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 $X_1, X_2, \ldots$, 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:

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

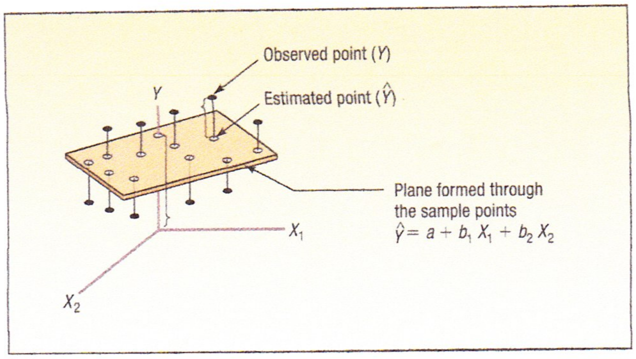

Observed point (Y)
Estimated point ($\hat{Y}$)
Y
Plane formed through the sample points
$\hat{Y} = a + b_1 X_1 + b_2 X_2$
X_1
X_2

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.

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

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.

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:

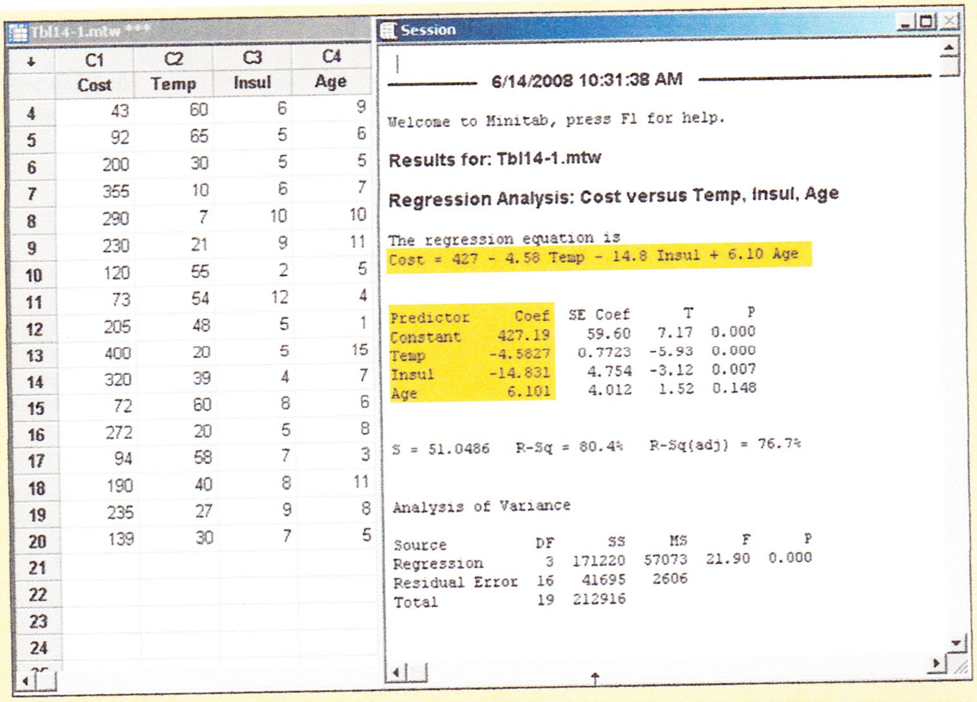
- The mean outside temperature in January, represented by X_1.
- The number of inches of insulation in the attic, represented by X_2.
- The age in years of the furnace, represented by X_3.

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

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

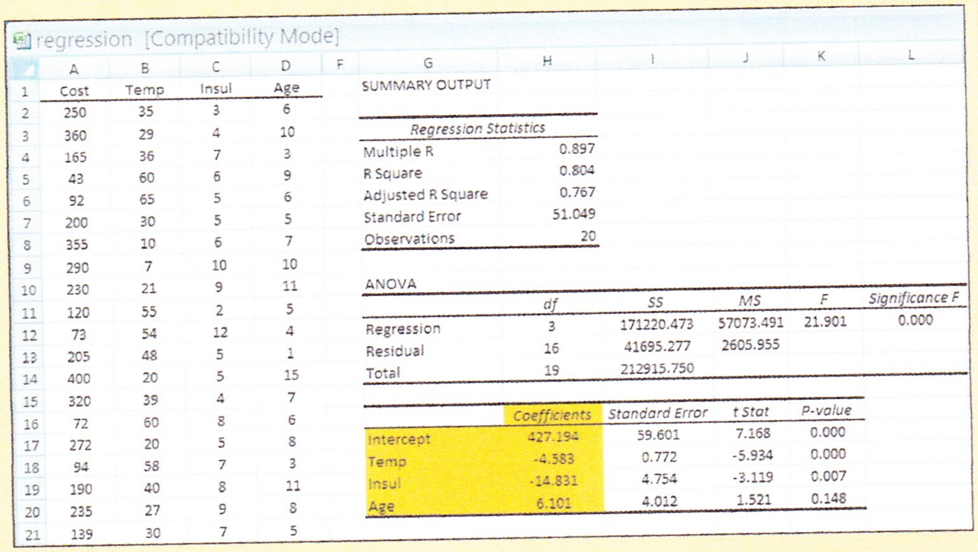
Now that we have defined the regression equation, we are ready to use either Excel or Minitab to compute all the statistics needed for the analysis. The outputs from the two software systems are shown below.

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.



↓	C1 Cost	C2 Temp	C3 Insul	C4 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				

```
———————— 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
1	Cost	Temp	Insul	Age
2	250	35	3	6
3	360	29	4	10
4	165	36	7	3
5	43	60	6	9
6	92	65	5	6
7	200	30	5	5
8	355	10	6	7
9	290	7	10	10
10	230	21	9	11
11	120	55	2	5
12	73	54	12	4
13	205	48	5	1
14	400	20	5	15
15	320	39	4	7
16	72	60	8	6
17	272	20	5	8
18	94	58	7	3
19	190	40	8	11
20	235	27	9	8
21	139	30	7	5

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.897
R Square	0.804
Adjusted R Square	0.767
Standard Error	51.049
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	171220.473	57073.491	21.901	0.000
Residual	16	41695.277	2605.955		
Total	19	212915.750			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.194	59.601	7.168	0.000
Temp	-4.583	0.772	-5.934	0.000
Insul	-14.831	4.754	-3.119	0.007
Age	6.101	4.012	1.521	0.148

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
X1            3.00     1.500     2.000
X2            4.00     3.000     1.333
X3           -3.00     0.20    -15.00
X4            0.20      .05      4.00
X5            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

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:

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

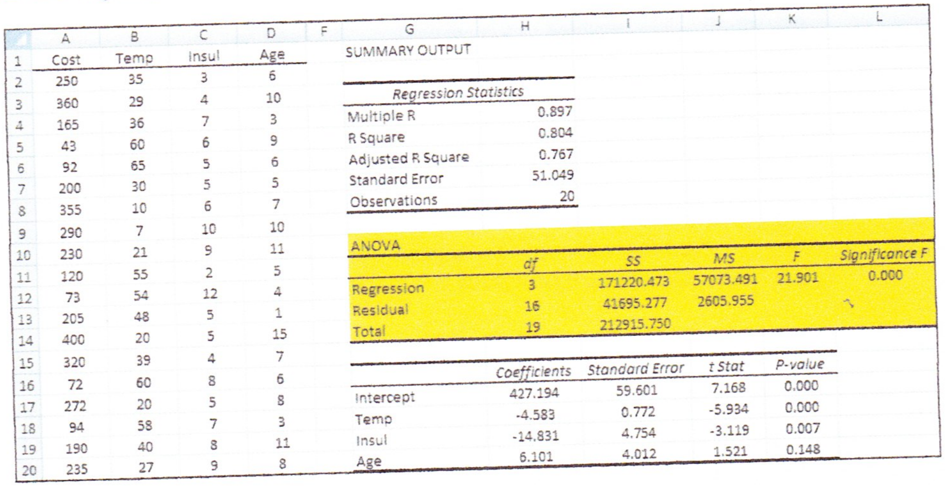
As we have seen, the total sum of squares is the sum of the regression and residual sum of squares. The regression sum of squares is the sum of the squared differences between the estimated or predicted values, $\hat{Y}$, and the overall mean of Y. The regression sum of squares is found by:

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

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
1	Cost	Temp	Insul	Age
2	250	35	3	6
3	360	29	4	10
4	165	36	7	3
5	43	60	6	9
6	92	65	5	6
7	200	30	5	5
8	355	10	6	7
9	290	7	10	10
10	230	21	9	11
11	120	55	2	5
12	73	54	12	4
13	205	48	5	1
14	400	20	5	15
15	320	39	4	7
16	72	60	8	6
17	272	20	5	8
18	94	58	7	3
19	190	40	8	11
20	235	27	9	8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.897
R Square	0.804
Adjusted R Square	0.767
Standard Error	51.049
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	171220.473	57073.491	21.901	0.000
Residual	16	41695.277	2605.955		
Total	19	212915.750			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.194	59.601	7.168	0.000
Temp	-4.583	0.772	-5.934	0.000
Insul	-14.831	4.754	-3.119	0.007
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

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

Chapter 14 Answers to Self-Review

14–1 a. \$389,500 or 389.5 (in \$000); found by
$2.5 + 3(40) + 4(72) - 3(10) + .2(20) + 1(5)$
$= 3895$

b. The b_2 of 4 shows profit will go up \$4,000 for each extra hour the restaurant is open (if none of the other variables change). The b_3 of -3 implies profit will fall \$3,000 for each added mile away from the central area (if none of the other variables change).

14–2 a. The total degrees of freedom ($n - 1$) is 25. So the sample size is 26.

b. There are 5 independent variables.

c. There is only 1 dependent variable (profit).

d. $S_{Y.12345} = 1.414$, found by $\sqrt{2}$. Ninety-five percent of the residuals will be between -2.828 and 2.828, found by $\pm 2(1.414)$.

e. $R^2 = .714$, found by 100/140. 71.4% of the deviation in profit is accounted for by these five variables.

f. $R^2_{\text{adj}} = .643$, found by

$$1 - \left[\frac{40}{(26 - (5 + 1))}\right] \Big/ \left[\frac{140}{(26 - 1)}\right]$$

14–3 a. H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
H_1: Not all of the β's are 0.

The decision rule is to reject H_0 if $F > 2.71$. The computed value of F is 10, found by 20/2. So, you reject H_0, which indicates at least one of the regression coefficients is different from zero.

Based on p-values, the decision rule is to reject the null hypothesis if the p-value is less than 0.05. The computed value of F is 10, found by 20/2, and has a p-value of 0.000. Thus, we reject the null hypothesis, which indicates that at least one of the regression coefficients is different from zero.

b. For variable 1: H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$
The decision rule is: Reject H_0 if $t < -2.086$ or $t > 2.086$. Since 2.000 does not go beyond either of those limits, we fail to reject the null hypothesis. This regression coefficient could be zero. We can consider dropping this variable. By parallel logic the null hypothesis is rejected for variables 3 and 4.

For variable 1, the decision rule is to reject H_0: $\beta_1 = 0$ if the p-value is less than 0.05. Because the p-value is 0.056, we cannot reject the null hypothesis. This regression coefficient could be zero. Therefore, we can consider dropping this variable. By parallel logic, we reject the null hypothesis for variables 3 and 4.

c. We should consider dropping variables 1, 2, and 5. Variable 5 has the smallest absolute value of t or largest p-value. So delete it first and refigure the regression analysis.

14–4 a. $\hat{Y} = 15.7625 + 0.4415X_1 + 3.8598X_2$
$\hat{Y} = 15.7625 + 0.4415(30) + 3.8598(1)$
$= 32.87$

b. Female agents make \$3,860 more than male agents.

c. H_0: $\beta_3 = 0$
H_1: $\beta_3 \neq 0$
$df = 17$, reject H_0 if $t < -2.110$ or $t > 2.110$

$$t = \frac{3.8598 - 0}{1.4724} = 2.621$$

The t-statistic exceeds the critical value of 2.110. Also, the p-value $= 0.0179$ and is less than 0.05. Reject H_0. Gender should be included in the regression equation.

A Review of Chapters 13 and 14

Simple regression and correlation examine the relationship between two variables.

This section is a review of the major concepts and terms introduced in Chapters 13 and 14. Chapter 13 noted that the strength of the relationship between the independent variable and the dependent variable can be measured by the *correlation coefficient.* The correlation coefficient is designated by the letter r. It can assume any value between -1.00 and $+1.00$ inclusive. Coefficients of -1.00 and $+1.00$ indicate a perfect relationship, and 0 indicates no relationship. A value near 0, such as $-.14$ or .14, indicates a weak relationship. A value near -1 or $+1$, such as $-.90$ or $+.90$, indicates a strong relationship. Squaring r gives the *coefficient of determination,* also called r^2. It indicates the proportion of the total variation in the dependent variable explained by the independent variable.

Likewise, the strength of the relationship between several independent variables and a dependent variable is measured by the *coefficient of multiple determination,* R^2. It measures the proportion of the variation in Y explained by two or more independent variables.