem (we know that these values exist, but we don't know what they are). To determine the numeric values of the population parameters, we would have to look at the entire population of Y at each level of X_1 —usually an impossible task. However, by taking a sample of Y values at selected levels of X_1 we can estimate the values of the population parameters. We will identify the estimated values of β_0 and β_1 as b_0 and b_1 , respectively. The remaining problem is to determine the best values of b_0 and b_1 from our sample data.

9.4 Defining “Best Fit”

An infinite number of values could be assigned to b_0 and b_1 . So, searching for the exact values for b_0 and b_1 to produce the line that best fits our sample data might seem like searching for a needle in a haystack—and it is certainly not something we want to do manually. To have the computer estimate the values for b_0 and b_1 that produce the line that best fits our data, we must give it some guidance and define what we mean by the best fit.

We will use the symbol \hat{Y}_i to denote our estimated, or fitted, value of Y_i , which is defined as:

$$\hat{Y}_i = b_0 + b_1 X_{1i} \quad 9.3$$

We want to find values for b_0 and b_1 that make all the *estimated* sales values (\hat{Y}_i) as close as possible to the *actual* sales values (Y_i). For example, the data in Figure 9.1 indicate that we spent \$30,000 on advertising ($X_{1i} = 30$) and observed sales of \$184,400 ($Y_1 = 184.4$). So in equation 9.3, if we let $X_{1i} = 30$, we want \hat{Y}_i to assume a value that is as close as possible to 184.4. Similarly, in the three instances in Figure 9.1 where \$70,000

was spent on advertising ($X_{16} = X_{17} = X_{18} = 70$), we observed sales of \$450,200, \$423,600, and \$410,200 ($Y_6 = 450.2$, $Y_7 = 423.6$, $Y_8 = 410.2$). So in equation 9.3, if we let $X_{1i} = 70$, we want \hat{Y}_i to assume a value that is as close as possible to 450.2, 423.6, and 410.2.

If we could find values for b_0 and b_1 so that all the estimated sales values were exactly the same as all the actual sales values ($\hat{Y}_i = Y_i$ for all observations i), we would have the equation of the straight line that passes through each data point—in other words, the line would fit our data perfectly. This is impossible for the data in Figure 9.2 because a straight line could not be drawn to pass through each data point in the graph. In most regression problems, it is impossible to find a function that fits the data perfectly because most data sets contain some amount of unsystematic variation.

Although we are unlikely to find values for b_0 and b_1 that will allow us to fit our data perfectly, we will try to find values that make the differences between the estimated values for the dependent variable and the actual values for the dependent variable ($Y_i - \hat{Y}_i$) as small as possible. We refer to the difference $Y_i - \hat{Y}_i$ as the **estimation error** for observation i because it measures how far away the estimated value \hat{Y}_i is from the actual value Y_i . The estimation errors in a regression problem are also referred to as *residuals*.

Although different criteria can be used to determine the best values for b_0 and b_1 , the most widely used method determines the values that minimize the sum of squared estimation errors—or *error sum of squares* (ESS) for short. That is, we will attempt to find values for b_0 and b_1 that minimize:

$$\text{ESS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1 X_{1i})]^2 \quad 9.4$$

Several observations should be made concerning ESS. Because each estimation error is squared, the value of ESS will always be nonnegative and, therefore, the smallest value ESS can assume is 0. The only way for ESS to equal 0 is for all the individual estimation errors to be 0 ($Y_i - \hat{Y}_i = 0$ for all observations), in which case the estimated regression line would fit our data perfectly. Thus, minimizing ESS seems to be a good objective to use in searching for the best values of b_0 and b_1 . Because regression analysis finds the values of the parameter estimates that minimize the sum of squared estimation errors, it is sometimes referred to as the **method of least squares**.

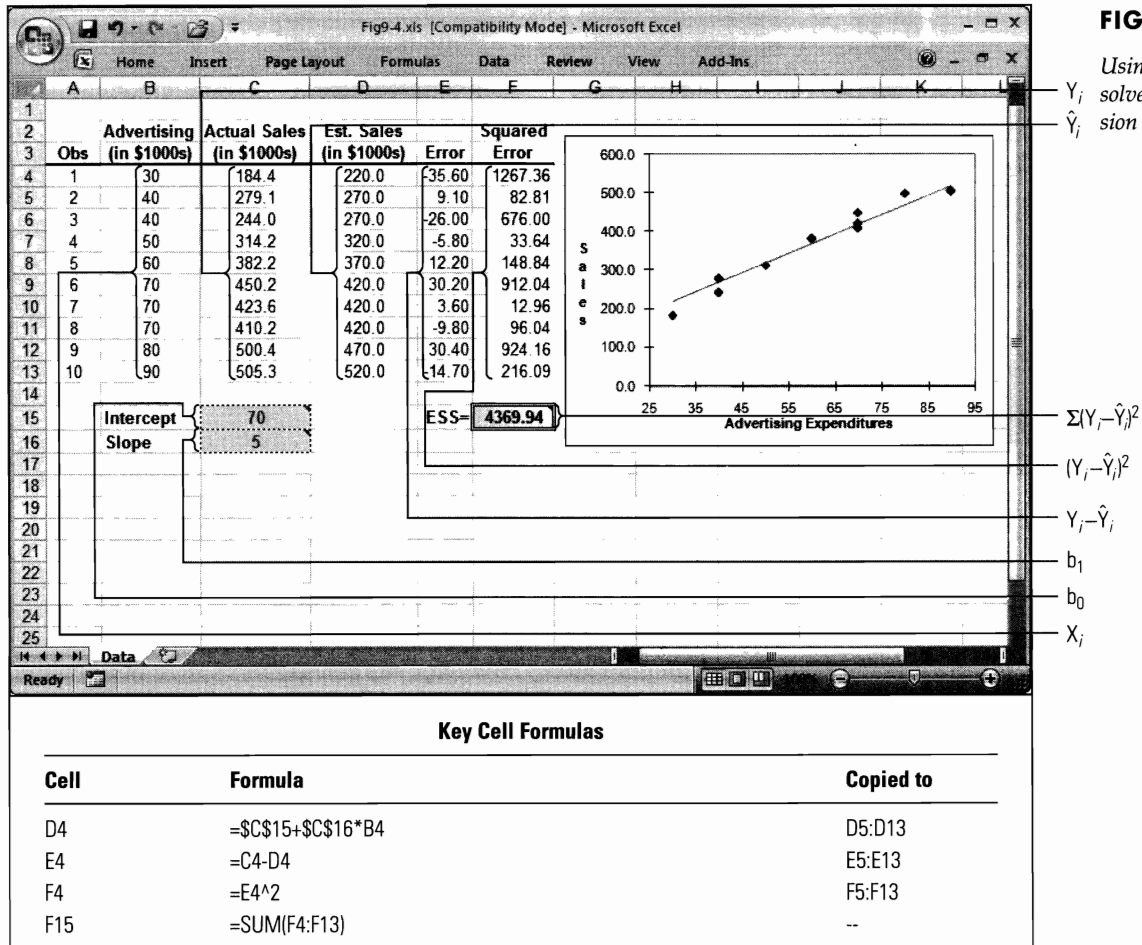
9.5 Solving the Problem Using Solver

We can calculate the optimal parameter estimates for a linear regression model in several ways. As in earlier chapters, we can use Solver to find the values for b_0 and b_1 that minimize the ESS quantity in equation 9.4.

Finding the optimal values for b_0 and b_1 in equation 9.4 is an unconstrained nonlinear optimization problem. Consider the spreadsheet in Figure 9.4.

In Figure 9.4, cells C15 and C16 represent the values for b_0 and b_1 , respectively. These cells are labeled Intercept and Slope because b_0 represents the intercept in equation 9.3 and b_1 represents the slope. Values of 70 and 5 were entered for these cells as rough guesses of their optimal values.

To use Solver to calculate the optimal values of b_0 and b_1 , we need to implement a formula in the spreadsheet that corresponds to the ESS calculation in equation 9.4. This formula represents the objective function to be minimized. To calculate the ESS, we first need to calculate the sales values estimated by the regression function in equation 9.3 for



each observation in our sample. These estimated sales values (\hat{Y}_i) were created in column D as:

$$\text{Formula for cell D4: } =\$C\$15+\$C\$16*B4$$

(Copy to D5 through D13.)

The estimation errors ($Y_i - \hat{Y}_i$) were calculated in column E as:

$$\text{Formula for cell E4: } =C4-D4$$

(Copy to E5 through E13.)

The squared estimation errors ($(Y_i - \hat{Y}_i)^2$) were calculated in column F as:

$$\text{Formula for cell F4: } =E4^2$$

(Copy to F5 through F13.)

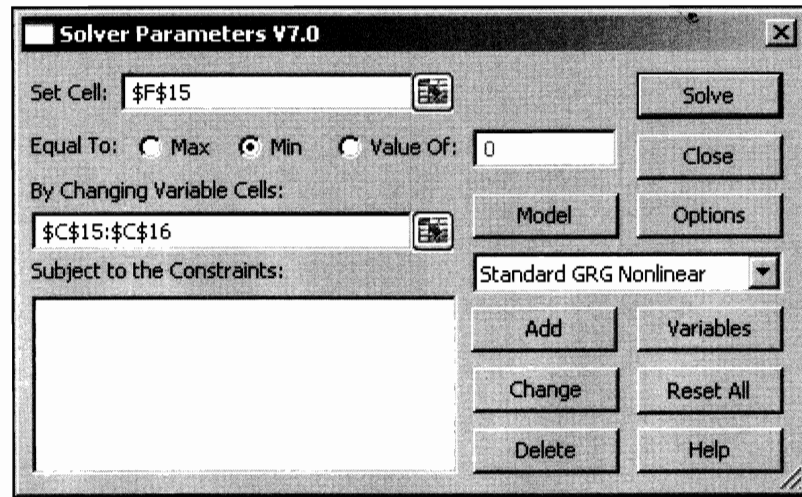
Finally, the sum of the squared estimation errors (ESS) was calculated in cell F15 as:

$$\text{Formula for cell F15: } =\text{SUM}(F4:F13)$$

Note that the formula in cell F15 corresponds exactly to equation 9.4.

FIGURE 9.5

Solver parameters for the regression problem



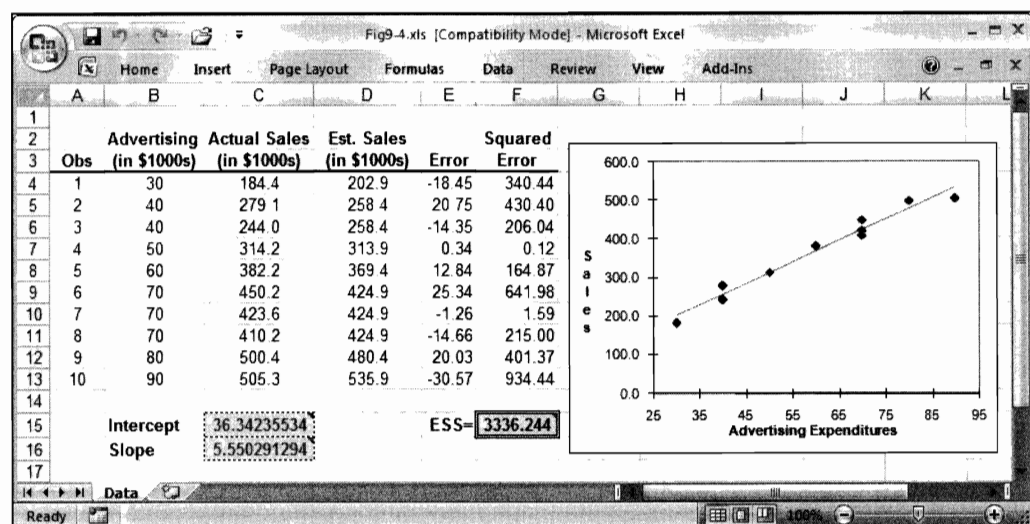
The graph in Figure 9.4 plots the line connecting the estimated sales values against the actual sales values. The intercept and slope of this line are determined by the values in C15 and C16. Although this line seems to fit our data fairly well, we do not know if this is the line that minimizes the ESS value. However, we can use the Solver parameters shown in Figure 9.5 to determine the values for C15 and C16 that minimize the ESS value in F15.

Figure 9.6 shows the optimal solution to this problem. In this spreadsheet, the intercept and slope of the line that best fits our data are $b_0 = 36.34235$ and $b_1 = 5.550293$, respectively. The ESS value of 3,336.244 associated with these optimal parameter estimates is better (or smaller) than the ESS value for the parameter estimates shown in Figure 9.4. No other values for b_0 and b_1 would result in an ESS value smaller than the one shown in Figure 9.6. Thus, the equation of the straight line that best fits our data according to the least squares criterion is represented by:

$$\hat{Y}_i = 36.34235 + 5.550291X_i \quad 9.5$$

FIGURE 9.6

Optimal solution to the regression problem



9.6 Solving the Problem Using the Regression Tool

In addition to Solver, Excel provides another tool for solving regression problems that is easier to use and provides more information about a regression problem. We will demonstrate the use of this regression tool by referring back to the original data for the current problem, shown in Figure 9.1. Before you can use the regression tool in Excel, you need to make sure that the Analysis ToolPak add-in is available. You can do this by completing the following steps:

1. Click Office button, Excel options, Add-Ins.
2. Locate and activate the Analysis ToolPak add-in. (If Analysis ToolPak is not listed among your available add-ins, you will need to install it from your Microsoft Office CD.)

After ensuring that the Analysis ToolPak is available, you can access the regression tool by completing the following steps:

1. Click the Data tab.
2. Click Data Analysis in the Analysis menu.
3. Select Regression and click OK.

After you choose the Regression command, the Regression dialog box appears, as shown in Figure 9.7. This dialog box presents many options and selections. At this point, we will focus on only three options: the Y-Range, the X-Range, and the Output-Range. The Y-Range corresponds to the range in the spreadsheet containing the sample

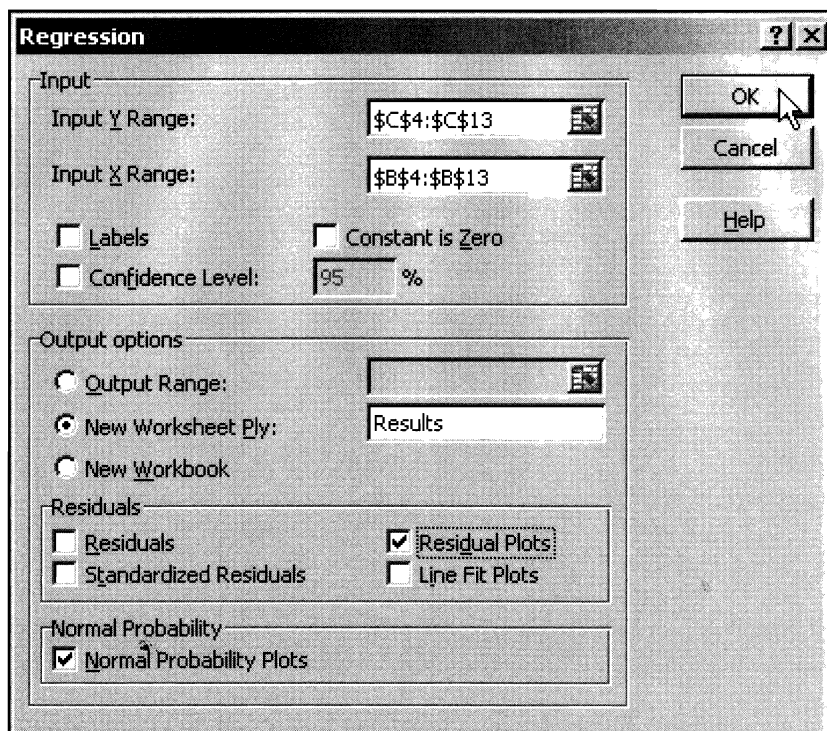


FIGURE 9.7

Regression dialog box

observations for the *dependent* variable (C4 through C13 for the example in Figure 9.1). The X-Range corresponds to the range in the spreadsheet containing the sample observations for the *independent* variable (B4 through B13 for the current example). We also need to specify the output range where we want the regression results to be reported. In Figure 9.7, we selected the New Worksheet Ply option to indicate that we want the regression results placed on a new sheet named "Results." With the dialog box selections complete, we can click the OK button and Excel will calculate the least squares values for b_0 and b_1 (along with other summary statistics).

Figure 9.8 shows the Results sheet for our example. For now, we will focus on only a few values in Figure 9.8. Note that the value labeled "Intercept" in cell B17 represents the optimal value for b_0 ($b_0 = 36.342$). The value representing the coefficient for "X Variable 1" in cell B18 represents the optimal value for b_1 ($b_1 = 5.550$). Thus, the estimated regression function is represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} = 36.342 + 5.550X_{1i} \quad 9.6$$

Equation 9.6 is essentially the same result we obtained earlier using Solver (see equation 9.5). Thus, we can calculate the parameter estimates for a regression function using either Solver or the regression tool shown in Figure 9.7. The advantage of the regression tool is that it does not require us to set up any special formulas or cells in the spreadsheet, and it produces additional statistical results about the problem under study.

FIGURE 9.8

Results for the regression calculations

Fig9-1.xls [Compatibility Mode] - Microsoft Excel									
	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.98444428							
5	R Square	0.96913053							
6	Adjusted R Square	0.96527185							
7	Standard Error	20.4213237							
8	Observations	10							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	104739.600	104739.600	251.156	0.000			
13	Residual	8	3336.244	417.030					
14	Total	9	108075.844						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	36.342	21.983	1.653	0.137	-14.351	87.036	-14.351	87.036
18	X Variable 1	5.550	0.350	15.848	0.000	4.743	6.358	4.743	6.358
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	Observation	Predicted Y	Residuals						
25	1	202.851	-18.451						
26	2	258.354	20.746						
27	3	258.354	-14.354						
28	4	313.857	0.343						
29	5	369.360	12.840						
30									
31	PROBABILITY OUTPUT								
32									
33	Percentile	Y							
34	5	184.400							
35	15	244.000							
36	25	279.100							
37	35	314.200							
38	45	382.200							

9.7 Evaluating the Fit

Our goal in the example problem is to identify the equation of a straight line that fits our data well. Having calculated the estimated regression line (using either Solver or the regression tool), we might be interested in determining how well the line fits our data. Using equation 9.6, we can compute the estimated or expected level of sales (\hat{Y}_i) for each observation in our sample. The \hat{Y}_i values could be calculated in column D of Figure 9.9 as follows:

Formula for cell D4: $=36.342+5.550*B4$
 (Copy to D5 through D13.)

However, we can also use the TREND() function in Excel to compute the \hat{Y}_i values in column D as follows:

Alternate Formula for cell D4: $=TREND(\$C\$4:\$C\$13,\$B\$4:\$B\$13,B4)$
 (Copy to D5 through D13.)

This TREND() function computes the least squares linear regression line using a Y-range of C4 through C13 and an X-range of B4 through B13. It then uses this regression function to estimate the value of Y using the value of X given in cell B4. Thus, using the TREND() function, we don't have to worry about typing the wrong values for the estimated intercept or slope. Notice that the resulting estimated sales values shown in column D in Figure 9.9 match the predicted Y values shown toward the bottom on column B in Figure 9.8.

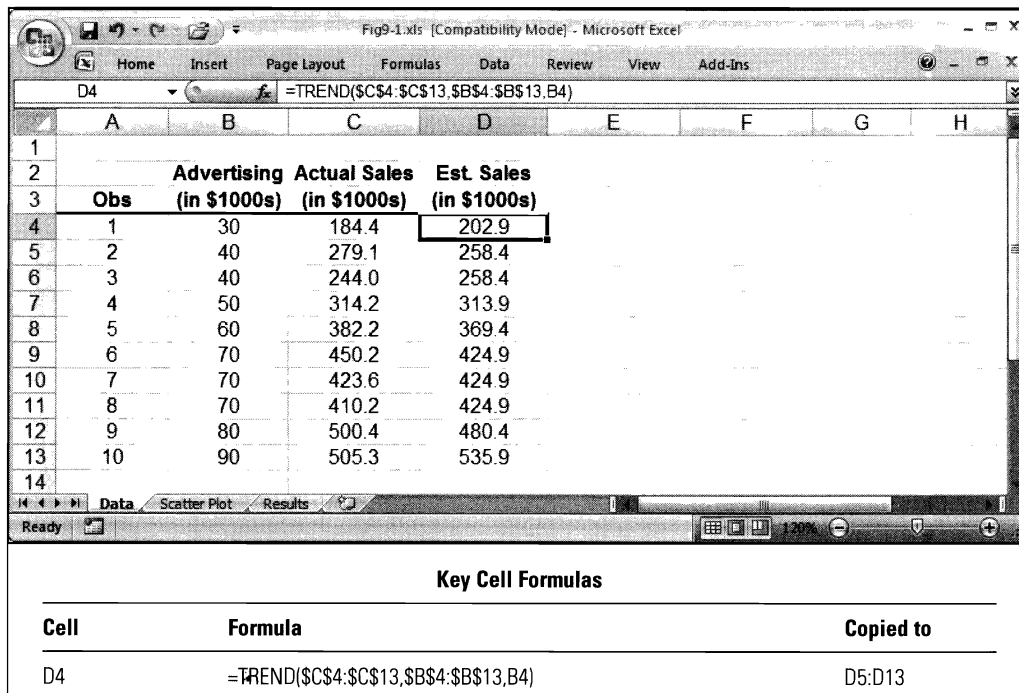


FIGURE 9.9

Estimated sales values at each level of advertising

A Note on the TREND() Function

The TREND() function can be used to calculate the estimated values for linear regression models. The format of the TREND() function is as follows:

TREND(Y-range, X-range, X-value for prediction)

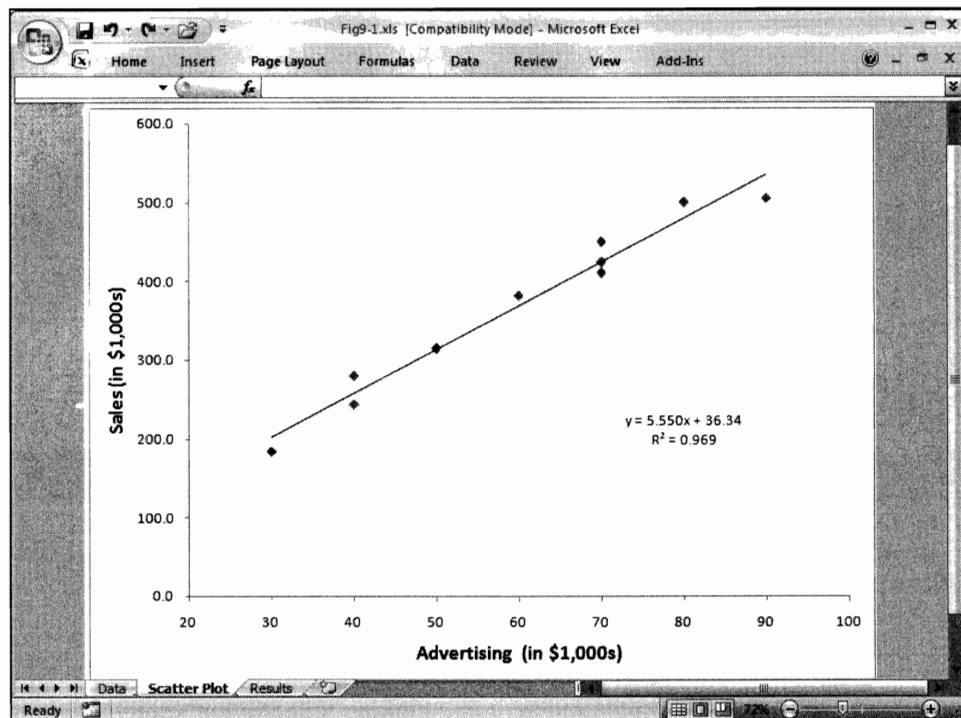
where Y-range is the range in the spreadsheet containing the dependent Y variable, X-range is the range in the spreadsheet containing the independent X variable(s), and X-value for prediction is a cell (or cells) containing the values for the independent X variable(s) for which we want an estimated value of Y. The TREND() function has an advantage over the regression tool in that it is dynamically updated whenever any inputs to the function change. However, it does not provide the statistical information provided by the regression tool. It is best to use these two different approaches to doing regression in conjunction with one another.

Figure 9.10 shows a graph of the estimated regression function along with the actual sales data. This function represents the expected amount of sales that would occur for each value of the independent variable (that is, each value in column D of Figure 9.9 falls on this line). To insert this estimated trend line on the existing scatter plot:

1. Right-click on any of the data points in the scatter plot to select the series of data.
2. Select Add Trendline.
3. Click Linear.
4. Select the "Display equation on chart" and "Display R-squared value on chart."
5. Click Close.

FIGURE 9.10

Graph of the regression line through the actual sales data



From this graph, we see that the regression function seems to fit the data reasonably well in this example. In particular, it seems that the actual sales values fluctuate around this line in a fairly unsystematic, or random, pattern. Thus, it appears that we have achieved our goal of identifying a function that accounts for most, if not all, of the systematic variation between the dependent and independent variables.

9.8 The R² Statistic

In Figure 9.8, the value labeled “R Square” in cell B5 (or “R²” in Figure 9.10) provides a goodness-of-fit measure. This value represents the R² statistic (also referred to as the coefficient of determination). This statistic ranges in value from 0 to 1 ($0 \leq R^2 \leq 1$) and indicates the proportion of the total variation in the dependent variable Y around its mean (average) that is accounted for by the independent variable(s) in the estimated regression function.

The total variation in the dependent variable Y around its mean is described by a measure known as the *total sum of squares* (TSS), which is defined as:

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad 9.7$$

The TSS equals the sum of the squared differences between each observation Y_i in the sample and the average value of Y, denoted in equation 9.7 by \bar{Y} . The difference between each observed value of Y_i and the average value \bar{Y} can be decomposed into two parts as:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \quad 9.8$$

Figure 9.11 illustrates this decomposition for a hypothetical data point. The value $Y_i - \hat{Y}_i$ in equation 9.8 represents the estimation error, or the amount of the total deviation between Y_i and \bar{Y} that is not accounted for by the regression function. The value $\hat{Y}_i - \bar{Y}$ in equation 9.8 represents the amount of the total deviation in Y_i from \bar{Y} that is accounted for by the regression function.

The decomposition of the individual deviation in equation 9.8 also applies to the TSS in equation 9.7. That is, the *total sum of squares* (TSS) can be decomposed into the following two parts:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad 9.9$$

$$TSS = ESS + RSS$$

ESS is the quantity that is minimized in least squares regression. ESS represents the amount of variation in Y around its mean that the regression function cannot account for, or the amount of variation in the dependent variable that is unexplained by the regression function. Therefore, the *regression sum of squares* (RSS) represents the amount of variation in Y around its mean that the regression function can account for, or the amount of variation in the dependent variable that is explained by the regression function. In Figure 9.8, cells C12, C13, and C14 contain the values for RSS, ESS, and TSS, respectively.

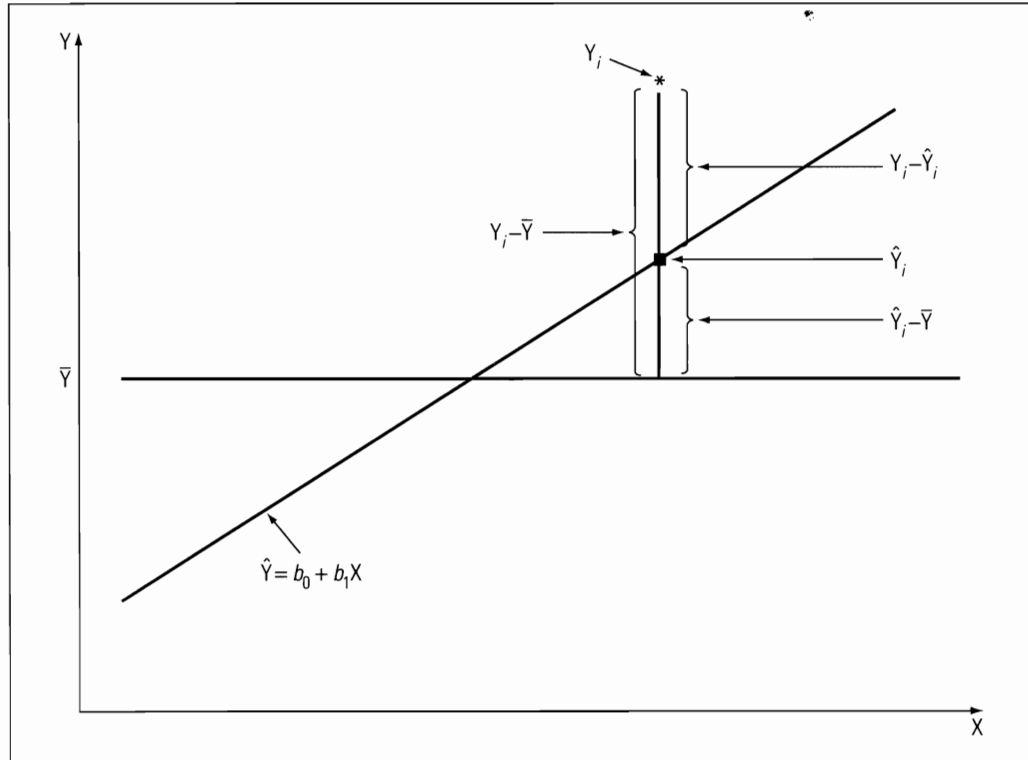
Now consider the following definitions of the R² statistic:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \quad 9.10$$

From the previous definition of TSS in equation 9.9, we can see that if $ESS = 0$ (which can occur only if the regression function fits the data perfectly), then $TSS = RSS$ and, therefore, $R^2 = 1$. On the other hand, if $RSS = 0$ (which means that the regression function was unable to explain any of the variation in the behavior of the dependent variable Y),

FIGURE 9.11

Decomposition of the total deviation into error and regression components



then $TSS = ESS$ and $R^2 = 0$. So, the closer the R^2 statistic is to the value 1, the better the estimated regression function fits the data.

From cell B5 in Figure 9.8, we observe that the value of the R^2 statistic is approximately 0.969. This indicates that approximately 96.9% of the total variation in our dependent variable around its mean has been accounted for by the independent variable in our estimated regression function. Because this value is fairly close to the maximum possible R^2 value (1), this statistic indicates that the regression function we have estimated fits our data well. This is confirmed by the graph in Figure 9.10.

The **multiple R** statistic shown in cell B4 of the regression output in Figure 9.8 represents the strength of the linear relationship between actual and estimated values for the dependent variable. As with the R^2 statistic, the multiple R statistic varies between 0 and 1 with values near 1 indicating a good fit. When a regression model includes only one independent variable, the multiple R statistic is equivalent to the square root of the R^2 statistic. We'll focus on the R^2 statistic because its interpretation is more apparent than that of the multiple R statistic.

9.9 Making Predictions

Using the estimated regression in equation 9.6, we can make predictions about the level of sales expected for different levels of advertising expenditures. For example, suppose that the company wants to estimate the level of sales that would occur if \$65,000 were spent on advertising in a given market. Assuming that the market in question is similar to those used in estimating the regression function, the expected sales level is estimated as:

$$\text{Estimated Sales} = b_0 + b_1 \times 65 = 36.342 + 5.550 \times 65 = 397.092$$

(in \$1000s)

So, if the company spends \$65,000 on advertising (in a market similar to those used to estimate the regression function), we would expect to observe sales of approximately \$397,092. The *actual* level of sales is likely to differ somewhat from this value due to other random factors influencing sales.

9.9.1 THE STANDARD ERROR

A measure of the accuracy of the prediction obtained from a regression model is given by the standard deviation of the estimation errors—also known as the standard error, S_e . If we let n denote the number of observations in the data set, and k denote the number of independent variables in the regression model, the formula for the standard error is represented by:

$$S_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - k - 1}} \quad 9.11$$

The **standard error** measures the amount of scatter, or variation, in the actual data around the fitted regression function. Cell B7 in Figure 9.8 indicates that the standard error for our example problem is $S_e = 20.421$.

The standard error is useful in evaluating the level of uncertainty in predictions we make with a regression model. As a *very* rough rule of thumb, there is approximately a 68% chance of the actual level of sales falling within ± 1 standard error of the predicted value \hat{Y}_i . Alternatively, the chance of the actual level of sales falling within ± 2 standard errors of the predicted value \hat{Y}_i is approximately 95%. In our example, if the company spends \$65,000 on advertising, we could be roughly 95% confident that the actual level of sales observed would fall somewhere in the range from \$356,250 to \$437,934 ($\hat{Y}_i \pm 2S_e$).

9.9.2 PREDICTION INTERVALS FOR NEW VALUES OF Y

To calculate a more accurate confidence interval for a prediction, or **prediction interval**, of a new value of Y when $X_1 = X_{1h}$, we first calculate the estimated value \hat{Y}_h as:

$$\hat{Y}_h = b_0 + b_1 X_{1h} \quad 9.12$$

A $(1 - \alpha)\%$ prediction interval for a new value of Y when $X_1 = X_{1h}$ is represented by:

$$\hat{Y}_h \pm t_{(1-\alpha/2, n-2)} S_p \quad 9.13$$

where $t_{(1-\alpha/2, n-2)}$ represents the $1 - \alpha/2$ percentile of a t -distribution with $n - 2$ degrees of freedom, and S_p represents the standard prediction error defined by:

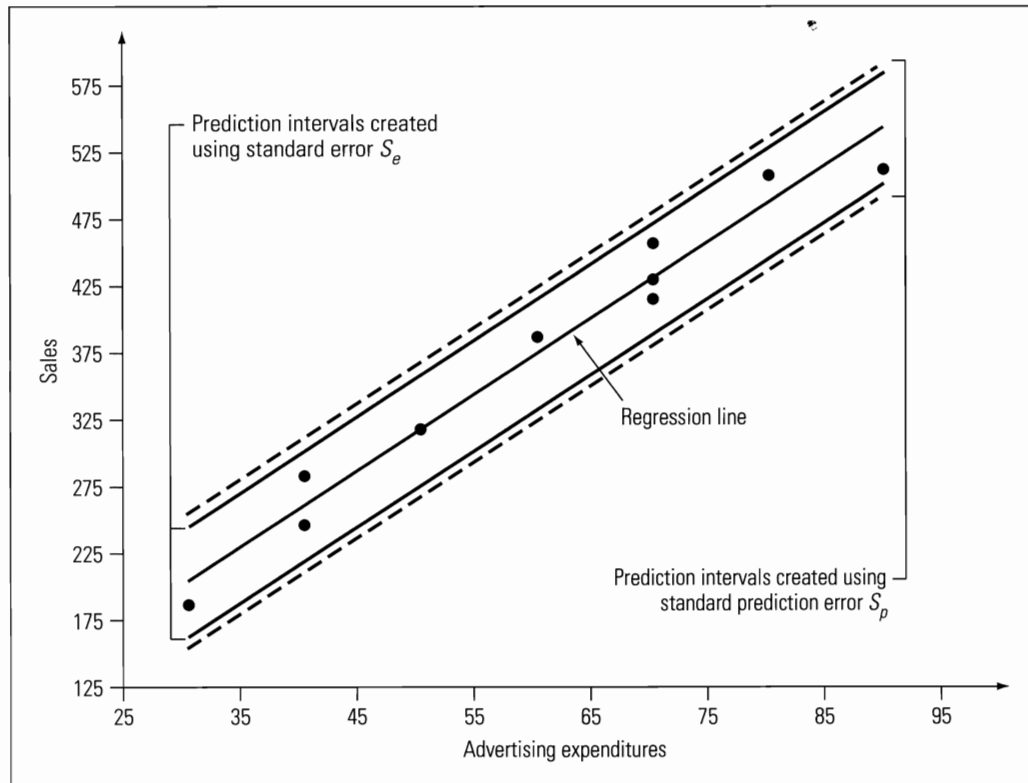
$$S_p = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_{1h} - \bar{X})^2}{\sum_{i=1}^n (X_{1i} - \bar{X})^2}} \quad 9.14$$

The rule of thumb presented earlier is a generalization of equation 9.13. Notice that S_p is always larger than S_e because the term under the square root symbol is always greater than 1. Also notice that the magnitude of the difference between S_p and S_e increases as the difference between X_{1h} and \bar{X} increases. Thus, the prediction intervals generated by the rule of thumb tend to underestimate the true amount of uncertainty involved in making predictions. This is illustrated in Figure 9.12.

As shown in Figure 9.12, for this example problem, there is not a lot of difference between the prediction intervals created using the rule of thumb and the more accurate prediction interval given in equation 9.13. In a situation requiring a precise prediction

FIGURE 9.12

Comparison of prediction intervals obtained using the rule of thumb and the more accurate statistical calculation



interval, the various quantities needed to construct the prediction interval in equation 9.13 can be computed easily in Excel. Figure 9.13 provides an example of a 95% prediction interval for a new value of sales when \$65,000 is spent on advertising.

To create this prediction interval, we first use the `TREND()` function to calculate the estimated sales level (\hat{Y}_h) when advertising equals \$65,000 ($X_{1h} = 65$). The value 65 is entered in cell B17 to represent X_{1h} and the estimated sales level (\hat{Y}_h) is calculated in cell D17 as:

Formula for cell D17: `=TREND(C4:C13,B4:B13,B17)`

The expected level of sales when \$65,000 is spent on advertising is approximately \$397,100. The standard error (S_e) shown in cell B19 is extracted from the Results sheet shown in Figure 9.8 as:

Formula for cell B19: `=Results!B7`

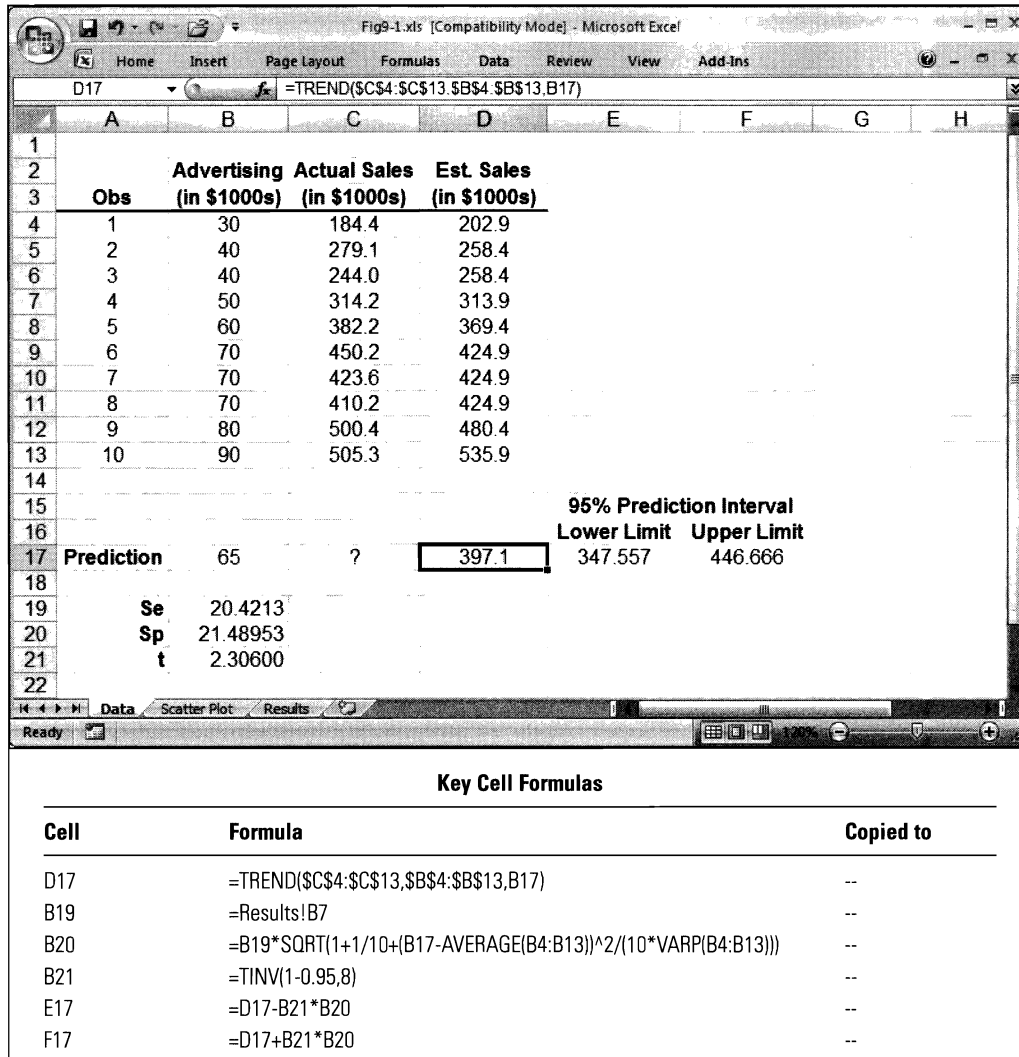
The standard prediction error (S_p) is calculated in cell B20 as:

Formula for cell B20: `=B19*SQRT(1+1/10+(B17-AVERAGE(B4:B13))^2/(10*VARP(B4:B13)))`

The value 10 appearing in the preceding formula corresponds to the sample size n in equation 9.14. The appropriate t -value for a 95% confidence (or prediction) interval is calculated in cell B21 as:

Formula for cell B21: `=TINV(1-0.95,8)`

The first argument in the preceding formula corresponds to 1 minus the desired confidence level (or $\alpha = 0.05$). The second argument corresponds to $n - 2$ ($10 - 2 = 8$). Cells E17 and F17 calculate the lower and upper limits of the prediction interval as:

**FIGURE 9.13**

Example of calculating a prediction interval

Formula for cell E17: =D17-B21*B20

Formula for cell F17: =D17+B21*B20

The results indicate that when \$65,000 is spent on advertising, we expect to observe sales of approximately \$397,100, but realize that the actual sales level is likely to deviate somewhat from this value. However, we can be 95% confident that the actual sales value observed will fall somewhere in the range from \$347,556 to \$446,666. (Notice that this prediction interval is somewhat wider than the range from \$356,250 to \$437,934 generated earlier using the rule of thumb.)

9.9.3 CONFIDENCE INTERVALS FOR MEAN VALUES OF Y

At times, you might want to construct a confidence interval for the average, or mean, value of Y when $X_1 = X_{1h}$. This involves a slightly different procedure from constructing

a prediction interval for a new individual value of Y when $X_1 = X_{1h}^*$. A $(1 - \alpha)\%$ confidence interval for the average value of Y when $X_1 = X_{1h}$ is represented by:

$$\hat{Y}_h \pm t_{(1-\alpha/2; n-2)} S_a \quad 9.15$$

where \hat{Y}_h is defined by equation 9.12, $t_{(1-\alpha/2; n-2)}$ represents the $1 - \alpha/2$ percentile of a t -distribution with $n - 2$ degrees of freedom, and S_a is represented by:

$$S_a = S_e \sqrt{\frac{1}{n} + \frac{(X_{1h} - \bar{X})^2}{\sum_{i=1}^n (X_{1i} - \bar{X})^2}} \quad 9.16$$

Comparing the definition of S_a in equation 9.16 with that of S_p in equation 9.14 reveals that S_a will always be smaller than S_p . Therefore, the confidence interval for the average value of Y when $X_1 = X_{1h}$ will be tighter (or cover a smaller range) than the prediction interval for a new value of Y when $X_1 = X_{1h}$. This type of confidence interval can be implemented in a similar way to that described earlier for prediction intervals.

9.9.4 A NOTE ABOUT EXTRAPOLATION

Predictions made using an estimated regression function might have little or no validity for values of the independent variable that are substantially different from those represented in the sample. For example, the advertising expenditures represented in the sample in Figure 9.1 range from \$30,000 to \$90,000. Thus, we cannot assume that our model will give accurate estimates of sales levels at advertising expenditures significantly above or below this range of values, because the relationship between sales and advertising might be quite different outside this range.

9.10 Statistical Tests for Population Parameters

Recall that the parameter β_1 in equation 9.2 represents the slope of the *true* regression line (or the amount by which the dependent variable Y is expected to change given a unit change in X_1). If no linear relationship exists between the dependent and independent variables, the true value of β_1 for the model in equation 9.2 should be 0. As mentioned earlier, we cannot calculate or observe the true value of β_1 but instead must estimate its value using the sample statistic b_1 . However, because the value of b_1 is based on a sample rather than on the entire population of possible values, its value probably is not exactly equal to the true (but unknown) value of β_1 . Thus, we might want to determine how different the true value of β_1 is from its estimated value b_1 . The regression results in Figure 9.8 provide a variety of information addressing this issue.

Cell B18 in Figure 9.8 indicates that the estimated value of β_1 is $b_1 = 5.550$. Cells F18 and G18 give the lower and upper limits of a 95% confidence interval for the true value of β_1 . That is, we can be 95% confident that $4.74 \leq \beta_1 \leq 6.35$. This indicates that for every \$1,000 increase in advertising, we would expect to see an increase in sales of approximately \$4,740 to \$6,350. Notice that this confidence interval does not include the value 0. Thus, we can be at least 95% confident that a linear relationship exists between advertising and sales ($\beta_1 \neq 0$). (If we want an interval other than a 95% confidence interval, we can use the Confidence Level option in the Regression dialog box, shown in Figure 9.7, to specify a different interval.)

The t -statistic and p -value listed in cells D18 and E18 in Figure 9.8 provide another way of testing whether $\beta_1 = 0$. According to statistical theory, if $\beta_1 = 0$, then the ratio of

b_1 to its standard error should follow a t -distribution with $n-2$ degrees of freedom. Thus, the t -statistic for testing if $\beta_1 = 0$ in cell D18 is:

$$t\text{-statistic in cell D18} = \frac{b_1}{\text{standard error of } b_1} = \frac{5.550}{0.35022} = 15.848$$

The p -value in cell E18 indicates the probability of obtaining an outcome that is more extreme than the observed test statistic value if $\beta_1 = 0$. In this case, the p -value is 0, indicating that there is virtually no chance that we will obtain an outcome as large as the observed value for b_1 if the true value of β_1 is 0. Therefore, we conclude that the true value of β_1 is not equal to 0. This is the same conclusion implied earlier by the confidence interval for β_1 .

The t -statistic, p -value, and confidence interval for the intercept β_0 are listed in Figure 9.8 in row 17, and would be interpreted in the same way as demonstrated for β_1 . Notice that the confidence interval for β_0 straddles the value 0 and, therefore, we cannot be certain that the intercept is significantly different from 0. The p -value for β_0 indicates that we have a 13.689% chance of obtaining an outcome more extreme than the observed value of b_0 if the true value of β_0 is 0. Both of these results indicate a fair chance that $\beta_0 = 0$.

9.10.1 ANALYSIS OF VARIANCE

The *analysis of variance* (ANOVA) results, shown in Figure 9.8, provide another way of testing whether or not $\beta_1 = 0$. The values in the MS column in the ANOVA table represent values known as the *mean squared regression* (MSR) and *mean squared error* (MSE), respectively. These values are computed by dividing the RSS and ESS values in C12 and C13 by the corresponding degrees of freedom values in cells B12 and B13.

If $\beta_1 = 0$, then the ratio of MSR to MSE follows an F-distribution. The statistic labeled “F” in cell E12 is:

$$F\text{-statistic in cell E12} = \frac{\text{MSR}}{\text{MSE}} = \frac{104739.6}{417.03} = 251.156$$

The value in F12 labeled “Significance F” is similar to the p -values described earlier, and indicates the probability of obtaining a value in excess of the observed value for the F-statistic if $\beta_1 = 0$. In this case, the significance of F is 0, indicating that there is virtually no chance that we would have obtained the observed value for b_1 if the true value of β_1 is 0. Therefore, we conclude that the true value of β_1 is not equal to 0. This is the same conclusion implied earlier by our previous analysis.

The F-statistic might seem a bit redundant, given that we can use the t -statistic to test whether or not $\beta_1 = 0$. However, the F-statistic serves a different purpose, which becomes apparent in multiple regression models with more than one independent variable. The F-statistic tests whether or not *all* of the β_i for *all* of the independent variables in a regression model are all simultaneously equal to 0. A simple linear regression model contains only one independent variable. In this case, the tests involving the F-statistic and the t -statistic are equivalent.

9.10.2 ASSUMPTIONS FOR THE STATISTICAL TESTS

The methods for constructing confidence intervals are based on important assumptions concerning the simple linear regression model presented in equation 9.2. Throughout this discussion, we assumed that the error terms ε_i are independent, normally distributed random variables with expected (or mean) values of 0 and constant variances. Thus, the statistical procedures for constructing intervals and performing t -tests apply

only when these assumptions are true for a given set of data. As long as these assumptions are not seriously violated, the procedures described offer good approximations of the desired confidence intervals and t -tests. Various diagnostic checks can be performed on the residuals ($Y_i - \hat{Y}_i$) to see whether or not our assumptions concerning the properties of the error terms are valid. These diagnostics are discussed in depth in most statistics books, but are not repeated in this text. Excel also provides basic diagnostics that can help determine whether assumptions about the error terms are violated.

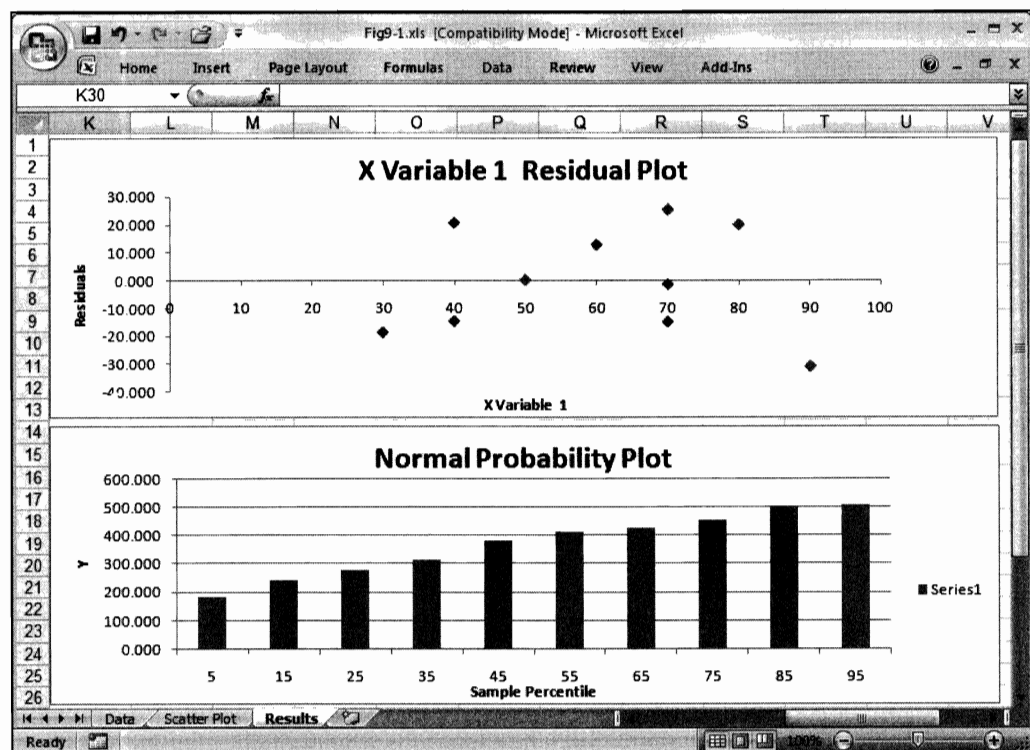
The Regression dialog box (shown in Figure 9.7) provides two options for producing graphs that highlight serious violations of the error term assumptions. These options are Residual Plots and Normal Probability Plots. Figure 9.14 shows the graphs produced by these two options for our example problem.

The first graph in Figure 9.14 results from the Residual Plots option. This graph plots the residuals (or estimation errors) versus each independent variable in the regression model. Our example problem involves one independent variable—therefore, we have one residual plot. If the assumptions underlying the regression model are met, the residuals should fall within a horizontal band centered on zero and should display no systematic tendency to be positive or negative. The residual plot in Figure 9.14 indicates that the residuals for our example problem fall randomly within a range from -30 to $+30$. Thus, no serious problems are indicated by this graph.

The second graph in Figure 9.14 results from the Normal Probability Plot option. If the error terms in equation 9.2 are normally distributed random variables, the dependent variable in equation 9.2 is a normally distributed random variable prior to sampling. Thus, one way to evaluate whether we can assume that the error terms are normally distributed is to determine if we can assume that the dependent variable is normally distributed. The normal probability plot provides an easy way to evaluate whether the sample values on the dependent variable are consistent with the normality

FIGURE 9.14

Residual plot and normal probability plot for the example problem



assumption. A plot with an approximately linear rate of increase (such as the one in Figure 9.14) supports the assumption of normality.

If the residual plot shows a systematic tendency for the residuals to be positive or negative, this indicates that the function chosen to model the systematic variation between the dependent and independent variables is inadequate and that another functional form would be more appropriate. An example of this type of residual plot is given in the first graph in Figure 9.15.

If the residual plot indicates that the magnitude of the residuals is increasing (or decreasing) as the value of the independent variable increases, we would question the validity of the assumption of constant error variances. An example of this type of residual plot is given in the second graph in Figure 9.15. (Note that checking for increasing or decreasing magnitude in the residuals requires multiple observations on Y at the same value of X and at various levels of X .) In some cases, a simple transformation of the

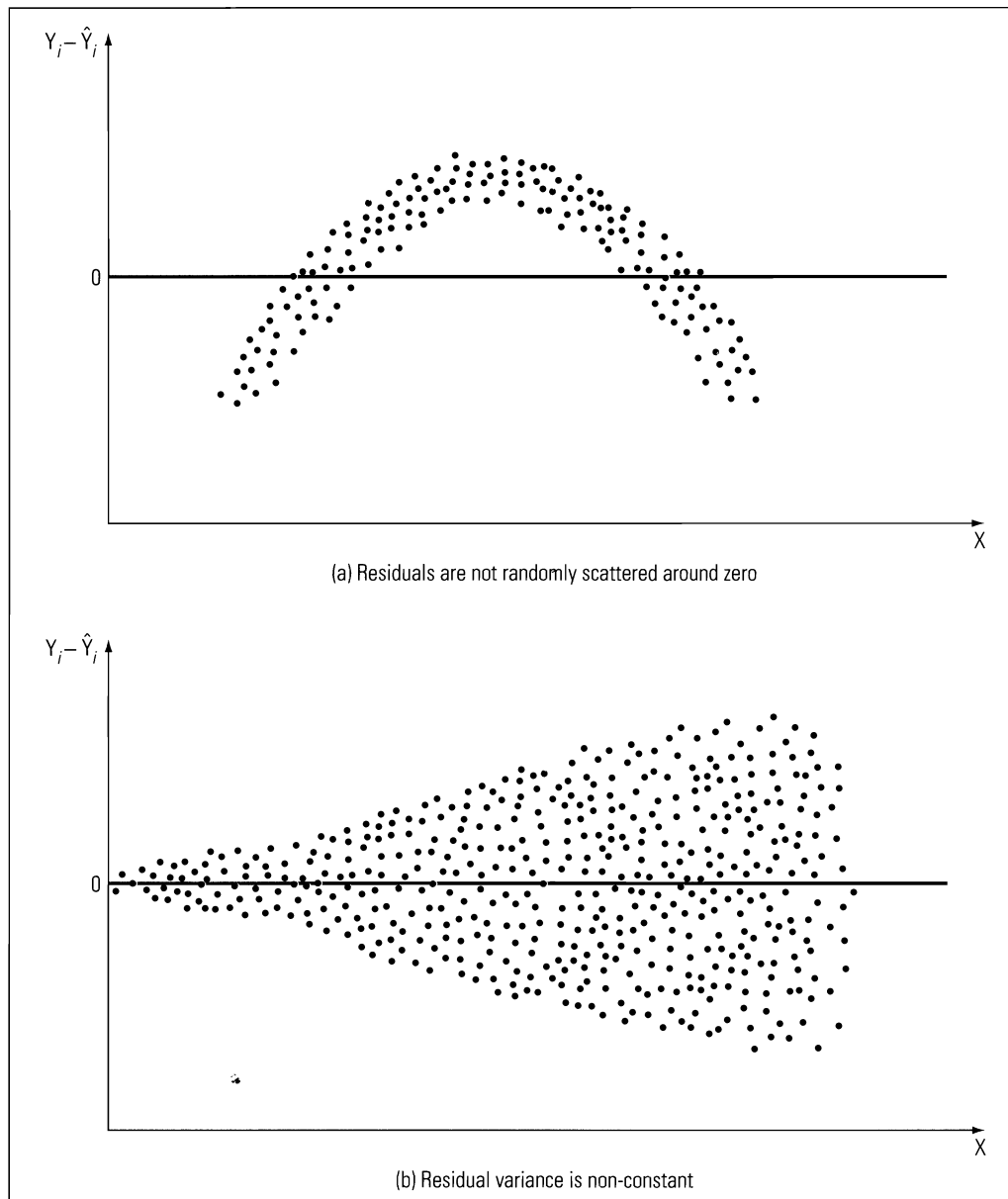


FIGURE 9.15

Residual plots indicating that the fitted regression model is not adequate

dependent variable can correct the problem of nonconstant error variances. Such transformations are discussed in more advanced texts on regression analysis.

9.10.3 A NOTE ABOUT STATISTICAL TESTS

Regardless of the form of the distribution of the error terms, least squares regression can always be used to fit regression curves to data to predict the value that the dependent variable will assume for a given level of the independent variables. Many decision makers never bother to look at residual plots or to construct confidence intervals for parameters in the regression models for the predictions they make. However, the accuracy of predictions made using regression models depends on how well the regression function fits the data. At the very least, we always should check to see how well a regression function fits a given data set. We can do so using residual plots, graphs of the actual data versus the estimated values, and the R^2 statistic.

9.11 Introduction to Multiple Regression

We have seen that regression analysis involves identifying a function that relates the *systematic* changes in a continuous dependent variable to the values of one or more independent variables. That is, our goal in regression analysis is to identify an appropriate representation of the function $f(\bullet)$ in:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon \quad 9.17$$

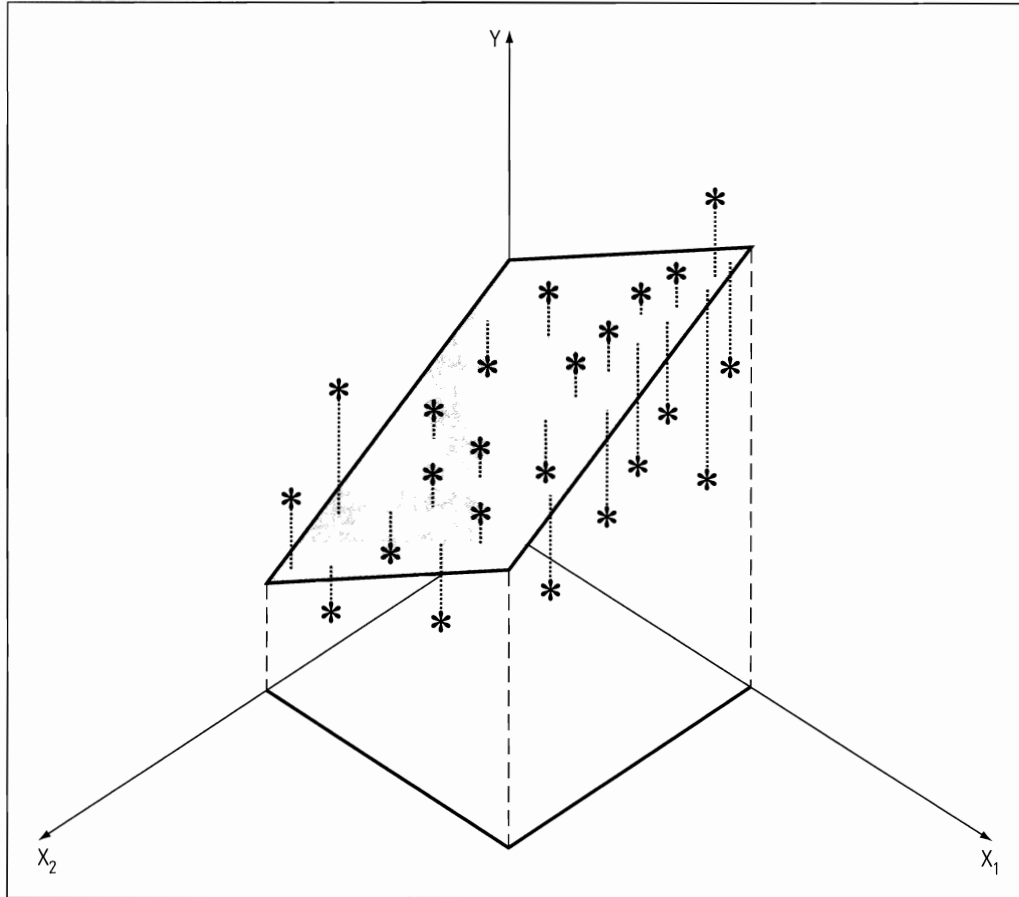
The previous sections in this chapter introduced some of the basic concepts of regression analysis by considering a special case of equation 9.17 that involves a *single* independent variable. Although such a model might be appropriate in some situations, a business person is far more likely to encounter situations involving more than one (or multiple) independent variables. We'll now consider how *multiple* regression analysis can be applied to these situations.

For the most part, multiple regression analysis is a direct extension of simple linear regression analysis. Although volumes have been written on this topic, we'll focus our attention on the multiple linear regression function represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + \dots + b_kX_{ki} \quad 9.18$$

The regression function in equation 9.18 is similar to the simple linear regression function except that it allows for more than one (or " k ") independent variables. Here again, \hat{Y}_i represents the estimated value for the i th observation in our sample whose actual value is Y_i . The symbols $X_{1i}, X_{2i}, \dots, X_{ki}$ represent the observed values of the independent variables associated with observation i . Assuming that each of these variables vary in a linear fashion with the dependent variable Y , the function in equation 9.18 might be applied appropriately to a variety of problems.

We can easily visualize the equation of a straight line in our earlier discussion of regression analysis. In multiple regression analysis, the concepts are similar but the results are more difficult to visualize. Figure 9.16 shows an example of the type of regression surface we might fit using equation 9.18 if the regression function involves only two independent variables. With two independent variables, we fit a *plane* to our data. With three or more independent variables, we fit a *hyperplane* to our data. It is difficult to visualize or draw graphs in more than three dimensions, so we cannot actually see what a hyperplane looks like. However, just as a **plane** is a generalization of a straight line into three dimensions, a **hyperplane** is a generalization of a plane into more than three dimensions.

**FIGURE 9.16**

Example of a regression surface for two independent variables

Regardless of the number of independent variables, the goal in multiple regression analysis is the same as the goal in a problem with a single independent variable. That is, we want to find the values for b_0, b_1, \dots, b_k in equation 9.18 that minimize the sum of squared estimation errors represented by:

$$ESS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

We can use the method of least squares to determine the values for b_0, b_1, \dots, b_k that minimize ESS. This should allow us to identify the regression function that best fits our data.

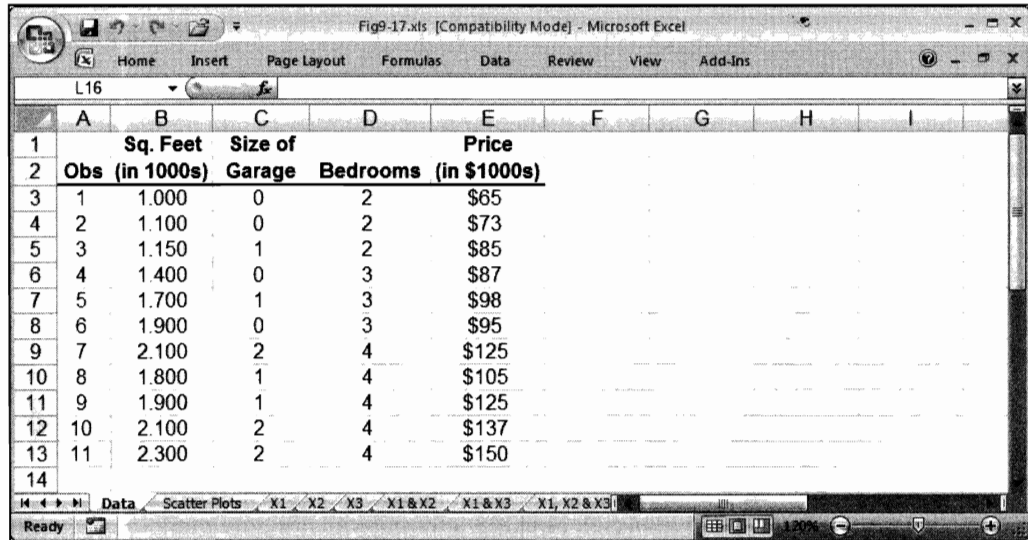
9.12 A Multiple Regression Example

The following example illustrates how to perform multiple linear regression.

A real estate appraiser is interested in developing a regression model to help predict the fair market value of houses in a particular town. She visited the county courthouse and collected the data shown in Figure 9.17 (and in the file Fig9-17.xls on your data disk). The appraiser wants to determine if the selling price of the houses can be accounted for by the total square footage of living area, the size of the garage (as measured by the number of cars that can fit in the garage), and the number of bedrooms in each house. (Note that a garage size of 0 indicates that the house has no garage.)

FIGURE 9.17

Data for the real estate appraisal problem

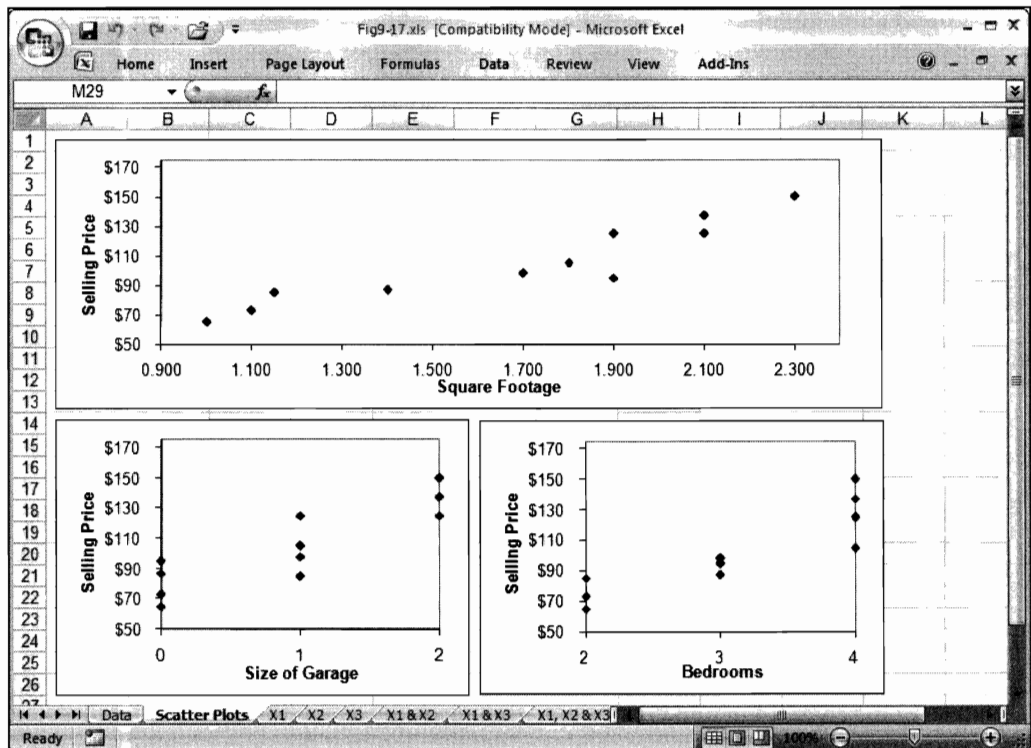


	A	B	C	D	E	F	G	H	I
1		Sq. Feet	Size of		Price				
2		Obs (in 1000s)	Garage	Bedrooms	(in \$1000s)				
3	1	1.000	0	2	\$65				
4	2	1.100	0	2	\$73				
5	3	1.150	1	2	\$85				
6	4	1.400	0	3	\$87				
7	5	1.700	1	3	\$98				
8	6	1.900	0	3	\$95				
9	7	2.100	2	4	\$125				
10	8	1.800	1	4	\$105				
11	9	1.900	1	4	\$125				
12	10	2.100	2	4	\$137				
13	11	2.300	2	4	\$150				
14									

In this example, the dependent variable Y represents the selling price of a house, and the independent variables X_1 , X_2 , and X_3 represent the total square footage, the size of the garage, and the number of bedrooms, respectively. To determine if the multiple linear regression function in equation 9.18 is appropriate for these data, we should first construct scatter plots between the dependent variable (selling price) and each independent variable, as shown in Figure 9.18. These graphs seem to indicate a

FIGURE 9.18

Scatter plots of the real estate appraisal problem



linear relationship between each independent variable and the dependent variable. Thus, we have reason to believe that a multiple linear regression function would be appropriate for these data.

9.13 Selecting the Model

In our discussion of modeling and problem-solving in Chapter 1, we noted that the best model is often the simplest model that accurately reflects the relevant characteristics of the problem being studied. This is particularly true in multiple regression models. The fact that a particular problem might involve numerous independent variables does not necessarily mean that all of the variables should be included in the regression function. If the data used to build a regression model represents a sample from a larger population of data, it is possible to over-analyze or *overfit* the data in the sample. That is, if we look too closely at a sample of data, we are likely to discover characteristics of the sample that are not representative (or which do not generalize) to the population from which the sample was drawn. This can lead to erroneous conclusions about the population being sampled. To avoid the problem of overfitting when building a multiple regression model, we should attempt to identify the *simplest* regression function that adequately accounts for the behavior of the dependent variable we are studying.

9.13.1 MODELS WITH ONE INDEPENDENT VARIABLE

With this idea of simplicity in mind, the real estate appraiser in our example problem might begin her analysis by trying to estimate the selling prices of the houses in the sample using a simple regression function with only one independent variable. The appraiser might first try to fit each of the following three simple linear regression functions to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i} \quad 9.19$$

$$\hat{Y}_i = b_0 + b_2X_{2i} \quad 9.20$$

$$\hat{Y}_i = b_0 + b_3X_{3i} \quad 9.21$$

In equations 9.19 through 9.21, \hat{Y}_i represents the estimated or fitted selling price for the i th observation in the sample, and X_{1i} , X_{2i} , and X_{3i} represent the total square footage, size of garage, and number of bedrooms for this same observation i , respectively.

To obtain the optimal values for the b_i in each regression function, the appraiser must perform three separate regressions. She would do so in the same way as described earlier in our example involving the prediction of sales from advertising expenditures. Figure 9.19 summarizes the results of these three regression functions.

The values of the R^2 statistic in Figure 9.19 indicate the proportion of the total variation in the dependent variable around its mean accounted for by each of the three simple linear regression functions. (We will comment on the adjusted- R^2 and S_e values

Independent Variable in the Model	R^2	Adjusted- R^2	S_e	Parameter Estimates
X_1	0.870	0.855	10.299	$b_0 = 9.503, b_1 = 56.394$
X_2	0.759	0.731	14.030	$b_0 = 78.290, b_2 = 28.382$
X_3	0.793	0.770	12.982	$b_0 = 16.250, b_3 = 27.607$

FIGURE 9.19

Regression results
for the three simple
linear regression
models

shortly.) The model that uses X_1 (square footage) as the independent variable accounts for 87% of the variation in Y (selling price). The model using X_2 (garage size) accounts for roughly 76% of the variation in Y , and the model that uses X_3 (number of bedrooms) as the independent variable accounts for about 79% of the variation in the selling price.

If the appraiser wants to use only one of the available independent variables in a simple linear regression model to predict the selling price of a house, it seems that X_1 would be the best choice because, according to the R^2 statistics, it accounts for more of the variation in selling price than either of the other two variables. In particular, X_1 accounts for about 87% of the variation in the dependent variable. This leaves approximately 13% of the variation in Y unaccounted for. Thus, the best linear regression function with one independent variable is represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} = 9.503 + 56.394 X_{1i} \quad 9.22$$

9.13.2 MODELS WITH TWO INDEPENDENT VARIABLES

Next, the appraiser might want to determine if one of the other two variables could be combined with X_1 in a *multiple* regression model to account for a significant portion of the remaining 13% variation in Y that was not accounted for by X_1 . To do this, the appraiser could fit each of the following multiple regression functions to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} \quad 9.23$$

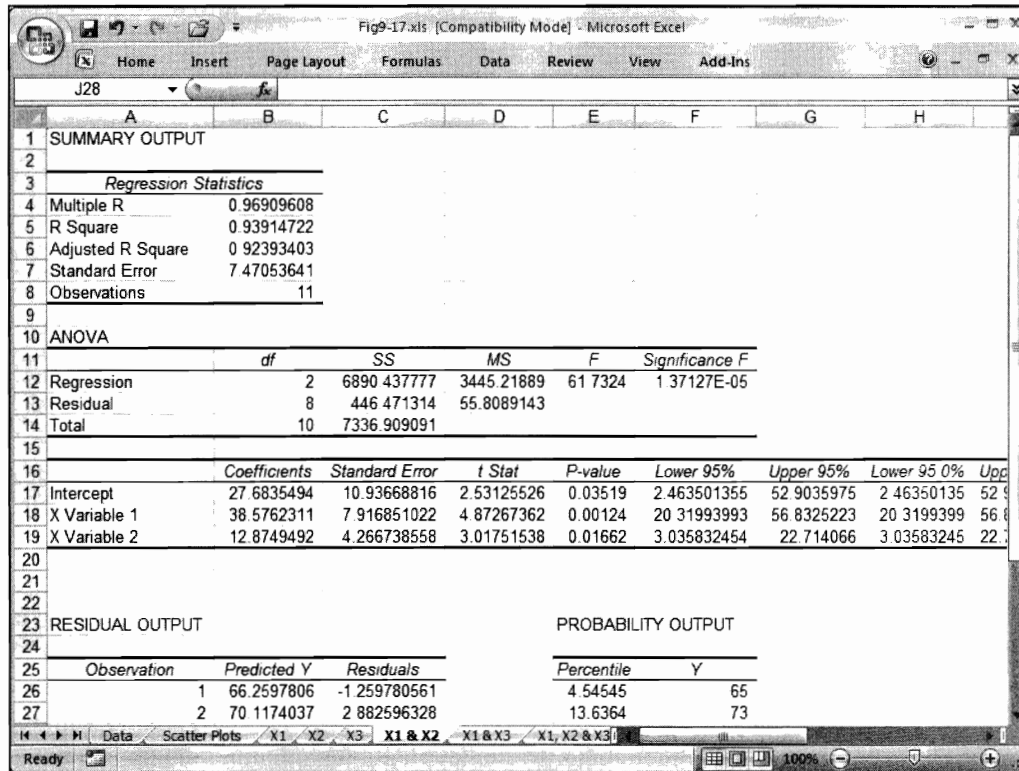
$$\hat{Y}_i = b_0 + b_1X_{1i} + b_3X_{3i} \quad 9.24$$

To determine the optimal values for the b_i in the regression model in equation 9.23, we would use the settings shown in the Regression dialog box in Figure 9.20. The Input X-Range in this dialog box is the range in Figure 9.17 that corresponds to the values for

FIGURE 9.20

Regression dialog box settings for the multiple regression model using square footage and garage size as independent variables

The screenshot shows the 'Regression' dialog box in Excel. The 'Input' section contains 'Input Y Range' set to '\$E\$3:\$E\$13' and 'Input X Range' set to '\$B\$3:\$B\$13'. There are checkboxes for 'Labels', 'Constant is Zero', and 'Confidence Level' (set to 95%). The 'Output options' section has radio buttons for 'Output Range', 'New Worksheet Ply:' (selected, with 'X1 New' in the text box), and 'New Workbook'. The 'Residuals' section has checkboxes for 'Residuals', 'Standardized Residuals', 'Residual Plots' (checked), and 'Line Fit Plots' (checked). The 'Normal Probability' section has a checked checkbox for 'Normal Probability Plots'. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

**FIGURE 9.21**

Results of the multiple regression model using square footage and garage size as independent variables

X_1 (total square footage) and X_2 (garage size). After we click the OK button, Excel performs the appropriate calculations and displays the regression results shown in Figure 9.21.

Figure 9.21 lists *three* numbers in the Coefficients column. These numbers correspond to the parameter estimates b_0 , b_1 , and b_2 . Note that the value listed for X Variable 1 is the coefficient for the first variable in the X Range (which, in some cases, might be X_2 or X_3 , depending on how the data are arranged in the spreadsheet). The value for X Variable 2 corresponds to the second variable in the X-Range (which might be X_3 or X_1 , depending on the arrangement of the data).

From the regression results in Figure 9.21, we know that when using X_1 (square footage) and X_2 (garage size) as independent variables, the estimated regression function is:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} = 27.684 + 38.576X_{1i} + 12.875X_{2i} \quad 9.25$$

Notice that adding the second independent variable caused the values of b_0 and b_1 to change from their earlier values shown in equation 9.22. Thus, the values assumed by the parameters in a regression model might vary depending on the number (and combination) of variables in the model.

We could obtain the values for the parameters in the second multiple regression model (in equation 9.24) in the same way. Note, however, that before issuing the Regression command again, we would need to rearrange the data in the spreadsheet so that the values for X_1 (total square footage) and X_3 (number of bedrooms) are located next to each other in one contiguous block. The regression tool in Excel (and in most other spreadsheet software packages) requires that the X-Range be represented by one contiguous block of cells.

FIGURE 9.22

Comparison of regression results for models with two independent variables versus the best model with one independent variable

Independent Variables in the Model	R ²	Adjusted-R ²	S _e	Parameter Estimates
X ₁	0.870	0.855	10.299	b ₀ = 9.503, b ₁ = 56.394
X ₁ and X ₂	0.939	0.924	7.471	b ₀ = 27.684, b ₁ = 38.576, b ₂ = 12.875
X ₁ and X ₃	0.877	0.847	10.609	b ₀ = 8.311, b ₁ = 44.313, b ₃ = 6.743

Important Software Note

When using the regression tool, the values for the independent variables *must* be listed in *adjacent* columns in the spreadsheet and cannot be separated by any intervening columns. That is, the Input X-Range option in the Regression dialog box must always specify a contiguous block of numbers.

Figure 9.22 compares the regression results for the model in equation 9.24 and the results for the model in equation 9.23 versus the earlier results of the best simple linear regression model in equation 9.22, where X₁ was the only independent variable in the model.

These results indicate that when using X₁ (square footage) and X₃ (number of bedrooms) as independent variables, the estimated regression function is:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_3X_{3i} = 8.311 + 44.313X_{1i} + 6.743X_{3i} \quad 9.26$$

The appraiser was hoping that the inclusion of a second independent variable in the models in equation 9.23 and equation 9.24 might help to explain a significant portion of the remaining 13% of the variation in the dependent that was not accounted for by the simple linear regression function in equation 9.22. How can we tell if this happened?

9.13.3 INFLATING R²

Figure 9.22 indicates that adding either X₂ or X₃ to the simple linear regression model caused the R² statistic to increase. This should not be surprising. As it turns out, the value of R² can never decrease as a result of adding an independent variable to a regression function. The reason for this is easy to see. From equation 9.10, recall that $R^2 = 1 - \text{ESS}/\text{TSS}$. Thus, the only way R² could decrease as the result of adding an independent variable (X_n) to the model would be if ESS *increased*. However, because the method of least squares attempts to minimize ESS, a new independent variable cannot cause ESS to increase because this variable simply could be ignored by setting b_n = 0. In other words, if adding the new independent variable does not help to reduce ESS, least squares regression simply would ignore the new variable.

When you add *any* independent variable to a regression function, the value of the R² statistic never can decrease, and usually will increase at least a little. Therefore, we can make the R² statistic arbitrarily large simply by including enough independent variables in the regression function—regardless of whether or not the new independent variables are related at all to the dependent variable. For example, the real estate appraiser probably could increase the value R² to some degree by including another

independent variable in the model that represents the height of the mailbox at each house—which probably has little to do with the selling price of a house. This results in a model that overfits our data and might not generalize well to other data not included in the sample being analyzed.

9.13.4 THE ADJUSTED- R^2 STATISTIC

The value of the R^2 statistic can be inflated artificially by including independent variables in a regression function that have little or no logical connection with the dependent variable. Thus, another goodness-of-fit measure, known as the **adjusted- R^2 statistic** (denoted by R_a^2), has been suggested which accounts for the number of independent variables included in a regression model. The adjusted- R^2 statistic is defined as:

$$R_a^2 = 1 - \left(\frac{\text{ESS}}{\text{TSS}} \right) \left(\frac{n - 1}{n - k - 1} \right) \quad 9.27$$

where n represents the number of observations in the sample, and k represents the number of independent variables in the model. As variables are added to a regression model, the ratio of ESS to TSS in equation 9.27 will decrease (because ESS decreases and TSS remains constant), but the ratio of $n - 1$ to $n - k - 1$ will increase (because $n - 1$ remains constant and $n - k - 1$ decreases). Thus, if we add a variable to the model that does not reduce ESS enough to compensate for the increase in k , the adjusted- R^2 value will decrease.

The adjusted- R^2 value can be used as a rule of thumb to help us decide if an additional independent variable enhances the predictive ability of a model or if it simply inflates the R^2 statistic artificially. However, using the adjusted- R^2 statistic in this way is not foolproof and requires a good bit of judgment on the part of the person performing the analysis.

9.13.5 THE BEST MODEL WITH TWO INDEPENDENT VARIABLES

As shown in Figure 9.22, when X_2 (garage size) is introduced to the model, the adjusted- R^2 *increases* from 0.855 to 0.924. We can conclude from this increase that the addition of X_2 to the regression model helps to account for a significant portion of the remaining variation in Y that was not accounted for by X_1 . On the other hand, when X_3 is introduced as an independent variable in the regression model, the adjusted- R^2 statistic in Figure 9.22 *decreases* (from 0.855 to 0.847). This indicates that adding this variable to the model does not help account for a significant portion of the remaining variation in Y if X_1 is already in the model. The best model with two independent variables is given in equation 9.25, which uses X_1 (total square footage) and X_2 (garage size) as predictors of selling price. According to the R^2 statistic in Figure 9.22, this model accounts for about 94% of the total variation in Y around its mean. This model leaves roughly 6% of the variation in Y unaccounted for.

9.13.6 MULTICOLLINEARITY

We should not be too surprised that no significant improvement was observed when X_3 (number of bedrooms) was added to the model containing X_1 (total square footage), because both of these variables represent similar factors. That is, the number of bedrooms in a house is closely related (or correlated) to the total square footage in the house. Thus,

if we already have used total square footage to help explain variations in the selling prices of houses (as in the first regression function), adding information about the number of bedrooms is somewhat redundant. Our analysis confirms this.

The term **multicollinearity** is used to describe the situation when the independent variables in a regression model are correlated among themselves. Multicollinearity tends to increase the uncertainty associated with the parameters estimates (b_i) in a regression model and should be avoided whenever possible. Specialized procedures for detecting and correcting multicollinearity can be found in advanced texts on regression analysis.

9.13.7 THE MODEL WITH THREE INDEPENDENT VARIABLES

As a final test, the appraiser might want to see if X_3 (number of bedrooms) helps to explain a significant portion of the remaining 6% variation in Y that was not accounted for by the model using X_1 and X_2 as independent variables. This involves fitting the following multiple regression function to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} \quad 9.28$$

Figure 9.23 shows the regression results for this model. The results of this model are also summarized for comparison purposes in Figure 9.24, along with the earlier results for the best model with one independent variable and the best model with two independent variables.

Figure 9.24 indicates that when X_3 is added to the model that contains X_1 and X_2 , the R^2 statistic increases slightly (from 0.939 to 0.943). However, the adjusted- R^2 drops from

FIGURE 9.23

Results of regression model using all three independent variables

The screenshot shows an Excel spreadsheet titled 'Fig9-17.xls [Compatibility Mode] - Microsoft Excel'. The data is organized as follows:

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.9708333
R Square	0.9425173
Adjusted R Square	0.91788186
Standard Error	7.76204445
Observations	11

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	6915.163752	2305.05458	38.2586	0.000103625
Residual	7	421.7453385	60.2493341		
Total	10	7336.909091			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	26.4403704	11.52795953	2.29358633	0.05551	-0.818922247	53.6996631	-0.81892225	53.6996631
X Variable 1	30.8034038	14.65878551	2.10136125	0.07372	-3.859115877	65.4659235	-3.85911588	65.4659235
X Variable 2	12.5674509	4.459141057	2.81835688	0.02583	2.0232578	23.111644	2.0232578	23.111644
X Variable 3	4.57596046	7.143016784	0.64062015	0.54216	-12.31459026	21.4665112	-12.3145903	21.4665112

RESIDUAL OUTPUT			PROBABILITY OUTPUT	
Observation	Predicted Y	Residuals	Percentile	Y
1	66.3956952	-1.395695157	4.54545	65

The bottom of the spreadsheet shows a row of tabs: 'Scatter Plots', 'X1', 'X2', 'X3', 'X1 & X2', 'X1 & X3', 'X1, X2 & X3', and 'Y'. The 'X1, X2 & X3' tab is currently selected.

Independent Variables in the Model	R ²	Adjusted-R ²	S _e	Parameter Estimates
X ₁	0.870	0.855	10.299	$b_0 = 9.503, b_1 = 56.394$
X ₁ and X ₂	0.939	0.924	7.471	$b_0 = 27.684, b_1 = 38.576, b_2 = 12.875$
X ₁ , X ₂ , and X ₃	0.943	0.918	7.762	$b_0 = 26.440, b_1 = 30.803, b_2 = 12.567, b_3 = 4.576$

FIGURE 9.24

Comparison of regression results for the model with three independent variables versus the best models with one and two independent variables

0.924 to 0.918. Thus, it does not appear that adding information about X₃ (number of bedrooms) helps to explain selling prices in any significant way when X₁ (total square footage) and X₂ (size of garage) are already in the model.

It is also interesting to note that the best model with two independent variables also has the smallest standard error S_e. This means that the confidence intervals around any predictions made with this model will be narrower (or more precise) than those of the other models. It can be shown that the model with the highest adjusted-R² always has the smallest standard error. For this reason, the adjusted-R² statistic is sometimes the sole criterion used to select which multiple regression model to use in a given problem. However, other procedures for selecting regression models exist and are discussed in advanced texts on regression analysis.

9.14 Making Predictions

On the basis of this analysis, the appraiser most likely would choose to use the estimated regression model in equation 9.25, which includes X₁ (total square footage) and X₂ (garage size) as independent variables. For a house with X_{1i} total square feet and space for X_{2i} cars in its garage, the estimated selling price \hat{Y}_i is:

$$\hat{Y}_i = 27.684 + 38.576X_{1i} + 12.875X_{2i}$$

For example, the expected selling price (or average market value) of a house with 2,100 square feet and a two-car garage is estimated as:

$$\hat{Y}_i = 27.684 + 38.576 \times 2.1 + 12.875 \times 2 = 134.444$$

or approximately \$134,444. Note that in making this prediction, we expressed the square footage of the house in the same units in which X₁ (total square footage variable) was expressed in the sample used to estimate the model. This should be done for all independent variables when making predictions.

The standard error of the estimation errors for this model is 7.471. Therefore, we should not be surprised to see prices for houses with 2,100 square feet and two-car garages varying within roughly ± 2 standard errors (or $\pm \$14,942$) of our estimate. That is, we expect prices on this type of house to be as low as \$119,502 or as high as \$149,386 depending on other factors not included in our analysis (such as age or condition of the roof, presence of a swimming pool, and so on).

As demonstrated earlier in the case of simple linear regression models, more accurate techniques exist for constructing prediction intervals using multiple regression models. In the case of a multiple regression model, the techniques used to construct prediction intervals require a basic knowledge of matrix algebra, which is not assumed in this text. The interested reader should consult advanced texts on multiple regression analysis for

a description of how to construct more accurate prediction intervals using multiple regression models. Keep in mind that the simple rule of thumb described earlier gives an underestimated (narrower) approximation of the more accurate prediction interval.

9.15 Binary Independent Variables

As just mentioned, the appraiser might want to include other independent variables in her analysis. Some of these, such as age of the roof, could be measured numerically and be included as an independent variable. But how would we create variables to represent the presence of a swimming pool or the condition of the roof?

The presence of a swimming pool can be included in the analysis with a binary independent variable coded as:

$$X_{p_i} = \begin{cases} 1, & \text{if house } i \text{ has a pool} \\ 0, & \text{otherwise} \end{cases}$$

The condition of the roof could also be modeled with binary variables. Here, however, we might need more than one binary variable to model all the possible conditions. If some qualitative variable can assume p possible values, we need $p - 1$ binary variables to model the possible outcomes. For example, suppose that the condition of the roof could be rated as good, average, or poor. There are three possible values for the variable representing the condition of the roof; therefore, we need two binary variables to model these outcomes. These binary variables are coded as:

$$X_{r_i} = \begin{cases} 1, & \text{if the roof of house } i \text{ is in good condition} \\ 0, & \text{otherwise} \end{cases}$$

$$X_{r+1_i} = \begin{cases} 1, & \text{if the roof of house } i \text{ is in average condition} \\ 0, & \text{otherwise} \end{cases}$$

It might appear that we left out a coding for a roof in poor condition. However, note that this condition is implied when $X_{r_i} = 0$ and $X_{r+1_i} = 0$. That is, if the roof is *not* in good condition (as implied by $X_{r_i} = 0$) *and* the roof is *not* in average condition (as implied by $X_{r+1_i} = 0$), then the roof must be in poor condition. Thus, we need only two binary variables to represent three possible roof conditions. For reasons that go beyond the scope of this text, the computer could not perform the least squares calculations if we included a third binary variable to indicate houses with roofs in poor condition. Also, it would be inappropriate to model the condition of the roof with a single variable coded as 1 for good, 2 for average, and 3 for poor because this implies that the average condition is twice as bad as the good condition, and that the poor condition is three times as bad as the good condition and 1.5 times as bad as the average condition.

As this example illustrates, we can use binary variables as independent variables in regression analysis to model a variety of conditions that are likely to occur. In each case, the binary variables would be placed in the X-Range of the spreadsheet and appropriate b_i values would be calculated by the regression tool.

9.16 Statistical Tests for the Population Parameters

Statistical tests for the population parameters in a multiple regression model are performed in much the same way as for the simple regression model. As described earlier, the F-statistic tests whether or not *all* of the β_i for *all* of the independent variables are *all* simultaneously equal to 0 ($\beta_1 = \beta_2 = \dots = \beta_k = 0$). The value in the regression results

labeled Significance of F indicates the probability of this condition being true for the data under consideration.

In the case of a multiple regression model, the t -statistics for each independent variable require a slightly different interpretation due to the possible presence of multicollinearity. Each t -statistic can be used to test whether or not the associated population parameter $\beta_i = 0$ *given all the other independent variables in the model*. For example, consider the t -statistics and p -values associated with the variable X_1 shown in Figures 9.21 and 9.23. The p -value for X_1 in cell E18 of Figure 9.21 indicates only a 0.123% chance that $\beta_1 = 0$ when X_2 is the only other independent variable in the model. The p -value for X_1 in cell E18 of Figure 9.23 indicates a 7.37% chance that $\beta_1 = 0$ when X_2 and X_3 are also in the model. This illustrates one of the potential problems caused by multicollinearity. Because X_1 and X_3 are highly correlated, it is less certain that X_1 plays a significant (nonzero) role in accounting for the behavior of the dependent variable Y when X_3 is also in the model.

In Figure 9.23, the p -value associated with X_3 indicates a 54.2% chance that $\beta_3 = 0$ given the other variables in the model. Thus, if we had started our analysis by including all three independent variables in the model, the p -value for X_3 in Figure 9.23 suggests that it might be wise to drop X_3 from the model because there is a fairly good chance that it contributes 0 ($\beta_3 = 0$) to explaining the behavior of the dependent variable, given the other variables in the model. In this case, if we drop X_3 from the model, we end up with the same model selected using the adjusted- R^2 criterion.

The statistical tests considered here are valid only when the underlying errors around the regression function are normally distributed random variables with constant means and variances. The graphical diagnostics described earlier apply equally to the case of multiple regression. However, the various statistics presented give reasonably accurate results if the assumptions about the distribution of the error terms are not violated too seriously. Furthermore, the R^2 and adjusted- R^2 statistics are purely descriptive in nature and do not depend in any way on the assumptions about the distribution of the error terms.

9.17 Polynomial Regression

When introducing the multiple linear regression function in equation 9.18, we noted that this type of model might be appropriate when the independent variables vary in a linear fashion with the dependent variable. Business problems exist where there is *not* a linear relationship between the dependent and independent variables. For example, suppose that the real estate appraiser in our earlier example had collected the data in Figure 9.25 (and in the file Fig9-25.xls on your data disk) showing the total square footage and selling price for a number of houses. Figure 9.26 shows a scatter plot of these data.

Figure 9.26 indicates a very strong relationship between total square footage and the selling price of the houses in this sample. However, this relationship is *not* linear. Rather, more of a *curvilinear* relationship exists between these variables. Does this mean that linear regression analysis cannot be used with these data? Not at all.

The data in Figure 9.25 (plotted in Figure 9.26) indicate a *quadratic* relationship between square footage and selling price. So, to account adequately for the variation in the selling price of houses, we need to use the following type of regression function:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{1i}^2 \quad 9.29$$

where \hat{Y}_i represents the estimated selling price of the i th house in our sample, and X_{1i} represents the total square footage in the house. Notice that the second independent variable in equation 9.29 is the first independent variable squared (X_1^2).

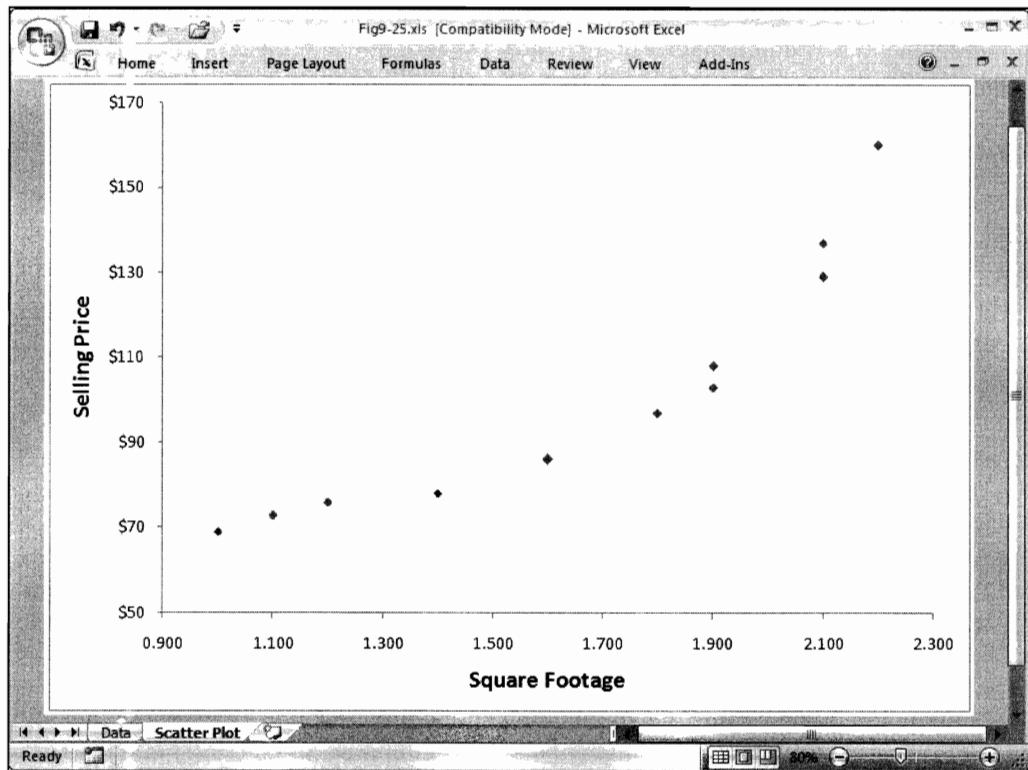
FIGURE 9.25

Data for nonlinear regression example

Obs	Sq. Feet (in 1000s)	Price (in \$1000s)
1	1.000	\$69
2	1.100	\$73
3	1.200	\$76
4	1.400	\$78
5	1.600	\$86
6	1.800	\$97
7	1.900	\$103
8	1.900	\$108
9	2.100	\$129
10	2.100	\$137
11	2.200	\$160

FIGURE 9.26

Scatter plot of data showing relationship between total square footage and selling price



9.17.1 EXPRESSING NONLINEAR RELATIONSHIPS USING LINEAR MODELS

Equation 9.29 is not a linear function because it contains the nonlinear variable X_1^2 . It is linear with respect to the parameters that the computer must estimate—namely, b_0 , b_1 , and b_2 . That is, none of the parameters in the regression function appear as an exponent

or are multiplied together. Thus, we can use least squares regression to estimate the optimal values for b_0 , b_1 , and b_2 . Note that if we define a new independent variable as $X_{2i} = X_{1i}^2$, then the regression function in equation 9.29 is equivalent to:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} \quad 9.30$$

Equation 9.30 is equivalent to the multiple linear regression function in equation 9.29. As long as a regression function is linear with respect to its parameters, we can use Excel's regression analysis tool to find the least squares estimates for the parameters.

To fit the regression function in equation 9.30 to our data, we must create a second independent variable to represent the values of X_{2i} , as shown in Figure 9.27.

Because the X-Range for the Regression command must be represented as one contiguous block, we inserted a new column between the square footage and selling price columns and placed the values of X_{2i} in this column. Note that $X_{2i} = X_{1i}^2$ in column C in Figure 9.27:

Formula for cell C3: =B3^2
(Copy to C4 through C13.)

The regression results are generated with a Y-Range of D3:D13 and an X-Range of B3:C13. Figure 9.28 shows the regression results.

In Figure 9.28, the estimated regression function is represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} = 194.9714 - 203.3812X_{1i} + 83.4063X_{2i} \quad 9.31$$

According to the R^2 statistic, this function accounts for 97.0% of the total variation in selling prices, so we expect that this function fits our data well. We can verify this by plotting the prices that would be estimated by the regression function in equation 9.31 for each observation in our sample against the actual selling prices.

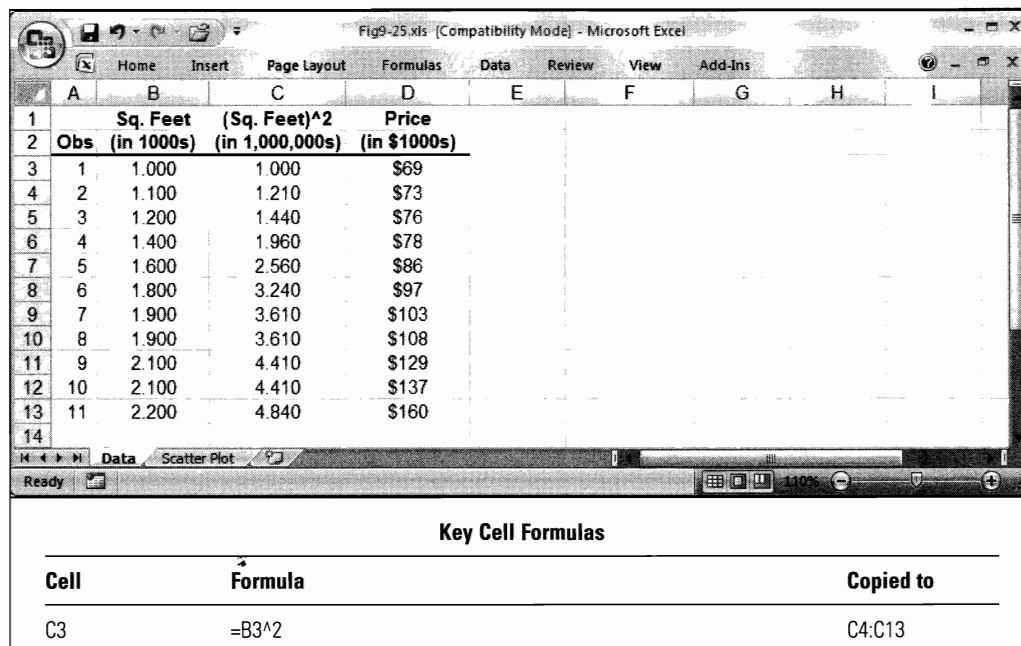


FIGURE 9.27

Modification of data to include squared independent variable

FIGURE 9.28

Regression results
for nonlinear
example problem

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.984952458								
R Square	0.970131344								
Adjusted R Square	0.96266418								
Standard Error	5.736768013								
Observations	11								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	2	8551.443215	4275.721607	129.919651	7.95908E-07				
Residual	8	263.2840579	32.91050723						
Total	10	8814.727273							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	194.9714091	34.23724921	5.69471595	0.00045716	116.020171	273.9226473	116.020171	273.9226473	
X Variable 1	-203.381237	45.08699888	-4.51086215	0.00197356	-307.3520423	-99.4104308	-307.3520423	-99.4104308	
X Variable 2	83.40635268	14.04028535	5.940502675	0.00034564	51.02939663	115.7833087	51.02939663	115.7833087	

To calculate the estimated selling prices, we applied the formula in equation 9.31 to each observation in the sample, as shown in Figure 9.29 where the following formula was entered in cell E3, then copied to cells E4 through E20:

Formula for cell E3: =TREND(\$D\$3:\$D\$13,\$B\$3:\$C\$13,B3:C3)
(Copy to E4 through E13.)

Figure 9.30 shows a curve representing the estimated prices calculated in column E of Figure 9.29. This curve was added to our previous scatter plot as follows:

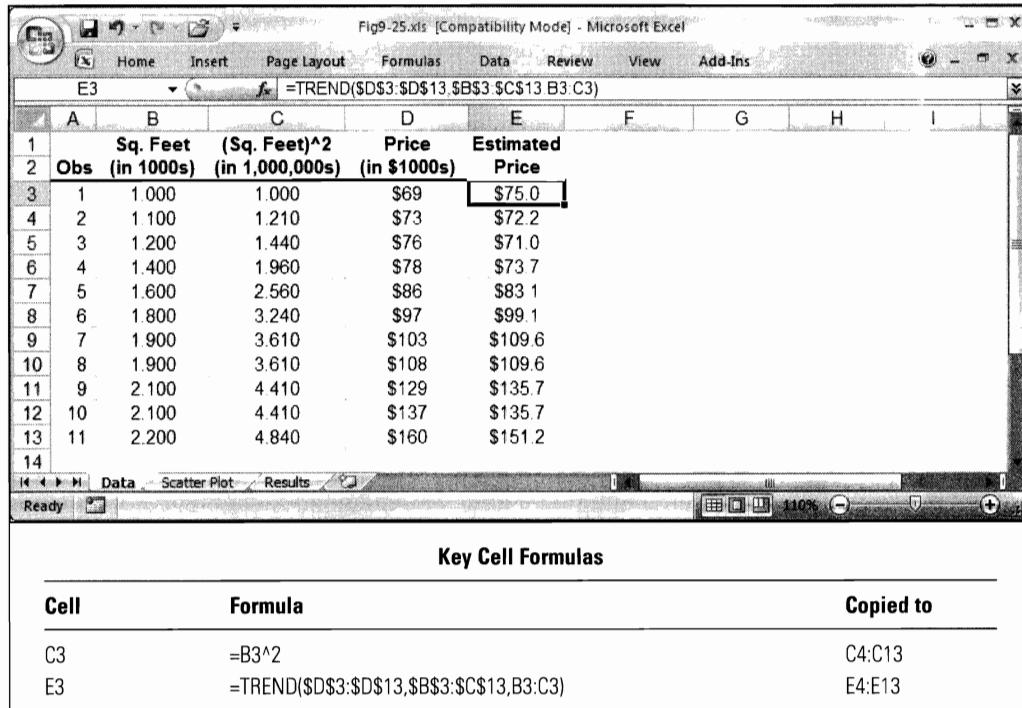
1. Right-click on any of the data points in the scatter plot to select the series of data.
2. Click Add Trendline.
3. Click Polynomial and use an Order value of 2.
4. Select Display Equation on Chart and Display R-squared Value on Chart.
5. Click Close.

This graph indicates that our regression model accounts for the nonlinear, quadratic relationship between the square footage and selling price of a house in a reasonably accurate manner.

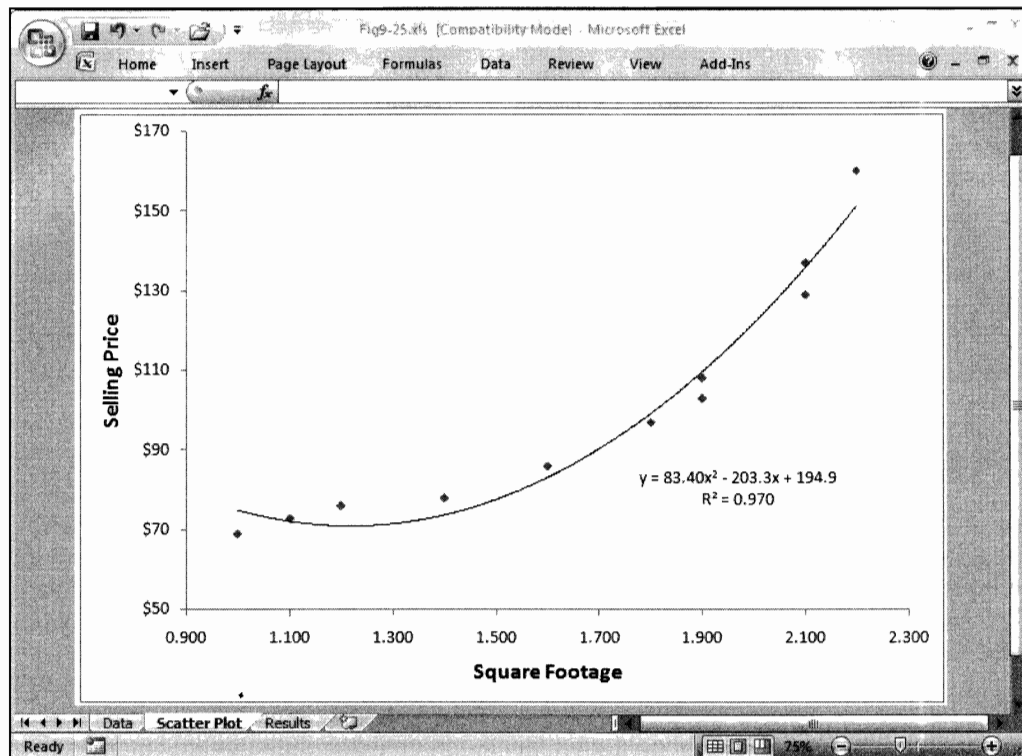
Figure 9.31 shows the result obtained by fitting a third-order polynomial model to our data of the form:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{1i}^2 + b_3X_{1i}^3 \quad 9.32$$

This model appears to provide an even better fit than the model shown in Figure 9.30. As you might imagine, we could continue to add higher order terms to the model and further increase the value of the R^2 statistic. Here again, the adjusted- R^2 statistic could help us select a model that provides a good fit to our data without overfitting the data.

**FIGURE 9.29**

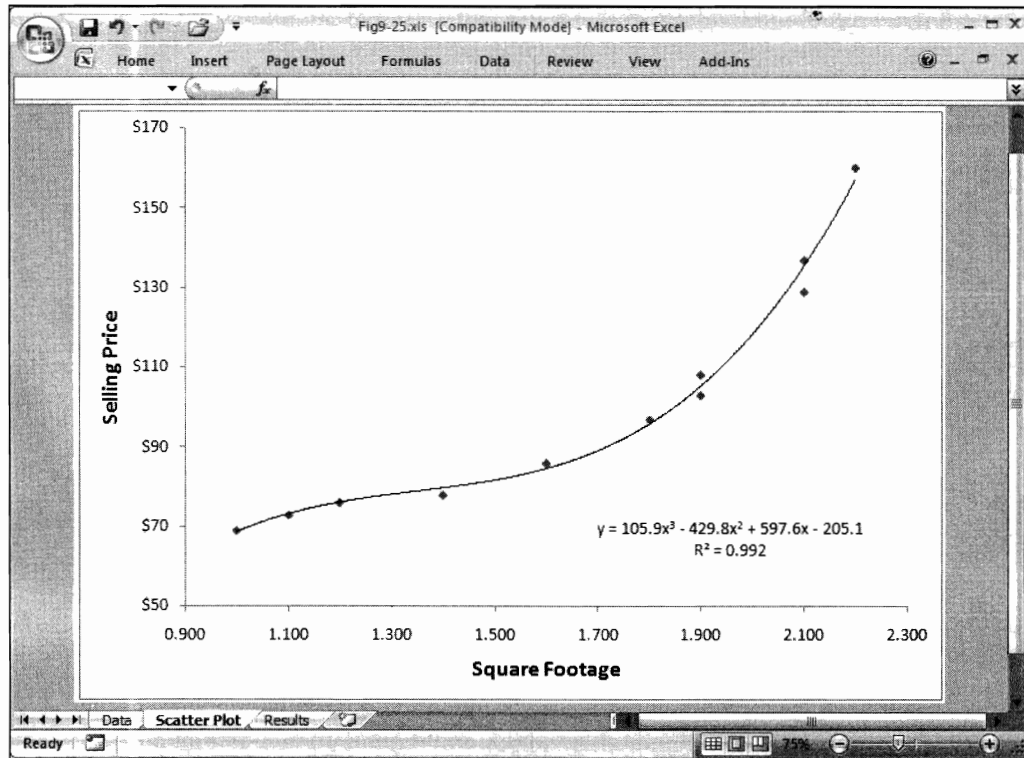
Estimated selling prices using a second order polynomial model

**FIGURE 9.30**

Plot of estimated regression function versus actual data

FIGURE 9.31

Plot of estimated regression function using a third order polynomial model



9.17.2 SUMMARY OF NONLINEAR REGRESSION

This brief example of a polynomial regression problem highlights the fact that regression analysis can be used not only in fitting straight lines or hyperplanes to linear data, but also in fitting other types of curved surfaces to nonlinear data. An in-depth discussion of nonlinear regression is beyond the intended scope of this book, but a wealth of information is available on this topic in numerous texts devoted solely to regression analysis.

This example should help you appreciate the importance of preparing scatter plots of each independent variable against the dependent variable in a regression problem to see if the relationship between the variables is linear or nonlinear. Relatively simple nonlinear relationships, such as the one described in the previous example, often can be accounted for by including squared or cubed terms in the model. In more complicated cases, sophisticated transformations of the dependent or independent variables might be required.

9.18 Summary

Regression analysis is a statistical technique that can be used to identify and analyze the relationship between one or more independent variables and a continuous dependent variable. This chapter presented an overview of some key issues involved in performing regression analysis and demonstrated some of the tools and methods available in Excel to assist managers in performing regression analysis.

The goal in regression analysis is to identify a function of the independent variables that adequately accounts for the behavior of the dependent variable. The method of least squares provides a way to determine the best values for the parameters in a regression model for a given sample of data. After identifying such a function, it can be used to predict what value the dependent variable will assume given specific values for

the independent variables. Various statistical techniques are available for evaluating how well a given regression function fits a data set and for determining which independent variables are most helpful in explaining the behavior of the dependent variable. Although regression functions can assume a variety of forms, this chapter focused on linear regression models where a linear combination of the independent variables is used to model the dependent variable. Simple transformations of the independent variables can allow this type of model to fit both linear and nonlinear data sets.

9.19 References

- Montgomery D. and E. Peck. *Introduction to Linear Regression Analysis*. New York: Wiley, 1991.
Neter, J., W. Wasserman, and M. Kutner. *Applied Linear Statistical Models*. Homewood, IL: Irwin, 1996.
Younger, M. *A First Course in Linear Regression*. Boston: Duxbury Press, 1985.

THE WORLD OF MANAGEMENT SCIENCE

Better Predictions Create Cost Savings for Ohio National Bank

The Ohio National Bank in Columbus must process checks for clearing in a timely manner, to minimize float. This had been difficult because of wide and seemingly unpredictable variations in the volume of checks received.

As checks pass through the processing center, they are encoded with the dollar amount in magnetic ink at the bottom of the check. This operation requires a staff of clerks, whose work schedules must be planned so that staffing is adequate during peak times. Because the bank could not accurately predict these peaks, deadlines often were missed and the clerks often were required to work overtime.

The variations in check volume seemed to be caused by changes in business activity brought about by the calendar—that is, volume was influenced by certain months, days of the week, days of the month, and proximity to certain holidays. A linear regression model was developed to predict staffing needs using a set of binary (dummy) independent variables representing these calendar effects. The regression study was very successful. The resulting model had a coefficient of determination (R^2) of 0.94 and a mean absolute percentage error of 6%. The bank then used these predictions as input to an LP shift-scheduling model that minimized the number of clerks needed to cover the predicted check volumes.

The planning process required data on check volumes and productivity estimates from the line supervisors in the encoding department. Initial reluctance of the supervisors to supply this information presented an obstacle to implementing the system. Eventually, this was overcome by taking time to explain the reasons for the data collection to the supervisors.

The new system provides estimated savings of \$700,000 in float costs and \$300,000 in labor costs. The close-out time of 10 pm is now met 98% of the time; previously, it was rarely met. Management has performed sensitivity analysis with the model to study the effects of productivity improvements associated with employing experienced full-time encoding clerks instead of part-time clerks.

Source: Krajewski, L. J. and L. P. Ritzman, "Shift Scheduling in Banking Operations: A Case Application." *Interfaces*, vol. 10, no. 2, April 1980, pp. 1–6.

Questions and Problems

- Members of the Roanoke Health and Fitness Club pay an annual membership fee of \$250 plus \$3 each time they use the facility. Let X denote the number of times a person visits the club during the year. Let Y denote the total annual cost for membership in the club.
 - What is the mathematical relationship between X and Y ?
 - Is this a functional or statistical relationship? Explain your answer.
- In comparing two different regression models that were developed using the same data, we might say that the model with the higher R^2 value will provide the most accurate predictions. Is this true? Why or why not?
- Show how R_a^2 and S_e are related algebraically (identify the function $f(\cdot)$ such that $R_a^2 = f(S_e)$).
- Least squares regression finds the estimated values for the parameters in a regression model to minimize $ESS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. Why is it necessary to square the estimation errors? What problem might be encountered if we attempt to minimize just the sum of the estimation errors?
- Suppose that you are interested in creating a prediction interval for Y at a particular value of X_1 (denoted by X_{1h}) using a simple linear regression model, and that data has not yet been collected. For a given sample size n , how you would attempt to collect the sample data to make the most accurate prediction? (*Hint*: Consider equation 9.14.)
- An accounting firm that specializes in auditing mining companies collected the data found in the file Dat9-6.xls on your data disk describing the long-term assets and long-term debt of its 12 clients.
 - Prepare a scatter plot of the data. Does there appear to be a linear relationship between these variables?
 - Develop a simple linear regression model that can be used to predict long-term debt from long-term assets. What is the estimated regression equation?
 - Interpret the value of R^2 .
 - Suppose that the accounting firm has a client with total assets of \$50,000,000. Construct an approximate 95% confidence interval for the amount of long-term debt that the firm expects this client to have.
- The IRS wants to develop a method for detecting whether or not individuals have overstated their deductions for charitable contributions on their tax returns. To assist in this effort, the IRS supplied data found in the file Dat9-7.xls on your data disk listing the adjusted gross income (AGI) and charitable contributions for 11 taxpayers whose returns were audited and found to be correct.
 - Prepare a scatter plot of the data. Does there appear to be a linear relationship between these variables?
 - Develop a simple linear regression model that can be used to predict the level of charitable contributions from a return's AGI. What is the estimated regression equation?
 - Interpret the value of R^2 .
 - How might the IRS use the regression results to identify returns with unusually high charitable contributions?
- Roger Gallagher owns a used car lot that deals solely in used Corvettes. He wants to develop a regression model to help predict the price he can expect to receive for the cars he owns. He collected the data found in the file Dat9-8.xls on your data disk describing the mileage, model year, presence of a T-top, and selling price of a number of cars he has sold in recent months. Let Y represent the selling price, X_1 the mileage, X_2 the model year, and X_3 the presence (or absence) of a T-top.

- a. If Roger wants to use a simple linear regression function to estimate the selling price of a car, which X variable do you recommend that he use?
- b. Determine the parameter estimates for the regression function represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$$

What is the estimated regression function? Does X_2 help to explain the selling price of the cars if X_1 is also in the model? What might be the reason for this?

- c. Set up a binary variable (X_{3i}) to indicate whether or not each car in the sample has a T-top. Determine the parameter estimates for the regression function represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_3X_{3i}$$

Does X_3 help to explain the selling price of the cars if X_1 is also in the model? Explain.

- d. According to the previous model, on average, how much does a T-top add to the value of a car?
- e. Determine the parameter estimates for the regression function represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i}$$

What is the estimated regression function?

- f. Of all the regression functions considered here, which do you recommend that Roger use?
9. Refer to question 8. Prepare scatter plots of the values of X_1 and X_2 against Y .
 - a. Do these relationships seem to be linear or nonlinear?
 - b. Determine the parameter estimates for the regression function represented by:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i}$$

where $X_{4i} = X_{2i}^2$. What is the estimated regression function?

- c. Consider the p -values for each β_i in this model. Do these values indicate that any of the independent variables should be dropped from the model?
10. Golden Years Easy Retirement Homes owns several adult care facilities throughout the southeast United States. A budget analyst for Golden Years has collected the data found in the file Dat9-10.xls on your data disk describing for each facility: the number of beds (X_1), the annual number of medical in-patient days (X_2), the total annual patient days (X_3), and whether or not the facility is in a rural location (X_4). The analyst would like to build a multiple regression model to estimate the annual nursing salaries (Y) that should be expected for each facility.
 - a. Prepare three scatter plots showing the relationship between the nursing salaries and each of the independent variables. What sort of relationship does each plot suggest?
 - b. If the budget analyst wanted to build a regression model using only one independent variable to predict the nursing salaries, what variable should be used?
 - c. If the budget analyst wanted to build a regression model using only two independent variables to predict the nursing salaries, what variables should be used?
 - d. If the budget analyst wanted to build a regression model using three independent variables to predict the nursing salaries, what variables should be used?
 - e. What set of independent variables results in the highest value for the adjusted R^2 statistic?
 - f. Suppose the personnel director chooses to use the regression function with all independent variables X_1 , X_2 and X_3 . What is the estimated regression function?

- g. In your spreadsheet, calculate an estimated annual nursing salary for each facility using the regression function identified in part f. Based on this analysis which facilities, if any, should the budget analyst be concerned about? Explain your answer.
11. The O-rings in the booster rockets on the space shuttle are designed to expand when heated to seal different chambers of the rocket so that solid rocket fuel is not ignited prematurely. According to engineering specifications, the O-rings expand by some amount, say at least 5%, to ensure a safe launch. Hypothetical data on the amount of O-ring expansion and the atmospheric temperature in Fahrenheit at the time of several different launches are given in the file Dat9-11.xls on your data disk.
- Prepare a scatter plot of the data. Does there appear to be a linear relationship between these variables?
 - Obtain a simple linear regression model to estimate the amount of O-ring expansion as a function of atmospheric temperature. What is the estimated regression function?
 - Interpret the R^2 statistic for the model you obtained.
 - Suppose that NASA officials are considering launching a space shuttle when the temperature is 29 degrees. What amount of O-ring expansion should they expect at this temperature, according to your model?
 - On the basis of your analysis of these data, would you recommend that the shuttle be launched if the temperature is 29 degrees? Why or why not?
12. An analyst for Phidelity Investments wants to develop a regression model to predict the annual rate of return for a stock based on the price-earnings (PE) ratio of the stock and a measure of the stock's risk. The data found in the file Dat9-12.xls were collected for a random sample of stocks.
- Prepare scatter plots for each independent variable versus the dependent variable. What type of model do these scatter plots suggest might be appropriate for the data?
 - Let Y = Return, X_1 = PE Ratio, and X_2 = Risk. Obtain the regression results for the following regression model:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$$

Interpret the value of R^2 for this model.

- Obtain the regression results for the following regression model:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i}$$

where $X_{3i} = X_{1i}^2$ and $X_{4i} = X_{2i}^2$. Interpret the value of R^2 for this model.

- Which of the previous two models would you recommend that the analyst use?
13. Oriented Strand Board (OSB) is manufactured by gluing woodchips together to form panels. Several panels are then bonded together to form a board. One of the factors influencing the strength of the final board is the amount of glue used in the production process. An OSB manufacturer conducted a test to determine the breaking point of a board based on the amount of glue used in the production process. In each test, a board was manufactured using a given amount of glue. Weight was then applied to determine the point at which the board would fail (or break). This test was performed 27 times using various amounts of glue. The data obtained from this testing may be found in the file Dat9-13.xls on your data disk.
- Prepare a scatter plot of this data.
 - What type of regression function would you use to fit this data?

- c. Estimate the parameters of the regression function. What is the estimated regression function?
 - d. Interpret the value of the R^2 statistic.
 - e. Suppose the company wants to manufacture boards that will withstand up to 110 lbs. of pressure per square inch. How much glue should they use?
14. When interest rates decline, Patriot Bank has found that they get inundated with requests to refinance home mortgages. To better plan its staffing needs in the mortgage processing area of its operations, Patriot wants to develop a regression model to help predict the total number of mortgage applications (Y) each month as a function of the prime interest rate (X_1). The bank collected the data shown in the file Dat9-14.xls on your data disk representing the average prime interest rate and total number of mortgage applications in 20 different months.
- a. Prepare a scatter plot of these data.
 - b. Fit the following regression model to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i}$$

Plot the number of monthly mortgage applications that are estimated by this model along with the actual values in the sample. How well does this model fit the data?

- c. Using the previous model, develop a 95% prediction interval for the number of mortgage applications Patriot could expect to receive in a month in which the interest rate is 6%. Interpret this interval.
 - d. Fit the following regression model to the data:
- $$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$$
- where $X_{2i} = X_{1i}^2$. Plot the number of monthly mortgage applications that are estimated by this model along with the actual values in the sample. How well does this model fit the data?
- e. Using the previous model, develop a 95% prediction interval for the number of mortgage applications that Patriot could expect to receive in a month in which the interest rate is 6%. Interpret this interval.
 - f. Which model would you suggest that Patriot Bank use, and why?
15. Creative Confectioners is planning to introduce a new brownie. A small-scale “taste test” was conducted to assess consumers’ preferences (Y) with regard to moisture content (X_1) and sweetness (X_2). Data from the taste test may be found in the file Dat9-15.xls on your data disk.

- a. Prepare a scatter plot of moisture content versus preference. What type of relationship does your plot suggest?
- b. Prepare a scatter plot of sweetness versus preference. What type of relationship does your plot suggest?
- c. Estimate the parameters for the following regression function:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{1i}^2 + b_3X_{2i} + b_4X_{2i}^2$$

What is the estimated regression function?

- d. Using the estimated regression function in part c, what is the expected preference rating of a brownie recipe with a moisture content of 7 and a sweetness rating of 9.5?
16. AutoReports is a consumer magazine that reports on the cost of maintaining various types of automobiles. The magazine collected the data found in the file Dat9-16.xls on your data disk describing the annual maintenance cost of a certain type of luxury imported automobile along with the age of the car (in years).

- a. Prepare a scatter plot of these data.
- b. Let Y = Maintenance Cost and X = Age. Fit the following regression model to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i}$$

Plot the maintenance costs that are estimated by this model along with the actual costs in the sample. How well does this model fit the data?

- c. Fit the following regression model to the data:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$$

where $X_{2i} = X_{1i}^2$. Plot the maintenance costs that are estimated by this model along with the actual costs in the sample. How well does this model fit the data?

- d. Fit the following regression model to this data:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i}$$

where $X_{2i} = X_{1i}^2$ and $X_{3i} = X_{1i}^3$. Plot the maintenance costs that are estimated by this model along with the actual costs in the sample. How well does this model fit the data?

17. Duque Power Company wants to develop a regression model to help predict its daily peak power demand. This prediction is useful in determining how much generating capacity needs to be available (or purchased from competitors) on a daily basis. The daily peak power demand is influenced primarily by the weather and the day of the week. The file Dat9-17.xls on your data disk contains data summarizing Duque's daily peak demand and maximum daily temperature during the month of July last year.
 - a. Build a simple linear regression model to predict peak power demand using maximum daily temperature. What is the estimated regression equation?
 - b. Prepare a line chart plotting the actual peak demand data against the values predicted by this regression equation. How well does the model fit the data?
 - c. Interpret the R^2 statistic for this model.
 - d. Build a multiple linear regression model to predict peak power demand using maximum daily temperature and the day of the week as independent variables. (Note: This model will have seven independent variables.) What is the estimated regression equation?
 - e. Prepare a line chart plotting the actual peak demand data against the values predicted by this regression equation. How well does the model fit the data?
 - f. Interpret the R^2 statistic for this model.
 - g. Using the model you developed in part d above, what is the estimated peak power demand Duque should expect on a Wednesday in July when the daily high temperature is forecast to be 94?
 - h. Compute a 95% prediction interval for the estimate in the previous question. Explain the managerial implications of this interval for Duque.
18. An appraiser collected the data found in file Dat9-18.xls describing the auction selling price, diameter (in inches), and item type of several pieces of early 20th century metal tableware manufactured by a famous artisan. The item type variable is coded as follows: B=bowl, C=casserole pan, D=dish, T=tray, and P=plate. The appraiser wants to build a multiple regression model for this data to predict average selling prices of similar items.
 - a. Construct a multiple regression model for this problem. (Hint: Create binary independent variables to represent the item type data). What is the estimated regression function?

- b. Interpret the value of the R^2 statistic for this model.
 - c. Construct an approximate 95% prediction interval for the expected selling price of an 18-inch diameter casserole pan. Interpret this interval.
 - d. What other variables not included in the model might help explain the remaining variation in auction selling prices for these items?
19. The personnel director for a small manufacturing company has collected the data found in the file Dat9-19.xls on your data disk describing the salary (Y) earned by each machinist in the factory along with the average performance rating (X_1) over the past 3 years, the years of service (X_2), and the number of different machines each employee is certified to operate (X_3).

The personnel director wants to build a regression model to estimate the average salary an employee should expect to receive based on his or her performance, years of service, and certifications.

- a. Prepare three scatter plots showing the relationship between the salaries and each of the independent variables. What sort of relationship does each plot suggest?
 - b. If the personnel director wanted to build a regression model using only one independent variable to predict the salaries, what variable should be used?
 - c. If the personnel director wanted to build a regression model using only two independent variables to predict the salaries, what two variables should be used?
 - d. Compare the adjusted- R^2 statistics obtained in parts b and c with that of a regression model using all three independent variables. Which model would you recommend that the personnel director use?
 - e. Suppose the personnel director chooses to use the regression function with all three independent variables. What is the estimated regression function?
 - f. Suppose the company considers an employee's salary to be "fair" if it is within 1.5 standard errors of the value estimated by the regression function in part e. What salary range would be appropriate for an employee with 12 years of service, who has received average reviews of 4.5, and is certified to operate 4 pieces of machinery?
20. **Caveat Emptor, Inc. is a home inspection service that provides prospective home buyers with a thorough assessment of the major systems in a house prior to the execution of the purchase contract. Prospective home buyers often ask the company for an estimate of the average monthly heating cost of the home during the winter. To answer this question, the company wants to build a regression model to help predict the average monthly heating cost (Y) as a function of the average outside temperature in winter (X_1), the amount of attic insulation in the house (X_2), the age of the furnace in the house (X_3), and the size of the house measured in square feet (X_4). Data on these variables for a number of homes was collected and may be found in the file Dat9-20.xls on your data disk.**
- a. Prepare scatter plots showing the relationship between the average heating cost and each of the potential independent variables. What sort of relationship does each plot suggest?
 - b. If the company wanted to build a regression model using only one independent variable to predict the average heating cost of these houses, what variable should be used?
 - c. If the company wanted to build a regression model using only two independent variables to predict the average heating cost of these houses, what variables should be used?
 - d. If the company wanted to build a regression model using only three independent variables to predict the average heating cost of these houses, what variables should be used?

- e. Suppose the company chooses to use the regression function with all four independent variables. What is the estimated regression function?
 - f. Suppose the company decides to use the model with the highest adjusted R^2 statistic. Develop a 95% prediction interval for the average monthly heating cost of a house with 4 inches of attic insulation, a 5-year-old furnace, 2500 square feet, and in a location with an average outside winter temperature of 40. Interpret this interval.
21. Throughout our discussion of regression analysis, we used the Regression command to obtain the parameter estimates that minimize the sum of squared estimation errors. Suppose that we want to obtain parameter estimates that minimize the sum of the absolute value of the estimation errors, or:

$$\text{MIN: } \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

- a. Use Solver to obtain the parameter estimates for a simple linear regression function that minimizes the sum of the absolute value of the estimation errors for the data in question 9.
 - b. What advantages, if any, do you see in using this alternate objective to solve a regression problem?
 - c. What disadvantages, if any, do you see in using this alternate objective to solve a regression problem?
22. Throughout our discussion of regression analysis, we used the Regression command to obtain the parameter estimates that minimize the sum of squared estimation errors. Suppose that we want to obtain parameter estimates that minimize the absolute value of the maximum estimation error, or:

$$\text{MIN: MAX } (|Y_1 - \hat{Y}_1|, |Y_2 - \hat{Y}_2|, \dots, |Y_n - \hat{Y}_n|)$$

- a. Use Solver to obtain the parameter estimates for a simple linear regression function that minimizes the absolute value of the maximum estimation error for the data in question 9.
- b. What advantages, if any, do you see in using this alternate objective to solve a regression problem?
- c. What disadvantages, if any, do you see in using this alternate objective to solve a regression problem?

CASE 9.1

Diamonds Are Forever

(Inspired from actual events related by former Virginia Tech MBA student Brian Ellyson.)

With Christmas coming, Ryan Bellison was searching for the perfect gift for his wife. Ryan leaned back in his chair at the office and tried to think, after several years of marriage, of the one thing his wife had wanted during the years they pinched pennies to get through graduate school. Then he remembered the way her eyes had lit up last week when they walked by the jewelry store windows at the mall and she saw the diamond earrings. He knew he wanted to see that same look on her face Christmas morning. And so his hunt began for the perfect set of diamond earrings.

Ryan's first order of business was to educate himself about the things to look for when buying diamonds. After perusing the Web, he learned about the "4Cs" of diamonds: cut, color, clarity, and carat (see: <http://www.adiamondisforever.com>). He knew his wife wanted round-cut earrings mounted in white gold settings, so he

immediately narrowed his focus to evaluating color, clarity, and carat for that style earring.

After a bit of searching, Ryan located a number of earring sets that he would consider purchasing. But he knew the pricing of diamonds varied considerably and he wanted to make sure he didn't get ripped off. To assist in his decision making, Ryan decided to use regression analysis to develop a model to predict the retail price of different sets of round-cut earrings based on their color, clarity, and carat scores. He assembled the data in the file *Diamonds.xls* on your data disk for this purpose. Use this data to answer the following questions for Ryan.

- a. Prepare scatter plots showing the relationship between the earring prices (Y) and each of the potential independent variables. What sort of relationship does each plot suggest?
- b. Let X_1 , X_2 , and X_3 represent diamond color, clarity, and carats, respectively. If Ryan wanted to build a linear regression model to estimate earring prices using these variables, which variables would you recommend that he use? Why?
- c. Suppose Ryan decides to use clarity (X_2) and carats (X_3) as independent variables in a regression model to predict earring prices. What is the estimated regression equation? What is the value of the R^2 and adjusted- R^2 statistics?
- d. Use the regression equation identified in the previous question to create estimated prices for each of the earring sets in Ryan's sample. Which sets of earrings appear to be overpriced and which appear to be bargains? Based on this analysis, which set of earrings would you suggest that Ryan purchase?
- e. Ryan now remembers that it is sometimes helps to perform a square root transformation on the dependent variable in a regression problem. Modify your spreadsheet to include a new dependent variable that is the square root on the earring prices (use Excel's $\text{SQRT}()$ function). If Ryan wanted to build a linear regression model to estimate the square root of earring prices using the same independent variables as before, which variables would you recommend that he use? Why?
- f. Suppose Ryan decides to use clarity (X_2) and carats (X_3) as independent variables in a regression model to predict the square root of the earring prices. What is the estimated regression equation? What is the value of the R^2 and adjusted- R^2 statistics?
- g. Use the regression equation identified in the previous question to create estimated prices for each of the earring sets in Ryan's sample. (Remember, your model estimates the square root of the earring prices. So you must square the model's estimates to convert them to actually price estimates.) Which sets of earring appears to be overpriced and which appear to be bargains? Based on this analysis, which set of earrings would you suggest that Ryan purchase?
- h. Ryan now also remembers that it sometimes helps to include interaction terms in a regression model—where you create a new independent variable as the product of two of the original variables. Modify your spreadsheet to include three new independent variables, X_4 , X_5 , and X_6 , representing interaction terms where: $X_4 = X_1 \times X_2$, $X_5 = X_1 \times X_3$, and $X_6 = X_2 \times X_3$. There are now six potential independent variables. If Ryan wanted to build a linear regression model to estimate the square root of earring prices using the same independent variables as before, which variables would you recommend that he use? Why?
- i. Suppose Ryan decides to use clarity (X_1), carats (X_3) and the interaction terms X_4 and X_5 as independent variables in a regression model to predict the square root of the earring prices. What is the estimated regression equation? What is the value of the R^2 and adjusted- R^2 statistics?

- j. Use the regression equation identified in the previous question to create estimated prices for each of the earring sets in Ryan's sample. (Remember, your model estimates the square root of the earring prices. So you must square the model's estimates to convert them to actual price estimates.) Which sets of earrings appear to be overpriced and which appear to be bargains? Based on this analysis, which set of earrings would you suggest that Ryan purchase?

CASE 9.2

Fiasco in Florida

The 2000 U.S. Presidential election was one of the most controversial in history, with the final outcome ultimately being decided in a court of law rather than in the voting booth. At issue were the election results in Palm Beach, Florida. Palm Beach County used a so-called "butterfly" ballot where the candidates' names were arranged to the left and right of a center row of holes. Voters were to specify their preference by "punching" the appropriate hole next to the desired candidate. According to several news accounts, many voters in Palm Beach, Florida, claimed that they were confused by the ballot structure and might have voted inadvertently for Pat Buchanan when in fact they intended to vote for Al Gore. This allegedly contributed to Gore not obtaining enough votes to overtake George Bush's slim margin of victory in Florida—and ultimately cost Gore the election.

The file *Votes.xls* on your data disk contains the original vote totals by Florida county for Gore, Bush, and Buchanan as of November 8, 2000. (These data reflect the results prior to the hand recount that was done due to other problems with the election in Florida (e.g., the "hanging chad" problem.)) Use the data in this file to answer the following questions.

- What was George Bush's margin of victory in Florida?
- Prepare a scatter plot showing the relationship between the number of votes received by Gore (X-axis) and Buchanan (Y-axis) in each county. Do there appear to be any outliers? If so, for what counties?
- Estimate the parameters for a simple linear regression model for predicting the number of votes for Buchanan in each county (excluding Palm Beach County) as a function of the number of votes for Gore. What is the estimated regression equation?
- Interpret the value for R^2 obtained using the equation from question 3.
- Using the regression results from question 3, develop a 99% prediction interval for the number of votes you expect Buchanan to receive in Palm Beach County. What are the upper and lower limits of that interval? How does this compare with the actual number of votes reported for Buchanan in Palm Beach County?
- Prepare a scatter plot showing the relationship between the number of votes received by Bush (X-axis) and Buchanan (Y-axis) in each county. Do there appear to be any outliers? If so, for what counties?
- Estimate the parameters for a simple linear regression model for predicting the number of votes for Buchanan in each county (excluding Palm Beach County) as a function of the number of votes for Bush. What is the estimated regression equation?
- Interpret the value for R^2 obtained using the equation from question 7.
- Using the regression results from question 7, develop a 99% prediction interval for the number of votes you expect Buchanan to receive in Palm Beach County. What are the upper and lower limits of that interval? How does this compare with the actual number of votes reported for Buchanan in Palm Beach County?
- What do these results suggest? What assumptions are being made by using regression analysis in this way?

The Georgia Public Service Commission

CASE 9.3

(Inspired by discussions with Mr. Nolan E. Ragsdale of Banks County, Georgia.)

Nolan Banks is an auditor for the Public Service Commission for the state of Georgia. The Public Service Commission is a government agency responsible for ensuring that utility companies throughout the state manage their operations efficiently so that they can provide quality services to the public at fair prices.

Georgia is the largest state east of the Mississippi River, and various communities and regions throughout the state have different companies that provide water, power, and phone service. These companies have a monopoly in the areas they serve and, therefore, could take unfair advantage of the public. One of Nolan's jobs is to visit the companies and audit their financial records to detect whether or not any abuse is occurring.

A major problem Nolan faces in his job is determining whether the expenses reported by the utility companies are reasonable. For example, when he reviews a financial report for a local phone company, he might see line maintenance costs of \$1,345,948, and he needs to determine if this amount is reasonable. This determination is complicated because companies differ in size—so he cannot compare the costs of one company directly to those of another. Similarly, he cannot come up with a simple ratio to determine costs (such as 2% for the ratio of line maintenance costs to total revenue) because a single ratio might not be appropriate for companies of different sizes.

To help solve this problem, Nolan wants you to build a regression model to estimate what level of line maintenance expense would be expected for companies of different sizes. One measure of size for a phone company is the number of customers it has. Nolan collected the data in the file PhoneService.xls on your data disk representing the number of customers and line maintenance expenses of 12 companies that he audited in the past year and determined were being run in a reasonably efficient manner.

- Enter the data in a spreadsheet.
- Create a scatter diagram of these data.
- Use regression to estimate the parameters for the following linear equation for the data.

$$\hat{Y} = b_0 + b_1X_1$$

What is the estimated regression equation?

- Interpret the value for R^2 obtained using the equation from question 3.
- According to the equation in question 3, what level of line maintenance expense would be expected for a phone company with 75,000 customers? Show how you arrive at this value.
- Suppose that a phone company with 75,000 customers reports a line maintenance expense of \$1,500,000. Based on the results of the linear model, should Nolan view this amount as reasonable or excessive?
- In your spreadsheet, calculate the estimated line maintenance expense that would be predicted by the regression function for each company in the sample. Plot the predicted values you calculate on your graph (connected with a line) along with the original data. Does it appear that a linear regression model is appropriate?
- Use regression to estimate the parameters for the following quadratic equation for the data:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_1^2$$

To do this, you must insert a new column in your spreadsheet next to the original X values. In this new column, calculate the values X_1^2 . What is the new estimated regression equation for this model?

- i. Interpret the value for R^2 obtained using the equation in question 8.
- j. What is the value for the adjusted- R^2 statistic? What does this statistic tell you?
- k. What level of line maintenance expense would be expected for a phone company with 75,000 customers according to this new estimated regression function? Show how you arrive at this value.
- l. In your spreadsheet, calculate the estimated line maintenance expense that would be predicted by the quadratic regression function for each company in the sample. Plot these values on your graph (connected with a line) along with the original data and the original regression line.
- m. Suppose that a phone company with 75,000 customers reports a line maintenance expense of \$1,500,000. Based on the results of the quadratic model, should Nolan view this amount as reasonable or excessive?
- n. Which of the two regression functions would you suggest that Nolan use for prediction purposes?