# 14

# Multiple Regression and Correlation Analysis

A mortgage department of a large bank is studying its recent loans. A random sample of 25 recent loans is obtained, searching for factors such as the value of the home, education level of borrower, age, monthly mortgage payment and gender relate to the family income. Are these variables effective predictors of the income of the household? (See Exercise 26 and Goal 1.)

## GOALS

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

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

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

3 Compute and interpret the *multiple standard error of estimate,* the *coefficient of multiple determination,* and the *adjusted coefficient of multiple determination.*

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

5 Conduct a test of hypothesis on each of the regression coefficients.

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

7 Evaluate the effects of correlated independent variables.

8 Use and understand qualitative independent variables.

9 Understand and interpret the *stepwise regression method.*

10 Understand and interpret possible interaction among independent variables.

## Introduction

In Chapter 13 we described the relationship between a pair of interval- or ratio-scaled variables. We began the chapter by studying the coefficient of correlation, 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.

## 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:

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

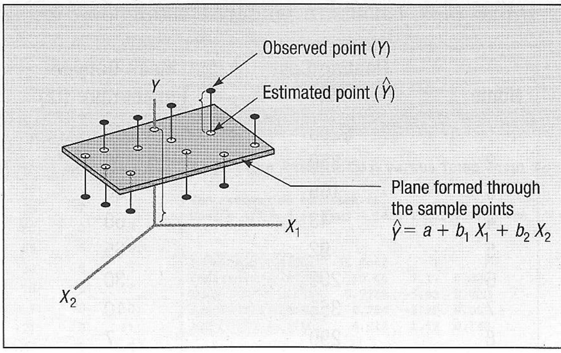

Observed point ($Y$)
$Y$
Estimated point ($\hat{Y}$)
Plane formed through the sample points
$\hat{Y} = a + b_1 X_1 + b_2 X_2$
$X_1$
$X_2$

**CHART 14–1** Regression Plane with Ten 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. 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 using 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

**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 |

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.

The data in Table 14–1 is available in both Excel and MINITAB formats on the student CD. 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:

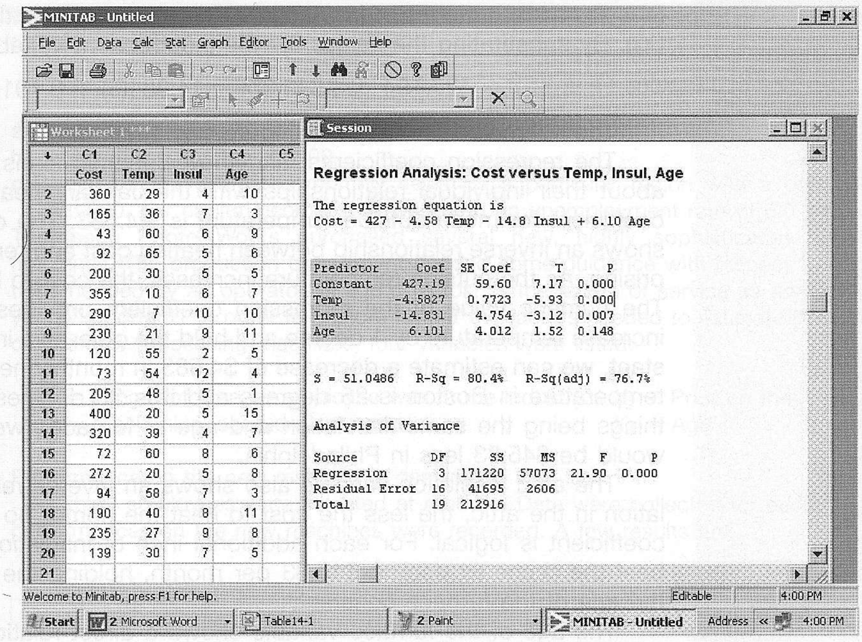
- The mean outside temperature in January, represented by $X_1$.
- The number of inches of insulation in the attic, represented by $X_2$.
- The age in years of the furnace, represented by $X_3$.

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

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

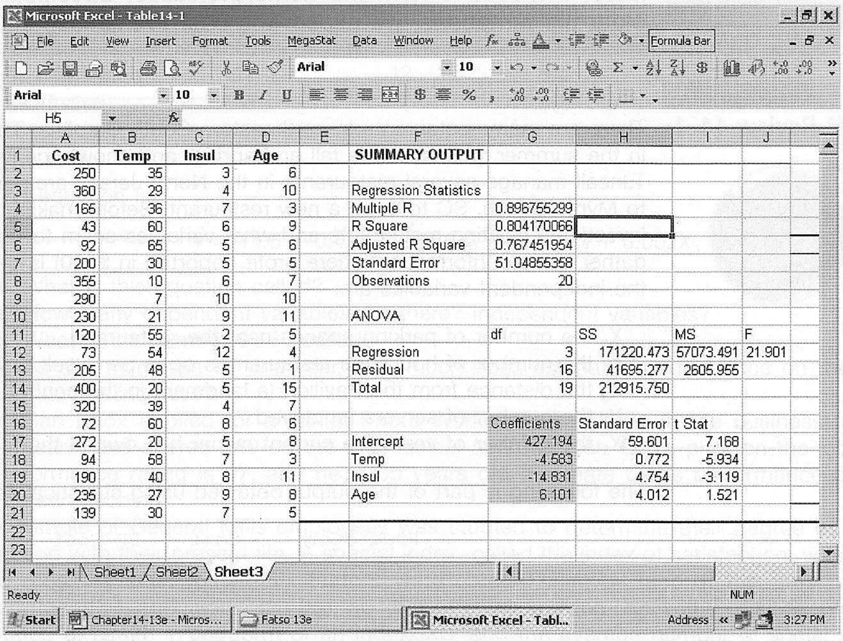
Now that we have defined the regression equation, we are ready to use either Excel or MINITAB to compute all the statistics needed for the analysis. The outputs from the two software systems are shown below.

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



| ↓ | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| | Cost | Temp | Insul | Age | |
| 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 | |
| 21 | | | | | |

```
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
```


Microsoft Excel - Table14-1
File Edit View Insert Format Tools MegaStat Data Window Help
Formula Bar
Arial
10
H5
Sheet1
Sheet2
Sheet3
Ready
NUM
Start
Chapter14-13e - Micros...
Fatso 13e
Microsoft Excel - Tabl...
Address
3:27 PM

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 0.896755299 | | | |
| 5 | 43 | 60 | 6 | 9 | | R Square | 0.804170066 | | | |
| 6 | 92 | 65 | 5 | 6 | | Adjusted R Square | 0.767451954 | | | |
| 7 | 200 | 30 | 5 | 5 | | Standard Error | 51.04855358 | | | |
| 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 |
| 12 | 73 | 54 | 12 | 4 | | Regression | 3 | 171220.473 | 57073.491 | 21.901 |
| 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 | |
| 17 | 272 | 20 | 5 | 8 | | Intercept | 427.194 | 59.601 | 7.168 | |
| 18 | 94 | 58 | 7 | 3 | | Temp | -4.583 | 0.772 | -5.934 | |
| 19 | 190 | 40 | 8 | 11 | | Insul | -14.831 | 4.754 | -3.119 | |
| 20 | 235 | 27 | 9 | 8 | | Age | 6.101 | 4.012 | 1.521 | |
| 21 | 139 | 30 | 7 | 5 | | | | | | |
| 22 | | | | | | | | | | |
| 23 | | | | | | | | | | |

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.

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:

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

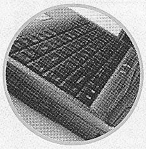
The estimated January heating cost is $276.56.

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 the Pavilion (a landmark in the central area) 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 and that is open 72 hours per week, is 10 miles from the Pavilion, 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. $X_3$ = Prior on-the-job rating.
$X_2$ = Mechanical aptitude test score. $X_4$ = 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 present 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?

# How Well Does the Equation Fit the Data?

Once you have the multiple regression equation, it is natural to ask "how well does the equation fit the data?" In linear regression, discussed in the previous chapter, you used summary statistics such as the standard error of estimate and the coefficient of determination to describe how effectively a single independent variable explained the variation of the dependent variable. The same procedures, broadened to additional independent variables, are used in multiple regression.

## Multiple Standard Error of Estimate

We begin with the **multiple standard error of estimate.** Recall that the standard error of estimate is comparable to the standard deviation. The standard deviation uses squared deviations from the mean, $(Y - \overline{Y})^2$, whereas the standard error of estimate utilizes squared deviations from the regression line, $(Y - \hat{Y})^2$. 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 514. 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 item sampled. Our next step is to square this difference, that is find $(Y - \hat{Y})^2 = (250 - 258.90)^2 = (-8.90)^2 = 79.21$. We repeat these operations for the other 19 observations and total these squared values. This value is the numerator of the multiple standard error of estimate. The denominator is the degrees of freedom, that is $n - (k + 1)$. The formula for the standard error is:

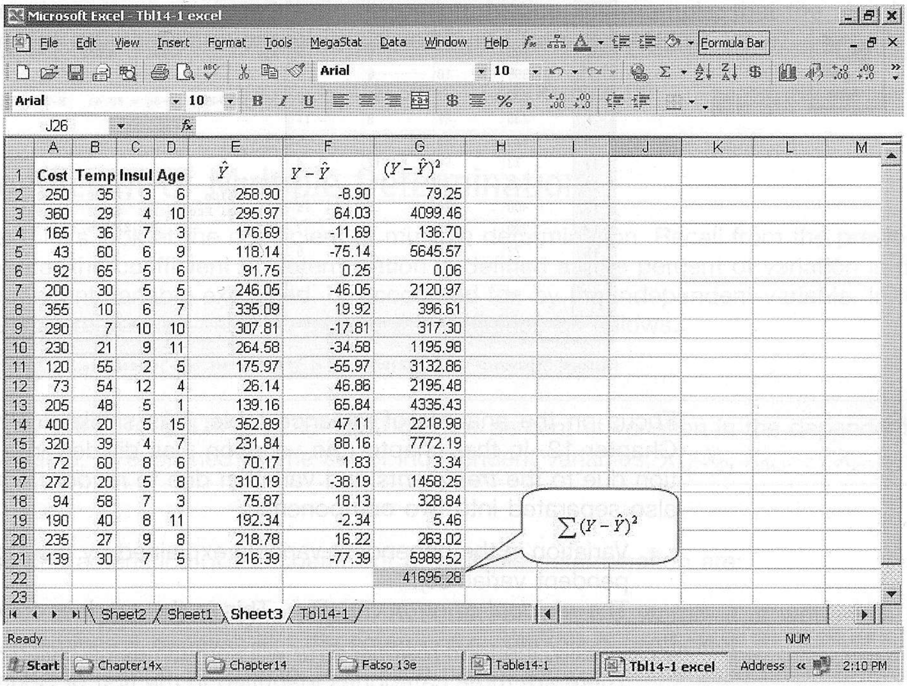
**MULTIPLE STANDARD ERROR OF ESTIMATE**

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

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.

In this example $n = 20$ and $k = 3$ (three independent variables) and we use the Excel software system to find the term $\Sigma(Y - \hat{Y})^2$. Note: There are small discrepancies due to rounding.


Microsoft Excel - Tbl14-1 excel
File Edit View Insert Format Tools MegaStat Data Window Help
Formula Bar
Arial
10
J26
Sheet2 Sheet1 Sheet3 Tbl14-1
Ready
NUM
Start Chapter14x Chapter14 Fatso 13e Table14-1 Tbl14-1 excel
Address
2:10 PM

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | **Cost** | **Temp** | **Insul** | **Age** | $\hat{Y}$ | $Y - \hat{Y}$ | $(Y - \hat{Y})^2$ |
| 2 | 250 | 35 | 3 | 6 | 258.90 | -8.90 | 79.25 |
| 3 | 360 | 29 | 4 | 10 | 295.97 | 64.03 | 4099.46 |
| 4 | 165 | 36 | 7 | 3 | 176.69 | -11.69 | 136.70 |
| 5 | 43 | 60 | 6 | 9 | 118.14 | -75.14 | 5645.57 |
| 6 | 92 | 65 | 5 | 6 | 91.75 | 0.25 | 0.06 |
| 7 | 200 | 30 | 5 | 5 | 246.05 | -46.05 | 2120.97 |
| 8 | 355 | 10 | 6 | 7 | 335.09 | 19.92 | 396.61 |
| 9 | 290 | 7 | 10 | 10 | 307.81 | -17.81 | 317.30 |
| 10 | 230 | 21 | 9 | 11 | 264.58 | -34.58 | 1195.98 |
| 11 | 120 | 55 | 2 | 5 | 175.97 | -55.97 | 3132.86 |
| 12 | 73 | 54 | 12 | 4 | 26.14 | 46.86 | 2195.48 |
| 13 | 205 | 48 | 5 | 1 | 139.16 | 65.84 | 4335.43 |
| 14 | 400 | 20 | 5 | 15 | 352.89 | 47.11 | 2218.98 |
| 15 | 320 | 39 | 4 | 7 | 231.84 | 88.16 | 7772.19 |
| 16 | 72 | 60 | 8 | 6 | 70.17 | 1.83 | 3.34 |
| 17 | 272 | 20 | 5 | 8 | 310.19 | -38.19 | 1458.25 |
| 18 | 94 | 58 | 7 | 3 | 75.87 | 18.13 | 328.84 |
| 19 | 190 | 40 | 8 | 11 | 192.34 | -2.34 | 5.46 |
| 20 | 235 | 27 | 9 | 8 | 218.78 | 16.22 | 263.02 |
| 21 | 139 | 30 | 7 | 5 | 216.39 | -77.39 | 5989.52 |
| 22 | | | | | | | 41695.28 |
| 23 | | | | | | | |

$\sum(Y - \hat{Y})^2$

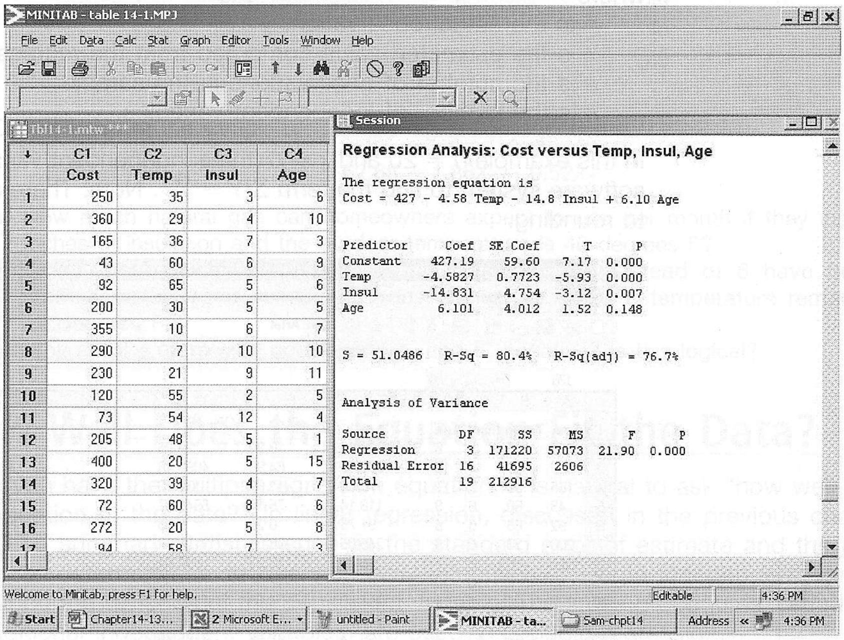
Since we have 3 independent variables, we identify the multiple standard error as $s_{Y.123}$. The subscripts indicate that three independent variables are being used to estimate $Y$.

$$s_{Y.123} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - (k + 1)}} = \sqrt{\frac{41{,}695.28}{20 - (3 + 1)}} = 51.05$$

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) = ±\$102.10. Refer to column F of the Excel output, headed $Y - \hat{Y}$. Of the 20 values in this column, 14 (or 70 percent) are less than ±\$51.05 and all are within ±\$102.10, which is very close to the guidelines of 68 percent and 95 percent.

## The ANOVA Table

As we said before, the multiple regression computations are long. Luckily, many statistical software systems do the calculations. Most of them report the results in a standard format. The outputs from Excel and MINITAB on page 515 are typical. In particular, they include an analysis of variance (ANOVA) table. The output from MINITAB is repeated here.



| ↓ | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| | Cost | Temp | Insul | Age |
| 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 |

**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
```

Focus on the analysis of variance table. It is similar to the ANOVA table used in Chapter 12. In that chapter the variation was divided into two components: variation due to the *treatments* and variation due to *random error.* Here total variation is also separated into two components:

- Variation in the dependent variable explained by the *regression* model (the independent variables).
- The *residual or error* variation. This is the random error due to sampling.

Incidentally, the term *residual error* will sometimes be called *random error* or just *error.*

There are three categories identified in the first or Source column in the ANOVA table; namely, the regression or explained variation, the residual or unexplained variation, and the total variation.

The second column is labeled *df* in the ANOVA table. It is the degrees of freedom. The degrees of freedom in the "Regression" row is the number of independent variables. We let $k$ represent the number of independent variables, so $k = 3$. The degrees of freedom in the "Error" is $n - (k + 1) = 20 - (3 + 1) = 16$. In this example, there are 20 observations so $n = 20$. The total degrees of freedom is $n - 1 = 20 - 1 = 19$.

The heading SS in the third column of the ANOVA table is the sum of squares or the variation.

$$\text{Total variation} = \text{SS total} = \Sigma(\hat{Y} - \overline{Y})^2 = 212{,}916$$

$$\text{Residual error or error variance} = \text{SSE} = \Sigma(Y - \hat{Y})^2 = 41{,}695$$

$$\text{Regression variation} = \text{SSR} = \Sigma(\hat{Y} - \overline{Y})^2 = \text{SS total} - \text{SSE}$$

$$= 212{,}916 - 41{,}695 = 171{,}220$$

(There is a small "round off" difference of one unit, which will have no effect on later calculations.)

The fourth column heading, MS or mean square, is obtained by dividing the SS quantity by the matching *df*. Thus MSR, the mean square regression, is equal to SSR/*k*. Similarly, MSE, the mean square error, is SSE/($n - (k + 1)$).

The following ANOVA table summarizes the process.

| **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 | | |

Each value in the ANOVA table plays an important role in the evaluation and interpretation of a multiple regression equation. Notice, for example, that the standard error of estimate can very easily be computed from the ANOVA table.

$$s_{Y.123} = \sqrt{\text{MSE}} = \sqrt{2{,}606} = 51.05$$

## 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, $X_1, X_2, X_3, \ldots X_k$.

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 MINITAB output, and use the regression sum of squares, SSR, then divide by the total sum of squares, SS total.

**COEFFICIENT OF MULTIPLE DETERMINATION**

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

The ANOVA portion of the MINITAB output in the heating cost example is repeated below.

```
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
```

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 say 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**

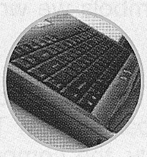
$$R^2_{adj} = 1 - \frac{\dfrac{\text{SSE}}{n - (k + 1)}}{\dfrac{\text{SS total}}{n - 1}} \qquad \textbf{[14–4]}$$

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:

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

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.

# 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 $\alpha$ and $\beta_j$. For example, the sample regression coefficient $b_2$ is a point estimate of the population parameter $\beta_2$. 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

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 $\beta_1$, $\beta_2$, and $\beta_3$. We now test whether the net 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:

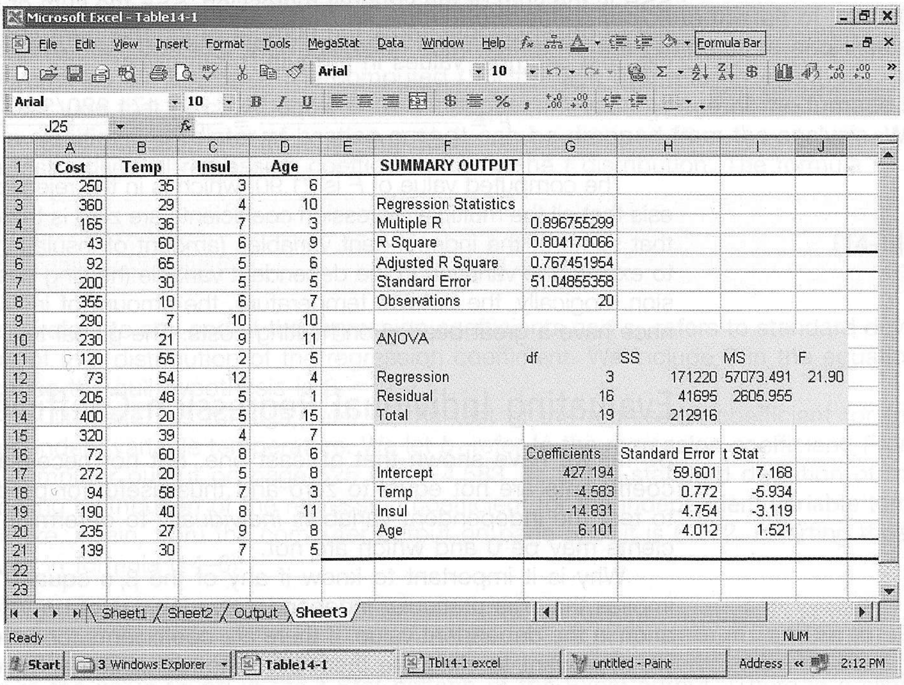
$H_1$: Not all the $\beta_i$'s are 0.

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.

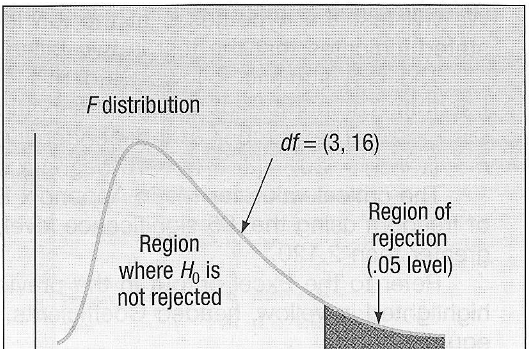
The degrees of freedom for the numerator and the denominator may be found in the Excel ANOVA table that follows. The ANOVA output is highlighted in light green. The top number in the column marked "df" is 3, indicating there are 3 degrees of freedom in the numerator. This value corresponds to the number of independent variables. The middle number in the "df" column (16) indicated there are 16 degrees of freedom in the denominator. The number 16 is found by $(n - (k + 1)) = 20 - (3 + 1) = 16$.



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 0.896755299 | | | |
| 5 | 43 | 60 | 6 | 9 | | R Square | 0.804170066 | | | |
| 6 | 92 | 65 | 5 | 6 | | Adjusted R Square | 0.767451954 | | | |
| 7 | 200 | 30 | 5 | 5 | | Standard Error | 51.04855358 | | | |
| 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 |
| 12 | 73 | 54 | 12 | 4 | | Regression | 3 | 171220 | 57073.491 | 21.90 |
| 13 | 205 | 48 | 5 | 1 | | Residual | 16 | 41695 | 2605.955 | |
| 14 | 400 | 20 | 5 | 15 | | Total | 19 | 212916 | | |
| 15 | 320 | 39 | 4 | 7 | | | | | | |
| 16 | 72 | 60 | 8 | 6 | | | Coefficients | Standard Error | t Stat | |
| 17 | 272 | 20 | 5 | 8 | | Intercept | 427.194 | 59.601 | 7.168 | |
| 18 | 94 | 58 | 7 | 3 | | Temp | -4.583 | 0.772 | -5.934 | |
| 19 | 190 | 40 | 8 | 11 | | Insul | -14.831 | 4.754 | -3.119 | |
| 20 | 235 | 27 | 9 | 8 | | Age | 6.101 | 4.012 | 1.521 | |
| 21 | 139 | 30 | 7 | 5 | | | | | | |
| 22 | | | | | | | | | | |
| 23 | | | | | | | | | | |

Sheet1 / Sheet2 / Output / Sheet3

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 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 value of $F$ is found from the following equation.

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

SSR is the sum of the squares regression, SSE the sum of squares error, $n$ the number of observations, and $k$ the number of independent variables. Inserting the heating cost example values in formula (14–5) gives:

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{171{,}220/3}{41{,}695/[20 - (3 + 1)]} = 21.90$$

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 some of the independent variables (amount of insulation, etc.) do have the ability to explain the variation in the dependent variable (heating cost). We expected this decision. Logically, the outside temperature, the amount of insulation, and age of the furnace have a great bearing on heating costs. The global test assures us that they do.

## Evaluating Individual Regression Coefficients

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 $\beta_i$'s equal 0? If a $\beta$ 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. The way the alternate hypothesis is stated indicates that the test is two-tailed.

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 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 on the Excel output labeled Standard Error indicates the standard error of the sample regression coefficient. Recall that Salsberry Realty selected a sample of 20 homes along the east coast of the United States. If it was to select a second random sample and compute the regression coefficients of that sample, the values would not be exactly the same. If it repeated the sampling process many times, however, we could design a sampling distribution of these regression coefficients. The column labeled Standard Error estimates the variability of these regression coefficients. The sampling distribution of Coefficients/Standard Error follows 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 computed $t$ ratio is $-5.934$ for temperature 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 $b_3$ 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 test individual regression coefficients using the $t$ distribution. The formula is:

**TESTING INDIVIDUAL REGRESSION COEFFICIENTS**

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

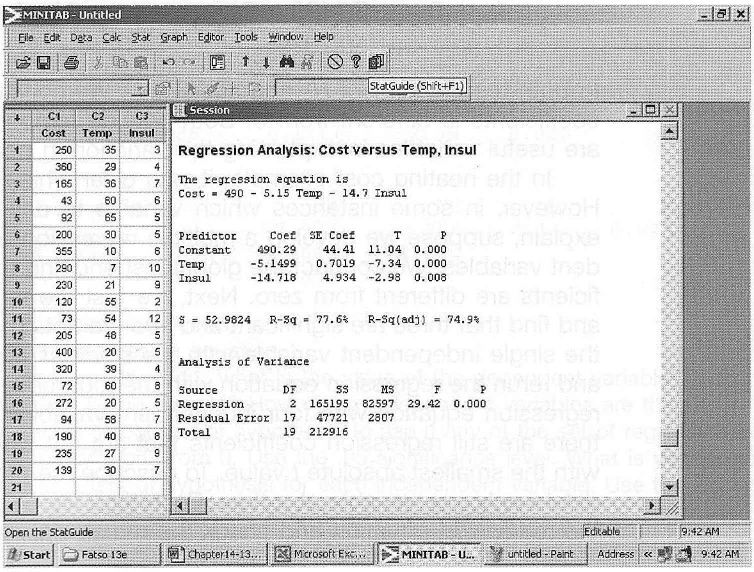
The $b_i$ refers to any one of the regression coefficients and $s_{b_i}$ refers to standard deviation of that distribution of the regression coefficient. We include 0 in the equation because the null hypothesis is $\beta_i = 0$.

To illustrate this formula, refer to the test of the regression coefficient for the independent variable temperature. We let $b_1$ refer to the regression coefficient. From the computer output on page 525 it is $-4.583$. $s_{b_i}$ is the standard deviation of the sampling distribution of the regression coefficient for the independent variable temperature. Again, from the computer output on page 525, it is 0.772. Inserting these values in formula (14–6):

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-4.583 - 0}{0.772} = -5.936$$

This is the value found in the $t$ Stat column of the Excel output. [There is a slight difference due to rounding.]

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.



| ↓ | 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 |
| 19 | 235 | 27 | 9 |
| 20 | 139 | 30 | 7 |
| 21 | | | |

**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
```

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 $\beta_i$'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.

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                Insulation

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-5.1499 - 0}{0.7019} = -7.34 \qquad t = \frac{b_2 - 0}{s_{b_2}} = \frac{-14.718 - 0}{4.934} = -2.98$$

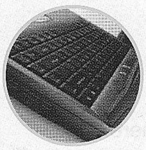
In both tests we reject $H_0$ and accept $H_1$. We conclude that each of the regression coefficients is different from 0. Both outside temperature and amount of insulation are useful variables in explaining the variation in heating costs.

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. 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
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

Analysis of Variance
Source           DF      SS     MS
Regression        5     100     20
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

**7.** Given the following regression output,

```
Predictor      Coef     SE Coef        T       P
Constant     84.998      1.863     45.61   0.000
X1            2.391      1.200      1.99   0.051
X2          -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
Constant        7.987      2.967       2.69
X1            0.12242    0.03121       3.92
X2           -0.12166    0.05353      -2.27
X3           -0.06281    0.03901      -1.61
X4            0.5235     0.1420        3.69
X5           -0.06472    0.03999      -1.62

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
```

$X_1$ 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?
**e.** Outline a strategy for deleting independent variables in this case.

## 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. It should be mentioned, however, that in practice strict adherence to the following assumptions is not always possible. Fortunately the statistical techniques discussed in this chapter appear to 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 page 480.) 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 $\hat{Y}$.** The 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.

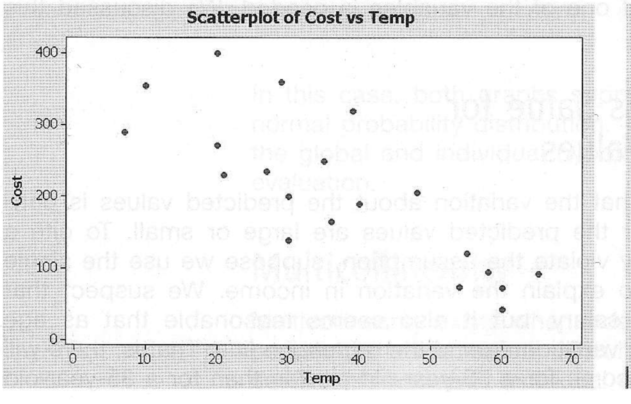
In this section we present a brief discussion of each of these assumptions. In addition, we provide methods to validate these assumptions and indicate the consequences if these assumptions cannot be met. For those interested in additional discussion, Kutner, Nachtscheim, and Neter, *Applied Linear Regression Models*, 4th ed. (McGraw-Hill: 2004), is an excellent reference.

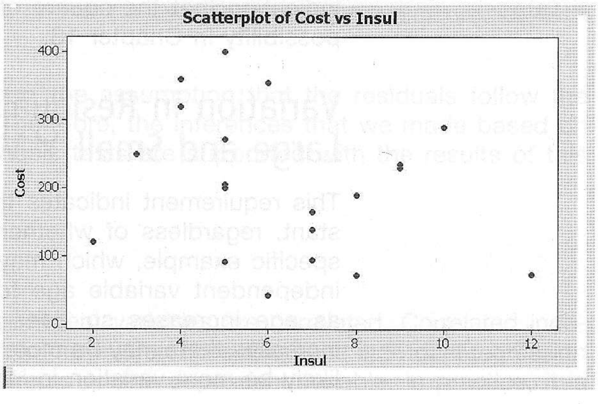
## Linear Relationship

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 513. We can evaluate this assumption with scatter diagrams and residual plots.

**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



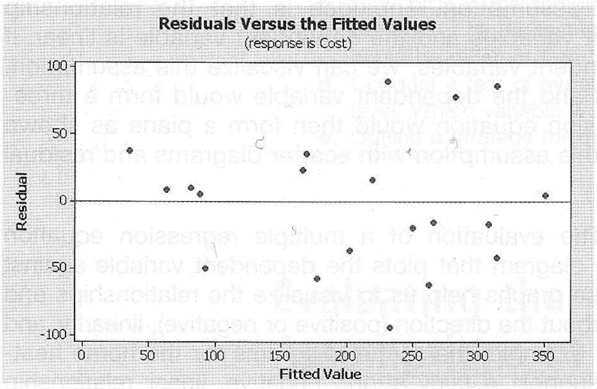


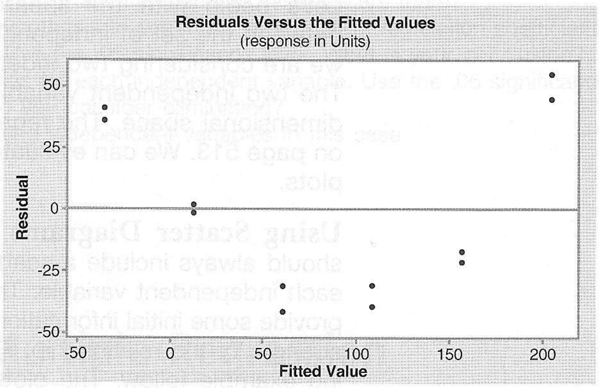
discussed the idea that the best regression line passed through the center of the data in a scatter plot. In this case, you would find a good number of the observations above the regression line (these residuals would have a positive sign), and a good number of the observations below the line (these 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 show 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 perhaps the equation is quadratic, indicating that the square of one of the variables is needed. We discussed this possibility 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 532.) Based on the scatter diagram in that software output, it is reasonable to conclude that this assumption has not been violated.

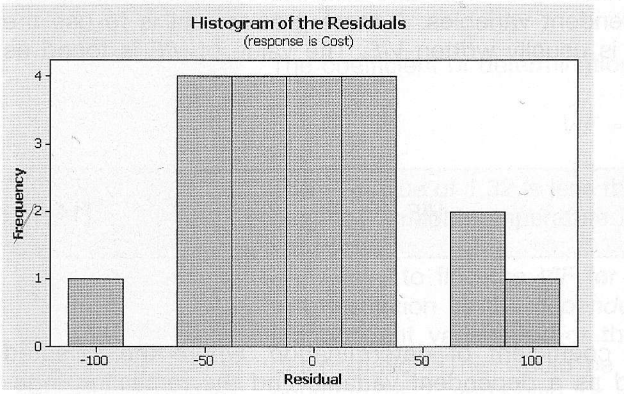
## Distribution of Residuals

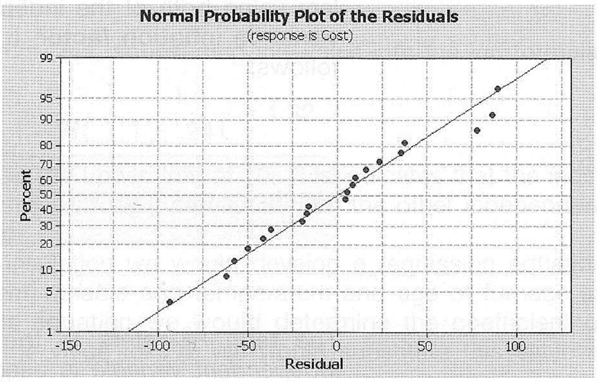
To be sure that the inferences that we make in the global and individual hypotheses tests are valid, we evaluate the distribution of residuals. Ideally, the residuals should follow a normal probability distribution.

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. Without detailing the calculations, 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.

Minitab





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

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 multicollinearity, 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

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

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 four 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
```

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     VIF
Constant       57.99      12.35    4.70   0.000
Insul         -0.509      1.488   -0.34   0.737     1.0
Age           -2.509      1.103   -2.27   0.036     1.0

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.

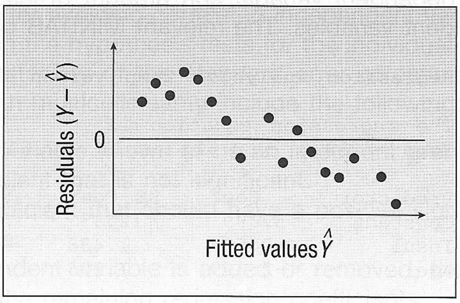
Again, to find the *VIF* for insulation we would develop a regression equation with insulation as the *dependent variable* and temperature and age of furnace as independent variables. For this equation we would determine the coefficient of determination. This would be the value for $R_2^2$. We would 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* of the MINITAB output. Both of these values are 1.0, hence we conclude there is not a problem with multicollinearity in this example.

## Independent Observations

The fifth assumption about regression and correlation analysis is that successive residuals should be independent. This means that there is not a pattern to the residuals, the residuals are not highly correlated, and there are not long runs of positive or negative residuals. When successive residuals are correlated we refer to this condition as **autocorrelation.**

Autocorrelation frequently occurs when the data are collected over a period of time. For example, we wish to predict yearly sales of Ages Software, Inc., based on the time and the amount spent on advertising. The dependent variable is yearly sales and the independent variables are time and amount spent on advertising. It is likely that for a period of time the actual points will be above the regression plane (remember there are two independent variables) and then for a period of time the points will be below the regression plane. The graph below shows the residuals plotted on the vertical axis and the fitted values $\hat{Y}$ on the horizontal axis. Note the run of residuals above the mean of the residuals, followed by a run below the mean. A scatter plot such as this would indicate possible autocorrelation.



There is a test for autocorrelation, called the Durbin-Watson. We present the details of this test in Chapter 16.

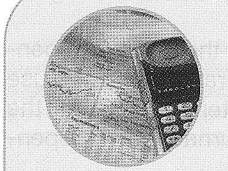
## Qualitative Independent Variables

In the previous example regarding heating cost, the two independent variables outside temperature and insulation were quantitative; that is, numerical in nature. Frequently we wish to use nominal-scale variables—such as gender, whether the home has a swimming pool, or whether the sports team was the home or the visiting team—in our analysis. These are called **qualitative variables** because they describe a particular quality, such as male or female. To use a qualitative variable in regression analysis, we use a scheme of **dummy variables** in which one of the two possible conditions is coded 0 and the other 1.

**DUMMY VARIABLE** A variable in which there are only two possible outcomes. For analysis, one of the outcomes is coded a 1 and the other a 0.

For example, we might be interested in estimating an executive's salary on the basis of years of job experience and whether he or she graduated from college. "Graduation from college" can take on only one of two conditions: yes or no. Thus, it is considered a qualitative variable.

Suppose in the Salsberry Realty example that the independent variable "garage" is added. For those homes without an attached garage, 0 is used; for homes with an attached garage, a 1 is used. We will refer to the "garage" variable as $X_4$. The data from Table 14–2 are entered into the MINITAB system.



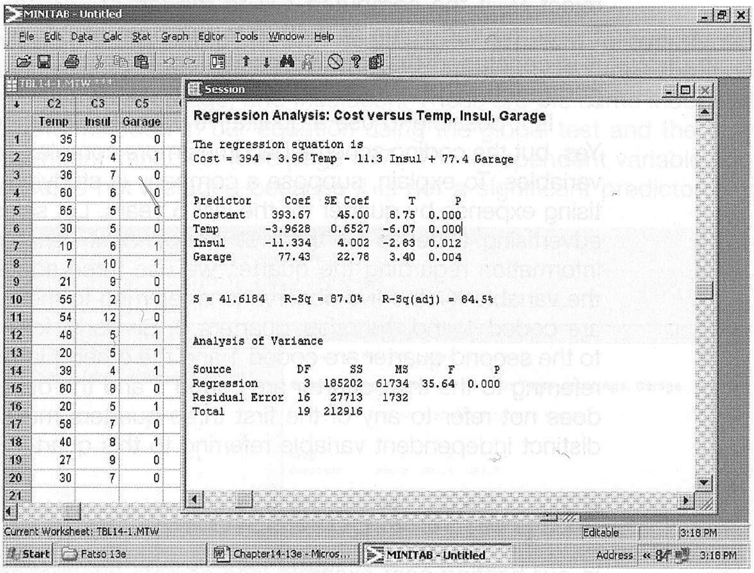
### Statistics in Action

In recent years, multiple regression has been used in a variety of legal proceedings. It is particularly useful in cases alleging discrimination by gender or race. As an example, suppose that a woman alleges that Company X's wage rates are unfair to women. To support the claim, the plaintiff produces data showing that, on the average, women earn less than men. In response, Company X argues that its wage rates are based on experience, training, and skill and that its female employees, on the average, are younger and less experienced than the male employees. In fact, the company might further argue that the current situation is actually due to its recent successful efforts to hire more women.

**TABLE 14–2** Home Heating Costs, Temperature, Insulation, and Presence of a Garage for a Sample of 20 Homes

| Cost, $Y$ | Temperature, $X_1$ | Insulation, $X_2$ | Garage, $X_4$ |
|---|---|---|---|
| $250 | 35 | 3 | 0 |
| 360 | 29 | 4 | 1 |
| 165 | 36 | 7 | 0 |
| 43 | 60 | 6 | 0 |
| 92 | 65 | 5 | 0 |
| 200 | 30 | 5 | 0 |
| 355 | 10 | 6 | 1 |
| 290 | 7 | 10 | 1 |
| 230 | 21 | 9 | 0 |
| 120 | 55 | 2 | 0 |
| 73 | 54 | 12 | 0 |
| 205 | 48 | 5 | 1 |
| 400 | 20 | 5 | 1 |
| 320 | 39 | 4 | 1 |
| 72 | 60 | 8 | 0 |
| 272 | 20 | 5 | 1 |
| 94 | 58 | 7 | 0 |
| 190 | 40 | 8 | 1 |
| 235 | 27 | 9 | 0 |
| 139 | 30 | 7 | 0 |

The output from MINITAB is:


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
TBL14-1.MTW ***
Session
Current Worksheet: TBL14-1.MTW
Editable
3:18 PM
Start
Fatso 13e
Chapter14-13e - Micros...
MINITAB - Untitled
Address
3:18 PM

| ↓ | C2 | C3 | C5 |
|---|---|---|---|
| | Temp | Insul | Garage |
| 1 | 35 | 3 | 0 |
| 2 | 29 | 4 | 1 |
| 3 | 36 | 7 | 0 |
| 4 | 60 | 6 | 0 |
| 5 | 65 | 5 | 0 |
| 6 | 30 | 5 | 0 |
| 7 | 10 | 6 | 1 |
| 8 | 7 | 10 | 1 |
| 9 | 21 | 9 | 0 |
| 10 | 55 | 2 | 0 |
| 11 | 54 | 12 | 0 |
| 12 | 48 | 5 | 1 |
| 13 | 20 | 5 | 1 |
| 14 | 39 | 4 | 1 |
| 15 | 60 | 8 | 0 |
| 16 | 20 | 5 | 1 |
| 17 | 58 | 7 | 0 |
| 18 | 40 | 8 | 1 |
| 19 | 27 | 9 | 0 |
| 20 | 30 | 7 | 0 |
| 21 | | | |

**Regression Analysis: Cost versus Temp, Insul, Garage**

```
The regression equation is
Cost = 394 - 3.96 Temp - 11.3 Insul + 77.4 Garage


Predictor       Coef  SE Coef      T      P
Constant      393.67    45.00   8.75  0.000
Temp         -3.9628   0.6527  -6.07  0.000
Insul        -11.334    4.002  -2.83  0.012
Garage         77.43    22.78   3.40  0.004


S = 41.6184   R-Sq = 87.0%   R-Sq(adj) = 84.5%


Analysis of Variance

Source          DF      SS     MS      F      P
Regression       3  185202  61734  35.64  0.000
Residual Error  16   27713   1732
Total           19  212916
```

What is the effect of the garage variable? Should it be included in the analysis? To show the effect of the variable, suppose we have two houses exactly alike next to each other in Buffalo, New York; one has an attached garage, and the other does not. Both homes have 3 inches of insulation, and the mean January temperature in Buffalo is 20 degrees. For the house without an attached garage, a 0 is substituted for $X_4$ in the regression equation. The estimated heating cost is \$280.90, found by:

$$\begin{aligned}\hat{Y} &= 394 - 3.96X_1 - 11.3X_2 + 77.4X_4\\ &= 394 - 3.96(20) - 11.3(3) + 77.4(0) = 280.90\end{aligned}$$

For the house with an attached garage, a 1 is substituted for $X_4$ in the regression equation. The estimated heating cost is \$358.30, found by:

$$\begin{aligned}\hat{Y} &= 394 - 3.96X_1 - 11.3X_2 + 77.4X_4\\ &= 394 - 3.96(20) - 11.3(3) + 77.4(1) = 358.30\end{aligned}$$

The difference between the estimated heating costs is \$77.40 (\$358.30 −\$280.90). Hence, we can expect the cost to heat a house with an attached garage to be \$77.40 more than the cost for an equivalent house without a garage.

We have shown the difference between the two types of homes to be \$77.40, but is the difference significant? We conduct the following test of hypothesis.

$$H_0: \beta_4 = 0$$

$$H_1: \beta_4 \neq 0$$

The information necessary to answer this question is on the MINITAB output above. The net regression coefficient for the independent variable garage is 77.43, and the standard deviation of the sampling distribution is 22.78. We identify this as the fourth independent variable, so we use a subscript of 4. Finally, we insert these values in formula (14–6).

$$t = \frac{b_4 - 0}{s_{b_4}} = \frac{77.43 - 0}{22.78} = 3.40$$

There are three independent variables in the analysis, so there are $n - (k + 1) = 20 - (3 + 1) = 16$ degrees of freedom. The critical value from Appendix B.2 is 2.120. The decision rule, using a two-tailed test and the .05 significance level, is to reject $H_0$ if the computed $t$ is to the left of −2.120 or to the right of 2.120. Since the computed value of 3.40 is to the right of 2.120, the null hypothesis is rejected. It is concluded that the regression coefficient is not zero. The independent variable garage should be included in the analysis.

Is it possible to use a qualitative variable with more than two possible outcomes? Yes, but the coding scheme becomes more complex and will require a series of dummy variables. To explain, suppose a company is studying its sales as they relate to advertising expense by quarter for the last 5 years. Let sales be the dependent variable and advertising expense be the first independent variable, $X_1$. To include the qualitative information regarding the quarter, we use three additional independent variables. For the variable $X_2$, the five observations referring to the first quarter of each of the 5 years are coded 1 and the other quarters 0. Similarly, for $X_3$ the five observations referring to the second quarter are coded 1 and the other quarters 0. For $X_4$ the five observations referring to the third quarter are coded 1 and the other quarters 0. An observation that does not refer to any of the first three quarters must refer to the fourth quarter, so a distinct independent variable referring to this quarter is not necessary.

## Stepwise Regression

In our heating cost example (see sample information in Table 14–1 and Table 14–2) we considered four independent variables: the mean outside temperature, the amount of insulation in the home, the age of the furnace, and whether or not there was an attached garage. To obtain the equation, we first ran a global or "all at once" test to determine if any of the regression coefficients were significant. When we found at least one to be significant, we tested the regression coefficients individually to determine which were important. We left out the independent variables that did not have significant regression coefficients and kept the others. By retaining the independent variables with significant coefficients, we found the regression equation that used the fewest independent variables. This made the regression equation easy to interpret and explained as much variation in the dependent variable as possible.

We are now going to describe a technique called **stepwise regression,** which is more efficient in building the regression equation.

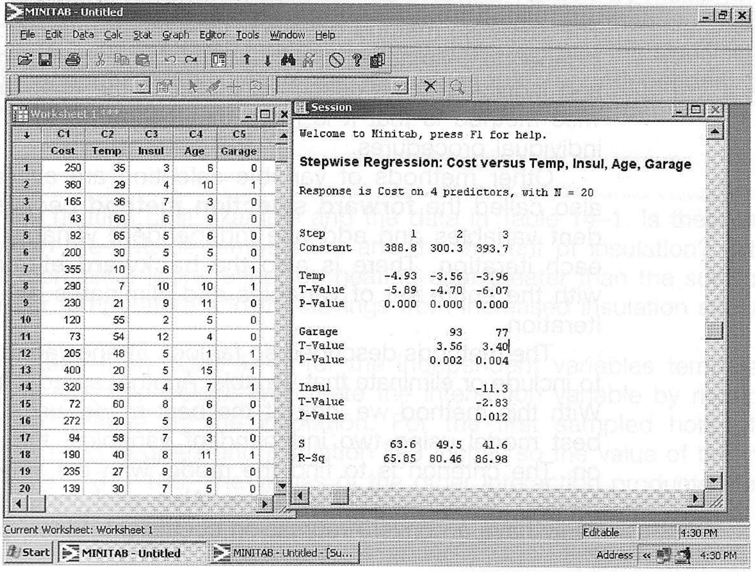
**STEPWISE REGRESSION** A step-by-step method to determine a regression equation that begins with a single independent variable and adds or deletes independent variables one by one. Only independent variables with nonzero regression coefficients are included in the regression equation.

In the stepwise method, we develop a sequence of equations. The first equation contains only one independent variable. However, this independent variable is the one from the set of proposed independent variables that explains the most variation in the dependent variable. Stated differently, if we compute all the simple correlations between each independent variable and the dependent variable, the stepwise method first selects the independent variable with the strongest correlation with the dependent variable.

Next, the stepwise method looks at the remaining independent variables and then selects the one that will explain the largest percentage of the variation yet unexplained. We continue this process until all the independent variables with significant regression coefficients are included in the regression equation. The advantages to the stepwise method are:

1. Only independent variables with significant regression coefficients are entered into the equation.
2. The steps involved in building the regression equation are clear.
3. It is efficient in finding the regression equation with only significant regression coefficients.
4. The changes in the multiple standard error of estimate and the coefficient of determination are shown.

The stepwise MINITAB output for the heating cost problem follows. Note that the final equation, which is reported in the column labeled 3 includes the independent variables temperature, garage, and insulation. These are the same independent variables that were included in our equation using the global test and the test for individual independent variables. (See page 537.) The independent variable age, for age of the furnace, is not included because it is not a significant predictor of cost.


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Worksheet 1 ***
Session
Current Worksheet: Worksheet 1
Start
MINITAB - Untitled
MINITAB - Untitled - [Su...
Editable
4:30 PM
Address
4:30 PM

| ↓ | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
|  | Cost | Temp | Insul | Age | Garage |
| 1 | 250 | 35 | 3 | 6 | 0 |
| 2 | 360 | 29 | 4 | 10 | 1 |
| 3 | 165 | 36 | 7 | 3 | 0 |
| 4 | 43 | 60 | 6 | 9 | 0 |
| 5 | 92 | 65 | 5 | 6 | 0 |
| 6 | 200 | 30 | 5 | 5 | 0 |
| 7 | 355 | 10 | 6 | 7 | 1 |
| 8 | 290 | 7 | 10 | 10 | 1 |
| 9 | 230 | 21 | 9 | 11 | 0 |
| 10 | 120 | 55 | 2 | 5 | 0 |
| 11 | 73 | 54 | 12 | 4 | 0 |
| 12 | 205 | 48 | 5 | 1 | 1 |
| 13 | 400 | 20 | 5 | 15 | 1 |
| 14 | 320 | 39 | 4 | 7 | 1 |
| 15 | 72 | 60 | 8 | 6 | 0 |
| 16 | 272 | 20 | 5 | 8 | 1 |
| 17 | 94 | 58 | 7 | 3 | 0 |
| 18 | 190 | 40 | 8 | 11 | 1 |
| 19 | 235 | 27 | 9 | 8 | 0 |
| 20 | 139 | 30 | 7 | 5 | 0 |

```
Welcome to Minitab, press F1 for help.

Stepwise Regression: Cost versus Temp, Insul, Age, Garage

Response is Cost on 4 predictors, with N = 20


Step            1      2      3
Constant    388.8  300.3  393.7

Temp        -4.93  -3.56  -3.96
T-Value     -5.89  -4.70  -6.07
P-Value     0.000  0.000  0.000

Garage                93     77
T-Value             3.56   3.40
P-Value            0.002  0.004

Insul                     -11.3
T-Value                   -2.83
P-Value                   0.012

S            63.6   49.5   41.6
R-Sq        65.85  80.46  86.98
```

Reviewing the steps and interpreting output:

1. The stepwise procedure selects the independent variable temperature first. This variable explains more of the variation in heating cost than any of the other three proposed independent variables. Temperature explains 65.85 percent of the variation in heating cost. The regression equation is:

$$\hat{Y} = 388.8 - 4.93X_1$$

There is an inverse relationship between heating cost and temperature. For each degree the temperature increases, heating cost is reduced by \$4.93.

2. The next independent variable to enter the regression equation is garage. When this variable is added to the regression equation, the coefficient of determination is increased from 65.85 percent to 80.46 percent. That is, by adding garage as an independent variable, we increase the coefficient of determination by 14.61 percent. The regression equation after step 2 is:

$$\hat{Y} = 300.3 - 3.56X_1 + 93.0X_2$$

Usually the regression coefficients will change from one step to the next. In this case the coefficient for temperature retained its negative sign, but it changed from −4.93 to −3.56. This change is reflective of the added influence of the independent variable garage. Why did the stepwise method select the independent variable garage instead of either insulation or age? The increase in $R^2$, the coefficient of determination, is larger if garage is included rather than either of the other two variables.

3. At this point there are two unused variables remaining, insulation and age. Notice on the third step the procedure selects insulation and then stops. This indicates the variable insulation explains more of the remaining variation in heating cost than the age variable does. After the third step, the regression equation is:

$$\hat{Y} = 393.7 - 3.96X_1 + 77.0X_2 - 11.3X_3$$

At this point 86.98 percent of the variation in heating cost is explained by the three independent variables temperature, garage, and insulation. This is the same $R^2$ value and regression equation we found on page 537 except for rounding differences.

4. At this point the stepwise procedure stops. This means the independent variable age does not add significantly to the coefficient of determination.

The stepwise method developed the same regression equation, selected the same independent variables, and found the same coefficient of determination as the global and individual tests described earlier in the chapter. The advantages to the stepwise method is that it is more direct than using a combination of the global and individual procedures.

Other methods of variable selection are available. The stepwise method is also called the **forward selection method** because we begin with no independent variables and add one independent variable to the regression equation at each iteration. There is also the **backward elimination method,** which begins with the entire set of variables and eliminates one independent variable at each iteration.

The methods described so far look at one variable at a time and decide whether to include or eliminate that variable. Another approach is the **best-subset regression.** With this method we look at the best model using one independent variable, the best model using two independent variables, the best model with three and so on. The criterion is to find the model with the largest $R^2$ value, regardless of the number of independent variables. Also, each independent variable does not necessarily have a nonzero regression coefficient. Since each independent variable could either be included or not included, there are $2^k - 1$ possible models, where $k$ refers to the number of independent variables. In our heating cost example we considered four independent variables so there are 15 possible regression models, found by $2^4 - 1 = 16 - 1 = 15$. We would examine all regression models using one independent variable, all combinations using two variables, all combinations using three independent variables, and the possibility of using all four independent variables. The advantages to the best-subset method is it may examine combinations of independent variables not considered in the stepwise method. The process is available in MINITAB and MegaStat.

## Regression Models with Interaction

In Chapter 12 we discussed interaction among independent variables. To explain, suppose we are studying weight loss and assume, as the current literature suggests, that diet and exercise are related. So the dependent variable is amount of change in weight and the independent variables are: diet (yes or no) and exercise (none, moderate, significant). We are interested in whether there is interaction among the independent variables. That is, if those studied maintain their diet and exercise significantly, will that increase the mean amount of weight lost? Is total weight loss more than the sum of the loss due to the diet effect and the loss due to the exercise effect?

We can expand on this idea. Instead of having two nominal-scale variables, diet and exercise, we can examine the effect (interaction) of several ratio-scale variables. For example, suppose we want to study the effect of room temperature (68, 72, 76, or 80 degrees Fahrenheit) and noise level (60, 70, or 80 decibels) on the number of units produced. To put it another way, does the combination of noise level in the room and the temperature of the room have an effect on the productivity of the workers? Would the workers in a quiet, cool room produce more units than those in a hot, noisy room?

In regression analysis, interaction can be examined as a separate independent variable. An interaction prediction variable can be developed by multiplying the data values in one independent variable by the values in another independent variable, thereby creating a new independent variable. A two-variable model that includes an interaction term is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

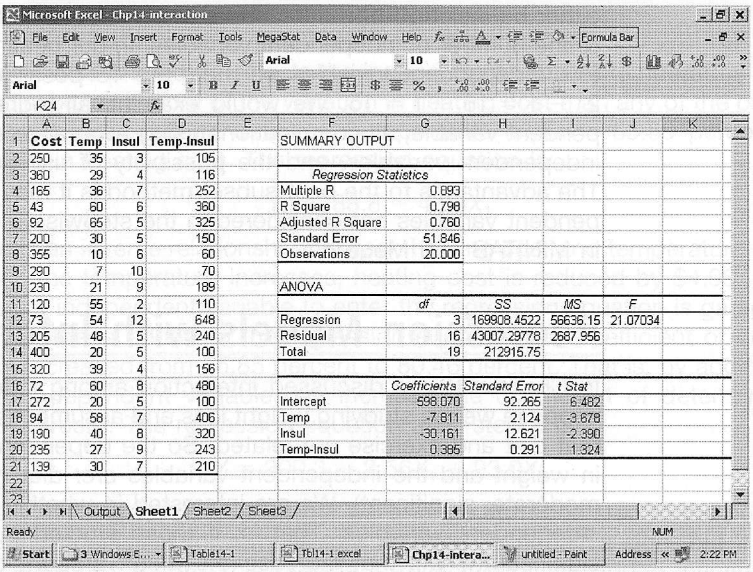
The term $X_1X_2$ is the *interaction term.* We create this variable by multiplying the values of $X_1$ and $X_2$ to create the third independent variable. We then develop a regression equation using the three independent variables and test the significance of the third independent variable using the individual test for independent variables, described earlier in the chapter. An example will illustrate the details.

**Example**

Refer to the heating cost example and the data in Table 14–1. Is there an interaction between the outside temperature and the amount of insulation? If both variables are increased, is the effect on heating cost greater than the sum of savings from warmer temperature and the savings from increased insulation separately?

**Solution**

The information from Table 14–1 for the independent variables temperature and insulation is repeated below. We create the interaction variable by multiplying the temperature variable by the insulation. For the first sampled home the value temperature is 35 degrees and insulation is 3 inches so the value of the interaction variable is 35 × 3 = 105. The values of the other interaction products are found in a similar fashion.


Microsoft Excel - Chp14-interaction
File Edit View Insert Format Tools MegaStat Data Window Help
Formula Bar
Arial 10
K24
Output Sheet1 Sheet2 Sheet3
Ready
NUM
Start 3 Windows E... Table14-1 Tbl14-1 excel Chp14-intera... untitled - Paint Address 2:22 PM

| Cost | Temp | Insul | Temp-Insul |
|---|---|---|---|
| 250 | 35 | 3 | 105 |
| 360 | 29 | 4 | 116 |
| 165 | 36 | 7 | 252 |
| 43 | 60 | 6 | 360 |
| 92 | 65 | 5 | 325 |
| 200 | 30 | 5 | 150 |
| 355 | 10 | 6 | 60 |
| 290 | 7 | 10 | 70 |
| 230 | 21 | 9 | 189 |
| 120 | 55 | 2 | 110 |
| 73 | 54 | 12 | 648 |
| 205 | 48 | 5 | 240 |
| 400 | 20 | 5 | 100 |
| 320 | 39 | 4 | 156 |
| 72 | 60 | 8 | 480 |
| 272 | 20 | 5 | 100 |
| 94 | 58 | 7 | 406 |
| 190 | 40 | 8 | 320 |
| 235 | 27 | 9 | 243 |
| 139 | 30 | 7 | 210 |

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.893 |
| R Square | 0.798 |
| Adjusted R Square | 0.760 |
| Standard Error | 51.846 |
| Observations | 20.000 |

ANOVA

| | *df* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| Regression | 3 | 169908.4522 | 56636.15 | 21.07034 |
| Residual | 16 | 43007.29778 | 2687.956 | |
| Total | 19 | 212915.75 | | |

| | *Coefficients* | *Standard Error* | *t Stat* |
|---|---|---|---|
| Intercept | 598.070 | 92.265 | 6.482 |
| Temp | -7.811 | 2.124 | -3.678 |
| Insul | -30.161 | 12.621 | -2.390 |
| Temp-Insul | 0.385 | 0.291 | 1.324 |

We find the multiple regression using temperature, insulation, and the interaction of temperature and insulation as independent variables. The regression equation is reported below.

$$\hat{Y} = 598.070 - 7.811X_1 - 30.161X_2 + 0.385X_1X_2$$

The question we wish to answer is whether the interaction variable is significant. We will use the .05 significance level. In terms of a hypothesis:

$H_0$: $\beta_3 = 0$
$H_1$: $\beta_3 \neq 0$

There is $n - (k + 1) = 20 - (3 + 1) = 16$ degrees of freedom. Using the .05 significance level and a two-tailed test, the critical values of $t$ are $-2.120$ and $2.120$. We reject the null hypothesis if $t$ is less than $-2.120$ or $t$ is greater than $2.120$. From the output, $b_3 = 0.385$ and $s_{b_3} = 0.291$. To find the value of $t$ we use formula (14–6).

$$t = \frac{b_3 - 0}{s_{b_3}} = \frac{0.385 - 0}{0.291} = 1.324$$

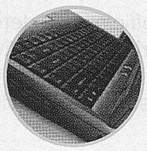
Because the computed value of 1.324 is less than the critical value of 2.120, we do not reject the null hypothesis. We conclude that there is not a significant interaction between temperature and insulation.

There are other situations that can occur when studying interaction among independent variables.

1. It is possible to have a three-way interaction among the independent variables. In our heating example, we might have considered the three-way interaction between temperature, insulation, and age of the furnace.
2. It is possible to have an interaction where one of the independent variables is nominal scale. In our heating cost example, we could have studied the interaction between temperature and garage.

Studying all possible interactions can become very complex. However, careful consideration to possible interactions among independent variables can often provide useful insight into the regression models.

**Self-Review 14–4**



A study by the American Realtors Association investigated the relationship between the commissions earned by sales associates last year and the number of months since the associates earned their real estate licenses. Also of interest in the study is the gender of the sales associate. Below is a portion of the regression output. The dependent variable is commissions, which is reported in \$000, and the independent variables are months since the license was earned and gender (female = 1 and male = 0).

```
Regression Analysis
         R²  0.642
Adjusted R²  0.600                n    20
          R  0.801                k    2
Std.  Error  3.219          Dep.  Var.   Commissions

ANOVA table
Source            SS     df          MS         F      p-value
Regression   315.9291     2    157.9645     15.25        .0002
Residual     176.1284    17     10.3605
Total        492.0575    19

Regression output
Variables     coefficients     std.  error    t  (df = 17)     p-value     95%  lower      95%  upper
Intercept          15.7625          3.0782         5.121       .0001         9.2680         22.2570
   Months           0.4415          0.0839         5.263       .0001         0.2645          0.6186
   Gender           3.8598          1.4724         2.621       .0179         0.7533          6.9663
```

(a) Write out the regression equation. How much commission would you expect a female agent to make who earned her license 30 months ago?

(b) Do the female agents on the average make more or less than the male agents? How much more?

(c) Conduct a test of hypothesis to determine if the independent variable gender should be included in the analysis. Use the 0.05 significance level. What is your conclusion?

## Exercises

9. The production manager of High Point Sofa and Chair, a large furniture manufacturer located in North Carolina, is studying the job performance ratings of a sample of 15 electrical repairmen employed by the company. An aptitude test is required by the human resources department to become an electrical repairman. The production manager was able to get the score for each repairman in the sample. In addition, he determined which of the repairmen were union members (code = 1) and which were not (code = 0). The sample information is reported below.

| Worker | Job Performance Score | Aptitude Test Score | Union Membership |
|---|---|---|---|
| Abbott | 58 | 5 | 0 |
| Anderson | 53 | 4 | 0 |
| Bender | 33 | 10 | 0 |
| Bush | 97 | 10 | 0 |
| Center | 36 | 2 | 0 |
| Coombs | 83 | 7 | 0 |
| Eckstine | 67 | 6 | 0 |
| Gloss | 84 | 9 | 0 |
| Herd | 98 | 9 | 1 |
| Householder | 45 | 2 | 1 |
| Iori | 97 | 8 | 1 |
| Lindstrom | 90 | 6 | 1 |
| Mason | 96 | 7 | 1 |
| Pierse | 66 | 3 | 1 |
| Rohde | 82 | 6 | 1 |

**a.** Use a statistical software package to develop a multiple regression equation using the job performance score as the dependent variable and aptitude test score and union membership as independent variables.

**b.** Comment on the regression equation. Be sure to include the coefficient of determination and the effect of union membership. Are these two variables effective in explaining the variation in job performance?

**c.** Conduct a test of hypothesis to determine if union membership should be included as an independent variable.

**d.** Repeat the analysis considering possible interaction terms.

**10.** Cincinnati Paint Company sells quality brands of paints through hardware stores throughout the United States. The company maintains a large sales force whose job it is to call on existing customers as well as look for new business. The national sales manager is investigating the relationship between the number of sales calls made and the miles driven by the sales representative. Also, do the sales representatives who drive the most miles and make the most calls necessarily earn the most in sales commissions? To investigate, the vice president of sales selected a sample of 25 sales representatives and determined:

- The amount earned in commissions last month ($Y$).
- The number of miles driven last month ($X_1$)
- The number of sales calls made last month ($X_2$)

The information is reported below.

| Commissions ($000) | Calls | Driven | Commissions ($000) | Calls | Driven |
|---|---|---|---|---|---|
| 22 | 139 | 2,371 | 38 | 146 | 3,290 |
| 13 | 132 | 2,226 | 44 | 144 | 3,103 |
| 33 | 144 | 2,731 | 29 | 147 | 2,122 |
| 38 | 142 | 3,351 | 38 | 144 | 2,791 |
| 23 | 142 | 2,289 | 37 | 149 | 3,209 |
| 47 | 142 | 3,449 | 14 | 131 | 2,287 |
| 29 | 138 | 3,114 | 34 | 144 | 2,848 |
| 38 | 139 | 3,342 | 25 | 132 | 2,690 |
| 41 | 144 | 2,842 | 27 | 132 | 2,933 |
| 32 | 134 | 2,625 | 25 | 127 | 2,671 |
| 20 | 135 | 2,121 | 43 | 154 | 2,988 |
| 13 | 137 | 2,219 | 34 | 147 | 2,829 |
| 47 | 146 | 3,463 | | | |

Develop a regression equation including an interaction term. Is there a significant interaction between the number of sales calls and the miles driven?

**11.** An art collector is studying the relationship between the selling price of a painting and two independent variables. The two independent variables are the number of bidders at the particular auction and the age of the painting, in years. A sample of 25 paintings revealed the following sample information.

| Painting | Auction Price | Bidders | Age | Painting | Auction Price | Bidders | Age |
|---|---|---|---|---|---|---|---|
| 1 | 3,470 | 10 | 67 | 14 | 4,020 | 6 | 79 |
| 2 | 3,500 | 8 | 56 | 15 | 4,190 | 4 | 83 |
| 3 | 3,700 | 7 | 73 | 16 | 4,130 | 3 | 71 |
| 4 | 3,860 | 4 | 71 | 17 | 4,130 | 9 | 89 |
| 5 | 3,920 | 12 | 99 | 18 | 4,370 | 5 | 103 |
| 6 | 3,900 | 10 | 87 | 19 | 4,450 | 3 | 106 |
| 7 | 3,830 | 11 | 78 | 20 | 4,390 | 8 | 93 |
| 8 | 3,940 | 8 | 83 | 21 | 4,380 | 8 | 88 |
| 9 | 3,880 | 13 | 90 | 22 | 4,540 | 4 | 96 |
| 10 | 3,940 | 13 | 98 | 23 | 4,660 | 5 | 94 |
| 11 | 4,200 | 0 | 91 | 24 | 4,710 | 3 | 88 |
| 12 | 4,060 | 7 | 93 | 25 | 4,880 | 1 | 84 |
| 13 | 4,200 | 2 | 97 | | | | |

a. Develop a multiple regression equation using the independent variables number of bidders and age of painting to estimate the dependent variable auction price. Discuss the equation. Does it surprise you that there is an inverse relationship between the number of bidders and the price of the painting?

b. Create an interaction variable and include it in the regression equation. Explain the meaning of the interaction. Is this variable significant?

c. Use the stepwise method and the independent variables for the number of bidders, the age of the painting, and the interaction between the number of bidders and the age of the painting. Which variables would you select?

12. A real estate developer wishes to study the relationship between the size of home a client will purchase (in square feet) and other variables. Possible independent variables include the family income, family size, whether there is a senior adult parent living with the family (1 for yes, 0 for no), and the total years of education beyond high school for the husband and wife. The sample information is reported below.

| Family | Square Feet | Income (000s) | Family Size | Senior Parent | Education |
|---|---|---|---|---|---|
| 1 | 2,240 | 60.8 | 2 | 0 | 4 |
| 2 | 2,380 | 68.4 | 2 | 1 | 6 |
| 3 | 3,640 | 104.5 | 3 | 0 | 7 |
| 4 | 3,360 | 89.3 | 4 | 1 | 0 |
| 5 | 3,080 | 72.2 | 4 | 0 | 2 |
| 6 | 2,940 | 114 | 3 | 1 | 10 |
| 7 | 4,480 | 125.4 | 6 | 0 | 6 |
| 8 | 2,520 | 83.6 | 3 | 0 | 8 |
| 9 | 4,200 | 133 | 5 | 0 | 2 |
| 10 | 2,800 | 95 | 3 | 0 | 6 |

Develop an appropriate multiple regression equation. Which independent variables would you include in the final regression equation? Use the stepwise method.

## Chapter Summary

I. The general form of a multiple regression equation is:

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

where $a$ is the $Y$-intercept when all $X$'s are zero, $b_i$ refers to the sample regression coefficients, and $X_i$ refers to the value of the various independent variables.

**A.** There can be any number of independent variables.

**B.** The least squares criterion is used to develop the regression equation.

**C.** A statistical software package is needed to perform the calculations.

II. There are two measures of the effectiveness of the regression equation.

**A.** The multiple standard error of estimate is similar to the standard deviation.

1. It is measured in the same units as the dependent variable.
2. It is based on squared deviations from the regression equation.
3. It ranges from 0 to plus infinity.
4. It is calculated from the following equation.

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

**B.** The coefficient of multiple determination reports the percent of the variation in the dependent variable explained by the set of independent variables.

1. It may range from 0 to 1.
2. It is also based on squared deviations from the regression equation.

3. It is found by the following equation.

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

4. When the number of independent variables is large, we adjust the coefficient of determination for the degrees of freedom as follows.

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

**III.** An ANOVA table summarizes the multiple regression analysis.

**A.** It reports the total amount of the variation in the dependent variable and divides this variation into that explained by the set of independent variables and that not explained.

**B.** It reports the degrees of freedom associated with the independent variables, the error variation, and the total variation.

**IV.** A correlation matrix shows all possible simple correlation coefficients between pairs of variables.

**A.** It shows the correlation between each independent variable and the dependent variable.

**B.** It shows the correlation between each pair of independent variables.

**V.** A global test is used to investigate whether any of the independent variables have significant regression coefficients.

**A.** The null hypothesis is: All the regression coefficients are zero.

**B.** The alternate hypothesis is: At least one regression coefficient is not zero.

**C.** The test statistic is the *F* distribution with *k* (the number of independent variables) degrees of freedom in the numerator and $n - (k + 1)$ degrees of freedom in the denominator, where *n* is the sample size.

**D.** The formula to calculate the value of the test statistic for the global test is:

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

**VI.** The test for individual variables determines which independent variables have nonzero regression coefficients.

**A.** The variables that have zero regression coefficients are usually dropped from the analysis.

**B.** The test statistic is the *t* distribution with $n - (k + 1)$ degrees of freedom.

**C.** The formula to calculate the value of the test statistic for the individual test is:

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

**VII.** Dummy variables are used to represent qualitative variables and can assume only one of two possible outcomes.

**VIII.** There are five assumptions to use multiple regression analysis.

**A.** The relationship between the dependent variable and the set of independent variables must be linear.

1. To verify this assumption develop a scatter diagram and plot the residuals on the vertical axis and the fitted values on the horizontal axis.
2. If the plots appear random, we conclude the relationship is linear.

**B.** The variation is the same for both large and small values of $\hat{Y}$.

1. Homoscedasticity means the variation is the same for all fitted values of the dependent variable.
2. This condition is checked by developing a scatter diagram with the residuals on the vertical axis and the fitted values on the horizontal axis.
3. If there is no pattern to the plots—that is, they appear random—the residuals meet the homoscedasticity requirement.

**C.** The residuals follow the normal probability distribution.
   1. This condition is checked by developing a histogram of the residuals to see if they follow a normal distribution.
   2. The mean of the distribution of the residuals is 0.

**D.** The independent variables are not correlated.
   1. A correlation matrix will show all possible correlation among independent variables. Signs of trouble are correlations larger than 0.70 or less than −0.70.
   2. Signs of correlated independent variables include when an important predictor variable is found insignificant, when an obvious reversal occurs in signs in one or more of the independent variables, or when a variable is removed from the solution, there is a large change in the regression coefficients.
   3. The variance inflation factor is used to identify correlated independent variables.

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

**E.** Each residual is independent of other residuals.
   1. Autocorrelation occurs when successive residuals are correlated.
   2. When autocorrelation exists, the value of the standard error will be biased and will return poor results for tests of hypothesis regarding the regression coefficients.

**IX.** Several techniques help build a regression model.

**A.** A dummy or qualitative independent variable can assume one of two possible outcomes.
   1. A value of 1 is assigned to one outcome and 0 the other.
   2. Use formula (14–6) to determine if the dummy variable should remain in the equation.

**B.** Stepwise regression is a step-by-step process to find the regression equation.
   1. Only independent variables with nonzero regression coefficients enter the equation.
   2. Independent variables are added one at a time to the regression equation.

**C.** Interaction is the case in which one independent variable (such as $X_2$) affects the relationship with another independent variable ($X_1$) and the dependent variable ($Y$).

## Pronunciation Key

| SYMBOL | MEANING | PRONUNCIATION |
|---|---|---|
| $b_1$ | Regression coefficient for the first independent variable | *b sub 1* |
| $b_k$ | Regression coefficient for any independent variable | *b sub k* |
| $s_{y.12...k}$ | Multiple standard error of estimate | *s sub y dot 1, 2 . . . k* |

## Chapter Exercises

**13.** A multiple regression equation yields the following partial results.

| Source | Sum of Squares | *df* |
|---|---|---|
| Regression | 750 | 4 |
| Error | 500 | 35 |

**a.** What is the total sample size?
**b.** How many independent variables are being considered?
**c.** Compute the coefficient of determination.
**d.** Compute the standard error of estimate.
**e.** Test the hypothesis that none of the regression coefficients is equal to zero. Let $\alpha = .05$.

**14.** In a multiple regression equation two independent variables are considered, and the sample size is 25. The regression coefficients and the standard errors are as follows.

$$b_1 = 2.676 \qquad s_{b_1} = 0.56$$

$$b_2 = -0.880 \qquad s_{b_2} = 0.71$$

Conduct a test of hypothesis to determine whether either independent variable has a coefficient equal to zero. Would you consider deleting either variable from the regression equation? Use the .05 significance level.

**15.** The following output was obtained.

```
Analysis of variance

SOURCE         DF        SS          MS
Regression      5       100          20
Error          20        40           2
Total          25       140

Predictor     Coef     StDev    t-ratio
Constant      3.00      1.50       2.00
   X1         4.00      3.00       1.33
   X2         3.00      0.20      15.00
   X3         0.20      0.05       4.00
   X4        -2.50      1.00      -2.50
   X5         3.00      4.00       0.75
```

**a.** What is the sample size?

**b.** Compute the value of $R^2$.

**c.** Compute the multiple standard error of estimate.

**d.** Conduct a global test of hypothesis to determine whether any of the regression coefficients are significant. Use the .05 significance level.

**e.** Test the regression coefficients individually. Would you consider omitting any variable(s)? If so, which one(s)? Use the .05 significance level.

**16.** In a multiple regression equation $k = 5$ and $n = 20$, the MSE value is 5.10, and SS total is 519.68. At the .05 significance level, can we conclude that any of the regression coefficients are not equal to 0?

**17.** The district manager of Jasons, a large discount electronics chain, is investigating why certain stores in her region are performing better than others. She believes that three factors are related to total sales: the number of competitors in the region, the population in the surrounding area, and the amount spent on advertising. From her district, consisting of several hundred stores, she selects a random sample of 30 stores. For each store she gathered the following information.

$Y$ = total sales last year (in $ thousands).
$X_1$ = number of competitors in the region.
$X_2$ = population of the region (in millions).
$X_3$ = advertising expense (in $ thousands).

The sample data were run on MINITAB, with the following results.

```
Analysis of variance

SOURCE        DF        SS            MS
Regression     3     3050.00      1016.67
Error         26     2200.00        84.62
Total         29     5250.00

Predictor     Coef      StDev    t-ratio
Constant     14.00       7.00       2.00
   X1        -1.00       0.70      -1.43
   X2        30.00       5.20       5.77
   X3         0.20       0.08       2.50
```

a. What are the estimated sales for the Bryne store, which has four competitors, a regional population of 0.4 (400,000), and advertising expense of 30 ($30,000)?
b. Compute the $R^2$ value.
c. Compute the multiple standard error of estimate.
d. Conduct a global test of hypothesis to determine whether any of the regression coefficients are not equal to zero. Use the .05 level of significance.
e. Conduct tests of hypotheses to determine which of the independent variables have significant regression coefficients. Which variables would you consider eliminating? Use the .05 significance level.

**18.** Suppose that the sales manager of a large automotive parts distributor wants to estimate as early as April the total annual sales of a region. On the basis of regional sales, the total sales for the company can also be estimated. If, based on past experience, it is found that the April estimates of annual sales are reasonably accurate, then in future years the April forecast could be used to revise production schedules and maintain the correct inventory at the retail outlets.

Several factors appear to be related to sales, including the number of retail outlets in the region stocking the company's parts, the number of automobiles in the region registered as of April 1, and the total personal income for the first quarter of the year. Five independent variables were finally selected as being the most important (according to the sales manager). Then the data were gathered for a recent year. The total annual sales for that year for each region were also recorded. Note in the following table that for region 1 there were 1,739 retail outlets stocking the company's automotive parts, there were 9,270,000 registered automobiles in the region as of April 1 and so on. The sales for that year were $37,702,000.

| Annual Sales ($ millions), $Y$ | Number of Retail Outlets, $X_1$ | Number of Automobiles Registered (millions), $X_2$ | Personal Income ($ billions), $X_3$ | Average Age of Automobiles (years), $X_4$ | Number of Supervisors, $X_5$ |
|---|---|---|---|---|---|
| 37.702 | 1,739 | 9.27 | 85.4 | 3.5 | 9.0 |
| 24.196 | 1,221 | 5.86 | 60.7 | 5.0 | 5.0 |
| 32.055 | 1,846 | 8.81 | 68.1 | 4.4 | 7.0 |
| 3.611 | 120 | 3.81 | 20.2 | 4.0 | 5.0 |
| 17.625 | 1,096 | 10.31 | 33.8 | 3.5 | 7.0 |
| 45.919 | 2,290 | 11.62 | 95.1 | 4.1 | 13.0 |
| 29.600 | 1,687 | 8.96 | 69.3 | 4.1 | 15.0 |
| 8.114 | 241 | 6.28 | 16.3 | 5.9 | 11.0 |
| 20.116 | 649 | 7.77 | 34.9 | 5.5 | 16.0 |
| 12.994 | 1,427 | 10.92 | 15.1 | 4.1 | 10.0 |

a. Consider the following correlation matrix. Which single variable has the strongest correlation with the dependent variable? The correlations between the independent variables outlets and income and between cars and outlets are fairly strong. Could this be a problem? What is this condition called?

```
             sales    outlets       cars     income        age
outlets      0.899
cars         0.605      0.775
income       0.964      0.825      0.409
age         -0.323     -0.489     -0.447     -0.349
bosses       0.286      0.183      0.395      0.155      0.291
```

b. The output for all five variables is on the following page. What percent of the variation is explained by the regression equation?

```
The regression equation is
sales = -19.7 - 0.00063 outlets + 1.74 cars + 0.410 income
       + 2.04 age - 0.034 bosses

            Predictor        Coef          StDev        t-ratio
            Constant      -19.672          5.422          -3.63
            outlets     -0.000629       0.002638          -0.24
            cars           1.7399         0.5530           3.15
            income        0.40994        0.04385           9.35
            age            2.0357         0.8779           2.32
            bosses        -0.0344         0.1880          -0.18

Analysis of Variance
            SOURCE          DF         SS             MS
            Regression       5    1593.81         318.76
            Error            4       9.08           2.27
            Total            9    1602.89
```

**c.** Conduct a global test of hypothesis to determine whether any of the regression coefficients are not zero. Use the .05 significance level.

**d.** Conduct a test of hypothesis on each of the independent variables. Would you consider eliminating "outlets" and "bosses"? Use the .05 significance level.

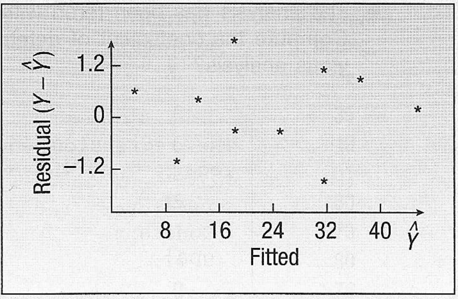
**e.** The regression has been rerun below with "outlets" and "bosses" eliminated. Compute the coefficient of determination. How much has $R^2$ changed from the previous analysis?

```
The regression equation is
sales = -18.9 + 1.61 cars + 0.400 income + 1.96 age

            Predictor        Coef          StDev        t-ratio
            Constant      -18.924          3.636          -5.20
            cars           1.6129         0.1979           8.15
            income        0.40031        0.01569          25.52
            age            1.9637         0.5846           3.36
Analysis of Variance
            SOURCE          DF         SS             MS
            Regression       3    1593.66         531.22
            Error            6       9.23           1.54
            Total            9    1602.89
```

**f.** Following is a histogram and a stem-and-leaf chart of the residuals. Does the normality assumption appear reasonable?

```
Histogram of residual N = 10        Stem-and-leaf of residual N = 10
                                    Leaf Unit = 0.10
Midpoint  Count
   -1.5      1    *                 1  -1   7
   -1.0      1    *                 2  -1   2
   -0.5      2    **                2  -0
   -0.0      2    **                5  -0   440
    0.5      2    **                5   0   24
    1.0      1    *                 3   0   68
    1.5      1    *                 1   1
                                    1   1   7
```

**g.** Following is a plot of the fitted values of $Y$ (i.e., $\hat{Y}$) and the residuals. Do you see any violations of the assumptions?



**19.** The administrator of a new paralegal program at Seagate Technical College wants to estimate the grade point average in the new program. He thought that high school GPA, the verbal score on the Scholastic Aptitude Test (SAT), and the mathematics score on the SAT would be good predictors of paralegal GPA. The data on nine students are:

| Student | High School GPA | SAT Verbal | SAT Math | Paralegal GPA |
|---|---|---|---|---|
| 1 | 3.25 | 480 | 410 | 3.21 |
| 2 | 1.80 | 290 | 270 | 1.68 |
| 3 | 2.89 | 420 | 410 | 3.58 |
| 4 | 3.81 | 500 | 600 | 3.92 |
| 5 | 3.13 | 500 | 490 | 3.00 |
| 6 | 2.81 | 430 | 460 | 2.82 |
| 7 | 2.20 | 320 | 490 | 1.65 |
| 8 | 2.14 | 530 | 480 | 2.30 |
| 9 | 2.63 | 469 | 440 | 2.33 |

**a.** Consider the following correlation matrix. Which variable has the strongest correlation with the dependent variable? Some of the correlations among the independent variables are strong. Does this appear to be a problem?

```
              legal       gpa      verbal
gpa           0.911
verbal        0.616      0.609
math          0.487      0.636      0.599
```

**b.** Consider the following output. Compute the coefficient of multiple determination.

```
The regression equation is
legal  =  -0.411  +  1.20  gpa  +  0.00163  verbal  -  0.00194  math

Predictor            Coef               StDev         t-ratio
Constant          -0.4111              0.7823           -0.53
gpa                1.2014              0.2955            4.07
verbal           0.001629            0.002147            0.76
math            -0.001939            0.002074           -0.94

Analysis  of  Variance
SOURCE          DF          SS             MS
Regression       3      4.3595         1.4532
Error            5      0.7036         0.1407
Total            8      5.0631
```

**c.** Conduct a global test of hypothesis from the preceding output. Does it appear that any of the regression coefficients are not equal to zero?

**d.** Conduct a test of hypothesis on each independent variable. Would you consider eliminating the variables "verbal" and "math"? Let $\alpha = .05$.

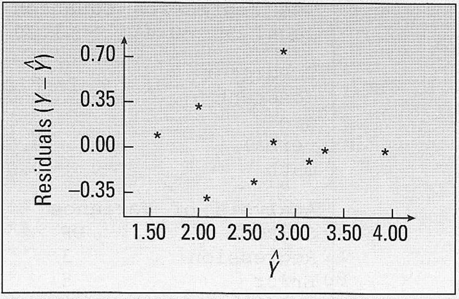
**e.** The analysis has been rerun without "verbal" and "math." See the following output. Compute the coefficient of determination. How much has $R^2$ changed from the previous analysis?

```
The regression equation is
legal = -0.454 + 1.16 gpa

 Predictor         Coef        StDev      t-ratio
 Constant       -0.4542       0.5542        -0.82
  gpa            1.1589       0.1977         5.86

Analysis of Variance
SOURCE          DF         SS          MS
Regression       1     4.2061      4.2061
Error            7     0.8570      0.1224
Total            8     5.0631
```

**f.** Following are a histogram and a stem-and-leaf diagram of the residuals. Does the normality assumption for the residuals seem reasonable?

```
Histogram of residual  N = 9

Midpoint          Count
    -0.4              1  *
    -0.2              3  ***
     0.0              3  ***
     0.2              1  *
     0.4              0
     0.6              1  *

Stem-and-leaf of residual  N = 9
Leaf unit = 0.10
  1    -0   4
  2    -0   2
 (3)   -0   110
  4     0   00
  2     0
  1     0
  1     0   6
```

**g.** Following is a plot of the residuals and the $\hat{Y}$ values. Do you see any violation of the assumptions?



**20.** Mike Wilde is president of the teachers' union for Otsego School District. In preparing for upcoming negotiations, he would like to investigate the salary structure of classroom teachers in the district. He believes there are three factors that affect a teacher's salary: years of experience, a rating of teaching effectiveness given by the principal, and whether the teacher has a master's degree. A random sample of 20 teachers resulted in the following data.

| Salary ($ thousands), $Y$ | Years of Experience, $X_1$ | Principal's Rating, $X_2$ | Master's Degree,* $X_3$ |
|---|---|---|---|
| 31.1 | 8 | 35 | 0 |
| 33.6 | 5 | 43 | 0 |
| 29.3 | 2 | 51 | 1 |
| 43.0 | 15 | 60 | 1 |
| 38.6 | 11 | 73 | 0 |
| 45.0 | 14 | 80 | 1 |
| 42.0 | 9 | 76 | 0 |
| 36.8 | 7 | 54 | 1 |
| 48.6 | 22 | 55 | 1 |
| 31.7 | 3 | 90 | 1 |
| 25.7 | 1 | 30 | 0 |
| 30.6 | 5 | 44 | 0 |
| 51.8 | 23 | 84 | 1 |
| 46.7 | 17 | 76 | 0 |
| 38.4 | 12 | 68 | 1 |
| 33.6 | 14 | 25 | 0 |
| 41.8 | 8 | 90 | 1 |
| 30.7 | 4 | 62 | 0 |
| 32.8 | 2 | 80 | 1 |
| 42.8 | 8 | 72 | 0 |

*1 = yes, 0 = no.

a. Develop a correlation matrix. Which independent variable has the strongest correlation with the dependent variable? Does it appear there will be any problems with multicollinearity?

b. Determine the regression equation. What salary would you estimate for a teacher with five years' experience, a rating by the principal of 60, and no master's degree?

c. Conduct a global test of hypothesis to determine whether any of the regression coefficients differ from zero. Use the .05 significance level.

d. Conduct a test of hypothesis for the individual regression coefficients. Would you consider deleting any of the independent variables? Use the .05 significance level.

e. If your conclusion in part (d) was to delete one or more independent variables, run the analysis again without those variables.

f. Determine the residuals for the equation of part (e). Use a stem-and-leaf chart or a histogram to verify that the distribution of the residuals is approximately normal.

g. Plot the residuals computed in part (f) in a scatter diagram with the residuals on the $Y$-axis and the $\hat{Y}$ values on the $X$-axis. Does the plot reveal any violations of the assumptions of regression?

21. The district sales manager for a major automobile manufacturer is studying car sales. Specifically, he would like to determine what factors affect the number of cars sold at a dealership. To investigate, he randomly selects 12 dealers. From these dealers he obtains the number of cars sold last month, the minutes of radio advertising purchased last month, the number of full-time salespeople employed in the dealership, and whether the dealer is located in the city. The information is as follows:

| Cars Sold Last Month, $Y$ | Advertising, $X_1$ | Sales Force, $X_2$ | City, $X_3$ | Cars Sold Last Month, $Y$ | Advertising, $X_1$ | Sales Force, $X_2$ | City, $X_3$ |
|---|---|---|---|---|---|---|---|
| 127 | 18 | 10 | Yes | 161 | 25 | 14 | Yes |
| 138 | 15 | 15 | No | 180 | 26 | 17 | Yes |
| 159 | 22 | 14 | Yes | 102 | 15 | 7 | No |
| 144 | 23 | 12 | Yes | 163 | 24 | 16 | Yes |
| 139 | 17 | 12 | No | 106 | 18 | 10 | No |
| 128 | 16 | 12 | Yes | 149 | 25 | 11 | Yes |

**a.** Develop a correlation matrix. Which independent variable has the strongest correlation with the dependent variable? Does it appear there will be any problems with multicollinearity?

**b.** Determine the regression equation. How many cars would you expect to be sold by a dealership employing 20 salespeople, purchasing 15 minutes of advertising, and located in a city?

**c.** Conduct a global test of hypothesis to determine whether any of the regression coefficients differ from zero. Let $\alpha = .05$.

**d.** Conduct a test of hypothesis for the individual regression coefficients. Would you consider deleting any of the independent variables? Let $\alpha = .05$.

**e.** If your conclusion in part (d) was to delete one or more independent variables, run the analysis again without those variables.

**f.** Determine the residuals for the equation of part (e). Use a stem-and-leaf chart or a histogram to verify that the distribution of the residuals is approximately normal.

**g.** Plot the residuals computed in part (f) in a scatter diagram with the residuals on the $Y$-axis and the $\hat{Y}$ values on the $X$-axis. Does the plot reveal any violations of the assumptions of regression?

**22.** Fran's Convenience Marts are located throughout metropolitan Erie, Pennsylvania. Fran, the owner, would like to expand into other communities in northwestern Pennsylvania and southwestern New York, such as Jamestown, Corry, Meadville, and Warren. To prepare her presentation to the local bank, she would like to better understand the factors that make a particular outlet profitable. She must do all the work herself, so she will not be able to study all her outlets. She selects a random sample of 15 marts and records the average daily sales ($Y$), the floor space (area), the number of parking spaces, and the median income of families in that ZIP code region for each. The sample information is reported below.

| Sampled Mart | Daily Sales | Store Area | Parking Spaces | Income ($ thousands) |
|---|---|---|---|---|
| 1 | $1,840 | 532 | 6 | 44 |
| 2 | 1,746 | 478 | 4 | 51 |
| 3 | 1,812 | 530 | 7 | 45 |
| 4 | 1,806 | 508 | 7 | 46 |
| 5 | 1,792 | 514 | 5 | 44 |
| 6 | 1,825 | 556 | 6 | 46 |
| 7 | 1,811 | 541 | 4 | 49 |
| 8 | 1,803 | 513 | 6 | 52 |
| 9 | 1,830 | 532 | 5 | 46 |
| 10 | 1,827 | 537 | 5 | 46 |
| 11 | 1,764 | 499 | 3 | 48 |
| 12 | 1,825 | 510 | 8 | 47 |
| 13 | 1,763 | 490 | 4 | 48 |
| 14 | 1,846 | 516 | 8 | 45 |
| 15 | 1,815 | 482 | 7 | 43 |

**a.** Determine the regression equation.

**b.** What is the value of $R^2$? Comment on the value.

**c.** Conduct a global hypothesis test to determine if any of the independent variables are different from zero.

**d.** Conduct individual hypothesis tests to determine if any of the independent variables can be dropped.

**e.** If variables are dropped, recompute the regression equation and $R^2$.

**23.** Great Plains Roofing and Siding Company, Inc., sells roofing and siding products to home repair retailers, such as Lowe's and Home Depot, and commercial contractors. The owner is interested in studying the effects of several variables on the value of shingles sold ($000). The marketing manager is arguing that the company should spend more money on advertising, while a market researcher suggests it should focus more on making its brand and product more distinct from its competitors.

The company has divided the United States into 26 marketing districts. In each district it collected information on the following variables: volume of sales (in thousands of dollars), advertising dollars (in thousands), number of active accounts, number of competing brands, and a rating of district potential.

| Sales (000s) | Advertising Dollars (000s) | Number of Accounts | Number of Competitors | Market Potential |
|---|---|---|---|---|
| 79.3 | 5.5 | 31 | 10 | 8 |
| 200.1 | 2.5 | 55 | 8 | 6 |
| 163.2 | 8.0 | 67 | 12 | 9 |
| 200.1 | 3.0 | 50 | 7 | 16 |
| 146.0 | 3.0 | 38 | 8 | 15 |
| 177.7 | 2.9 | 71 | 12 | 17 |
| 30.9 | 8.0 | 30 | 12 | 8 |
| 291.9 | 9.0 | 56 | 5 | 10 |
| 160.0 | 4.0 | 42 | 8 | 4 |
| 339.4 | 6.5 | 73 | 5 | 16 |
| 159.6 | 5.5 | 60 | 11 | 7 |
| 86.3 | 5.0 | 44 | 12 | 12 |
| 237.5 | 6.0 | 50 | 6 | 6 |
| 107.2 | 5.0 | 39 | 10 | 4 |
| 155.0 | 3.5 | 55 | 10 | 4 |
| 291.4 | 8.0 | 70 | 6 | 14 |
| 100.2 | 6.0 | 40 | 11 | 6 |
| 135.8 | 4.0 | 50 | 11 | 8 |
| 223.3 | 7.5 | 62 | 9 | 13 |
| 195.0 | 7.0 | 59 | 9 | 11 |
| 73.4 | 6.7 | 53 | 13 | 5 |
| 47.7 | 6.1 | 38 | 13 | 10 |
| 140.7 | 3.6 | 43 | 9 | 17 |
| 93.5 | 4.2 | 26 | 8 | 3 |
| 259.0 | 4.5 | 75 | 8 | 19 |
| 331.2 | 5.6 | 71 | 4 | 9 |

Conduct a multiple regression analysis to find the best predictors of sales.

**a.** Draw a scatter diagram comparing sales volume with each of the independent variables. Comment on the results.

**b.** Develop a correlation matrix. Do you see any problems? Does it appear there are any redundant independent variables?

**c.** Develop a regression equation. Conduct the global test. Can we conclude that some of the independent variables are useful in explaining the variation in the dependent variable?

**d.** Conduct a test of each of the independent variables. Are there any that should be dropped?

**e.** Refine the regression equation so the remaining variables are all significant.

**f.** Develop a histogram of the residuals and a normal probability plot. Are there any problems?

**g.** Determine the variance inflation factor for each of the independent variables. Are there any problems?

**24.** The *Times-Observer* is a daily newspaper in Metro City. Like many city newspapers, the *Times-Observer* is suffering through difficult financial times. The circulation manager is studying other papers in similar cities in the United States and Canada. She is particularly interested in what variables relate to the number of subscriptions to the paper. She is able to obtain the following sample information on 25 newspapers in similar cities. The following notation is used:

Sub = Number of subscriptions (in thousands).
Popul = The metropolitan population (in thousands).
Adv = The advertising budget of the paper (in $ hundreds).
Income = The median family income in the metropolitan area (in $ thousands).

| Paper | Sub | Popul | Adv | Income | Paper | Sub | Popul | Adv | Income |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 37.95 | 588.9 | 13.2 | 35.1 | 14 | 38.39 | 586.5 | 15.4 | 35.5 |
| 2 | 37.66 | 585.3 | 13.2 | 34.7 | 15 | 37.29 | 544.0 | 11.0 | 34.9 |
| 3 | 37.55 | 566.3 | 19.8 | 34.8 | 16 | 39.15 | 611.1 | 24.2 | 35.0 |
| 4 | 38.78 | 642.9 | 17.6 | 35.1 | 17 | 38.29 | 643.3 | 17.6 | 35.3 |
| 5 | 37.67 | 624.2 | 17.6 | 34.6 | 18 | 38.09 | 635.6 | 19.8 | 34.8 |
| 6 | 38.23 | 603.9 | 15.4 | 34.8 | 19 | 37.83 | 598.9 | 15.4 | 35.1 |
| 7 | 36.90 | 571.9 | 11.0 | 34.7 | 20 | 39.37 | 657.0 | 22.0 | 35.3 |
| 8 | 38.28 | 584.3 | 28.6 | 35.3 | 21 | 37.81 | 595.2 | 15.4 | 35.1 |
| 9 | 38.95 | 605.0 | 28.6 | 35.1 | 22 | 37.42 | 520.0 | 19.8 | 35.1 |
| 10 | 39.27 | 676.3 | 17.6 | 35.6 | 23 | 38.83 | 629.6 | 22.0 | 35.3 |
| 11 | 38.30 | 587.4 | 17.6 | 34.9 | 24 | 38.33 | 680.0 | 24.2 | 34.7 |
| 12 | 38.84 | 576.4 | 22.0 | 35.4 | 25 | 40.24 | 651.2 | 33.0 | 35.8 |
| 13 | 38.14 | 570.8 | 17.6 | 35.0 | | | | | |

**a.** Determine the regression equation.

**b.** Conduct a global test of hypothesis to determine whether any of the regression coefficients are not equal to zero.

**c.** Conduct a test for the individual coefficients. Would you consider deleting any coefficients?

**d.** Determine the residuals and plot them against the fitted values. Do you see any problems?

**e.** Develop a histogram of the residuals. Do you see any problems with the normality assumption?

**25.** How important is GPA in determining the starting salary of recent business school graduates? Does graduating from a business school increase the starting salary? The director of undergraduate studies at a major university wanted to study these questions. She gathered the following sample information on 15 graduates last spring to investigate these questions.

| Student | Salary | GPA | Business | Student | Salary | GPA | Business |
|---|---|---|---|---|---|---|---|
| 1 | $31.5 | 3.245 | 0 | 9 | $34.7 | 3.355 | 1 |
| 2 | 33.0 | 3.278 | 0 | 10 | 32.5 | 3.080 | 0 |
| 3 | 34.1 | 3.520 | 1 | 11 | 31.5 | 3.025 | 0 |
| 4 | 35.4 | 3.740 | 1 | 12 | 32.2 | 3.146 | 0 |
| 5 | 34.2 | 3.520 | 1 | 13 | 34.0 | 3.465 | 1 |
| 6 | 34.0 | 3.421 | 1 | 14 | 32.8 | 3.245 | 0 |
| 7 | 34.5 | 3.410 | 1 | 15 | 31.8 | 3.025 | 0 |
| 8 | 35.0 | 3.630 | 1 | | | | |

The salary is reported in $000, GPA on the traditional 4-point scale. A 1 indicates the student graduated from a school of business; a 0 indicates that the student graduated from one of the other schools.

**a.** Develop a correlation matrix. Do you see any problems with multicollinearity?

**b.** Determine the regression equation. Discuss the regression equation. How much does graduating from a college of business add to a starting salary? What starting salary would you estimate for a student with a GPA of 3.00 who graduated from a college of business?

**c.** What is the value of $R^2$? Can we conclude that this value is greater than 0?

**d.** Would you consider deleting either of the independent variables?

**e.** Plot the residuals in a histogram. Is there any problem with the normality assumption?

**f.** Plot the fitted values against the residuals. Does this plot indicate any problems with homoscedasticity?

**26.** A mortgage department of a large bank is studying its recent loans. Of particular interest is how such factors as the value of the home (in thousands of dollars), education level of the head of the household, age of the head of the household, current monthly mortgage payment (in dollars), and gender of the head of the household (male = 1, female = 0) relate to the family income. Are these variables effective predictors of the income of the household? A random sample of 25 recent loans is obtained.

| Income ($ thousands) | Value ($ thousands) | Years of Education | Age | Mortgage Payment | Gender |
|---|---|---|---|---|---|
| $40.3 | $190 | 14 | 53 | $230 | 1 |
| 39.6 | 121 | 15 | 49 | 370 | 1 |
| 40.8 | 161 | 14 | 44 | 397 | 1 |
| 40.3 | 161 | 14 | 39 | 181 | 1 |
| 40.0 | 179 | 14 | 53 | 378 | 0 |
| 38.1 | 99 | 14 | 46 | 304 | 0 |
| 40.4 | 114 | 15 | 42 | 285 | 1 |
| 40.7 | 202 | 14 | 49 | 551 | 0 |
| 40.8 | 184 | 13 | 37 | 370 | 0 |
| 37.1 | 90 | 14 | 43 | 135 | 0 |
| 39.9 | 181 | 14 | 48 | 332 | 1 |
| 40.4 | 143 | 15 | 54 | 217 | 1 |
| 38.0 | 132 | 14 | 44 | 490 | 0 |
| 39.0 | 127 | 14 | 37 | 220 | 0 |
| 39.5 | 153 | 14 | 50 | 270 | 1 |
| 40.6 | 145 | 14 | 50 | 279 | 1 |
| 40.3 | 174 | 15 | 52 | 329 | 1 |
| 40.1 | 177 | 15 | 47 | 274 | 0 |
| 41.7 | 188 | 15 | 49 | 433 | 1 |
| 40.1 | 153 | 15 | 53 | 333 | 1 |
| 40.6 | 150 | 16 | 58 | 148 | 0 |
| 40.4 | 173 | 13 | 42 | 390 | 1 |
| 40.9 | 163 | 14 | 46 | 142 | 1 |
| 40.1 | 150 | 15 | 50 | 343 | 0 |
| 38.5 | 139 | 14 | 45 | 373 | 0 |

**a.** Determine the regression equation.

**b.** What is the value of $R^2$? Comment on the value.

**c.** Conduct a global hypothesis test to determine whether any of the independent variables are different from zero.

**d.** Conduct individual hypothesis tests to determine whether any of the independent variables can be dropped.

**e.** If variables are dropped, recompute the regression equation and $R^2$.

**27.** Fred G. Hire is the manager of human resources at Crescent Tool and Die, Inc. As part of his yearly report to the CEO, he is required to present an analysis of the salaried employees. Because there are over 1,000 employees, he does not have the staff to gather information on each salaried employee, so he selects a random sample of 30. For each employee, he records monthly salary; service at Crescent, in months; gender (1 = male, 0 = female); and whether the employee has a technical or clerical job. Those working technical jobs are coded 1, and those who are clerical 0.

| Sampled Employee | Monthly Salary | Length of Service | Age | Gender | Job |
|---|---|---|---|---|---|
| 1 | $1,769 | 93 | 42 | 1 | 0 |
| 2 | 1,740 | 104 | 33 | 1 | 0 |
| 3 | 1,941 | 104 | 42 | 1 | 1 |
| 4 | 2,367 | 126 | 57 | 1 | 1 |
| 5 | 2,467 | 98 | 30 | 1 | 1 |
| 6 | 1,640 | 99 | 49 | 1 | 1 |
| 7 | 1,756 | 94 | 35 | 1 | 0 |
| 8 | 1,706 | 96 | 46 | 0 | 1 |
| 9 | 1,767 | 124 | 56 | 0 | 0 |
| 10 | 1,200 | 73 | 23 | 0 | 1 |

*continued*

| Sampled Employee | Monthly Salary | Length of Service | Age | Gender | Job |
|---|---|---|---|---|---|
| 11 | $1,706 | 110 | 67 | 0 | 1 |
| 12 | 1,985 | 90 | 36 | 0 | 1 |
| 13 | 1,555 | 104 | 53 | 0 | 0 |
| 14 | 1,749 | 81 | 29 | 0 | 0 |
| 15 | 2,056 | 106 | 45 | 1 | 0 |
| 16 | 1,729 | 113 | 55 | 0 | 1 |
| 17 | 2,186 | 129 | 46 | 1 | 1 |
| 18 | 1,858 | 97 | 39 | 0 | 1 |
| 19 | 1,819 | 101 | 43 | 1 | 1 |
| 20 | 1,350 | 91 | 35 | 1 | 1 |
| 21 | 2,030 | 100 | 40 | 1 | 0 |
| 22 | 2,550 | 123 | 59 | 1 | 0 |
| 23 | 1,544 | 88 | 30 | 0 | 0 |
| 24 | 1,766 | 117 | 60 | 1 | 1 |
| 25 | 1,937 | 107 | 45 | 1 | 1 |
| 26 | 1,691 | 105 | 32 | 0 | 1 |
| 27 | 1,623 | 86 | 33 | 0 | 0 |
| 28 | 1,791 | 131 | 56 | 0 | 1 |
| 29 | 2,001 | 95 | 30 | 1 | 1 |
| 30 | 1,874 | 98 | 47 | 1 | 0 |

**a.** Determine the regression equation, using salary as the dependent variable and the other four variables as independent variables.

**b.** What is the value of $R^2$? Comment on this value.

**c.** Conduct a global test of hypothesis to determine whether any of the independent variables are different from 0.

**d.** Conduct an individual test to determine whether any of the independent variables can be dropped.

**e.** Rerun the regression equation, using only the independent variables that are significant. How much more does a man earn per month than a woman? Does it make a difference whether the employee has a technical or a clerical job?

**28.** Many regions along the coast in North and South Carolina and Georgia have experienced rapid population growth over the last 10 years. It is expected that the growth will continue over the next 10 years. This has motivated many of the large grocery store chains building new stores in the region. The Kelley's Super Grocery Stores, Inc., chain is no exception. The director of planning for Kelley's Super Grocery Stores wants to study adding more stores in this region. He believes there are two main factors that indicate the amount families spend on groceries. The first is their income and the other is the number of people in the family. The director gathered the following sample information.

| Family | Food | Income | Size | Family | Food | Income | Size |
|---|---|---|---|---|---|---|---|
| 1 | $5.04 | $ 73.98 | 4 | 14 | $4.92 | $ 171.36 | 2 |
| 2 | 4.08 | 54.90 | 2 | 15 | 6.60 | 82.08 | 9 |
| 3 | 5.76 | 94.14 | 4 | 16 | 5.40 | 141.30 | 3 |
| 4 | 3.48 | 52.02 | 1 | 17 | 6.00 | 36.90 | 5 |
| 5 | 4.20 | 65.70 | 2 | 18 | 5.40 | 56.88 | 4 |
| 6 | 4.80 | 53.64 | 4 | 19 | 3.36 | 71.82 | 1 |
| 7 | 4.32 | 79.74 | 3 | 20 | 4.68 | 69.48 | 3 |
| 8 | 5.04 | 68.58 | 4 | 21 | 4.32 | 54.36 | 2 |
| 9 | 6.12 | 165.60 | 5 | 22 | 5.52 | 87.66 | 5 |
| 10 | 3.24 | 64.80 | 1 | 23 | 4.56 | 38.16 | 3 |
| 11 | 4.80 | 138.42 | 3 | 24 | 5.40 | 43.74 | 7 |
| 12 | 3.24 | 125.82 | 1 | 25 | 4.80 | 48.42 | 5 |
| 13 | 6.60 | 77.58 | 7 | | | | |

Food and income are reported in thousands of dollars per year, and the variable size refers to the number of people in the household.

**a.** Develop a correlation matrix. Do you see any problems with multicollinearity?
**b.** Determine the regression equation. Discuss the regression equation. How much does an additional family member add to the amount spent on food?
**c.** What is the value of $R^2$? Can we conclude that this value is greater than 0?
**d.** Would you consider deleting either of the independent variables?
**e.** Plot the residuals in a histogram. Is there any problem with the normality assumption?
**f.** Plot the fitted values against the residuals. Does this plot indicate any problems with homoscedasticity?

**29.** An investment advisor is studying the relationship between a common stock's price to earnings (P/E) ratio and factors that she thinks would influence it. She has the following data on the earnings per share (EPS) and the dividend percentage (Yield) for a sample of 20 stocks.

| Stock | P/E | EPS | Yield | Stock | P/E | EPS | Yield |
|---|---|---|---|---|---|---|---|
| 1 | 20.79 | $2.46 | 1.42 | 11 | 1.35 | $2.93 | 2.59 |
| 2 | 3.03 | 2.69 | 4.05 | 12 | 25.43 | 2.07 | 1.04 |
| 3 | 44.46 | −0.28 | 4.16 | 13 | 22.14 | 2.19 | 3.52 |
| 4 | 41.72 | −0.45 | 1.27 | 14 | 24.21 | −0.83 | 1.56 |
| 5 | 18.96 | 1.60 | 3.39 | 15 | 30.91 | 2.29 | 2.23 |
| 6 | 18.42 | 2.32 | 3.86 | 16 | 35.79 | 1.64 | 3.36 |
| 7 | 34.82 | 0.81 | 4.56 | 17 | 18.99 | 3.07 | 1.98 |
| 8 | 30.43 | 2.13 | 1.62 | 18 | 30.21 | 1.71 | 3.07 |
| 9 | 29.97 | 2.22 | 5.10 | 19 | 32.88 | 0.35 | 2.21 |
| 10 | 10.86 | 1.44 | 1.17 | 20 | 15.19 | 5.02 | 3.50 |

**a.** Develop a multiple linear regression with P/E as the dependent variable.
**b.** Are either of the two independent variables an effective predictor of P/E?
**c.** Interpret the regression coefficients.
**d.** Do any of these stocks look particularly undervalued?
**e.** Plot the residuals and check the normality assumption. Plot the fitted values against the residuals.
**f.** Does there appear to be any problems with homoscedasticity?
**g.** Develop a correlation matrix. Do any of the correlations indicate multicollinearity?

**30.** The Conch Café, located in Gulf Shores, Alabama, features casual lunches with a great view of the Gulf of Mexico. To accommodate the increase in business during the summer vacation season, Fuzzy Conch, the owner, hires a large number of servers as seasonal help. When he interviews a prospective server he would like to provide data on the amount a server can earn in tips. He believes that the amount of the bill and the number of diners are both related to the amount of the tip. He gathered the following sample information.

| Customer | Amount of Tip | Amount of Bill | Number of Diners | Customer | Amount of Tip | Amount of Bill | Number of Diners |
|---|---|---|---|---|---|---|---|
| 1 | $7.00 | $48.97 | 5 | 16 | $3.30 | $23.59 | 2 |
| 2 | 4.50 | 28.23 | 4 | 17 | 3.50 | 22.30 | 2 |
| 3 | 1.00 | 10.65 | 1 | 18 | 3.25 | 32.00 | 2 |
| 4 | 2.40 | 19.82 | 3 | 19 | 5.40 | 50.02 | 4 |
| 5 | 5.00 | 28.62 | 3 | 20 | 2.25 | 17.60 | 3 |
| 6 | 4.25 | 24.83 | 2 | 21 | 5.50 | 44.47 | 4 |
| 7 | 0.50 | 6.24 | 1 | 22 | 3.00 | 20.27 | 2 |
| 8 | 6.00 | 49.20 | 4 | 23 | 1.25 | 19.53 | 2 |
| 9 | 5.00 | 43.26 | 3 | 24 | 3.25 | 27.03 | 3 |
| 10 | 4.75 | 31.36 | 4 | 25 | 3.00 | 21.28 | 2 |
| 11 | 5.25 | 32.87 | 4 | 26 | 6.25 | 43.38 | 4 |
| 12 | 6.00 | 34.99 | 3 | 27 | 5.60 | 28.12 | 4 |
| 13 | 4.00 | 33.91 | 4 | 28 | 2.50 | 26.25 | 2 |
| 14 | 3.35 | 23.06 | 2 | 29 | 9.25 | 56.81 | 5 |
| 15 | 0.75 | 4.65 | 1 | 30 | 8.25 | 50.65 | 5 |

a. Develop a multiple regression equation with the amount of tips as the dependent variable and the amount of the bill and the number of diners as independent variables. Write out the regression equation. How much does another diner add to the amount of the tips?

b. Conduct a global test of hypothesis to determine if at least one of the independent variables is significant. What is your conclusion?

c. Conduct an individual test on each of the variables. Should one or the other be deleted?

d. Use the equation developed in part (c) to determine the coefficient of determination. Interpret the value.

e. Plot the residuals. Is it reasonable to assume they follow the normal distribution?

f. Plot the residuals against the fitted values. Is it reasonable to conclude they are random?

**31.** The president of Blitz Sales Enterprises sells kitchen products through television commercials, often called infomercials. He gathered data from the last 15 weeks of sales to determine the relationship between sales and the number of infomercials.

| Infomercials | Sales ($000s) | Infomercials | Sales ($000s) |
|---|---|---|---|
| 20 | 3.2 | 22 | 2.5 |
| 15 | 2.6 | 15 | 2.4 |
| 25 | 3.4 | 25 | 3.0 |
| 10 | 1.8 | 16 | 2.7 |
| 18 | 2.2 | 12 | 2.0 |
| 18 | 2.4 | 20 | 2.6 |
| 15 | 2.4 | 25 | 2.8 |
| 12 | 1.5 | | |

a. Determine the regression equation. Are the sales predictable from the number of commercials?

b. Determine the residuals and plot a histogram. Does the normality assumption seem reasonable?

**32.** The director of special events for Sun City believed that the amount of money spent on fireworks displays on the 4th of July was predictive of attendance at the Fall Festival held in October. She gathered the following data to test her suspicion.

| 4th of July ($000) | Fall Festival (000) | 4th of July ($000) | Fall Festival (000) |
|---|---|---|---|
| 10.6 | 8.8 | 9.0 | 9.5 |
| 8.5 | 6.4 | 10.0 | 9.8 |
| 12.5 | 10.8 | 7.5 | 6.6 |
| 9.0 | 10.2 | 10.0 | 10.1 |
| 5.5 | 6.0 | 6.0 | 6.1 |
| 12.0 | 11.1 | 12.0 | 11.3 |
| 8.0 | 7.5 | 10.5 | 8.8 |
| 7.5 | 8.4 | | |

Determine the regression equation. Is the amount spent on fireworks related to attendance at the Fall Festival? Conduct a hypothesis test to determine if there is a problem with autocorrelation.

**33.** You are a new hire at Laurel Woods Real Estate which specializes in selling foreclosed homes via public auction. Your boss has asked you to use the following data (mortgage balance, monthly payments, payments made before default, and final auction price) on a random sample of recent sales in order to estimate what the actual auction price will be.

| Loan | Monthly Payments | Payments Made | Auction Price | Loan | Monthly Payments | Payments Made | Auction Price |
|---|---|---|---|---|---|---|---|
| $ 85,600 | $ 985.87 | 1 | $16,900 | $105,200 | $ 915.24 | 34 | $52,600 |
| 115,300 | 902.56 | 33 | 75,800 | 105,900 | 905.67 | 38 | 51,900 |
| 103,100 | 736.28 | 6 | 43,900 | 94,700 | 810.70 | 25 | 43,200 |
| 84,600 | 945.45 | 9 | 16,600 | 105,600 | 891.33 | 20 | 52,600 |
| 97,600 | 821.07 | 24 | 40,700 | 104,100 | 864.38 | 7 | 42,700 |
| 104,400 | 983.27 | 26 | 63,100 | 85,700 | 1074.73 | 30 | 22,200 |
| 113,800 | 1075.54 | 19 | 72,600 | 113,600 | 871.61 | 24 | 77,000 |
| 116,400 | 1087.16 | 35 | 72,300 | 119,400 | 1021.23 | 58 | 69,000 |
| 100,000 | 900.01 | 33 | 58,100 | 90,600 | 836.46 | 3 | 35,600 |
| 92,800 | 683.11 | 36 | 37,100 | 104,500 | 1056.37 | 22 | 63,000 |

**a.** Carry out a global test of hypothesis to verify if any of the regression coefficients are different from zero.

**b.** Do an individual test of the independent variables. Would you remove any of the variables?

**c.** If it seems one or more of the independent variables is not needed, remove it and work out the revised regression equation.

**34.** Think about the figures from the previous exercise. Add a new variable that describes the potential interaction between the loan amount and the number of payments made. Then do a test of hypothesis to check if the interaction is significant.

## exercises.com

**35.** The National Institute of Standards and Technology provides several data sets to allow any user to test the accuracy of its statistical software. Go to the website: http://www.itl.nist.gov/div898/strd. Select the **Dataset Archives** section and, within that, the **Linear Regression** section. You will find the names of 11 small data sets stored in ASCII format on this page. Select one and run the data through your statistical software. Compare your results with the "official" results of the federal government.

**36.** As described in the exercises in Chapters 12 and 13, many real estate companies and rental agencies now publish their listings on the Web. One example is Dunes Realty Company, located in Garden City and Surfside Beaches in South Carolina. Go to the Web site http://www.dunes.com, select **Vacation Rentals,** then **Beach Home Search,** then indicate 5 bedrooms, accommodations for 14 people, oceanfront, and no pool or floating dock, select a period in July and August, indicate that you are willing to spend $10,000 per week, and then click on **Search the Beach Homes.** The output should include details on the cottages that meet your criteria. Develop a multiple linear regression equation using the rental price per week as the dependent variable and number of bedrooms, number of bathrooms, and how many people the cottage will accommodate as independent variables. Analyze the regression equations. Would you consider deleting any independent variables? What is the coefficient of determination? If you delete any of the variables, rerun the regression equation and discuss the new equation.

## Data Set Exercises

**37.** Refer to the Real Estate data, which report information on homes sold in the Denver, Colorado, area during the last year. Use the selling price of the home as the dependent variable and determine the regression equation with number of bedrooms, size of the house, whether there is a pool, whether there is an attached garage, distance from the center of the city, and number of bathrooms as independent variables.

**a.** Write out the regression equation. Discuss each of the variables. For example, are you surprised that the regression coefficient for distance from the center of the city is negative? How much does a garage or a swimming pool add to the selling price of a home?

- **b.** Determine the value of $R^2$. Interpret.
- **c.** Develop a correlation matrix. Which independent variables have strong or weak correlations with the dependent variable? Do you see any problems with multicollinearity?
- **d.** Conduct the global test on the set of independent variables. Interpret.
- **e.** Conduct a test of hypothesis on each of the independent variables. Would you consider deleting any of the variables? If so, which ones?
- **f.** Rerun the analysis until only significant regression coefficients remain in the analysis. Identify these variables.
- **g.** Develop a histogram or a stem-and-leaf display of the residuals from the final regression equation developed in part (f). Is it reasonable to conclude that the normality assumption has been met?
- **h.** Plot the residuals against the fitted values from the final regression equation developed in part (f) against the fitted values of $Y$. Plot the residuals on the vertical axis and the fitted values on the horizontal axis.

**38.** Refer to the Global Financial Performance data set that reports information on 148 corporations. Let the dependent variable be Return on Assets. Let the independent variables be Sales, Cost of Sales, Total Taxes, and Net Income.

- **a.** Develop a correlation matrix. Which independent variables have strong correlations with the dependent variable? Do you see any problems with multicollinearity? Explore how the cost of sales, total taxes, net income, and return on assets are calculated. Is multiple regression an appropriate application for this data set? Explain.

**39.** Refer to the Wage data, which report information on annual wages for a sample of 100 workers. Also included are variables relating to industry, years of education, and gender for each worker. Determine the regression equation using annual wage as the dependent variable and years of education, gender, years of work experience, age in years, and whether or not the worker is a union member.

- **a.** Write out the regression equation. Discuss each variable.
- **b.** Determine and interpret the $R^2$ value.
- **c.** Develop a correlation matrix. Which independent variables have strong or weak correlations with the dependent variable? Do you see any problems with multicollinearity?
- **d.** Conduct a global test of hypothesis on the set of independent variables. Interpret your findings. Is it reasonable to continue with the analysis or should you stop here?
- **e.** Conduct a test of hypothesis on each of the independent variables. Would you consider deleting any of these variables? If so, which ones?
- **f.** Rerun the analysis deleting any of the independent variables that are not significant. Delete the variables one at a time.
- **g.** Develop a histogram or a stem-and-leaf chart of the residuals from the final regression equation. Is it reasonable to conclude that the normality assumption has been met?
- **h.** Plot the residuals against the fitted values from the final regression equation. Plot the residuals on the vertical axis and the fitted values on the horizontal axis.

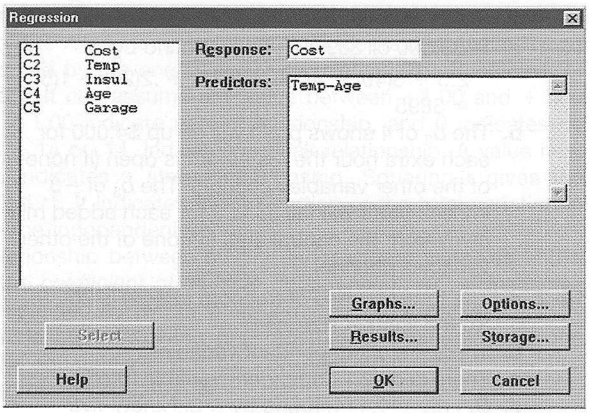
**40.** Refer to the CIA data which reports demographic and economic information on 62 countries. Let unemployment be the dependent variable and percent of the population over 65, life expectancy, and literacy be the independent variables.

- **a.** Determine the regression equation using a software package. Write out the regression equation.
- **b.** What is the value of the coefficient of determination? Interpret.
- **c.** Check the independent variables for multicollinearity. Describe your findings.
- **d.** Conduct a global regression analysis.
- **e.** Conduct a stepwise analysis to select the most appropriate variables for the regression model.
- **f.** Analyze the residuals by plotting the results versus the fitted values and using a normal probability plot. Are there any problems indicated by the graphs?

## Software Commands

*Note:* We do not show steps for all the statistical software used in this chapter. Below are the first two, which show the basic steps.

1. The MINITAB commands for the multiple regression output on page 515 are:
   **a.** Import the data from the CD. The file name is **Tbl14–1.**
   **b.** Select **Stat, Regression,** and then click on **Regression.**
   **c.** Select *Cost* as the **Response** variable, and *Temp, Insul,* and *Age* as the **Predictors,** then click on **OK.**



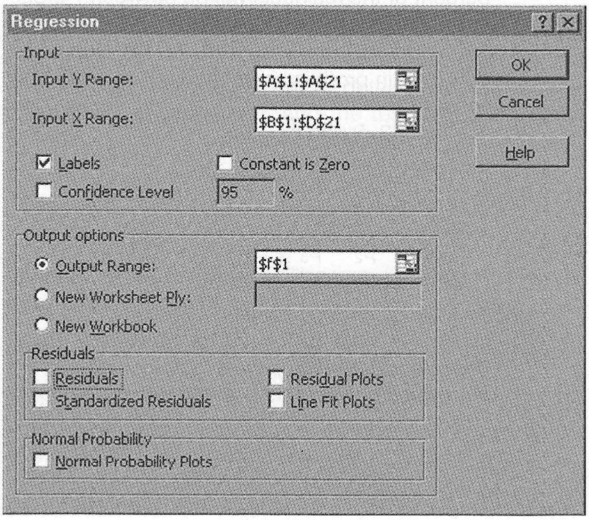
Regression

C1 Cost
C2 Temp
C3 Insul
C4 Age
C5 Garage

Response: Cost

Predictors: Temp-Age

2. The Excel commands to produce the multiple regression output on page 515 are:
   **a.** Import the data from the CD. The file name is **Tbl14.**
   **b.** Select **Tools,** then **Data Analysis,** highlight **Regression,** and click **OK.**
   **c.** Make the **Input Y Range** *A1:A21,* the **Input X Range** *B1:D21,* check the **Labels** box, the **Output Range** is *F1,* then click **OK.**



Regression

Input

Input Y Range: $A$1:$A$21

Input X Range: $B$1:$D$21

☑ Labels □ Constant is Zero

□ Confidence Level 95 %

Output options

● Output Range: $f$1

○ New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

# 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

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

**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.

**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.

**c.** We should consider dropping variables 1, 2 and 5. Variable 5 has the smallest absolute value of $t$. 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 $t > 2.110$

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

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 *coefficient of correlation.* The coefficient of correlation 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 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.

Multiple regression and correlation is concerned with relationship between two or more independent variables and the dependent 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.

The linear relationship in the simple case involving one independent variable and one dependent variable is described by the equation $\hat{Y} = a + bx$. For three independent variables, $X_1$, $X_2$, and $X_3$, the same multiple regression equation is

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

Computer invaluable in multiple regression and correlation.

Solving for $b_1, b_2, b_3, \cdots, b_k$ would involve tedious calculations. Fortunately, this type of problem can be quickly solved using one of the many statistical software packages and spreadsheet packages. Various measures, such as the coefficient of determination, the multiple standard error of estimate, the results of the global test, and the test of the individual variables, are reported in the output of most computer software programs.

## Glossary

### Chapter 13

**Coefficient of correlation** A measure of the strength of association between two variables.

**Coefficient of determination** The proportion of the total variation in the dependent variable that is explained by the independent variable. It can assume any value between 0 and +1.00 inclusive. A coefficient of .82 indicates that 82 percent of the variation in *Y* is accounted for by *X.* This coefficient is computed by squaring the coefficient of correlation, *r.*

**Correlation analysis** A group of statistical techniques used to measure the strength of the relationship between two variables.

**Covariance** The variance of *X* and *Y* together.

**Dependent variable** The variable that is being predicted or estimated.

**Independent variable** A variable that provides the basis for estimation.

**Least squares method** A technique used to arrive at the regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y* values and the predicted *Y* values.

**Linear regression equation** A mathematical equation that defines the relationship between two variables. It has the form $\hat{Y} = a + bX$. It is used to predict *Y* based on a selected *X* value. *Y* is the dependent variable and *X* the independent variable.

**Scatter diagram** A chart that visually depicts the relationship between two variables.

**Standard error of estimate** Measures the dispersion of the actual *Y* values about the regression line. It is reported in the same units as the dependent variable.

***t* test of significance of *r*** A formula to answer the question: Is the correlation in the population from which the sample was selected zero? The test statistic is *t,* and the number of degrees of freedom is $n - 2$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \textbf{[13–2]}$$

### Chapter 14

**Autocorrelation** Correlation of successive residuals. This condition frequently occurs when time is involved in the analysis.

**Correlation matrix** A listing of all possible simple coefficients of correlation. A correlation matrix includes the correlations between each of the independent variables and the dependent variable, as well as those among all the independent variables.

**Dummy variable** A qualitative variable. It can assume only one of two possible outcomes.

**Global test** A test used to determine if any of the set of independent variables have regression coefficients different from zero.

**Homoscedasticity** The standard error of estimate is the same for all fitted values of the dependent variable.

**Individual test** A test to determine if a particular independent variable has a regression coefficient different from zero.

**Interaction** The case in which one independent variable (such as $X_2$) affects the relationship between another independent variable ($X_1$) and the dependent variable (*Y*).

**Multicollinearity** A condition that occurs in multiple regression analysis if the independent variables are themselves correlated.

**Multiple regression equation** The relationship in the form of a mathematical equation between several independent variables and a dependent variable. The general form is $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_K$. It is used to estimate *Y* given *h* independent variables, $X_i$.

**Qualitative variables** A nominal-scale variable that is coded to assume only one of two possible outcomes. For example, a person is considered either employed or unemployed.

**Residual** The difference between the actual value of the dependent variable and the estimated value of the dependent variable, that is, $Y - \hat{Y}$.

**Stepwise regression** A step-by-step process for finding the regression equation. Only independent variables with nonzero regression coefficients enter the regression equation. Independent variables are added one at a time to the regression equation.

**Variance inflation factor** A test used to detect correlation among independent variables.

## Exercises

### Part I—Multiple Choice

1. The strength of the association between a set of independent variables $X$ and a dependent variable $Y$ is measured by the
   - a. Coefficient of correlation.
   - b. Coefficient of determination.
   - c. Standard error of estimate.
   - d. All of the above.
2. The percent of total variation of the dependent variable $Y$ explained by the set of independent variables $X$ is measured by the
   - a. Coefficient of correlation.
   - b. Coefficient of determination.
   - c. Standard error of estimate.
   - d. Multicollinearity.
3. A coefficient of correlation is computed to be $-0.90$. This result means:
   - a. The relationship between two variables is weak.
   - b. The relationship between two variables is strong and positive.
   - c. The relationship between two variables is strong and negative.
   - d. The relationship between four variables is strong.
4. The coefficient of determination was computed to be .38 in a problem involving one independent variable and one dependent variable. This result means
   - a. The relationship between the two variables is negative.
   - b. The correlation coefficient is also .38.
   - c. 38 percent of the total variation is explained by the independent variable.
   - d. 38 percent of the total variation is explained by the dependent variable.
5. What is the relationship between the coefficient of correlation and the coefficient of determination?
   - a. They are unrelated.
   - b. The coefficient of determination is the coefficient of correlation squared.
   - c. The coefficient of determination is the square root of the coefficient of correlation.
   - d. They are equal.
6. Multicollinearity exists when
   - a. Independent variables are correlated less than $-0.70$ or more than 0.70.
   - b. An independent variable is strongly associated with a dependent variable.
   - c. There is only one independent variable.
   - d. The relationship between the dependent and independent variables is nonlinear.
7. If "time" is used as the independent variable in a simple linear regression analysis, which of the following assumptions could be violated?
   - a. There is a linear relationship between the independent and dependent variables.
   - b. The residual variation is the same for all fitted values of $Y$.
   - c. The residuals are normally distributed.
   - d. Successive observations of the dependent variable are uncorrelated.
8. In multiple regression, when the global test of significance is rejected, we can conclude:
   - a. All of the net sample regression coefficients are equal to zero.
   - b. All of the sample regression coefficients are not equal to zero.
   - c. At least one sample regression coefficient is not equal to zero.
   - d. The regression equation intersects the $Y$-axis at zero.
9. A residual is defined as:
   - a. $Y - \hat{Y}$.
   - b. Error sum of squares.
   - c. Regression sum of squares.
   - d. Type I error.
10. What test statistic is used for a global test of significance?
    - a. $z$ statistic.
    - b. $t$ statistic.

**c.** Chi-square statistic.
**d.** *F* statistic.

## Part II—Problems

**11.** The accounting department at Crate and Barrel wishes to estimate the profit for each of the chain's many stores on the basis of the number of employees in the store, overhead costs, average markup, and theft loss. A few statistics from the stores are:

| Store | Net Profit ($ thousands) | Number of Employees | Overhead Cost ($ thousands) | Average Markup (percent) | Theft Loss ($ thousands) |
|---|---|---|---|---|---|
| 1 | $846 | 143 | $79 | 69% | $52 |
| 2 | 513 | 110 | 64 | 50 | 45 |

**a.** The dependent variable is _____.
**b.** The general equation for this problem is _____.
**c.** The multiple regression equation was computed to be $\hat{Y} = 67 + 8X_1 - 10X_2 + 0.004X_3 - 3X_4$. What are predicted sales for a store with 112 employees, an overhead cost of $65,000, a markup rate of 50 percent, and a loss from theft of $50,000?
**d.** Suppose $R^2$ was computed to be .86. Explain.
**e.** Suppose that the multiple standard error of estimate was 3 (in $ thousands). Explain what this means in this problem.

**12.** Quick-print firms in a large downtown business area spend most of their advertising dollars on displays on bus benches. A research project involves predicting monthly sales based on the annual amount spent on placing ads on bus benches. A sample of quick-print firms revealed these advertising expenses and sales:

| Firm | Annual Bus Bench Advertising ($ thousands) | Monthly Sales ($ thousands) |
|---|---|---|
| A | 2 | 10 |
| B | 4 | 40 |
| C | 5 | 30 |
| D | 7 | 50 |
| E | 3 | 20 |

**a.** Draw a scatter diagram.
**b.** Determine the coefficient of correlation.
**c.** What is the coefficient of determination?
**d.** Compute the regression equation.
**e.** Estimate the monthly sales of a quick-print firm that spends $4,500 on bus bench advertisements.
**f.** Summarize your findings.

**13.** The following ANOVA output is given.

```
SOURCE        Sum of Squares       DF          MS
Regression        1050.8            4      262.70
Error               83.8           20        4.19
Total             1134.6           24

  Predictor          Coef     St.Dev.     t-ratio
Constant            70.06       2.13        32.89
X1                   0.42       0.17         2.47
X2                   0.27       0.21         1.29
X3                   0.75       0.30         2.50
X4                   0.42       0.07         6.00
```

**a.** Compute the coefficient of determination.
**b.** Compute the multiple standard error of estimate.
**c.** Conduct a test of hypothesis to determine whether any of the regression coefficients are different from zero.
**d.** Conduct a test of hypothesis on the individual regression coefficients. Can any of the variables be deleted?

# Cases

## A. The Century National Bank

Refer to the Century National Bank data. Using checking account balance as the dependent variable and using as independent variables the number of ATM transactions, the number of other services used, whether the individual has a debit card, and whether interest is paid on the particular account, write a report indicating which of the variables seem related to the account balance and how well they explain the variation in account balances. Should all of the independent variables proposed be used in the analysis or can some be dropped?

## B. Terry and Associates: The Time to Deliver Medical Kits

Terry and Associates is a specialized medical testing center in Denver, Colorado. One of the firm's major sources of revenue is a kit used to test for elevated amounts of lead in the blood. Workers in auto body shops, those in the lawn care industry, and commercial house painters are exposed to large amounts of lead and thus must be randomly tested. It is expensive to conduct the test, so the kits are delivered on demand to a variety of locations throughout the Denver area.

Kathleen Terry, the owner, is concerned about setting appropriate costs for each delivery. To investigate, Ms. Terry gathered information on a random sample of 50 recent deliveries. Factors thought to be related to the cost of delivering a kit were:

Prep The time in minutes between when the customized order is phoned into the company and when it is ready for delivery.

Delivery The actual travel time in minutes from Terry's plant to the customer.

Mileage The distance in miles from Terry's plant to the customer.

| Sample Number | Cost | Prep | Delivery | Mileage |
|---|---|---|---|---|
| 1 | $32.60 | 10 | 51 | 20 |
| 2 | 23.37 | 11 | 33 | 12 |
| 3 | 31.49 | 6 | 47 | 19 |
| 4 | 19.31 | 9 | 18 | 8 |
| 5 | 28.35 | 8 | 88 | 17 |
| 6 | 22.63 | 9 | 20 | 11 |
| 7 | 22.63 | 9 | 39 | 11 |
| 8 | 21.53 | 10 | 23 | 10 |
| 9 | 21.16 | 13 | 20 | 8 |
| 10 | 21.53 | 10 | 32 | 10 |
| 11 | 28.17 | 5 | 35 | 16 |
| 12 | 20.42 | 7 | 23 | 9 |
| 13 | 21.53 | 9 | 21 | 10 |
| 14 | 27.55 | 7 | 37 | 16 |
| 15 | 23.37 | 9 | 25 | 12 |
| 16 | 17.10 | 15 | 15 | 6 |
| 17 | 27.06 | 13 | 34 | 15 |

| Sample Number | Cost | Prep | Delivery | Mileage |
|---|---|---|---|---|
| 18 | $15.99 | 8 | 13 | 4 |
| 19 | 17.96 | 12 | 12 | 4 |
| 20 | 25.22 | 6 | 41 | 14 |
| 21 | 24.29 | 3 | 28 | 13 |
| 22 | 22.76 | 4 | 26 | 10 |
| 23 | 28.17 | 9 | 54 | 16 |
| 24 | 19.68 | 7 | 18 | 8 |
| 25 | 25.15 | 6 | 50 | 13 |
| 26 | 20.36 | 9 | 19 | 7 |
| 27 | 21.16 | 3 | 19 | 8 |
| 28 | 25.95 | 10 | 45 | 14 |
| 29 | 18.76 | 12 | 12 | 5 |
| 30 | 18.76 | 8 | 16 | 5 |
| 31 | 24.29 | 7 | 35 | 13 |
| 32 | 19.56 | 2 | 12 | 6 |
| 33 | 22.63 | 8 | 30 | 11 |
| 34 | 21.16 | 5 | 13 | 8 |
| 35 | 21.16 | 11 | 20 | 8 |
| 36 | 19.68 | 5 | 19 | 8 |
| 37 | 18.76 | 5 | 14 | 7 |
| 38 | 17.96 | 5 | 11 | 4 |
| 39 | 23.37 | 10 | 25 | 12 |
| 40 | 25.22 | 6 | 32 | 14 |
| 41 | 27.06 | 8 | 44 | 16 |
| 42 | 21.96 | 9 | 28 | 9 |
| 43 | 22.63 | 8 | 31 | 11 |
| 44 | 19.68 | 7 | 19 | 8 |
| 45 | 22.76 | 8 | 28 | 10 |
| 46 | 21.96 | 13 | 18 | 9 |
| 47 | 25.95 | 10 | 32 | 14 |
| 48 | 26.14 | 8 | 44 | 15 |
| 49 | 24.29 | 8 | 34 | 13 |
| 50 | 24.35 | 3 | 33 | 12 |

1. Develop a multiple linear regression equation that describes the relationship between the cost of delivery and the other variables. Do these three variables explain a reasonable amount of the variation in the dependent variable? Estimate the delivery cost for a kit that takes 10 minutes for preparation, takes 30 minutes to deliver, and must cover a distance of 14 miles.
2. Test to determine that at least one net regression coefficient differs from zero. Also test to see whether any of the variables can be dropped from the analysis. If some of the variables can be dropped, rerun the regression equation until only significant variables are included. Write a brief report interpreting the final regression equation.

15

# Index Numbers

## GOALS

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

1 Describe the term *index.*

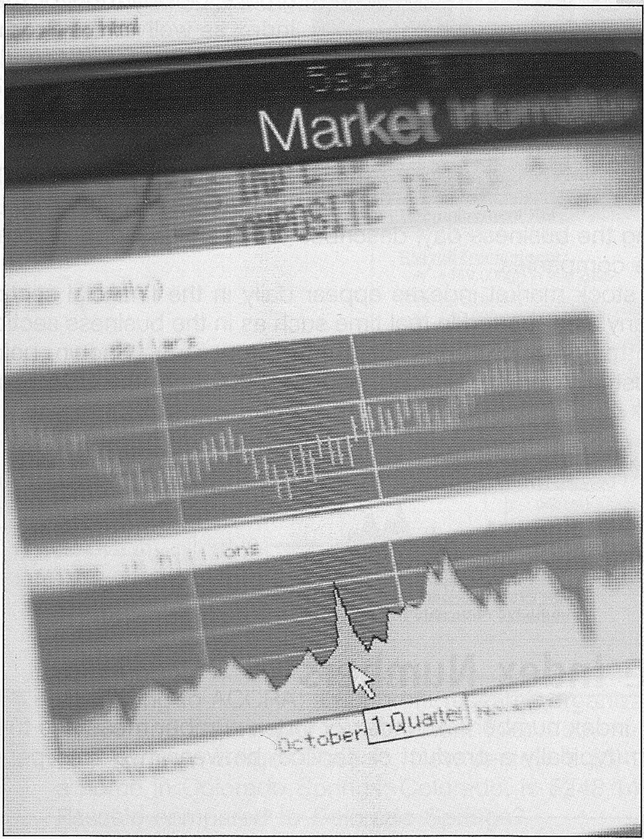
2 Understand the difference between a weighted and an unweighted index.

3 Construct and interpret a *Laspeyres price index.*

4 Construct and interpret a *Paasche price index.*

5 Construct and interpret a *value index.*

6 Explain how the Consumer Price Index is constructed and interpreted.



Information for food items for the years 2000 and 2006 are provided in Exercise 27. Compute a simple price index for each of the four items, using 2000 as the base period. (See Exercise 27 and Goal 2.)