Chapter FIFTEEN

Multiple Regression Analysis and Model Building

CHAPTER OUTCOMES

After studying the material in Chapter 15, you should be able to:

1. Understand the general concepts behind model building using multiple regression analysis.
2. Apply multiple regression analysis to business decision-making situations.
3. Analyze the computer output for a multiple regression model and interpret the regression results.
4. Test hypotheses about the significance of a multiple regression model and test the significance of the independent variables in the model.
5. Recognize potential problems when using multiple regression analysis and take steps to correct the problems.
6. Incorporate qualitative variables into a regression model by using dummy variables.
7. Apply regression analysis to situations where the relationship between the independent variable(s) and the dependent variable is nonlinear.
8. Understand the uses of stepwise regression.

PREPARING FOR CHAPTER FIFTEEN

- Review the methods for testing a null hypothesis using the *t*-distribution in Chapter 9.
- Review confidence intervals discussed in Chapter 8.
- Make sure you review the discussion about scatter plots in Chapters 2 and 14.
- Review the concepts associated with simple linear regression and correlation analysis presented in Chapter 14.
- In Chapter 14, review the steps involved in using the *t*-distribution for testing the significance of a correlation coefficient and a regression coefficient.

WHY YOU NEED TO KNOW

Chapter 14 pointed out that decision-makers often need to consider the relationship between two variables when analyzing a problem. Simple linear regression and correlation analyses provide a basis for analyzing the relationship between two variables. If the two variables are correlated, there is a linear relationship between them, and linear regression analysis can be used to model that relationship.

As you might expect, business problems are not limited to linear relationships involving only two variables. Many practical situations involve analyzing the relationships among three or more variables, and these relationships may be nonlinear. For example, a vice president of planning for an automobile manufacturer would be interested in the relationship between her company's automobile sales and the variables that influence those sales. Included in her analysis might be such independent or explanatory variables as automobile price, competitors' sales, and advertising, as well as such economic variables as disposable personal income, the inflation rate, and the unemployment rate.

When multiple independent variables are to be included in an analysis simultaneously, the technique introduced in this chapter—multiple linear regression—is very useful. When a relationship between variables is nonlinear, we may be able to apply variable transformations that allow us to use multiple linear regression analysis to construct a model. This chapter examines the general topic of model building by extending the concepts of simple linear regression analysis. The background information provided in Chapter 14 will be very helpful in understanding and applying multiple regression analysis to business decision-making situations.

15.1 Introduction to Multiple Regression Analysis

Chapter 14 introduced the concept of simple linear regression analysis. The simple regression model is characterized by two variables: y, the *dependent variable*, and x, the *independent*, or *explanatory*, *variable*. The single independent variable explains some variation in the dependent variable, but unless x and y are perfectly correlated, the proportion explained will be less than 100%.

In multiple regression analysis, additional independent variables are added to the regression model to explain some of the yet-unexplained variation in the dependent variable. You will note as we proceed that multiple regression is merely an extension of simple regression analysis. However, as we expand the model for the population from one independent variable to two or more, there are some new considerations.

The general format of a *multiple regression model for the population* is given by Equation 15.1.

Multiple Regression Model (Population Model)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (15.1)$$

where:

β_0 = Population's regression constant
β_j = Population's regression coefficient for each variable x_j; = 1, 2, . . . , k
k = Number of independent variables
ε = Model error

Four assumptions similar to those that apply to the simple linear multiple regression model apply to the multiple regression model.

Assumptions

1. Individual residuals, ε, are statistically independent of one another, and these values represent a random sample from the population of possible residuals at each level of x.
2. For a given value of x, there can exist many values of y, and therefore many possible values for ε. Further, the distribution of possible ε-values for any level of x is normally distributed.
3. The distributions of possible ε-values have equal variances at each level of x.
4. The means of the dependent variable, y, for all specified values of x can be connected with a line called the population regression model.

TABLE 15.1 Sample Data to Illustrate the Difference Between Simple and Multiple Regression Models

(A) One Independent Variable		*(B) Two Independent Variable*		
y	x_1	y	x_1	x_2
564.99	50	564.99	50	10
601.06	60	601.06	60	13
560.11	40	560.11	40	14
616.41	50	616.41	50	12
674.96	60	674.96	60	15
630.58	45	630.58	45	16
554.66	53	554.66	53	14

Equation 15.1 represents the multiple regression model for the population. However, in most instances, you will be working with a random sample from the population. Given the preceding assumptions, the estimated multiple regression model, based on the sample data, is of the form shown in Equation 15.2.

Estimated Multiple Regression Model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \quad \textbf{(15.2)}$$

This estimated model is an extension of an estimated simple regression model. The principal difference is that, whereas the estimated simple regression model is the equation for a straight line in a two-dimensional space, the estimated multiple regression model forms a hyperplane (or response surface) through multidimensional space. Each regression coefficient represents a different slope. Therefore, for a decision maker, using Equation 15.2, a value of the dependent variable can be estimated using values of two or more independent variables. The **regression hyperplane** represents the relationship between the dependent variable and the k independent variables.

Regression Hyperplane
The multiple regression equivalent of the simple regression line. The plane typically has a different slope for each independent variable.

For example, Table 15.1a shows sample data for a dependent variable, y, and one independent variable, x_1. Figure 15.1 shows a scatter plot and the regression line for the simple

FIGURE 15.1
Simple Regression Line

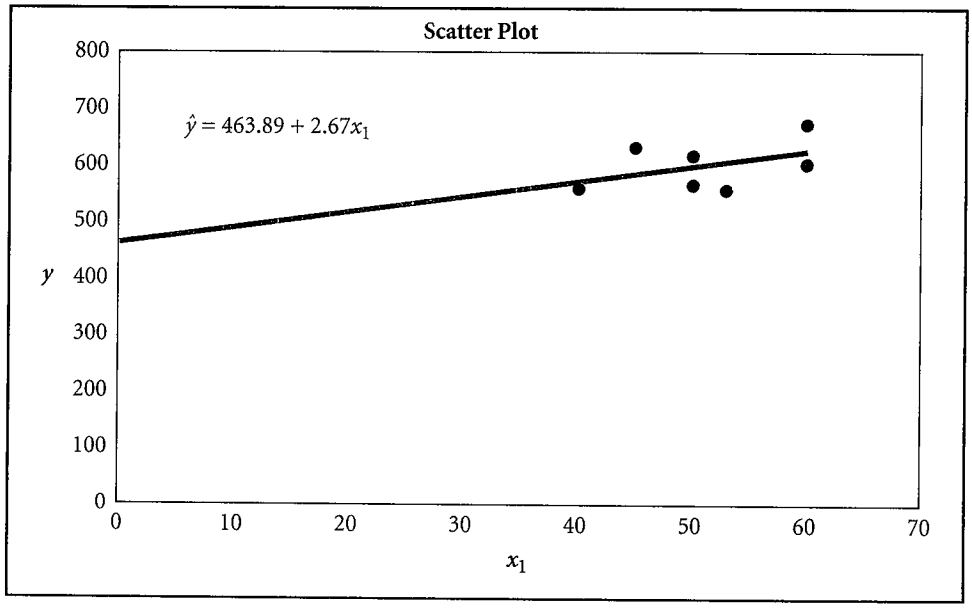
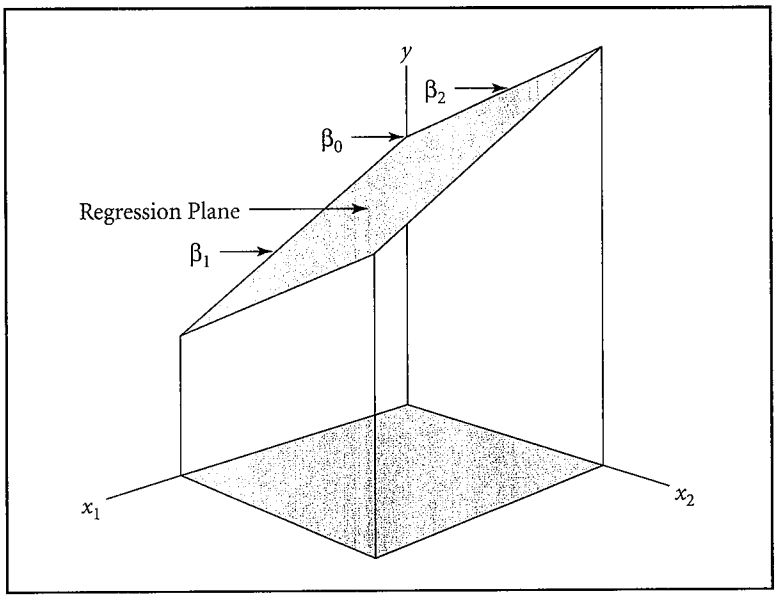


FIGURE 15.2

Multiple Regression Hyperplane for Population



regression analysis for y and x_1. The points are plotted in two-dimensional space, and the regression model is represented by a line through the points such that the sum of squares of error [$SSE = \Sigma(y - \hat{y})^2$] is minimized.

If we add variable x_2 to the model, as shown in Table 15.1B, the resulting multiple regression equation becomes

$$\hat{y} = 307.71 + 2.85x_1 + 10.94x_2$$

For the time being don't worry about how this equation was computed. That will be discussed shortly. Note, however, that the (y, x_1, x_2) points form a three-dimensional space, as shown in Figure 15.2. The regression equation forms a slice (hyperplane) through the data such that $\Sigma(y - \hat{y})^2$ is minimized. This is the same *least squares criterion* that is used with simple linear regression.

The mathematics for developing the least squares regression equation for simple linear regression involves differential calculus. The same is true for the multiple regression equation. Because the least squares regression coefficients are determined using matrix algebra, the mathematical derivation is beyond the scope of this text.[1]

Multiple regression analysis is virtually always performed with the aid of a computer and appropriate software. Both Minitab and Excel contain procedures for performing multiple regression. Minitab has a far more complete regression procedure. However, the PHStat Excel add-ins on your CD-ROM expand Excel's capabilities. Each software package presents the output in a slightly different format; however, the same basic information will appear in all regression output.

CHAPTER OUTCOME #1

Basic Model-Building Concepts

Model
A representation of an actual system using either a physical or a mathematical portrayal.

An important activity in business decision making is referred to as **model** building. Models are often used to test changes in a system without actually having to change the real system. Models are also used to help describe a system or to predict the output of a system based on certain specified inputs. You are probably quite aware of physical models. Airlines use flight simulators to train pilots. Wind tunnels are used to determine the aerodynamics of automobile designs. Golf ball makers use a physical model of a golfer called "Iron Mike" that can be set to swing golf clubs in a very controlled manner to determine how far a golf

[1] For a complete treatment of the matrix algebra approach for estimating multiple regression coefficients, consult *Applied Linear Statistical Models* by Kutner et al.

ball will fly. Although physical models are very useful in business decision making, our emphasis in this chapter is on mathematical models. In particular, we are interested in statistical models that are developed using multiple regression analysis.

People involved in model building frequently conclude that it is both an art and a science. Determining an appropriate model is a challenging task, but it can be made manageable by employing a model-building process consisting of the following three components: model specification, model fitting, and model diagnosis.

Model Specification

Model specification, or model identification, is the process of determining the dependent variable, deciding which independent variables should be included in the model, and obtaining the sample data for all variables. As with any statistical tool, the larger the sample size the better, because the potential for extreme sampling error is reduced when the sample size is large. However, at a minimum, the sample size required to compute a regression model must be at least one greater than the number of independent variables.[2] If we are thinking of developing a regression model with five independent variables, the absolute minimum number of cases required is six. Otherwise, the computer software will indicate an error has been made or will print out meaningless values. However, as a practical matter, the sample size should be at least four times the number of independent variables. Thus, if we had five independent variables ($k = 5$), we would want at least 20 cases to develop the regression model.

SELECT (IDENTIFY) VARIABLES THAT GO UP AND DOWN

SUMMARY Model Specification

In the context of the statistical models discussed in this chapter, this component involves the following three steps:

1. Decide what question you want to ask. The question being asked usually indicates the dependent variable. In the previous chapter, we discussed how simple linear regression analysis could be used to describe the relationship between a dependent and an independent variable.
2. List the potential independent variables for your model. Here, your knowledge of the situation you are modeling guides you in identifying potential independent variables.
3. Gather the sample data (observations) for all variables.

Model Building *Model building* is the process of actually constructing a mathematical equation in which some or all of the independent variables are used in an attempt to explain the variation in the dependent variable.

Model Diagnosis *Model diagnosis* is the process of analyzing the quality of the model you have constructed by determining how well a specified model fits the data you just gathered. You will examine such output values as R-squared and the standard error of the model. At this stage, you will also assess the extent to which the model's assumptions appear to be satisfied. (Section 15.5 is devoted to examining whether a model meets the regression analysis assumptions.) If the model is unacceptable in any of these areas, you will be forced to revert to the model-specification step and begin again. However, you will be the final judge of whether the model provides acceptable results, and you will always be constrained by time and cost considerations.

An important consideration in practical situations is to use the simplest available model that will meet your needs. The objective of model building is to help you make better decisions. You do not need to feel that a sophisticated model is better if a simpler one will provide acceptable results.

[2] There are mathematical reasons for this sample-size requirement that are beyond the scope of this text. In essence, the regression coefficient in Equation 15.2 can't be computed if the sample size is not at least one larger than the number of independent variables.

SUMMARY Developing a Multiple Regression Model

The following steps are employed in developing a multiple regression model:

1. Specify the model by determining the dependent variable, potential independent variables and select the sample data.
2. Formulate the model. This is done by computing the correlation coefficients for the dependent variable and each independent variable and for each independent variable with all other independent variables. The multiple regression equation is also computed. The computations are performed using computer software such as Excel or Minitab.
3. Perform diagnostic checks on the model to determine how well the specified model fits the data and how well the model appears to meet the multiple regression assumptions.

EXAMPLE 15-1 Developing a Multiple Regression Model

TRY PROBLEM 15.3

CHAPTER OUTCOME #2

Excel and Minitab Tutorial

First City Real Estate First City Real Estate executives wish to build a model to predict sales prices for residential property. Such a model will be valuable when working with potential sellers who might list their homes with First City. This can be done using the following steps:

Step 1 Specify the model.

The question being asked is how can the real estate firm determine the selling price for a house? Thus, the dependent variable is the sales price. This is what the managers want to be able to predict. The managers met in a brainstorming session to determine a list of possible independent (explanatory) variables. Some variables, such as "condition of the house," were eliminated because of lack of data. Others such as "curb appeal" (the appeal of the house to people as they drive by) were eliminated because the values for these variables would be too subjective and difficult to quantify. From a wide list of possibilities, the managers selected the following variables as good candidates:

x_1 = Home size (in square free)
x_2 = Age of house
x_3 = Number of bedrooms
x_4 = Number of bathrooms
x_5 = Garage size (number of cars)

Data were obtained for a sample of 319 residential properties that had sold within the previous two months in an area served by two of First City's offices. For each house in the sample, the sales price and values for each potential independent variable were collected. The data are in the CD-ROM file **First City**.

Step 2 Formulate the model.

The regression model is developed by including independent variables from among those for which you have complete data. There is no way to determine whether an independent variable will be a good predictor variable by analyzing the individual variable's descriptive statistics, such as the mean and standard deviation. Instead, we need to look at the correlation between the independent variables and the dependent variable, which is measured by the **correlation coefficient**.

Correlation Coefficient
A quantitative measure of the strength of the linear relationship between two variables. The correlation coefficient, r, ranges from -1.0 to $+1.0$.

Correlation Matrix
A table showing the pairwise correlations between all variables (dependent and independent).

When we have multiple independent variables and one dependent variable, we can look at the correlation between all pairs of variables by developing a **correlation matrix**. Each correlation is computed using one of the equations in Equation 15.3. The appropriate formula is determined by whether the correlation is being calculated for an independent variable and the dependent variable or for two independent variables, respectively.

Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} \quad \text{or} \quad r = \frac{\Sigma(x_i-\bar{x}_i)(x_j-\bar{x}_j)}{\sqrt{\Sigma(x_i-\bar{x}_i)^2\,\Sigma(x_j-\bar{x}_j)^2}} \qquad (15.3)$$

One x variable with y — One x variable with another x

The actual calculations are done using Excel's correlation tool or Minitab's correlation command, and the results are shown in Figure 15.3a and Figure 15.3b. The output provides the correlation between y and each x variable and between each pair of independent variables.[3] Recall that in Chapter 14, a t-test (see Equation 14-3) was used to test whether the correlation coefficient is statistically significant.

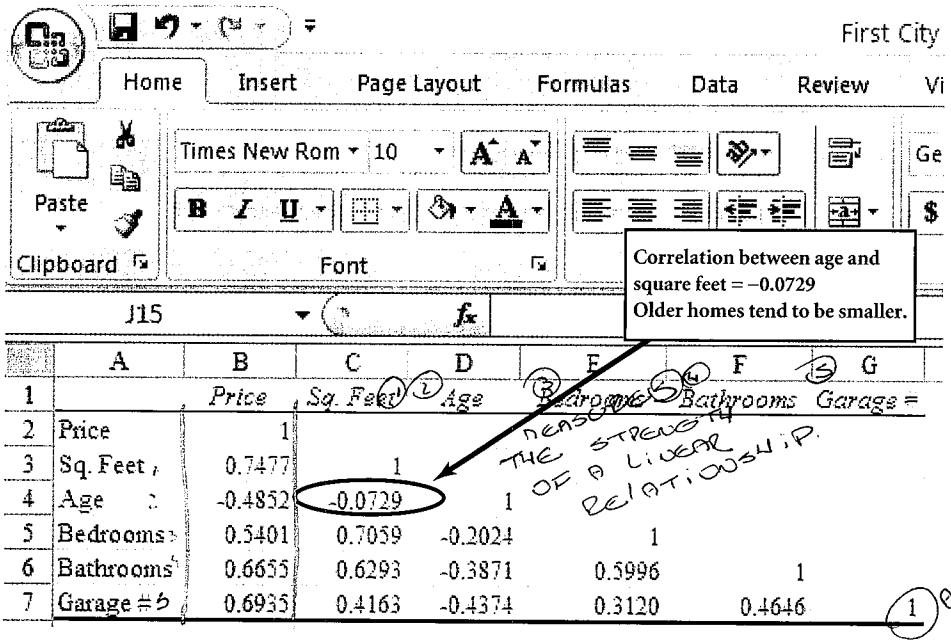
$$H_0: \rho = 0 \qquad H_A: \rho \neq 0$$

We will conduct the test with a significance level of

$$\alpha = 0.05$$

FIGURE 15.3A

Excel 2007 Results Showing First City Real Estate Correlation Matrix



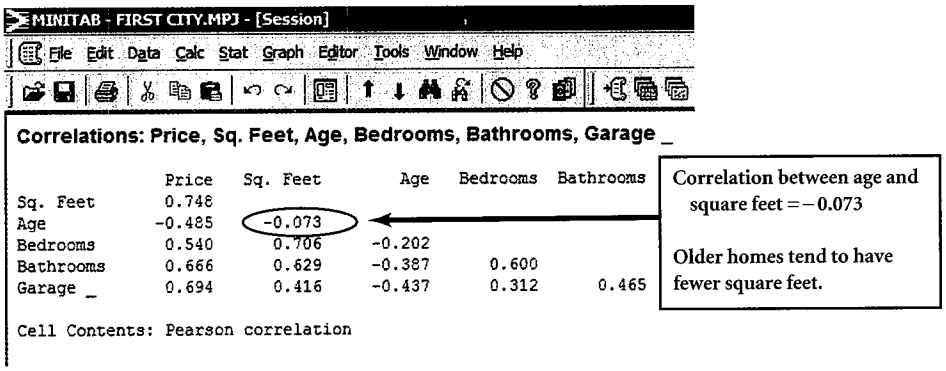
	A	B	C	D	E	F	G
1		Price	Sq. Feet	Age	Bedrooms	Bathrooms	Garage =
2	Price	1					
3	Sq. Feet	0.7477	1				
4	Age	-0.4852	-0.0729	1			
5	Bedrooms	0.5401	0.7059	-0.2024	1		
6	Bathrooms	0.6655	0.6293	-0.3871	0.5996	1	
7	Garage = 5	0.6935	0.4163	-0.4374	0.3120	0.4646	1

Excel 2007 Instructions:
1. Open file: First City.xls.
2. Select Home Sample 1 worksheet.
3. Click on **Data > Data Analysis.**
4. Select **Correlation.**
5. Define y variable range (all rows and columns).
6. Click on **Labels.**
7. Click **OK.**

[3] Minitab, in addition to providing the correlation matrix, can provide the p-values for each correlation. If the p-value is less than the specified alpha, the correlation is statistically significant.

FIGURE 15.3B

Minitab Results Showing First CityReal Estate Correlation Matrix



Minitab Instructions:

1. Open file: First City.MTW.
2. Choose **Stat > Basic Statistics > Correlation.**
3. In **Variables**, enter variable columns.
4. Click **OK.**

Given degrees of freedom equal to

$$n - 2 = 319 - 2 = 317$$

the critical t (see Appendix E) for a two-tailed test is approximately[4]

$$t_{0.025} = 1.96$$

Any correlation coefficient generating a t-value > 1.96 or less than -1.96 is determined to be significant.

For now, we will focus on the correlations in the first column in Figures 15.3a and 15.3b, which measure the strength of the linear relationship between each independent variable and the dependent variable, sales price. For example, the t statistic for price and square feet is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7477}{\sqrt{\frac{1-0.7477^2}{319-2}}} = 20.048$$

Because

$$t = 20.048 > 1.96$$

we reject H_0 and conclude that the correlation between sales price and square feet is statistically significant.

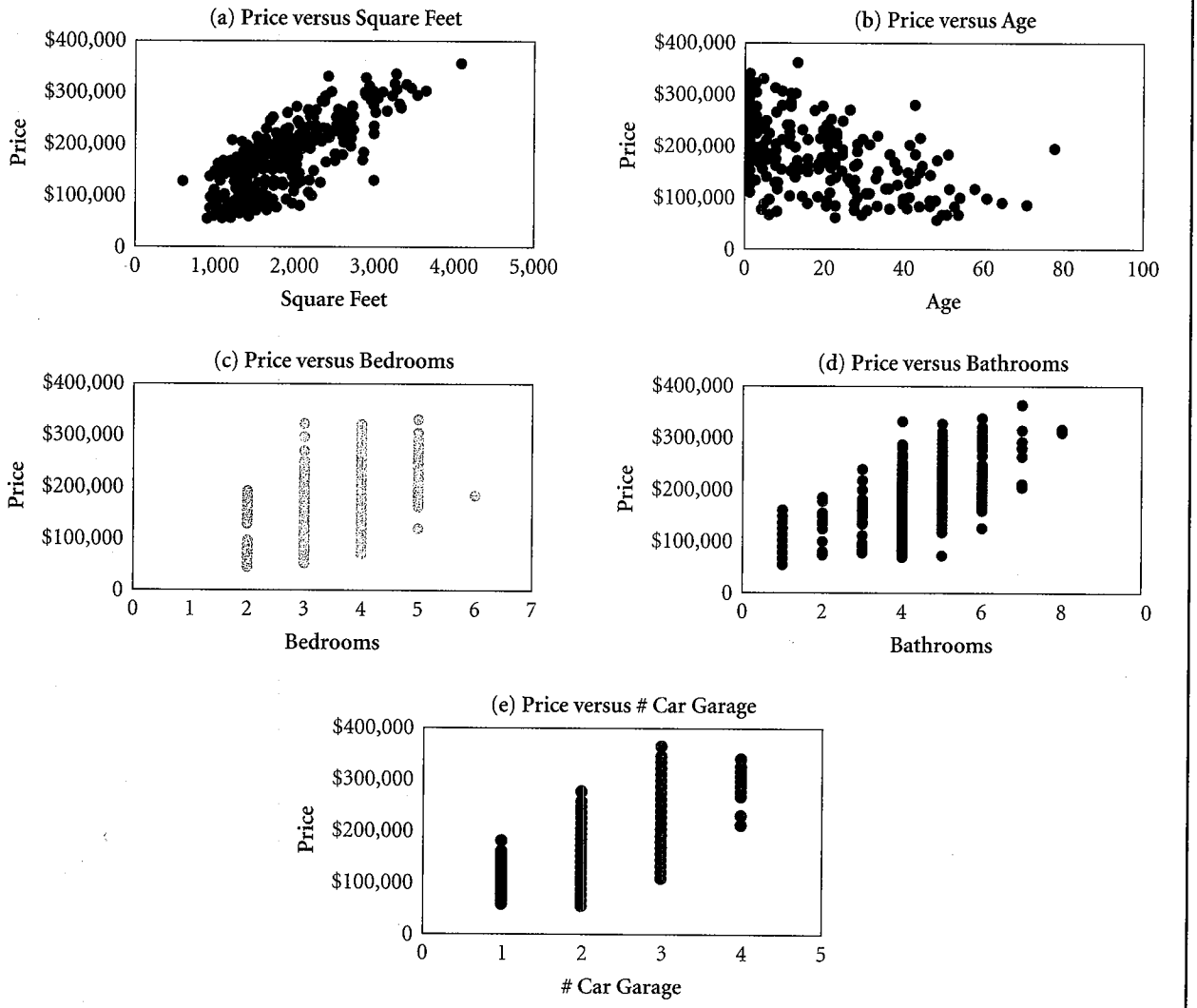
Similar calculations for the other independent variables with price show that all variables are statistically correlated with price. This indicates that a significant linear relationship exists between each independent variable and sales price. Variable x_1, square feet, has the highest correlation at 0.748. Variable x_2, age of the house, has the lowest correlation at -0.485. The negative correlation implies that older homes tend to have lower sales prices.

As we discussed in Chapter 14, it is always a good idea to develop scatter plots to visually see the relationship between two variables. Figure 15.4 shows the scatter plots for each independent variable and the dependent variable, sales price. In each case, the plots indicate a linear relationship between the independent variable and the dependent

[4] You can use the Excel TINV function to get the precise t-value, which is 1.967.

FIGURE 15.4

First City Real Estate Scatter Plots



variable. Note that several of the independent variables (bedrooms, bathrooms, garage size) are quantitative but discrete. The scatter plots for these variables show points at each level of the independent variable rather than over a continuum of values.

CHAPTER OUTCOME #3

Computing the Regression Equation First City's goal is to develop a regression model to predict the appropriate selling price for a home, using certain measurable characteristics. The first attempt at developing the model will be to run a multiple regression computer program using all available independent variables. The regression outputs from Excel and Minitab are shown in Figure 15.5a and Figure 15.5b.

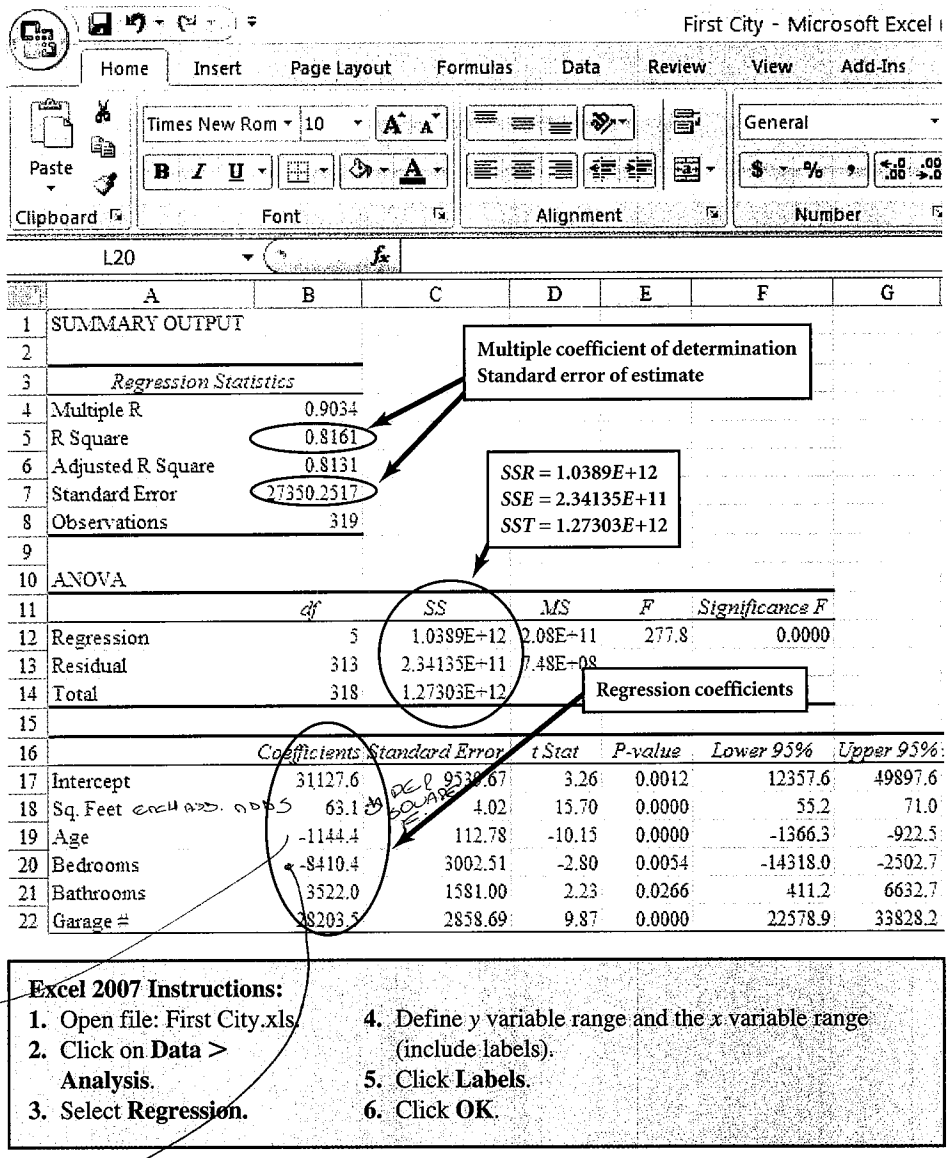
The estimate of the multiple regression model given in Figure 15.5a is

$$\hat{y} = 31{,}127.6 + 63.1(\text{sq. ft.}) - 1{,}144.4(\text{age}) - 8{,}410.4(\text{bedrooms}) + 3{,}522.0(\text{bathrooms}) + 28{,}203.5(\text{garage})$$

The coefficients for each independent variable represent an estimate of the average change in the dependent variable for a 1-unit change in the independent variable, holding all other

FIGURE 15.5A

Excel 2007 Multiple Regression Model Results for First City Real Estate



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9034					
5	R Square	0.8161					
6	Adjusted R Square	0.8131					
7	Standard Error	27350.2517					
8	Observations	319					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	1.0389E+12	2.08E+11	277.8	0.0000	
13	Residual	313	2.34135E+11	7.48E+08			
14	Total	318	1.27303E+12				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	31127.6	9539.67	3.26	0.0012	12357.6	49897.6
18	Sq. Feet	63.1	4.02	15.70	0.0000	55.2	71.0
19	Age	-1144.4	112.78	-10.15	0.0000	-1366.3	-922.5
20	Bedrooms	-8410.4	3002.51	-2.80	0.0054	-14318.0	-2502.7
21	Bathrooms	3522.0	1581.00	2.23	0.0266	411.2	6632.7
22	Garage	28203.5	2858.69	9.87	0.0000	22578.9	33828.2

Excel 2007 Instructions:

1. Open file: First City.xls.
2. Click on **Data** > **Analysis**.
3. Select **Regression**.
4. Define *y* variable range and the *x* variable range (include labels).
5. Click **Labels**.
6. Click **OK**.

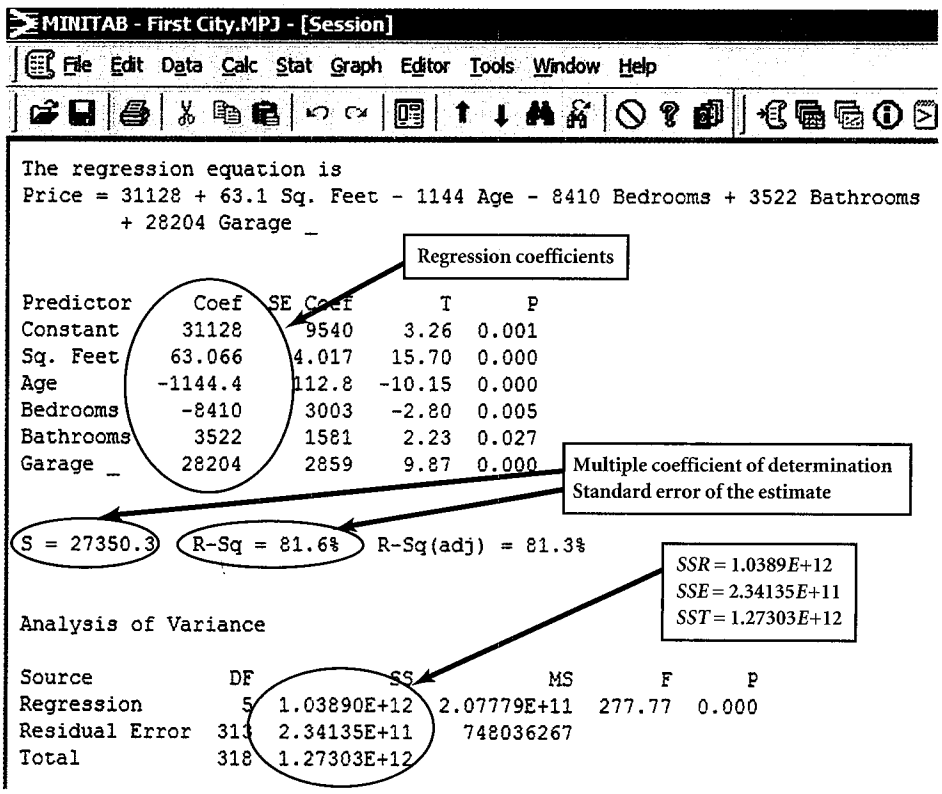
independent constant. For example, for houses of the same age, with the same number of bedrooms, baths, and garages, a 1-square-foot increase in the size of the house is estimated to increase its price by an average of $63.10. Likewise, for houses with the same square feet, bedrooms, bathrooms, and garages, a 1-year increase in the age of the house is estimated to result in an average drop in sales price of $1,144.40. The other coefficients are interpreted in the same way. Note, in each case, we are interpreting the regression coefficient for one independent variable while holding the other variables constant.

To estimate the value of a residential property, First Real Estate brokers would substitute values for the independent variables into the regression equation. For example, suppose a house with the following characteristics is considered:

x_1 = Square feet = 2,100
x_2 = Age = 15
x_3 = Number of bedrooms = 4
x_4 = Number of baths = 3
x_5 = Size of garage = 2

FIGURE 15.5B

Minitab Multiple Regression Model Results for First City Real Estate



Minitab Instructions:

1. Open file: First City. MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Click **OK**.

The point estimate for the sales price is

$$\hat{y} = 31{,}127.6 + 63.1(\text{sq. ft.}) - 1{,}144.4(\text{age}) - 8{,}410.4(\text{bathrooms}) + 3{,}522.0(\text{bathrooms}) + 28{,}203.5(\text{garage})$$
$$\hat{y} = 31{,}127.6 + 63.1(2{,}100) - 1{,}144.4(15) - 8{,}410.4(4) + 3{,}522.0(3) + 28{,}203.5(2)$$
$$\hat{y} = \$179{,}802.70$$

The Coefficient of Determination You learned in Chapter 14 that the *coefficient of determination*, R^2, measures the proportion of variation in the dependent variable that can be explained by the dependent variable's relationship to a single independent variable. When there are multiple independent variables in a model, R^2 is also used to determine the proportion of variation in the dependent variable that is explained by the dependent variable's relationship to all the independent variables in the model. However, R^2 is now called the **multiple coefficient of determination**. Equation 15.4 is used to compute R^2 for a multiple regression model.

Multiple Coefficient of Determination (R^2)

$$R^2 = \frac{\text{Sum of squares regression}}{\text{Total sum of squares}} = \frac{SSR}{SST} \tag{15.4}$$

As shown in Figure 15.5a, $R^2 = 0.8161$. Both *SSR* and *SST* are also included in the output. Therefore, you can use Equation 15.4 to get R^2, as follows:

$$\frac{SSR}{SST} = \frac{1.0389E+12}{1.27303E+12} = 0.8161$$

More than 81% of the variation in sales price can be explained by the linear relationship of the five independent variables in the regression model to the dependent variable. However, as we shall shortly see, not all independent variables are equally important to the model's ability to explain this variation.

Step 3 Diagnose the model.

Before First City actually uses this regression model to estimate the sales price of a house, there are several questions that should be answered.

1. Is the overall model significant?
2. Are the individual variables significant?
3. Is the standard deviation of the model error too large to provide meaningful results?
4. Is multicollinearity a problem?
5. Have the regression analysis assumptions been satisfied?

We shall answer the first four questions in order. We will have to wait until Section 15.5 before we have the tools to answer the fifth important question.

CHAPTER OUTCOME #4

Is the Model Significant? You should keep in mind that the regression model we constructed is based on a sample of data from the population and is subject to sampling error. Therefore, we need to test the statistical significance of the overall regression model. We have previously discussed the multiple coefficient of determination, R^2, which is a measure of how much of the variation in the dependent variable can be explained by the regression model. Because R^2 is a sample statistic, it can be used to make inferences about whether the overall model is statistically significant in explaining the variation in the dependent variable. The specific null and alternative hypotheses tested for First City Real Estate are

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_A: \text{At least one } \beta_i \neq 0.$$

If the null hypothesis is true and all the slope coefficients are simultaneously equal to zero, the overall regression model is not useful for predictive or descriptive purposes.

The *F*-test is a method for testing whether the regression model explains a significant proportion of the variation in the dependent variable (and whether the overall model is significant). The *F*-test statistic for a multiple regression model is shown in Equation 15.5.

F-Test Statistic

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} \tag{15.5}$$

where:

SSR = Sum of squares regression = $\sum(\hat{y}-\bar{y})^2$
SSE = Sum of squares error = $\sum(y-\hat{y})^2$
n = Number of data points
k = Number of independent variables
Degrees of freedom = $D_1 = k$ and $D_2 = (n-k-1)$

The ANOVA portion of the output shown in Figure 15.5a contains values for *SSR*, *SSE*, and the *F*-value. The general format of the ANOVA table in a regression analysis is as follows:

ANOVA

Source	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	k	*SSR*	$MSR = SSR/k$	*MSR/MSE*	computed *p*-value
Residual	$n-k-1$	*SSE*	$MSE = SSE/(n-k-1)$		
Total	$n-1$	*SST*			

The ANOVA portion of the output from Figure 15.5a is as follows:

ANOVA

Source	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	5	1.04*E* +12	2.08*E* + 11	277.8	0.0000
Residual	313	2.34*E* +11	7.48*E* + 08		
Total	318	1.27303*E* +12			

We can test the model's significance.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_A: \text{At least one } \beta_i \neq 0.$$

by either comparing the calculated F-value, 277.8, with a critical value for a given alpha level

$$\alpha = 0.01$$

and $k = 5$ and $n - k - 1 = 313$ degrees of freedom using Excel's **FINV** function, ($F_{0.01} = 3.079$), or compare the p-value in the output with a specified alpha level. Because

$$F = 277.8 > 3.079, \text{ reject } H_0$$

or because

$$p\text{-value} \approx 0.0 < 0.01, \text{ reject } H_0$$

Therefore, we should conclude that the regression model *does* explain a significant proportion of the variation in sales price. Thus, the overall model is statistically significant. This means we can conclude that at least one of the regression slope coefficients is not equal to zero.

Excel and Minitab also provide a measure called the R-sq(adj), which is the **adjusted R-squared** value (see Figure 15.5a and 15.5b). It is calculated by Equation 15.6.

Adjusted R-Squared

A measure of the percentage of explained variation in the dependent variable that takes into account the relationship between the sample size and the number of independent variables in the regression model.

$$R\text{-sq(adj)} = R_A^2 = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right) \tag{15.6}$$

where:

n = Sample size
k = Number of independent variables
R^2 = Coefficient of determination

Adding independent variables to the regression model will always increase R^2, even if these variables have no relationship to the dependent variable. Therefore, as the number of independent variables is increased (regardless of the quality of the variables), R^2 will increase. However, each additional variable results in the loss of 1 degree of freedom. This is viewed as part of the cost of adding the specified variable. The addition to R^2 may not justify the reduction in degrees of freedom. The R_A^2 value takes into account this cost and adjusts the R_A^2 value accordingly. R_A^2 will always be less than R^2. When a variable is added that does not contribute its fair share to the explanation of the variation in the dependent variable, the R_A^2 may actually decline, even though R^2 will always increase. The adjusted R-squared is a particularly important measure when the number of independent variables is large relative to the sample size. It takes into account the relationship between sample size and number of variables. R^2 may appear artificially high if the number of variables is large compared with the sample size.

In this example, in which the sample size is quite large relative to the number of independent variables, the adjusted R-squared is 81.3%, only slightly less than $R^2 = 81.6\%$.

CHAPTER OUTCOME #4

Are the Individual Variables Significant? We have concluded that the overall model is significant. This means *at least* one independent variable explains a significant proportion of the variation in sales price. This does not mean that *all* the variables are significant, however. To determine which variables are significant, we test the following hypotheses:

$$H_0: \beta_j = 0$$
$$H_A: \beta_j \neq 0 \qquad \text{for all } j$$

We can test the significance of each independent variable using significance level

$$\alpha = 0.05$$

and a t-test, as discussed in Chapter 14. The calculated t-values should be compared to the critical t-value with

$$n - k - 1 = 319 - 5 - 1 = 313$$

degrees of freedom, which is approximately

$$t_{0.025} \approx 1.96$$

for $\alpha = 0.05$. The calculated t-value for each variable is provided on the computer printout in Figures 15.5a and 15.5b. Recall that the t statistic is determined by dividing the regression coefficient by the estimator of the standard deviation of the regression coefficient, as shown in Equation 15.9.

t-Test for Significance of Each Regression Coefficient

$$t = \frac{b_j - 0}{s_{b_j}} \qquad df = n - k - 1 \tag{15.7}$$

where:

b_j = Sample slope coefficient for the jth independent variable
s_{b_j} = Estimate of the standard error for the jth sample slope coefficient

For example, the t-value for square feet shown in Figure 15.5a is 15.70. This was computed using Equation 15.7, as follows:

$$t = \frac{b_j - 0}{s_{b_j}} = \frac{63.1 - 0}{4.02} = 15.70$$

Because

$$t = 15.70 > 1.96, \text{ we reject } H_0,$$

and conclude that the regression slope for square feet is not zero.

We can also look at the Excel or Minitab output and compare the p-value for each regression slope coefficient with alpha. If the p-value is less than alpha, we reject the null hypothesis and conclude that the independent variable is statistically significant in the model. Both the t-test and the p-value techniques will give the same results.

You should consider that these t-tests are *conditional* tests. This means the null hypothesis is that *the value of each slope coefficient is 0, given that the other independent variables are already in the model.*[5] Figure 15.6 shows the hypothesis tests for each independent variable using a 0.05 significance level. We conclude that all five independent variables in the model are significant. When a regression model is to be used for prediction, the model should contain no insignificant variables. If insignificant variables are present, they should be dropped and a new regression equation obtained before the model is used for prediction purposes. We will have more to say about this later.

Is the Standard Deviation of the Regression Model Too Large? The purpose of developing the First City regression model is to be able to determine values of the dependent variable when corresponding values of the independent variables are known. An indication of how good the regression model is can be found by looking at the relationship between the measured values of the dependent variable and those values that would be predicted by the regression model. The standard deviation of the regression model (also called the *standard error of the estimate),* measures the dispersion of observed home sale values, y, around values predicted by the regression model. The standard error of the estimate is shown in Figure 15.5a, and can be computed using Equation 15.8.

Standard Error of the Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{MSE} \tag{15.8}$$

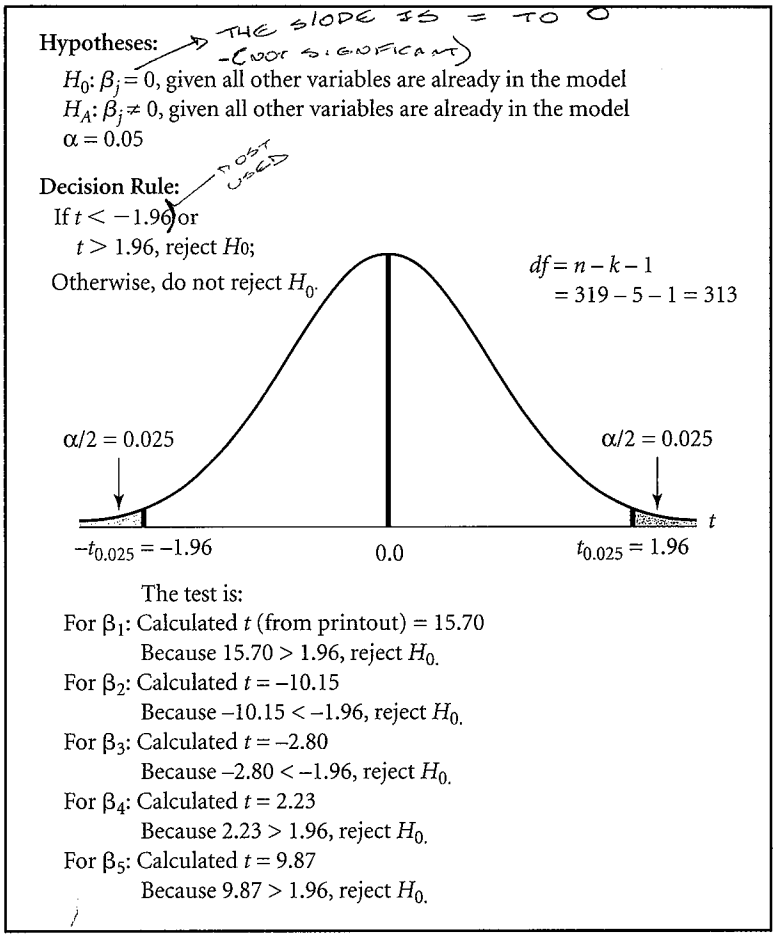
where:

SSE = Sum of squares error (residual)
n = Sample size
k = Number of independent variables

[5] Note that the t-tests may be affected if the independent variables in the model are themselves correlated. A procedure known as the *sum of squares drop F-test*, discussed by Kutner et al. in *Applied Linear Statistical Models*, should be used in this situation. Each t-test considers only the marginal contribution of the independent variables and may indicate that none of the variables in the model are significant, even though the ANOVA procedure indicates otherwise.

FIGURE 15.6

Significance Tests for Each Independent Variable in the First City Real Estate Example



Note: The degrees of freedom for the t-distribution is $(n - k - 1)$, where k is the total number of independent variables in the model.

Examining Equation 15.8 closely, we see that this standard error of the estimate is the square root of the mean square error of the residuals found in the analysis of variance table.

Sometimes, even though a model has a high R^2, the standard error of the estimate will be too large to provide adequate precision for confidence and prediction intervals. A rule of thumb that we have found useful is to examine the range $\pm 2s_\varepsilon$. If this range is acceptable from a practical viewpoint, the standard error of the estimate might be considered acceptable.[6]

In this First City Real Estate Company example, the model standard error, shown in Figure 15.5a, is $27,350. Thus, the rough prediction range for the price of an individual home is

$\pm 2(\$27,350)$

$\pm \$54,700$

[6] The actual confidence interval for prediction of a new observation requires the use of matrix algebra. However, when the sample size is large and dependent variable values near the means of the dependent variables are used, the rule of thumb given here is a close approximation. Refer to *Applied Linear Statistical Models* by Kutner et al. for further discussion.

From a practical viewpoint, a potential error of $54,700 above or below the true value is probably not acceptable. Not many homeowners would be willing to have their appraisal value set by a model with a possible error this large. The company needs to take steps to reduce the standard deviation of the model error. Subsequent sections of this chapter discuss some ways we can attempt to reduce it.

CHAPTER OUTCOME #5

Is Multicollinearity a Problem? Even if the overall regression model is significant and each independent variable is significant, decision makers should still examine the regression model to determine whether it appears reasonable. This is referred to as checking for *face validity*. Specifically, you should check to see that signs on the regression coefficients are consistent with the signs on the correlation coefficients between the independent variables and the dependent variable. Does any regression coefficient have an unexpected sign?

Before answering this question for the First City Real Estate example, we should review what the regression coefficients mean. First, the constant term, b_0, is the estimate of the model's y intercept. If the data used to develop the regression model contain values of x_1, x_2, x_3, x_4, and x_5 that are simultaneously 0 (such as would be the case for vacant land), b_0 is the mean value of y, given that x_1, x_2, x_3, x_4, and x_5 all equal 0. Under these conditions b_0 would estimate the average value of a vacant lot. However, in the First City example, no vacant land was in the sample, so b_0 has no particular meaning.

The coefficient for square feet, b_1, estimates the average change in sales price corresponding to a change in house size of 1 square foot, holding the other independent variables constant. The value shown in Figure 15.5a for b_1 is 63.1. The coefficient is positive, indicating that an increase in square footage is associated with an increase in sales price. This relationship is expected. All other things being equal, bigger houses should sell for more money.

Likewise, the coefficient for x_5, the size of the garage, is positive, indicating that an increase in size is also associated with an increase in price. This is expected. The coefficient for x_2, the age of the house is negative, indicating that an older house is worth less than a similar younger house. This also seems reasonable. However, the coefficient for variable x_3, the number of bedrooms, is −$8,410.4 meaning that, if we hold the other variables constant but increase the number of bedrooms by one, the average price will *drop* by $8,410.40. This would appear to run counter to conventional thinking about the housing market. Finally, variable x_4 for bathrooms has the expected positive sign.

Referring to the correlation matrix that was shown earlier in Figure 15.3, the correlation between variable x_3, bedrooms, and y, the sales price, is +0.540. This indicates that, without considering the other independent variables, the linear relationship between number of bedrooms and sales price is positive. But why does the regression coefficient turn out negative in the model? The answer lies in what is called **multicollinearity**.

Multicollinearity
A high correlation between two independent variables such that the two variables contribute redundant information to the model. When highly correlated independent variables are included in the regression model, they can adversely affect the regression results.

Multicollinearity occurs when independent variables overlap with respect to the information they provide in explaining the variation in the dependent variable. For example, x_3 and the other independent variables have the following correlations (see Figure 15.3b):

$$r_{x_3,x_1} = 0.706$$
$$r_{x_3,x_2} = -0.202$$
$$r_{x_3,x_4} = 0.600$$
$$r_{x_3,x_5} = 0.312$$

All four correlations have t-values indicating a significant linear relationship. Refer to the correlation matrix in Figure 15.3 to see that other independent variables are also correlated with each other.

The problems caused by multicollinearity, and how to deal with them, continue to be of prime concern to statisticians. From a decision maker's viewpoint, you should be aware that multicollinearity can (and usually does) exist and recognize the basic problems it can cause. Some of the most obvious problems and indications of severe multicollinearity are the following:

1. Incorrect signs on the coefficients.
2. A sizable change in the values of the previously estimated coefficients when a new variable is added to the model.
3. A variable that was previously significant in the regression model becomes insignificant when a new independent variable is added.
4. The estimate of the standard deviation of the model error increases when a variable is added to the model.

Mathematical approaches exist for dealing with multicollinearity and reducing its impact. Although these procedures are beyond the scope of this text, one suggestion is to eliminate the variables that are the chief cause of the multicollinearity problems.

If the independent variables in a regression model are correlated and multicollinearity is present, another potential problem is that the t-tests for the significance of the individual independent variables may be misleading. That is, a t-test may indicate that the variable is not statistically significant when in fact it is.

Variance Inflation Factor
A measure of how much the variance of an estimated regression coefficient increases if the independent variables are correlated. A *VIF* equal to 1.0 for a given independent variable indicates that this independent variable is not correlated with the remaining independent variables in the model. The greater the multicollinearity, the larger the *VIF*.

One method of measuring multicollinearity is known as the **variance inflation factor (*VIF*)**. Equation 15.9 is used to compute the *VIF* for each independent variable.

Variance Inflation Factor

$$VIF = \frac{1}{(1 - R_j^2)} \tag{15.9}$$

where:

R_j^2 = Coefficient of determination when the jth independent variable is regressed against the remaining $k - 1$ independent variables.

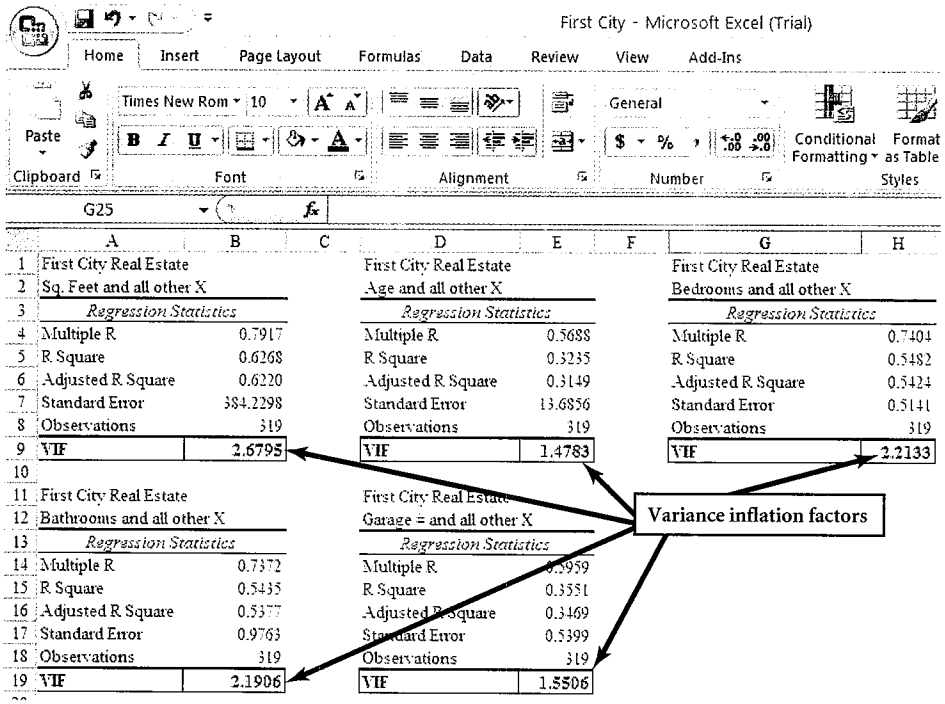
Both the PHStat add-ins to Excel and Minitab contain options that provide *VIF* values.[7]

Figure 15.7 shows the Excel (PHStat) output of the variance inflation factors for the First City Real Estate example. The effect of multicollinearity is to decrease the test statistic, thus reducing the probability that the variable will be declared significant. A related impact is to increase the width of the confidence interval estimate of the slope coefficient in the regression model. Generally, if the $VIF < 5$ for a particular independent variable, multicollinearity is not considered a problem for that variable. *VIF* values ≥ 5 imply that the correlation between the independent variables is too extreme and should be dealt with by dropping variables from the model. As Figure 15.7 illustrates, the *VIF* values for each independent variable are less than 5, so based on variance inflation

[7] Excel's Regression procedure in the Data Analysis Tools area does not provide *VIF* values directly. Without PHStat, you would need to compute each regression analysis individually and record the R-squared value to compute the *VIF*.

FIGURE 15.7

Excel 2007 (PHStat) Multiple Regression Model Results for First City Real Estate with Variance Inflation Factors



	A	B	C	D	E	F	G	H
1	First City Real Estate			First City Real Estate			First City Real Estate	
2	Sq. Feet and all other X			Age and all other X			Bedrooms and all other X	
3	*Regression Statistics*			*Regression Statistics*			*Regression Statistics*	
4	Multiple R	0.7917		Multiple R	0.5688		Multiple R	0.7404
5	R Square	0.6268		R Square	0.3235		R Square	0.5482
6	Adjusted R Square	0.6220		Adjusted R Square	0.3149		Adjusted R Square	0.5424
7	Standard Error	384.2298		Standard Error	13.6856		Standard Error	0.5141
8	Observations	319		Observations	319		Observations	319
9	VIF	2.6795		VIF	1.4783		VIF	2.2133
10								
11	First City Real Estate			First City Real Estate				
12	Bathrooms and all other X			Garage = and all other X				
13	*Regression Statistics*			*Regression Statistics*				
14	Multiple R	0.7372		Multiple R	0.5959			
15	R Square	0.5435		R Square	0.3551			
16	Adjusted R Square	0.5377		Adjusted R Square	0.3469			
17	Standard Error	0.9763		Standard Error	0.5399			
18	Observations	319		Observations	319			
19	VIF	2.1906		VIF	1.5506			

Excel 2007 Instructions:

1. Open file: First City.xls.
2. Click on **Add-Ins PHStat**.
3. Select **Regression >Multiple Regression**.
4. Define *y* variable range and the *x* variable range.
5. Select **Regression Statistics Table** and **ANOVA** and **Coefficients Table**.
6. Select **Various Inflation Factor (VIF)**.
7. Click **OK**. Note VIFs consolidated to one page for display in Figure 15.7.

Minitab Instructions (for similar results):

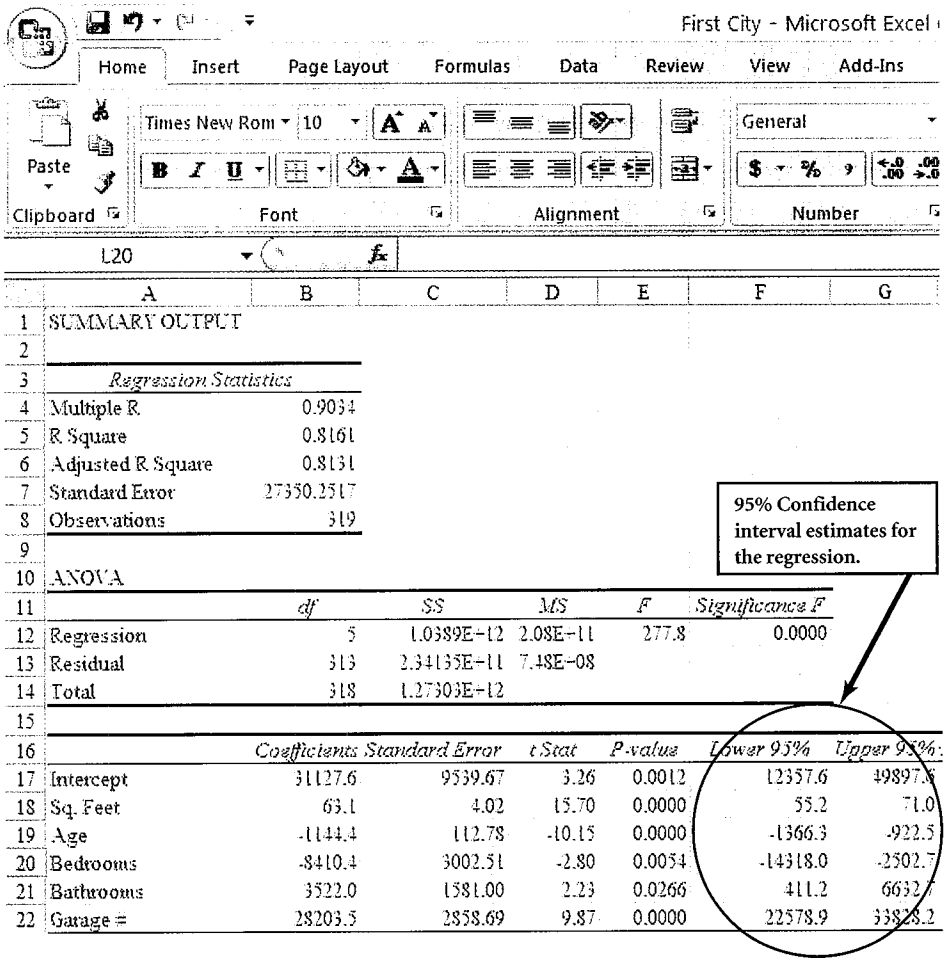
1. Open file: First City.MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response** enter dependent (*y*) variable.
4. In **Predictors**, enter independent (*x*) variables.
5. Click **Options**.
6. In **Display**, select Variance Inflation factors.
7. Click **OK**. **OK**.

factors, even though the sign on the variable, bedrooms, is unexpected, the other multicollinearity issues do not exist among these independent variables.

Confidence Interval Estimation for Regression Coefficients Previously we showed how to determine whether the regression coefficients are statistically significant. This was necessary because the estimates of the regression coefficients are developed from sample data and are subject to sampling error. The issue of sampling error also comes into play when interpreting the slope coefficients.

FIGURE 15.8A

Excel 2007 Multiple Regression Model Results for First City Real Estate



First City - Microsoft Excel

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins

Times New Rom | 10 | General

Paste | Clipboard | Font | Alignment | Number

L20

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9034					
5	R Square	0.8161					
6	Adjusted R Square	0.8131					
7	Standard Error	27350.2517					
8	Observations	319					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	1.0389E+12	2.08E+11	277.8	0.0000	
13	Residual	313	2.34135E+11	7.48E+08			
14	Total	318	1.27303E+12				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	31127.6	9539.67	3.26	0.0012	12357.6	49897.6
18	Sq. Feet	63.1	4.02	15.70	0.0000	55.2	71.0
19	Age	-1144.4	112.78	-10.15	0.0000	-1366.3	-922.5
20	Bedrooms	-8410.4	3002.51	-2.80	0.0054	-14318.0	-2502.7
21	Bathrooms	3522.0	1581.00	2.23	0.0266	411.2	6632.7
22	Garage =	28203.5	2858.69	9.87	0.0000	22578.9	33828.2

95% Confidence interval estimates for the regression.

Excel 2007 Instructions:

1. Open file: First City.xls.
2. Click on **Data > Data Analysis**.
3. Select **Regression**.
4. Define *y* variable range and the *x* variable range (include labels).
5. Click **Labels**.
6. Click **OK**.

Consider again the regression models for First City Real Estate shown in Figures 15.8a and 15.8b. The regression coefficients shown are *point estimates* for the true regression coefficients. For example, the coefficient for the variable square feet is $b_1 = 63.1$. We interpret this to mean that, holding the other variables constant, for each increase in the size of a home by 1 square foot, the price of a house is estimated to increase by \$63.1. But like all point estimates, this is subject to sampling error. In Chapter 14 you were introduced to the concept of confidence interval estimates for the regression coefficients. That same concept applies in multiple regression models. Equation 15.10 is used to develop the confidence interval estimate for the regression coefficients.

Confidence Interval Estimate for the Regression Slope

$$b_j \pm t s_{b_j} \quad \textbf{(15.10)}$$

where:

b_j = Point estimate for the regression coefficient for x_j
t = Critical t-value for the specified confidence level
s_{b_j} = The standard error of the jth regression coefficient

The Excel output in Figure 15.8a provides the confidence interval estimates for each regression coefficient. For example, the 95% interval estimate for square feet is

$55.2 -------- $71.0

Minitab does not have a command to generate confidence intervals for the individual regression parameters. However, statistical quantities are provided on the Minitab output in Figure 15.8b to allow the manual calculation of these confidence intervals. As an example, the confidence interval for the coefficient associated with the square feet variable can be computed using Equation 15.10 as[8]

FIGURE 15.8B

Minitab Multiple Regression Model Results for the First City Real Estate

MINITAB - First City.MPJ - [Session]

File Edit Data Calc Stat Graph Editor Tools Window Help

```
The regression equation is
Price = 31128 + 63.1 Sq. Feet - 1144 Age - 8410 Bedrooms + 3522 Bathrooms
        + 28204 Garage _

                                  s_b1

Predictor      Coef  SE Coef      T      P  VIF
Constant      31128     3540   3.26  0.001
Sq. Feet     63.066    4.017  15.70  0.000  2.7
Age         -1144.4    112.8 -10.15  0.000  1.5
Bedrooms      -8410     3003  -2.80  0.005  2.2
Bathrooms      3522     1581   2.23  0.027  2.2
Garage        28204     2859   9.87  0.000  1.6

  b1 = 63.1

S = 27350.3   R-Sq = 81.6%   R-Sq(adj) = 81.3%


Analysis of Variance

Source           DF          SS          MS       F      P
Regression        5  1.03890E+12  2.07779E+11  277.77  0.000
Residual Error  313  2.34135E+11    748036267
Total           318  1.27303E+12
```

Minitab Instructions:

1. Open file: First City. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variable.
5. Click **OK.**

[8] Note, we used Excel's TINV function to get the precise t-value of 1.967.

$$b_1 \pm ts_{b_1}$$
$$63.1 \pm 1.967(4.017)$$
$$63.1 \pm 7.90$$
$$\$55.2 \text{ -------- } \$71.0$$

We interpret this interval as follows: Holding the other variables constant, using a 95% confidence level, a change in square feet by 1 foot is estimated to generate an average change in home price of between \$55.20 and \$71.00. Each of the other regression coefficients can be interpreted in the same manner.

15-1: Exercises

Skill Development

15-1. You are given the following estimated regression equation involving a dependent and two independent variables:

$$\hat{y} = 12.67 + 4.14x_1 + 8.72x_2$$

a. Interpret the values of the slope coefficients in the equation.
b. Estimate the value of the dependent variable when $x_1 = 4$ and $x_2 = 9$.

15-2. In working for a local retail store you have developed the following estimated regression equation:

$$\hat{y} = 22{,}167 - 412x_1 + 818x_2 - 93x_3 - 71x_4$$

where:
y = Weekly sales
x_1 = Local unemployment rate
x_2 = Weekly average high temperature
x_3 = Number of activities in the local community
x_4 = Average gasoline price

a. Interpret the values of b_1, b_2, b_3, and b_4 in this estimated regression equation.
b. What is the estimated sales if the unemployment rate is 5.7%, the average high temperature is 61°, there were 14 activities, and gasoline average price was \$1.39?

15-3. The following output is associated with a multiple regression model with three independent variables:

	df	*SS*	*MS*	*F*	Significance *F*
Regression	3	16,646.091	5,548.697	5.328	0.007
Residual	21	21,871.669	1,041.508		
Total	24	38,517.760			

	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	87.790	25.468	3.447	0.002
x_1	−0.970	0.586	−1.656	0.113
x_2	0.002	0.001	3.133	0.005
x_3	−8.723	7.495	−1.164	0.258

	Lower 95%	Upper 95%	Lower 90%	Upper 90%
Intercept	34.827	140.753	43.966	131.613
x_1	−2.189	0.248	−1.979	0.038
x_2	0.001	0.004	0.001	0.004
x_3	−24.311	6.864	−21.621	4.174

a. What is the regression model associated with this data?
b. Is the model statistically significant?
c. How much of the variation in the dependent variable can be explained by the model?
d. Are all of the independent variables in the model significant? If not, which are not and how can you tell?
e. How much of a change in the dependent variable will be associated with a one unit change in x_2? In x_3?
f. Do any of the 95% confidence interval estimates of the slope coefficients contain zero. If so, what does this indicate?

15-4. The following correlation matrix is associated with the same data used to build the regression model in Problem 15-3:

	y	x_1	x_2	x_3
y	1			
x_1	−0.406	1		
x_2	0.459	0.051	1	
x_3	−0.244	0.504	0.272	1

Does this output indicate any potential multicollinearity problems with the analysis?

15-5. Consider the following set of data:

x_1	29	48	28	22	28	42	33	26	48	44
x_2	15	37	24	32	47	13	43	12	58	19
y	16	46	34	26	49	11	41	13	47	16

a. Obtain the estimated regression equation.
b. Develop the correlation matrix for this set of data. Select the independent variable whose correlation magnitude is the smallest with the dependent variable. Determine if its correlation with the dependent variable is significant.
c. Determine if the overall model is significant. Use a significance level of 0.05.
d. Calculate the variance inflation factor for each of the independent variables. Indicate if multicollinearity exists between the two independent variables.

15-6. Consider the following set of data:

x_2	10	8	11	7	10	11	6
x_1	50	45	37	32	44	51	42
y	103	85	115	73	97	102	65

a. Obtain the estimated regression equation.
b. Examine the coefficient of determination and the adjusted coefficient of determination. Does it seem that either of the independent variables' addition to R^2 does not justify the reduction in degrees of freedom that results from their addition to the regression model? Support your assertions.
c. Conduct a hypothesis test to determine if the dependent variable increases when x_2 increases. Use a significance level of 0.025 and the p-value approach.
d. Construct a 95% confidence interval for the coefficient of x_1.

Computer Database Exercises

15-7. An article in *BusinessWeek* ("Hot Growth Companies," June 5, 2006) presents a list of the 100 companies perceived as having "hot growth" characteristics. A company's rank on the list is the sum of 0.5 times its rank in return on total capital plus 0.25 times its sales and profit-growth ranks. The file entitled **Growth** contains sales ($million), sales increase (%), return on capital, market value ($million), and recent stock price of the top 20 ranked companies.
a. Produce a correlation matrix for the variables contained in the file entitled **Growth**.
b. Select the two variables that are most highly correlated with the recent stock price and produce the regression equation to predict the recent stock price as a function of the two variables you chose.
c. Determine if the overall model is significant. Use a significance level of 0.10.
d. Examine the coefficient of determination and the adjusted coefficient of determination. Does it seem that either of the independent variables' addition to R^2 does not justify the reduction in degrees of freedom that results from their addition to the regression model? Support your assertions.
e. Select the variable that is most correlated with the stock price and test to see if it is a significant predictor of the stock price. Use a significance level of 0.10 and the p-value approach.

15-8. Refer to Exercise 15-7, which referenced an article in *BusinessWeek* ("Hot Growth Companies," June 5, 2006) that presented a list of the 100 companies perceived as having "hot growth" characteristics. The file entitled **Logrowth** contains sales ($million), sales increase (%), return on capital, market value ($million), and recent stock price of the companies ranked from 81 to 100. In Exercise 15.7, stock prices were the focus. Here examine the sales of the companies.
a. Produce a regression equation that will predict the sales as a function of the other variables.
b. Determine if the overall model is significant. Use a significance level of 0.05.
c. Conduct a test of hypothesis to discover if market value should be removed from this model.
d. To see that a variable can be insignificant in one model but very significant in another, construct a regression equation in which sales is the dependent variable and market value is the independent variable. Test the hypothesis that market

value is a significant predictor of sales for those companies ranked from 81 to 100. Use a significance level of 0.05 and the *p*-value approach.

15-9. An investment analyst collected data about 20 randomly chosen companies. The data consisted of the 52-week-high stock prices, PE ratio, and the market value of the company. This data is in the file entitled **Investment**.

a. Produce a regression equation to predict the market value using 52-week-high stock price and the PE ratio of the company.
b. Determine if the overall model is significant. Use a significance level of 0.05.
c. OmniVision Technologies (Sunnyvale, CA) in April 2006 had a 52-week-high stock price of 31 and a PE ratio of 19. Estimate its market value for that time period. (*Note*: Its actual market value for that time period was $1,536.)

15-10. The National Association of Theatre Owners is the largest exhibition trade organization in the world, representing more than 26,000 movie screens in all 50 states and in more than 20 countries worldwide. Its membership includes the largest cinema chains and hundreds of independent theatre owners. It publishes statistics concerning the movie sector of the economy. The file entitled **Flicks** contains data on total U.S. box office grosses ($billion), total number of admissions (billion), average U.S. ticket price ($), and number of movie screens.

a. Construct a regression equation in which total U.S. box office grosses are predicted using the other variables.
b. Determine if the overall model is significant. Use a significance level of 0.05.
c. Determine the range of plausible values for the change in box office grosses if the average ticket price were to be increased by $1. Use a confidence level of 95%.
d. Calculate the variance inflation factor for each of the independent variables. Indicate if multicollinearity exists between any two independent variables.
e. Produce the regression equation suggested by your answer to part d.

15-11. The athletic director of State University is interested in developing a multiple regression model that might be used to explain the variation in attendance at football games at his school. A sample of 16 games was selected from home games played during the past 10 seasons. Data for the following factors were determined:

y = Game attendance
x_1 = Team win/loss percentage to date
x_2 = Opponent win/loss percentage to date
x_3 = Games played this season
x_4 = Temperature at game time

The data collected are in the file called **Football**.

a. Produce scatter plots for each independent variable versus the dependent variable. Based on the scatter plots, produce a model that you believe represents the relationship between the dependent variable and the group of predictor variables represented in the scatter plots.
b. Based on the correlation matrix developed from these data, comment on whether you think a multiple regression model will be effectively developed from these data.
c. Use the sample data to estimate the multiple regression model that contains all four independent variables.
d. What percentage of the total variation in the dependent variable is explained by the four independent variables in the model?
e. Test to determine whether the overall model is statistically significant. Use $\alpha = 0.05$.
f. Which, if any, of the independent variables is statistically significant? Use a significance level of $\alpha = 0.08$ and the *p*-value approach to conduct these tests.
g. Estimate the standard deviation of the model error and discuss whether this regression model is acceptable as a means of predicting the football attendance at State University at any given game.
h. Define the term *multicollinearity* and indicate the potential problems that multicollinearity can cause for this model. Indicate what, if any, evidence there is of multicollinearity problems with this regression model. Use the variance inflation factor to assist you in this analysis.
i. Develop a 95% confidence interval estimate for each of the regression coefficients and interpret each estimate. Comment on whether the interpretation of the intercept is relevant in this situation.

CHAPTER OUTCOME #6

15.2 Using Qualitative Independent Variables

In Example 15-1 involving the First City Real Estate Company, the independent variables were quantitative and ratio level. However, you will encounter many situations in which you may wish to use a qualitative, lower-level variable as an explanatory variable.

If a variable is nominal, and numerical codes are assigned to the categories, you already know not to perform mathematical calculations using those data. The results would be meaningless. Yet, we may wish to use a variable such as marital status, gender, or geographical location as an independent variable in a regression model. If the variable of interest is coded as an ordinal variable, such as education level or job performance ranking, computing means and variances is also inappropriate. Then how are these variables incorporated into a multiple regression analysis? The answer lies in using what are called **dummy** (or indicator) **variables**.

Dummy Variables
A variable that is assigned a value equal to either 0 or 1, depending on whether the observation possesses a given characteristic.

For instance, consider the variable gender, which can take on two possible values: male or female. Gender can be converted to a dummy variable as follows:

$$x_1 = 1 \text{ if female}$$
$$x_1 = 0 \text{ if male}$$

Thus, a data set consisting of males and females will have corresponding values for x_1 equal to 0s and 1s, respectively. Note that it makes no difference which gender is coded 1 and which is coded 0.

If a categorical variable has more than two mutually exclusive outcome possibilities, multiple dummy variables must be created. Consider the variable marital status, with the following possible outcomes:

never married married divorced widowed

In this case, marital status has four values. To account for all the possibilities, you would create three dummy variables, one less than the number of possible outcomes for the original variable. They could be coded as follows:

$$x_1 = 1 \text{ if never married, 0 if not}$$
$$x_2 = 1 \text{ if married, 0 if not}$$
$$x_3 = 1 \text{ if divorced, 0 if not}$$

Note that we don't need the fourth variable because we would know that a person is widowed if $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$. If the person isn't single, married, or divorced, he or she must be widowed. *Always use one fewer dummy variables than categories.* The mathematical reason that the number of dummy variables must be one less than the number of possible responses is called the *dummy variable trap*. Perfect multicollinearity is introduced, and the least squares regression estimates cannot be obtained, if the number of dummy variables equals the number of possible categories.

EXAMPLE 15-2 Incorporating Dummy Variables

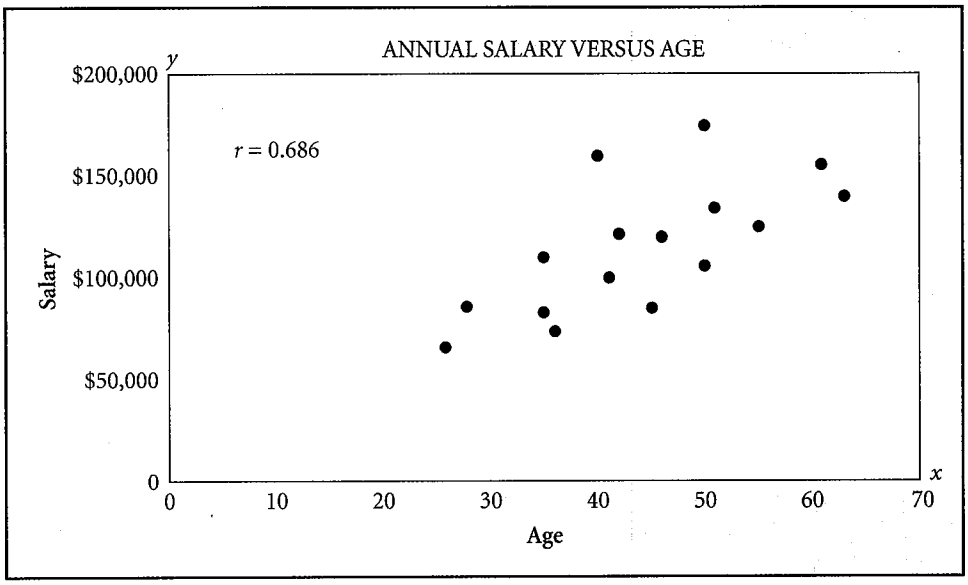
TRY PROBLEM 15.17

Business Executive Salaries To illustrate the effect of incorporating dummy variables into a regression model, consider the sample data displayed in the scatter plot in Figure 15.9. The population from which the sample was selected consists of executives between the ages of 24 and 60 who are working in U.S. manufacturing businesses. Data for annual salary (y) and age (x_1) are available. The objective is to determine whether a model can be generated to explain the variation in annual salary for business executives. Even though age and annual salary are significantly correlated ($r = 0.686$) at the $\alpha = 0.05$ level, the coefficient of determination is only 47%. Therefore, we would likely search for other independent variables that could help us to further explain the variation in annual salary.

TABLE 15.2
Executive Salary Data Including MBA Variable

Salary	Age	MBA
$ 65,000	26	0
85,000	28	1
74,000	36	0
83,000	35	0
110,000	35	1
160,000	40	1
100,000	41	0
122,000	42	1
85,000	45	0
120,000	46	1
105,000	50	0
135,000	51	1
125,000	55	0
175,000	50	1
156,000	61	1
140,000	63	0

FIGURE 15.9
Executive Salary Data—Scatter Plot



Suppose we can determine which of the 16 people in the sample had an MBA degree. Figure 15.10 shows the scatter plot for these same data, with the MBA data represented by triangles. To incorporate a qualitative variable into the analysis, use the following steps:

Step 1 **Code the qualitative variable as a dummy variable.**
Create a new variable, x_2, which is a dummy variable coded as

$$x_2 = 1 \text{ if MBA, } 0 \text{ if not}$$

The data with the new variable are shown in Table 15.2.

Step 2 **Develop a multiple regression model with the dummy variables incorporated as independent variables.**
The two-variable population multiple regression model has the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Using either Excel or Minitab, we get the following regression equation as an estimate of the population model:

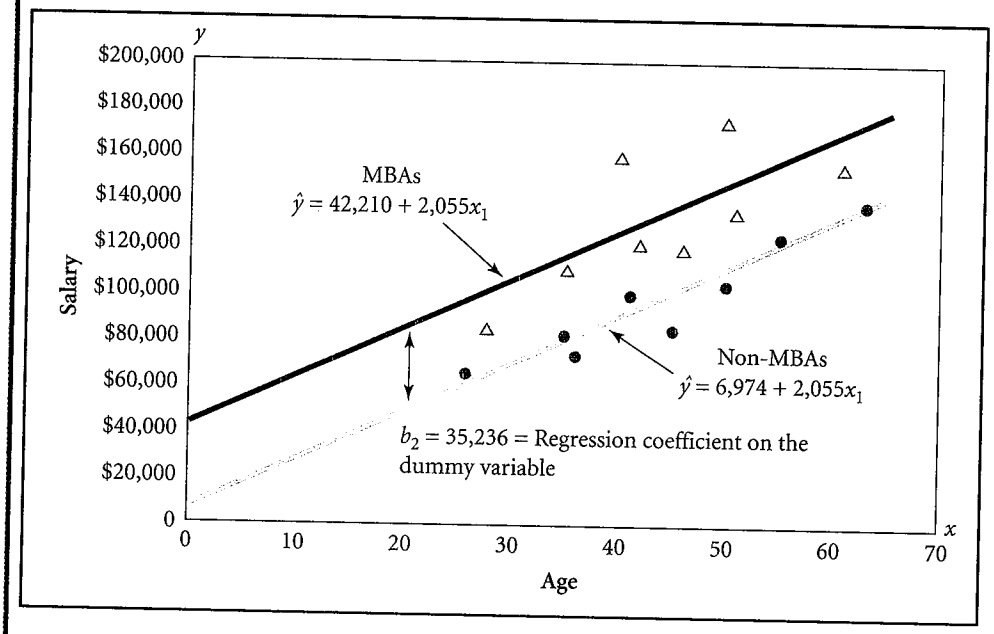
$$\hat{y} = 6{,}974 + 2{,}055x_1 + 35{,}236x_2$$

Because the dummy variable, x_2, has been coded 0 or 1 depending on MBA status, incorporating it into the regression model is like having two simple linear regression lines with the same slopes, but different intercepts. For instance, when $x_2 = 0$, the regression equation is

$$\begin{aligned}\hat{y} &= 6{,}974 + 2{,}055x_1 + 35{,}236(0) \\ &= 6{,}974 + 2{,}055x_1\end{aligned}$$

FIGURE 15.10

Impact of a Dummy Variable



This line is shown in Figure 15.10.

However, when $x_2 = 1$ (the executive has an MBA), the regression equation is

$$\hat{y} = 6,974 + 2,055x_1 + 35,236(1)$$
$$= 42,210 + 2,055x_1$$

This regression line is also shown in Figure 15.10. As you can see, incorporating the dummy variable affects the regression intercept. In this case, the intercept for executives with an MBA degree is $35,236 higher than for those without an MBA. We interpret the regression coefficient on this dummy variable as follows: Based on these data, and holding age (x_1) constant, we estimate that executives with an MBA degree make an average of $35,236 per year more in salary than their non-MBA counterparts.

Business Application

FIRST CITY REAL ESTATE (*CONTINUED*) The regression model developed in Example 15-1 for First City Real Estate showed potential because the overall model was statistically significant. Looking back at Figure 15.8, we see that the model explained nearly 82% of the variation in sales prices for the homes in the sample. All of the independent variables were significant, given that the other independent variables were in the model. However, the standard error of the estimate was quite large.

Excel and Minitab Tutorial

The managers have decided to try to improve the model. First, they have decided to add a new variable: area. However, at this point, the only area variable they can get is whether the house is in the foothills. Because this is a categorical variable with two possible outcomes (foothills or flatland), a dummy variable can be created as follows:

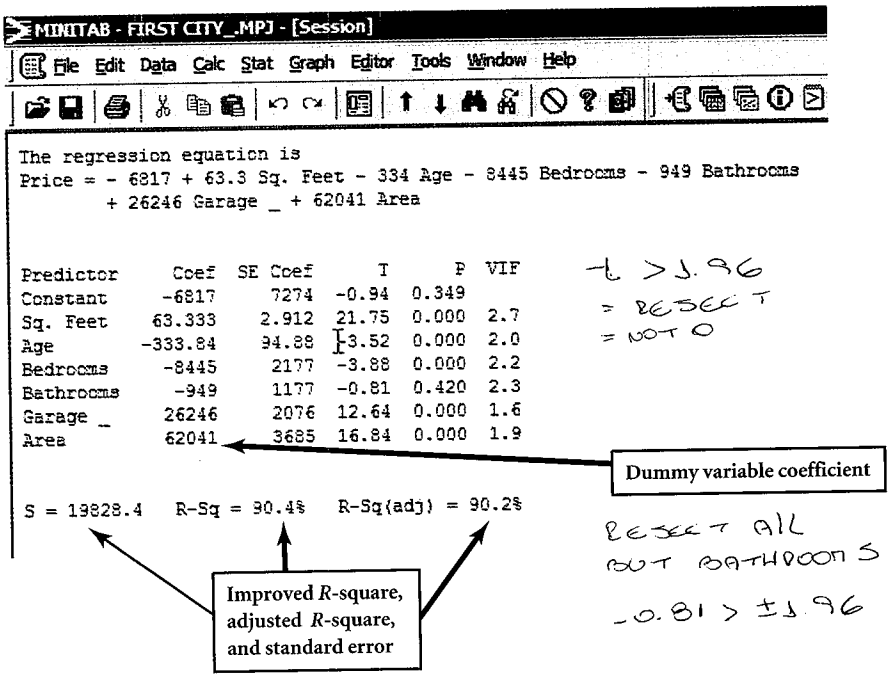
$$x_6 \text{ (area)} = 1 \text{ if foothills, } 0 \text{ if not}$$

Of the 319 homes in the sample, 249 were homes in the foothills and 70 were in the flatland. Figure 15.11 shows the revised Minitab multiple regression with the variable, area, added.

FIGURE 15.11

Minitab Output—First City Real Estate Revised Regression Model

Minitab Instructions:
1. Open file: First City. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (*y*) variable.
4. In **Predictors**, enter Independent (*x*) variables.
5. Click **Options.**
6. In **Display**, select **Variance inflation Factors.**
7. Click **OK. OK.**



This model is an improvement over the original model because the adjusted R-squared has increased from 81.3% to 90.2% and the standard error of the estimate has decreased from \$27,350 to \$19,828. The conditional t-tests show that all of the regression model's slope coefficients, except that for the variable bathrooms, differ significantly from 0. The Minitab output shows that the variance inflation factors are all less than 5.0, so we don't need to be too concerned about the t-tests understating the significance of the regression coefficients. (See the Excel Tutorial on the CD-ROM for this example to get the full *VIF* output from PHStat.)

The resulting regression model is

$$\hat{y} = -6,817 + 63.3(\text{sq. ft.}) - 334(\text{age}) - 8,445(\text{bedrooms}) - 949(\text{bathrooms}) + 26,246\ (\text{garage}) + 62,041(\text{area})$$

Because the variable, bathrooms, is not significant in the presence of the other variables, we can remove the variable and rerun the multiple regression. The resulting model is

$$\text{Price} = -7,050 + 62.5(\text{sq. ft.}) - 322(\text{age}) - 8,830(\text{bedrooms}) + 26,054(\text{garage}) + 61,370(\text{area})$$

Based on the sample data and this regression model, we estimate that a house with the same characteristics (square feet, age, bedrooms, and garages) is worth an average of \$61,370 more if it is located in the foothills (based on how the dummy variable was coded).

There are still signals of multicollinearity problems. The coefficient on the independent variable, bedrooms, is negative, when we would expect homes with more bedrooms to sell for more. Also, the standard error of the estimate is still too large (\$19,817) to provide the precision the managers need to set prices for homes. More work needs to be done before the model is complete.

Possible Improvements to the First City Appraisal Model Because the standard error of the estimate is still too high, we look to improve the model. We could start by identifying possible problems:

1. We may be missing useful independent variables.
2. Independent variables may have been included that should not have been included.

There is no sure way of determining the correct model specification. However, a recommended approach is for the decision maker to try adding variables or removing variables from the model.

We begin by removing the bedrooms variable, which has an unexpected sign on the regression slope coefficient. (*Note*: If the regression model's sole purpose is for prediction, independent variables with unexpected signs do not automatically pose a problem and do not necessarily need to be deleted. However, insignificant variables should be deleted.) The resulting model is shown in Figures 15.12a and 15.12b. Now all the variables in the model have the expected signs. However, the estimate of the model's standard error has increased slightly.

Adding other explanatory variables might help. For instance, consider whether the house has central air conditioning, which might affect sales. If we can identify whether a house has air conditioning, we could add a dummy variable coded as follows:

$$\text{If air conditioning, } x_7 = 1$$
$$\text{If no air conditioning, } x_7 = 0$$

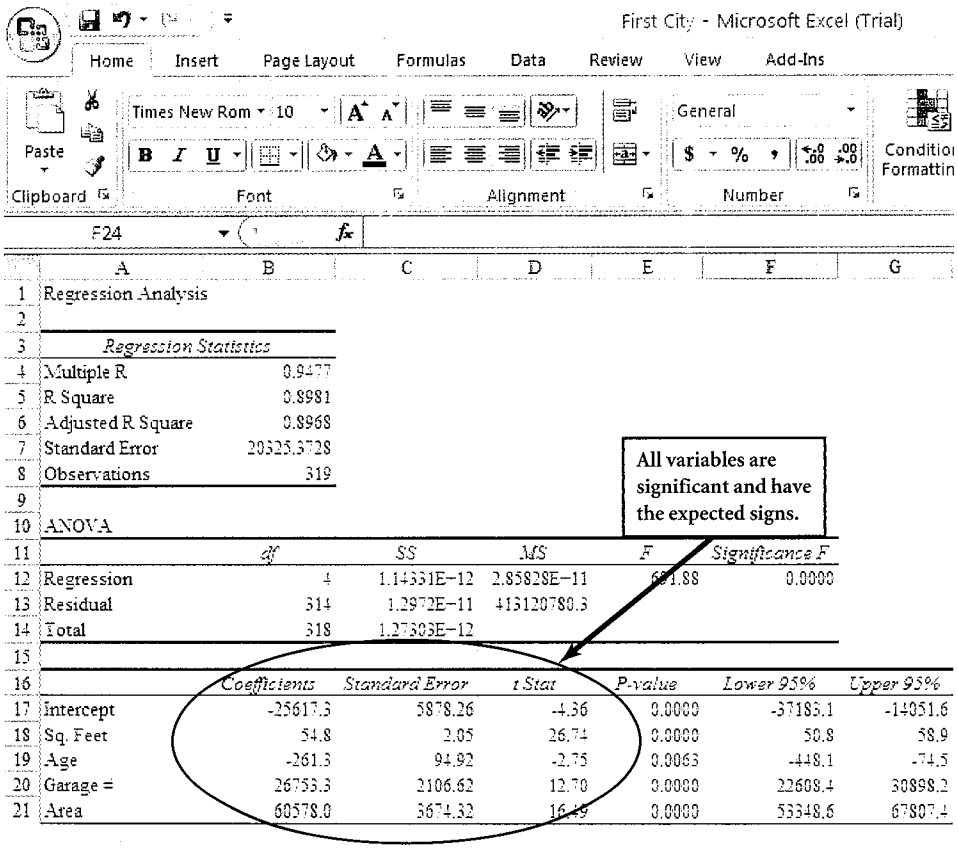
Other potential independent variables might include a more-detailed location variable, a measure of the physical condition, or whether the house has one or two stories. Can you think of others?

The First City example illustrates that even though a regression model may pass the statistical tests of significance, it may not be functional. Good appraisal models can be developed using multiple regression analysis, provided more detail is available about such characteristics as finish quality, landscaping, location, neighborhood characteristics, and so forth. The cost and effort required to obtain these data can be relatively high.

Developing a multiple regression model is more of an art than a science. The real decisions revolve around how to select the best set of independent variables for the model.

FIGURE 15.12A

Excel 2007 Output for the First City Real Estate Revised Model



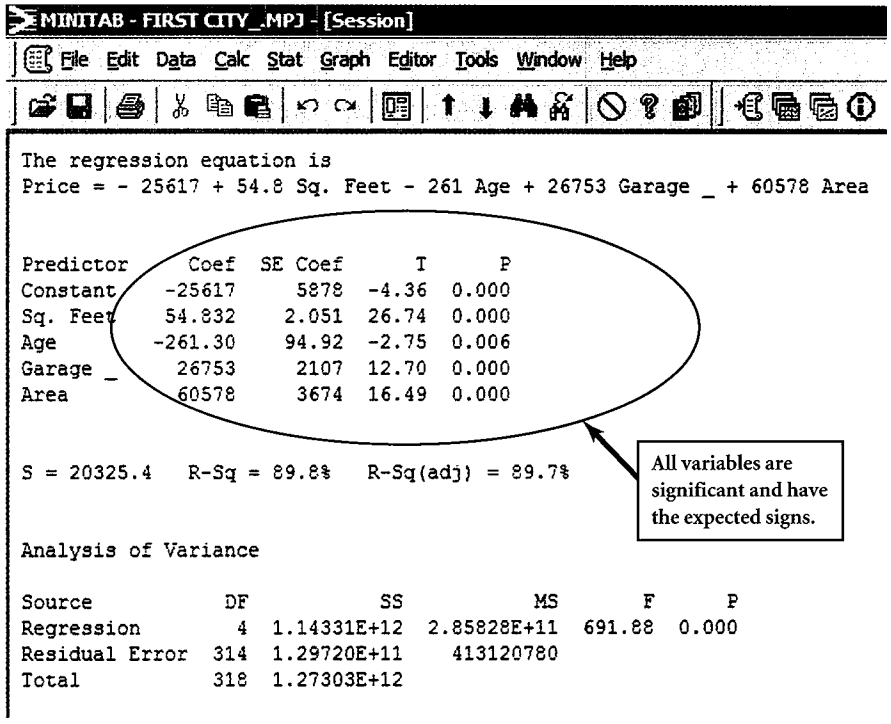
	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.9477					
5	R Square	0.8981					
6	Adjusted R Square	0.8968					
7	Standard Error	20325.3728					
8	Observations	319					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	1.14331E+12	2.85828E+11	691.88	0.0000	
13	Residual	314	1.2972E+11	413120780.3			
14	Total	318	1.27303E+12				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-25617.3	5878.26	-4.36	0.0000	-37183.1	-14051.6
18	Sq. Feet	54.8	2.05	26.74	0.0000	50.8	58.9
19	Age	-261.3	94.92	-2.75	0.0063	-448.1	-74.5
20	Garage =	26753.3	2106.62	12.70	0.0000	22608.4	30898.2
21	Area	60578.0	3674.32	16.49	0.0000	53348.6	67807.4

Excel 2007 Instructions:

1. Open file: First City.xls (worksheet: Homes-Sample-2).
2. Click on **Data** tab—the click on **Data Analysis**.
3. Select **Regression**.
4. Define y variable range and the x variables range.
5. Click **OK**.

FIGURE 15.12B

Minitab Output for the First City Real Estate Revised Model



Minitab Instructions:
1. Open file: First City. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter Independent (x) variables.
5. Click **Options.**
6. In **Display**, select **Variance inflation Factors.**
7. Click **OK. OK.**

15-2: Exercises

Skill Development

15-12. You are considering developing a regression equation relating a dependent variable to two independent variables. One of the variables can be measured on a ratio scale but the other is a categorical variable with two possible levels.
a. Write a multiple regression equation relating the dependent variable to the independent variables.
b. Interpret the meaning of the coefficients in the regression equation.

15-13. You are considering developing a regression equation relating a dependent variable to two independent variables. One of the variables can be measured on a ratio scale but the other is a categorical variable with four possible levels.
a. How many dummy variables are needed to represent the categorical variable?
b. Write a multiple regression equation relating the dependent variable to the independent variables.
c. Interpret the meaning of the coefficients in the regression equation.

15-14. Consider the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

$$x_1 = \text{a quantitative variable}$$

$$x_2 = \begin{cases} 1 \text{ if } x_1 < 20 \\ 0 \text{ if } x_1 \geq 20 \end{cases}$$

The following estimated regression equation was obtained from a sample of 30 observations:

$$\hat{y} = 24.1 + 5.8x_1 + 7.9x_2$$

a. Provide the estimated regression equation for instances in which $x_1 < 20$.
b. Determine the value of $\hat{y}$ when $x_1 = 10$.
c. Provide the estimated regression equation for instances in which $x_1 \geq 20$.
d. Determine the value of $\hat{y}$ when $x_1 = 30$.

15-15. A real estate agent wishes to estimate the monthly rental for apartments based upon the size (square feet) and the location of the apartments. She chose the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

$$x_1 = \text{square footage of the apartment}$$

$$x_2 = \begin{cases} 1 \text{if located in town center} \\ 0 \text{ if not located in town center} \end{cases}$$

This linear regression model was fitted to a sample of size 50 to produce the following regression equation:

$$\hat{y} = 145 + 1.2x_1 + 300x_2$$

a. Predict the average monthly rental for an apartment located in the town center that has 1,500 square feet.
b. Predict the average monthly rent for an apartment located in the suburbs that has 1,500 square feet.
c. Interpret b_2 in the context of this exercise.

Business Applications

15-16. The Polk Utility Corporation is developing a multiple regression model that it plans to use to predict customers' utility usage. The analyst currently has three quantitative variables (x_1, x_2, and x_3) in the model, but she is dissatisfied with the R-squared and the estimate of the standard deviation of the model's error. Two variables that she thinks might be useful are whether the house has a gas water heater or an electric water heater and whether the house was constructed after the 1974 energy crisis or before.

Provide the model she should use to predict customers' utility usage. Specify the dummy variables to be used, the values these variables could assume, and what each value will represent.

15-17. A study was recently performed by the American Automobile Association in which it attempted to develop a regression model to explain variation in EPA mileage ratings of new cars. At one stage of the analysis, the estimate of the model took the following form:

$$\hat{y} = 34.20 - 0.003x_1 + 4.56x_2$$

where:

$$x_1 = \text{Vehicle weight}$$
$$x_2 = \begin{cases} 1, & \text{if standard transmission} \\ 0, & \text{if automatic transmission} \end{cases}$$

a. Interpret the regression coefficient for variable x_1.
b. Interpret the regression coefficient for variable x_2.
c. Present an estimate of a model that would predict the average EPA mileage rating for an automobile with standard transmission as a function of the vehicle's weight.
d. Cadillac's STS-V with automatic transmission weighs approximately 4,394 pounds. Provide an estimate of the average highway mileage you would expect to obtain from this model.
e. Discuss the effect of a dummy variable being incorporated in a regression equation like this one. Use a graph if it is helpful.

15-18. A real estate agent wishes to determine the selling price of residences using the size (square feet) and whether the residence is a condominium or a single-family home. A sample of 20 residences was obtained with the following results:

Price	Type	Square Feet	Price	Type	Square Feet
199,700	Family	1,500	200,600	Condo	1,375
211,800	Condo	2,085	208,000	Condo	1,825
197,100	Family	1,450	210,500	Family	1,650
228,400	Family	1,836	233,300	Family	1,960
215,800	Family	1,730	187,200	Condo	1,360
190,900	Condo	1,726	185,200	Condo	1,200
312,200	Family	2,300	284,100	Family	2,000
313,600	Condo	1,650	207,200	Family	1,755
239,000	Family	1,950	258,200	Family	1,850
184,400	Condo	1,545	203,100	Family	1,630

a. Produce a regression equation to predict the selling price for residences using a model of the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

$$x_1 = \text{square footage} \quad \text{and} \quad x_2 = \begin{cases} 1 \text{ if a condo} \\ 0 \text{ if a single-family home} \end{cases}$$

b. Interpret the parameters β_1 and β_2 in the model given in part a.
c. Produce an equation that describes the relationship between the selling price and the square footage of (1) condominiums, and (2) single-family homes.
d. Conduct a test of hypothesis to determine if the relationship between the selling price and the square footage is different between condominiums and single-family homes.

15-19. J.D. Power and Associates reported ["2004 Initial Quality Study (IQS)," April 28, 2004] on the initial quality of Korean-branded, European-, and domestic-branded vehicles. The 2004 Initial Quality Study was based on responses from more than 62,000 purchasers and lessors of new-model-year cars and trucks, who were surveyed after 90 days of ownership. Initial quality is measured by the number of problems per 100 vehicles (PP100). The PP100 data from the interval 1998–2004 follow:

	1998	1999	2000	2001	2002	2003	2004
Korean	272	227	222	214	172	152	117
Domestic	182	177	164	153	137	135	123
European	158	171	154	141	137	136	122

a. Produce a regression equation to predict the PP100 for vehicles in the model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

$$x_1 = \begin{cases} 1 \text{ if Domestic} \\ 0 \text{ if not Domestic} \end{cases} \quad \text{and} \quad x_2 = \begin{cases} 1 \text{ if European} \\ 0 \text{ if not European} \end{cases}$$

b. Interpret the parameters β_0, β_1, and β_2 in the model given in part a.

c. Conduct a test of hypothesis using the model in part a to determine if the average PP100 is the same for the three international automobile production regions.

Computer Database Exercises

15-20. The Gilmore Accounting firm collected the following data in an effort to explain variation in client profitability. The data are in the data file called **Gilmore**.

y	x_1	x_2
2,345	45	1
4,200	56	2
278	26	3
1,211	56	2
1,406	24	2
500	23	3
−700	34	3
3,457	45	1
2,478	47	1
1,975	24	2
206	32	3

where:

y = Net profit earned from the client
x_1 = Number of hours spent working with the client
x_2 = Type of client:
1, if manufacturing
2, if service
3, if governmental

a. Develop a scatter plot of each independent variable against the client income variable. Comment on what, if any, relationship appears to exist in each case.

b. Run a simple linear regression analysis using only variable x_1 as the independent variable. Describe the resulting estimate fully.

c. Test to determine if the number of hours spent working with the client is useful in predicting the net profit earned by a client.

15-21. Using the data from the Gilmore Accounting firm found in the data file **Gilmore** (see Exercise 15-20),

a. Incorporate the client type into the regression analysis using dummy variables. Describe the resulting multiple regression estimate.

b. Test to determine if this model is useful in predicting the net profit earned from the client.

c. Test to determine if the number of hours spent working with the client is useful in this model in predicting the net profit earned from a client.

d. Considering the tests you have performed, construct a model and its estimate for predicting the net profit earned from the client.

e. Predict the average difference in profit if the client is governmental versus one who is in manufacturing. Also state this in terms of a 95% confidence interval estimate.

15-22. The Energy Information Administration (EIA), created by Congress in 1977, is a statistical agency of the U.S. Department of Energy. It provides data, forecasts, and analyses to promote sound policymaking and public understanding regarding energy and its interaction with the economy and the environment. One of the most important areas of analysis is petroleum. The file entitled **Crude** contains data for the period 1991–2006 concerning the price, supply, and demand for fuel. It has been conjectured that the pricing structure of gasoline changed at the turn of the century.

a. Produce a regression equation to predict the selling price of gasoline

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon$$

where:

$$x_1 = \begin{cases} 1 \text{ if is in 21st century} \\ 0 \text{ if is in 20th century} \end{cases}$$

b. Conduct a hypothesis test to address the conjecture. Use a significance level of 0.05 and the test statistic approach.

c. Produce a 95% confidence interval to estimate the change of the average selling price of gasoline between the 20th and the 21st centuries.

15-23. In a press release entitled "College Board Offers Glimpse of New SAT with Writing for Upcoming Class of '06," August 30, 2005, the College Board announced SAT scores for students in the class of 2005. The file entitled **MathSAT** contains the average SAT scores for the interval 1967 to 2005. The class of 2005 was the last to take the former version of the SAT featuring math and verbal sections. There has been conjecture indicating a relationship between the average math SAT score and the average verbal SAT score and the gender

of the student taking the SAT examination. Consider the following relationship:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

x_1 = average verbal SAT score and $x_2 = \begin{cases} 1 \text{ if a female} \\ 0 \text{ if a male} \end{cases}$

a. Compute the linear regression equation to predict the average math SAT score using the gender and the average verbal SAT score of the students taking the SAT examination.
b. Interpret the parameters in the model.
c. Conduct a hypothesis test to determine if the gender of the student taking the SAT examination is a significant predictor of the student's average math SAT score for a given average verbal SAT score.
d. Predict the average math SAT score of female students with an average verbal SAT score of 500.

CHAPTER OUTCOME #7

15.3 Working with Nonlinear Relationships

Section 14.1 in Chapter 14 showed there are a variety of ways in which two variables can be related. Correlation and regression analysis techniques are tools for measuring and modeling linear relationships between variables. Many situations in business have a linear relationship between two variables, and regression equations that model that relationship will be appropriate to use in these situations. However, there are also many instances in which the relationship between two variables will be curvilinear, rather than linear. For instance, demand for electricity has grown at an almost exponential rate relative to the population growth in some areas. Advertisers believe that a diminishing returns relationship will occur between sales and advertising if advertising is allowed to grow too large. These two situations are shown in Figures 15.13 and 15.14, respectively. They represent just two of the great many possible curvilinear relationships that could exist between two variables.

As you will soon see, models with nonlinear relationships become more complicated than models showing only linear relationships. Although complicated models are sometimes

FIGURE 15.13

Exponential Relationship of Increased Demand for Electricity versus Population Growth

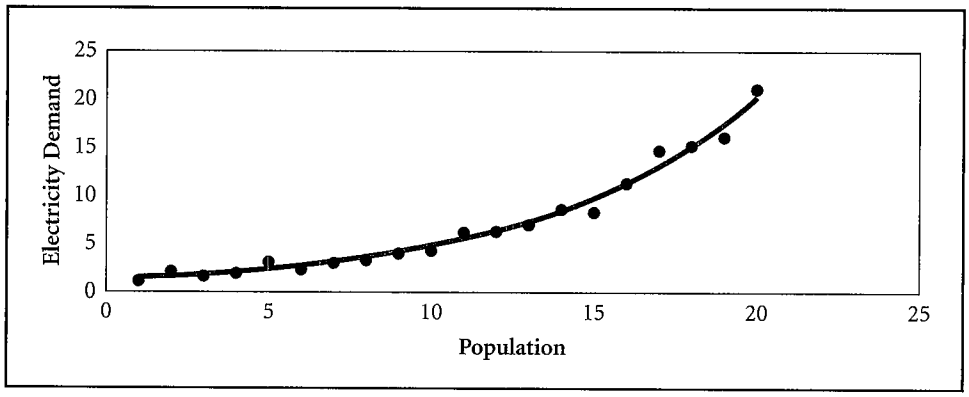
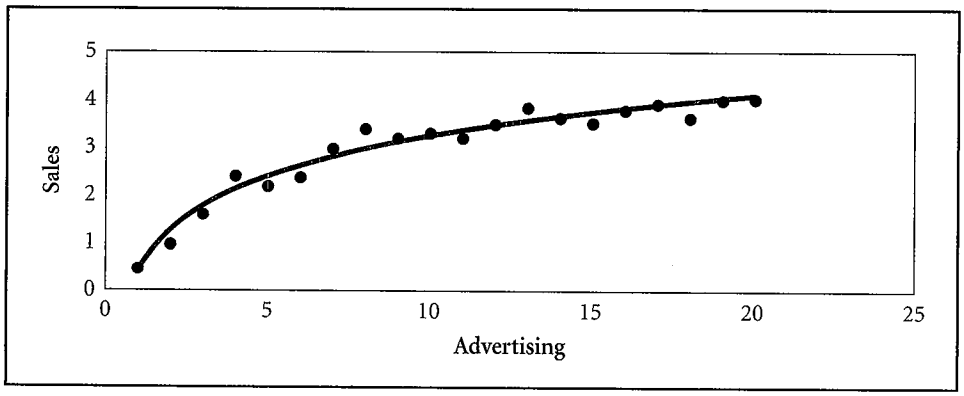


FIGURE 15.14

Diminishing Returns Relationship of Advertising versus Sales



necessary, decision makers should use them with caution for several reasons. First, management researchers and authors have written that people use decision aids they understand and don't use those they don't understand. So, the more complicated a model is, the less likely it is to be used. Second, the scientific principle of parsimony suggests using the simplest model possible that provides a reasonable fit of the data, because complex models typically do not reflect the underlying phenomena that produce the data in the first place.

This section provides a brief introduction into how linear regression analysis can be used in dealing with curvilinear relationships. In order to model such curvilinear relationships, we must incorporate terms into the multiple regression model that will create "curves" in the model we are building. Including terms whose independent variable has an exponent larger than 1 generates these curves. When a model possesses such terms we refer to it as a *polynomial model.* The general equation for a polynomial with one independent variable is given in Equation 15.11.

Polynomial Population Regression Model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon \qquad \textbf{(15.11)}$$

where:

β_0 = Population regression's constant
β_j = Popultion's regression coefficient for variable x x^j; $j = 1, 2, \ldots, p$
p = Order (or degree) of the polynomial
ε = Model error

The order, or degree, of the model is determined by the largest exponent of the independent variable in the model. For instance, the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

is a second-order polynomial because the largest exponent in any term of the polynomial is 2. You will note that this model contains terms of all orders less than or equal to 2. A polynomial with this property is said to be a *complete* polynomial. Therefore, the previous model would be referred to as a complete *second-order regression model.* A second-order model produces a parabola. The parabola either opens upward ($\beta_2 > 0$) or downward ($\beta_2 < 0$), shown in Figure 15.15. You will notice that the models in Figures 15.13, 15.14 and 15.15 possess a single curve.

As more curves appear in the data, the order of the polynomial must be increased. A general (complete) third-order polynomial is given by the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

FIGURE 15.15

Second-Order Regression Models

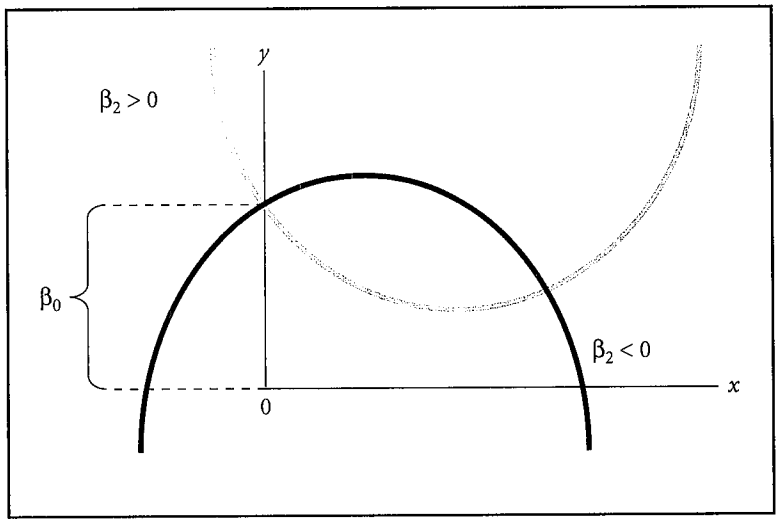
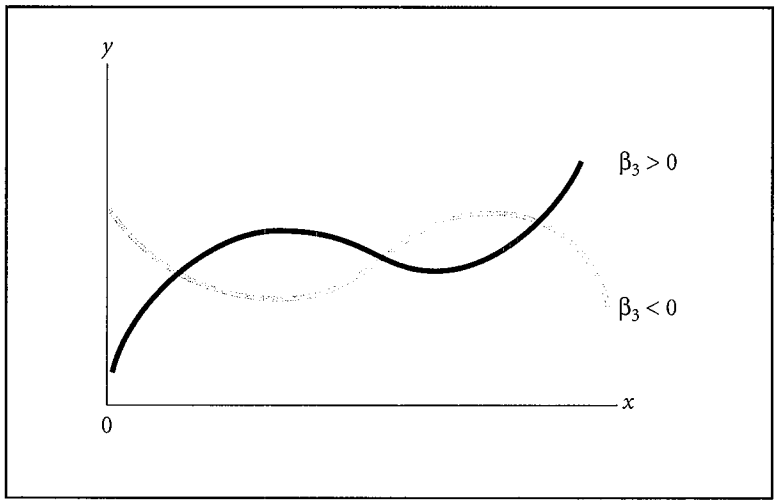


FIGURE 15.16

Third-Order Regression Models



This model produces a curvilinear model that reverses the direction of the initial curve to produce a second curve, as shown in Figure 15.16. Note that there are two curves in the third-order model. In general, a pth-order polynomial will exhibit $p - 1$ curves.

Although polynomials of all orders exist in the business sector, perhaps second-order polynomials are the most common. Sharp reversals in the curvature of a relationship between variables in the business environment usually point to some unexpected or, perhaps, severe changes that were not foreseen. The vast majority of organizations try to avoid such reverses. For this reason, and the fact that this is an introductory business statistics course, we will direct most of our attention to second-order polynomials.

The following examples illustrate two of the most common instances in which curvilinear relationships can be used in decision making. They should give you an idea of how to approach similar situations.

EXAMPLE 15-3 Modeling Curvilinear Relationships

TRY PROBLEM 15.24

CHAPTER OUTCOME #7

Excel and Minitab Tutorial

Ashley Investment Services Ashley Investment Services was severely shaken by the downturn in the stock market after the September 11th attack. To maintain profitability and save as many jobs as possible, since then everyone has been extra busy analyzing new investment opportunities. The director of personnel has noticed an increased number of people suffering from "burnout," in which physical and emotional fatigue hurt job performance. Although he cannot change the job's pressures, he has read that the more time a person spends socializing with coworkers away from the job, the more likely a higher degree of burnout. With the help of the human resources lab at the local university, the personnel director has administered a questionnaire to company employees. A burnout index has been computed from the responses to the survey. Likewise, the survey responses are used to determine quantitative measures of socialization. Sample data from questionnaires are contained in the file **Ashley** on the CD-ROM. The following steps can be used to model the relationship between the socialization index and the burnout index for Ashley employees:

Step 1 **Specify the model by determining the dependent and potential independent variables.**

The dependent variable is the burnout index. The company wishes to explain the variation in burnout level. One potential independent variable is the socialization index.

Step 2 **Formulate the model.**

We begin by proposing that a linear relationship exists between the two variables. Figures 15.17a and 15.17b show the linear regression analysis

results using Excel and Minitab. The correlation between the two variables is $r = 0.818$, which is statistically different from zero at any reasonable significance level. The estimate of the population linear regression model shown in Figures 15.17a is

$$\hat{y} = -66.164 + 9.589x$$

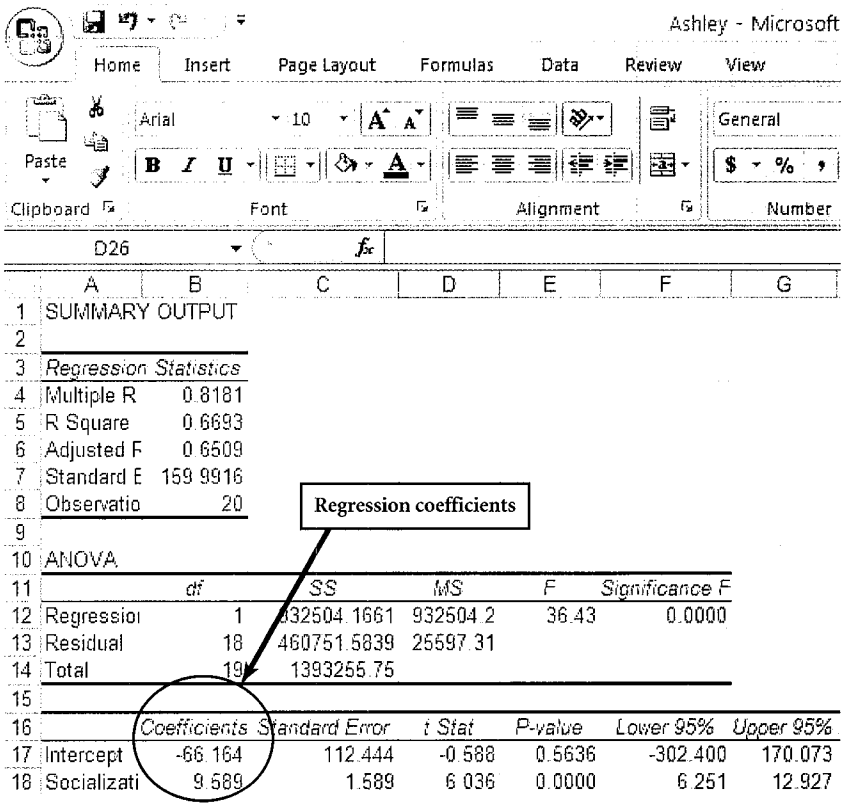
Step 3 Perform diagnostic checks on the model.

The sample data and the regression line are plotted in Figure 15.18. The line appears to fit the data. However, a closer inspection reveals instances where several consecutive points lie above or below the line. The points are not randomly dispersed around the regression line, as should be the case given the regression analysis assumptions. (In Chapter 14 we briefly discussed the concept of residual analysis. Section 15.5 expands the residual analysis discussion.)

As you will recall from earlier discussions, we can use an F-test to test whether a regression model explains a significant amount of variation in the dependent variable.

FIGURE 15.17A

Excel 2007 Output of a Simple Linear Regression for Ashley Investment Services



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8181					
5	R Square	0.6693					
6	Adjusted F	0.6509					
7	Standard E	159.9916					
8	Observatio	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regressio	1	932504.1661	932504.2	36.43	0.0000	
13	Residual	18	460751.5839	25597.31			
14	Total	19	1393255.75				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-66.164	112.444	-0.588	0.5636	-302.400	170.073
18	Socializati	9.589	1.589	6.036	0.0000	6.251	12.927

Excel 2007 Instructions:

1. Open file: Ashley.xls.
2. Select **Data > Data Analysis**.
3. Select **Regression**.
4. Specify y variable range and x variable range (include labels).
5. Check **Labels** option.
6. Specify output location.
7. Click **OK**.

FIGURE 15.17B

Minitab Output of a Simple Linear Regression for Ashley Investment Services

Minitab Instructions:
1. Open file: Ashley.MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response**, enter the y variable column.
4. In **Predictors**, enter the x variable column.
5. Click **OK**.

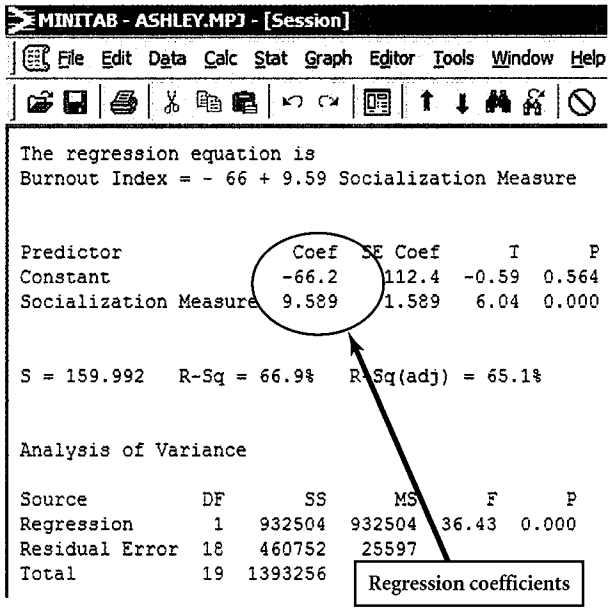
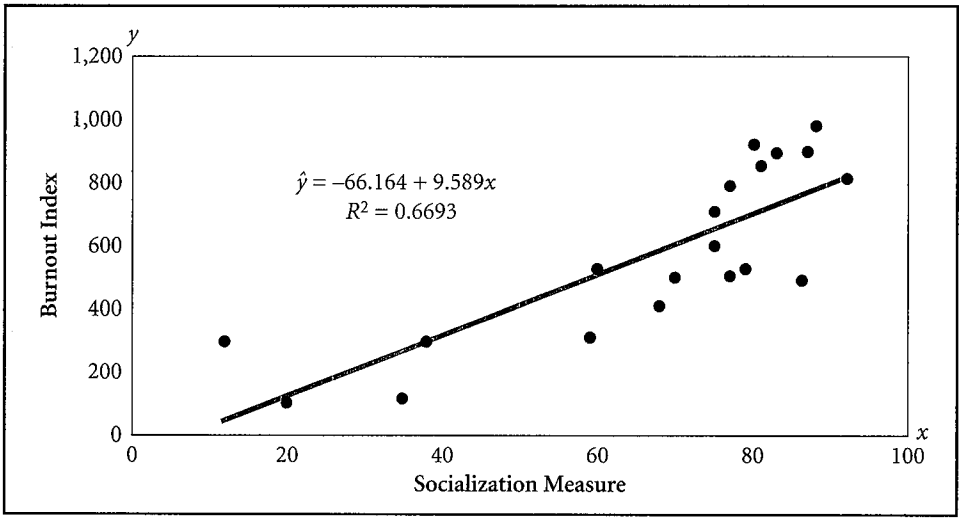


FIGURE 15.18

Plot of Regression Line for the Ashley Investment Services Example



$$H_0: \rho^2 = 0$$
$$H_A: \rho^2 > 0$$

From the output in Figure 15.17a,

$$F = 36.43$$

which has a p-value ≈ 0.0000.

Thus, we conclude that the simple linear model is statistically significant. However, we should also examine the data to determine if any curvilinear relationships may be present.

Step 4 Model the curvilinear relationship.

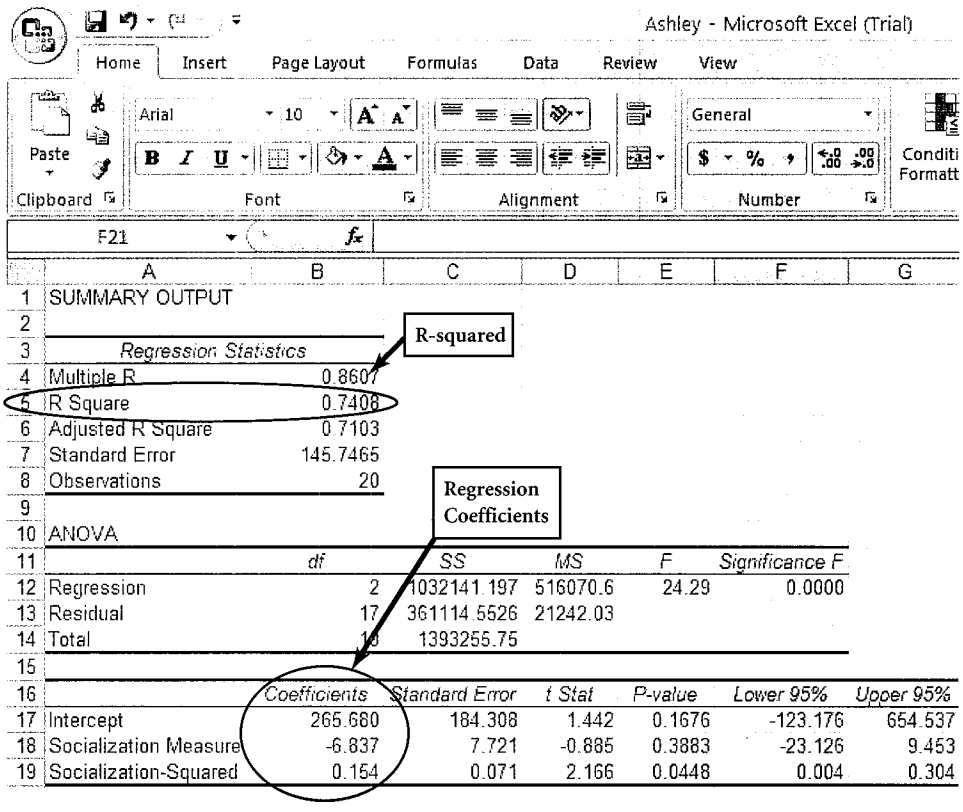
Finding instances of nonrandom patterns in the residuals for a regression model indicates the possibility of using a curvilinear relationship rather than a linear one. One possible approach to modeling the curvilinear nature of the data in the Ashley Investments example is with the use of polynomials. From Figure 15.18, we can see that there appears to be one curve in the data. This suggests fitting the second-order polynomial

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Before fitting the estimate for this population model, you will need to create the new independent variable by squaring the socialization measure variable. In Excel, use the formula option, or in Minitab, use the **Calc > Calculator** command to create the new variable. Figures 15.19a and 15.19b show the output after fitting this second-order polynomial model.

FIGURE 15.19A

Excel 2007 Output of a Second-Order Polynomial Fit for Ashley Investment



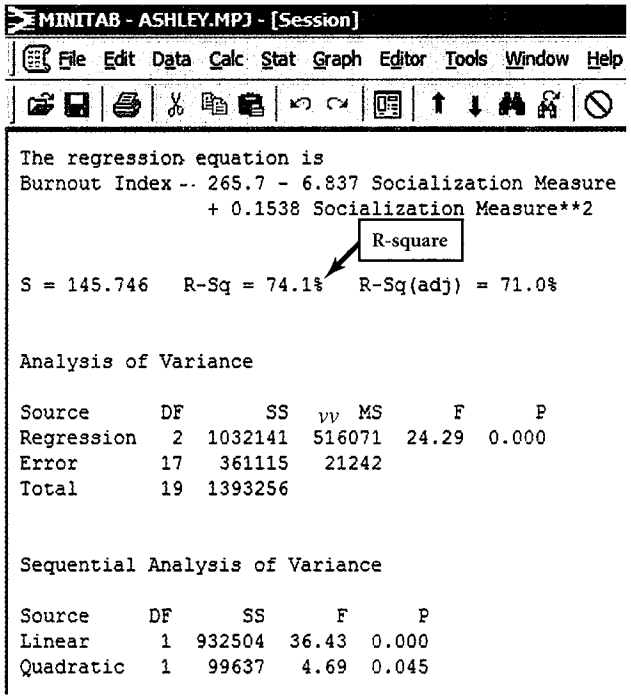
	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8607					
5	R Square	0.7408					
6	Adjusted R Square	0.7103					
7	Standard Error	145.7465					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	1032141.197	516070.6	24.29	0.0000	
13	Residual	17	361114.5526	21242.03			
14	Total	19	1393255.75				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	265.680	184.308	1.442	0.1676	-123.176	654.537
18	Socialization Measure	-6.837	7.721	-0.885	0.3883	-23.126	9.453
19	Socialization-Squared	0.154	0.071	2.166	0.0448	0.004	0.304

Excel 2007 Instructions:

1. Open file: Ashley.xls.
2. Use Excel equations to create the new variable in column C (i.e. for the first data value use = **A2^2**. Then copy down).
3. Select **Data > Data Analysis**.
4. Select **Regression**.
5. Specify *y* variable range and *x* variable range (include the new variable and the labels).
6. Check **Labels** option.
7. Specify ouput location.
8. Click **OK**.

FIGURE 15.19B

Minitab Output of a Second-Order Polynomial Fit for Ashley Investment



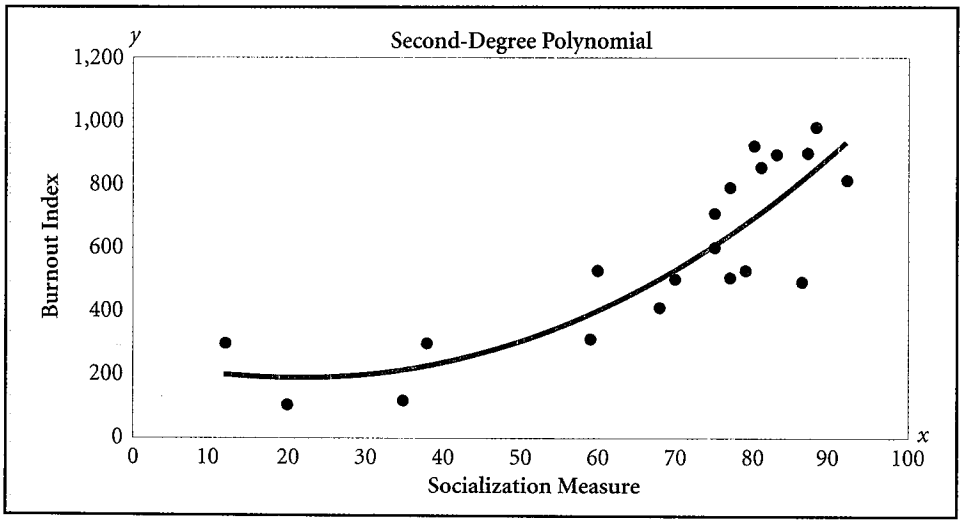
Minitab Instructions:

1. Open file: Ashley.MTW.
2. Choose **Stat > Regression > Fitted Line Plot.**
3. In **Response**, enter *y* variable.
4. In **Predictor**, enter *x* variable.
5. Under **Type of Regression Model**, Choose **Quadratic.**
6. Click **OK.**

Step 5 Perform diagnostics on the revised curvilinear model.
Notice the second-order polynomial provides a model whose estimated regression equation has an R^2 of 74.1%. This is higher than the R^2 of 66.9% for the linear model. Figure 15.20 shows the plot of the second-order polynomial model. Comparing Figure 15.20 with Figure 15.18, we can see that the polynomial model does appear to fit the sample data better than the linear model.

FIGURE 15.20

Plot of Second-Order Polynomial Model for Ashley Investment



Analyzing Interaction Effects

Business Application

Excel and Minitab Tutorial

ASHLEY INVESTMENT SERVICES (*CONTINUED*) Referring to Example 15-3 involving Ashley Investment Services, the director of personnel wondered if the effects of burnout differ among male and female workers. He therefore identified the gender of the previously surveyed employees (see the CD-ROM file **Ashley-2**). A multiple scatter plot of the data appears as Figure 15.21.

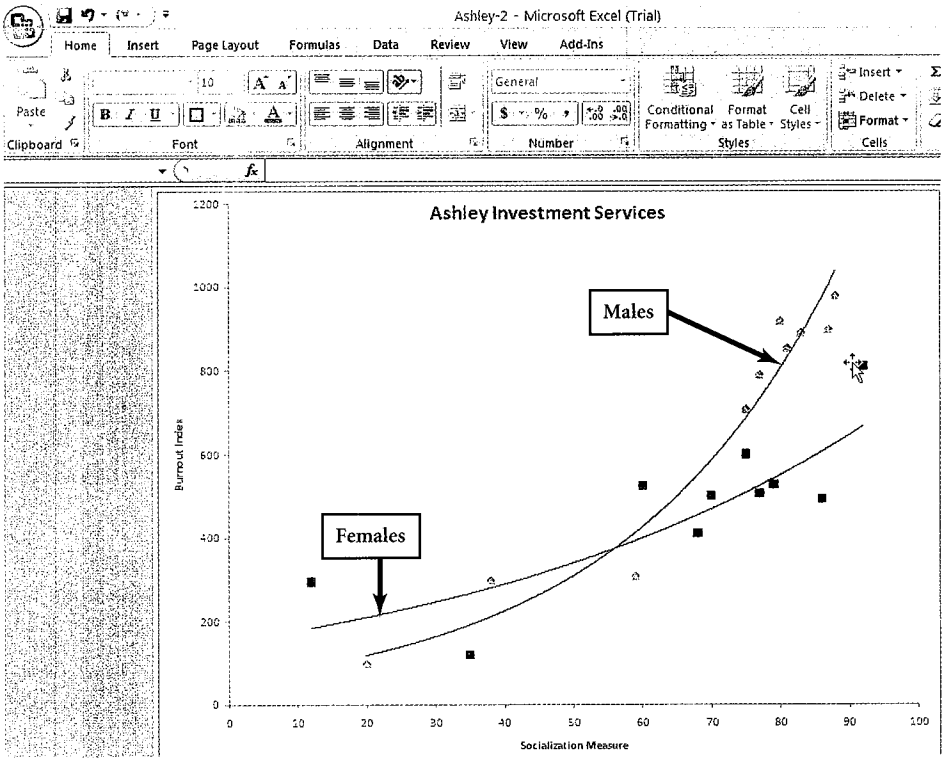
The personnel director tried to determine the relationship between the burnout index and socialization measure for men and women. The graphical result is presented in Figure 15.21. Note that both relationships appear to be curvilinear with a similarly shaped curve. As we showed earlier, curvilinear shapes often can be modeled by the second-order polynomial

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

However, the regression equations that estimate this second-order polynomial for men and women are not the same. The two equations seem to have different locations and different rates of curvature. Whether an employee is a man or woman seems to change the basic relationship between burnout index (y) and socialization measure (x_1). In order to

FIGURE 15.21

Excel 2007 Multiple Scatter Plot for Ashley Investment Services



Excel 2007 Instructions:

1. Open file: Ashley-2.xls.
2. Select the Socialization Measure and Burnout Index Columns.
3. Select the **Insert** tab.
4. Select the **XY (Scatter)**.
5. Select **Chart** and click the right mouse button—choose **Select Data**.
6. Click on **Add** on the **Legend Entry (Series)** section.
7. Enter **Series Name**—(Females)—For **Series X Values** select data from Socialization column for row corresponding to females (rows 2–11). For **Series Y Values** select data from the Burnout column corresponding to females (rows 2–11).
8. Repeat step 7 for males.
9. Click on **layout** tab to remove legend and to add chart and axis titles.
10. Select data points for males—right click and select **Add Trendline > Exponential**.
11. Repeat Step 10 for females.

represent this difference, the equation's coefficients b_0, b_1, and b_2 must be different for men and women employees. Thus, we could use two models, one for each gender. Alternatively, we could use one model for both males and females by incorporating a dummy independent variable with two levels, which is shown as

$$x_2 = 1 \text{ if a male, } 0 \text{ if a female}$$

As x_2 changes values from 0 to 1, it affects the values of the coefficients b_0, b_1, and b_2. Suppose the director fitted the second-order model for the female employees only. He obtained the following regression equation:

$$\hat{y} = 291.70 - 4.62x_1 + 0.102x_1^2$$

The equation for the male employees only was

$$\hat{y} = 149.59 - 4.40x_1 + 0.160x_1^2$$

Interaction
The case in which one independent variable (such as x_2) affects the relationship between another independent variable (x_1) and a dependent variable (y).

To explain how a change in gender can cause this kind of change, we must introduce the concept of **interaction**. Therefore, in our example, gender (x_2) interacts with the relationship between socialization measure (x_1) and burnout index (y). The question is how do we obtain the interaction terms to model such a relationship? To answer this question, we first obtain the model for the basic relationship between the x_1 and the y variables. The population model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

To obtain the interaction terms, multiply the terms on the right-hand side of this equation by the variable that is interacting with this relationship between y and x_1. In this case, that interacting variable is x_2. Then the interaction terms would be

$$\beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2$$

Notice that we have changed the coefficient subscripts so we do not duplicate those in the original model. Then the interaction terms are added to the original model to produce the **composite model**.

Composite Model
The model that contains both the basic terms and the interaction terms.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

Note, the model for women is obtained by substituting $x_2 = 0$ into the composite model. This gives

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1^2(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \end{aligned}$$

Similarly, for men we substitute the value of $x_2 = 1$. The model then becomes

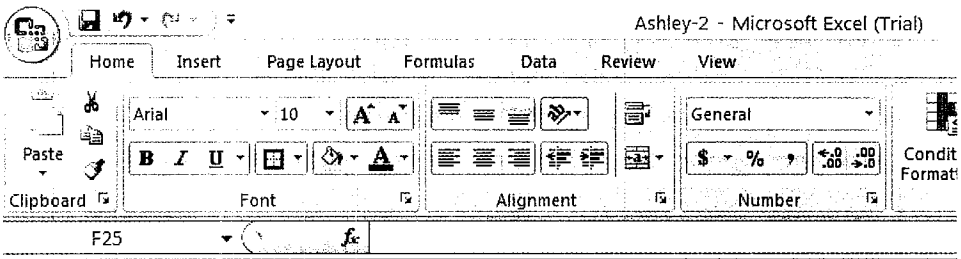
$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3(1) + \beta_4 x_1(1) + \beta_5 x_1^2(1) + \varepsilon \\ &= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) x_1 + (\beta_2 + \beta_5) x_1^2 + \varepsilon \end{aligned}$$

This illustrates how the coefficients are changed for different values of x_2, and therefore, how x_2 is interacting with the relationship between x_1 and y. Once we know β_3, β_4, and β_5, we know the effect of the interaction of gender on the original relationship between the burnout index (y) and the socialization measure (x_1). In order to estimate the composite model, we need to create the required variables, as shown in Figure 15.22. Figures 15.23a and 15.23b show the regression for the composite model. The estimate for the composite model is

$$\hat{y} = 291.706 - 4.615x_1 + 0.102x_1^2 - 142.113x_2 + 0.215x_1x_2 + 0.058x_1^2x_2$$

FIGURE 15.22

Excel 2007 Data Preparation for Estimating Interactive Effects for Second-Order Model for Ashley Investment



	A	B	C	D	E	F
1	Socialization Measure x_1	Burnout Index y	Socialization Squared x_1^2	Gender x_2	x_1x_2	$x_1^2x_2$
2	12	296	144	0	0	0
3	35	120	1225	0	0	0
4	60	525	3600	0	0	0
5	68	410	4624	0	0	0
6	70	501	4900	0	0	0
7	75	600	5625	0	0	0
8	77	506	5929	0	0	0
9	79	527	6241	0	0	0
10	86	493	7396	0	0	0
11	92	810	8464	0	0	0
12	20	100	400	1	20	400
13	38	300	1444	1	38	1444
14	59	310	3481	1	59	3481
15	75	709	5625	1	75	5625
16	77	791	5929	1	77	5929
17	80	920	6400	1	80	6400
18	81	855	6561	1	81	6561
19	83	892	6889	1	83	6889
20	87	900	7569	1	87	7569
21	88	980	7744	1	88	7744

Excel 2007 Instructions:
1. Open file: Ashley-2.xls.
2. Use Excel formulas to create new variables in columns C, E and F.

We obtain the model for females by substituting $x_2 = 0$, giving

$$\hat{y} = 291.706 - 4.615x_1 + 0.102x_1^2 - 142.113(0) + 0.215x_1(0) + 0.058x_1^2(0)$$
$$\hat{y} = 291.706 - 4.615x_1 + 0.102x_1^2$$

For males, we substitute $x_2 = 1$, giving

$$\hat{y} = 291.706 - 4.615x_1 + 0.102x_1^2 - 142.113(1) + 0.215x_1(1) + 0.058x_1^2(1)$$
$$\hat{y} = 149.593 - 4.40x_1 + 0.160x_1^2$$

Note that these equations for males and females are the same as those we found earlier when we generated two separate regression models, one for each gender.

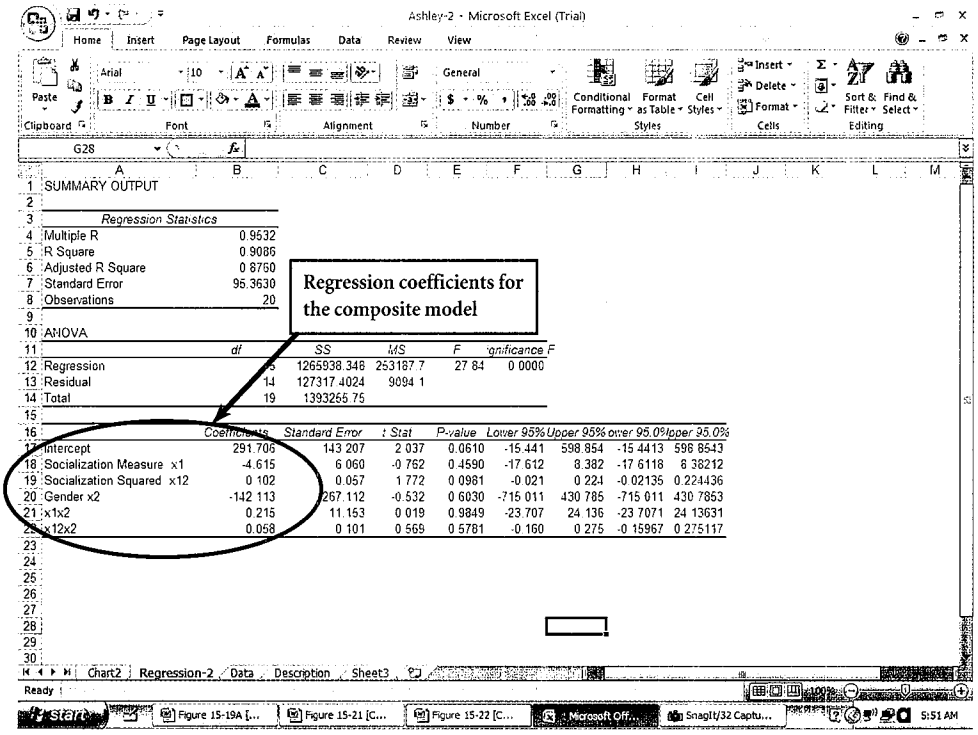
In this example we have looked at a case in which a dummy variable interacts with the relationship between another independent variable and the dependent variable. However, the interacting variable need not be a dummy variable. It can be any independent variable. Also, strictly speaking, interaction is not said to exist if the only effect of the interaction variable is to change the y intercept of the equation relating another independent variable to the dependent variable. Therefore, when you search a scatter plot to detect interaction, you are trying to determine if the relationships produced, when the interaction variable changes values, are parallel or not. If the relationships are parallel, that indicates that only the y intercept is being affected by the change of the interacting variable and that interaction does not exist. Figure 15.24 demonstrates this concept graphically.

FIGURE 15.23A

Excel 2007 Composite Model for Ashley Investment Services

Excel 2007 Instructions:

1. Open file: Ashley-2.xls.
2. Create new variables (see Figure 15.22 Excel 2007 Instructions).
3. Rearrange columns so all x variables are contiguous.
4. Select **Data > Data Analysis**.
5. Select **Regression**.
6. Specify y variable range and x variable range (include the new variables and the labels).
7. Check **Labels** option.
8. Specify output location.
9. Click **OK**.



SUMMARY OUTPUT

Regression Statistics

Multiple R	0.9532
R Square	0.9086
Adjusted R Square	0.8760
Standard Error	95.3630
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1265938.348	253187.7	27.84	0.0000
Residual	14	127317.4024	9094.1		
Total	19	1393256.75			

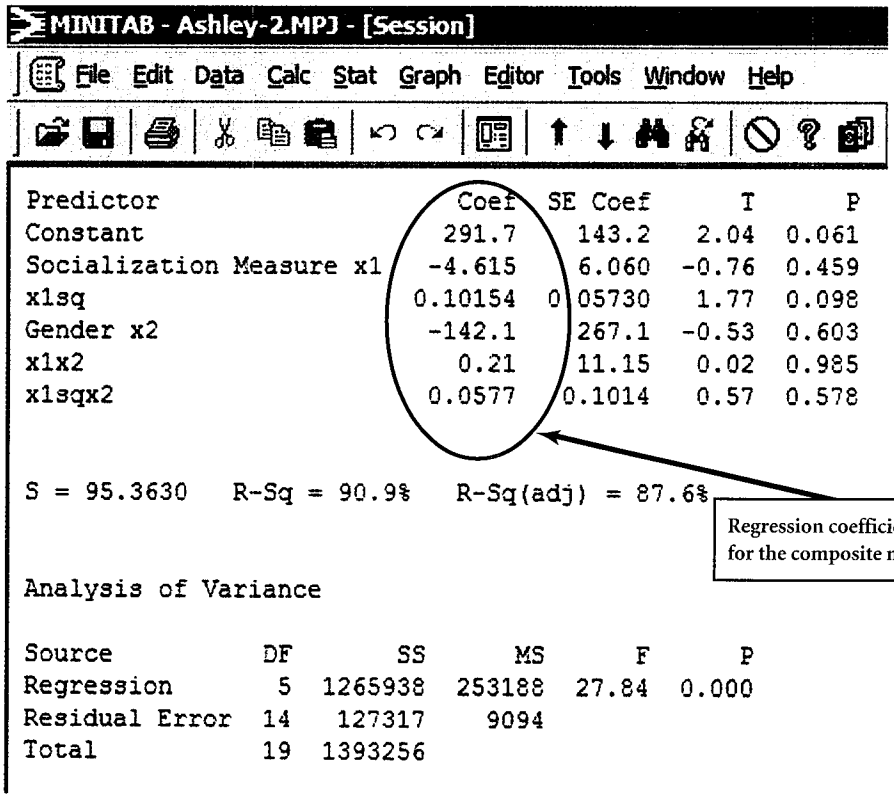
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	291.706	143.207	2.037	0.0610	-15.441	598.854	-15.4413	598.8543
Socialization Measure x1	-4.615	6.060	-0.762	0.4590	-17.612	8.382	-17.6118	8.38212
Socialization Squared x12	0.102	0.057	1.772	0.0981	-0.021	0.224	-0.02135	0.224436
Gender x2	-142.113	267.112	-0.532	0.6030	-715.011	430.785	-715.011	430.7853
x1x2	0.215	11.153	0.019	0.9849	-23.707	24.136	-23.7071	24.13631
x12x2	0.058	0.101	0.569	0.5781	-0.160	0.275	-0.15967	0.275117

FIGURE 15.23B

Minitab Composite Model for Ashley Investment Services

Minitab Instructions:

1. Continue from Figure 15.19b.
2. Choose **Stat > Regression > Regression**.
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Click **OK**.



Predictor	Coef	SE Coef	T	P
Constant	291.7	143.2	2.04	0.061
Socialization Measure x1	-4.615	6.060	-0.76	0.459
x1sq	0.10154	0.05730	1.77	0.098
Gender x2	-142.1	267.1	-0.53	0.603
x1x2	0.21	11.15	0.02	0.985
x1sqx2	0.0577	0.1014	0.57	0.578

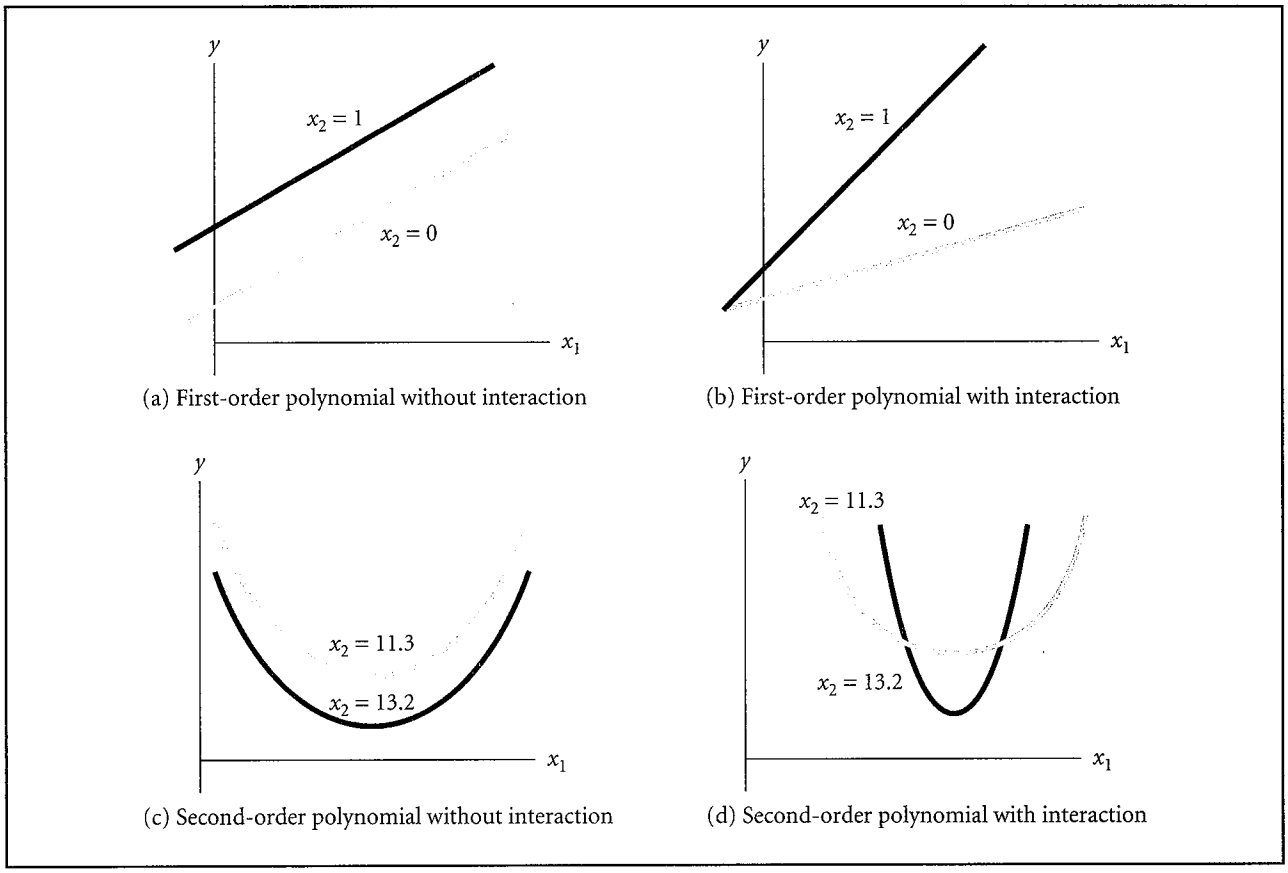
S = 95.3630 R-Sq = 90.9% R-Sq(adj) = 87.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	1265938	253188	27.84	0.000
Residual Error	14	127317	9094		
Total	19	1393256			

FIGURE 15.24

Graphical Evidence of Interaction



15-3: Exercises

Skill Development

15-24. Consider the following values for the dependent and independent variables:

x	y
5	10
15	15
40	25
50	44
60	79
80	112

a. Develop a scatter plot of the data. Does the plot suggest a linear or nonlinear relationship between the dependent and independent variables?

b. Develop an estimated linear regression equation for the data. Is the relationship significant? Test at an $\alpha = 0.05$ level.

c. Develop a regression equation of the form $\hat{y} = b_0 + b_1x + b_2x^2$. Does this equation provide a better fit to the data than that found in part b?

15-25. Consider the following values for the dependent and independent variables:

x	y
6	5
9	20
14	28
18	30
22	33
27	35

a. Develop a scatter plot of the data. Does the plot suggest a linear or nonlinear relationship between the dependent and independent variables?

b. Develop an estimated linear regression equation for the data. Is the relationship significant? Test at an $\alpha = 0.05$ level.
c. Develop a regression equation of the form $\hat{y} = b_0 + b_1 \ln(x)$. Does this equation provide a better fit to the data than that found in part b?

15-26. Examine the following data:

x	1	4	5	7	8	12	11	14	19	20
y	1	54	125	324	512	5530	5331	5740	7058	7945

a. Construct a scatter plot of the data. Determine the order of the polynomial that is represented by the data.
b. Obtain an estimate of the model identified in part a.
c. Conduct a test of hypothesis to determine if a third-order, as opposed to a second-order, polynomial is a better representation of the relationship between y and x. Use a significance level of 0.05 and the p-value approach.

15-27. Examine the following two sets of data:

When $x_2 = 1$		When $x_2 = 2$	
x_1	y	x_1	y
1	2	2	3
4	15	3	9
5	23	6	5
7	52	7	10
8	60	9	48
12	154	10	50
11	122	14	87
14	200	13	51
19	381	16	63
20	392	21	202

a. Produce a distinguishable scatter plot for each of the data sets on the same graph. Does it appear that there is interaction between x_2 and the relationship between y and x_1? Support your assertions.
b. Consider the following model to represent the relationship among y, x_1, and x_2:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 x_2 + \varepsilon$$

Produce the estimated regression equation for this model.
c. Conduct a test of hypothesis for each interaction term. Use a significance level of 0.05 and the p-value approach.
d. Based upon the two hypothesis tests in part c, does it appear that there is interaction between x_2 and the relationship between y and x_1? Support your assertions.

Computer Database Exercises

15-28. The Gilmore Accounting firm collected the following in an effort to explain variation in client profitability. The data are also in the data file called **Gilmore**.

y	x_1	x_2
2,345	45	1
4,200	56	2
278	26	3
1,211	56	2
1,406	24	2
500	23	3
−700	34	3
3,457	45	1
2,478	47	1
1,975	24	2
206	32	3

where:
y = Net profit earned from the client
x_1 = Number of hours spent working with the client
x_2 = Type of client:
1, if manufacturing
2, if service
3, if governmental

Gilmore has asked if it needs the client type in addition to the number of hours spent working with the client to predict the net profit earned from the client. You are asked to provide this information.
a. Fit a model to the data that incorporates the number of hours spent working with the client and the type of client as independent variables. (*Hint*: Client type has three levels.)
b. Fit a second-order model to the data, again using dummy variables for client type. Does this model provide a better fit than that found in part a? Which model would you recommend be used?

15-29. McCullom's International Grains is constantly searching out areas in which to expand its market. Such markets present different challenges since tastes in the international market are often different from domestic tastes. India is one country on which McCullom's has recently focused. Paddy is a grain used widely in India but its characteristics are unknown to McCullom's. Charles Walters has been

assigned to take charge of the handling of this grain. He has researched its various characteristics. During his research he came across an article, "Determination of Biological Maturity and Effect of Harvesting and Drying conditions on Milling Quality of Paddy" [*Journal of Agricultural Engineering Research* (1975), pp. 353–361], which examines the relationship between y, the yield (kg/ha) of paddy, as a function of x, the number of days after flowering at which harvesting took place. The accompanying data appeared in the article and are in a file called **Paddy** on the CD-ROM.

y	x	y	x
2,508	16	3,823	32
2,518	18	3,646	34
3,304	20	3,708	36
3,423	22	3,333	38
3,057	24	3,517	40
3,190	26	3,241	42
3,500	28	3,103	44
3,883	30	2,776	46

a. Construct a scatter plot of the yield (kg/ha) of paddy as a function of the number of days after flowering at which harvesting took place. Display at least two models that would explain the relationship you see in the scatter plot.
b. Conduct tests of hypotheses to determine if the models you selected are useful in predicting the yield of paddy.
c. Consider the model that includes the second-order term x^2. Would a simple linear regression model be preferable to the model containing the second-order term? Conduct a hypothesis test using the p-value approach to arrive at your answer.
d. Which model should Charles use to predict the yield of paddy? Explain your answer.

15-30. The National Association of Realtors Existing-Home Sales Series provides a measurement of the residential real estate market. One of the measurements it produces is the Housing Affordability Index (HAI). It is a measure of the financial ability of U.S. families to buy a house: 100 means that families earning the national median income have just the amount of money needed to qualify for a mortgage on a median-priced home; higher than 100 means they have more than enough, and lower than 100 means they have less than enough. The file entitled **Index** contains the HAI and associated variables.

a. Construct a scatter plot relating the HAI to the median family income.
b. Determine the order of the polynomial that is suggested by the scatter plot produced in part a. Obtain the estimated regression equation of the polynomial selected.
c. Determine if monthly principle and interest interacts with the relationship between the HAI and the median family income indicated in part b. (*Hint*: You may need to conduct more than one hypothesis test.) Use a significance level of 0.05 and the p-value approach.

15-31. American men have closed the gap with women on life span according to a *USA Today* article (Kim Painter, "Male life span increasing," June 12, 2006). Male life expectancy attained a record 75.2 years and women's reached 80.4 in 2004. The National Center for Health Statistics provided the data given in the file entitled **Life**.

a. Produce a scatter plot depicting the relationship between the life expectancy of women and men.
b. Determine the order of the polynomial that is represented on the scatter plot obtained in part a. Produce the estimated regression equation that represents this relationship.
c. Determine if women's average life expectancy can be used in a second-order polynomial to predict the average life expectancy of men. Use a significance level of 0.05.
d. Use the estimated regression equation computed in part b to predict the average length of life of men when women's length of life equals 100. What does this tell you about the wisdom (or lack thereof) of extrapolation in regression models?

15-32. Badeaux Brothers Louisiana Treats ships packages of Louisiana coffee, cakes, and Cajun spices to individual customers around the United States. The cost to ship these products depends primarily on the weight of the package being shipped. Badeaux charges the customers for shipping and then ships the product itself. As a part of a study of whether it is economically feasible to continue to ship products itself, Badeaux sampled 20 recent shipments to determine what, if any, relationship exists between shipping costs and package weight. These data are in the file called **Badeaux**.

a. Develop a scatter plot of the data with the dependent variable, cost, on the vertical axis and the independent variable, weight, on the horizontal axis. Does there appear to be a relationship between the two variables? Is the relationship linear?
b. Compute the sample correlation coefficient between the two variables. Conduct a test, using a significance level of 0.05, to determine whether the population correlation coefficient is significantly different from zero.

c. Badeaux Brothers has been using a simple linear regression equation to predict the cost of shipping various items. Would you recommend it use a second-order polynomial model instead? Is the second-order polynomial model a significant improvement on the simple linear regression equation?
d. Badeaux Brothers has made a decision to stop shipping products if the shipping charges exceed $100. The company has asked you to determine the maximum weight for future shipments. Do this for both the first- and second-order models you have developed.

15-33. The National Association of Theatre Owners is the largest exhibition trade organization in the world, representing more than 26,000 movie screens in all 50 states and in more than 20 countries worldwide. Its membership includes the largest cinema chains and hundreds of independent theatre owners. It publishes statistics concerning the movie sector of the economy. The file entitled **Flicks** contains data on total U.S. box office grosses ($billion), total number of admissions (billion), average U.S. ticket price ($), and number of movie screens. One concern is the effect the increasing ticket prices have upon the number of individuals who go to the theatres to view movies.
a. Construct a scatter plot depicting the relationship between the total number of admissions and the U.S. ticket price.
b. Determine the order of the polynomial that is suggested by the scatter plot produced in part a. Obtain the estimated regression equation of the polynomial selected.
c. Examine the p-value associated with the F-test for the polynomial you have selected in part a. Relate these results to those of the t-tests for the individual parameters and the adjusted coefficient of determination. To what is this attributed?
d. Conduct t-tests to remove higher-order components until no components can be removed.

15-34. The Energy Information Administration (EIA), created by Congress in 1977, is a statistical agency of the U.S. Department of Energy. It provides data, forecasts, and analyses to promote sound policymaking and public understanding regarding energy and its interaction with the economy and the environment. One of the most important areas of analysis is petroleum. The file entitled **Crude** contains data for the period 1991–2006 concerning the price, supply, and demand for fuel. One concern has been the increase in imported oil into the United States.
a. Examine the relationship between price of gasoline and the annual amount of imported crude oil. Construct a scatter plot depicting this relationship.
b. Determine the order of the polynomial that would fit the data displayed in part a. Express "Imports" in millions of gallons, i.e., 3,146,454/1,000,000 = 3.146454. Produce an estimate of this polynomial.
c. Is a linear or quadratic model more appropriate for predicting the price of gasoline using the annual quantity of imported oil? Conduct the appropriate hypothesis test to substantiate your answer.

CHAPTER OUTCOME #8

15.4 Stepwise Regression

One option in regression analysis is to bring all possible independent variables into the model in one step. This is what we have done in the previous sections. We use the term *full regression* to describe this approach. Another method for developing a regression model is called *stepwise regression*. Stepwise regression, as the name implies, develops the least squares regression equation in steps, either through *forward selection, backward elimination*, or *standard stepwise* regression.

Forward Selection

The forward selection procedure begins by selecting a single independent variable from all those available. The independent variable selected at Step 1 is the variable that is most highly correlated with the dependent variable. A t-test is used to determine if this variable explains a significant amount of the variation in the dependent variable. At Step 1, if the variable does explain a significant amount of the dependent variable's variation, it is selected to be part of the final model used to predict the dependent variable. If it does not, the process is terminated. If no variables are found to be significant, the researcher will have to search for different independent variables than the ones already tested.

At Step 2, a second independent variable is selected based on its ability to explain the remaining unexplained variation in the dependent variable. The independent variable

Coefficient of Partial Determination
The measure of the marginal contribution of each independent variable, given that other independent variables are in the model.

selected in the second, and each subsequent, step is the variable with the highest **coefficient of partial determination**.

Recall that the coefficient of determination (R^2) measures the proportion of variation explained by all of the independent variables in the model. Thus, after the first variable (say, x_1) is selected, R^2 will indicate the percentage of variation explained by this variable. The forward selection routine will then compute all possible two-variable regression models, with x_1 included, and determine the R^2 for each model. The coefficient of partial determination at Step 2 is the proportion of the yet unexplained variation (after x_1 is in the model) that is explained by the additional variable. The independent variable that adds the most to R^2, given the variable(s) already in the model, is the one selected. Then a t-test is conducted to determine if the newly added variable is significant. This process continues until either all independent variables have been entered or the remaining independent variables do not add appreciably to R^2. For the forward selection procedure, the model begins with no variables. Variables are entered one at a time, and after a variable is entered, it cannot be removed.

Backward Elimination

Backward elimination is just the reverse of the forward selection procedure. In the backward elimination procedure, all variables are forced into the model to begin the process. Variables are removed one at a time until no more insignificant variables are found. Once a variable has been removed from the model, it cannot be reentered.

EXAMPLE 15-4 Applying Forward Selection Stepwise Regression Analysis

TRY PROBLEM 15.35

B. T. Longmont Company The B. T. Longmont Company operates a large retail department store in Macon, Georgia. Like other department stores, Longmont has incurred heavy losses due to shoplifting and employee pilferage. The store's security manager wants to develop a regression model to explain the monthly dollar loss. The following steps can be used when developing a multiple regression model using stepwise regression:

Step 1 Specify the model by determining the dependent variable and potential independent variables.

The dependent variable (y) is the monthly dollar losses due to shoplifting and pilferage. The security manager has identified the following potential independent variables:

x_1 = Average monthly temperature (degrees Fahrenheit)
x_2 = Number of sales transactions
x_3 = Dummy variable for holiday month (1 if holiday during month, 0 if not)
x_4 = Number of persons on the store's monthly payroll

The data are contained in a CD-ROM file called **Longmont**.

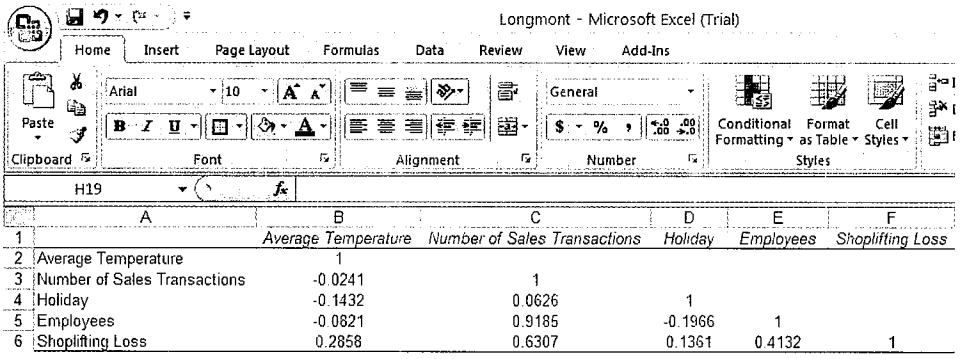
Step 2 Formulate the regression model.

The correlation matrix for the data is presented in Figure 15.25. The forward selection procedure will select the independent variable most highly correlated with the dependent variable. By examining the bottom row in the correlation matrix in Figure 15.25, you can see the variable x_2, number of sales transactions, is most highly correlated ($r = 0.6307$) with dollars lost. Once this variable is entered into the model, the remaining independent variables will be entered based on their ability to explain the remaining variation in the dependent variable.

Figure 15.26a shows the PHStat stepwise regression output, and Figure 15.26b has the Minitab output. At Step 1, variable x_2, number of monthly sales transactions, enters the model.

FIGURE 15.25

Excel 2007 Correlation Matrix Output for Longmont



	A	B	C	D	E	F
1		*Average Temperature*	*Number of Sales Transactions*	*Holiday*	*Employees*	*Shoplifting Loss*
2	Average Temperature	1				
3	Number of Sales Transactions	-0.0241	1			
4	Holiday	-0.1432	0.0626	1		
5	Employees	-0.0821	0.9185	-0.1966	1	
6	Shoplifting Loss	0.2858	0.6307	0.1361	0.4132	1

Excel Instructions:

1. Open file: Longmont.xls.
2. Select **Data tab**.
3. Select **Data Analysis > Correlation**.
4. Specify data range (include labels but do not include the Month column).
5. Click **Labels**.
6. Specify output location.
7. Click **OK**.

Step 3 Perform diagnostic checks on the model.

Although PHStat does not provide R^2 or the estimate of the standard error of the estimate directly, they can be computed from the output in the ANOVA section of the printout. Recall from Chapter 14 that R^2 is computed as

$$R^2 = \frac{SSR}{SST} = \frac{1,270,172.193}{3,192,631.529} = 0.398$$

This one independent variable explains 39.8% ($R^2 = 0.398$) of the variation in the dependent variable. The estimate of the standard error of the estimate is the square root of the mean square residual.

$$s_\varepsilon = \sqrt{MSE} = \sqrt{MS\ \text{Residual}} = \sqrt{128,163.96} = 358$$

Now at Step 1, we test the following:

$$H_0: \beta_2 = 0 \text{ (slope for variable } x_2 = 0)$$
$$H_A: \beta_2 \neq 0$$
$$\alpha = 0.05$$

As shown in Figure 15.26a, the calculated t-value is 3.1481. We compare this to the critical value from the t-distribution for $\alpha/2 = (0.05/2) = 0.025$ and degrees of freedom equal to

$$n - k - 1 = 17 - 1 - 1 = 15$$

This critical value is

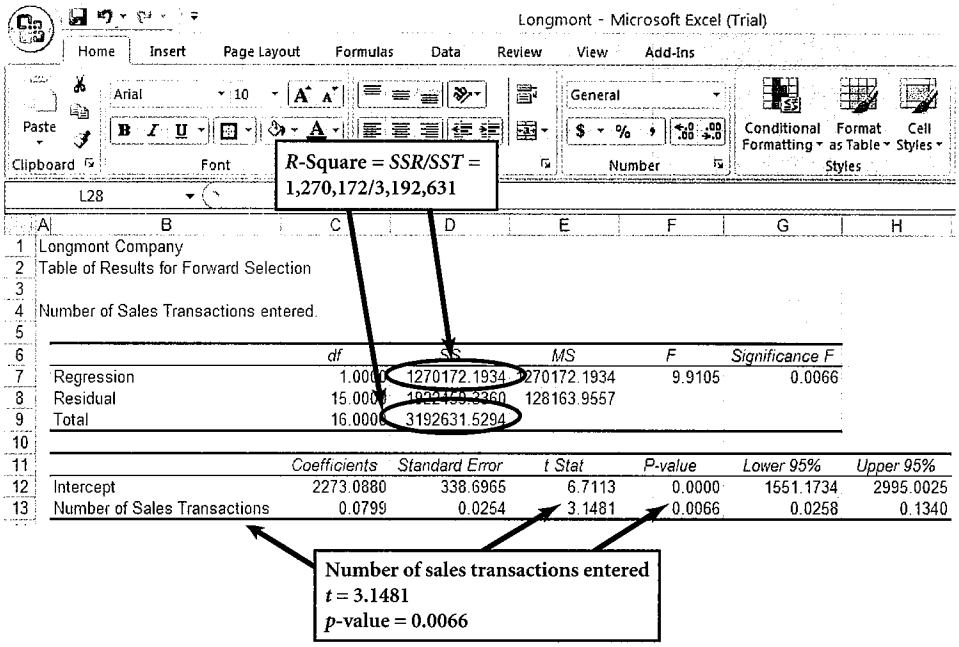
$$t_{0.025} = 2.1315$$

Because

$$t = 3.1481 > 2.1315$$

FIGURE 15.26A

Excel 2007 (PHStat) Forward Selection Results for Longmont Co.—Step 1



Excel 2007 (PHStat) Instructions:

1. Open file: Longmont.xls.
2. Select **Add-Ins**.
3. Select **PHStat**.
4. Select **Regression > Stepwise Regression**.
5. Define data range for y variable and data range for x variables.
6. Check ***p*-value** criteria.
7. Select **Forward Selection**.
8. Click **OK**.

we reject the null hypothesis and conclude that the regression slope coefficient for the variable, number of sales transactions, is not zero. Note also, because the

$$p\text{-value} = 0.0066 < \alpha = 0.05$$

we would reject the null hypothesis.

Step 4 Continue to formulate and diagnose the model by adding other independent variables.

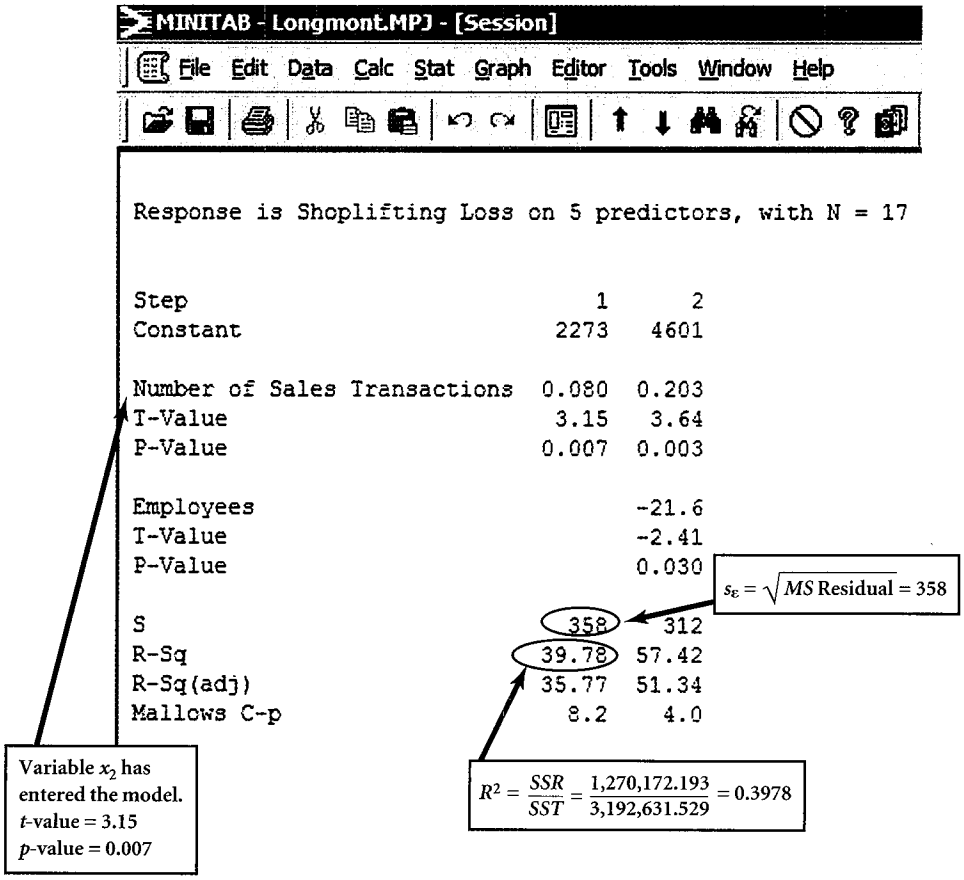
The next variable to be selected will be the one that can do the most to increase R^2. If you were doing this manually, you would try each variable to see which one yields the highest R^2, given that the transactions variable is already in the model. Both the PHStat add-in software and Minitab do this automatically. As shown in Figure 15.26a and Figure 15.27, the variable selected in Step 2 is x_4, number of employees. Using the ANOVA section, we can determine R^2 and s_ε as before.

$$R^2 = \frac{SSR}{SST} = \frac{1,833,270.524}{3,192,631.529} = 0.5742 \quad \text{and}$$

$$s_\varepsilon = \sqrt{MS \text{ Residual}} = \sqrt{97,097.22} = 311.6$$

FIGURE 15.26B

Minitab Forward Selection Results for Longmont Co.—Step 1



Minitab Instructions:

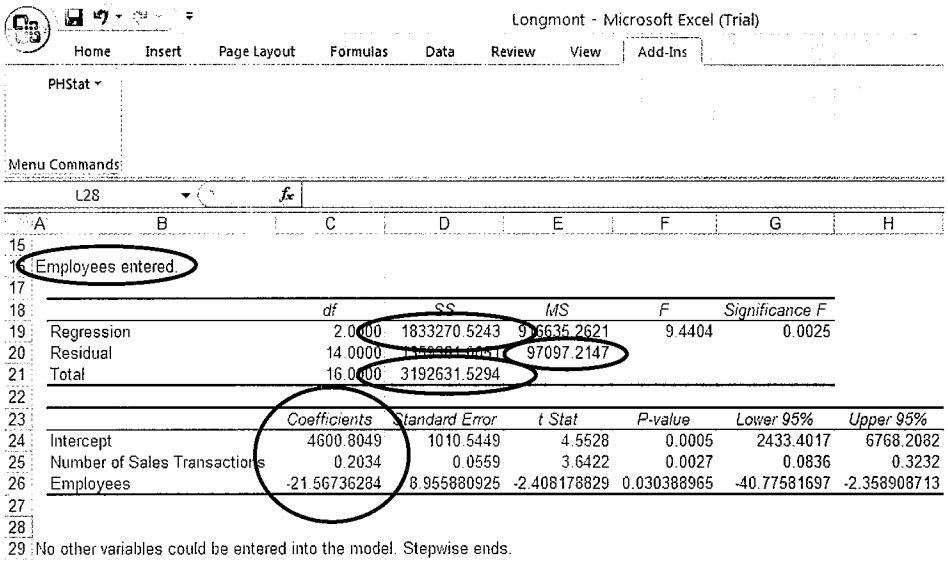
1. Open file: Longmont.MTW.
2. Choose **Stat > Regression > Stepwise**.
3. In **Response**, enter dependent variable (y).
4. In **Predictors**, enter independent variable (x).
5. Select **Methods**.
6. Select **Forward selection**, enter α in **Alpha to enter** and F in **F to enter**.
7. Click **OK**.

The model now explains 57.42% of the variation in the dependent variable. The t-values for both slope coefficients exceed $t = 2.145$ (the critical value from the t-distribution table with a one-tailed area equal to 0.025 and $17 - 2 - 1 = 14$ degrees of freedom), so we conclude that both variables are significant in explaining the variation in the dependent variable, shoplifting loss.

The forward selection routine continues to enter variables as long as each additional variable explains a significant amount of the remaining variation in the dependent variable. Note that PHStat allows you to set the significance level in terms of a p-value or in terms of the t statistic. Then as long as the calculated p-value for an incoming variable is less than your limit, the variable is allowed to enter the model. Likewise, if the calculated statistic exceeds your t limit, the variable is allowed to enter.

FIGURE 15.27

PHStat Forward Selection Results for Longmont Co.—Step 2



	df	SS	MS	F	Significance F
Regression	2.0000	1833270.5243	916635.2621	9.4404	0.0025
Residual	14.0000	[illegible]	97097.2147		
Total	16.0000	3192631.5294			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4600.8049	1010.5449	4.5528	0.0005	2433.4017	6768.2082
Number of Sales Transactions	0.2034	0.0559	3.6422	0.0027	0.0836	0.3232
Employees	-21.56736284	8.955880925	-2.408178829	0.030388965	-40.77581697	-2.358908713

In this example, with the p-value limit set at 0.05, neither of the two remaining independent variables would explain a significant amount of the remaining variation in the dependent variable. The procedure is, therefore, terminated. The resulting regression equation provided by forward selection is

$$\hat{y} = 4600.8 + 0.203x_2 - 21.57x_4$$

Note that the dummy variables for holiday and temperature did not enter the model. This implies that, given the other variables, knowing whether the month in question has a holiday or knowing its average temperature does not add significantly to the model's ability to explain the variation in the dependent variable.

The Longmont Company can now use this regression model to explain variation in shoplifting and pilferage losses based on knowing the number of sales transactions and the number of employees.

Standard Stepwise Regression

The standard stepwise procedure (sometimes referred to as forward stepwise regression—not to be confused with forward selection) combines attributes of both backward elimination and forward selection. The standard stepwise method serves one more important function. If two or more variables overlap, a variable selected in an early step may become insignificant when other variables are added at later steps. The standard stepwise procedure will drop this insignificant variable from the model. Standard stepwise regression also offers a means of observing multicollinearity problems, because we can see how the regression model changes as each new variable is added to it.

The standard stepwise procedure is widely used in decision-making applications and is generally recognized as a useful regression method. However, care should be exercised when using this procedure because it is easy to rely too heavily on the automatic selection process. Remember, the order of variable selection is conditional, based on the variables already in the model. There is no guarantee that stepwise regression will lead you to the best set of independent variables from those available. Decision makers still must use common sense in applying regression analysis to make sure they have usable regression models.

Best Subsets Regression

Another method for developing multiple regression models is called the *best subsets* method. As the name implies, the best subsets method works by trying possible subsets from the list of possible independent variables. The user can then select the "best" model based on such measures as R-squared or the estimate of the standard deviation of the model error. Both Minitab and PHStat contain procedures for performing best subsets regression.

EXAMPLE 15-5 Applying Best Subsets Regression

TRY PROBLEM 15.44

Winston Investment Advisors Charles L. Winston, founder and CEO at Winston Investment Advisors in Burbank, California, is interested in developing a regression model to explain the variation in dividends paid per share by U.S. companies. Such a model would be useful in advising his clients. The following steps can be used to develop such a model using the best subsets regression approach:

Step 1 Specify the model.

Some publicly traded companies pay higher dividends than others. The CEO is interested in developing a multiple regression model to explain the variation in dividends per share paid to shareholders. The dependent variable will be dividends per share. The CEO met with other analysts in his firm to identify potential independent variables for which data would be readily available. The following list of potential independent variables was selected:

x_1 = Return on equity (net income/equity)
x_2 = Earnings per share
x_3 = Current assets in millions of dollars
x_4 = Year-ending stock price
x_5 = Current ratio (current assets/current liabilities)

A random sample of 35 publicly traded U.S. companies was selected. For each company in the sample, data were obtained for the dividends per share paid last year and for year-ending data on the five independent variables. These data are contained in the data file **Company Financials**.

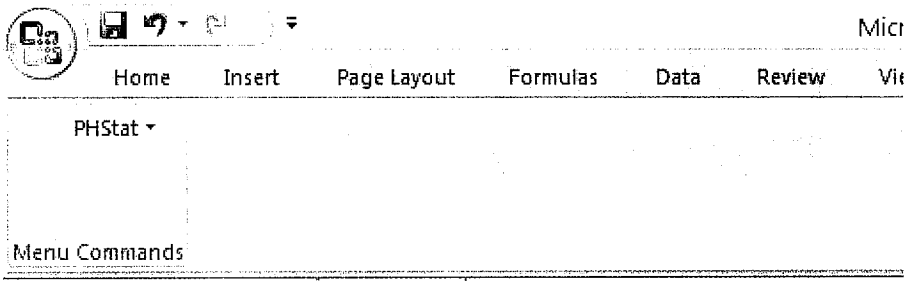
Step 2 Formulate the regression model.

The CEO is interested in developing the "best" regression model for explaining the variation in the dependent variable, dividends per share. The approach taken is to use best subsets, which requires that multiple regression models be developed, each containing a different mix of variables. The models tried will contain from one to five independent variables. The resulting models will be evaluated by comparing values for R-square and the standard error of the estimate. High R-squares and low standard errors are desirable. Another statistic, identified as C_p, is sometimes used to evaluate regression models. This statistic measures the difference between the estimated model and the true population model. Values of C_p close to $p = k + 1$ (k is the number of independent variables in the model) are preferred.

Both the PHStat Excel add-ins and Minitab can be used to perform best subsets regression analysis. Figure 15.28 shows the partial output

FIGURE 15.28

Best Subsets Regression Output for Winston Investment Advisors



Company Financials

	A	B	C	D	E	F
10	Model	Cp	k+1	R Square	Adj. R Square	Std. Error
11	X1	42.306	2	0.0816	0.0331	0.7519
12	X2	20.662	2	0.3387	0.3186	0.6312
13	X3	7.423	2	0.5081	0.4932	0.5444
14	X4	20.684	2	0.3384	0.3183	0.6314
15	X5	46.792	2	0.0042	-0.0260	0.7746
16	X1X2	20.853	3	0.3818	0.3219	0.6297
17	X1X3	7.745	3	0.5296	0.5002	0.5406
18	X1X4	22.657	3	0.3387	0.2974	0.6410
19	X1X5	43.934	3	0.0663	0.0080	0.7616
20	X2X3	1.449	3	0.6102	0.5858	0.4921
21	X2X4	17.751	3	0.4015	0.3641	0.6098
22	X2X5	22.434	3	0.3416	0.3004	0.6396
23	X3X4	3.244	3	0.5872	0.5614	0.5064
24	X3X5	9.419	3	0.5082	0.4774	0.5528
25	X4X5	22.464	3	0.3412	0.3000	0.6398
26	X1X2X3	3.054	4	0.6153	0.5780	0.4967
27	X1X2X4	17.690	4	0.4279	0.3725	0.6057
28	X1X2X5	22.670	4	0.3841	0.3026	0.6386
29	X1X3X4	5.197	4	0.5878	0.5480	0.5141
30	X1X3X5	9.735	4	0.5297	0.4842	0.5492
31	X1X4X5	24.414	4	0.3418	0.2781	0.6497
32	X2X3X4	2.543	4	0.6218	0.5852	0.4925
33	X2X3X5	3.438	4	0.6104	0.5726	0.4999
34	X2X4X5	19.741	4	0.4016	0.3437	0.6185
35	X3X4X5	5.080	4	0.5893	0.5498	0.5132
36	X1X2X3X4	4.028	5	0.6284	0.5789	0.4963
37	X1X2X3X5	5.045	5	0.6154	0.5641	0.5049
38	X1X2X4X5	19.661	5	0.4283	0.3521	0.6156
39	X1X3X4X5	7.052	5	0.5897	0.5350	0.5215
40	X2X3X4X5	4.525	5	0.6220	0.5716	0.5005
41	X1X2X3X4X5	6.000	6	0.6288	0.5648	0.5045

Excel 2007 (PHStat)
Instructions:

1. Open file: Company Financials.xls.
2. Select **Add-Ins**.
3. Select **PHStat**.
4. Select **Regression > Best Subsets Regression**.
5. Define data range for y variable and data range for x variables.
6. Click **OK**.

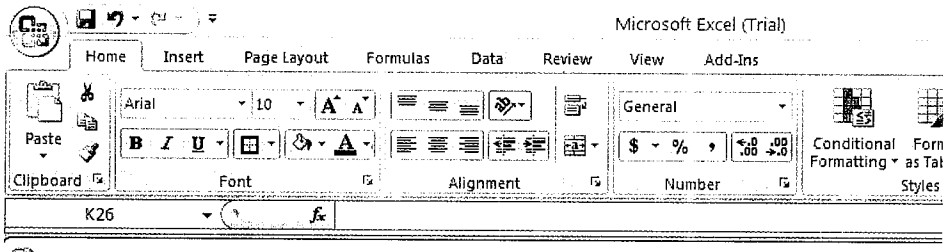
from PHStat. Notice that all possible combinations of models with $k = 1$ to $k = 5$ independent variables are included. Several models appear to be good candidates based on R-square, adjusted R-square, standard error of the estimate, and C_p values. These are:

Model	C_p	$k+1$	R-square	Adj. R-square	Std. Error
X1X2X3X4	**4.0**	**5**	0.628	0.579	0.496
X1X2X3X4X5	**6.0**	**6**	0.629	0.565	0.505
X1X2X3X5	**5.0**	**5**	0.615	0.564	0.505
X2X3	**1.4**	**3**	0.610	0.586	0.492
X2X3X4	**2.5**	**4**	0.622	0.585	0.493
X2X3X4X5	**4.5**	**5**	0.622	0.572	0.500
X2X3X5	**3.4**	**4**	0.610	0.573	0.500

There is little difference in these seven models in terms of the statistics shown. We can examine any of them in more detail by looking at further PHStat output. For instance, the model containing variables x_1, x_2, x_3, and x_5 is shown in Figure 15.29. Note that although this model is among the best with respect to R-square, adjusted R-square, standard error of the estimate, and C_p value, two of the four variables (x_1 = ROE and x_5 = current ratio) have statistically insignificant regression coefficients. Figure 15.30 shows the regression model with the two statistically significant variables remaining. The R-square value is 0.61, the adjusted R^2 has increased, and the overall model is statistically significant.

FIGURE 15.29

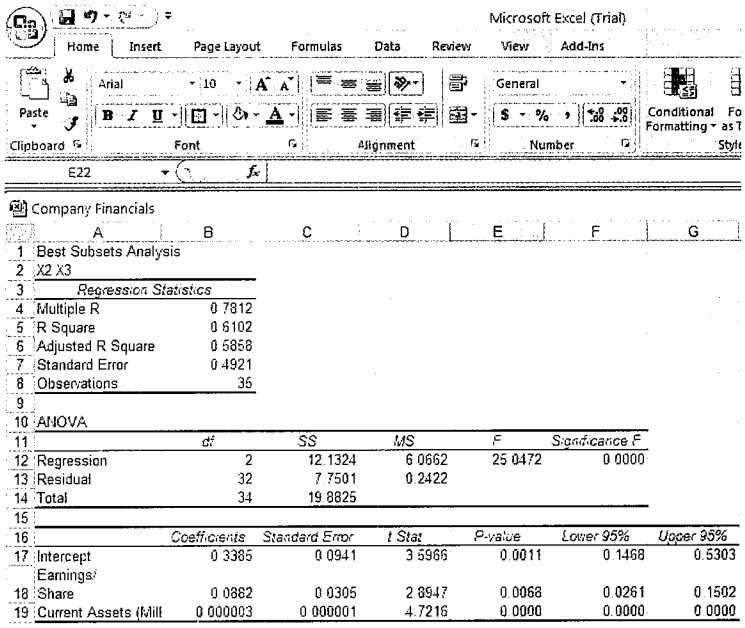
One Potential Model for Winston Investment Advisors



	A	B	C	D	E	F	G
1	Best Subsets Analysis						
2	X1 X2 X3 X5						
3	*Regression Statistics*						
4	Multiple R	0.7845					
5	R Square	0.6154					
6	Adjusted R Square	0.5641					
7	Standard Error	0.5049					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	12.2352	3.0588	11.9996	0.0000	
13	Residual	30	7.6473	0.2549			
14	Total	34	19.8825				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	0.3236	0.1203	2.6906	0.0115	0.0780	0.5691
18	ROE	-0.1189	0.1899	-0.6261	0.5360	-0.5067	0.2689
19	Earnings/ Share	0.1041	0.0403	2.5845	0.0149	0.0218	0.1864
20	Current Assets (Mill	0.0000	0.0000	4.4267	0.0001	0.0000	0.0000
21	Current Ratio	-0.0028	0.0305	-0.0906	0.9284	-0.0651	0.0595

FIGURE 15.30

A Final Model for Winston Financial Advisors



Microsoft Excel (Trial)

Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins

Company Financials

	A	B	C	D	E	F	G
1	Best Subsets Analysis						
2	X2 X3						
3	*Regression Statistics*						
4	Multiple R	0.7812					
5	R Square	0.6102					
6	Adjusted R Square	0.5856					
7	Standard Error	0.4921					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	12.1324	6.0662	25.0472	0.0000	
13	Residual	32	7.7501	0.2422			
14	Total	34	19.8825				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	0.3385	0.0941	3.5966	0.0011	0.1468	0.5303
18	Earnings/ Share	0.0882	0.0305	2.8947	0.0068	0.0261	0.1502
19	Current Assets (Mill	0.000003	0.000001	4.7216	0.0000	0.0000	0.0000

15-4: Exercises

Skill Development

15-35. You are given the following set of data:

y	x_1	x_2	x_3
33	9	192	40
44	11	397	47
34	10	235	37
60	13	345	61
20	11	245	23
30	7	235	35
45	12	296	52
25	9	235	27
53	10	295	57
45	13	335	50
37	11	243	41
44	13	413	51

a. Determine the appropriate correlation matrix and use it to predict which variable will enter in the first step of a stepwise regression model.

b. Use standard stepwise regression to construct a model, entering all significant variables.

c. Construct a full regression model. What are the differences in the model? Which model explains the most variation in the dependent variable?

15-36. You are given the following set of data:

y	x_1	x_2	x_3
45	40	41	39
41	31	41	35
43	45	49	39
38	43	41	41
50	42	42	51
39	48	40	42
50	44	44	41
45	42	39	37
43	37	52	41
34	40	47	40
49	35	44	44
45	39	40	45
40	43	30	42
43	53	34	34

a. Determine the appropriate correlation matrix and use it to predict which variable will enter in the first step of a stepwise regression model.
b. Use standard stepwise regression to construct a model, entering all significant variables.
c. Construct a full regression model. What are the differences in the model? Which model explains the most variation in the dependent variable?

15-37. Suppose you have four potential independent variables, x_1, x_2, x_3 and x_4, from which you want to develop a multiple regression model. Using stepwise regression, x_2 and x_4 entered the model.
a. Why did only two variables enter the model? Discuss.
b. Suppose a full regression with only variables x_2 and x_4 had been run. Would the resulting model be different from the stepwise model that included only these two variables? Discuss.
c. Now suppose a full regression model had been developed, with all four independent variables in the model. Which would have the higher R^2 value, the full regression model or the stepwise model? Discuss.

15-38. The following data represent a dependent variable and four independent variables:

y	x_1	x_2	x_3	x_4
61	37	13	10	21
37	25	7	11	32
22	23	6	7	18
48	12	8	8	30
66	34	15	2	33
69	35	19	9	23
24	23	14	7	31
68	35	17	3	33
65	37	11	17	19
45	30	9	24	31

a. Use the standard stepwise regression to produce an estimate of a multiple regression model to predict y. Use 0.15 as the alpha to enter and to remove.
b. Change the alpha to enter to 0.01. Repeat the standard stepwise procedure.
c. Change the alpha to remove to 0.35, leaving alpha to enter to be 0.15. Repeat the standard stepwise procedure.
d. Change the alpha to remove to 0.15, leaving alpha to enter to be 0.05. Repeat the standard stepwise procedure.
e. Change the alpha to remove and to enter to 0.05. Repeat the standard stepwise procedure.
f. Compare the three estimated regression equations developed in parts a, c, and d.

15-39. Consider the following set of data:

y	x_1	x_2	x_3	x_4
61	37	18	2	13
37	25	5	4	10
22	23	12	7	4
48	12	6	2	15
66	34	14	3	25
69	35	21	2	20
24	23	7	6	9
68	35	15	3	14
65	37	19	2	19
45	30	12	3	12

a. Use standard stepwise regression to produce an estimate of a multiple regression model to predict y.
b. Use forward selection stepwise regression to produce an estimate of a multiple regression model to predict y.
c. Use backwards elimination stepwise regression to produce an estimate of a multiple regression model to predict y.
d. Use best subsets regression to produce an estimate of a multiple regression model to predict y.

Computer Database Exercises

15-40. The Western State Tourist Association gives out pamphlets, maps, and other tourist-related information to people who call a toll-free number and request the information. The association orders the packets of information from a document printing company and likes to have enough available to meet the immediate need without having too many sitting around taking up space. The marketing manager decided to develop a multiple regression model to be used in predicting the number of calls that will be received in the coming week. A random sample of 12 weeks is selected, with the following variables:

y = Number of calls
x_1 = Number of advertisements placed the previous week
x_2 = Number of calls received the previous week
x_3 = Number of airline tour bookings into Western cities for the current week

These data are in the data file called **Western States**.

a. Develop the multiple regression model for predicting the number of calls received, using backward elimination stepwise regression.
b. At the final step of the analysis, how many variables are in the model?
c. Discuss why the variables were removed from the model in the order shown by the stepwise regression.

15-41. Refer to Problem 15-40.
a. Develop the correlation matrix that includes all independent variables and the dependent variable. Predict the order that the variables will be selected into the model if forward selection stepwise regression is used.
b. Use forward selection stepwise regression to develop a model for predicting the number of calls that the company will receive. Write a report that describes what has taken place at each step of the regression process.
c. Compare the results of the forward selection stepwise regression results in part b and the backward elimination results determined in Problem 15-40. Which model would you choose? Explain your answer.

15-42. The U.S. Energy Information Administration publishes summary statistics concerning the energy sector of the U.S. economy. The electric power industry continues to grow. Electricity generation and sales rose to record levels. The file entitled **Energy** presents the revenue from retail sales ($million) and the net generation by energy source for the period 1993–2004.
a. Produce the correlation matrix of all the variables. Predict the variables that will remain in the estimated regression equation if standard stepwise regression is used to predict the revenues from retail sales of energy.
b. Use standard stepwise regression to develop an estimate of a model that is to predict the revenue from retail sales of energy ($million).
c. Compare the results of parts a and b. Explain any difference between the two models.

15-43. An investment analyst collected data of 20 randomly chosen companies. The data consisted of the 52-week high stock prices, PE ratio, and the market value of the company. These data are in the file entitled **Investment**. The analyst wishes to produce a regression equation to predict the market value using 52-week-high stock price and the PE ratio of the company. He creates a complete second-degree polynomial.
a. Produce two scatter plots: (1) market value versus stock price, and (2) market value versus PE ratio. Do the scatter plots support the analyst's decision to produce a second-order polynomial? Support your assertion with statistical reasoning.
b. Use forward selection stepwise regression to eliminate any unneeded components from the analyst's model.
c. Does forward selection stepwise regression support the analyst's decision to produce a second-order polynomial? Support your assertion with statistical reasoning.

15-44. A variety of sources suggest that individuals assess their health, at least in part, by estimating their percentage of body fat. A widely accepted measure of body fat uses an underwater weighing technique. There are, however, more convenient methods using only a scale and a measuring tape. An article in the *Journal of Statistics Education* 4, no. 1 (1996) (Roger W. Johnson, "Fitting Percentage of Body Fat to Simple Body Measurements") explored regression models to predict body fat. The file entitled **Bodyfat** lists a portion of the data presented in the cited article.
a. Use best subsets stepwise regression to establish the relationship between body fat and the variables in the specified file.
b. Predict the body fat of an individual whose age is 21, weight is 170 pounds, height is 70 inches, chest circumference is 100 centimeters, abdomen is 90 centimeters, hip is 105, and whose thigh is 60 centimeters around.

15-45. The consumer price index (CPI) is a measure of the average change in prices over time in a fixed market basket of goods and services typically purchased by consumers. One of the items in this market basket that affects the CPI is the price of oil and its derivatives. The file entitled **Consumer** contains the price of the derivatives of oil and the CPI adjusted to 2005 levels.
a. Use forward selection stepwise regression to determine which combination of the oil derivative prices drive the CPI. If you encounter difficulties in completing this task, explain what caused the difficulties.
b. Eliminate the source of the difficulties in part a by producing a correlation matrix to determine where the difficulty lies.
c. Delete one of the variables indicated in part b and complete the instructions in part a.

15.5 Determining the Aptness of the Model

In Section 15.1 we discussed the basic steps involved in building a multiple regression model. These are

1. specify the model
2. build the model
3. perform diagnostic checks on the model

The final step is the diagnostic step in which you examine the model to determine how well it performs. In Section 15.2, we discussed several statistics that you need to consider when performing the diagnostic step, including analyzing R^2, adjusted R^2, and the standard error of the estimate. In addition, we discussed the concept of multicollinearity and the impacts that can occur when multicollinearity is present. Section 15.3 introduced another diagnostic step that involves looking for potential curvilinear relationships between the independent variables and the dependent variable. We presented some basic data transformation techniques for dealing with curvilinear situations. However, a major part of the diagnostic process involves an analysis of how well the model fits the regression analysis assumptions.

The assumptions of multiple regression include the following:

Assumptions

1. Individual residuals, ε, are statistically independent of one another, and these values represent a random sample from the population of possible residuals at each level of x.
2. For a given value of x, there can exist many values of y, and therefore many possible values for ε. Further, the distribution of possible ε-values for any level of x is normally distributed.
3. The distributions of possible ε-values have equal variances at each level of x.
4. The means of the dependent variable, y, for all specified values of x can be connected with a line called the population regression model.

The degree to which a regression model satisfies these assumptions is called *aptness*.

Analysis of Residuals

Residual
The difference between the actual value of the dependent variable and the value predicted by the regression model.

The **residual** is computed using Equation 15.12.

Residual

$$e_i = y_i - \hat{y}_i \tag{15.12}$$

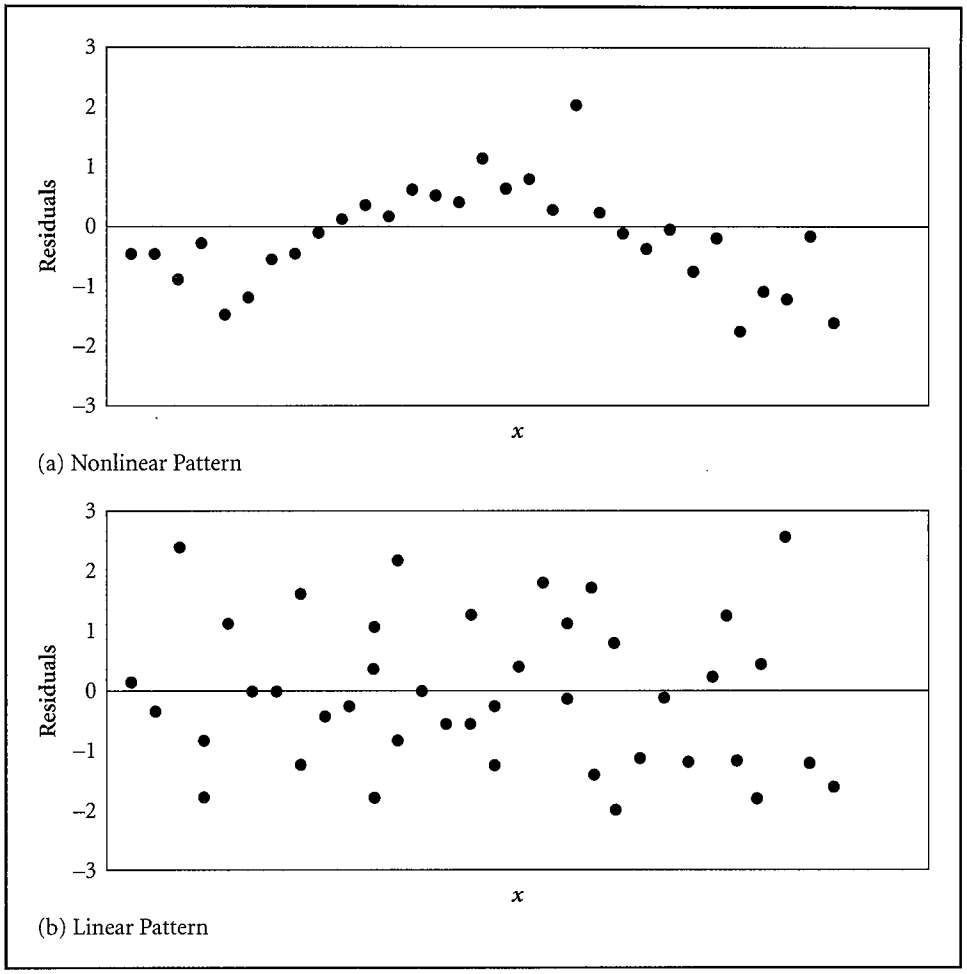
A residual value can be computed for each observation in the data set. A great deal can be learned about the aptness of the regression model by analyzing the residuals. The principal means of residual analysis is a study of residual plots. The following problems can be inferred through graphical analysis of residuals:

1. The regression function is not linear.
2. The residuals do not have a constant variance.
3. The residuals are not independent.
4. The residual terms are not normally distributed.

We will address each of these in order. The regression options in both Minitab and Excel provide extensive residual analysis.

FIGURE 15.31

Residual Plots Showing Linear and Nonlinear Patterns



Checking for Linearity A plot of the residuals (on the vertical axis) against the independent variable (on the horizontal axis) is useful for detecting whether a linear function is the appropriate regression function. Figure 15.31 illustrates two different residual plots. Figure 15.31a shows residuals that systematically depart from 0. When x is small, the residuals are negative. When x is in the midrange, the residuals are positive; and for large x-values, the residuals are negative again. This type of plot suggests that the relationship between y and x is nonlinear. Figure 15.31b shows a plot in which residuals do not show a systematic variation from 0, implying that the relationship between x and y is linear.

If a linear model is appropriate, we expect the residuals to band around 0 with no systematic pattern displayed. If the residual plot shows a systematic pattern, it may be possible to transform the independent variable (refer to Section 15.3) so that the revised model will produce residual plots that will not systematically vary from 0.

Business Application

FIRST CITY REAL ESTATE (*CONTINUED*) We have been using First City Real Estate to introduce multiple regression tools throughout this chapter. Remember, the managers wish to develop a multiple regression model for predicting the sales prices of homes in their market. Suppose that the most current model incorporates a transformation of the lot size variable as log of lot size. The output for this model is shown in Figure 15.32. Notice the model now has a R^2 value of 96.9%.

Excel and Minitab Tutorial

There are currently four independent variables in the model: square feet, bedrooms, garage, and the log of lot size. Both Minitab and Excel provide procedures for automatically producing residual plots. Figure 15.33 shows the plots of the residuals against each of the independent variables. The transformed variable, log lot size, has a residual pattern that shows a systematic pattern. The residuals are positive for small values of log lot size,

FIGURE 15.32

Minitab Output of First City Real Estate Appraisal Model

MINITAB - First City-3.MPJ - [Session]

File Edit Data Calc Stat Graph Editor Tools Window Help

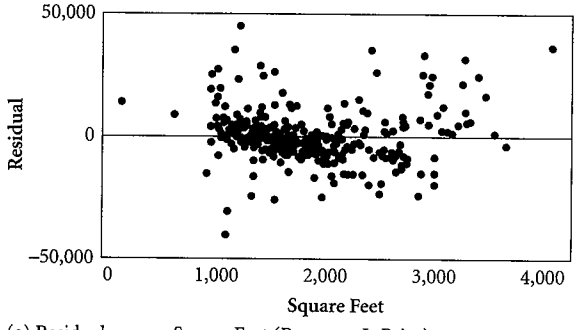
Regression Analysis: Price versus Sq. Feet, Bedrooms, ...

```
The regression equation is
Price = - 521920 + 17.9 Sq. Feet - 2319 Bedrooms + 6013 Garage _
        + 159050 log Lot Size


Predictor        Coef  SE Coef       T      P
Constant       -521920    10798  -48.34  0.000
Sq. Feet         17.858    1.682   10.62  0.000
Bedrooms         -2319     1174   -1.98  0.049
Garage _          6013     1299    4.63  0.000
log Lot Size    159050     3286   48.40  0.000


S = 11249.9  R-Sq = 96.9%  R-Sq(adj) = 96.8%


Analysis of Variance

Source           DF          SS           MS        F      P
Regression        4  1.23329E+12  3.08323E+11  2436.16  0.000
Residual Error  314  39740194153    126561128
Total           318  1.27303E+12
```

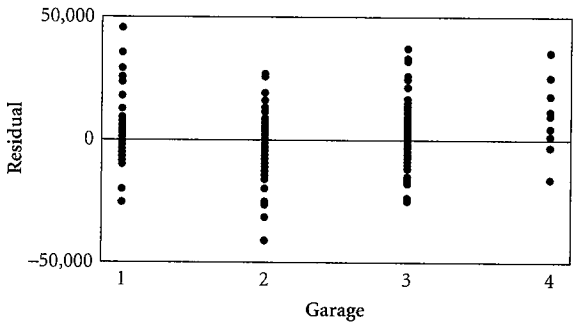
Minitab Instructions:
1. Open file: First City-3 MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Click **OK.**

FIGURE 15.33

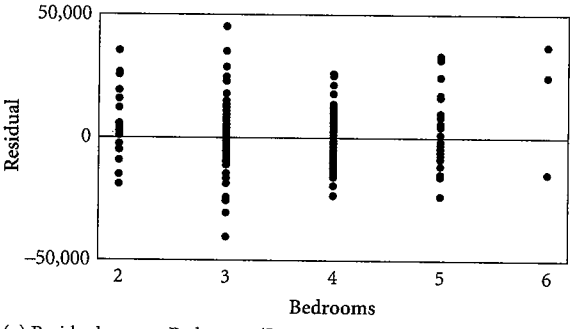
First City Real Estate Residual Plots versus the Independent Variables



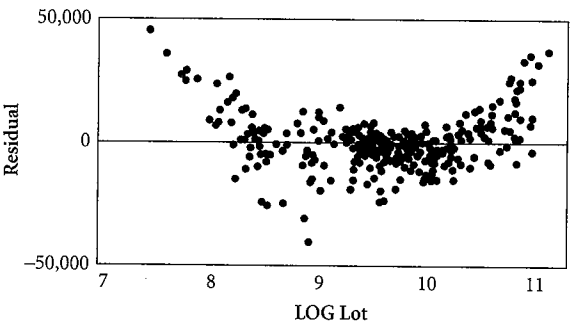
(a) Residuals versus Square Feet (Response Is Price)



(b) Residuals versus Garage (Response Is Price)



(c) Residuals versus Bedrooms (Response Is Price)



(d) Residuals versus LOG Lot (Response Is Price)

negative for intermediate values of log lot size, and positive again for large values of log lot size. This pattern suggests that the curvature of the relationship between sales prices of homes and lot size is even more pronounced than the logarithm implies. Potentially, a second- or third-degree polynomial in the lot size should be pursued.

Do the Residuals Have Equal Variances at all Levels of Each *x* Variable? Residual plots also can be used to determine whether the residuals have a constant variance. Consider Figure 15.34, in which the residuals are plotted against an independent variable. The plot in Figure 15.34a shows an example in which, as x increases, the residuals become less variable. Figure 15.34b shows the opposite situation. When x is small, the residuals are tightly packed around 0, but as x increases, the residuals become more variable. Figure 15.34c shows an example in which the residuals exhibit a constant variance around the zero mean.

When a multiple regression model has been employed, we can analyze the equal variance assumption by plotting the residuals against the fitted ($\hat{y}$) values. When the residual plot is cone-shaped, as in Figure 15.35, it suggests that the assumption of equal variance has been violated.

Figure 15.36 shows the residuals plotted against the $\hat{y}$-values for First City Real Estate's appraisal model. We have drawn a band around the residuals that shows that the variance of the residuals stays quite constant over the range of the fitted values.

Are the Residuals Independent? If the data used to develop the regression model are measured over time, a plot of the residuals against time is used to determine whether the residuals are correlated. Figure 15.37a shows an example in which the residual plot against time suggests independence. The residuals in Figure 15.37a appear to be randomly distributed around the mean of zero over time. However, in Figure 15.37b, the plot suggests that the residuals are not independent, because in the early time periods the residuals are negative and in later time periods the residuals are positive. This, or any other nonrandom pattern in the residuals over time, indicates that the assumption of independent residuals has been violated. Generally, this means some variable associated with the passage of time has been omitted from the model. Often, time is used as a surrogate for other time-related variables in a regression model. Chapter 16 will discuss time-series data analysis and forecasting techniques in more detail and will address the issue of incorporating the time variable into the model. In Chapter 16, we introduce a procedure called the Durbin-Watson test to determine whether residuals are correlated over time.

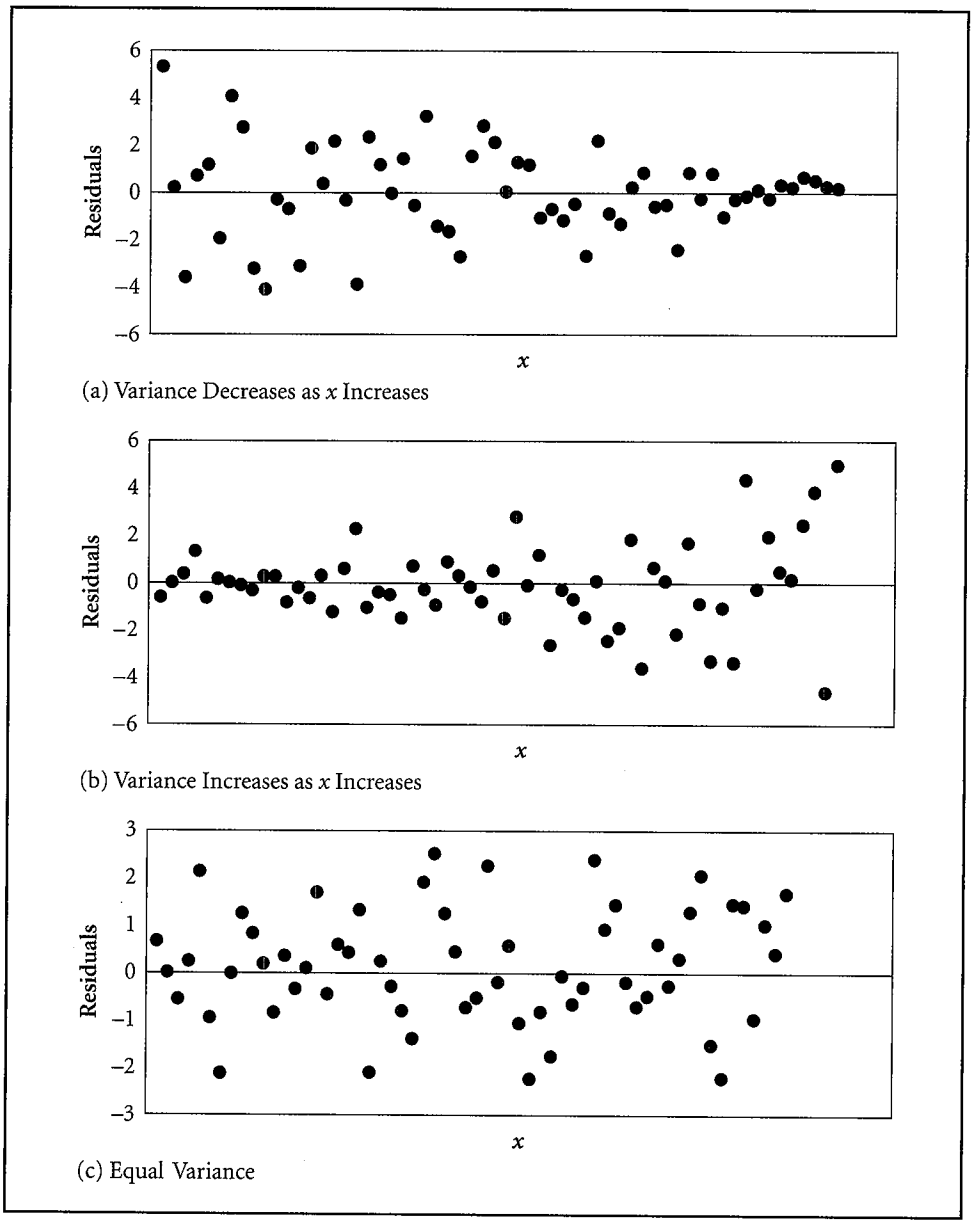
Checking for Normally Distributed Error Terms The need for normally distributed model errors occurs when we want to test a hypothesis about the regression model. Small departures from normality do not cause serious problems. However, if the model errors depart dramatically from a normal distribution, there is cause for concern. Examining the sample residuals will allow us to detect such dramatic departures. One method for graphically analyzing the residuals is to form a frequency histogram of the residuals to determine whether the general shape is normal. The chi-square goodness-of-fit test presented in Chapter 13 can be used to test whether the residuals fit a normal distribution.

Another method for determining normality is to calculate and plot the *standardized residuals*. In Chapter 3 you learned that a random variable is standardized by subtracting its mean and dividing the result by its standard deviation. The mean of the residuals is zero. Therefore, dividing each residual by an estimate of its standard deviation gives the standardized residual.[9] Although the proof is beyond the scope of this text, it can be shown that the standardized residual for any particular observation for a simple linear regression model is found using Equation 15.13.

[9] The standardized residual is also referred to as the studentized residual.

FIGURE 15.34

Residual Plots Showing Constant and Nonconstant Variances



Standardized Residual—Simple Linear Regression

$$s_{e_i} = \frac{e_i}{s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}} \tag{15.13}$$

where:

e_i = ith residual value
s_ε = Standard error of the estimate
x_i = Value of x used to generate the predicted y-value for the ith observation

Computing the standardized residual for an observation in a multiple regression model is too complicated to be done by hand. However, the standardized residuals are generated from most statistical software, including Minitab and Excel. The Excel and Minitab tutorials on

FIGURE 15.35

Residual Plots Against the Fitted ($\hat{y}$) Values

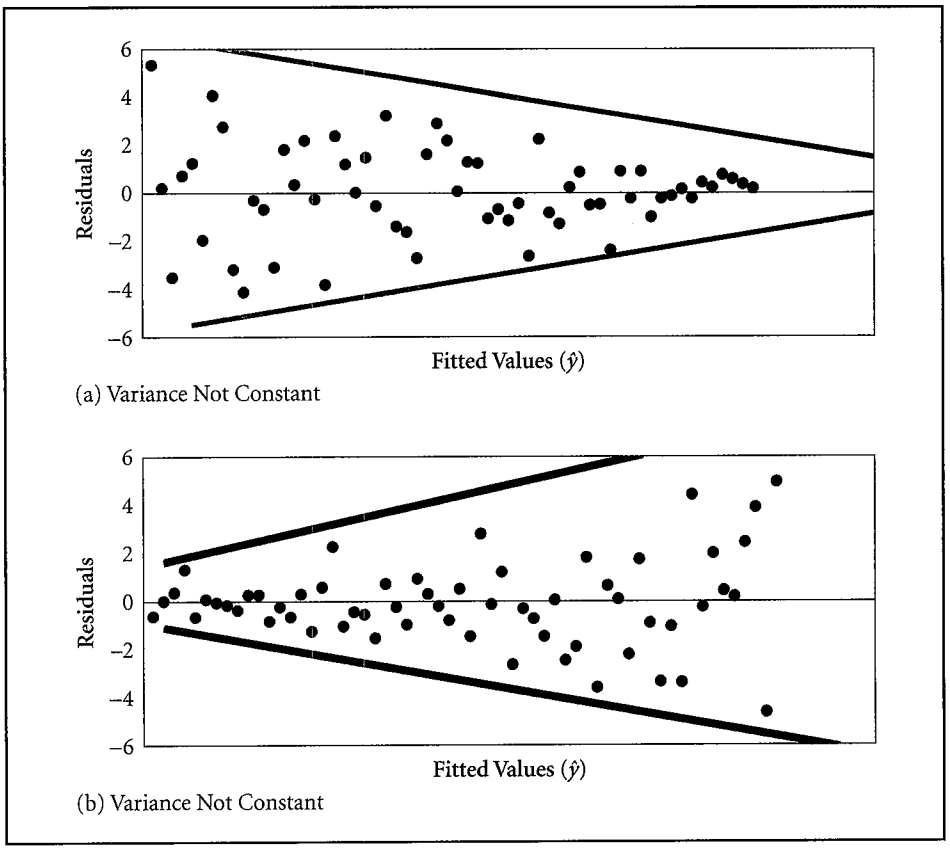
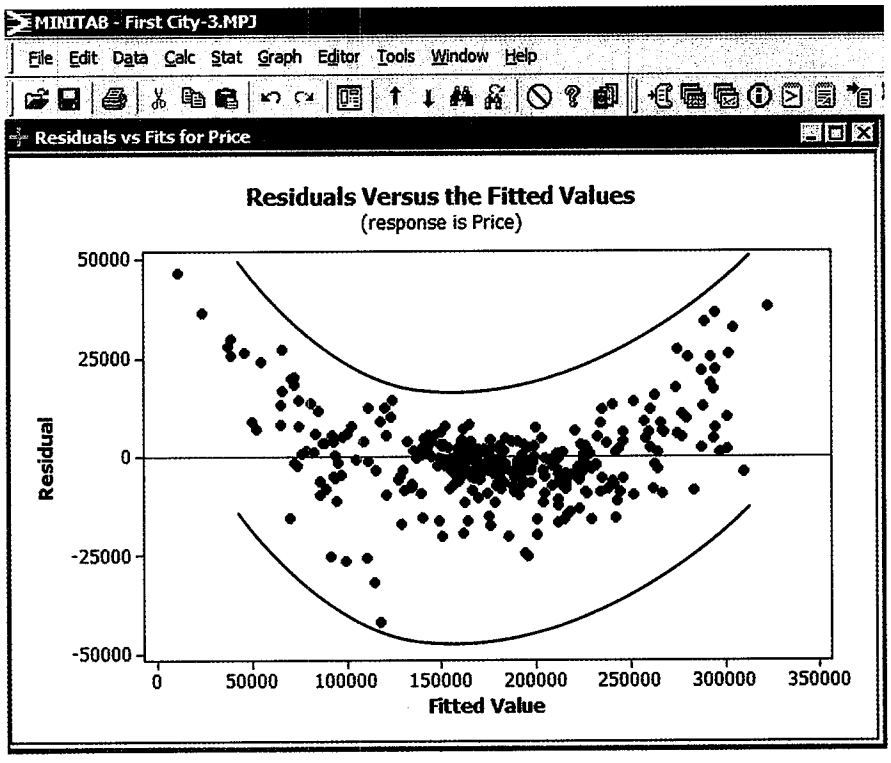


FIGURE 15.36

Minitab Plot of Residuals versus Fitted Values for First City Real Estate



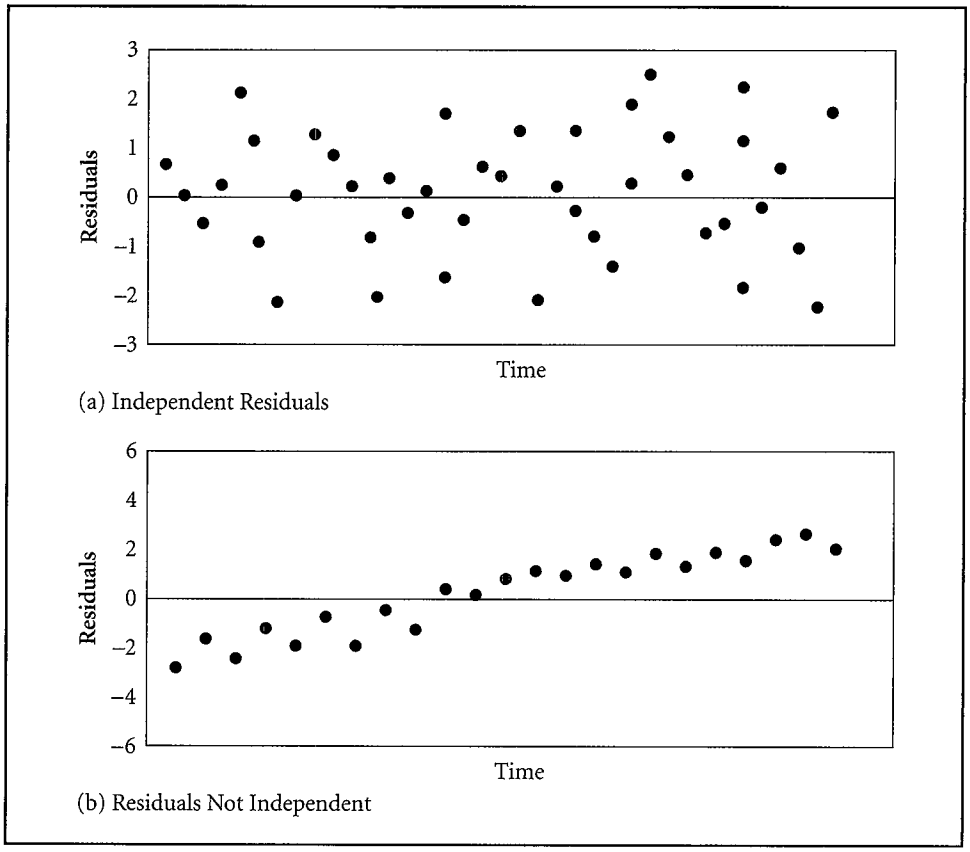
Minitab Instructions:

1. Open file: First City-3. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Choose **Graphs.**
6. Under **Residual Plots**, select **Residuals versus fits.**
7. Click **OK. OK.**

your CD-ROM illustrate the methods required to generate the standardized residuals and residual plots. Because other problems such as nonconstant variance and nonindependent residuals can result in residuals that seem to be abnormal, you should check these other factors before addressing the normality assumption.

FIGURE 15.37

Plot of Residuals Against Time



Recall that for a normal distribution, approximately 68% of the values will fall within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and virtually all values will fall within 3 standard deviations of the mean.

Figure 15.38 illustrates the histogram of the residuals for the First City Real Estate example. The distribution of residuals looks to be close to a normal distribution. Figure 15.39 shows the histogram for the standardized residuals, which will have the same basic shape as the residual distribution in Figure 15.38.

Another approach for checking for normality of the residuals is to form a *probability plot*. We start by arranging the residuals in numerical order from smallest to largest. The standardized residuals are plotted on the horizontal axis, and the corresponding expected value for the standardized residual is plotted on the vertical axis. Although we won't delve into how the expected value is computed, you can examine the normal probability plot to see whether the plot forms a straight line. The closer the line is to linear, the closer the residuals are to being normally distributed. Figure 15.40 shows the normal probability plot for the First City Real Estate Company example.

You should be aware that Minitab and Excel format their residual plots slightly differently. However, the same general information is conveyed, and you can look for the same signs of problems with the regression model.

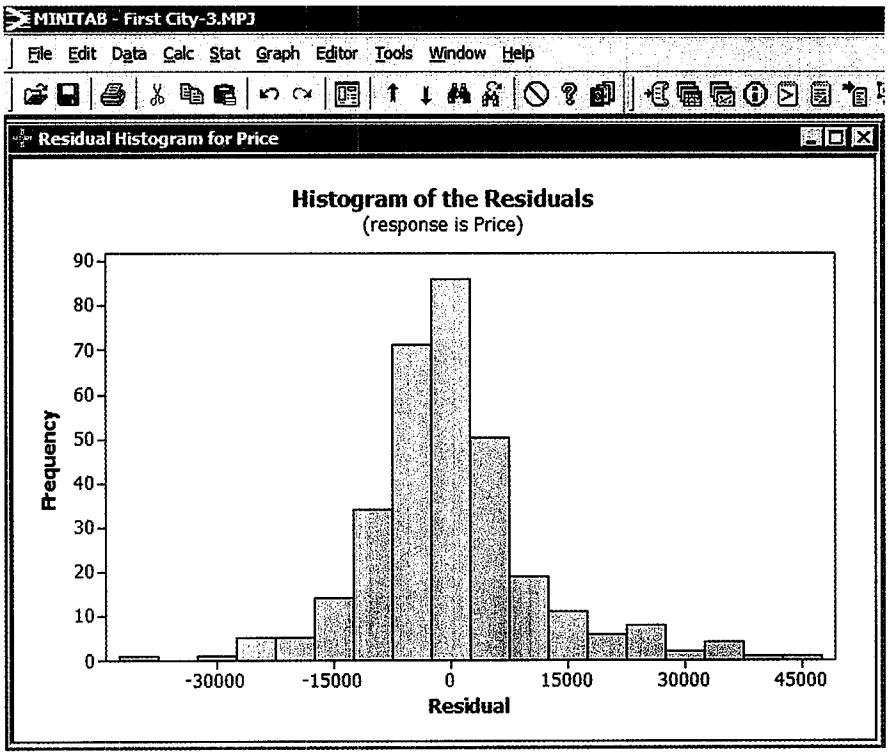
Corrective Actions

If, based on analyzing the residuals, you decide the model constructed is not appropriate, but you still want a regression-based model, some corrective action may be warranted. There are three approaches that may work: transform some of the existing independent variables; remove some variables from the model; or start over in the development of the regression model.

Earlier in this chapter, we discussed a basic approach involved in variable transformation. In general, the transformations of the independent variables (such as raising x to a power, taking the square root of x, or taking the log of x) are used to make the data better

FIGURE 15.38

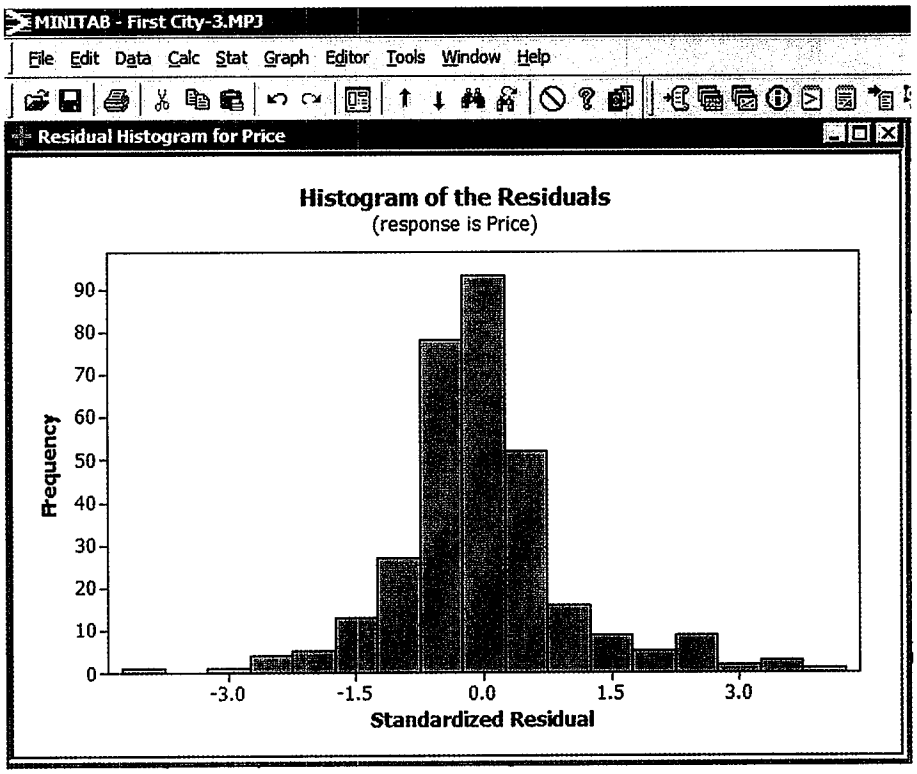
Minitab Histogram of Residuals for First City Real Estate



Minitab Instructions:
1. Open file: First City-3. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Choose **Graphs.**
6. Under **Residual Plots**, select **Histogram of residuals.**
7. Click **OK. OK.**

FIGURE 15.39

Minitab Histogram of Standardized Residuals for First City Real Estate

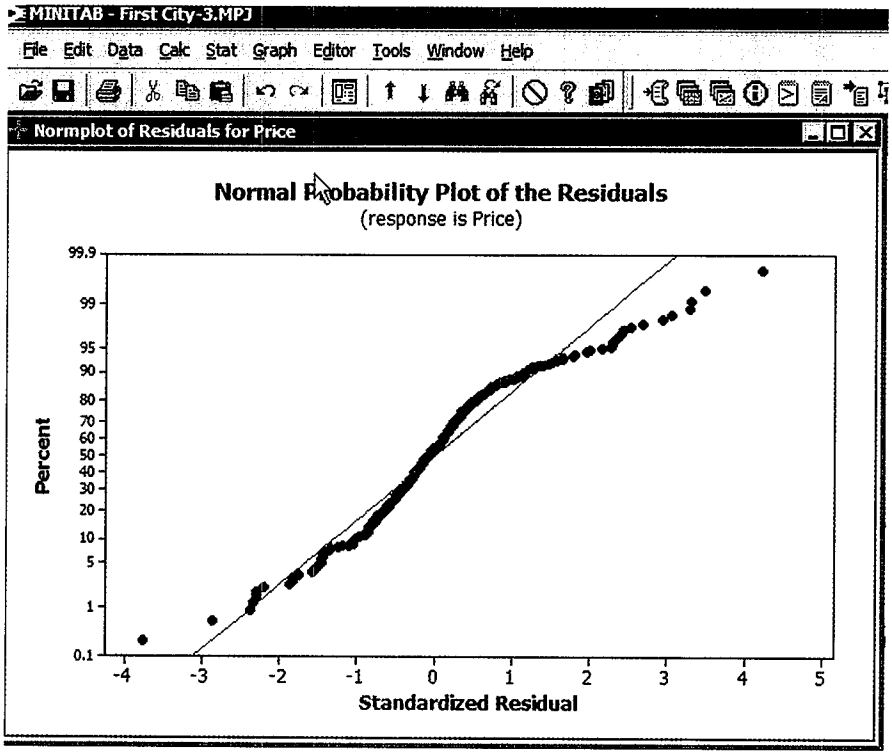


Minitab Instructions:
1. Open file: First City-3. MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Choose **Graphs.**
6. Under **Residual for Plots**, select **Standadized.**
7. Under **Residual Plots**, select **Histogram of residuals.**
8. Click **OK. OK.**

conform to a linear relationship. If the model suffers from nonlinearity and if the residuals have a nonconstant variance, you may want to transform both the independent and dependent variables. In cases in which the normality assumption is not satisfied, transforming the dependent variable is often useful. In many instances, a log transformation works. In some instances, a transformation involving the product of two independent variables will help.

FIGURE 15.40

Minitab Normal Probability Plot of Residuals for First City Real Estate



Minitab Instructions:

1. Open file: First City-3.MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter dependent (y) variable.
4. In **Predictors**, enter independent (x) variables.
5. Choose **Graphs.**
6. Under **Residual** Plots, select **Normal plot of residuals.**
7. Click **OK. OK.**

A more detailed discussion is beyond the scope of this text. However, you can read more about this subject in the Kutner et al. reference listed at the end of the chapter.

The alternative of using a different regression model means that we respecify the model to include new independent variables or remove existing variables from the model. In most modeling situations, we are in a continual state of model respecification. We are always seeking to improve the regression model by finding new independent variables.

15-5: Exercises

Skill Development

15-46. Consider the following values for an independent and dependent variable:

x	y
6	5
9	20
14	28
18	30
22	33
27	35
33	45

a. Determine the estimated linear regression equation relating the dependent and independent variables.

b. Is the regression equation you found significant? Test at the $\alpha = 0.05$ level.

c. Determine both the residuals and standardized residuals. Is there anything about the residuals that would lead you to question whether the assumptions necessary to use regression analysis are satisfied? Discuss.

15-47. Consider the following values for an independent and dependent variable:

x	y
6	5
9	20
14	28
18	15
22	27
27	31
33	32
50	60
61	132
75	160

a. Determine the estimated linear regression equation relating the dependent and independent variables.
b. Is the regression equation you found significant? Test at the $\alpha = 0.05$ level.
c. Determine both the residuals and standardized residuals. Is there anything about the residuals that would lead you to question whether the assumptions necessary to use regression analysis are satisfied?

15-48. Examine the following data set

y	x
25	10
35	10
14	10
45	20
52	20
41	20
65	30
63	30
68	30

a. Determine the estimated regression equation for this data set.
b. Calculate the residuals for this regression equation.
c. Produce the appropriate residual plot to determine if the linear function is the appropriate regression function for this data set.
d. Use a residual plot to determine if the residuals have a constant variance.
e. Produce a residual plot to determine if the residuals are independent. Assume the order of appearance is the time order of the data.
f. Use a probability plot to determine if the error terms are normally distributed.

15-49. Examine the following data set:

y	x_1	x_2
25	5	25
35	5	5
14	5	5
45	25	40
52	25	5
41	25	25
65	30	30
63	30	30
68	30	25
75	40	30

a. Determine the estimated regression equation for this data set.
b. Calculate the residuals and the standardized residuals for this regression equation.
c. Produce the appropriate residual plot to determine if the linear function is the appropriate regression function for this data set.
d. Use a residual plot to determine if the residuals have a constant variance.
e. Produce the appropriate residual plot to determine if the residuals are independent.
f. Construct a probability plot to determine if the error terms are normally distributed.

Computer Database Exercises

15-50. The Western State Tourist Association gives out pamphlets, maps, and other tourist-related information to people who call a toll-free number and request the information. The association orders the packets of information from a document printing company and likes to have enough available to meet the immediate need without having too many sitting around taking up space. The marketing manager decided to develop a multiple regression model to be used in predicting the number of calls that will be received in the coming week. A random sample of 12 weeks is selected, with the following variables:

y = Number of calls
x_1 = Number of advertisements placed the previous week
x_2 = Number of calls received the previous week
x_3 = Number of airline tour bookings into Western cities for the current week

The data are in the file called **Western States**.

a. Construct a multiple regression model using all three independent variables. Write a short report discussing the model.
b. Based on the appropriate residual plots, what can you conclude about the constant variance assumption? Discuss.
c. Based on the appropriate residual analysis, does it appear that the residuals are independent? Discuss.
d. Use an appropriate analysis of the residuals to determine whether the regression model meets the assumption of normally distributed error terms. Discuss.

15-51. The athletic director of State University is interested in developing a multiple regression model that might be used to explain the variation in attendance at football games at his school. A sample of 16 games was selected from home games played during the past 10 seasons. Data for the following factors were determined:

y = Game attendance
x_1 = Team win/loss percentage to date

x_2 = Opponent win/loss percentage to date
x_3 = Games played this season
x_4 = Temperature at game time

The sample data are in the file called **Football**.

a. Build a multiple regression model using all four independent variables. Write a short report that outlines the characteristics of this model.
b. Develop a table of residuals for this model. What is the average residual value? Why do you suppose it came out to this value? Discuss.
c. Based on the appropriate residual plot, what can you conclude about the constant variance assumption? Discuss.
d. Based on the appropriate residual analysis, does it appear that the model errors are independent? Discuss.
e. Can you conclude, based on the appropriate method of analysis, that the model error terms are approximately normally distributed?

15-52. Refer to Exercise 15-7 which referenced an article in *BusinessWeek* ("Hot Growth Companies," June 5, 2006) that presented a list of the 100 companies perceived as having "hot growth" characteristics. The file entitled **Logrowth** contains sales ($million), sales increase (%), return on capital, market value ($million), and recent stock price of the companies ranked from 81 to 100. In Exercise 15-8, a regression equation was constructed in which the sales of the companies was predicted using their market value.

a. Determine the estimated regression equation for this data set.
b. Calculate the residuals and the standardized residuals for this regression equation.
c. Produce the appropriate residual plot to determine if the linear function is the appropriate regression function for this data set.
d. Use a residual plot to determine if the residuals have a constant variance.
e. Produce the appropriate residual plot to determine if the residuals are independent. Assume the data were extracted in the order listed.
f. Construct a probability plot to determine if the error terms are normally distributed.

15-53. The consumer price index (CPI) is a measure of the average change in prices over time in a fixed market basket of goods and services typically purchased by consumers. One of the items in this market basket that affects the CPI is the price of oil and its derivatives. The file entitled **Consumer** contains the price of the derivatives of oil and the CPI adjusted to 2005 levels. In Exercise 15-45, backward elimination stepwise regression was used to determine the relationship between CPI and two independent variables: the price of heating oil and of diesel fuel.

a. Construct an estimate of the regression equation using the same variables.
b. Produce the appropriate residual plots to determine if the linear function is the appropriate regression function for this data set.
c. Use a residual plot to determine if the residuals have a constant variance.
d. Produce the appropriate residual plot to determine if the residuals are independent. Assume the data were extracted in the order listed.
e. Construct a probability plot to determine if the error terms are normally distributed.

15-54. A variety of sources suggest that individuals assess their health, at least in part, by estimating their percentage of body fat. A widely accepted measure of body fat uses an underwater weighing technique. There are, however, more convenient methods using only a scale and a measuring tape. An article in the *Journal of Statistics Education* 4, no. 1 (1996), (Roger W. Johnson), "Fitting Percentage of Body Fat to Simple Body Measurements") explored regression models to predict body fat. The file entitled **Bodyfat** lists a portion of the data presented in the cited article. Exercise 15-43 utilized best subsets stepwise regression to establish the relationship between body fat and the independent variables weight, abdomen circumference, and thigh circumference.

a. Construct an estimate of the regression equation using the same variables.
b. Produce the appropriate residual plots to determine if the linear function is the appropriate regression function for this data set.
c. Use a residual plot to determine if the residuals have a constant variance.
d. Produce the appropriate residual plot to determine if the residuals are independent. Assume the data were extracted in the order listed.
e. Construct a probability plot to determine if the error terms are normally distributed.

15-55. The National Association of Theatre Owners is the largest exhibition trade organization in the world, representing more than 26,000 movie screens in all 50 states and in more than 20 countries worldwide. Its membership includes the largest cinema chains and hundreds of independent theatre owners. It publishes statistics concerning the movie sector of the economy. The file entitled **Flicks** contains data on total U.S. box office grosses ($billion), total number of admissions (billion), average U.S. ticket price ($), and number of movie screens.

a. Construct a regression equation in which total U.S. box office grosses are predicted using the other variables.

b. Produce the appropriate residual plots to determine if the linear function is the appropriate regression function for this data set.
c. Square each of the independent variables and add them to the model upon which the regression equation in part a was built. Produce the new regression equation.
d. Use a residual plot to determine if the quadratic model in part c alleviates the problem identified in part b.
e. Construct a probability plot to determine if the error terms are normally distributed for the updated model.

Summary and Conclusions

Multiple regression is an extension of simple regression analysis. In multiple regression, two or more independent variables are used to explain the variation in the dependent variable. Just as a manager searches for the best combination of employees to perform a job, the decision maker using multiple regression analysis searches for the best combination of independent variables to explain variation in the dependent variable.

The presentation of multiple regression analysis has largely been an analysis of computer printouts. As a decision maker, you will almost assuredly not be required to manually develop the regression model, but you will have to judge its applicability based on a computer printout. The Excel and Minitab software we have used in Chapters 14 and 15 are representative of the many software packages that are available. You no doubt will encounter printouts that look somewhat different from those shown in this text and some of the terms used may differ slightly. However, the basic information will be the same, as will be the inferences you can make from the model.

This chapter has discussed the difference between R^2 and adjusted R^2, as well as the difference between statistical significance and practical significance. As a decision maker, you must recognize that a regression model can be statistically significant yet have no practical use because the standard error of the estimate is too large or multicollinearity impacts too heavily.

As you continue your study of business, you will find that multiple regression is one of the most widely used statistical tools. You will find it applied particularly to the areas of production, finance, accounting, and economics.

Equations

Population Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad \textbf{(15.1)}$$

Estimated Multiple Regression Model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \quad \textbf{(15.2)}$$

Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}}$$

$$\text{or} \quad r = \frac{\Sigma(x_i-\bar{x}_i)(x_j-\bar{x}_j)}{\sqrt{\Sigma(x_i-\bar{x}_i)^2\,\Sigma(x_j-\bar{x}_j)^2}} \quad \textbf{(15.3)}$$

Multiple Coefficient of Determination (R^2)

$$R^2 = \frac{\text{Sum of squares regression}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad \textbf{(15.4)}$$

F-Test Statistic

$$F = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}} \quad \textbf{(15.5)}$$

Adjusted R-Squared

$$R\text{-sq(adj)} = R_A^2 = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right) \quad \textbf{(15.6)}$$

t-Test for Significance of Each Regression Coefficient

$$t = \frac{b_j - 0}{s_{b_j}} \qquad df = n-k-1 \quad \textbf{(15.7)}$$

Standard Error of the Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE} \quad \textbf{(15.8)}$$

Variance Inflation Factor

$$VIF = \frac{1}{(1-R_j^2)} \quad \textbf{(15.9)}$$

Confidence Interval Estimate for the Regression Slope

$$b_j \pm s_{b_j} \quad \textbf{(15.10)}$$

Polynomial Population Regression Model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon \quad \textbf{(15.11)}$$

Residual

$$e_i = y_i - \hat{y}_i \quad \textbf{(15.12)}$$

Standardized Residual

$$s_{e_i} = \frac{e_i}{s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}} \quad \textbf{(15.13)}$$

Key Terms

Aptness 736
Adjusted R-squared (R_A^2) 692
Coefficient of partial determination 726
Composite model 719
Correlation coefficient 684
Correlation matrix 685
Dummy variables 703
Interaction 719
Model 682
Multicollinearity 695
Multiple coefficient of determination (R^2) 689
Multiple regression model for the population 680
Polynomial model 712
Regression hyperplane 681
Residual 737
Second-order regression model 712
Standard error of the estimate 693
Standardized residual 740
Variance inflation factor (*VIF*) 696

Chapter Exercises

Conceptual Questions

15-56. Discuss in your own terms the similarities and differences between simple linear regression analysis and multiple regression analysis.

15-57. Discuss what is meant by the least squares criterion as it pertains to multiple regression analysis. Is the least squares criterion any different for simple regression analysis? Discuss.

15-58. List the basic assumptions of regression analysis and discuss in your own terms what each means.

15-59. What does it mean if we have developed a multiple regression model and have concluded that the model is apt?

15-60. Go to the library, or use the Internet, to locate three articles using a regression model with more than one independent variable. For each article write a short summary covering the following points:
purpose for using the model
how the variables in the model were selected
how the data in the model were selected
any possible violations of the needed assumptions
the conclusions drawn from using the model

15-61. Consider the following model:

$$\hat{y} = 5 + 3x_1 + 5x_2$$

a. Provide an interpretation of the coefficient of x_1.
b. Is the interpretation provided in part a true regardless of the value of x_2? Explain.
c. Now consider the model $\hat{y} = 5 + 3x_1 + 5x_2 + 4x_1x_2$. Let $x_2 = 1$. Give an interpretation of the coefficient of x_1 when $x_2 = 1$.
d. Repeat part c when $x_2 = 2$. Is the interpretation provided in part a true regardless of the value of x_2? Explain.
e. Considering your answers to parts b and c, what type of regression components has conditional interpretations?

Computer Database Exercises

The following information applies to Exercises 15-62, 15-63, and 15-64.

A publishing company in New York is attempting to develop a model that it can use to help predict textbook sales for books it is considering for future publication. The marketing department has collected data on several variables from a random sample of 15 books. These data are given in the file **Textbooks**.

15-62. Develop the correlation matrix showing the correlation between all possible pairs of variables. Test statistically to determine which independent

variables are significantly correlated with the dependent variable, book sales. Use a significance level of 0.05.

15-63. Develop a multiple regression model containing all four independent variables. Show clearly the regression coefficients. Write a short report discussing the model. In your report make sure you cover the following issues:

a. How much of the total variation in book sales can be explained by these four independent variables? Would you conclude that the model is significant at the 0.05 level?
b. Develop a 95% confidence interval for each regression coefficient and interpret these confidence intervals.
c. Which of the independent variables can you conclude to be significant in explaining the variation in book sales? Test using $\alpha = 0.05$.
d. How much of the variation in the dependent variable is explained by the dependent variable? Is the model statistically significant at the $\alpha = 0.01$ level? Discuss.
e. How much, if at all, does adding one more page to the book impact the sales volume of the book? Develop and interpret a 95% confidence interval estimate to answer this question.
f. Perform the appropriate analysis to determine the aptness of this regression model. Discuss your results and conclusions.

15-64. The publishing company recently came up with some additional data for the 15 books in the original sample. Two new variables, production expenditures (x_5) and number of prepublication reviewers (x_6), have been added. These additional data are as follows:

Book	x_5	x_6	Book	x_5	x_6
1	$38,000	5	9	$51,000	4
2	86,000	8	10	34,000	6
3	59,000	3	11	20,000	2
4	80,000	9	12	80,000	5
5	29,500	3	13	60,000	5
6	31,000	3	14	87,000	8
7	40,000	5	15	29,000	3
8	69,000	4			

Incorporating this additional data, calculate the correlation between each of these additional variables and the dependent variable, book sales.

a. Test the significance of the correlation coefficients, using $\alpha = 0.05$. Comment on your results.
b. Develop a multiple regression model that includes all six independent variables. Which, if any, variables would you recommend be retained if this model is going to be used to predict book sales for the publishing company? For any statistical tests you might perform, use a significance level of 0.05. Discuss your results.
c. Use the F-test approach to test the null hypothesis that all slope coefficients are 0. Test with a significance level of 0.05. What do these results mean? Discuss.
d. Do multicollinearity problems appear to be present in the model? Discuss the potential consequences of multicollinearity with respect to the regression model.
e. Discuss whether the standard error of the estimate is small enough to make this model useful for predicting the sales of textbooks.
f. Plot the residuals against the predicted value of y and comment on what this plot means relative to the aptness of the model.
g. Compute the standardized residuals and form these into a frequency histogram. What does this indicate about the normality assumption?
h. Comment on the overall aptness of this model and indicate what might be done to improve the model.

The following information applies to Exercises 15-65 through 15-74.

The J. J. McCracken Company has authorized its marketing research department to make a study of customers who have been issued a McCracken charge card. The marketing research department hopes to be able to identify the significant variables that explain the variation in purchases. Once these variables are determined, the department intends to try to attract new customers who would be predicted to make a high volume of purchases.

Twenty-five customers were selected at random and values for the following variables were recorded in the file called **McCracken**:

y = Average monthly purchases (in dollars) at McCracken
x_1 = Customer age
x_2 = Customer family income
x_3 = Family size

15-65. A first step in regression analysis often involves developing a scatter plot of the data. Develop the scatter plots of all the possible pairs of variables, and with a brief statement indicate what each plot says about the relationship between the two variables.

15-66. Compute the correlation matrix for these data. Develop the decision rule for testing the significance of each coefficient. Which, if any, correlations are not significant? Use $\alpha = 0.05$.

15-67. Use forward selection stepwise regression to develop the multiple regression model. The variable x_2, family income, was brought into the model. Discuss why this happened.

15-68. Test the significance of the regression model at Step 1 of the computer printout. Justify the significance level you have selected.

15-69. Develop a 95% confidence level for the slope coefficient for the family income variable at Step 1 of the model. Be sure to interpret this confidence interval.

15-70. Describe the regression model at Step 2 of the analysis. In your discussion, be sure to discuss the effect of adding a new variable on the standard error of the estimate and on R^2.

15-71. Referring to Problem 15-70, suppose the manager of McCracken's marketing department questions the appropriateness of adding a second variable. How would you respond to her question?

15-72. Looking carefully at the stepwise regression model, you can see that the value of the slope coefficient for variable x_2, family income, changes each time a new variable is added to the regression model. Discuss why this change takes place.

15-73. Analyze the stepwise regression model at Step 3 and the intermediate results at Steps 1 and 2. Write a report to the marketing manager pointing out the strengths and weaknesses of the model. Be sure to comment on the department's goal of being able to use the model to predict which customers will purchase high volumes from McCracken.

15-74. Plot the residuals against the predicted value of y and comment on what this plot means relative to the aptness of the model.

a. Compute the standardized residuals and form these in a frequency histogram. What does this indicate about the normality assumption?

b. Comment on the overall aptness of this model and indicate what might be done to improve the model.

15-75. Amazon.com has become one of the most successful online merchants. Two measures of its success are sales and net income/loss figures. In an article entitled "Amazon CEO takes long view," *USA Today* (Byron Acohido, July 6, 2005) presented these figures (in $million) for the period 1995 to 2004. They are given here.

Net income/ loss	−0.3	−5.7	−27.5	−124.5	−719.9
Sales	0.5	15.7	147.7	609.8	1,639.8
Net income/ loss	−1,411.2	−567.3	−149.1	35.3	588.5
Sales	2,761.9	3,122.9	3,932.9	5,263.7	6,921.1

a. Produce a scatter plot for Amazon's net income/loss and sales figures for the period 1995 to 2004. Determine the order (or degree) of the polynomial that could be used to predict Amazon's net income/loss using sales figures for the period 1995 to 2004.

b. Produce the polynomial indicated in part a.

c. Test to determine whether the overall model is statistically significant. Use a significance level of 0.10 and the test statistic approach.

d. Conduct a hypothesis test to determine if curvature exists in the model that predicts Amazon's net income/loss using sales figures. Use a significance level of 0.02 and the test statistic approach.

15-76. The National Association of Realtors Existing-Home Sales Series provides a measurement of the residential real estate market. One of the measurements it produces is the Housing Affordability Index (HAI). It is a measure of the financial ability of U.S. families to buy a house. A value of 100 means that families earning the national median income have just the amount of money needed to qualify for a mortgage on a median-priced home; higher than 100 means they have more than enough and lower than 100 means they have less than enough. The file entitled **Index** contains the HAI and associated variables.

a. Produce the correlation matrix of all the variables. Predict the variables that will remain in the estimated regression equation if standard stepwise regression is used.

b. Use standard stepwise regression to develop an estimate of a model that is to predict the HAI from the associated variables found in the file entitled **Index**.

c. Compare the results of parts a and b. Explain any difference between the two models.

15-77. An investment analyst collected data from 20 randomly chosen companies. The data consisted of the 52-week-high stock prices, PE ratio, and the market value of the company. This data are in the file entitled **Investment**. The analyst wishes to produce a regression equation to predict the market value using the 52-week-high stock price and the PE ratio of the company. He creates a complete second-degree polynomial.

a. Construct an estimate of the regression equation using the indicated variables.
b. Produce the appropriate residual plots to determine if the polynomial function is the appropriate regression function for this data set.
c. Use a residual plot to determine if the residuals have a constant variance.
d. Produce the appropriate residual plot to determine if the residuals are independent. Assume the data were extracted in the order listed.
e. Construct a probability plot to determine if the error terms are normally distributed.

15-78. The consumer price index (CPI) is a measure of the average change in prices over time in a fixed market basket of goods and services typically purchased by consumers. One of the items in this market basket that affects the CPI is the price of oil and its derivatives. The file entitled **Consumer** contains the price of the derivatives of oil and the CPI adjusted to 2005 levels.
a. Produce a multiple regression equation depicting the relationship between the CPI and the price of the derivatives of oil.
b. Conduct a t-test on the coefficient which has the highest p-value. Use a significance level of 0.02 and the p-value approach.
c. Produce a multiple regression equation depicting the relationship between the CPI and the price of the derivatives of oil leaving out the variable tested in part b.
d. Referring to the regression results in part c, repeat the tests indicated in part b.
e. Perform a test of hypothesis to determine if the resulting overall model is statistically significant. Use a significance level of 0.02 and the p-value approach.

15-79. Refer to the State Department of Transportation data set called **Liabins**. The department was interested in determining the rate of compliance with the state's mandatory liability insurance law, as well as other things. *Assume the data were collected using a simple random sampling process.* Develop the best possible linear regression model using vehicle year as the dependent variable and any or all of the other variables as potential independent variables. Assume that your objective is to develop a predictive model. Write a report that discusses the steps you took to develop the final model. Include a correlation matrix and all appropriate statistical tests. Use an $\alpha = 0.05$. If you are using a nominal or ordinal variable, remember that you must make sure it is in the form of one or more dummy variables.

CASE 15.1

Dynamic Scales, Inc.

In 1985, Stanley Ahlon and three financial partners formed Dynamic Scales, Inc. The company was based on an idea Stanley had for developing a scale to weigh trucks in motion and thus eliminate the need for every truck to stop at weigh stations along highways. This dynamic scale would be placed in the highway approximately one-quarter mile from the regular weigh station. The scale would have a minicomputer that would automatically record truck speed, axle weights, and climate variables, including temperature, wind, and moisture. Stanley Ahlon and his partners believed that state transportation departments in the United States would be the primary market for such a scale.

As with many technological advances, developing the dynamic scale has been difficult. When the scale finally proved accurate for trucks traveling 40 miles per hour, it would not perform for trucks traveling at higher speeds. However, eight months ago, Stanley announced that the dynamic scale was ready to be field-tested by the Nebraska State Department of Transportation under a grant from the federal government. Stanley explained to his financial partners, and to Nebraska transportation officials, that the dynamic weight would not exactly equal the static weight (truck weight on a static scale). However he was sure a statistical relationship between dynamic weight and static weight could be determined, which would make the dynamic scale useful.

Nebraska officials, along with people from Dynamic Scales, installed a dynamic scale on a major highway in Nebraska. Each month for six months data were collected for a random sample of trucks weighed on both the dynamic scale and a static scale. Table 15.3 presents these data.

Once the data were collected, the next step was to determine whether, based on this test, the dynamic scale measurements could be used to predict static weights. A complete report will be submitted to the U.S. government and to Dynamic Scales.

TABLE 15.3 Test Data for the Dynamic Scales Example

Month	Front-Axle Static Weight	Front-Axle Dynamic Weight	Truck Speed	Temperature	Moisture
January	1,800 lb.	1,625 lb.	52 mph	21°F	0.00%
	1,311	1,904	71	17	0.15
	1,504	1,390	48	13	0.40
	1,388	1,402	50	19	0.10
	1,250	1,100	61	24	0.00
February	2,102	1,950	55	26	0.10
	1,410	1,475	58	32	0.20
	1,000	1,103	59	38	0.15
	1,430	1,387	43	24	0.00
	1,073	948	59	18	0.40
March	1,502	1,493	62	34	0.00
	1,721	1,902	67	36	0.00
	1,113	1,415	48	42	0.21
	978	983	59	29	0.32
	1,254	1,149	60	48	0.00
April	994	1,052	58	37	0.00
	1,127	999	52	34	0.21
	1,406	1,404	59	40	0.40
	875	900	47	48	0.00
	1,350	1,275	68	51	0.00
May	1,102	1,120	55	52	0.00
	1,240	1,253	57	57	0.00
	1,087	1,040	62	63	0.00
	993	1,102	59	62	0.10
	1,408	1,400	67	68	0.00
June	1,420	1,404	58	70	0.00
	1,808	1,790	54	71	0.00
	1,401	1,396	49	83	0.00
	933	1,004	62	88	0.40
	1,150	1,127	64	81	0.00

CASE 15.2

Glaser Machine Works

Glaser Machine Works had experienced a significant change in its business operations over the past 50 years. Glaser started business as a machine shop that produced specialty tools and products for the timber and lumber industry. This was a logical fit, given its location in the southern part of the United States. However, over the years Glaser looked to expand its offerings beyond the lumber and timber industry. Initially, its small size coupled with its rural location made it difficult to attract the attention of large companies that could use its products. All of that began to change as Glaser developed the ability to not only fabricate parts and tools but also to assemble products for customers who needed special components in large quantities. Glaser's business really took off when first foreign, and then domestic, automakers began to build automobile plants in the southern United States. Glaser was able to provide quality parts quickly for firms that expected high quality and responsive delivery. Many of Glaser's customers operated with little inventory and required that suppliers be able to provide shipments with short lead times.

As part of its relationship with the automobile industry, Glaser was expected to buy into the lean manufacturing and quality improvement initiatives of its customers. Glaser had always prided itself on its quality, but as the number and variety of its products increased, along with ever higher expectations by its customers, Glaser knew that it would have to respond by ensuring its quality and operations were continually improving. Of recent concern was the performance of its manufacturing line 107B. This line produced a component part for a Japanese automobile company. The Japanese firm had initially been pleased with Glaser's performance, but lately the number of defects was approaching an unacceptable level. Managers of the 107B line knew the line and its workers had been asked to ramp up production to meet increased demand and that some workers were concerned with the amount of overtime being required. There was also concern about the second shift now being run at 107B. Glaser had initially run only one shift, but when demand for its product became so high that there was not sufficient capacity with one shift, additional workers were hired to operate a night shift.

Management was wondering if the new shift had been stretched beyond its capabilities. Glaser plant management asked Kristi Johnson, the assistant production supervisor for line 107B, to conduct an analysis of product defects for the line. Kristi randomly selected several days of output and counted the number of defective parts produced on the 107B line. This information, along with other data, is contained in the file **Glaser Machine Works**. Kristi promised to have a full report for the management team by the end of the month.

CASE 15.3

Hawlins Manufacturing

Ross Hawlins had done it all at Hawlins Manufacturing, a company founded by his grandfather 63 years ago. Among his many duties, Ross oversaw all the plant's operations, a task that had grown in responsibility given the company's rapid growth over the past three decades. When Ross's grandfather founded the company, there were only two manufacturing sites. Expansion and acquisition of competitors over the years had caused that number to grow to over 50 manufacturing plants in 18 states.

Hawlins had a simple process that produced only two products, but the demand for these products was strong and Ross had spent millions of dollars upgrading his facilities over the past decade. Consequently, most of the company's equipment was less than 10 years old on average. Hawlins's two products were produced for local markets, as prohibitive shipping costs prevented shipping the product long distances. Product demand was sufficiently strong to support two manufacturing shifts (day and night) at every plant, and every plant had the capability to produce both products sold by Hawlins. Recently, the management team at Hawlins noticed that there were differences in output levels across the various plants. They were uncertain what, if anything, might explain these differences. Clearly, if some plants were more productive than others, there might be some meaningful insights that could be standardized across plants to boost overall productivity.

Because his duties required him to be involved in a variety of activities, Ross Hawlins has asked Lisa Chandler, an industrial engineer at the company's headquarters, to conduct a study of the plant's productivity. Lisa has randomly sampled 159 weeks of output from various plants together with the number of plant employees working that week, the plants average age in years, the product mix produced that week (either product A or B), and whether the output was from the day or night shift. The sampled data are contained in the file **Hawlins Manufacturing**. The Hawlins management team is expecting a written report and a presentation by Lisa when they meet again next Tuesday.

CASE 15.4

Sapphire Coffee—Part 2

Jennie Garcia could not believe that her career had moved so far so fast. When she left graduate school with a master's degree in anthropology, she intended to work at a local coffee shop until something else came along that was more related to her academic background. But after a few months she came to enjoy the business, and in a little over a year she was promoted to store manager. When the company for whom she worked continued to grow, Jennie was given oversight over a few stores.

Now, eight years after she started as a barista, Jennie was in charge of operations and planning for the company's southern region. As a part of her responsibilities, Jennie tracks store revenues and forecasts coffee demand. Historically, Sapphire Coffee would base its demand forecast on the number of stores, believing that each store sold approximately the same amount of coffee. This approach seemed to work well when

the company had shops of similar size and layout, but as the company grew, stores became more varied. Now, some stores had drive-thru windows, a feature that top management added to some stores believing that it would increase coffee sales for customers who wanted a cup of coffee on their way to work but who were too rushed to park and enter the store.

Jennie noticed that weekly sales seemed to be more variable across stores in her region and was wondering what, if anything, might explain the differences. The company's financial vice president had also noticed the increased differences in sales across stores and was wondering what might be happening. In an e-mail to Jennie he stated that weekly store sales are expected to average $5.00 per square foot. Thus, a 1,000-square-foot store would have average weekly sales of $5,000. He asked that Jennie analyze the stores in her region to see if this rule of thumb was a reliable measure of a store's performance.

Jennie had been in the business long enough to know that a store's size, while an important factor, was not the only thing that might influence sales. She had never been convinced of the efficacy of the drive-thru window, believing that it detracted from the coffee house experience that so many of Sapphire Coffee customers had come to expect. The VP of finance was expecting the analysis to be completed by the weekend. Jennie decided to randomly select weekly sales records for 53 stores, along with each store's size, whether it was located close to a college, and whether it had a drive-thru window. The data are in the file **Sapphire Coffee-2**. A full analysis would need to be sent to the corporate office by Friday.

CASE 15.5

Wendell Motors

Wendell Motors manufactures and ships small electric motors and drives to a variety of industrial and commercial customers in and around St. Louis. Wendell is a small operation with a single manufacturing plant. Wendell's products are different from other motor and drive manufacturers because it only produces small motors (25 horsepower or less) and because its products are used in a variety of industries and businesses that appreciate Wendell's quality and speed of delivery. Because it has only one plant, Wendell ships motors directly from the plant to its customers. Wendell's reputation for quality and speed of delivery allows it to maintain low inventories of motors and ship make-to-order products directly.

As part of its ongoing commitment to lean manufacturing and continuous process improvement, Wendell carefully monitors the cost associated with both production and shipping. The manager of shipping for Wendell, Tyler Jenkins, regularly reports the shipping costs to Wendell's management team. Because few finished goods inventories are maintained, competitive delivery times often require that Wendell expedite shipments. This is almost always the case for those customers who operate their business around the clock every day of the week. Such customers might maintain their own backup safety stock of a particular motor or drive, but circumstances often result in cases where replacement products have to be rushed through production and then expedited to the customer.

Wendell's management team wondered if these special orders were too expensive to handle in this way and if it might be less expensive to produce and hold certain motors as finished goods inventory, enabling off-the-shelf delivery using less expensive modes of shipping. This might especially be true for orders that must be filled on a holiday, incurring an additional shipping charge. At the last meeting of the management team, Tyler Jenkins was asked to analyze expedited shipping costs and to develop a model that could be used to estimate the cost of expediting a customer's order.

Donna Layton, an industrial engineer in the plant, was asked to prepare an inventory cost analysis to determine the expenses of holding additional finished goods inventory. To begin his analysis, Tyler randomly selected 45 expedited shipping records. The sampled data can be found in the file **Wendell Motors**. The management team expects a full report in five days. Tyler knew he would need a model for explaining shipping costs for expedited orders and that he would also need to answer the questions as to what effect, if any, shipping on a holiday had on costs.

References

Berenson, Mark L., and David M. Levine, *Basic Business Statistics: Concepts and Applications*, 10th ed. (Upper Saddle River, NJ: Prentice Hall, 2006).

Bowerman, Bruce L., and Richard T. O'Connell, *Linear Statistical Models: An Applied Approach*, 2nd ed. (Belmont, CA: Duxbury Press, 1990).

Cryer, Jonathan D., and Robert B. Miller, *Statistics for Business: Data Analysis and Modeling*, 2nd ed. (Belmont, CA: Duxbury Press, 1994).

Demmert, Henry, and Marshall Medoff, "Game-Specific Factors and Major League Baseball Attendance: An Econometric Study," *Santa Clara Business Review* (1977), pp. 49–56.

Dielman, Terry E., *Applied Regression Analysis: A Second Course in Business and Economic Statistics*, 4th ed. (Belmont, CA: Duxbury Press, 2005).

Draper, Norman R., and Harry Smith, *Applied Regression Analysis*, 3rd ed. (New York: John Wiley and Sons, 1998).

Frees, Edward W., *Data Analysis Using Regression Models: The Business Perspective* (Upper Saddle River, NJ: Prentice Hall, 1996).

Gloudemans, Robert J., and Dennis Miller, "Multiple Regression Analysis Applied to Residential Properties." *Decision Sciences 7* (April 1976) pp. 294–304.

Kleinbaum, David G., Lawrence L. Kupper, Keith E. Muller, and Azhar Nizam, *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed. (Belmont, CA: Duxbury Press, 1998).

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill Irwin, 2005).

Microsoft Excel 2007 (Redmond, WA: Microsoft Corp. 2007).

Minitab for Windows Version 14 (State College, PA: Minitab, 2005).