14

Multiple Regression Analysis

Learning Objectives

When you have completed this chapter, you will be able to:

LO1 Describe the relationship between several independent variables and a dependent variable using multiple regression analysis.

LO2 Set up, interpret, and apply an ANOVA table.

LO3 Compute and interpret measures of association in multiple regression.

LO4 Conduct a hypothesis test to determine whether a set of regression coefficients differ from zero.

LO5 Conduct a hypothesis test of each regression coefficient.

LO6 Use residual analysis to evaluate the assumptions of multiple regression analysis.

LO7 Evaluate the effects of correlated independent variables.

LO8 Evaluate and use qualitative independent variables.

LO9 Explain the possible interaction among independent variables.

LO10 Explain stepwise regression.

The mortgage department of the Bank of New England is studying data from recent loans. Of particular interest is how such factors as the value of the home being purchased, education level of the head of the household, age of the head of the household, current monthly mortgage payment, and gender of the head of the household relate to the family income. Are the proposed variables effective predictors of the dependent variable family income? (See the Example/Solution in Section 14.9 and LO1.)

14.1 Introduction

In Chapter 13, we described the relationship between a pair of interval- or ratio-scaled variables. We began the chapter by studying the correlation coefficient, which measures the strength of the relationship. A coefficient near plus or minus 1.00 (−.88 or .78, for example) indicates a very strong linear relationship, whereas a value near 0 (−.12 or .18, for example) means that the relationship is weak. Next we developed a procedure to determine a linear equation to express the relationship between the two variables. We referred to this as a *regression line.* This line describes the relationship between the variables. It also describes the overall pattern of a dependent variable (Y) to a single independent or explanatory variable (X).

In multiple linear correlation and regression, we use additional independent variables (denoted X_1, X_2, . . . , and so on) that help us better explain or predict the dependent variable (Y). Almost all of the ideas we saw in simple linear correlation and regression extend to this more general situation. However, the additional independent variables do lead to some new considerations. Multiple regression analysis can be used either as a descriptive or as an inferential technique.

14.2 Multiple Regression Analysis

LO1 Describe the relationship between several independent variables and a dependent variable using multiple regression analysis.

The general descriptive form of a multiple linear equation is shown in formula (14–1). We use k to represent the number of independent variables. So k can be any positive integer.

GENERAL MULTIPLE REGRESSION EQUATION

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k \qquad \textbf{[14–1]}$$

where:

a is the intercept, the value of Y when all the X's are zero.

b_j is the amount by which Y changes when that particular X_j increases by one unit, with the values of all other independent variables held constant. The subscript j is simply a label that helps to identify each independent variable; it is not used in any calculations. Usually the subscript is an integer value between 1 and k, which is the number of independent variables. However, the subscript can also be a short or abbreviated label. For example, age could be used as a subscript.

In Chapter 13, the regression analysis described and tested the relationship between a dependent variable, $\hat{Y}$, and a single independent variable, X. The relationship between $\hat{Y}$ and X was graphically portrayed by a line. When there are two independent variables, the regression equation is

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

Because there are two independent variables, this relationship is graphically portrayed as a plane and is shown in Chart 14–1. The chart shows the residuals as the difference between the actual Y and the fitted $\hat{Y}$ on the plane. If a multiple regression analysis includes more than two independent variables, we cannot use a graph to illustrate the analysis since graphs are limited to three dimensions.

To illustrate the interpretation of the intercept and the two regression coefficients, suppose a vehicle's mileage per gallon of gasoline is directly related to the octane rating of the gasoline being used (X_1) and inversely related to the weight of the automobile (X_2). Assume that the regression equation, calculated using statistical software, is:

$$\hat{Y} = 6.3 + 0.2X_1 - 0.001X_2$$

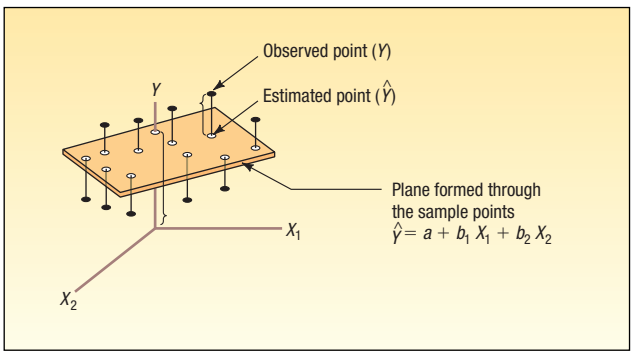


CHART 14–1 Regression Plane with 10 Sample Points

The intercept value of 6.3 indicates the regression equation intersects the Y-axis at 6.3 when both X_1 and X_2 are zero. Of course, this does not make any physical sense to own an automobile that has no (zero) weight and to use gasoline with no octane. It is important to keep in mind that a regression equation is not generally used outside the range of the sample values.

The b_1 of 0.2 indicates that for each increase of 1 in the octane rating of the gasoline, the automobile would travel 2/10 of a mile more per gallon, *regardless of the weight of the vehicle.* The b_2 value of -0.001 reveals that for each increase of one pound in the vehicle's weight, the number of miles traveled per gallon decreases by 0.001, *regardless of the octane of the gasoline being used.*

As an example, an automobile with 92-octane gasoline in the tank and weighing 2,000 pounds would travel an average 22.7 miles per gallon, found by:

$$\hat{Y} = a + b_1X_1 + b_2X_2 = 6.3 + 0.2(92) - 0.001(2{,}000) = 22.7$$

The values for the coefficients in the multiple linear equation are found by using the method of least squares. Recall from the previous chapter that the least squares method makes the sum of the squared differences between the fitted and actual values of Y as small as possible, that is, the term $\Sigma(Y - \hat{Y})^2$ is minimized. The calculations are very tedious, so they are usually performed by a statistical software package, such as Excel or Minitab.

In the following example, we show a multiple regression analysis using three independent variables employing Excel and Minitab. Both packages report a standard set of statistics and reports. However, Minitab also provides advanced regression analysis techniques that we will use later in the chapter.

Example

Salsberry Realty sells homes along the east coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at Salsberry has been asked to develop some guidelines regarding heating costs for single-family homes. Three variables are thought to relate to the heating costs: (1) the mean daily outside temperature, (2) the number of inches of insulation in the attic, and (3) the age in years of the furnace. To investigate, Salsberry's research department selected a random sample of 20 recently sold homes. It determined the cost to heat each home last January, as well as the January outside temperature in the region, the number of inches of insulation in the attic, and the age of the furnace. The sample information is reported in Table 14–1.

Statistics in Action

Many studies indicate a woman will earn about 70 percent of what a man would for the same work. Researchers at the University of Michigan Institute for Social Research found that about one-third of the difference can be explained by such social factors as differences in education, seniority, and work interruptions. The remaining two-thirds is not explained by these social factors.

TABLE 14–1 Factors in January Heating Cost for a Sample of 20 Homes

Home	Heating Cost ($)	Mean Outside Temperature (°F)	Attic Insulation (inches)	Age of Furnace (years)
1	$250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

The data in Table 14–1 is available in both Excel and Minitab formats at the textbook website, www.mhhe.com/lind15e. The basic instructions for using Excel and Minitab for this data are in the Software Commands section at the end of this chapter.

Determine the multiple regression equation. Which variables are the independent variables? Which variable is the dependent variable? Discuss the regression coefficients. What does it indicate if some coefficients are positive and some coefficients are negative? What is the intercept value? What is the estimated heating cost for a home if the mean outside temperature is 30 degrees, there are 5 inches of insulation in the attic, and the furnace is 10 years old?

Solution

We begin the analysis by defining the dependent and independent variables. The dependent variable is the January heating cost. It is represented by Y. There are three independent variables:

- The mean outside temperature in January, represented by X_1.
- The number of inches of insulation in the attic, represented by X_2.
- The age in years of the furnace, represented by X_3.

Given these definitions, the general form of the multiple regression equation follows. The value $\hat{Y}$ is used to estimate the value of Y.

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$$

Now that we have defined the regression equation, we are ready to use either Excel or Minitab to compute all the statistics needed for the analysis. The outputs from the two software systems are shown below.

To use the regression equation to predict the January heating cost, we need to know the values of the regression coefficients, b_j. These are highlighted in the software reports. Note that the software used the variable names or labels associated with each independent variable. The regression equation intercept, a, is labeled as "constant" in the Minitab output and "intercept" in the Excel output.

Tbl14-1.mtw ***

	C1	C2	C3	C4
	Cost	Temp	Insul	Age
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5
21				
22				
23				
24				

Session

```
6/14/2008 10:31:38 AM

Welcome to Minitab, press F1 for help.

Results for: Tbl14-1.mtw

Regression Analysis: Cost versus Temp, Insul, Age

The regression equation is
Cost = 427 - 4.58 Temp - 14.8 Insul + 6.10 Age


Predictor      Coef  SE Coef      T      P
Constant     427.19    59.60   7.17  0.000
Temp        -4.5827   0.7723  -5.93  0.000
Insul       -14.831    4.754  -3.12  0.007
Age           6.101    4.012   1.52  0.148


S = 51.0486   R-Sq = 80.4%   R-Sq(adj) = 76.7%


Analysis of Variance

Source           DF      SS     MS      F      P
Regression        3  171220  57073  21.90  0.000
Residual Error   16   41695   2606
Total            19  212916
```

regression [Compatibility Mode]

	A	B	C	D	F	G	H	I	J	K	L
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		*Regression Statistics*					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	
21	139	30	7	5							

In this case, the estimated regression equation is:

$$\hat{Y} = 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3$$

We can now estimate or predict the January heating cost for a home if we know the mean outside temperature, the inches of insulation, and the age of the furnace. For an example home, the mean outside temperature for the month is 30 degrees (X_1), there are 5 inches of insulation in the attic (X_2), and the furnace is 10 years old (X_3). By substituting the values for the independent variables:

$$\hat{Y} = 427.194 - 4.583(30) - 14.831(5) + 6.101(10) = 276.56$$

The estimated January heating cost is $276.56.

The regression coefficients, and their algebraic signs, also provide information about their individual relationships with the January heating cost. The regression coefficient for mean outside temperature is −4.583. The coefficient is negative and shows an inverse relationship between heating cost and temperature. This is not surprising. As the outside temperature increases, the cost to heat the home decreases. The numeric value of the regression coefficient provides more information. If we increase temperature by 1 degree and hold the other two independent variables constant, we can estimate a decrease of $4.583 in monthly heating cost. So if the mean temperature in Boston is 25 degrees and it is 35 degrees in Philadelphia, all other things being the same (insulation and age of furnace), we expect the heating cost would be $45.83 less in Philadelphia.

The attic insulation variable also shows an inverse relationship: the more insulation in the attic, the less the cost to heat the home. So the negative sign for this coefficient is logical. For each additional inch of insulation, we expect the cost to heat the home to decline $14.83 per month, holding the outside temperature and the age of the furnace constant.

The age of the furnace variable shows a direct relationship. With an older furnace, the cost to heat the home increases. Specifically, for each additional year older the furnace is, we expect the cost to increase $6.10 per month.

Self-Review 14–1

There are many restaurants in northeastern South Carolina. They serve beach vacationers in the summer, golfers in the fall and spring, and snowbirds in the winter. Bill and Joyce Tuneall manage several restaurants in the North Jersey area and are considering moving to Myrtle Beach, SC, to open a new restaurant. Before making a final decision, they wish to investigate existing restaurants and what variables seem to be related to profitability. They gather sample information where profit (reported in $000) is the dependent variable and the independent variables are:

X_1 the number of parking spaces near the restaurant.
X_2 the number of hours the restaurant is open per week.
X_3 the distance from Peaches Corner, a landmark in Myrtle Beach.
X_4 the number of servers employed.
X_5 the number of years the current owner has owned the restaurant.

The following is part of the output obtained using statistical software.

Predictor	Coef	SE Coef	T
Constant	2.50	1.50	1.667
X_1	3.00	1.500	2.000
X_2	4.00	3.000	1.333
X_3	−3.00	0.20	−15.00
X_4	0.20	.05	4.00
X_5	1.00	1.50	0.667

(a) What is the amount of profit for a restaurant with 40 parking spaces that is open 72 hours per week, is 10 miles from Peaches Corner, has 20 servers, and has been open 5 years?
(b) Interpret the values of b_2 and b_3 in the multiple regression equation.

Exercises

connect™

1. The director of marketing at Reeves Wholesale Products is studying monthly sales. Three independent variables were selected as estimators of sales: regional population, per capita income, and regional unemployment rate. The regression equation was computed to be (in dollars):

$$\hat{Y} = 64{,}100 + 0.394X_1 + 9.6X_2 - 11{,}600X_3$$

a. What is the full name of the equation?
b. Interpret the number 64,100.
c. What are the estimated monthly sales for a particular region with a population of 796,000, per capita income of $6,940, and an unemployment rate of 6.0 percent?

2. Thompson Photo Works purchased several new, highly sophisticated processing machines. The production department needed some guidance with respect to qualifications needed by an operator. Is age a factor? Is the length of service as an operator (in years) important? In order to explore further the factors needed to estimate performance on the new processing machines, four variables were listed:

X_1 = Length of time an employee was in the industry
X_2 = Mechanical aptitude test score
X_3 = Prior on-the-job rating
X_4 = Age

Performance on the new machine is designated *Y*.

Thirty employees were selected at random. Data were collected for each, and their performances on the new machines were recorded. A few results are:

Name	Performance on New Machine, Y	Length of Time in Industry, X_1	Mechanical Aptitude Score, X_2	Prior On-the-Job Performance, X_3	Age, X_4
Mike Miraglia	112	12	312	121	52
Sue Trythall	113	2	380	123	27

The equation is:

$$\hat{Y} = 11.6 + 0.4X_1 + 0.286X_2 + 0.112X_3 + 0.002X_4$$

a. What is this equation called?
b. How many dependent variables are there? Independent variables?
c. What is the number 0.286 called?
d. As age increases by one year, how much does estimated performance on the new machine increase?
e. Carl Knox applied for a job at Photo Works. He has been in the business for six years, and scored 280 on the mechanical aptitude test. Carl's prior on-the-job performance rating is 97, and he is 35 years old. Estimate Carl's performance on the new machine.

3. A sample of General Mills employees was studied to determine their degree of satisfaction with their quality of life. A special index, called the index of satisfaction, was used to measure satisfaction. Six factors were studied, namely, age at the time of first marriage (X_1), annual income (X_2), number of children living (X_3), value of all assets (X_4), status of health in the form of an index (X_5), and the average number of social activities per week—such as bowling and dancing (X_6). Suppose the multiple regression equation is:

$$\hat{Y} = 16.24 + 0.017X_1 + 0.0028X_2 + 42X_3 + 0.0012X_4 + 0.19X_5 + 26.8X_6$$

a. What is the estimated index of satisfaction for a person who first married at 18, has an annual income of $26,500, has three children living, has assets of $156,000, has an index of health status of 141, and has 2.5 social activities a week on the average?
b. Which would add more to satisfaction, an additional income of $10,000 a year or two more social activities a week?

4. Cellulon, a manufacturer of home insulation, wants to develop guidelines for builders and consumers on how the thickness of the insulation in the attic of a home and the outdoor temperature affect natural gas consumption. In the laboratory, it varied the insulation thickness and temperature. A few of the findings are:

Monthly Natural Gas Consumption (cubic feet), Y	Thickness of Insulation (inches), X_1	Outdoor Temperature (°F), X_2
30.3	6	40
26.9	12	40
22.1	8	49

On the basis of the sample results, the regression equation is:

$$\hat{Y} = 62.65 - 1.86X_1 - 0.52X_2$$

a. How much natural gas can homeowners expect to use per month if they install 6 inches of insulation and the outdoor temperature is 40 degrees F?
b. What effect would installing 7 inches of insulation instead of 6 have on the monthly natural gas consumption (assuming the outdoor temperature remains at 40 degrees F)?
c. Why are the regression coefficients b_1 and b_2 negative? Is this logical?

14.3 Evaluating a Multiple Regression Equation

Many statistics and statistical methods are used to evaluate the relationship between a dependent variable and more than one independent variable. Our first step was to write the relationship in terms of a multiple regression equation. The next step follows on the concepts presented in Chapter 13 by using the information in an ANOVA table to evaluate how well the equation fits the data.

The ANOVA Table

LO2 Set up, interpret, and apply an ANOVA table.

As in Chapter 13, the statistical analysis of a multiple regression equation is summarized in an ANOVA table. To review, the total variation of the dependent variable, *Y,* is divided into two components: (1) *regression,* or the variation of *Y* explained by all the independent variables and (2) *the error or residual,* or unexplained variation of *Y.* These two categories are identified in the first column of an ANOVA table below. The column headed "*df*" refers to the degrees of freedom associated with each category. The total number of degrees of freedom is $n - 1$. The number of degrees of freedom in the regression is equal to the number of independent variables in the multiple regression equation. We call the regression degrees of freedom *k.* The number of degrees of freedom associated with the error term is equal to the total degrees of freedom minus the regression degrees of freedom. In multiple regression, the degrees of freedom are $n - (k + 1)$.

Source	*df*	SS	MS	*F*
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Residual or error	$n - (k + 1)$	SSE	MSE = SSE/$[n - (k + 1)]$	
Total	$n - 1$	SS total		

The term "SS" located in the middle of the ANOVA table refers to the sum of squares. Notice that there is a sum of squares for each source of variation. The sum of squares column shows the amount of variation attributable to each source. The total variation of the dependent variable, *Y,* is summarized in SS total. You should

note that this is simply the numerator of the usual formula to calculate any variation—in other words, the sum of the squared deviations from the mean. It is computed as:

$$\text{Total Sum of Squares} = \text{SS total} = \Sigma(Y - \overline{Y})^2$$

As we have seen, the total sum of squares is the sum of the regression and residual sum of squares. The regression sum of squares is the sum of the squared differences between the estimated or predicted values, $\hat{Y}$, and the overall mean of Y. The regression sum of squares is found by:

$$\text{Regression Sum of Squares} = \text{SSR} = \Sigma(\hat{Y} - \overline{Y})^2$$

The residual sum of squares is the sum of the squared differences between the observed values of the dependent variable, Y, and their corresponding estimated or predicted values, $\hat{Y}$. Notice that this difference is the error of estimating or predicting the dependent variable with the multiple regression equation. It is calculated as:

$$\text{Residual or Error Sum of Squares} = \text{SSE} = \Sigma(Y - \hat{Y})^2$$

We will use the ANOVA table information from the previous example to evaluate the regression equation to estimate January heating costs.

	A	B	C	D	F	G	H	I	J	K	L
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		*Regression Statistics*					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	

Multiple Standard Error of Estimate

LO3 Compute and interpret measures of association in multiple regression.

We begin with the **multiple standard error of estimate.** Recall that the standard error of estimate is comparable to the standard deviation. To explain the details of the standard error of estimate, refer to the first sampled home in Table 14–1 in the previous example on page 515. The actual heating cost for the first observation, Y, is \$250, the outside temperature, X_1, is 35 degrees, the depth of insulation, X_2, is 3 inches, and the age of the furnace, X_3, is 6 years. Using the regression equation developed in the previous section, the estimated heating cost for this home is:

$$\begin{aligned}\hat{Y} &= 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3 \\ &= 427.194 - 4.583(35) - 14.831(3) + 6.101(6) \\ &= 258.90\end{aligned}$$

So we would estimate that a home with a mean January outside temperature of 35 degrees, 3 inches of insulation, and a 6-year-old furnace would cost \$258.90 to heat. The actual heating cost was \$250, so the residual—which is the difference between the actual value and the estimated value—is $Y - \hat{Y} = 250 - 258.90 = -8.90$. This difference of \$8.90 is the random or unexplained error for the first home sampled. Our next step is to square this difference—that is, find $(Y - \hat{Y})^2 = (250 - 258.90)^2 = (-8.90)^2 = 79.21$.

If we repeat this calculation for the other 19 observations and sum all 20 squared differences, the total will be the residual or error sum of squares from the ANOVA table. Using this information, we can calculate the multiple standard error of the estimate as:

MULTIPLE STANDARD ERROR OF ESTIMATE

$$s_{Y.123...k} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k + 1)}} = \sqrt{\frac{\text{SSR}}{n - (k + 1)}} \qquad [14\text{–}2]$$

where:

Y is the actual observation.
$\hat{Y}$ is the estimated value computed from the regression equation.
n is the number of observations in the sample.
k is the number of independent variables.
SSR is the Residual Sum of Squares from an ANOVA table.

There is still more information in the ANOVA table that can be used to compute the multiple standard error of the estimate. Note that the next column in the ANOVA table is labeled MS, or mean square. For the regression and residual sources of variation, the mean squares are calculated as the sum of squares divided by its corresponding degrees of freedom. In the case of the multiple standard error of the mean, the multiple standard error of the estimate is the square root of the residual mean square.

$$s_{Y.123...K} = \sqrt{\text{MSE}} = \sqrt{2605.995} = \$51.05$$

How do we interpret the standard error of estimate of 51.05? It is the typical "error" when we use this equation to predict the cost. First, the units are the same as the dependent variable, so the standard error is in dollars, $51.05. Second, we expect the residuals to be approximately normally distributed, so about 68 percent of the residuals will be within ±$51.05 and about 95 percent within ±2(51.05) or ±$102.10. As before with similar measures of dispersion, such as the standard error of estimate in Chapter 13, a smaller multiple standard error indicates a better or more effective predictive equation.

Coefficient of Multiple Determination

Next, let's look at the coefficient of multiple determination. Recall from the previous chapter the coefficient of determination is defined as the percent of variation in the dependent variable explained, or accounted for, by the independent variable. In the multiple regression case, we extend this definition as follows.

COEFFICIENT OF MULTIPLE DETERMINATION The percent of variation in the dependent variable, Y, explained by the set of independent variables, X_1, X_2, X_3, . . . X_k.

The characteristics of the coefficient of multiple determination are:

1. **It is symbolized by a capital *R* squared.** In other words, it is written as R^2 because it behaves like the square of a correlation coefficient.
2. **It can range form 0 to 1.** A value near 0 indicates little association between the set of independent variables and the dependent variable. A value near 1 means a strong association.
3. **It cannot assume negative values.** Any number that is squared or raised to the second power cannot be negative.
4. **It is easy to interpret.** Because R^2 is a value between 0 and 1, it is easy to interpret, compare, and understand.

We can calculate the coefficient of determination from the information found in the ANOVA table. We look in the sum of squares column, which is labeled SS in the Excel output, and use the regression sum of squares, SSR, then divide by the total sum of squares, SS total.

COEFFICIENT OF MULTIPLE DETERMINATION

$$R^2 = \frac{\text{SSR}}{\text{SS total}} \qquad \textbf{[14–3]}$$

Using the residual and total sum of squares from the ANOVA table, we can use formula (14–3) to calculate the coefficient of multiple determination.

$$R^2 = \frac{\text{SSR}}{\text{SS total}} = \frac{171{,}220}{212{,}916} = .804$$

How do we interpret this value? We conclude that the independent variables (outside temperature, amount of insulation, and age of furnace) explain, or account for, 80.4 percent of the variation in heating cost. To put it another way, 19.6 percent of the variation is due to other sources, such as random error or variables not included in the analysis. Using the ANOVA table, 19.6 percent is the error sum of squares divided by the total sum of squares. Knowing that the SSR + SSE = SS total, the following relationship is true.

$$1 - R^2 = 1 - \frac{\text{SSR}}{\text{SS total}} = \frac{\text{SSE}}{\text{SS total}} = \frac{41{,}695}{212{,}916} = .196$$

Adjusted Coefficient of Determination

The number of independent variables in a multiple regression equation makes the coefficient of determination larger. Each new independent variable causes the predictions to be more accurate. That, in turn, makes SSE smaller and SSR larger. Hence, R^2 increases only because of the total number of independent variables and not because the added independent variable is a good predictor of the dependent variable. In fact, if the number of variables, *k,* and the sample size, *n,* are equal, the coefficient of determination is 1.0. In practice, this situation is rare and would also be ethically questionable. To balance the effect that the number of independent variables has on the coefficient of multiple determination, statistical software packages use an *adjusted* coefficient of multiple determination.

ADJUSTED COEFFICIENT OF DETERMINATION

$$R^2_{\text{adj}} = 1 - \frac{\dfrac{\text{SSE}}{n - (k + 1)}}{\dfrac{\text{SS total}}{n - 1}} \qquad \textbf{[14–4]}$$

The error and total sum of squares are divided by their degrees of freedom. Notice especially the degrees of freedom for the error sum of squares includes *k,* the number of independent variables. For the cost of heating example, the adjusted coefficient of determination is:

$$R^2_{\text{adj}} = 1 - \frac{\dfrac{41{,}695}{20 - (3 + 1)}}{\dfrac{212{,}916}{20 - 1}} = 1 - \frac{2{,}606}{11{,}206.0} = 1 - .23 = .77$$

If we compare the R^2 (0.80) to the adjusted R^2 (0.77), the difference in this case is small.

Self-Review 14–2

Refer to Self-Review 14–1 on the subject of restaurants in Myrtle Beach. The ANOVA portion of the regression output is presented below.

```
Analysis of Variance
Source            DF     SS    MS
Regression         5    100    20
Residual Error    20     40     2
Total             25    140
```

(a) How large was the sample?
(b) How many independent variables are there?
(c) How many dependent variables are there?
(d) Compute the standard error of estimate. About 95 percent of the residuals will be between what two values?
(e) Determine the coefficient of multiple determination. Interpret this value.
(f) Find the coefficient of multiple determination, adjusted for the degrees of freedom.

Exercises

connect™

5. Consider the ANOVA table that follows.

```
Analysis of Variance
Source            DF         SS        MS       F        P
Regression         2     77.907    38.954    4.14    0.021
Residual Error    62    583.693     9.414
Total             64    661.600
```

a. Determine the standard error of estimate. About 95 percent of the residuals will be between what two values?
b. Determine the coefficient of multiple determination. Interpret this value.
c. Determine the coefficient of multiple determination, adjusted for the degrees of freedom.

6. Consider the ANOVA table that follows.

```
Analysis of Variance
Source            DF         SS        MS        F
Regression         5    3710.00    742.00    12.89
Residual Error    46    2647.38     57.55
Total             51    6357.38
```

a. Determine the standard error of estimate. About 95 percent of the residuals will be between what two values?
b. Determine the coefficient of multiple determination. Interpret this value.
c. Determine the coefficient of multiple determination, adjusted for the degrees of freedom.

14.4 Inferences in Multiple Linear Regression

Thus far, multiple regression analysis has been viewed only as a way to describe the relationship between a dependent variable and several independent variables. However, the least squares method also has the ability to draw inferences or generalizations about the relationship for an entire population. Recall that when you create confidence intervals or perform hypothesis tests as a part of inferential statistics, you view the data as a random sample taken from some population.

In the multiple regression setting, we assume there is an unknown population regression equation that relates the dependent variable to the k independent

variables. This is sometimes called a **model** of the relationship. In symbols we write:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

This equation is analogous to formula (14–1) except the coefficients are now reported as Greek letters. We use the Greek letters to denote *population parameters*. Then under a certain set of assumptions, which will be discussed shortly, the computed values of a and b_j are sample statistics. These sample statistics are point estimates of the corresponding population parameters α and β_j. For example, the sample regression coefficient b_2 is a point estimate of the population parameter β_2. The sampling distribution of these point estimates follows the normal probability distribution. These sampling distributions are each centered at their respective parameter values. To put it another way, the means of the sampling distributions are equal to the parameter values to be estimated. Thus, by using the properties of the sampling distributions of these statistics, inferences about the population parameters are possible.

Global Test: Testing the Multiple Regression Model

LO4 Conduct a hypothesis test to determine whether a set of regression coefficients differ from zero.

We can test the ability of the independent variables $X_1, X_2, \ldots, X_k$ to explain the behavior of the dependent variable Y. To put this in question form: Can the dependent variable be estimated without relying on the independent variables? The test used is referred to as the **global test.** Basically, it investigates whether it is possible all the independent variables have zero regression coefficients.

To relate this question to the heating cost example, we will test whether the independent variables (amount of insulation in the attic, mean daily outside temperature, and age of furnace) effectively estimate home heating costs. In testing a hypothesis, we first state the null hypothesis and the alternate hypothesis. In the heating cost example, there are three independent variables. Recall that b_1, b_2, and b_3 are sample regression coefficients. The corresponding coefficients in the population are given the symbols β_1, β_2, and β_3. We now test whether the regression coefficients in the population are all zero. The null hypothesis is:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

The alternate hypothesis is:

$$H_1: \text{Not all the } \beta_i\text{'s are 0.}$$

If the null hypothesis is true, it implies the regression coefficients are all zero and, logically, are of no use in estimating the dependent variable (heating cost). Should that be the case, we would have to search for some other independent variables—or take a different approach—to predict home heating costs.

To test the null hypothesis that the multiple regression coefficients are all zero, we employ the F distribution introduced in Chapter 12. We will use the .05 level of significance. Recall these characteristics of the F distribution:

1. **There is a family of *F* distributions.** Each time the degrees of freedom in either the numerator or the denominator changes, a new F distribution is created.
2. **The *F* distribution cannot be negative.** The smallest possible value is 0.
3. **It is a continuous distribution.** The distribution can assume an infinite number of values between 0 and positive infinity.
4. **It is positively skewed.** The long tail of the distribution is to the right-hand side. As the number of degrees of freedom increases in both the numerator and the denominator, the distribution approaches the normal probability distribution. That is, the distribution will move toward a symmetric distribution.
5. **It is asymptotic.** As the values of X increase, the F curve will approach the horizontal axis, but will never touch it.

The F-statistic to test the global hypothesis follows. As in Chapter 12, it is the ratio of two variances. In this case, the numerator is the regression sum of squares

divided by its degrees of freedom, k. The denominator is the residual sum of squares divided by its degrees of freedom, $n - (k + 1)$. The formula follows.

GLOBAL TEST

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} \qquad \textbf{[14–5]}$$

Using the ANOVA table, the F-statistic is

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{\text{MSR}}{\text{MSE}} = 21.90$$

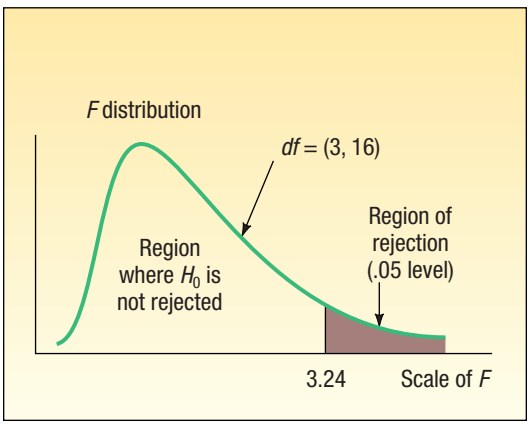
Remember that the F-statistic tests the basic null hypothesis that two variances or, in this case, two mean squares are equal. Also remember that we always put the larger of the two variances in the numerator. In our global multiple regression hypothesis test, we will reject the null hypothesis, H_0, that all regression coefficients are zero when the regression mean square is larger in comparison to the residual mean square. If this is true, the F-statistic will be relatively large and in the far right tail of the F-distribution, and the p-value will be small, that is, less than our choice of our significance level of 0.05. Thus, we will reject the null hypothesis.

As with other hypothesis-testing methods, the decision rule can be based on either of two methods: (1) comparing the test statistic to a critical value or (2) calculating a p-value based on the test statistic and comparing the p-value to the significance level. Using the critical value method, we first find the critical value of F that requires three pieces of information: (1) the numerator degrees of freedom, (2) the denominator degrees of freedom, and (3) the significance level. The degrees of freedom for the numerator and the denominator are reported in the Excel ANOVA table that follows. The ANOVA output is highlighted in light green. The top number in the column marked "*df*" is 3, indicating there are 3 degrees of freedom in the numerator. This value corresponds to the number of independent variables. The middle number in the "*df*" column (16) indicates that there are 16 degrees of freedom in the denominator. The number 16 is found by $n - (k - 1) = 20 - (3 - 1) = 16$.

regression [Compatibility Mode]

	A	B	C	D	F	G	H	I	J	K	L
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		*Regression Statistics*					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	
21	139	30	7	5							

The critical value of F is found in Appendix B.4. Using the table for the .05 significance level, move horizontally to 3 degrees of freedom in the numerator, then down to 16 degrees of freedom in the denominator, and read the critical value. It is 3.24. The region where H_0 is not rejected and the region where H_0 is rejected are shown in the following diagram.



Continuing with the global test, the decision rule is: Do not reject the null hypothesis, H_0, that all the regression coefficients are 0 if the computed value of F is less than or equal to 3.24. If the computed F is greater than 3.24, reject H_0 and accept the alternate hypothesis, H_1.

The computed value of F is 21.90, which is in the rejection region. The null hypothesis that all the multiple regression coefficients are zero is therefore rejected. This means that at least one of the independent variables has the ability to explain the variation in the dependent variable (heating cost). We expected this decision. Logically, the outside temperature, the amount of insulation, or age of the furnace have a great bearing on heating costs. The global test assures us that they do.

Testing the null hypothesis can also be based on a p-value, which is reported in the computer software output for all hypothesis tests. In the case of the F-statistic, the p-value is defined as the probability of observing an F-value as large or larger than the F test statistic, assuming the null hypothesis is true. If the p-value is less than our selected significance level, then we decide to reject the null hypothesis. The ANOVA shows the F-statistic's p-value is equal to 0.000. It is clearly less than our significance level of 0.05. Therefore, we decide to reject the global null hypothesis and conclude that at least one of the regression coefficients is not equal to zero.

The decision is the same as when we used the critical value approach. The advantage to using the p-value approach is that the p-value gives us a "flavor" of the decision. The computed p-value is much smaller than our significance level (.000 versus .05). We reject the null hypothesis that all the regression coefficients are 0 and, on the basis of the p-value, conclude that there is little likelihood this hypothesis is true.

Evaluating Individual Regression Coefficients

LO5 Conduct a hypothesis test of each regression coefficient.

So far we have shown that at least one, but not necessarily all, of the regression coefficients are not equal to zero and thus useful for predictions. The next step is to test the independent variables *individually* to determine which regression coefficients may be 0 and which are not.

Why is it important to know if any of the β_i's equal 0? If a β could equal 0, it implies that this particular independent variable is of no value in explaining any variation in the dependent value. If there are coefficients for which H_0 cannot be rejected, we may want to eliminate them from the regression equation.

We will now conduct three separate tests of hypothesis—for temperature, for insulation, and for the age of the furnace.

For temperature:	For insulation:	For furnace age:
H_0: $\beta_1 = 0$	H_0: $\beta_2 = 0$	H_0: $\beta_3 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_2 \neq 0$	H_1: $\beta_3 \neq 0$

We will test the hypotheses at the .05 level. Note that these are two-tailed tests.

The test statistic follows Student's t distribution with $n - (k + 1)$ degrees of freedom. The number of sample observations is n. There are 20 homes in the study, so $n = 20$. The number of independent variables is k, which is 3. Thus, there are $n - (k + 1) = 20 - (3 + 1) = 16$ degrees of freedom.

The critical value for t is in Appendix B.2. For a two-tailed test with 16 degrees of freedom using the .05 significance level, H_0 is rejected if t is less than -2.120 or greater than 2.120.

Refer to the Excel output in the previous section. (See page 525.) The column highlighted in yellow, headed Coefficients, shows the values for the multiple regression equation:

$$\hat{Y} = 427.194 - 4.583X_1 - 14.831X_2 + 6.101X_3$$

Interpreting the term $-4.583X_1$ in the equation: For each degree the temperature increases, it is expected that the heating cost will decrease about \$4.58, holding the two other variables constant.

The column in the Excel output labeled "Standard Error" shows the standard error of the sample regression coefficients. Recall that Salsberry Realty selected a sample of 20 homes along the East Coast of the United States. If Salsberry Realty selected a second random sample and computed the regression coefficients for that sample, the values would not be exactly the same. If the sampling process was repeated many times, we could construct a sampling distribution for each of these regression coefficients. The column labeled "Standard Error" estimates the variability for each of these regression coefficients. The sampling distributions of the coefficients follow the t distribution with $n - (k + 1)$ degrees of freedom. Hence, we are able to test the independent variables individually to determine whether the net regression coefficients differ from zero. The formula is:

TESTING INDIVIDUAL REGRESSION COEFFICIENTS

$$t = \frac{b_i - 0}{s_{b_i}} \qquad \textbf{[14–6]}$$

The b_i refers to any one of the regression coefficients, and s_{b_i} refers to the standard deviation of that distribution of the regression coefficient. We include 0 in the equation because the null hypothesis is $\beta_i = 0$.

To illustrate this formula, refer to the test of the regression coefficient for the independent variable temperature. From the computer output on page 525, the regression coefficient for temperature is -4.583. The standard deviation of the sampling distribution of the regression coefficient for the independent variable temperature is 0.772. Inserting these values in formula (14–6):

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-4.583 - 0}{0.772} = -5.937$$

Applying the formula, the computed t ratio is -5.937 for temperature (the small difference between the computed value and that shown on the Excel output is due to rounding) and -3.119 for insulation. Both of these t-values are in the rejection region to the left of -2.120. Thus, we conclude that the regression coefficients for the temperature and insulation variables are *not* zero. The computed t for the age of the furnace is 1.521, so we conclude that could equal 0. The independent variable age of the furnace is not a significant predictor of heating cost. It can be dropped from the analysis.

We can also use p-values to test the individual regression coefficients. Again, these are commonly reported in computer software output. The computed t ratio for temperature on the Excel output is -5.934 and has a p-value of 0.000. Because the p-value is less than 0.05, the regression coefficient for the independent variable

temperature is not equal to zero and should be included in the equation to predict heating costs. For insulation, the t ratio is -3.119 and has a p-value of 0.007. As with temperature, the p-value is less than 0.05, so we conclude that the insulation regression coefficient is not equal to zero and should be included in the equation to predict heating cost. In contrast to temperature and insulation, the p-value to test the "age of the furnace" regression coefficient is 0.148. It is clearly greater than 0.05, so we conclude that the "age of furnace" regression coefficient could equal 0. Further, as an independent variable it is not a significant predictor of heating cost. Thus, age of furnace should not be included in the equation to predict heating costs.

At this point, we need to develop a strategy for deleting independent variables. In the Salsberry Realty case, there were three independent variables, and one (the age of the furnace) had a regression coefficient that did not differ from 0. It is clear that we should drop that variable and rerun the regression equation. Below is the Minitab output where heating cost is the dependent variable and outside temperature and amount of insulation are the independent variables.

Tbl14-1.mtw ***

↓	C1	C2	C3
	Cost	Temp	Insul
1	250	35	3
2	360	29	4
3	165	36	7
4	43	60	6
5	92	65	5
6	200	30	5
7	355	10	6
8	290	7	10
9	230	21	9
10	120	55	2
11	73	54	12
12	205	48	5
13	400	20	5
14	320	39	4
15	72	60	8
16	272	20	5
17	94	58	7
18	190	40	8

Session

Regression Analysis: Cost versus Temp, Insul

The regression equation is
Cost = 490 - 5.15 Temp - 14.7 Insul

Predictor	Coef	SE Coef	T	P
Constant	490.29	44.41	11.04	0.000
Temp	-5.1499	0.7019	-7.34	0.000
Insul	-14.718	4.934	-2.98	0.008

S = 52.9824 R-Sq = 77.6% R-Sq(adj) = 74.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	165195	82597	29.42	0.000
Residual Error	17	47721	2807		
Total	19	212916			

Source	DF	Seq SS
Temp	1	140215
Insul	1	24980

Summarizing the results from this new Minitab output:

1. The new regression equation is:

$$\hat{Y} = 490.29 - 5.1499X_1 - 14.718X_2$$

Notice that the regression coefficients for outside temperature (X_1) and amount of insulation (X_2) are similar to but not exactly the same as when we included the independent variable age of the furnace. Compare the above equation to that in the Excel output on page 525. Both of the regression coefficients are negative as in the earlier equation.

2. The details of the global test are as follows:

H_0: $\beta_1 = \beta_2 = 0$
H_1: Not all of the β_i's $= 0$

The F distribution is the test statistic and there are $k = 2$ degrees of freedom in the numerator and $n - (k + 1) = 20 - (2 + 1) = 17$ degrees of freedom in the denominator. Using the .05 significance level and Appendix B.4, the decision

rule is to reject H_0 if F is greater than 3.59. We compute the value of F as follows:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{165{,}195/2}{47{,}721/[20 - (2 + 1)]} = 29.42$$

Because the computed value of F (29.42) is greater than the critical value (3.59), the null hypothesis is rejected and the alternate accepted. We conclude that at least one of the regression coefficients is different from 0.

Using the *p*-value, the F test statistic (29.42) has a *p*-value (0.000) that is clearly less than 0.05. Therefore, we reject the null hypothesis and accept the alternate. We conclude that at least one of the regression coefficients is different from 0.

3. The next step is to conduct a test of the regression coefficients individually. We want to find out if one or both of the regression coefficients are different from 0. The null and alternate hypotheses for each of the independent variables are:

Outside Temperature	Insulation
H_0: $\beta_1 = 0$	H_0: $\beta_2 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_2 \neq 0$

The test statistic is the t distribution with $n - (k + 1) = 20 - (2 + 1) = 17$ degrees of freedom. Using the .05 significance level and Appendix B.2, the decision rule is to reject H_0 if the computed value of t is less than -2.110 or greater than 2.110.

Outside Temperature

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-5.1499 - 0}{0.7019} = -7.34$$

Insulation

$$t = \frac{b_2 - 0}{s_{b_2}} = \frac{-14.718 - 0}{4.934} = -2.98$$

In both tests, we reject H_0 and accept H_1. We conclude that each of the regression coefficients is different from 0. Both outside temperature and amount of insulation are useful variables in explaining the variation in heating costs.

Using *p*-values, the *p*-value for the temperature *t*-statistic is 0.000 and the *p*-value for the insulation t-statistic is 0.008. Both *p*-values are less than 0.05, so in both tests we reject the null hypothesis and conclude that each of the regression coefficients is different from 0. Both outside temperature and amount of insulation are useful variables in explaining the variation in heating costs.

In the heating cost example, it was clear which independent variable to delete. However, in some instances which variable to delete may not be as clear-cut. To explain, suppose we develop a multiple regression equation based on five independent variables. We conduct the global test and find that some of the regression coefficients are different from zero. Next, we test the regression coefficients individually and find that three are significant and two are not. The preferred procedure is to drop the single independent variable with the *smallest absolute* t *value* or *largest* p-*value* and rerun the regression equation with the four remaining variables, then, on the new regression equation with four independent variables, conduct the individual tests. If there are still regression coefficients that are not significant, again drop the variable with the smallest absolute *t* value or the largest, nonsignificant *p*-value. To describe the process in another way, we should delete only one variable at a time. Each time we delete a variable, we need to rerun the regression equation and check the remaining variables.

This process of selecting variables to include in a regression model can be automated, using Excel, Minitab, MegaStat, or other statistical software. Most of the software systems include methods to sequentially remove and/or add independent variables and at the same time provide estimates of the percentage of variation

explained (the *R*-square term). Two of the common methods are **stepwise regression** and **best subset regression.** It may take a long time, but in the extreme we could compute every regression between the dependent variable and any possible subset of the independent variables.

Unfortunately, on occasion, the software may work "too hard" to find an equation that fits all the quirks of your particular data set. The suggested equation may not represent the relationship in the population. A judgment is needed to choose among the equations presented. Consider whether the results are logical. They should have a simple interpretation and be consistent with your knowledge of the application under study.

Self-Review 14–3 The regression output about eating places in Myrtle Beach is repeated below (see earlier self-reviews).

Predictor	Coef	SE Coef	T	*p*-value
Constant	2.50	1.50	1.667	—
X_1	3.00	1.500	2.000	0.056
X_2	4.00	3.000	1.333	0.194
X_3	−3.00	0.20	−15.00	0.000
X_4	0.20	.05	4.00	0.000
X_5	1.00	1.50	0.667	0.511

Analysis of Variance

Source	DF	SS	MS	F	*p*-value
Regression	5	100	20	10	0.000
Residual Error	20	40	2		
Total	25	140			

(a) Perform a global test of hypothesis to check if any of the regression coefficients are different from 0. What do you decide? Use the .05 significance level.
(b) Do an individual test of each independent variable. Which variables would you consider eliminating? Use the .05 significance level.
(c) Outline a plan for possibly removing independent variables.

Exercises

connect™

7. Given the following regression output,

Predictor	Coef	SE Coef	T	P
Constant	84.998	1.863	45.61	0.000
X_1	2.391	1.200	1.99	0.051
X_2	−0.4086	0.1717	−2.38	0.020

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	77.907	38.954	4.14	0.021
Residual Error	62	583.693	9.414		
Total	64	661.600			

answer the following questions:

a. Write the regression equation.
b. If X_1 is 4 and X_2 is 11, what is the value of the dependent variable?
c. How large is the sample? How many independent variables are there?
d. Conduct a global test of hypothesis to see if any of the set of regression coefficients could be different from 0. Use the .05 significance level. What is your conclusion?
e. Conduct a test of hypothesis for each independent variable. Use the .05 significance level. Which variable would you consider eliminating?
f. Outline a strategy for deleting independent variables in this case.

8. The following regression output was obtained from a study of architectural firms. The dependent variable is the total amount of fees in millions of dollars.

Predictor	Coef	SE Coef	T	*p*-value
Constant	7.987	2.967	2.69	—
X_1	0.12242	0.03121	3.92	0.000
X_2	−0.12166	0.05353	−2.27	0.028
X_3	−0.06281	0.03901	−1.61	0.114
X_4	0.5235	0.1420	3.69	0.001
X_5	−0.06472	0.03999	−1.62	0.112

Analysis of Variance

Source	DF	SS	MS	F	*p*-value
Regression	5	3710.00	742.00	12.89	0.000
Residual Error	46	2647.38	57.55		
Total	51	6357.38			

X_1 is the number of architects employed by the company.
X_2 is the number of engineers employed by the company.
X_3 is the number of years involved with health care projects.
X_4 is the number of states in which the firm operates.
X_5 is the percent of the firm's work that is health care–related.

a. Write out the regression equation.
b. How large is the sample? How many independent variables are there?
c. Conduct a global test of hypothesis to see if any of the set of regression coefficients could be different from 0. Use the .05 significance level. What is your conclusion?
d. Conduct a test of hypothesis for each independent variable. Use the .05 significance level. Which variable would you consider eliminating first?
e. Outline a strategy for deleting independent variables in this case.

14.5 Evaluating the Assumptions of Multiple Regression

In the previous section, we described the methods to statistically evaluate the multiple regression equation. The results of the test let us know if at least one of the coefficients was not equal to zero and we described a procedure of evaluating each regression coefficient. We also discussed the decision-making process for including and excluding independent variables in the multiple regression equation.

It is important to know that the validity of the statistical global and individual tests rely on several assumptions. That is, if the assumptions are not true, the results might be biased or misleading. However, strict adherence to the following assumptions is not always possible. Fortunately, the statistical techniques discussed in this chapter work well even when one or more of the assumptions are violated. Even if the values in the multiple regression equation are "off" slightly, our estimates using a multiple regression equation will be closer than any that could be made otherwise. Usually the statistical procedures are robust enough to overcome violations of some assumptions.

In Chapter 13, we listed the necessary assumptions for regression when we considered only a single independent variable. (See Section 13.8 on page 490.) The assumptions for multiple regression are similar.

1. **There is a linear relationship.** That is, there is a straight-line relationship between the dependent variable and the set of independent variables.

2. **The variation in the residuals is the same for both large and small values of $\hat{Y}$.** To put it another way, $(Y - \hat{Y})$ is unrelated to whether $\hat{Y}$ is large or small.
3. **The residuals follow the normal probability distribution.** Recall the residual is the difference between the actual value of Y and the estimated value $\hat{Y}$. So the term $(Y - \hat{Y})$ is computed for every observation in the data set. These residuals should approximately follow a normal probability distribution. In addition, the mean of the residuals should be 0.
4. **The independent variables should not be correlated.** That is, we would like to select a set of independent variables that are not themselves correlated.
5. **The residuals are independent.** This means that successive observations of the dependent variable are not correlated. This assumption is often violated when time is involved with the sampled observations.

In this section, we present a brief discussion of each of these assumptions. In addition, we provide methods to validate these assumptions and indicate the consequences if these assumptions cannot be met. For those interested in additional discussion, Kutner, Nachtscheim, Neter, and Li, *Applied Linear Statistical Models,* 5th ed., (McGraw-Hill: 2005), is an excellent reference.

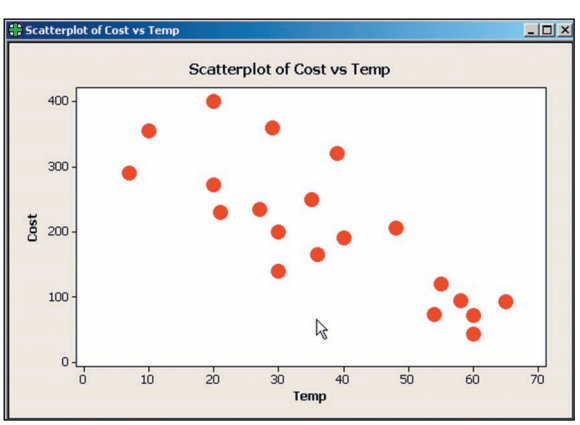
Linear Relationship

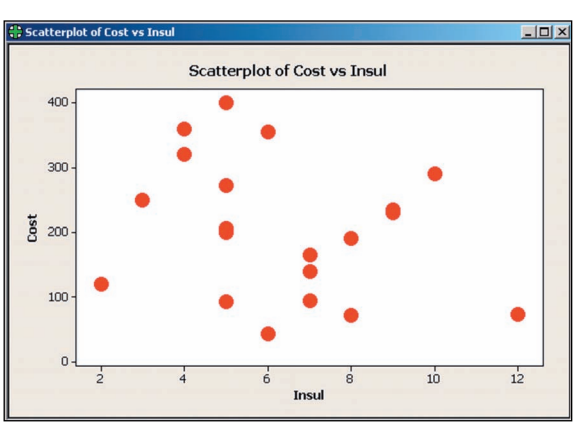
Let's begin with the linearity assumption. The idea is that the relationship between the set of independent variables and the dependent variable is linear. If we are considering two independent variables, we can visualize this assumption. The two independent variables and the dependent variable would form a three-dimensional space. The regression equation would then form a plane as shown on page 514. We can evaluate this assumption with scatter diagrams and residual plots.

LO6 Use residual analysis to evaluate the assumptions of multiple regression analysis.

Using Scatter Diagrams The evaluation of a multiple regression equation should always include a scatter diagram that plots the dependent variable against each independent variable. These graphs help us to visualize the relationships and provide some initial information about the direction (positive or negative), linearity, and strength of the relationship. For example, the scatter diagrams for the home heating example follow. The plots suggest a fairly strong negative, linear relationship between heating cost and temperature, and a negative relationship between heating cost and insulation.

Using Residual Plots Recall that a residual $(Y - \hat{Y})$ can be computed using the multiple regression equation for each observation in a data set. In Chapter 13, we



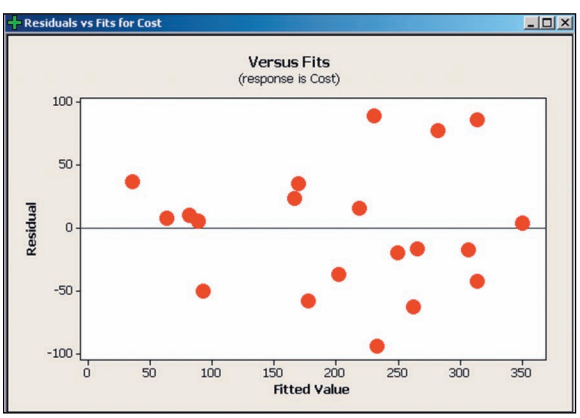


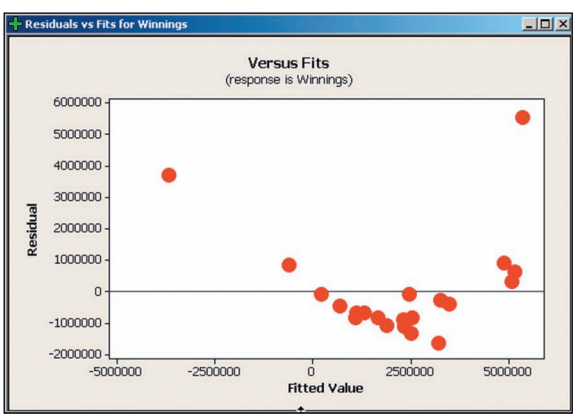
discussed the idea that the best regression line passed through the center of the data in a scatter plot. In this case, you would find a good number of the observations above the regression line (these residuals would have a positive sign), and a good number of the observations below the line (these residuals would have a negative sign). Further, the observations would be scattered above and below the line over the entire range of the independent variable.

The same concept is true for multiple regression, but we cannot graphically portray the multiple regression. However, plots of the residuals can help us evaluate the linearity of the multiple regression equation. To investigate, the residuals are plotted on the vertical axis against the predictor variable, $\hat{Y}$. The graph on the left below shows the residual plots for the home heating cost example. Notice the following:

- The residuals are plotted on the vertical axis and are centered around zero. There are both positive and negative residuals.
- The residual plots show a random distribution of positive and negative values across the entire range of the variable plotted on the horizontal axis.
- The points are scattered and there is no obvious pattern, so there is no reason to doubt the linearity assumption.

This plot supports the assumption of linearity.





If there is a pattern to the points in the scatter plot, further investigation is necessary. The points in the graph on the right above show nonrandom residuals. See that the residual plot does *not* show a random distribution of positive and negative values across the entire range of the variable plotted on the horizontal axis. In fact, the graph shows a curvature to the residual plots. This indicates the relationship may not be linear. In this case, we would evaluate different transformations of the equation as discussed in Chapter 13.

Variation in Residuals Same for Large and Small $\hat{Y}$ Values

This requirement indicates that the variation about the predicted values is constant, regardless of whether the predicted values are large or small. To cite a specific example, which may violate the assumption, suppose we use the single independent variable age to explain the variation in income. We suspect that as age increases so does salary, but it also seems reasonable that as age increases there may be more variation around the regression line. That is, there will likely be more variation in income for a 50-year-old person than for a 35-year-old

person. The requirement for constant variation around the regression line is called **homoscedasticity.**

HOMOSCEDASTICITY The variation around the regression equation is the same for all of the values of the independent variables.

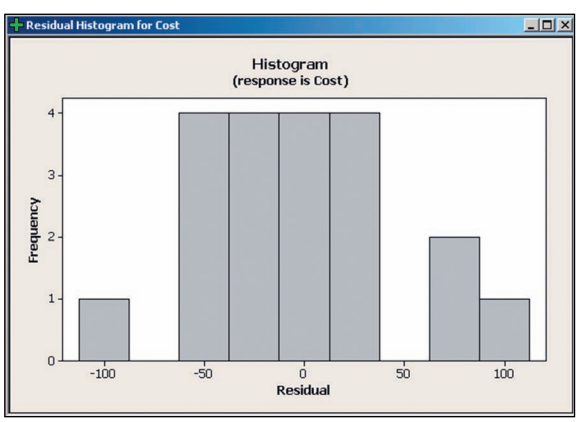
To check for homoscedasticity the residuals are plotted against the fitted values of *Y*. This is the same graph that we used to evaluate the assumption of linearity. (See page 533.) Based on the scatter diagram in that software output, it is reasonable to conclude that this assumption has not been violated.

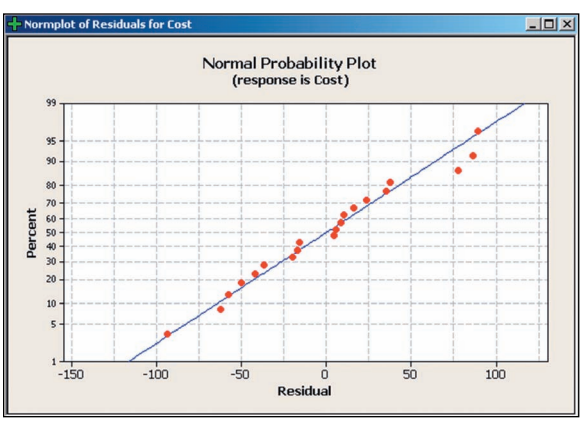
Distribution of Residuals

To be sure that the inferences we make in the global and individual hypotheses tests are valid, we evaluate the distribution of residuals. Ideally, the residuals should follow a normal probability distribution.

To evaluate this assumption, we can organize the residuals into a frequency distribution. The Minitab histogram of the residuals is shown following on the left for the home heating cost example. Although it is difficult to show that the residuals follow a normal distribution with only 20 observations, it does appear the normality assumption is reasonable.

Both Minitab and Excel offer another graph that helps to evaluate the assumption of normally distributed residuals. It is a called a **normal probability plot** and is shown to the right of the histogram. We describe this graph further in Section 17.6 starting on page 663. Basically, the normal probability plot supports the assumption of normally distributed residuals if the plotted points are fairly close to a straight line drawn from the lower left to the upper right of the graph.





In this case, both graphs support the assumption that the residuals follow the normal probability distribution. Therefore, the inferences that we made based on the global and individual hypothesis tests are supported with the results of this evaluation.

Multicollinearity

LO7 Evaluate the effects of correlated independent variables.

Multicollinearity exists when independent variables are correlated. Correlated independent variables make it difficult to make inferences about the individual regression coefficients and their individual effects on the dependent variable. In practice, it is

nearly impossible to select variables that are completely unrelated. To put it another way, it is nearly impossible to create a set of independent variables that are not correlated to some degree. However, a general understanding of the issue of multicollinearity is important.

First, we should point out that multicollinearity does not affect a multiple regression equation's ability to predict the dependent variable. However, when we are interested in evaluating the relationship between each independent variable and the dependent variable, multicollinearity may show unexpected results.

For example, if we use two highly multicollinear variables, high school GPA and high school class rank, to predict the GPA of incoming college freshmen (dependent variable), we would expect that both independent variables would be positively related to the dependent variable. However, because the independent variables are highly correlated, one of the independent variables may have an unexpected and inexplicable negative sign. In essence, these two independent variables are redundant in that they explain the same variation in the dependent variable.

A second reason for avoiding correlated independent variables is they may lead to erroneous results in the hypothesis tests for the individual independent variables. This is due to the instability of the standard error of estimate. Several clues that indicate problems with multicollinearity include the following:

1. An independent variable known to be an important predictor ends up having a regression coefficient that is not significant.
2. A regression coefficient that should have a positive sign turns out to be negative, or vice versa.
3. When an independent variable is added or removed, there is a drastic change in the values of the remaining regression coefficients.

In our evaluation of a multiple regression equation, an approach to reducing the effects of multicollinearity is to carefully select the independent variables that are included in the regression equation. A general rule is if the correlation between two independent variables is between -0.70 and 0.70, there likely is not a problem using both of the independent variables. A more precise test is to use the **variance inflation factor.** It is usually written *VIF*. The value of *VIF* is found as follows:

VARIANCE INFLATION FACTOR

$$VIF = \frac{1}{1 - R_j^2} \qquad \textbf{[14–7]}$$

The term R_j^2 refers to the coefficient of determination, where the selected *independent variable* is used as a dependent variable and the remaining independent variables are used as independent variables. A *VIF* greater than 10 is considered unsatisfactory, indicating that the independent variable should be removed from the analysis. The following example will explain the details of finding the *VIF*.

Example

Refer to the data in Table 14–1, which relates the heating cost to the independent variables outside temperature, amount of insulation, and age of furnace. Develop a correlation matrix for all the independent variables. Does it appear there is a problem with multicollinearity? Find and interpret the variance inflation factor for each of the independent variables.

Solution

We begin by using the Minitab system to find the correlation matrix for the dependent variable and the three independent variables. A portion of that output follows:

```
             Cost      Temp    Insul
Temp      −0.812
Insul     −0.257    −0.103
Age        0.537    −0.486    0.064

Cell Contents: Pearson correlation
```

The highlighted area indicates the correlation among the independent variables. None of the correlations among the independent variables exceed −.70 or .70, so we do not suspect problems with multicollinearity. The largest correlation among the independent variables is −0.486 between age and temperature.

To confirm this conclusion, we compute the *VIF* for each of the three independent variables. We will consider the independent variable temperature first. We use Minitab to find the multiple coefficient of determination with temperature as the *dependent variable* and amount of insulation and age of the furnace as independent variables. The relevant Minitab output follows.

```
Regression Analysis: Temp versus Insul, Age

The regression equation is
Temp = 58.0 − 0.51 Insul − 2.51 Age

Predictor       Coef    SE Coef        T        P
Constant       57.99      12.35     4.70    0.000
Insul         −0.509      1.488    −0.34    0.737
Age           −2.509      1.103    −2.27    0.036

S = 16.0311   R-Sq = 24.1%   R-Sq(adj) = 15.2%

Analysis of Variance
Source            DF       SS       MS       F        P
Regression         2    1390.3    695.1    2.70    0.096
Residual Error    17    4368.9    257.0
Total             19    5759.2
```

The coefficient of determination is .241, so inserting this value into the *VIF* formula:

$$VIF = \frac{1}{1 - R_1^2} = \frac{1}{1 - .241} = 1.32$$

The *VIF* value of 1.32 is less than the upper limit of 10. This indicates that the independent variable temperature is not strongly correlated with the other independent variables.

Again, to find the *VIF* for insulation we would develop a regression equation with insulation as the *dependent variable* and temperature and age of furnace as independent variables. For this equation, we would determine the coefficient of determination. This would be the value for R_2^2. We would substitute this value in equation 14–7, and solve for *VIF*.

Fortunately, Minitab will generate the *VIF* values for each of the independent variables. These values are reported in the right-hand column under the heading *VIF* in the following Minitab output. All these values are less than 10. Hence, we conclude there is not a problem with multicollinearity in this example.

```
The regression equation is
Cost = 427 − 4.58 Temp − 14.8 Insul + 6.10 Age

Predictor       Coef    SE Coef        T        P      VIF
Constant      427.19      59.60     7.17    0.000
Temp        −4.5827     0.7723    −5.93    0.000    1.318
Insul       −14.831      4.754    −3.12    0.007    1.011
Age          −6.101      4.012     1.52    0.148    1.310
```